UNIVERESITY OF CALIFORNIA, SAN DIEGO

Connecting the dots between DNA, proteins, and disease: Identifying genetic variants and proteins relevant for studying venous thromboembolism

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biomedical Science

by

Terry Solomon

Committee in Charge:

Professor Kelly Frazer, Chair
Professor Mohit Jain
Professor Abraham Palmer
Professor Bing Ren
Professor Sanford Shattil

2018

The Dissertation of Terry Solomon is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2018

DEDICATION

I dedicate this to my darling, Nicolas Mar,

whose support and effort during my doctorate was invaluable.

Without him, this would not have been possible.

Additionally, I dedicate this to my roommates, friends, and family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTAL FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Kelly Frazer for mentoring me during my time in her laboratory. I would like to thank all of the Frazer laboratory members and my collaborators for their help and support.

Chapter 2, in full, is a reprint of material as it appears in Circulation: Cardiovascular Genetics, 2016, Terry Solomon, Erin Smith, Hiroko Matsui, Sigrid Braekkan, Tom Wilsgaard, Inger Njølstad, Ellisiv Mathiesen, John-Bjarne Hansen, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication. Terry Solomon, John Lapek, Søren Beck Jensen, Hiroko Matsui, Kristian Hindberg, William Greenwald, Nadezhda Latysheva, Sigrid Braekkan, David Gonzalez, Kelly A. Frazer, Erin Smith, John-Bjarne Hansen. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, is currently being prepared for submission for publication. Søren Beck Jensen, Kristian Hindberg, Terry Solomon, Erin Smith, John Lapek, David Gonzalez, Nadezhda Latysheva, Kelly A. Frazer, Sigrid Braekkan, John-Bjarne Hansen. The dissertation author worked on the proteomics dataset and was a co-author of this paper.

VITA

2013 Bachelor of Science, Rochester Institute of Technology

2018 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Solomon T, Lapek J, Jensen SB, Greenwald WW, Hindberg K, Matsui H, Latysheva N, Braekkan SK, Gonzalez DJ, Frazer KA, Smith EN, Hansen JB. “Identifying putative functional variants that underlie plasma protein quantitative trait loci”. (Submitted)

Jensen SB, Hindberg K, **Solomon T**, Smith EN, Lapek J, Gonzalez DJ, Latysheva N, Frazer KA, Braekkan SK, Hansen JB. “Discovery of novel plasma biomarkers for future incident venous thromboembolism by untargeted SPS-MS3 proteomics”. (Submitted)

Horvei LD, Braekkan SK, Smith EN, **Solomon T**, Hindberg K, Frazer KA, Rosendaal FR, Hansen JB. Joint effects of prothrombotic genotypes and body height on the risk of venous thromboembolism: The Tromsø study. Journal of Thrombosis and Haemostasis 2017, 16(1).

Solomon T, Smith EN, Matsui H, Braekkan SK, Wilsgaard T, Njølstad I, Mathiesen EB, Hansen JB, Frazer KA, INVENT Consortium. Associations Between Common and Rare Exonic Genetic Variants and Serum Levels of Twenty Cardiovascular-Related Proteins: The Tromsø Study. Circulation: Cardiovascular Genetics 2016.

Gran OV, Smith EN, Braekkan SK, Jensvoll H, **Solomon T**, Hindberg K, Wilsgaard T, Rosendaal FR, Frazer KA, Hansen JB. Joint effects of cancer and variants in the Factor 5 gene on the risk of venous thromboembolism. Haematologica 2016, 101(9).

Brandler WM, Antaki D, Gujral M, Noor A, Rosanio G, Chapman TR, Barrera DJ, Lin GN, Malhotra D, Watts AC, Wong LC, Estabillo JA, Gadomski TE, Hong O, Fuentes Fajardo KV, Bhandari A, Owen R, Baughn M, Yuan J, **Solomon T**, Moyzis AG, MAile MS, Sanders SJ, Reiner GE, Vaux KK, Strom CM, Zhang K, Muotri AR, Akshomoff N, Leal SM, Pierce K, Courchesne E, Iakoucheva LM, Corsello C, Sebat J. Frequency and complexity of de novo structural mutation in autism. American Journal of Human Genetics 2016, 98(4).

Hasenkamp N, **Solomon T**, Tautz D. Selective sweeps versus introgression - population genetic dynamics of the murine leukemia virus receptor *Xpr1* in wild populations of the house mouse (*Mus musculus*). BMC Evolutionary Biology 2015, 15.

FIELDS OF STUDY

Major Field: Biomedical Science

Genetics
Professor Kelly Frazer

ABSTRACT OF THE DISSERTATION

Connecting the dots between DNA, proteins, and disease:
Identifying genetic variants and proteins relevant for studying venous thromboembolism

by

Terry Solomon

Doctor of Philosophy in Biomedical Science

University of California, San Diego, 2018

Professor Kelly Frazer, Chair

In order to prospectively identify individuals at risk for disease, it is important to identify markers that can be reliably measured and to understand the relation of these markers to the disease. For venous thromboembolism (VTE), large-scale genetic studies have had limited success identifying genetic variants and proteins

that contribute to disease risk. An alternative method is to focus on the intermediate steps between genetics and disease, such as protein levels. In this dissertation, we set out to identify genetic variants that contribute to blood protein levels and then identify blood proteins that can be used as biomarkers for venous thromboembolism. For the whole dissertation, we utilize the Tromsø Study, a single-center, prospective, study of the inhabitants of Tromsø, Norway. These individuals allow us to identify genetic variants and proteins that are associated with VTE before any symptoms of the disease start, which is key when trying to pre-emptively identify and treat individuals that are at risk for developing VTE. In chapter two, we measure cardiovascular-relevant serum proteins using enzyme-linked immunosorbent assays. We then identify common and rare genetic variation that is associated with the levels of these proteins. In chapter three, we measure the plasma proteome using tandem-mass-tagged mass spectrometry. We then identify common and rare genetic variation that is associated with the levels of these proteins. We further investigate the underlying mechanisms of how genetic variation regulates levels of plasma proteins. In chapter four, we utilize the same plasma proteome that was measured in chapter three in order to identify proteins that are associated with risk of venous thromboembolism. Together, this work advances our understanding of how genetic variants ultimately result in diseases, via their effects on intermediate protein levels.

Chapter 1: Introduction

Chapter 1.1: Background and Introduction

Venous thromboembolism (VTE) is comprised of deep vein thrombosis, where blood clots form in the deep veins, and pulmonary embolism, where these clots travel to the lungs, obstruct blood flow, and can result in death. VTE has an annual incidence rate of 1-2 per 1000 persons. Treatment involves prescribing blood thinners to prevent clotting, although this can lead to complications such as internal bleeding. Therefore, it is highly important to stratify individuals based on their molecular risk for VTE in order to decrease the economic, health, and morbidity burden that this disease has on society.

Often, stratifying high-risk patients is done by identifying protein relevant for the disease. In venous thromboembolism, the key proteins known to be involved are part of the coagulation cascade, a step-wise enzymatic procedure that results in the formation of a fibrin clot (Figure 1.1). The enzymes, cofactors, and activators of this pathway all circulate in the bloodstream inactive until an injury is sensed and the entire cascade is set into motion. Beyond the proteins involved in coagulation, there are only a few other proteins known to be involved in the etiology of VTE, such as GDF-15 and D-Dimer that have been identified as predictive biomarkers. Previous efforts to identify biomarker have been limited by the number of proteins tested and the limited number of studies that prospectively study VTE.

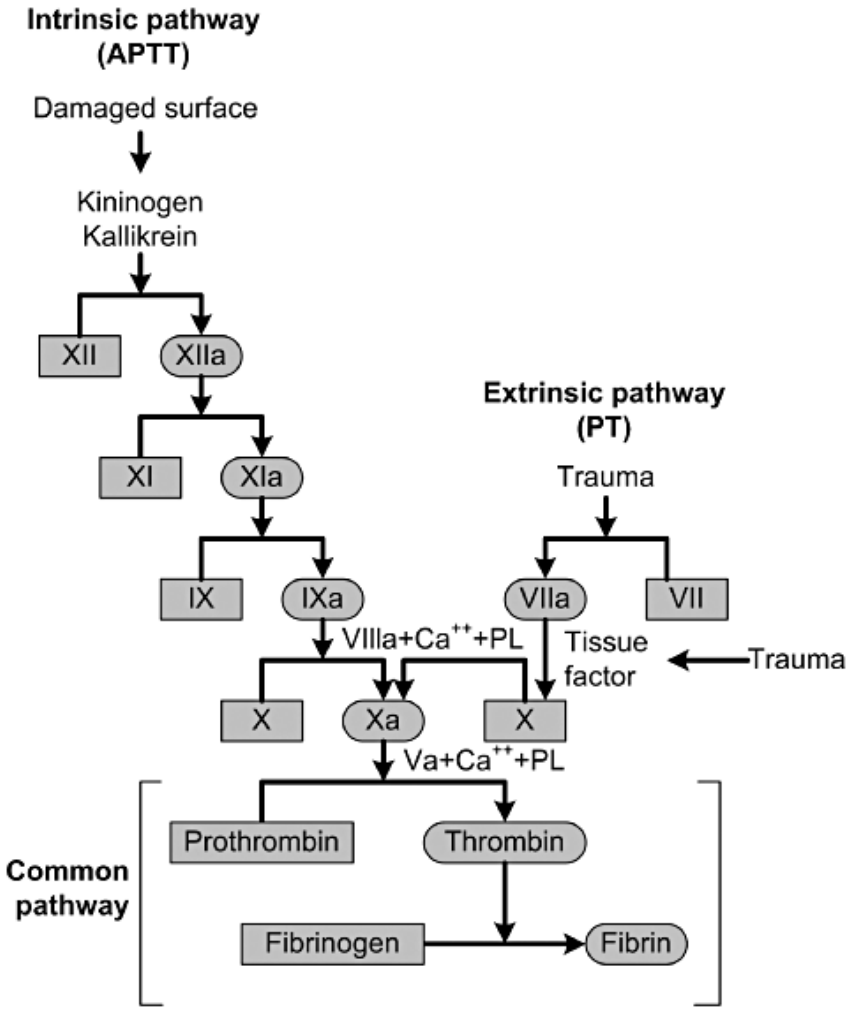


Figure 1.1 Coagulation Cascade from (Adams & Bird, 2009)

An alternative to regularly drawing blood in order to identify individuals with a high-risk for VTE is to identify genetic risk factors. This would enable patients to get genotyped for predictive variants once, and then based on their results, be stratified into high-, medium-, or low-risk categories that determine biomarker measurement and treatment schedules. Traditionally, large-scale studies such as genome-wide association studies (GWAS) have been performed to identify any genetic variants that are associated with having VTE. To date, 23 genetic variants

that are located in 15 loci have been found to be associated with VTE risk (Tregouet & Morange, 2018). From the largest VTE GWAS to date (7,500 cases and 50,000 controls), Germain et al. identified 8 of these loci that had common genetic variants associated with VTE (Germain et al., 2015). It is thought that it would take vastly larger studies and variants with large effect sizes to be able to find associations using a GWAS approach (Tregouet et al., 2016). Thus, researchers should complement GWAS studies with other study types to identify risks for VTE.

One method that requires fewer samples is to study an intermediate phenotype instead of the final phenotype of venous thromboembolism. This technique enables researchers to study the genetic regulation of the underlying mechanisms of the disease, such as gene expression, protein levels, and biomarker or pathway-level impacts. Investigating these intermediate traits can be done utilizing quantitative trait loci (QTL) studies. There has been a flood of studies that focus on the effects of genetic variants on gene expression (expression quantitative trait loci – eQTLs) due to the relative ease of measuring the entire transcriptome. Recently, there have been a handful of studies that focus instead on blood protein levels, a step closer to disease than gene expression while still retaining the advantages of QTL studies. These type of studies have been limited by the relative difficulty in measuring a vast amount of proteins in a fiscally feasible manner.

The work of this dissertation focuses on the genetic regulation and disease contributions of serum and plasma proteins. In the second chapter, I investigate how genetic variation contributes to the serum levels of cardiovascular-related proteins.

In the third chapter, I investigate how genetic variation contributes to levels of the entire plasma proteome and the mechanisms by which these genetic variants are acting. In the fourth chapter, I identify plasma proteins that are biomarkers for venous thromboembolism risk. Together this work advances our understanding of how genetic variants ultimately result in diseases, via their effects on intermediate protein levels.

Chapter 1.2: References

Adams, R. L., & Bird, R. J. (2009). Review article: Coagulation cascade and therapeutics update: relevance to nephrology. Part 1: Overview of coagulation, thrombophilias and history of anticoagulants. *Nephrology (Carlton), 14*(5), 462-470. doi:10.1111/j.1440-1797.2009.01128.x

Germain, M., Chasman, D. I., de Haan, H., Tang, W., Lindstrom, S., Weng, L. C., de Andrade, M., de Visser, M. C., Wiggins, K. L., Suchon, P., Saut, N., Smadja, D. M., Le Gal, G., van Hylckama Vlieg, A., Di Narzo, A., Hao, K., Nelson, C. P., Rocanin-Arjo, A., Folkersen, L., Monajemi, R., Rose, L. M., Brody, J. A., Slagboom, E., Aissi, D., Gagnon, F., Deleuze, J. F., Deloukas, P., Tzourio, C., Dartigues, J. F., Berr, C., Taylor, K. D., Civelek, M., Eriksson, P., Psaty, B. M., Houwing-Duitermaat, J., Goodall, A. H., Cambien, F., Kraft, P., Amouyel, P., Samani, N. J., Basu, S., Ridker, P. M., Rosendaal, F. R., Kabrhel, C., Folsom, A. R., Heit, J., Reitsma, P. H., Tregouet, D. A., Smith, N. L., & Morange, P. E. (2015). Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet, 96*(4), 532-542. doi:10.1016/j.ajhg.2015.01.019

Tregouet, D. A., Delluc, A., Roche, A., Derbois, C., Olaso, R., Germain, M., de Andrade, M., Tang, W., Chasman, D. I., van Hylckama Vlieg, A., Reitsma, P. H., Kabrhel, C., Smith, N., & Morange, P. E. (2016). Is there still room for additional common susceptibility alleles for venous thromboembolism? *J Thromb Haemost, 14*(9), 1798-1802. doi:10.1111/jth.13392

Tregouet, D. A., & Morange, P. E. (2018). What is currently known about the genetics of venous thromboembolism at the dawn of next generation sequencing technologies. *Br J Haematol, 180*(3), 335-345. doi:10.1111/bjh.15004

Chapter 2: Associations between common and rare exonic genetic variants and serum levels of twenty cardiovascular-related proteins - the Tromsø Study

Chapter 2.1: Abstract

Background: Genetic variation can be used to study causal relationships between biomarkers and diseases. Here, we identify new common and rare genetic variants associated with cardiovascular-related protein levels (protein quantitative trait loci, pQTLs). We functionally annotate these pQTLs, predict and experimentally confirm a novel molecular interaction and determine which pQTLs are associated with diseases and physiological phenotypes.

Methods and results: As part of a larger case/control study of VTE, serum levels of 51 proteins implicated in cardiovascular diseases were measured in 330 individuals from the Tromsø Study. Exonic genetic variation near each protein's respective gene (cis) was identified using sequencing and arrays. Using single site and gene-based tests, we identified 27 genetic associations between pQTLs and the serum levels of 20 proteins: 14 associated with common variation in cis, of which six are novel (i.e. not previously reported); seven associations with rare variants in cis, of which four are novel; and six associations in trans. Of the 20 proteins, 15 were associated with single sites and seven with rare variants. cis-pQTLs for kallikrein and F12 also show trans associations for proteins (uPAR, kininogen) known to be cleaved by kallikrein as well as with NTproBNP. We experimentally demonstrate that kallikrein can cleave proBNP (NTproBNP precursor) in vitro. Nine of the pQTLs have previously identified associations with 17 diseases and/or physiological phenotypes.

Conclusions: We have identified cis and trans genetic variation associated with the serum levels of 20 proteins and utilized these pQTLs to study molecular mechanisms underlying diseases and/or physiological phenotypes.

Chapter 2.2 Clinical Perspective

Cardiovascular diseases, including coronary artery disease and venous thromboembolism, are the leading cause of death worldwide. Biomarkers are important tools to diagnose or measure risk of disease, but the causal relationship between biomarkers and diseases is often not clear. Genetic variants that affect levels of protein biomarkers could be used to examine causal relationships between biomarkers and diseases and to provide mechanistic insight into disease. In this study, we investigated whether genetic variants were associated with the levels of 51 serum proteins, 17 of which we had previously identified as predictors for myocardial infarction in the Tromsø Study. We analyzed genotype data from exome sequencing and exome arrays and investigated whether common and rare genetic variation located near the gene (cis) that coded for each protein was associated with protein levels. We identified 13 proteins associated with common cis variants and 7 proteins associated with rare cis variation; 8 of these proteins we had previously identified as biomarkers. To identify pathway-level regulation, we tested whether these significantly associated cis variants were also associated in trans with the levels of the other 50 proteins in this study. We identified that genetic variation affecting the levels of kallikrein, a protease involved in coagulation, also affect the levels of NTproBNP, a known biomarker for heart failure. We experimentally show that kallikrein can cleave proBNP into NTproBNP and BNP. Our study shows that identifying genetic

variants that affect protein levels can provide novel insights and expand our knowledge of the mechanisms of disease.

Chapter 2.3: Introduction

Recent advances in genetics have yielded an unprecedented number of loci associated with disease and are beginning to yield mechanistic insight, such as with the IRX3/5 association with BMI, which revealed brown adipose as an important regulator of body weight (Claussnitzer et al., 2015). Genetic variation underlying molecular phenotypes, such as proteins and transcript expression levels, can be important tools in constructing the effects of genetic variations into pathways, ultimately resulting in physiological understanding of diseases (Schadt, 2009). Protein levels in particular may be more informative for understanding disease because there is often a poor correlation between transcript and protein levels (Anderson & Seilhamer, 1997). Several prior studies (Johansson et al., 2013; Lourdusamy et al., 2012; Melzer et al., 2008) have systematically identified genetic variations associated with protein levels and isoforms (protein quantitative trait loci or pQTLs). While most studies have focused on common variation (minor allele frequency ≥5%), rare variants, which can show strong loss of function effects, can be useful in understanding causality and pinpointing drug targets, such as deletion mutations in *PSCK9* that abolish the PSCK9 protein and reduce LDL cholesterol levels (Cohen, Boerwinkle, Mosley, & Hobbs, 2006). Systematic screening for rare variation influencing a wide variety of proteins, however, has not yet been performed.

Genetic variation is also useful in identifying causal relationships between biomarkers and diseases using tools such as Mendelian randomization (Lawlor, Harbord,

Sterne, Timpson, & Davey Smith, 2008) and could be used to ascertain how risk factors differentially affect various diseases, as well as trace causal pathways between risk loci and disease. We are investigating risk factors for cardiovascular diseases, including myocardial infarction (MI) and venous thromboembolism (VTE) in the Tromsø Study (Jacobsen, Eggen, Mathiesen, Wilsgaard, & Njølstad, 2012), a longitudinal prospective cohort study. We previously assayed 51 cardiovascular-related proteins in 419 first-ever MI cases and 398 controls in serum collected years prior to the MI event (Wilsgaard et al., 2015). Of the proteins measured, 17 were predictors for MI when considered individually after adjusting for traditional risk factors. Genetic variation associated with these protein levels could be used to study underlying mechanisms of cardiovascular diseases.

Here, using whole exome sequencing data and HumanCoreExome BeadChips, we investigate if genetic variants are associated with the serum levels of the same 51 cardiovascular-related proteins in 330 individuals chosen from the Tromsø Study because they did or did not go on to develop VTE during the 18 years of follow-up (mean time to VTE of 9 years). The serum samples were collected at study entry enabling us to identify pQTLs associated with baseline protein levels. We perform both common and rare variation association analyses in order to identify *cis*-pQTLs. Further characterization of the *cis*-pQTLs to determine if they also act as *trans*-pQTLs with any of the other 51 cardiovascular-related proteins, recapitulated well-established physiological relationships between F12, kallikrein, uPAR, kininogen, and a recent genetic association with NTproBNP. We experimentally confirmed an inferred physiological interaction from the *trans*-pQTLs by showing that kallikrein cleaves proBNP *in vitro*. We then examine genetic

associations from genome-wide association studies on coronary artery disease (CAD) and VTE as well as published literature to identify physiological and disease associations.

Chapter 2.4: Methods

Chapter 2.4.1: The Tromsø Study

The Tromsø Study is a prospective, single-site, cohort study of the inhabitants of Tromsø, Norway. In 1994-1995, 27,158 individuals filled out epidemiological surveys and donated (non-fasting) blood to the National CONOR Biobank (Jacobsen et al., 2012). These individuals were followed until 2013, with repeated surveys and identified in national registries that report various diseases and causes of death. In 2013, we identified individuals who between 1995 and 2013 had had an incident of VTE or death due to VTE, regardless of other comorbidities. We chose age and sex-matched controls randomly from the cohort. These samples were chosen for a currently ongoing case/control study of VTE. DNA and protein levels were ascertained from the blood collected in 1994.

For this specific study, blood and non-fasting serum samples were collected from 330 healthy individuals (166 males, 164 females) aged 45-75 (Supplemental Table 2.1). There were 196 individuals diagnosed with VTE between study entry (1994-'95) and the eighteen-year follow-up period (2013) and 134 controls without development of VTE during this period. Aspirin usage and other medication information were not collected for the Tromsø study. DNA was isolated from the blood for genotyping and serum samples were used to assay protein levels. The regional committee for medical and health

research ethics in North Norway approved the study, and all participants gave informed written consent.

Chapter 2.4.2: Protein Quantification

Protein levels were quantified using the same methods and at the same time as our previous MI study (Wilsgaard et al., 2015), but the samples from people that went on to develop VTE were not included in that study. Briefly, the literature was searched to create a list of over 900 cardiovascular-related proteins that might be potential biomarkers for myocardial infarction and atherosclerosis. This list was then prioritized to 165 candidate proteins, of which 51 had sufficient commercially available reagents (two antibodies and purified protein for control) in order for Tethys Bioscience, Inc (Emeryville, CA) to perform successful sandwich ELISAs (Supplemental Table 2.2). Normal Human Serum from VWR (Radnor, PA), a pool made from 10-16% of Tromsø study samples, and dilution buffer were used as controls. Each anti-protein antibody was either directly conjugated to an AlexaFluor 647 or was biotinylated and detected with a streptavidin-conjugated AlexaFluor 647. Each protein underwent 8 serial dilutions. All samples were performed in triplicate. The eight-point standard curve was measured in six replicates per plate. The AlexaFluor 647-labeled antibodies were detected using the Erenna System (Singulex, Inc., Alameda, CA). Emission from each labeled antibody is measured with a photon detector. The photon detector transmits an electronic pulse for each photon detected, and pulses are counted in 1-ms bins. Binned pulses that exceed a six standard deviation threshold above background are counted. Pulses are recorded as photons/minute. For each protein, it must be detected in >70% of samples, there must be

more than 2 logs of standard curve linear range in the ELISA, and there was less than 20% of variance between within-plate replicates for the assay to be considered successful. All protein levels were quantile normalized and mapped to the normal distribution using qnorm in R and significance was tested using Z-scores.

Chapter 2.4.3: Variant Identification and Annotation

Genotypes were determined using either the Illumina Infinium HD HumanExome BeadChip (N=87) or whole-exome sequencing (N=243) using Agilent SureSelect 50 Mb or V4 capture kits and Illumina TruSeq paired-end 100bp cluster kits. Sequence reads were mapped to the reference human genome (hg19) using BWA (version 0.7.10-r789) (Li & Durbin, 2009) with default parameters and then processed using Picard (version 1.115, tool Mark Duplicates) (http://broadinstitute.github.io/picard) and GATK (version 3.3-0, tools RealignerTargetCreator, IndelRealigner, BaseRecalibrator, PrintReads, and HaplotypeCaller) (Van der Auwera et al., 2013). We previously showed that the concordance of the exome sequencing and array genotyping data used in this study is 99.33%[2], therefore we felt confident that we could combine the genotypes from both platforms. Using the array data or information from both on and off-target reads[3] from the sequencing data, genotypes were imputed to the whole genome using Beagle (version 4.0, r1398) (Browning & Browning, 2016) and haplotypes from unrelated individuals from the European (EUR) and East Asian (EAS) superpopulations of the 1000 Genomes Project Phase 3 (Abecasis et al., 2012) for sites with a combined MAF >1%. Due to the

difference in coverage and imputation quality between sites that were exome sequenced or assayed by array, we only used sites that had a call rate of >90% in their respective datasets, and then added in additional imputation sites passing QC thresholds (allelic r^2 of >0.3). These two datasets were combined to get a final VCF with imputed and genotyped sites for both exome sequenced and exome genotyped individuals. Because the Tromsø Study is a population-based cohort study, it naturally includes some proportion of related individuals. Of the 330 individuals assayed, 20 were related to another individual in the study at an identity-by-descent value of 0.1 for exome sequenced individuals or 0.2 for arrayed individuals, based on genome-wide data.

All significant common variants were annotated for functional effects using variant effect predictor (VEP)[5], RefSeq genes, the hg19 reference genome, GeneVisible (http://genevisible.com/search) and ROADMAP[6] data of the 28-state chromHMM for Liver (E066), HepG2 (E118), and Monocyte (E029) cells. All rare variants (MAF≤5%) were annotated using VEP, RefSeq, and hg19.

Chapter 2.4.4: Statistical Analysis

Associations were performed using EPACTS software (Hyuan Min Kang, 2014). We used sex, age at study entry, BMI at study entry, genotyping platform, and VTE case/control status as covariates. Three covariates (age at serum collection, sex and BMI at serum collection) were associated respectively with ten, ten, and thirteen of the phenotypes (the 51 protein serum levels) when performing linear regressions, defined as having an FDR-adjusted P-value <0.05, and were included for consistency.

For common variants (MAF≥1%) we used EMMAX (H. M. Kang et al., 2010) (a mixed model implemented in the EPACTS software package (Hyuan Min Kang, 2014)), using q.emmax to test for single-site association. For *cis* associations we included any imputed common variants located within the interval surrounding and including the gene (+/- 500kb from transcript start and stop positions) that encodes the protein(s) being tested (C3 and C3b share the same locus). For *cis*-acting-in-*trans* associations we tested all significantly associated common *cis* variants against each of the other 50 phenotypes. For *trans* associations we tested the 100,378 common variants found in the 50 intervals against each of the 51 phenotypes (Figure 2.1).

SKAT-O (Lee, Wu, & Lin, 2012) was used to test clusters of rare variants (MAF≤5%) for association as implemented in EPACTS, using the skat-o version of the mmskat test. Rare variants were classified in three ways: 1) MAF≤5%: all rare variants located within the gene body and 2kb upstream; 2) Deleterious: all rare variants located in the gene body and the 2kb upstream region that were annotated as stop-gain, stop-loss, start-loss, essential splice site disruption, frame-shift causing, or nonsynonymous using VEP annotations; and 3) CADD-score: all rare variants in the gene or the 2kb upstream region with a PHRED-scaled c-score >10, as determined by Kircher *et al*. (Kircher et al., 2014).

We corrected for multiple testing by permuting the phenotype-genotype relationship 1000 times and for each permutation performing all variant-phenotype tests for each association type separately (e.g. *cis*, *cis*-acting-in-*trans*, or *trans*) (Hirschhorn & Daly, 2005). We obtained the lowest P-value from each permutation across all phenotypes and created a null distribution of minimum P-values. An association was

considered significant (family-wise P<0.05) if the nominal P-value was smaller than 95% of the null distribution (Supplemental Table 2.3).

To test for multiple, independent variants in the same locus, the top variant was included as a covariate until there was no longer a significant association (family-wise P<0.05) detected for that protein.

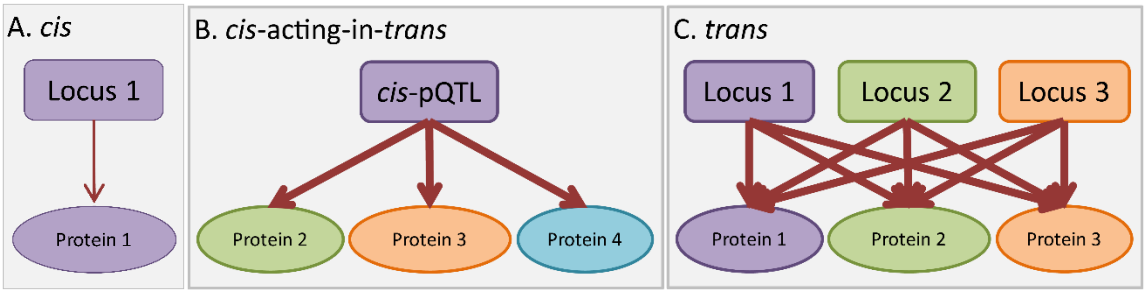


Figure 2.1 Overview of the three stages of association analyses
A) cis: for each of the 51 phenotypes (protein levels), we tested the variants located in the cis gene loci for associations with their respective protein level, B) cis-acting-in-trans: we tested the significant cis-pQTLs from stage 1 for trans effects against each of the 50 other protein levels, and C) trans: we tested all variants in the 50 cis loci (C3 and C3b share the same locus) for association with each of the 51 protein levels.

Chapter 2.4.5: Power Calculations

We calculated power using an equation from the Abecasis laboratory (http://genome.sph.umich.edu/wiki/Power_Calculations:_Quantitative_Traits) for common variants and the SKAT R package (Wu et al., 2011) for rare variants. Power for the common variant analyses was calculated using a sample size of 300 individuals, a phenotypic variance (R2) from 0.0 to 1.0, and alpha levels of 6.97 x 10-7 for the cis association, 7.29 x 10-5 for the cis-acting-in-trans analysis, and 1.25 x 10-8 for the trans analysis (Supplemental Table 2.4, Supplemental Figure 2.1). Power for the rare variant analysis was calculated using a sample size of 300 individuals, an effect size (β in

standard deviations) from 0.0 to 5.0, the default haplotypes (European) for the SKAT package, a causal MAF cutoff of 5%, a sampling subregion length of 3kb, and alpha levels of 3.72 x 10-4 for the cis association, 5.30 x 10-5 for the cis-acting-in-trans, and 9.21 x 10-6 for the trans associations (Supplemental Table 2.4, Supplemental Figure 2.2). We had 80% power to detect effects (R^2) down to 0.113 for the *cis*, common variant analysis and effects (beta) of 1.25 for the *cis*, rare variant analysis (assuming that 50% of the variants are causal), which is comparable to other pQTL studies (Garge et al., 2010; Johansson et al., 2013; Kim et al., 2013; Lourdusamy et al., 2012; Melzer et al., 2008).

Chapter 2.4.6: Clinical and Molecular Phenotype Association

Significant pQTLs from this study were queried against the eQTLs found by Schadt *et al.* (Schadt et al., 2008) in liver cells and the GTEx database (Consortium, 2013) (version 4, build 200, accessed at http://www.gtexportal.org/home/) for all tissue types. Additionally, we determined if they (or a variant in LD) overlapped any of the variants identified as pQTLs in five similarly-sized independent studies that investigated protein levels in serum (Melzer et al., 2008) or plasma (Johansson et al., 2013; Kim et al., 2013; Liu et al., 2015; Lourdusamy et al., 2012; Melzer et al., 2008). We examined pQTLs for clinical significance by determining if the variant has been previously identified and submitted to OMIM (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005), the GWAS Catalog (Welter et al., 2014), or GRASP v2.0 (Eicher et al., 2015). We identified pQTLs that were also significant in large meta-analyses of individuals of European descent for CAD or VTE. Data on CAD was downloaded from www.cardiogramplusc4d.org. For this analysis, we only used the CARDIoGRAM GWAS results (Schunkert et al., 2011), as

these individuals are of European descent. Data on VTE was shared by the INVENT consortium (Germain et al., 2015).

Chapter 2.4.7: In Vitro Assay of proBNP Cleavage

We obtained native kallikrein from human plasma from EMD-Millipore (Darmstadt, Germany; cat no. 420307); recombinant proBNP from Abcam (Cambridge, Ma; cat no. ab151881); the kallikrein inhibitor, PPACK II, from Santa Cruz Biotechnology (Dallas, Tx; cat no. sc-203215). 354ng (374 nM) of kallikrein was incubated with 80ng (606 nM) of proBNP with and without 26.5ng (36.4 μM) of PPACK II for 30min, 60min, and 90min at 37^{o}C. The reactions were stopped by adding 4X LDS sample buffer and DTT, and heating them for 2min at 85^{o}C. The proteins were run on a Tricine-SDS-page gel from ThermoFisher (Waltham, Ma), and either detected using the SilverQuest^{TM} Silver Staining Kit from ThermoFisher (Waltham, Ma) or transferred to a PVDF membrane and detected using an anti-BNP antibody from Novus Biologicals (Littleton, Co; cat no. NB100-62133) and chemiluminescence.

Chapter 2.5: Results

Chapter 2.5.1: Study Overview

The subjects were chosen as a sub-study from an ongoing case-control study examining the genetics of VTE, and include 196 individuals that developed VTE during the 18 year follow-up and 134 individuals that did not (Supplemental Table 2.1). Serum was assayed for the levels of 51 proteins using ELISAs (Supplemental Table 2.2). On

average, we obtained high quality protein measurements for 311 individuals per phenotype. We investigated if any of the protein levels were associated with VTE case/control status and found no significant associations. Knowing that the protein levels weren't statistically associated with VTE enabled us to combine the VTE cases and controls in order to explore the effects of genetic variation on baseline protein levels.

We performed high coverage (~100X) exome sequencing on DNA from blood samples for 243 individuals and assayed an additional 87 with HumanCoreExome Beadchips. We identified 158,137 variants (direct genotyping and imputation) in the 50 intervals that encode the 51 proteins (Supplemental Table 2.5). The majority of imputed variants were intergenic or intronic because these were variants not already captured by the genotyping array or were outside of the exome-sequencing target regions (Supplemental Table 2.6). There was an average of 1,122 variants per locus with the *AGER* locus having the most variants (3,523) and the *CD40LG* locus having the fewest (441) (Supplemental Table 2.2).

Chapter 2.5.2: Identifying *cis*-pQTLs from Common Variants

To identify genetic variation associated with serum protein levels, we tested for association between variants within the gene's *cis* locus and the normalized protein level for each of the 51 protein levels, adjusting for sample relatedness and population structure using a kinship matrix and including age, sex, BMI, genotype platform, and subsequent VTE status as covariates. Because of the high likelihood of linkage disequilibrium at the *cis* loci and slight correlations among protein levels, we accounted

for multiple testing by performing permutations to obtain a family-wise error rate. We identified significant associations (adjusted P<0.05, nominal P<$6.97x10^{-7}$) (Table 2.1, Figure 2.2) for thirteen of the 51 phenotypes. To test for multiple, independent associations we performed sequential conditioning on the most highly associated variant, and found two independent *cis* associations for LP(a). Of the fourteen *cis*-pQTLs that we report, we have replicated eight known pQTLs and identified six novel pQTLs. The same variant or a variant in LD (r^2>0.5 in EUR) has been previously reported for eight proteins with the same direction of effect that we found: AGT (Kim et al., 2013), C3 (Johansson et al., 2013), C3b (Johansson et al., 2013), CHIT1 (Lourdusamy et al., 2012), F12 (Liu et al., 2015; Lourdusamy et al., 2012), LBP (Lourdusamy et al., 2012), one of the variants for LP(a) (Kyriakou et al., 2014), and MMP3 (Zhu, Odeberg, Hamsten, & Eriksson, 2006) (Supplemental Table 2.7). Of the six novel pQTLs that we identified, four proteins have not previously been reported to have a *cis*-pQTL (a2-AP, ANG, KLKB1, and MMP8) and two proteins have been previously associated with a pQTL, but the variant identified here is not in LD with the previous variant (KNG1 (Liu et al., 2015), (Lourdusamy et al., 2012) and LP(a) (Kyriakou et al., 2014)). rs3373402 in *KLKB1* was previously reported to affect KLKB1 binding with kininogen (KNG1) but not affect KLKB1 levels in plasma (Katsuda, Maruyama, Ezaki, Sawamura, & Ichihara, 2007); therefore while this variant has been previously functionally characterized this is a novel pQTL. We annotated the 14 pQTLs for functional effects and identified their chromatin state in the tissue that they are most highly expressed in (Supplemental Table 2.8). Ten of the thirteen proteins are predominantly secreted by the liver. Five of the top variants are missense variants, three are in the UTR regions and five lie in predicted regulatory regions based on chromatin

state annotations. These analyses suggest possible mechanisms of action for some of the *cis*-pQTLs.

Table 2.1 Significant cis-pQTLs from the common variant analysis.

Protein	Gene	Top Variant	Chr (b37)	Position (b37)	Alleles (Ref/Alt)	Alt Allele Frequency	Nominal *P* Value	Adjusted *P* Value	β	R^2
a2-AP	*SERPINF2*	rs8077638	17	1640793	C/T	0.19	5.4×10^{-37}	<0.001	−1.15	0.42
AGT	*AGT*	rs4762	1	230845977	G/A	0.14	4.4×10^{-22}	<0.001	1.08	0.25
ANG	*ANG*	rs3748338	14	21167576	A/T	0.11	1.9×10^{-12}	<0.001	0.86	0.16
C3	*C3*	rs11569415	19	6716279	G/A	0.15	6.9×10^{-10}	<0.001	−0.63	0.13
C3B	*C3*	rs2230199	19	6718387	G/C	0.23	1.2×10^{-12}	<0.001	−0.65	0.16
CHIT1	*CHIT1*	rs2486951	1	203174921	A/G	0.18	3.7×10^{-21}	<0.001	−1.01	0.26
F12	*F12*	rs1801020	5	176836532	A/G	0.76	2.5×10^{-32}	<0.001	0.99	0.38
KLKB1	*KLKB1*	rs3733402	4	187158034	G/A	0.53	4.4×10^{-12}	<0.001	−0.51	0.15
KNG1	*KNG1*	rs166479	3	186443250	T/C	0.41	1.7×10^{-10}	<0.001	−0.46	0.13
LBP	*LBP*	rs2232613	20	36997655	C/T	0.10	2.2×10^{-22}	<0.001	−1.20	0.27
LP(a)*	*APOA*	rs41272114	6	161006077	C/T	0.030	3.1×10^{-8}	0.002	−1.27	0.10
LP(a)*	*APOA*	rs56393506	6	161089307	C/T	0.083	1.7×10^{-7}	0.011	0.66	0.08
MMP3	*MMP3*	rs7926920	11	102698724	G/A	0.35	2.4×10^{-14}	<0.001	−0.41	0.17
MMP8	*MMP8*	rs35231465	11	102584135	G/A	0.036	1.9×10^{-7}	0.012	−1.10	0.09

β indicates effect size of association in standard deviation units per each copy of the alternate allele; Alt, alternate; pQTLs, protein quantitative trait loci; R^2, amount of phenotypic variation explained by the variant; and Ref, reference.

*LP(a) has 2 independent *cis*-pQTLs. rs56393506 was identified as an independent pQTL for LP(a) by performing the association analysis using genotypes from the top variant (rs41272114) as a covariate.

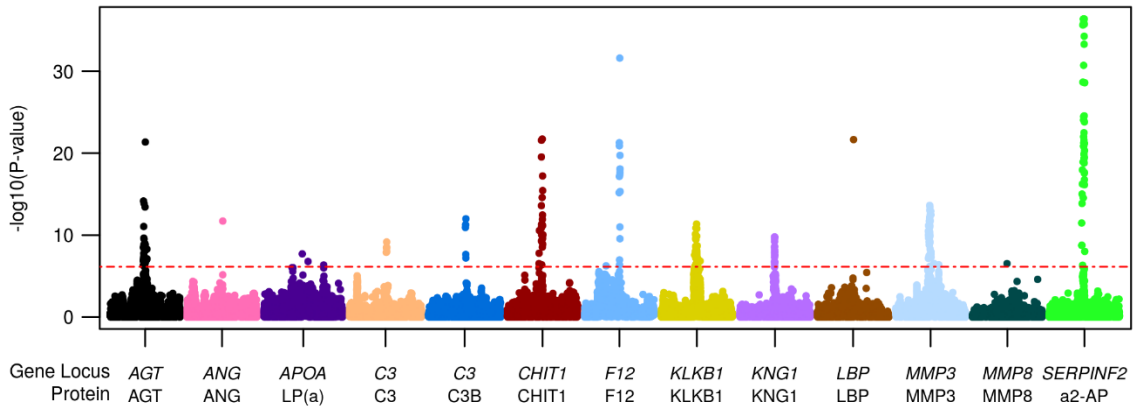


Figure 2.2 Association of cis variants with protein levels.
Modified Manhattan plot showing the –log10 P-values for association between variants in each cis locus (interval encoding protein +/- 500kb) and the respective protein levels. The red dashed line indicates the study-wide significant P-value cutoff when only examining cis regions (6.9x10-7) for a FWER <0.05.

Chapter 2.5.3: Identifying *cis*-pQTLs from Rare Variation

We next tested whether the combination of multiple rare variants at each *cis*-locus was associated with protein levels. There were 3,675 rare variants identified across all 50 loci. For rare variation association analyses, rare variants are grouped according to frequency or function and then jointly tested for association. Because functional prediction methods vary and it is currently unknown what method is superior (Santorico & Hendricks, 2016), we used three different classifications (MAF, Deleterious, and CADD-score – see Methods). Across all loci there was a range of 1 to 90 variants used for each method, with the MAF method having the most rare variants and CADD scores having the fewest. To account for multiple testing, we tested all three classifications in each round of permutations to determine the family-wise error rate P-value cutoff. We performed a SKAT-O association test using the same covariates as for the common variant association. We identified eight *cis*-pQTLs that were significant using one or more classifications (adjusted $P<0.05$, nominal $P<3.72\times10^{-4}$) (Table 2.2, Supplemental Table 2.9). Of these, *cis* rare variation has been associated with AGER (Hudson et al., 2008), Fetuin A (Yuasa & Umetsu, 1988), and LP(a) levels (Kyriakou et al., 2014); to our knowledge the other five associations are novel.

Of the eight proteins associated with rare variation, three were also associated with a common pQTL (CHIT1, LP(a), and MMP8). For LP(a) and MMP8, a common pQTL (with a MAF<5%) was also present on the list of rare variants and removal of these from the rare variant analysis made the rare association non-significant (CADD nominal P-value 0.148 and 0.469, respectively). For CHIT1, the common pQTL had a MAF of 18% and although not on the list of rare variants, when we included this variant as a covariate

in the rare variant analysis the association was nullified (nominal P-value 0.147). These results suggest that the rare variants in the *CHIT1* locus were associated with CHIT1 serum levels due to linkage disequilibrium with the common pQTL. Because the driving variant was common, we do not consider the CHIT1 association to be valid, resulting in seven proteins associated with rare variants.

Table 2.2 Rare variant cis-pQTLs that are significant using at least 1 of the 3 grouping methods.

	MAF≤5%		Deleterious		CADD10	
Protein	Nominal *P* Value	Adjusted *P* Value	Nominal *P* Value	Adjusted *P* Value	Nominal *P* Value	Adjusted *P* Value
AGER	3.2×10^{-4}	0.041	0.003	n.s.	0.006	n.s.
CD40L	0.042	n.s.	9.8×10^{-5}	0.009	0.003	n.s.
CHIT1*	4.3×10^{-8}	<0.001	0.108	n.s.	0.127	n.s.
Fetuin A	2.5×10^{-4}	0.026	1.5×10^{-5}	0.002	2.4×10^{-6}	<0.001
LP(a)	1.1×10^{-5}	0.002	2.6×10^{-8}	<0.001	4.4×10^{-7}	<0.001
MMP8	0.247	n.s.	7.7×10^{-4}	n.s.	6.3×10^{-6}	0.002
TAFI	0.014	n.s.	5.2×10^{-5}	0.003	0.002	n.s.
TIMP4	0.050	n.s.	2.4×10^{-4}	0.026	1.7×10^{-4}	0.018

n.s. indicates not significant; and pQTLs, protein quantitative trait loci.
*Not significant after adjusting for the common pQTL (rs2486951).

Chapter 2.5.4: Identifying *trans*-pQTLs

To characterize potential downstream effects of *cis*-pQTLs, we investigated whether any of the common *cis*-pQTLs might also have *trans* effects (*cis*-acting-in-*trans*) on any of the other 50 protein levels. After permutation to obtain adjusted P-values, we identified two *cis*-acting-in-*trans* loci, each of which was significantly associated with three proteins (adjusted P<0.05, nominal $P<7.29\times10^{-5}$) (Table 2.3). There was significant overlap in the proteins associated with the two loci and the associations were consistent with known physiological relationships between F12, KLKB1, KNG1, and uPAR, and the recently reported genetic relationship with NTproBNP (Musani et al., 2015) (Figure 2.3), despite none of the protein levels being strongly correlated (Supplemental Figure 2.3 and

Supplemental Table 2.10). We did not observe an association between the *cis*-pQTL for *KLKB1* and F12 protein levels despite the known physiological relationships of KLKB1 and F12 (Figure 2.3). Importantly, the genetic associations of *KLKB1* and *F12* with NTproBNP suggest that KLKB1 may physiologically cleave proBNP (the NTproBNP precursor). These findings illustrate how genetic variation can be used to identify potentially novel physiological relationships among proteins.

We further performed a full pairwise association (*trans*) between any of the variants located in the 50 regions encoding the proteins used in this study and all 51 protein levels. After permutation adjusting (adjusted $P<0.05$, nominal $P<1.25 \times 10^{-8}$) we did not find any additional *trans* associations and none of the *cis*-acting-in-*trans* associations remained significant; however 11 of the 14 *cis* associations remained significant.

Using a similar approach to the common variants, we tested if any of the rare variant *cis*-pQTLs were associated with any of the other 50 protein levels and did not observe any significant associations (adjusted $P<0.05$, nominal $P<5.30 \times 10^{-5}$). Additionally, we tested all 50 *cis* regions against all 51 protein levels in a pairwise manner, but did not identify additional associations (adjusted $P<0.05$, nominal $P<9.21 \times 10^{-6}$), although four of the eight rare *cis* associations were still significant at the more stringent threshold.

Table 2.3 cis-pQTLs that also act as trans-pQTLs

Variant	Protein	Nominal *P* Value	Adjusted *P* Value	β	R^2
rs1801020 in the *F12* locus	F12	2.5×10^{-32}	<0.001	0.985	0.382
	KLKB1	5.4×10^{-8}	<0.001	−0.488	0.092
	KNG1	1.1×10^{-7}	<0.001	−0.479	0.097
	NTproBNP	1.2×10^{-5}	0.002	−0.380	0.061
rs3733402 in the *KLKB1* locus	KLKB1	4.4×10^{-12}	<0.001	−0.506	0.152
	KNG1	5.2×10^{-5}	0.034	−0.309	0.049
	NTproBNP	4.2×10^{-8}	<0.001	−0.393	0.098
	uPAR	4.4×10^{-8}	<0.001	−0.401	0.097

β indicates effect size of association in standard deviation units; pQTLs, protein quantitative trait loci; and R^2, amount of phenotypic variation explained by the variant.

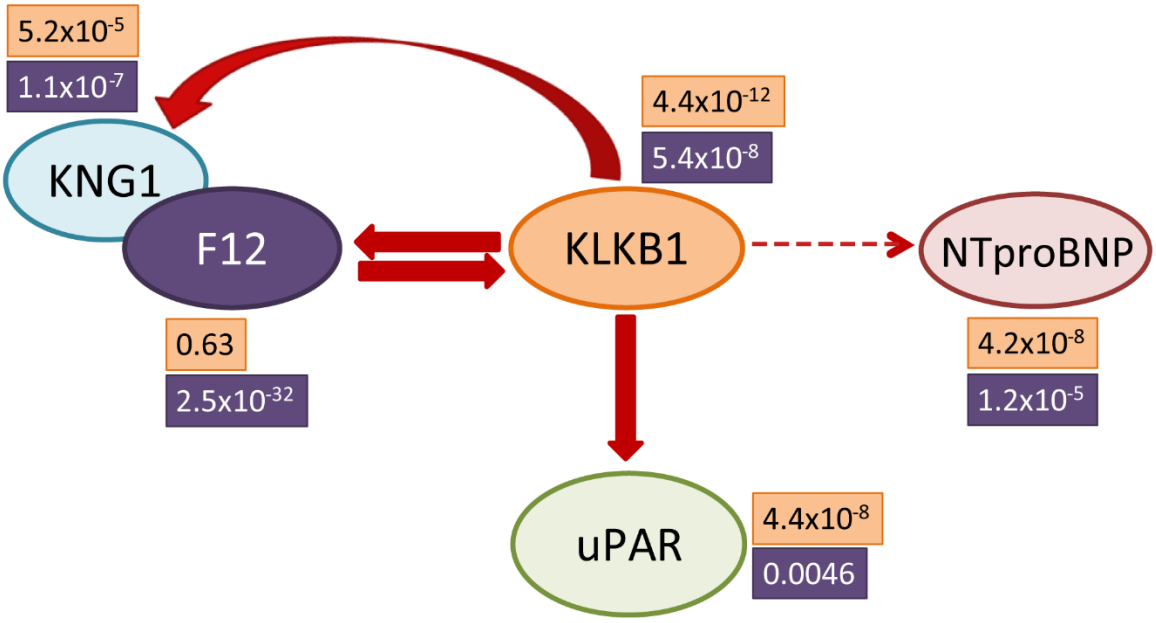


Figure 2.3 Schematic showing proteins with identified trans associations and their nominal associations with SNPs in F12 and KLKB1.
Previously known (solid) and proposed in this study (dashed) cleavage reactions are represented with arrows. Nominal P-values for the associations between protein levels and rs3733402 in the KLKB1 locus and rs1801020 in the F12 locus are shown respectively in orange and purple boxes next to the protein of interest.

Chapter 2.5.5: The Role of Kallikrein in proBNP Maturation

We experimentally tested the *cis*-acting-in-*trans* associations suggesting that kallikrein (KLKB1) may physiologically cleave proBNP. ProBNP is produced as a pro-peptide that may be cleaved intracellularly into BNP and NTproBNP, two biomarkers for heart failure (Clerico, Fontana, Zyw, Passino, & Emdin, 2007), before being secreted by cardiomyocytes in response to cardiac stress. Intracellularly, it is thought that furin or corin cleave proBNP (Semenov et al., 2010), but it is unclear which enzyme cleaves proBNP extracellularly when it is secreted intact (Tonne et al., 2011). To test whether kallikrein can cleave proBNP *in vitro,* we incubated increasing concentrations of kallikrein (74.8nM, 374nM, 748nM, and 1497nM) with proBNP for 1 hour at body temperature (37^{o}C) and saw progressive depletion of proBNP levels (Supplemental Figure 2.4). This depletion was prevented with the addition of PPACK II, a kallikrein-specific inhibitor. From this, we chose to incubate 374 nM of kallikrein with proBNP for 30, 60 or 90 minutes and again, we saw that the levels of proBNP decreased (Figure 2.4). These results suggest that kallikrein has the ability to cleave proBNP *in vivo*.

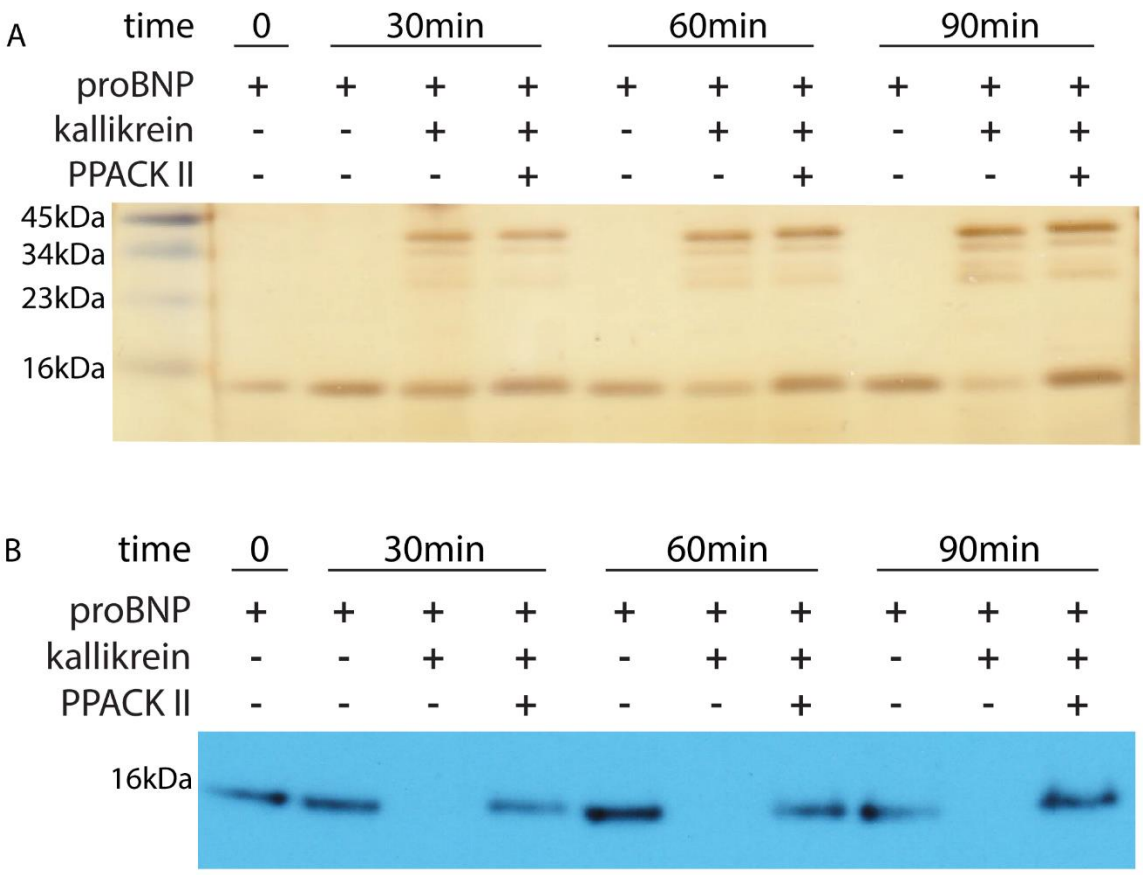


Figure 2.4 Kallikrein cleaves proBNP in vitro.
A) A silver stain of recombinant proBNP and kallikrein incubated together for 30, 60, and 90 minutes with and without a kallikrein-specific inhibitor (PPACK II) and B) a western blot of an identical experimental setup using an anti-BNP antibody. The silver stain binds all protein present and is a more sensitive procedure than using the anti-BNP antibody for the western blot. We believe that this explains why the amount of proBNP in the +/+/- wells visually appears to be different between the silver stain and western blot.

Chapter 2.5.6: Annotation of pQTLs Using Existing Databases and GWAS

We investigated whether the 14 common pQTLs that we identified were previously associated with gene expression levels (Supplemental Table 2.7) using eQTLs from the GTEx database (Consortium, 2013) as well as Schadt *et al.* (Schadt et al., 2008) to include additional data from liver samples, as many of the proteins studied are expressed in liver. In the GTEx database, the AGT pQTL was identified as an eQTL in ten tissues

(P-values from of $2.0x10^{-6}$ to $1.3x10^{-33}$), the CHIT1 pQTL is an eQTL in whole blood (P-value $4.2x10^{-8}$), the F12 pQTL is an eQTL in liver (P-value $2.3x10^{-10}$) and the pQTL in the *SERPINF2* locus (a2-AP protein) is an eQTL in six tissues (P-values from $5.3x10^{-7}$ to $8.8x10^{-18}$). Additionally, the pQTLs for a2-AP, AGT, CHIT1, F12, KLKB1, and MMP3 were also identified as eQTLs for other nearby genes. In the Schadt dataset rs3748338 in the *ANG* locus is in LD (r^2=0.24) with an eQTL for ANG (rs8008440). Thus, of the 14 common pQTLs, two have previously been identified as an eQTL for the *cis* gene, three as an eQTL for both the *cis* gene and other nearby genes, and three as an eQTL for nearby gene(s).

We also looked up whether there are any known disease associations with the fourteen pQTLs that we identified using the GWAS catalog (Welter et al., 2014). GRASP (Eicher et al., 2015), and OMIM (Hamosh et al., 2005) (Supplemental Table 2.7). The 8 known pQTLs along with the kallikrein pQTL are associated with a variety of phenotypes, including age-related macular degeneration (C3b), activated partial thromboplastin times (F12), serum metabolites (KLKB1), binding of LBP to LPS (LBP), and plasma plasminogen levels (LP(a)). In total, nine pQTLs (eight known and KLKB1) have been associated with 17 disease and/or physiological phenotypes.

Finally, to investigate if the pQTLs identified here are associated with VTE or CAD, we examined the results of two previously published meta-analyses. The INVENT (Germain et al., 2015) study is a large meta-analysis of 7,507 cases and 52,632 controls to identify variants associated with VTE. The CARDIoGRAM (Schunkert et al., 2011) study is a large meta-analysis of 22,233 cases and 64,762 controls designed to identify variants associated with CAD, which is predominantly comprised of MI. Of 14 common

pQTLs, ten (71.4%) could be tested in the INVENT and CARDIoGRAM datasets (Supplemental Table 2.11). The KLKB1 pQTL (rs3733402) is significantly associated with VTE; however this association becomes non-significant when the analysis is conditioned on the top six SNPs associated with VTE from the literature. The KLKB1 pQTL (rs3733402) is also nominally associated with CAD (P=0.0086). The KNG1 pQTL (rs166479) had a nominal P-value <0.05 in the INVENT consortium. While one of the pQTLs for LP(a) (rs41272114) has previously been associated with CAD (Kyriakou et al., 2014), it was not present in either dataset. Additionally, among the 17 protein biomarkers that we previously identified as being associated with first MI (Wilsgaard et al., 2015), we identified common *cis*-pQTLs for six (C3, C3b, KLKB1, LP(a), MMP3, MMP8) and rare *cis*-pQTLs for five (LP(a), MMP8, TAFI, and TIMP4). While we found pQTLs for these MI biomarkers, they weren't associated with CAD in the CARDIoGRAM study, which could indicate that the biomarkers are not causally related to CAD, but may also be a result of the relatively small sample size in the GWAS compared to typical Mendelian randomization studies. Thus, while CAD and VTE were not significantly associated with pQTLs, these loci could be used in further larger studies to elucidate functional mechanisms underlying disease.

Chapter 2.6: Discussion

Using a combination of exome sequencing and exome arrays in 330 individuals, we identified 27 genetic associations between pQTLs and the serum levels of 20 proteins: 14 associated with common variation in *cis*, of which six are novel and have not been previously reported; seven associations with rare variants in *cis*, of which four are novel;

and six associations in *trans*. Ultimately, 15 proteins were associated with single sites and seven were associated with rare variants. The strongest associations were identified for *cis* variation near the gene locus, but by directly testing the *cis*-pQTLs, we also identified two that acted in *trans*. Despite the limitations of our study (including a relatively small sample size and lack of a formal replication cohort) the presence of robust associations suggest that exome analysis is an effective tool to identify genetic variation associated with serum protein levels and that larger sample sizes would likely capture additional *trans* effects.

This is the first study, to the best of our knowledge, that uses exome data to investigate the effects of both common and rare variation on more than 50 protein levels and thus, provides insight into rare-variant association methods. For rare-variant analysis we used three different methods for grouping variants within a gene and accounted for the additional testing through permutation. Some associations were consistent across all three methods, such as LP(a), which carried a large number of variants (Supplemental Table 2.9) and for which rare variation has previously been associated with the protein level in the blood (Clarke et al., 2009). Others were only significant in one test, such as MMP8 when variants were grouped based on CADD score, which could be due to few variants with weak effects and would benefit from larger sample sizes to include more predicted functional sites. Variants with a MAF between 1% and 5% were tested in both the common and rare variant analyses. In two cases (LP(a) and MMP8) adjusting for the top common pQTL (with a MAF<5%) nullified the association. Additionally, for CHIT1, common variation (MAF>5%) was associated with rare variants through cryptic LD and

adjusting for the common variant also nullified the association. These data suggest that significant common and rare single sites may drive gene-based rare-variant associations.

Of the fourteen common pQTLs, four are missense variants in the relevant gene. Of the ten other variants, three are intronic, two are in the exons of nearby genes, and five lie in regions that are predicted to have regulatory functions, such as interrupting protein-binding sites or splicing (Supplemental Table 2.8). Analysis of the function of sequences harboring the pQTL can elucidate the mechanism of the variant. For example, it has been shown that rs1801020 in the 3' UTR of the *F12* locus prevents translation of F12 (Kanaji et al., 1998). The mechanisms of the other four regulatory pQTLs are not yet understood, but the results shown here point to plausible mechanisms. For instance, *ANG* and *RNASE4* are isoforms of the same gene with different functions and differential expression patterns that are influenced by CTCF (Sheng et al., 2014). The ANG pQTL is in the last exon of *RNASE4*, near a CTCF binding site which affects isoform expression levels (Sheng et al., 2014). This, and other potentially regulatory pQTLs, could be functionally tested using *in vitro* and *in vivo* assays for changes in gene or isoform expression. Thus, although we focused on exome sequences to generate genotypes for this analysis, imputation enabled us to identify many pQTLs with predicted regulatory effects.

pQTLs can be used to understand the relationship between proteins and disease, either through tracing molecular impacts through pathways or through studies of Mendelian randomization. By examining potential *trans* associations with *cis*-pQTLs, we recapitulated known and recently reported relationships between these proteins. The relationships between F12, kallikrein, and kininogen comprise the start of the intrinsic

coagulation pathway (Bhoola, Figueroa, & Worthy, 1992), the association between kallikrein and uPAR has been previously explored (Portelli et al., 2014), and the genetic relationship between kallikrein and NTproBNP was identified in a recent GWAS (Musani et al., 2015). We show that kallikrein is able to cleave proBNP *in vitro* using purified reagents, suggesting that extracellularly, kallikrein could be responsible for cleaving proBNP into NTproBNP and BNP, although further experiments are necessary to verify that this reaction occurs naturally in plasma. We also identified 17 reported disease and physiological phenotype associations with nine of the pQTLs (eight previously known and one novel). Interestingly, five of the six novel pQTLs were not implicated in GWAS studies. This could reflect a bias in GWAS phenotypes studied or candidate proteins chosen for pQTL studies and supports further work identifying downstream effects of these loci. We observed a nominal association between KLKB1 and CAD, which we previously identified as a biomarker for MI, supporting further examination of this relationship in larger studies. Overall, these findings support the use of pQTLs to identify molecular and phenotypic effects of proteins and help to elucidate underlying mechanisms of disease.

Chapter 2.7: Funding Sources

This work was supported by an independent grant from the K.G. Jebsen Foundation in Norway and partially funded by Tethys Bioscience. TS is supported by an institutional award to the UCSD Genetics Training Program from the National Institute for General Medical Sciences, T32 GM008666.

Chapter 2.8: Acknowledgments

Chapter 2, in full, is a reprint of material as it appears in Circulation: Cardiovascular Genetics, 2016, Terry Solomon, Erin Smith, Hiroko Matsui, Sigrid Braekkan, Tom Wilsgaard, Inger Njølstad, Ellisiv Mathiesen, John-Bjarne Hansen, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

The INVENT Consortium is comprised of Philippe Amouyel, Mariza de Andrade, Saonli Basu, Claudine Berr, Jennifer A Brody, Daniel I Chasman, Jean-Francois Dartigues, Aaron R Folsom, Marine Germain, Hugoline de Haan, John Heit, Jeanine Houwing-Duitermaat, Christopher Kabrhel, Peter Kraft, Grégoire Legal, Sara Lindström, Ramin Monajemi, Pierre-Emmanuel Morange, Bruce M Psaty, Pieter H Reitsma, Paul M Ridker, Lynda M Rose, Frits R Rosendaal, Noémie Saut, Eline Slagboom, David Smadja, Nicholas L Smith, Pierre Suchon, Weihong Tang, Kent D Taylor, David-Alexandre Trégouët, Christophe Tzourio, Marieke CH de Visser, Astrid van Hylckama Vlieg, Lu-Chen Weng, and Kerri L Wiggins.

Chapter 2.9: Disclosures of Conflict of Interest

The authors declare that they have nothing to disclose.

Chapter 2.10: Supplemental Tables

Supplemental Table 2.1 Cohort statistics

	Sex	N	Age	BMI	N Exome Sequenced	N Exome Arrayed
VTE Controls	Females	64	65.29 (50-74)	26.40 (17.6-38.6)	45 (70%)	19 (30%)
	Males	70	64.14 (46-74)	25.85 (20.3-34.6)	52 (74%)	18 (26%)
VTE Cases	Females	100	63.71 (50-75)	27.21 (18.6-41.8)	74 (74%)	26 (26%)
	Males	96	63.45 (45-74)	26.94 (19.7-36.4)	72 (75%)	24 (25%)
Total		330	64.03 (45-75)	26.69 (17.6-41.8)	243 (74%)	87 (26%)

N, number

Supplemental Table 2.2 The fifty-one proteins and 50 loci (C3 and C3b derive from the same gene and locus, but are considered two proteins here) used in this study.

Protein Symbol	Protein Name	Gene	Chr	Locus Start	Locus End	N *cis* Variants	N Individuals Measured	Phenotypic Mean	Range (min-max)	Units
a2-AP	alpha-2-antiplasmin	*SERPINF2*	17	1146129	2158559	1295	303	3.06	0.07-17.79	ug/mL
ACE	angiotensin-converting enzyme	*ACE*	17	61054421	62075741	1075	318	858.45	15.51-3344.92	ng/mL
ADIPOQ	adiponectin	*ADIPOQ*	3	186060462	187076252	766	329	6.81	0.74-99.32	ug/mL
AGER	advanced glycosylation end product-specific receptor	*AGER*	6	31648744	32652099	3523	303	0.40	0.08-1.67	ng/mL
AGT	angiotensinogen	*AGT*	1	230338271	231350336	678	330	1.37	0.01-13.71	ug/mL
ANG	angiogenin	*ANG*	14	20652335	21662345	1516	298	219.78	89.63-482.54	ng/mL
APOA1	apolipoprotein A-1	*APOA1*	11	116206468	117208338	721	302	1582.73	0.08-3704.94	ug/mL
APOB	apolipoprotein B-100	*APOB*	2	20724300	21766945	482	303	19.44	3.82-189.05	ug/mL
APOC3	apolipoprotein C-III	*APOC3*	11	116200623	117203787	715	294	255.52	108.25-960.93	ug/mL
BGLAP	osteocalcin	*BGLAP*	1	155711950	156713123	1564	328	6.27	0.60-63.58	ug/mL
BSG	basigin	*BSG*	19	71324	1083493	2306	328	42.99	15.03-87.72	ng/mL
C3	complement C3	*C3*	19	6177845	7220662	1714	284	233.77	71.76-631.00	mg/mL
C3b	complement component C3b	*C3*	19	6177845	7220662	1714	301	2.95	1.05-16.21	ug/mL
CCL5	C-C motif chemokine 5	*CCL5*	17	33698495	34707377	923	328	155.92	7.77-759.97	ng/mL
CD14	monocyte differentiation antigen CD14	*CD14*	5	139511312	140513286	1104	328	244.71	76.28-739.07	ng/mL
CD163	scavenger receptor cysteine-rich type 1 protein M130	*CD163*	12	7123411	8156414	930	328	180.84	14.71-4863.99	ng/mL
CD40-L	CD40 ligand	*CD40LG*	X	135230335	136242549	441	303	13.31	1.84-154.09	ng/mL
CHIT1	chitotriosidase-1	*CHIT1*	1	202685206	203698860	951	303	57.02	0.27-477.70	ng/mL
CRP	C-reactive protein	*CRP*	1	159182078	160184379	1050	303	199.32	2.50-4877.52	ng/mL
CST3	cystatin-C	*CST3*	20	23114293	24118574	459	302	678.73	239.86-2412.46	ng/mL
CTSG	cathepsin G	*CTSG*	14	24542723	25545466	1539	330	37.26	6.88-186.27	ng/mL
CXCL10	C-X-C motif chemokine 10	*CXCL10*	4	76442268	77444689	985	303	0.05	0.01-0.74	ng/mL
DCN	decorin	*DCN*	12	91039034	92073359	128	303	13.47	6.00-31.13	ng/mL
DPP4	dipeptidyl peptidase 4	*DPP4*	2	162348754	163431052	530	328	992.83	225.56-5879.35	ng/mL
F12	coagulation factor XII	*F12*	5	176329138	177336577	955	303	24.82	0.94-50.27	ug/mL
Fetuin A	alpha-2-HS-glycoprotein	*AHSG*	3	185830849	186839107	752	302	924.06	401.88-3961.18	ug/mL

Supplemental Table 2.2 The fifty-one proteins and 50 loci (C3 and C3b derive from the same gene and locus, but are considered two proteins here) used in this study, continued.

FTH1	ferritin heavy chain	*FTH1*	11	61231756	62235132	918	329	286.80	6.05-4905.20	ng/mL
HP	haptoglobin	*HP*	16	71588507	72594955	1000	301	799.22	0.06-4090.92	ug/mL
HSPA1B	heat shock 70kDa protein 1B	*HSPA1B*	6	31295511	32298031	3372	303	4.05	0.90-62.02	ng/mL
ICAM1	intercellular adhesion molecule 1	*ICAM1*	19	9881516	10897291	1920	328	34.59	4.31-97.50	ng/mL
KLKB1	plasma kallikrein	*KLKB1*	4	186648671	187679625	798	303	29494.68	2.46-59991.93	ng/mL
KNG1	kininogen-1	*KNG1*	3	185935097	186960678	809	302	84574.02	5.57-302245.75	ng/mL
LBP	lipopolysaccharide-binding protein	*LBP*	20	36474884	37505653	953	303	984.41	170.31-2056.94	ng/mL
LP(a)	apolipoprotein(A)	*APOA*	6	160452514	161587407	1080	303	343.47	12.11-3656.08	ng/mL
MMP3	stromelysin-1	*MMP3*	11	102206527	103214342	1056	328	11.26	2.21-75.17	ng/mL
MMP8	neutrophil collagenase	*MMP8*	11	102082525	103095685	971	302	14.34	1.47-68.09	ng/mL
MMP9	matrix metalloproteinase-9	*MMP9*	20	44137546	45145200	1312	303	458.41	106.99-2665.14	ng/mL
MPO	myeloperoxidase	*MPO*	17	55847216	56858296	1116	303	60.57	12.77-271.17	ng/mL
NTproBNP	n-terminus pro-brain natriuretic protein	*NPPB*	1	11417520	12418992	1380	303	0.30	0.02-6.38	ng/mL
PAI-1	plasminogen activator inhibitor 1	*SERPINE1*	7	100270378	101282547	2129	301	38.30	15.35-95.10	ng/mL
REN	renin	*REN*	1	203623943	204635465	1018	329	0.63	0.05-3.33	ng/mL
SHBG	sex hormone-binding globulin	*SHBG*	17	7017381	8036700	2576	303	1.79	0.21-7.28	ug/mL
TAFI	carboxypeptidase B2	*CPB2*	13	46127321	47179211	498	303	27.92	15.36-68.45	ug/mL
THBS1	thrombospondin-1	*THBS1*	15	39373279	40389668	413	320	44.55	4.05-190.52	ug/mL
THBS4	thrombospondin-4	*THBS4*	5	78831169	79879107	810	303	2.79	0.16-192.66	ug/mL
TIMP1	metalloproteinase inhibitor 1	*TIMP1*	X	46941689	47946190	501	328	61.42	7.27-176.27	ng/mL
TIMP4	metalloproteinase inhibitor 4	*TIMP4*	3	11694567	12700851	573	302	5.20	1.88-18.00	ng/mL
TNFRSF11B	tumor necrosis factor receptor superfamily member 11B	*TNFRSF11B*	8	119435795	120464383	175	328	49.80	11.71-152.77	ng/mL
TNFRSF1B	tumor necrosis factor receptor superfamily member 1B	*TNFRSF1B*	1	11727059	12769277	1355	328	19.79	7.56-49.06	ng/mL
uPAR	urokinase plasminogen activator surface receptor	*PLAUR*	19	43650246	44674498	1362	303	1.73	0.54-4.91	ng/mL
VCAM1	vascular cell adhesion protein 1	*VCAM1*	1	100685195	101704601	326	328	138.01	39.81-679.86	ng/mL

Supplemental Table 2.3 Number of tests performed for each type of association analysis and the P-value cutoffs using Bonferroni correction or permutations for a FWER < 0.05.

Variant Class	Analysis	Number of tests	Bonferroni cutoff	Permutation cutoff
Common, single site	*Cis*	100,378	4.98×10^{-7}	6.91×10^{-7}
	Cis-acting-in-*trans*	663	7.40×10^{-5}	7.29×10^{-5}
	trans	5,119,278	9.77×10^{-9}	1.25×10^{-8}
Rare, collapsed	*Cis*	153	3.21×10^{-4}	3.72×10^{-4}
	Cis-acting-in-*trans*	918	5.34×10^{-5}	5.30×10^{-5}
	trans	7,803	6.16×10^{-6}	9.21×10^{-6}

Supplemental Table 2.4 Amount of phenotypic variance explained (R^2) and effect size (β) detected for the various analyses when there is 80% power.

Variant Class	Analysis	Alpha	Variance Explained (R^2)
	cis	$6.91x10^{-7}$	0.113
Common, single-site	*cis*-as-*trans*	$7.29x10^{-5}$	0.078
	trans	$1.25x10^{-8}$	0.143
Variant Class	**Analysis**	**Alpha**	**Effect Size (β)**
	cis	3.72×10^{-4}	0.80
Rare, collapsed [100% causal]	*cis*-as-*trans*	5.30×10^{-5}	0.90
	trans	9.21×10^{-6}	1.0
	cis	3.72×10^{-4}	1.25
Rare, collapsed [50% causal]	*cis*-as-*trans*	5.30×10^{-5}	1.45
	trans	9.21×10^{-6}	1.75

Supplemental Table 2.5 Number of variants that were directly genotyped or imputed for the exome sequenced and exome arrayed individuals.

	Exome Sequenced (N = 243)	Exome Arrayed (N = 87)	Combined
Genotyped	24,008	2,563	24,915
Imputed	129,502	56,195	138,415
Total	153,510	58,758	158,137

N, number of individuals

Supplemental Table 2.6 Type of variants that were directly genotyped or imputed for the exome sequenced and exome arrayed individuals.

Variant type	Exome Sequenced (Genotyped)	Exome Arrayed (Genotyped)	Exome Sequenced (Imputed)	Exome Arrayed (Imputed)	% of total Exome Sequenced variants that were imputed	% of total Exome Arrayed variants that were imputed
Intergenic	1161	355	66432	25886	98.3%	98.6%
Non-coding RNA	694	30	783	470	53.0%	94.0%
Intronic	14700	459	59702	27640	80.2%	98.4%
UTR variant	1323	53	2017	1168	60.4%	95.7%
Synonymous	2488	73	203	600	7.5%	89.2%
Missense	3378	1568	313	380	8.5%	19.5%
Coding sequence indels	192	0	16	34	7.7%	100.0%
start or stop related variants	72	25	7	8	8.9%	24.2%

Supplemental Table 2.7 Reported disease associations of the significant cis-pQTLs.

Locus	Top Variant	Chr	Position	Known pQTL? (if reported direction of effect matches this study's direction of effect)	Schadt eQTL	GTEx eQTL	Genome-wide associations	OMIM
AGT	rs4762	1	230845977	Kim[8] (yes)		AGT sun-exposed skin, transformed fibroblasts, non-exposed skin, subcutaneous adipose, lung, colon, breast, testis, esophagus (2.0x10-6 to 1.3x10-33); RP11-99J16_A.2 transformed fibroblasts, sun-exposed skin, subcutaneous adipose (6.0x10-10 to 1.0x10-14)		in LD with rs699 and rs5051 (r^2 = 0.24, D' = 1) which are associated with hypertension, lower promoter activity and transcription amount
ANG	rs3748338	14	21167576	novel	rs8008440 (r^2=0.24) -log10p = 5.7144			
C3	rs11569415	19	6716279	Johansson[9] (yes)				in LD with rs2230199 (r^2 = 0.98, D' = 0.99) which is associated with age-related macular degeneracy
C3	rs2230199	19	6718387	Johansson[9] (yes)			age-related macular degeneration	age-related macular degeneracy, slow and fast in electrophoresis, kidney production of C3
CHIT1	rs2486951	1	203174921	Lourdusamy[10] (yes)		CHIT1 whole blood (4.2x10-8); ADORA1 transformed fibroblasts (1.4x10-6)		
F12	rs1801020	5	176836532	Liu[11] (yes)		F12 liver (2.3x10-10); MXD3 esophagus (4.2x10-6)	F12 levels, protective effect on acute coronary syndrome in people with stable CAD, activated partial thromplastin time, serum metabolite levels	

Supplemental Table 2.7 Reported disease associations of the significant cis-pQTLs, continued.

KLKB1	rs3733402	4	187158034	novel	F11 artery, esophagus, brain, and muscle (1.8-7.0x10-6)	B-type natriuretic peptide, midregional-proadrenomedullin and C-terminal-pro-endothelin-1, serum metabolite levels	PKD Sedi is two mutations (compound heterozygosity with rs121964952) that reduce binding to HMWK
KNG1	rs166479	3	186443250	novel		activated partial thromboplastin time	
LBP	rs2232613	20	36997655	Lourdusamy[10] (yes)		reduced binding capacity for LPS, homozygous has low serum concentrations, protease cleavage site, carriers have cleaved LBP which doesn't bind LPS and has low cytokine after LPS	
APOA	rs41272114	6	161006077	Kyriakou[12] (yes)		plasma plasminogen levels	
APOA	rs56393506	6	161089307	Novel			
MMP3	rs7926920	11	102698724	Zhu[13] (yes)	MMP1 transformed fibroblasts (6.8x10-9); WTAPP1 transformed fibroblasts, testis (1.9-8.0x10-7)	Serum MMP-1 levels	
MMP8	rs35231465	11	102584135	novel			
SERPINF2	rs2070863	17	1648502	novel	SERPINF2 skeletal muscle, sun-exposed skin, subcutaneous adipose, transformed fibroblasts, esophagus, testis (5.3x10-7 to 8.8x10-18); WRD81 esophagus, sun-exposed skin, breast (2x10-6 to 1.3x10-12)		

Supplemental Table 2.8 Functional annotations of the 14 significant cis-pQTLs using GeneVisble, variant effect predictor (VEP) and ROADMAP data of the 28-state chromHMM for Monocyte (E029), Liver (E066) and HepG2 (E118) cells.

Protein	Gene	Top Variant	Chr	Postion	Expressed in	macrophage ROADMAP	Liver ROADMAP	HepG2 ROADMAP	VEP Annotations
AGT	*AGT*	rs4762	1	230845977	liver	quiescent	downstream promoter TSS 2	Transcribed & regulatory	missense variant (*AGT*); TF binding site variant (Nrsf)
ANG	*ANG*	rs3748338	14	21167576	liver	quiescent	Weak transcription	quiescent	missense (*RNASE4*)
LP(a)	*APOA*	rs41272114	6	161006077	liver	quiescent	quiescent	quiescent	splice donor variant (*APOA*)
LP(a)	*APOA*	rs56393506	6	161089307		quiescent	quiescent	quiescent	upstream gene variant (*APOA*); regulatory region variant (CTCF binding site)
C3	*C3*	rs11569415	19	6716279	liver	quiescent	quiescent	Weak transcription	intronic (*C3*)
C3b	*C3*	rs2230199	19	6718387	liver	quiescent	downstream promoter TSS 2	downstream promoter TSS 2	missense (*C3*); TF binding site variant (Egr1)
CHIT1	*CHIT1*	rs2486951	1	203174921	macrophage	quiescent	quiescent	quiescent	regulatory region variant (promoter flanking region); intergenic
F12	*F12*	rs1801020	5	176836532	liver	quiescent	Poised promoter	Active transcription start site	5' UTR variant (*F12*)
KLKB1	*KLKB1*	rs3733402	4	187158034	liver	quiescent	quiescent	quiescent	missense (*KLKB1*)
KNG1	*KNG1*	rs1656921	3	186442833	liver	quiescent	quiescent	quiescent	intronic (*KNG1*)
LBP	*LBP*	rs2232613	20	36997655	liver	quiescent	Weak transcription	quiescent	missense (*LBP*)
MMP3	*MMP3*	rs2155013	11	102701858	joints	quiescent	quiescent	quiescent	intronic (*WTAPP1*)
MMP8	*MMP8*	rs35231465	11	102584135	bone marrow	quiescent	quiescent	quiescent	Stop gain (*MMP8*); 3' UTR variant (*MMP8*)
a2-AP	*SERPINF2*	rs8077638	17	1640793	liver	Transcribed 3' preferential	Transcribed 3' preferential	Transcribed 3' preferential	synonymous variant (*WDR81*); upstream gene variant; downstream gene variant

Supplemental Table 2.9 List of rare variants that comprise each significant rare cis-pQTL association and their P-values from the single-site associations.
An X means that the variant was used in the indicated clustering method. Gray regions mean that the collapsed region was not significant using that clustering method.

Protein	Chr	Start	rsID	Ref/Effect	MAF	Single-site P-value	MAF ≤ 5%	Deleterious	CADD10
AGER	6	32147044	rs142802704	AAG/A	0.03477	0.0005711	X		
	6	32147157	rs41268928,rs116420335	G/C	0.03465	0.0005482	X		
	6	32148724	rs41270464,rs114878357	C/T	0.04125	0.8907	X		
	6	32148814	.	GGGTTATACAGGAGAGA/G	0.0022	NA	X		
	6	32148909	rs201575255	CTG/C	0.00165	NA	X		
	6	32149065	rs181811810	C/T	0.00441	NA	X		
	6	32149140	rs3176931	C/T	0.00658	NA	X		
	6	32149471	rs114564020	G/A	0.0022	NA	X		
	6	32149571	.	C/A	0.01322	0.01551	X		
	6	32149801	rs9391855,rs116515025	C/T	0.02211	0.03112	X		
	6	32149883	rs204996,rs116334026	C/T	0.0297	0.1413	X		
	6	32150047	rs77170610	C/T	0.00704	NA	X		
	6	32150107	.	TGAGGCCCTATCTCAGG/T	0.0022	NA	X		
	6	32150303	rs144335694	T/C	0.0022	NA	X		
	6	32150523	.	T/C	0.0022	NA	X		
	6	32150872	.	G/A	0.0022	NA	X		
	6	32151443	rs2070600,rs114177847	C/T	0.0396	0.001871	X		
	6	32151458	rs80096349,rs116828224	G/A	0.00165	NA	X		
	6	32151539	.	G/A	0.0022	NA	X		
	6	32151882	rs115111668	C/T	0.0022	NA	X		
Fetuin A	3	186329282	rs111451953	G/A	0.03084	0.4621	X		
	3	186330883	.	G/A	0.00441	NA	X		
	3	186331119	.	T/A	0.00441	NA	X		
	3	186331138	rs150486317	C/T	0.01159	0.5768	X	X	
	3	186331245	rs190631595	G/T	0.0022	NA	X		
	3	186331298	.	A/C	0.00231	NA	X		
	3	186331299	.	G/C	0.00231	NA	X		
	3	186333378	.	C/A	0.0022	NA	X		
	3	186334343	rs79747711	T/G	0.00441	NA	X		
	3	186334932	.	A/G	0.0022	NA	X		
	3	186335056	rs140827890	G/A	0.00661	0.5564	X	X	X
	3	186335248	rs144616056	G/A	0.01762	0.5311	X		
	3	186337746	rs149819140	T/C	0.0022	NA	X		
	3	186337871	rs184392275	G/A	0.0022	NA	X		
	3	186338320	.	T/G	0.0022	NA	X		
	3	186338540	rs35799453	T/C	0.00441	NA	X		

Supplemental Table 2.7 Reported disease associations of the significant cis-pQTLs, continued.

	3	186338564	rs35457250	C/T	0.01821	1.48E-06	X	X	X
	3	186338869	rs11540663	C/T	0.01542	0.9813	X		
CD40LG	X	135741275	.	G/A	0.0022	NA		X	
	X	135741443	rs148594123	G/A	0.0165	0.001338		X	
CHIT1	1	203183825	rs2015402	G/A	0.04846	0.2873	X		
	1	203184018	rs946849	T/C	0.04846	0.2873	X		
	1	203184924	rs80241012	G/A	0.02643	0.1151	X		
	1	203185118	rs17532442	C/T	0.03744	0.2581	X		
	1	203186420	.	A/C	0.01106	0.667	X		
	1	203186666	rs41308417	C/G	0.02212	0.6183	X		
	1	203188379	.	CCCACTGGTTGTCCCGGAAGATGTAGGGCA/C	0.00661	0.02734	X		
	1	203189093	rs76499133	C/T	0.03111	0.02956	X		
	1	203189350	rs74969659	G/A	0.03304	0.2992	X		
	1	203189634	rs2486959	A/G	0.04626	3.46E-13	X		
	1	203191527	rs56152830	C/T	0.01322	0.4565	X		
	1	203191994	rs7512820	G/A	0.00889	0.3586	X		
	1	203192424	rs181385947	G/A	0.02643	0.9288	X		
	1	203192518	.	G/A	0.00661	0.9764	X		
	1	203193134	rs2486068	T/G	0.04405	1.32E-09	X		
	1	203194544	rs140940634	C/CATT	0.03965	0.5655	X		
	1	203194548	rs184545416	G/T	0.03965	0.5655	X		
	1	203194688	.	C/T	0.00441	NA	X		
	1	203194834	rs137852607	C/T	0.0033	NA	X		
	1	203195006	rs116389839	G/A	0.01322	0.2128	X		
	1	203195126	rs185483258	C/T	0.00229	NA	X		
	1	203195398	rs10920587	C/A	0.00881	0.9725	X		
	1	203195689	rs2486070	G/A	0.04405	9.21E-11	X		
	1	203196479	rs2486071	A/C	0.04626	9.39E-11	X		
	1	203196842	rs56035601	C/T	0.03965	0.5655	X		
	1	203198596	rs72739588	G/A	0.02477	0.2938	X		
LP(a)	6	160952621	rs41266381	A/C	0.00441	NA	X		
	6	160952667	rs73012273	C/G	0.01982	0.7021	X		
	6	160952780	rs186413938	C/T	0.0033	NA	X	X	X
	6	160952816	rs41267807	T/C	0.0132	0.4979	X	X	X
	6	160953642	rs41267809	A/G	0.0198	0.01013	X	X	
	6	160960892	.	AAG/A	0.0022	NA	X		
	6	160961137	rs3798220	T/C	0.00828	0.1252	X	X	X
	6	160962115	.	A/G	0.0022	NA	X		
	6	160962151	.	T/G	0.0022	NA	X	X	X
	6	160962185	.	A/G	0.00441	NA	X		
	6	160962190	.	C/A	0.0022	NA	X	X	X
	6	160962366	.	TC/T	0.02535	0.8887	X		
	6	160962368	rs116039216	G/T	0.02546	0.8848	X		
	6	160962370	rs116089584	G/T	0.02804	0.9047	X		

Supplemental Table 2.7 Reported disease associations of the significant cis-pQTLs, continued.

6	160963576	rs41265940	T/A	0.00221	NA	X		
6	160963648	rs6920765	G/A	0.0022	NA	X		
6	160963964	rs41265934	C/G	0.00441	NA	X		
6	160964135	rs41265930	T/C	0.00661	0.868	X		
6	160966559	rs139145675	G/A	0.00165	NA	X	X	X
6	160968863	rs149574804	C/T	0.01982	0.4082	X		
6	160968968	rs41264848	G/A	0.0022	NA	X		
6	160969075	rs41264844	C/T	0.03084	0.5789	X		
6	160969096	rs4708871	C/T	0.02423	0.2038	X		
6	160969113	rs145989243	G/A	0.0022	NA	X		
6	160971286	rs62441900	C/G	0.03965	0.2143	X		
6	160973905	rs62441901	C/G	0.03965	0.2143	X		
6	160976914	.	CA/C	0.00221	NA	X		
6	160978270	rs184372256	A/G	0.00442	NA	X		
6	160978686	rs149526393	A/G	0.00441	NA	X		
6	160985107	rs145783310	A/G	0.0495	0.7805	X		
6	160985438	rs79563112	C/T	0.0495	0.7805	X		
6	160985526	rs118039278	G/A	0.02632	0.0406	X		
6	160997118	rs74617384	A/T	0.02632	0.0406	X		
6	160998052	rs76602267	G/A	0.00441	NA	X		
6	160998143	.	C/T	0.0022	NA	X		
6	161005610	rs55730499	C/T	0.02632	0.0406	X		
6	161005898	.	C/T	0.00221	NA	X		
6	161005908	rs41272116	ACTT/A	0.00221	NA	X		
6	161006077	rs41272114*	C/T	0.0297	1.05E-08	X	X	X
6	161006084	rs76144756	G/A	0.00441	NA	X	X	X
6	161006105	rs41272112	C/T	0.00165	NA	X	X	
6	161007647	.	G/A	0.0022	NA	X		
6	161010118	rs10455872	A/G	0.01815	0.005497	X		
6	161010546	rs41267815	G/C	0.00441	NA	X		
6	161011907	rs74334585	C/T	0.01101	0.1663	X		
6	161012262	rs144958108	A/G	0.00667	0.8523	X		
6	161015301	rs41271036	A/G	0.01101	4.23E-06	X		
6	161016414	.	A/G	0.0022	NA	X		
6	161020526	rs41270998	A/G	0.0022	NA	X		
6	161020532	.	G/A	0.00661	0.417	X		X
6	161021800	.	G/T	0.00221	NA	X		
6	161022107	rs41259144	C/T	0.00662	0.003408	X	X	X
6	161022108	rs186072375	G/T	0.0022	NA	X		X
6	161025782	.	G/T	0.00455	NA	X		
6	161026197	rs117174672	G/A	0.03084	0.2908	X		
6	161026250	.	G/A	0.0022	NA	X		
6	161027256	.	A/G	0.00231	NA	X		
6	161027287	.	G/C	0.00221	NA	X		
6	161027430	rs144587038	G/T	0.01542	0.1217	X		

Supplemental Table 2.7 Reported disease associations of the significant cis-pQTLs, continued.

	6	161027821	rs75055004	A/T	0.0022	NA	X		
	6	161027895	.	TA/T	0.00457	NA	X		
	6	161032267	rs117949336	C/T	0.00461	NA	X		
	6	161032369	rs151298886	C/T	0.00221	NA	X		
	6	161032401	rs78893353	G/A	0.01106	0.1991	X		
	6	161032413	.	G/C	0.00221	NA	X		
	6	161032497	rs112092923	A/G	0.02212	0.09362	X		
	6	161055824	.	C/A	0.00243	NA	X		
	6	161055876	.	C/T	0.00444	NA	X		
	6	161055880	.	G/A	0.00222	NA	X		
	6	161055942	.	G/C	0.00441	NA	X		
	6	161055968	.	C/A	0.0022	NA	X		
	6	161055991	.	C/T	0.0022	NA	X		
	6	161056129	.	T/TAGA	0.0022	NA	X		
	6	161071476	rs200491482	T/G	0.00165	NA	X	X	X
	6	161071615	.	G/A	0.0022	NA	X		
	6	161071638	.	T/C	0.0022	NA	X		
	6	161087336	rs117643720	T/C	0.0022	NA	X		
	6	161087368	.	G/A	0.00881	0.06095	X		
	6	161087372	rs181060240	T/C	0.01322	0.03437	X		
MMP8	11	102584135	rs35231465*	G/A	0.03642	6.86E-07			X
	11	102585130	.	A/G	0.00224	NA			X
	11	102586142	rs61753779	A/G	0.00166	NA			X
	11	102592160	rs11602288	G/A	0.00661	0.7785			X
	11	102593266	rs112188995	C/T	0.00993	0.1649			X
TAFI	13	46627762	rs145067962	A/T	0.0066	0.003747		X	
	13	46648069	rs140446990	T/C	0.00495	0.0407		X	
	13	46656619	.	C/T	0.0022	NA		X	
TIMP4	3	12195137	.	G/C	0.00221	NA		X	X
	3	12195660	rs140022692	G/A	0.01106	0.1126			X
	3	12198889	.	A/G	0.00221	NA		X	X
	3	12200201	.	C/T	0.00221	NA		X	X
	3	12200210	.	G/A	0.00442	NA		X	X
	3	12200219	.	A/G	0.00221	NA		X	X
	3	12200269	.	C/T	0.00221	NA			X

* = variant was tested in both the common analysis and in the rare analysis.

Supplemental Table 2.10 Pearson's correlation between the proteins that were identified in the trans-pQTL analysis.

	F12	KLKB1	KNG1	NTproBNP	uPAR
F12	1.0	-0.18	0.09	-0.11	-0.06
KLKB1	-0.18	1.0	0.29	0.16	0.20
KNG1	0.09	0.29	1.0	0.09	0.32
NTproBNP	-0.11	0.16	0.09	1.0	0.28
uPAR	-0.06	0.20	0.32	0.28	1.0

Supplemental Table 2.11 Lookup of common cis-pQTLs for their associations in the CARDIoGRAM and INVENT meta-analyses.

Protein	Variant	Chr	Start	Ref/Alt	CARDIoGRAM Coronary Artery Disease			INVENT Venous Thromboembolism		
					P-value	β	SE	P-value	β	SE
a2-AP	rs8077638*	17	1640793	C/T	0.3399	0.156	0.017	Not in INVENT		
a2-AP	rs8065251*	17	1637458	G/A	0.3552	0.015	0.017	0.3298	-0.028	0.029
AGT	rs4762	1	230845977	G/A	0.5018	-0.143	0.021	0.1694	-0.048	0.035
ANG	rs3748338	14	21167576	A/T	0.4549	0.020	0.027	0.9519	0.002	0.039
C3/C3b	rs2230199†	19	6718387	G/C	0.1045	0.050	0.031	0.7668	0.010	0.034
CHIT1	rs2486951	1	203174921	A/G	0.7288	-0.006	0.017	0.8562	-0.005	0.028
F12	rs1801020	5	176836532	A/G	0.5115	-0.016	0.246	0.6957	0.011	0.028
KLKB1‡	**rs3733402**	4	187158034	G/A	**0.0086**	0.040	0.015	**$8.2x10^{-12}$**	-0.159	0.023
KNG1	**rs166479**	3	186443250	T/C	0.1692	0.020	0.014	**0.0436**	-0.046	0.023
LBP	rs2232613	20	36997655	C/T	0.3624	0.025	0.028	0.3531	0.050	0.054
MMP3	rs7926920	11	102698724	G/A	0.1016	0.023	0.014	0.7347	0.008	0.023

Bolded indicates a nominal P-value of <0.05. β, effect size. SE, standard error

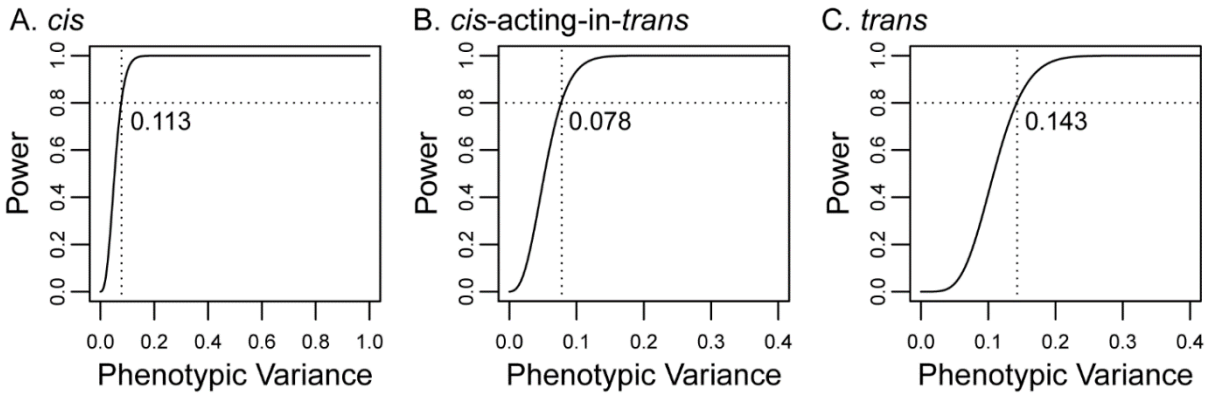
* rs8077638 was not present in the INVENT dataset, so the next most significant variant (rs8065251) was looked up in both studies as well.

† The top variant for C3 (rs11569415) was not present in either CARDIoGRAM or INVENT so the next most significant variant (rs2230199) was used.

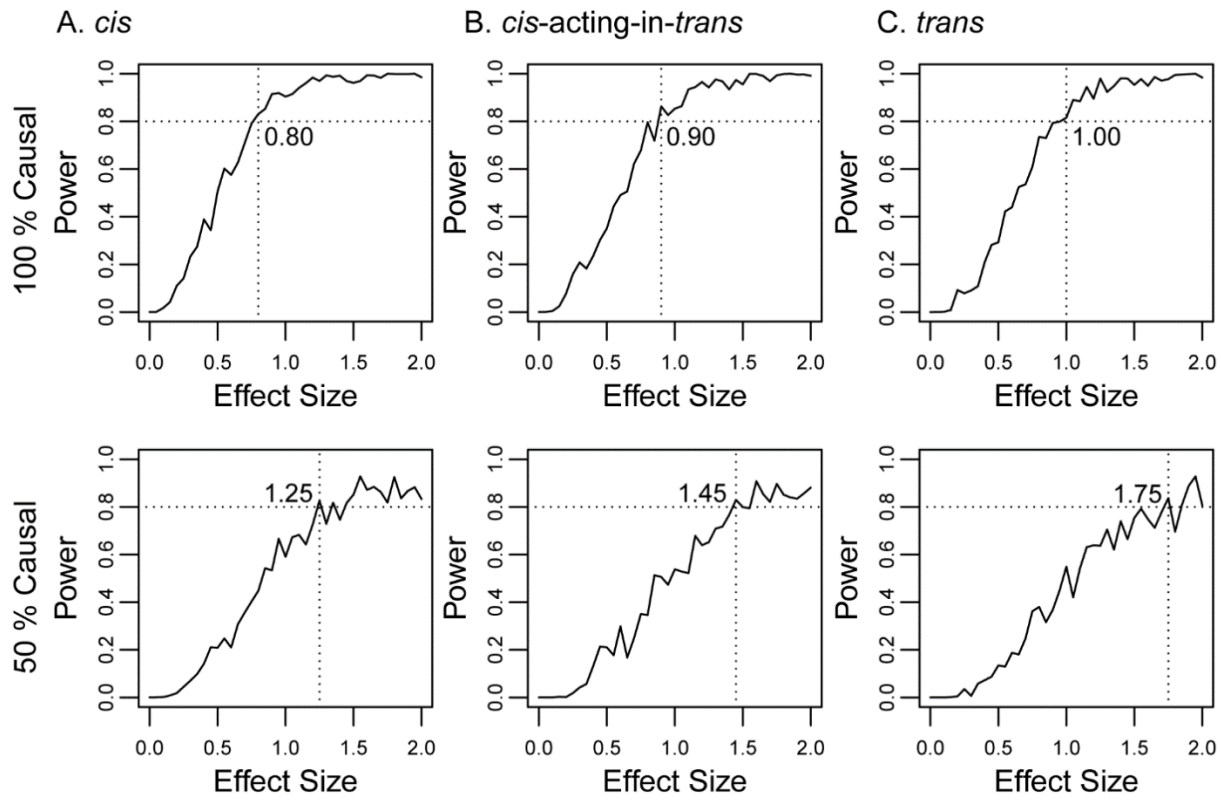
‡ The association with VTE was no longer significant after adjustment that included rs4253417.

No significant variants for LP(a) or MMP8 were present in either study.

Chapter 2.11: Supplemental Figures

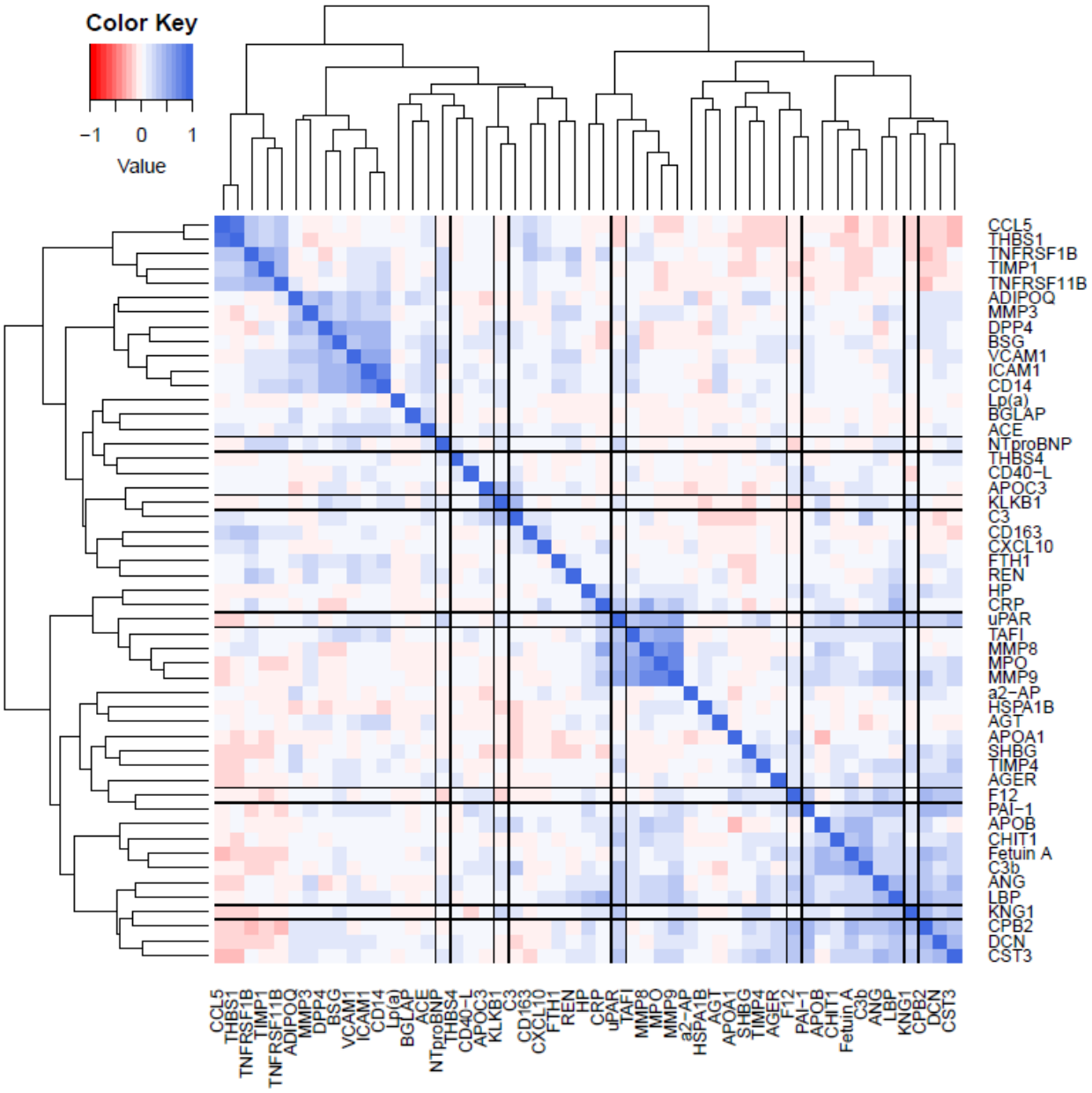


Supplemental Figure 2.1 Power to detect common variation pQTLs with varying effect sizes in the three stages of analysis.
A) Power curve for cis-pQTLs using the permutation cutoff of 6.91x10-7 as the alpha. B) Power curve for testing the cis-pQTLs acting-in-trans using the permutation cutoff of 7.29x10-5 as the alpha. C) Power curve for trans-pQTLs using the permutation cutoff of 1.25x10-8. The x-axis is measuring the amount of variance of the phenotype that a variant explains (R^2).

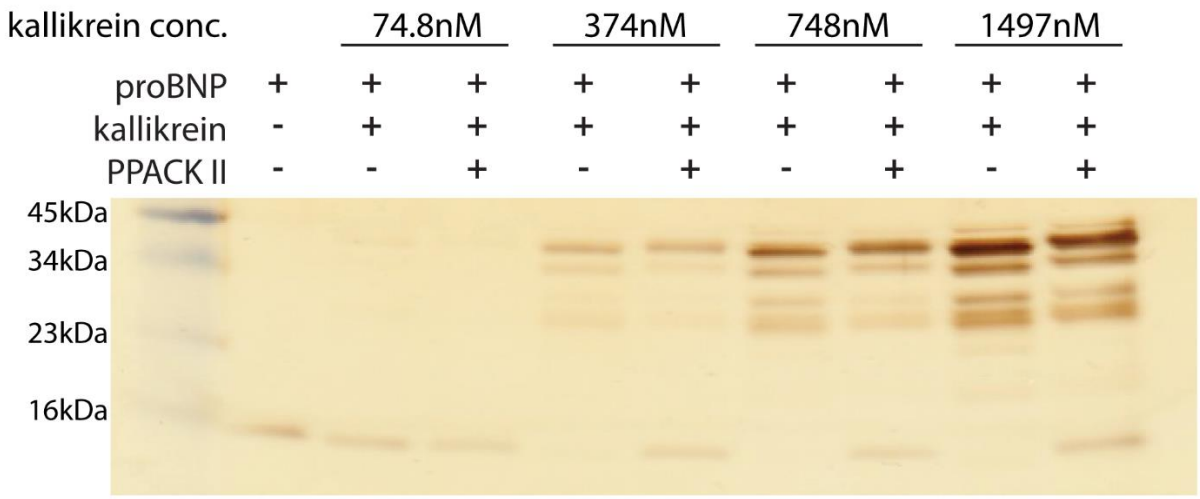


Supplemental Figure 2.2 Power to detect rare variation pQTLs with varying effect sizes in the three stages of analysis.

Effect size is measured in standard deviations (β). The top row assumes that all variants have an equal effect and that all variants are causal. The bottom row assumes that all variants have an equal effect and that half of the variants tested are causal. A) power to detect *cis* associations, alpha = 3.72×10^{-4}; B) power to detect *cis*-acting-in-*trans* pQTLs, alpha = 5.30×10^{-5}; C) power to detect *trans* associations, alpha = 9.21×10^{-6}. The x-axis is measuring the effect size (β) in standard deviations.



Supplemental Figure 2.3 Pearson's correlation of the protein levels.
Dendrogram shows clustering based on correlation. Black lines are outlining the proteins identified in the *cis*-acting-in-*trans* analysis: F12, KLKB1, KNG1, NTproBNP, uPAR; corresponding values can be found in Supplemental Table 2.10.



Supplemental Figure 2.4 Silver stain of proBNP incubated for 1 hour with varying concentrations of kallikrein, with and without a kallikrein-specific inhibitor, PPACK II. Kallikrein concentrations are 74.8nM, 374nM, 748nM, and 1497nM. The upper bands are the light and heavy chains of kallikrein. The lower band is proBNP. 374nM of kallikrein was chosen to perform the silver stain and western blot in Figure 2.4 of the paper.

Chapter 2.12: References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., McVean, G. A., & Consortium, G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi:10.1038/nature11632

Anderson, L., & Seilhamer, J. (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis, 18*(3-4), 533-537. doi:10.1002/elps.1150180333

Bhoola, K. D., Figueroa, C. D., & Worthy, K. (1992). Bioregulation of kinins: kallikreins, kininogens, and kininases. *Pharmacol Rev, 44*(1), 1-80.

Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet, 98*(1), 116-126. doi:10.1016/j.ajhg.2015.11.020

Clarke, R., Peden, J. F., Hopewell, J. C., Kyriakou, T., Goel, A., Heath, S. C., Parish, S., Barlera, S., Franzosi, M. G., Rust, S., Bennett, D., Silveira, A., Malarstig, A., Green, F. R., Lathrop, M., Gigante, B., Leander, K., de Faire, U., Seedorf, U., Hamsten, A., Collins, R., Watkins, H., Farrall, M., & Consortium, P. (2009). Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med, 361*(26), 2518-2528. doi:10.1056/NEJMoa0902604

Claussnitzer, M., Dankel, S. N., Kim, K. H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L., Puviindran, V., Abdennur, N. A., Liu, J., Svensson, P. A., Hsu, Y. H., Drucker, D. J., Mellgren, G., Hui, C. C., Hauner, H., & Kellis, M. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med, 373*(10), 895-907. doi:10.1056/NEJMoa1502214

Clerico, A., Fontana, M., Zyw, L., Passino, C., & Emdin, M. (2007). Comparison of the diagnostic accuracy of brain natriuretic peptide (BNP) and the N-terminal part of the propeptide of BNP immunoassays

in chronic and acute heart failure: a systematic review. *Clin Chem, 53*(5), 813-822. doi:10.1373/clinchem.2006.075713

Cohen, J. C., Boerwinkle, E., Mosley, T. H., & Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med, 354*(12), 1264-1272. doi:10.1056/NEJMoa054013

Consortium, G. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet, 45*(6), 580-585. doi:10.1038/ng.2653

Eicher, J. D., Landowski, C., Stackhouse, B., Sloan, A., Chen, W., Jensen, N., Lien, J. P., Leslie, R., & Johnson, A. D. (2015). GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Res, 43*(Database issue), D799-804. doi:10.1093/nar/gku1202

Garge, N., Pan, H., Rowland, M. D., Cargile, B. J., Zhang, X., Cooley, P. C., Page, G. P., & Bunger, M. K. (2010). Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. *Mol Cell Proteomics, 9*(7), 1383-1399. doi:10.1074/mcp.M900378-MCP200

Germain, M., Chasman, D. I., de Haan, H., Tang, W., Lindström, S., Weng, L. C., de Andrade, M., de Visser, M. C., Wiggins, K. L., Suchon, P., Saut, N., Smadja, D. M., Le Gal, G., van Hylckama Vlieg, A., Di Narzo, A., Hao, K., Nelson, C. P., Rocanin-Arjo, A., Folkersen, L., Monajemi, R., Rose, L. M., Brody, J. A., Slagboom, E., Aïssi, D., Gagnon, F., Deleuze, J. F., Deloukas, P., Tzourio, C., Dartigues, J. F., Berr, C., Taylor, K. D., Civelek, M., Eriksson, P., Psaty, B. M., Houwing-Duitermaat, J., Goodall, A. H., Cambien, F., Kraft, P., Amouyel, P., Samani, N. J., Basu, S., Ridker, P. M., Rosendaal, F. R., Kabrhel, C., Folsom, A. R., Heit, J., Reitsma, P. H., Trégouët, D. A., Smith, N. L., Morange, P. E., & Consortium, C. (2015). Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet, 96*(4), 532-542. doi:10.1016/j.ajhg.2015.01.019

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res, 33*(Database issue), D514-517. doi:10.1093/nar/gki033

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet, 6*(2), 95-108. doi:10.1038/nrg1521

Hudson, B. I., Carter, A. M., Harja, E., Kalea, A. Z., Arriero, M., Yang, H., Grant, P. J., & Schmidt, A. M. (2008). Identification, classification, and expression of RAGE gene splice variants. *FASEB J, 22*(5), 1572-1580. doi:10.1096/fj.07-9909com

Jacobsen, B. K., Eggen, A. E., Mathiesen, E. B., Wilsgaard, T., & Njølstad, I. (2012). Cohort profile: the Tromso Study. *Int J Epidemiol, 41*(4), 961-967. doi:10.1093/ije/dyr049

Johansson, Å., Enroth, S., Palmblad, M., Deelder, A. M., Bergquist, J., & Gyllensten, U. (2013). Identification of genetic variants influencing the human plasma proteome. *Proc Natl Acad Sci U S A, 110*(12), 4673-4678. doi:10.1073/pnas.1217238110

Kanaji, T., Okamura, T., Osaki, K., Kuroiwa, M., Shimoda, K., Hamasaki, N., & Niho, Y. (1998). A common genetic polymorphism (46 C to T substitution) in the 5'-untranslated region of the coagulation factor XII gene is associated with low translation efficiency and decrease in plasma factor XII level. *Blood, 91*(6), 2010-2014.

Kang, H. M. (2014). EPACTS (Version 3.2.5): University of Michigan Center for Statistical Genetics. Retrieved from http://www.sph.umich.edu/csg/kang/epacts/

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet, 42*(4), 348-354. doi:10.1038/ng.548

Katsuda, I., Maruyama, F., Ezaki, K., Sawamura, T., & Ichihara, Y. (2007). A new type of plasma prekallikrein deficiency associated with homozygosity for Gly104Arg and Asn124Ser in apple domain 2 of the heavy-chain region. *Eur J Haematol, 79*(1), 59-68. doi:10.1111/j.1600-0609.2007.00871.x

Kim, S., Swaminathan, S., Inlow, M., Risacher, S. L., Nho, K., Shen, L., Foroud, T. M., Petersen, R. C., Aisen, P. S., Soares, H., Toledo, J. B., Shaw, L. M., Trojanowski, J. Q., Weiner, M. W., McDonald, B. C., Farlow, M. R., Ghetti, B., Saykin, A. J., & (ADNI), A. s. D. N. I. (2013). Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. *PLoS One, 8*(7), e70269. doi:10.1371/journal.pone.0070269

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. doi:10.1038/ng.2892

Kyriakou, T., Seedorf, U., Goel, A., Hopewell, J. C., Clarke, R., Watkins, H., Farrall, M., & Consortium, P. (2014). A common LPA null allele associates with lower lipoprotein(a) levels and coronary artery disease risk. *Arterioscler Thromb Vasc Biol, 34*(9), 2095-2099. doi:10.1161/ATVBAHA.114.303462

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med, 27*(8), 1133-1163. doi:10.1002/sim.3034

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics, 13*(4), 762-775. doi:10.1093/biostatistics/kxs014

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Liu, Y., Buil, A., Collins, B. C., Gillet, L. C., Blum, L. C., Cheng, L. Y., Vitek, O., Mouritsen, J., Lachance, G., Spector, T. D., Dermitzakis, E. T., & Aebersold, R. (2015). Quantitative variability of 342 plasma proteins in a human twin population. *Mol Syst Biol, 11*, 786.

Lourdusamy, A., Newhouse, S., Lunnon, K., Proitsi, P., Powell, J., Hodges, A., Nelson, S. K., Stewart, A., Williams, S., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M., Vellas, B., Lovestone, S., Dobson, R., Consortium, A., & Initiative, A. s. D. N. (2012). Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum Mol Genet, 21*(16), 3719-3726. doi:10.1093/hmg/dds186

Melzer, D., Perry, J. R., Hernandez, D., Corsi, A. M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J. R., Paolisso, G., Rafiq, S., Simon-Sanchez, J., Lango, H., Scholz, S., Weedon, M. N., Arepalli, S., Rice, N., Washecka, N., Hurst, A., Britton, A., Henley, W., van de Leemput, J., Li, R., Newman, A. B., Tranah, G., Harris, T., Panicker, V., Dayan, C., Bennett, A., McCarthy, M. I., Ruokonen, A., Jarvelin, M. R., Guralnik, J., Bandinelli, S., Frayling, T. M., Singleton, A., & Ferrucci, L. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet, 4*(5), e1000072. doi:10.1371/journal.pgen.1000072

Musani, S. K., Fox, E. R., Kraja, A., Bidulescu, A., Lieb, W., Lin, H., Beecham, A., Chen, M. H., Felix, J. F., Fox, C. S., Kao, W. H., Kardia, S. L., Liu, C. T., Nalls, M. A., Rundek, T., Sacco, R. L., Smith, J., Sun, Y. V., Wilson, G., Zhang, Z., Mosley, T. H., Taylor, H. A., & Vasan, R. S. (2015). Genome-wide association analysis of plasma B-type natriuretic peptide in blacks: the Jackson Heart Study. *Circ Cardiovasc Genet, 8*(1), 122-130. doi:10.1161/circgenetics.114.000900

Portelli, M. A., Siedlinski, M., Stewart, C. E., Postma, D. S., Nieuwenhuis, M. A., Vonk, J. M., Nurnberg, P., Altmuller, J., Moffatt, M. F., Wardlaw, A. J., Parker, S. G., Connolly, M. J., Koppelman, G. H., & Sayers, I. (2014). Genome-wide protein QTL mapping identifies human plasma kallikrein as a post-translational regulator of serum uPAR levels. *FASEB J, 28*(2), 923-934. doi:10.1096/fj.13-240879

Santorico, S. A., & Hendricks, A. E. (2016). Progress in methods for rare variant association. *BMC Genet, 17 Suppl 2*, 6. doi:10.1186/s12863-015-0316-7

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature, 461*(7261), 218-223. doi:10.1038/nature08454

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J. D., Avila-

Campillo, I., Kruger, M. J., Johnson, J. M., Rohl, C. A., van Nas, A., Mehrabian, M., Drake, T. A., Lusis, A. J., Smith, R. C., Guengerich, F. P., Strom, S. C., Schuetz, E., Rushmore, T. H., & Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol, 6*(5), e107. doi:10.1371/journal.pbio.0060107

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., Absher, D., Aherrahrou, Z., Allayee, H., Altshuler, D., Anand, S. S., Andersen, K., Anderson, J. L., Ardissino, D., Ball, S. G., Balmforth, A. J., Barnes, T. A., Becker, D. M., Becker, L. C., Berger, K., Bis, J. C., Boekholdt, S. M., Boerwinkle, E., Braund, P. S., Brown, M. J., Burnett, M. S., Buysschaert, I., Carlquist, J. F., Chen, L., Cichon, S., Codd, V., Davies, R. W., Dedoussis, G., Dehghan, A., Demissie, S., Devaney, J. M., Diemert, P., Do, R., Doering, A., Eifert, S., Mokhtari, N. E., Ellis, S. G., Elosua, R., Engert, J. C., Epstein, S. E., de Faire, U., Fischer, M., Folsom, A. R., Freyer, J., Gigante, B., Girelli, D., Gretarsdottir, S., Gudnason, V., Gulcher, J. R., Halperin, E., Hammond, N., Hazen, S. L., Hofman, A., Horne, B. D., Illig, T., Iribarren, C., Jones, G. T., Jukema, J. W., Kaiser, M. A., Kaplan, L. M., Kastelein, J. J., Khaw, K. T., Knowles, J. W., Kolovou, G., Kong, A., Laaksonen, R., Lambrechts, D., Leander, K., Lettre, G., Li, M., Lieb, W., Loley, C., Lotery, A. J., Mannucci, P. M., Maouche, S., Martinelli, N., McKeown, P. P., Meisinger, C., Meitinger, T., Melander, O., Merlini, P. A., Mooser, V., Morgan, T., Mühleisen, T. W., Muhlestein, J. B., Münzel, T., Musunuru, K., Nahrstaedt, J., Nelson, C. P., Nöthen, M. M., Olivieri, O., Patel, R. S., Patterson, C. C., Peters, A., Peyvandi, F., Qu, L., Quyyumi, A. A., Rader, D. J., Rallidis, L. S., Rice, C., Rosendaal, F. R., Rubin, D., Salomaa, V., Sampietro, M. L., Sandhu, M. S., Schadt, E., Schäfer, A., Schillert, A., Schreiber, S., Schrezenmeir, J., Schwartz, S. M., Siscovick, D. S., Sivananthan, M., Sivapalaratnam, S., Smith, A., Smith, T. B., Snoep, J. D., Soranzo, N., Spertus, J. A., Stark, K., Stirrups, K., Stoll, M., Tang, W. H., Tennstedt, S., Thorgeirsson, G., Thorleifsson, G., Tomaszewski, M., Uitterlinden, A. G., van Rij, A. M., Voight, B. F., Wareham, N. J., Wells, G. A., Wichmann, H. E., Wild, P. S., Willenborg, C., Witteman, J. C., Wright, B. J., Ye, S., Zeller, T., Ziegler, A., Cambien, F., Goodall, A. H., Cupples, L. A., Quertermous, T., März, W., Hengstenberg, C., Blankenberg, S., Ouwehand, W. H., Hall, A. S., Deloukas, P., Thompson, J. R., Stefansson, K., Roberts, R., Thorsteinsdottir, U., O'Donnell, C. J., McPherson, R., Erdmann, J., Samani, N. J., Cardiogenics, & Consortium, C. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet, 43*(4), 333-338. doi:10.1038/ng.784

Semenov, A. G., Tamm, N. N., Seferian, K. R., Postnikov, A. B., Karpova, N. S., Serebryanaya, D. V., Koshkina, E. V., Krasnoselsky, M. I., & Katrukha, A. G. (2010). Processing of pro-B-type natriuretic peptide: furin and corin as candidate convertases. *Clin Chem, 56*(7), 1166-1176. doi:10.1373/clinchem.2010.143883

Sheng, J., Luo, C., Jiang, Y., Hinds, P. W., Xu, Z., & Hu, G. F. (2014). Transcription of angiogenin and ribonuclease 4 is regulated by RNA polymerase III elements and a CCCTC binding factor (CTCF)-dependent intragenic chromatin loop. *J Biol Chem, 289*(18), 12520-12534. doi:10.1074/jbc.M114.551762

Tonne, J. M., Campbell, J. M., Cataliotti, A., Ohmine, S., Thatava, T., Sakuma, T., Macheret, F., Huntley, B. K., Burnett, J. C., & Ikeda, Y. (2011). Secretion of glycosylated pro-B-type natriuretic peptide from normal cardiomyocytes. *Clin Chem, 57*(6), 864-873. doi:10.1373/clinchem.2010.157438

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics, 11*(1110), 11.10.11-11.10.33. doi:10.1002/0471250953.bi1110s43

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res, 42*(Database issue), D1001-1006. doi:10.1093/nar/gkt1229

Wilsgaard, T., Mathiesen, E. B., Patwardhan, A., Rowe, M. W., Schirmer, H., Løchen, M. L., Sudduth-Klinger, J., Hamren, S., Bønaa, K. H., & Njølstad, I. (2015). Clinically significant novel biomarkers for prediction of first ever myocardial infarction: the tromsø study. *Circ Cardiovasc Genet, 8*(2), 363-371. doi:10.1161/circgenetics.113.000630

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet, 89*(1), 82-93. doi:10.1016/j.ajhg.2011.05.029

Yuasa, I., & Umetsu, K. (1988). Genetic polymorphism of human alpha 2HS-glycoprotein: characterization and application to forensic hemogenetics. *Electrophoresis, 9*(8), 404-410. doi:10.1002/elps.1150090810

Zhu, C., Odeberg, J., Hamsten, A., & Eriksson, P. (2006). Allele-specific MMP-3 transcription under in vivo conditions. *Biochem Biophys Res Commun, 348*(3), 1150-1156. doi:10.1016/j.bbrc.2006.07.174

Chapter 3: Identification of common and rare genetic variation associated with plasma protein levels using whole exome sequencing and mass spectrometry

Chapter 3.1 Abstract

Background: Identifying genetic variation associated with plasma protein levels, and the mechanisms by which they act, could provide insight into alterable processes involved in regulation of protein levels. Genome sequencing has enabled the interrogation of common and rare genetic variants that affect protein levels, and advances in protein quantification (i.e. mass spectrometry) could reduce bias during protein quantification, and allow for the delineation of true associations from technical artifacts. Combining these techniques could enable the identification of common and rare genetic variation associated with plasma protein levels.

Methods and Results: We utilized TMT-mass spectrometry to measure the levels of 664 proteins in blood plasma from 165 participants of the Tromsø Study. Integrating whole exome sequencing data, we identified 110 independent, significant associations between common and rare genetic variation with peptide and protein levels. We then leveraged genotype data to identify technical artifacts, and excluded 50 of these associations. We describe rare variation associated with the complement pathway and platelet degranulation. We then use literature and database searches to identify putative functional variants for each pQTL, and show that, pQTLs act through diverse molecular mechanisms that affect both RNA and protein metabolism.

Conclusions: We show that, while the majority of pQTLs exert their effects by modulating a gene's RNA, many affect protein levels directly. Our work demonstrates the extent by which pQTL studies are affected by technical artifacts, and highlights how

identifying the functional variant in pQTL studies can lead to insights into the molecular steps by which the protein is regulated.

Chapter 3.2 Introduction

Blood plasma is comprised of proteins generated from cells involved in diverse processes including thrombosis, hemostasis, immunity, and hematopoiesis. As it contains proteins from a wide variety of cells, blood plasma is a source for many potential biomarkers (Jacobs et al., 2005), which if causally related to disease, may provide novel drug targets (Ong et al., 2009). Genetic variation that affects proteins can be used to assess the casual relationship between a particular biomarker and disease (Burgess, Timpson, Ebrahim, & Davey Smith, 2015), and the molecular function of the variant can provide insight into processes important to the protein's abundance. In particular, rare variation has proved to be an effective route to identifying drug targets (Cohen, Boerwinkle, Mosley, & Hobbs, 2006) as rare variants can have larger, wide reaching effects. By examining the effects of variants associated with a particular protein on the levels of other proteins (MacKeigan et al., 2003), it may be possible to identify the pathways the protein is involved in. Additionally, examining whether these genetic variants act by modulating RNA or protein levels, and identifying the specific molecular mechanisms by which they act, could provide insight into alterable processes involved in regulation of protein levels, thus elucidating insights into targeted therapeutics.

Recent advances in protein and genotype measurement, have enabled the interrogation of genetic variants that affect protein levels (protein quantitative trait loci, pQTLs) (Johansson et al., 2013; Kim et al., 2013; Liu et al., 2015; Melzer et al., 2008).

This advance has resulted in the identification of hundreds of plasma pQTLs in human samples, and is leading to insights into the proteomic consequences of risk for cardiovascular disease (Folkersen et al., 2017; Suhre et al., 2017). Previous pQTLs studies (Johansson et al., 2013; Kim et al., 2013; Liu et al., 2015; Melzer et al., 2008; Suhre et al., 2017), however, have been limited by utilizing assays which do not measure an entire protein (i.e. aptamer or antibody methods that only measure a single epitope), or by the range of genetic variation that they test (genotype arrays vs. sequencing). Additionally, these protein measurement methods can be affected by the presence of genetic variation that does not alter the protein's level, but rather affects the assay's quantification ability. By measuring the levels of multiple peptides in a protein through assays such as mass spectrometry, and identifying genotype data that includes complete coding information through exome sequencing, it could be possible to exclude pQTLs that are driven by artefactual associations and better identify potential underlying causal variants.

In this study, we utilized TMT-mass spectrometry to measure plasma levels of 664 proteins in 165 participants of the Tromsø Study who have high depth exome sequence data available. We identified 110 independent, significant associations between common and rare genetic variation with peptide and protein levels. Our subsequent analyses determined that while 60 of these were true associations, 50 were driven by previously unreported technical artifacts associated with the presence of genetic variation in coding exons. We examined common and rare associations for downstream effects on other proteins, and identified associations affecting the complement pathway and platelet degranulation. Using a combination of literature and database annotations, we identified

and described putative functional variants for each locus. We show that approximately half of the pQTLs could be explained by variants previously experimentally shown to influence the associated protein's level. Causal variants most often affected RNA metabolism, however, many affected protein metabolism and would therefore not be detected in studies that solely examined gene expression. These results illustrate the potential for pQTL studies to characterize the effects of rare variation, and highlight a need for high throughput studies of protein levels to take into account technical artifacts caused by exonic genetic variation.

Chapter 3.3 Methods

Chapter 3.3.1 Data Sharing

The Regional Committee of Medical and Health Research Ethics in North Norway approved this study, and all subjects gave their informed written consent to participate. The whole exome sequence and mass spectrometry data described in this study will not be made available, as the consent signed by the study participants does not allow the public release of these data.

Chapter 3.3.2 The Tromsø Study

The Tromsø Study (Jacobsen, Eggen, Mathiesen, Wilsgaard, & Njolstad, 2012) is a single-center, population-based cohort study of the inhabitants of Tromsø, Norway. 27,158 individuals participated in the fourth survey of the Tromsø Study between 1994-1995; baseline characteristics were collected using self-reported questionnaires, physical

examinations, and blood samples. Non-fasting blood was drawn from an antecubital vein to gather plasma and whole blood. Plasma was collected in 5ml vacutainer tubes containing EDTA as an anticoagulant, processed within 1 hour by centrifugation at 3,000g for 10 min, and collected and frozen at -70^{o}C. Whole blood was used to prepare archive quality DNA, and was stored at the HUNT Biobank in Levanger, Norway. In the immediate 4-12 weeks following the initial visit, 7,965 participants were invited for a follow-up for a more in-depth examination and additional blood sampling.

All 27,158 participants were followed from the date of enrollment through December 31, 2012. All cohort members that experienced an incident venous thromboembolism (VTE) during the study period were identified by searching the hospital discharge diagnosis registry, the autopsy registry, and the radiology procedure registry at the University Hospital of North Norway, the sole hospital in the Tromsø municipality (Braekkan et al., 2008). The VTE events were thoroughly validated by review of medical records as previously described in detail (Braekkan et al., 2008). Out of the 710 incident VTE cases that were identified, 100 cases were sampled for this study such that the time of blood collection occurred prior to incident VTE (range of time to VTE: 1 month to 7 years; average: 3.72 years). For each case, a paired control, matched on age and sex, was randomly sampled from the cohort.

Chapter 3.3.3 Sample Preparation and Mass Spectrometry

Plasma samples for this study were analyzed through TMT-multiplexed mass spectrometry by Proteome Sciences (London, England). Samples were visually inspected

for hemolysis, protein concentrations were determined using the Bradford assay, and samples were visualized on Coomassie stained SDS-PAGE 4-20% gradient gels. 25µL of each sample underwent albumin and IgG depletion using Qproteome Spin Columns from Qiagen (Hilden, Germany), with 100mM triethylammonium bicarbonate (TEAB) substituted for the buffer provided in the kit. Protein concentration was measured using a Bradford assay, and 17 samples were visualized on Coomassie stained SDS-PAGE 4-20% gradient gels for quality control purposes. Samples were run as 25 separate TMT10-plexes, with each 10-plex including: 1) four cases, 2) the respective four age- and sex-matched controls, and 3) two reference pools comprised of equal portions of all 100 cases or 100 controls, respectively. Specifically, 60 µg of protein from each depleted sample were brought to 0.1% SDS in 100 mM TEAB, reduced using tris(2-carboxyethyl)phosphine, alkylated with iodoacetamide, and trypsin digested to produce peptides. Peptides were then mixed with their respective TMT10-plex reagent (Thermo Fisher, Massachusetts, USA) and the reaction was terminated using hydroxylamine. Samples were pooled into their TMT10-plexes and diluted to an acetonitrile concentration of less than 5% before being purified via Oasis HLB cartridges. 300 µg of each TMT10-plex was fractionated into 8 fractions using HPLC (Waters Alliance 2695), desalted on Oasis HLB cartridges, and dried. Each fraction was run in duplicate for LC-MS/MS using an EASY-nLC 1000 system coupled to an Orbitrap Fusion Tribrid Mass Spectrometer (Thermo Fisher). Resuspended peptides were loaded onto a nanoViper C18 Acclaim PepMap 100 pre-column (Thermo Scientific), and resolved using an increasing gradient of 0.1% Formic acid in ACN through a 50 cm PepMap RSLC analytical column at a flow rate of 200 nL/min. Peptide mass spectra were acquired throughout the entire

chromatographic run (180 minutes), with FTMS scans at 120,000 resolving power at 400 m/z followed by a top 10 high collision induced dissociation (CID) method for FTMS2 scans at 30,000 resolving power at 400 m/z. For quantification, synchronous precursor selection was enabled, MS2 peaks were fragmented with higher energy collisional dissociation (HCD), and the TMT reporter ions were measured at 30,000 resolving power in the Orbitrap.

For initial quality control analysis, peptides and proteins were identified using Proteome Discoverer (PD) v1.4 (Thermo Scientific). The 400 raw data files (200 samples, each performed in duplicate) were submitted to PD v1.4 using the Spectrum Files node. Spectrum selector was set to default, while SEQUEST HT was set to search against the human FASTA UniProt-KB/Swiss-Prot database (August 2015). Spectra were identified in PD with the settings: 1% FDR; one Rank 1 peptide per protein. Processing, normalization, and filtering were done by Proteome Science using their in-house software. To identify batch effects, samples were hierarchically clustered using Spearman's correlations. It was observed that participants with plasma samples from the initial sample donation clustered together (N=176), while participants with plasma samples from the follow-up visit clustered together (N=24) (Jensen). The top proteins that had differential levels of expression between the two clusters were associated with blood clotting in a gene ontology analysis, suggesting the batch effect could be due to variation in plasma preparation between the first and second visit; thus, these 24 participants were removed from any further analysis.

Chapter 3.3.4 Peptide and Protein Identification

Peptide identification was performed on the 176 plasma samples collected from participants in the initial visit using Proteome Discoverer (PD) v2.1 (Thermo Fisher Scientific). MS2 data were searched against Gencode 19 (corresponding to Ensembl 75) (Harrow et al., 2012) and mapped to GRCh37 using the Sequest algorithm (Eng, McCormack, & Yates, 1994). A decoy search was also conducted with sequences in reverse order (Elias & Gygi, 2007). For the search, a precursor mass tolerance of 50 ppm was specified, and a 0.6 Da tolerance for MS2 fragments was specified. Static modifications of TMT10-plex tags on lysines, peptide n-termini (+299.162932 Da), and carbamidomethylation of cysteines (+57.02146 Da) were specified. Variable oxidation of methionine (+15.99492) was also specified in the search parameters. Data were filtered to 1% peptide and protein level false discovery rates using percolator (Kall, Canterbury, Weston, Noble, & MacCoss, 2007; Spivak, Weston, Bottou, Kall, & Noble, 2009).

TMT reporter ion intensities were extracted from MS3 spectra for quantitative analysis, and signal-to-noise ratios were used for quantification. Spectra were filtered out if they had either above 25% isolation interference, or an average signal-to-noise ratio across samples in a TMTplex of less than 10. Protein level quantification values were calculated by summing signal-to-noise ratios for all remaining peptides belonging to a given protein. Data were first normalized in a multi-step process as previously described (Lapek, Lewinski, Wozniak, Guatelli, & Gonzalez, 2017), following which they were quantile normalized to a standard normal distribution. In summary, for each sample, peptide levels were calculated as the normalized signal-to-noise ratios for each peptide,

and protein levels were calculated as the normalized sum of signal-to-noise ratios for all of the peptides belonging to the protein.

Chapter 3.3.5 Variant Identification and Annotation

Of the 176 participants who donated plasma during the first visit, 165 had genotype data available from whole exome sequencing generated as part of an ongoing study of the genetics of VTE (Carson et al., 2014). The samples used in this study were sequenced using the Agilent SureSelect 50Mb capture kit and the Illumina TruSeq paired-end 100bp cluster kit to an approximate depth of 100X on an Illumina HiSeq 2000. As previously described for 39 of these 165 exomes (Solomon et al., 2016), sequence reads were mapped to the reference human genome (hg19) using BWA (Li & Durbin, 2009) (version 0.7.10-r789) with default parameters, and processed using Picard (version 1.115) (http://broadinstitute.github.io/picard) and GATK (Van der Auwera et al., 2013) (version 3.3-0). Using the information from both on and off-target reads (Pasaniuc et al., 2012) from the sequencing data, genotypes were imputed to the whole genome using Beagle (Browning & Browning, 2016) (version 4.0, r1398) with reference haplotypes from the unrelated individuals in the European (EUR) and East Asian (EAS) superpopulations of the 1000 Genomes Project (Abecasis et al., 2012) Phase 3, at sites with a combined MAF > 1%. To obtain the final genotypes for this study, we: 1) preferentially used genotypes that had a call rate >90%, and 2) used imputation genotypes with a Beagle QC threshold allelic $r^2 > 0.7$. Finally, variants with a Hardy-Weinburg equilibrium p-value > $1x10^{-7}$ (as calculated in VCFtools (Danecek et al., 2011)) were included in the analysis. Variants were annotated using SNPEff v4.1 (Cingolani et al., 2012) and the highest impact

annotation reported for the canonical transcript was chosen. These annotations were then manually collapsed into larger categories (eg. stop-gain, stop-lost, and stop retained variant were all grouped into stop site) based on the first annotation listed per variant. Annotations of missense, synonymous, non-coding exonic, splice site, frameshift, stop site, and start site were considered to be exonic variants, while annotations of 3'UTR and 5' UTR were considered UTR variants.

Chapter 3.3.6 Genetic Associations

Associations between common genetic variants (MAF > 1%) and protein or peptide levels were calculated using EMMAX (H. M. Kang et al., 2010) from the EPACTS software package (Hyuan Min Kang, 2014). EMMAX is a linear mixed model which accounts for family relatedness and population stratification by including a kinship matrix. For common *cis* associations, genetic variants within 200kb +/- of the gene start and stop were tested for association with the protein encoded by that gene. While 8 subjects with peptide/protein quantifications and genotypes were required to obtain association statistics, significant associations were only observed with >70 measurements. Additionally, we modeled age, sex, BMI, smoking status, cancer status at the time of sample collection, VTE case-control status, and the TMT-multiplex experiment as covariates. Associations were considered significant if they had a peptide p-value less than 1.91×10^{-8} (0.05/ (466 variants per locus on average * 5608 peptides)) or a protein p-value less than 1.62×10^{-7} (0.05 / (466 variants per locus on average * 664 proteins)). For common *trans* associations, all genetic variants were tested against each peptide and protein level. Rare genetic variation (MAF < 5%) association was calculated using SKAT-

O (Lee, Wu, & Lin, 2012) from the EPACTS software package. SKAT-O collapses rare variants within a specified interval and performs both a burden test, and a kernel association test. Rare variants were collapsed using three methods: 1) **MAF < 5%**: all variants within the interval from 2kb upstream of the protein-coding gene to the transcription end of the gene; 2) **Deleterious**: all MAF <5% variants that were annotated using SNPEff v4.1 (Cingolani et al., 2012) as having a high or moderate effect impact; and 3) **CADD-score**: all MAF <5% variants with a PHRED-scaled CADD (Kircher et al., 2014) score greater than 10. For rare *cis* associations, rare variants were collapsed using all three methods, and tested for association with the level of the protein encoded by that gene. Associations were considered significant if they had a peptide p-value less than 2.97×10^{-6} (0.05 / (5608 peptides x 1 gene x 3 methods)) or a protein p-value less than 2.51×10^{-5} (0.05 / (664 proteins x 1 gene x 3 methods)). For common *trans* associations, associations were considered significant if they had a peptide p-value less than 8.91×10^{-12} (5×10^{-8} / 5608 peptides) or a protein p-value less than 7.53×10^{-11} (5×10^{-8} / 664 proteins). For rare *trans* associations, rare variants were collapsed using all three methods for every gene that encodes one of the parent proteins, and were tested for association (Bonferroni corrected $p < 0.05$) with all peptide and protein levels. Associations were considered significant if they had a peptide p-value less than 4.02×10^{-8} (0.05 / (5608 peptides x 654 genes x 3 methods)) or a protein p-value less than 3.78×10^{-8} (0.05 / (664 proteins x 654 genes x 3 methods)). Multiple independent associations occurring at a locus were identified by repeating the analysis with the top variant as a covariate. This process was repeated, including all independent associations as covariates, until no new significant associations were identified.

Chapter 3.3.7 Identification of pQTLs that were Technical Artifacts

Current mass spectrometry quantification techniques are limited to searching databases of known peptides (Wang & Zhang, 2013) that do not always contain alternate peptide sequences due to genetic variation. Therefore, genetic variants that alter peptide sequences could result in spectra that no longer match the database, and their absence could result in a false pQTL association. To identify pQTLs that were likely technical artifacts, we examined the exome sequence data to determine if the most strongly associated pQTL variant (sentinel variant), or variants in LD with the sentinel variant, resulted in a missense amino acid change in or near the associated peptide in our population. We then examined the impact of the missense variant and classified the artifact into three groups: homologue, digestion, or missense (Supplemental Table 3.1). If the missense amino acid change resulted in a peptide that was identical to a homologous protein through a BLASTp (Altschul, Gish, Miller, Myers, & Lipman, 1990) search, it was classified as a homologue artifact. If the missense amino acid flanked or fell within the digestion site of the peptide (1-4 amino acids from the peptide), it was classified as a digestion artifact. The remaining missense variants that disrupted the peptide itself were classified as missense artifacts.

Chapter 3.3.8 Putative Functional Variant Identification

To identify putative functional variants (PFVs) at each pQTL, we examined databases and published literature to find established research that linked protein levels to a variant in linkage disequilibrium (LD) with the sentinel variant through a specific

proposed or validated molecular mechanism (Supplemental Figure 3.1). For each PFV, we categorized the supporting evidence into three categories, which were based on whether the published evidence was experimentally validated through biological assays, associated statistically, or predicted based on functional characteristics (e.g. the variant was located in a promoter region):

1. known - the PFV had been experimentally shown to affect the parent protein's level through a particular molecular mechanism (Supplemental Figure 3.1A)
2. likely – either the PFV had been experimentally validated to have a molecular mechanism that was predicted to affect the parent protein's level, or the PFV was predicted to have a mechanism that had been experimentally shown to affect the parent protein's level (Supplemental Figure 3.1B).
3. suggestive - either the PFV was predicted to act through a specific molecular mechanism, and the PFV had been associated with the parent protein's level; the PFV was associated with the parent protein's level and a molecular mechanism had been previously predicted to affect the parent protein's level; or the PFV was predicted to act through a molecular mechanism, and the proposed mechanism had been associated with the parent protein's level (Supplemental Figure 3.1C)

To obtain this information, we performed sequential database and literature searches. First, we identified all variants in LD r^2>0.2 with the lead variant via HaploReg (version 4.1) (Ward & Kellis, 2012). We then examined whether any of these variants were documented in OMIM (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005), and whether the missense variants had been previously reported as associated with the

protein level (searching for the dbSNP identifier, the protein name, and the phrase “level”). For each protein, if there was at least one variant with mechanistic supporting evidence meeting “known” or “likely” criteria, we chose the variant with the most experimental evidence and labeled it the PFV. If no PFVs were identified for the protein, we next examined all missense, synonymous, 5’ UTR, and 3’ UTR variants. Using information from OMIM, genomic annotations from the UCSC Genome Browser (Kent et al., 2002), protein annotations from UniProt (The UniProt, 2017), and information regarding the position of the associated peptides relative to the protein from this study, we identified potential mechanisms by which the variant might act. We then performed literature searches of the dbSNP identifier, the protein name, and the potential mechanisms (alternative splicing, isoform, glycosylation, degradation, miRNA, promoter, enhancer, gene expression, maturation, cleavage, protein stability, or protein folding). Literature suggesting an established association was further investigated to determine the strength of the evidence. Within each protein, we labeled the variant with the best supporting evidence as the PFV, and categorized the strength of the evidence according to the three categories above. In the case where little or no mechanistic evidence had been established we labelled the sentinel variant as the PFV, and created a new category for this lack of evidence: “unknown”. We then annotated the PFVs with their variant type (e.g. missense, intergenic) using SnpEff v4.1 (Cingolani et al., 2012), and determined if the PFV was an expression quantitative trait locus (eQTL) in the GTEx (Consortium, 2013) Portal (accessed 03/08/2018). An eQTL was considered to be in the “same” tissue if it was identified in the tissue with the highest expression level of the associated gene, and in an “other” tissue if the eQTL was identified in a different tissue. Finally, the mechanism

of each PFV was classified based on the evidence gathered above into the following categories: 1) affecting RNA metabolism (promoter, isoform expression, nonsense-mediated decay, gene deletion, and miRNA processing), or 2) affecting protein metabolism (protein degradation, glycosylation, and secretion).

Chapter 3.4 Results

Chapter 3.4.1 Data Generation

We examined peptide and protein levels from plasma, and genotype data from whole exome sequencing of blood DNA, from 165 individuals from the Tromsø Study (Figure 3.1A). These individuals and data were part of an effort to identify predictive biomarkers for venous thromboembolism (Jensen et al., in preparation). To assess peptide and protein levels, we performed TMT-multiplexed mass spectrometry on blood plasma, identifying 5,608 peptides, corresponding to 664 proteins and 655 genes. Of the 5,608 peptides, 1,430 (25%) were present in all samples, 3,394 (61%) were identified in at least 50% (82 individuals), and 5,052 were identified in at least 5% (N=8, the minimum number required to perform genetic analysis) (Figure 3.2A). The identified peptides had an average length of 14.5 amino acids (range: 6 to 43) (Supplemental Figure 3.2A). We observed an average of 8.5 peptides mapping to each protein (range: 1 to 291) (Figure 3.2B); protein levels were calculated by summing these peptide measures. The functions of the proteins that were measured were consistent with their role in plasma, with the most enriched pathways (Reactome (Fabregat et al., 2018) pathway analysis FDR < 0.05) including the immune system and hemostasis (Figure 3.2C & Supplemental Figure 3.2B). From the whole exome sequencing data, we identified 501,682 genetic variants directly,

and an additional 2,647,181 variants through imputation. Of the 3,148,863 total variants, 2,624,979 were evaluated in common variant analyses (minor allele frequency (MAF) ≥ 1%), and 1,690,437 were evaluated in rare variant analyses (MAF <5%) (Figure 3.2D). While most variants were noncoding (intergenic or intronic) regions, a total of 182,828 (5.8%) were located in UTR and exonic regions (Figure 3.2E). Overall, these analyses generated information on 664 proteins and 3,148,863 variants for genetic association analyses.

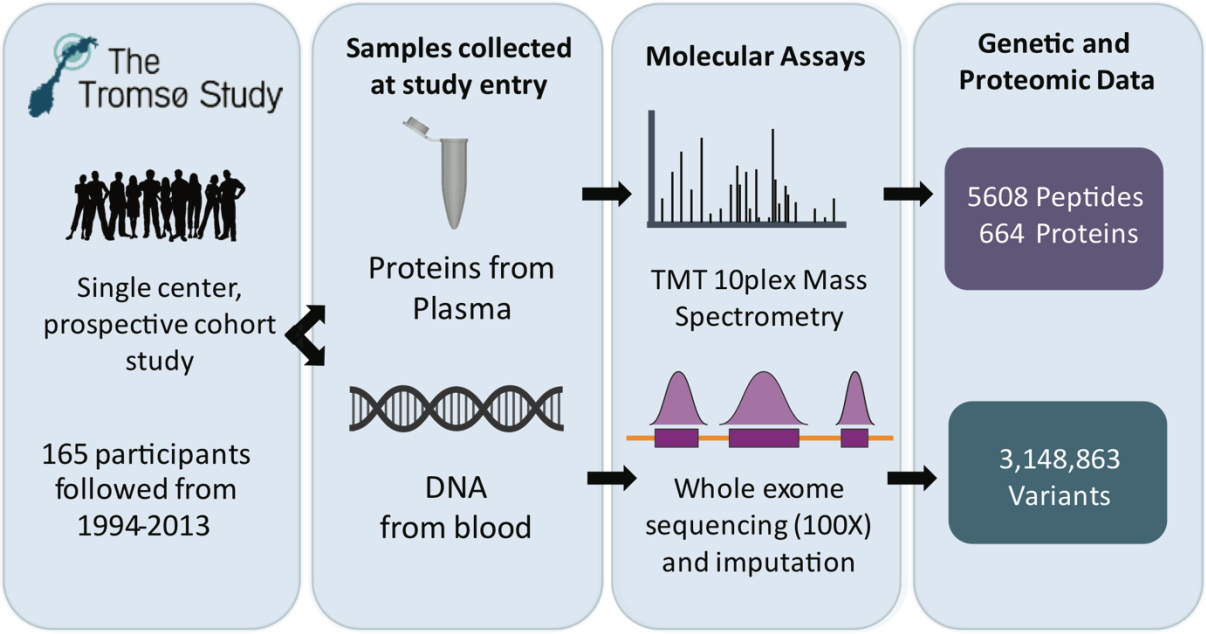


Figure 3.1 Study overview.
165 individuals from The Tromsø Study were followed from 1994-2013. Between 1994 and 1995, blood plasma and whole blood were collected; blood plasma and whole blood were processed and subsequently used for protein quantification by mass spectrometry and whole exome sequencing, respectively. These analyses identified 5,608 peptides and 664 proteins from plasma, and 3,148,863 variants from whole blood, across all individuals.

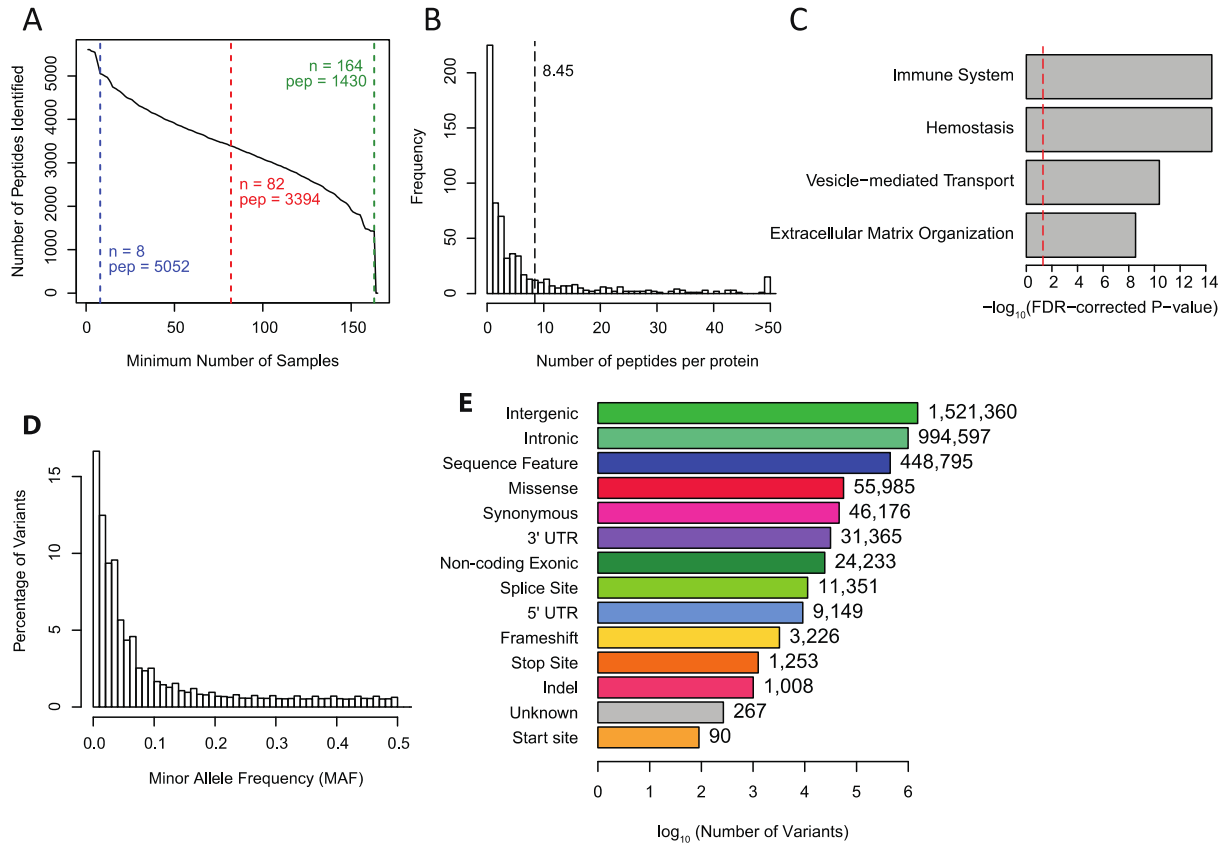


Figure 3.2 Description of protein and genotype data
(A) Cumulative distribution plot showing the number of peptides identified in at least N samples. 5,052 peptides were identified in at least 8 samples (blue), 3,394 peptides were identified in at least 82 samples (red), and 1,430 peptides were identified in all 165 samples (green). (B) Histogram showing the number of peptides identified for each of the 664 parent proteins. A mean of 8.45 peptides per parent protein were identified (dotted line). (C) Bar plot showing the q-values from Reactome pathway analysis that were enriched for plasma proteins. The significance threshold of –log10(0.05) is shown by the red dotted line. (D) Histogram of the minor allele frequencies in this study for all 3,148,863 genetic variants identified across individuals. (E) Bar plot of the number of identified genetic variants within each SnpEff annotation. The number of variants with each annotation is also listed next to each bar.

Chapter 3.4.2 Identification of Peptide and Protein *cis* pQTLs

We first identified *cis* pQTLs, i.e. those located near the gene encoding the plasma peptide and/or protein. We identified all variation within +/- 200 kb of the corresponding gene for each of the 5,608 peptides and 664 proteins. We tested for association between

genetic variants and peptide or protein levels using EMMAX, a linear mixed model that includes a kinship matrix to account for population structure and family relatedness. Additionally, we modeled age, sex, BMI, smoking status, cancer status at the time of sample collection, VTE case-control status, and the TMT-multiplex experiment as covariates (see Methods). We identified 148 peptides and 31 proteins with significant associations (Bonferroni adjusted $p < 0.05$) with 80 and 31 *cis* genetic variants, respectively. Next, we identified additional independent significant pQTLs for each of the 148 peptides and 31 proteins by performing a step-wise analysis conditioned on the most significant variant, and found six peptides and two proteins that had a second *cis* genetic variant. In total, we identified 33 pQTLs associated with the levels of 31 proteins and 154 pQTLs associated with the levels of 148 peptides (Supplemental Table 3.1).

Chapter 3.4.3 Integration of Peptide and Protein pQTLs

As we expected that the peptide pQTLs would also be protein QTLs for the parent protein, we investigated if differences between peptide and protein pQTLs could reflect technical artifacts introduced by genetic variants affecting the quantification process/pipeline. To examine the concordance between peptide and protein pQTLs, we determined the parent protein for all 154 peptide pQTLs and 33 protein pQTLs. We identified 67 unique parent proteins, of which 24 were associated with both a peptide pQTL and protein pQTL, 36 were only associated with peptide pQTL(s), and 7 were only associated with protein pQTL(s). For the 24 parent proteins with both peptide and protein pQTLs, we identified independent pQTL signals by examining whether the variants were the same or in linkage disequilibrium (LD; $r^2 > 0.2$). We created three classifications for

each independent pQTL: 1) those only associated with peptide levels (peptide-only pQTL), 2) those only associated with protein levels (protein-only pQTL), or 3) those associated with both peptide and protein levels (both pQTL). From this process, we obtained 91 independent pQTLs: 58 peptide-only pQTLs (43 parent proteins), 10 protein-only pQTLs, and 23 both pQTLs (22 parent proteins) (Supplemental Table 3.1). Using the exome sequencing data, we examined whether the peptides that were associated with the pQTLs (either directly or indirectly through LD with a polymorphic variant) affected the quantification process either by: 1) altering the sequence of the peptide, 2) altering the effectiveness of the trypsin digestion site, or 3) resulting in the association with a homologous protein (rather than the original parent protein). In total, 43 of the 91 independent pQTLs affected the quantification process by one of these three mechanisms and appeared to be technical artifacts. The majority of artifact pQTLs were peptide-only pQTLs (39 of the 43), however, we also found one protein-only pQTL, and three both pQTLs, to be technical artifacts. After removing these technical artifacts, the resulting data set had 48 independent associations: 9 protein-only pQTLs, 19 peptide-only pQTLs, and 20 both pQTLs. Of note, 32 of these associations were novel pQTLs (Johansson et al., 2013; Kim et al., 2013; Liu et al., 2015; Lourdusamy et al., 2012; Suhre et al., 2017; Sun et al., 2017) (Supplementary Table 3.1).

Chapter 3.4.4 Collapsing Variants to Identify Rare-variant cis pQTLs

To identify rare variants associated with protein levels, we tested the cumulative effects of sets of rare variants on peptide and protein levels. We collapsed rare variants using three different criteria: 1) MAF < 5%: all variants within the interval from 2kb

upstream of the protein-coding gene to the transcription end of the gene with a minor allele frequency <5%; 2) Deleterious: all MAF <5% variants that were annotated using SNPEff(Cingolani et al., 2012) as having an effect impact of high or moderate; and 3) CADD-score: all MAF <5% variants that have a PHRED-scaled CADD (Kircher et al., 2014) score greater than 10. For each peptide or protein measured, after collapsing the rare *cis* variants for their corresponding gene, we identified associations between the peptide, or protein, and rare variation using the optimal unified test SKAT-O (Lee et al., 2012) that combines a kernel test with a burden test. We identified 16 rare *cis* pQTLs (12 associations with peptides and 4 associations with proteins), of which 10 were independent: 6 peptide-only, 2 protein-only, and 2 both rare pQTLs (Supplemental Table 3.2). As with common variation, we examined the associations for technical artifacts, and found that all of the peptide-only pQTLs overlapped a rare missense mutation; they were therefore excluded. As the threshold used for identifying common variation was MAF > 1%, some variants were included in both common and rare tests; we removed these associations, resulting in a total of 3 independent rare *cis* pQTLs, of which 2 were previously reported (Brantly, Courtney, & Crystal, 1988; Stengaard-Pedersen et al., 2003). Thus, while genetic variation was associated with substantial artefactual pQTLs in *cis* rare variant analysis, the associations identified after filtering corresponded to established protein-level associations.

Chapter 3.4.5 Trans Associations

To identify downstream targets and pathways associated with pQTLs, and gain insight into the functional mechanism of the identified pQTLs, we tested for association in

trans. We first tested all 2.6 million variants with MAF >1% genome wide for association (*trans* pQTLs) with each of the 5,608 peptides and 664 proteins; this method did not find any *trans* pQTLs at genome-wide significance (peptide $P < 8.91 \times 10^{-12}$; protein $P < 7.54 \times 10^{-11}$). To increase our power, at each of the 655 loci encoding the measured proteins of this study, we performed association analyses using each of the three rare collapsing criteria to identify *trans* association with any of the peptides or proteins encoded at the other 654 loci. We identified 9 associations between rare variation and peptide levels (i.e. rare peptide-only *trans*-pQTLs) (Supplemental Table 3.3). One of the associations was a rare peptide-only *trans*-QTL between variation in *FCN3,* and levels of a peptide in the complement component C8 beta chain (C8B). FCN3 is an activator of the lectin complement pathway, and its pathway includes C8 in its final stages (Garred, Honore, Ma, Munthe-Fog, & Hummelshoj, 2009). Notably, this variation was just below the significance threshold for being a rare peptide-only *cis* QTL for *FCN3* (Figure 3.3A). We therefore examined the full established pathway of the lectin complement (Garred et al., 2009). We observed that rare variation in *FCN3* was associated with 8 other members of the lectin complement pathway at a nominal $P < 0.05$: C4a, C4b, C4BPa, C5, C6, C8b, C8a, and C8g (Supplemental Table 3.4), suggesting that the rare variation in *FCN3* was broadly associated with the levels of proteins in the complement pathway. We next examined the other rare *trans* pQTLs, and identified five loci associated with levels of SERPINA1 (alpha-1-antitrypsin): *CD109*, *CFL1*, *CLU*, *HYOU1*, and *RARRES2* (Figure 3.3B). Of the five genes, four encode proteins involved in platelet degranulation (Reactome (Fabregat et al., 2018) enrichment FDR = 7.2×10^{-6}). As alpha-1-antitrypsin is secreted into the plasma via platelet degranulation, these results suggest that rare

variation in proteins associated with platelet degranulation could be important modulators of alpha-1-antitrypsin levels. The fifth gene, *HYOU1,* has not been implicated in platelet degranulation, but is upregulated in response to hypoxia (Schofield & Ratcliffe, 2004), an important risk factor for blood clotting (Reitsma, Versteeg, & Middeldorp, 2012). Overall, these results suggest that rare variation in proteins can be associated with protein levels of downstream targets.

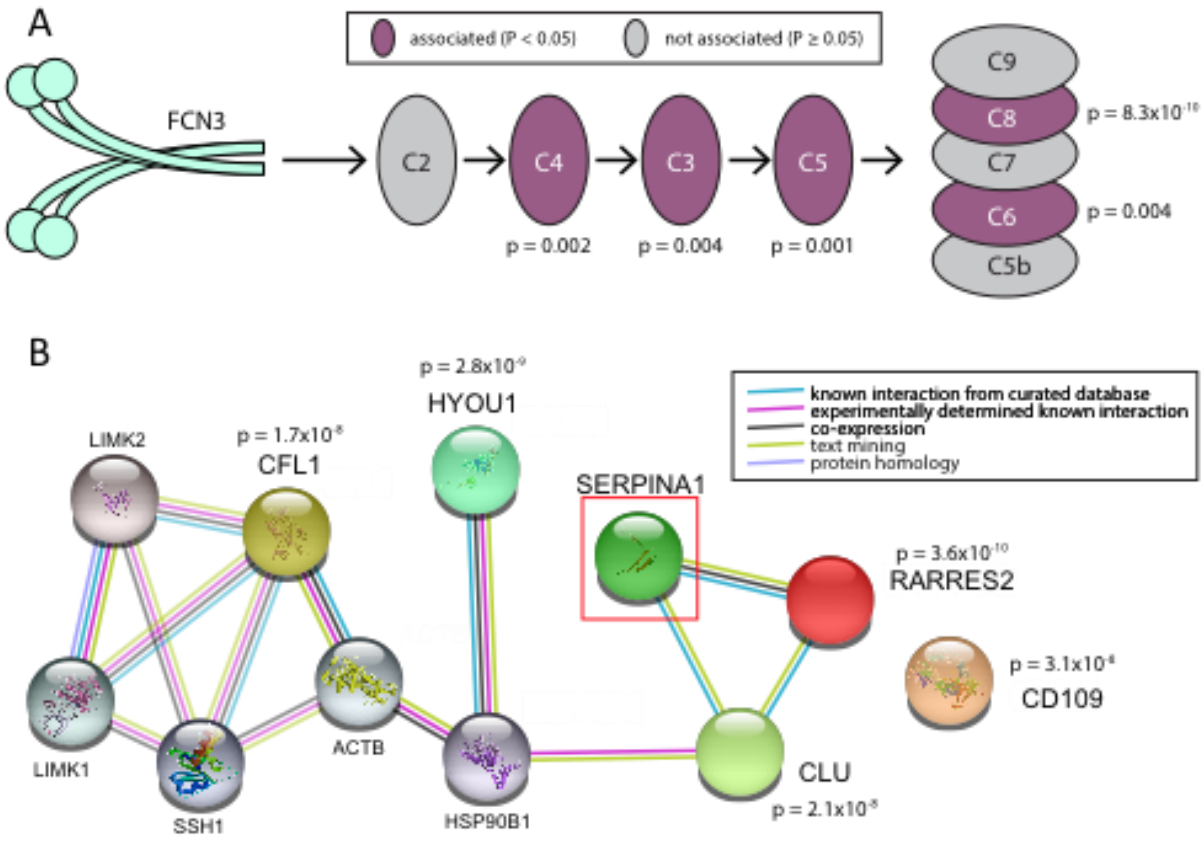


Figure 3.3 Pathways identified from rare variation analyses
(A) An overview of the lectin complement pathway showing the relationship between FCN3 (Ficolin 3; teal) and the complement pathway. Nominal p-values are shown for the association between rare variation at the FCN3 locus and levels of the complement pathway proteins. C4, C3, C5, C8, and C6 were associated at P < 0.05 (purple), C2, C9, C7, or C5b were not associated (gray). (B) STRING database diagram of the five proteins associated with rare SERPINA1 variation (each labeled with their nominal association p-value). Connections between proteins are colored based on their evidence (see legend and STRING documentation).

Chapter 3.4.6 Identifying Putative Functional Variants

Due to linkage disequilibrium (LD), the most strongly associated variant (sentinel variant) may not be the causal variant. To enable the examination of the distribution of functional mechanisms underlying the common pQTL associations, it is therefore necessary to examine variants in LD with the sentinel variant to identify the variants that could be driving the association (putative functional variants (PFVs)). Across all pQTLs, we observed an average of 151 variants in LD with the sentinel variant. Next, using a combination of database and literature searches, we identified candidate variants at each pQTL locus (Figure 3.4A; see methods). We categorized the strength of published evidence supporting a specific proposed or validated molecular mechanism according to four categories ordered by strength: 1) known; 2) likely; 3) suggestive; or 4) unknown (see methods). We selected the PFV at each locus as the variant with the strongest functional evidence (Supplemental Table 3.5). In total, we found 18 known, 5 likely, 5 suggestive, and 20 unknown PFVs; notably, 14 of the 23 PFVs with at least known or likely evidence were not the sentinel variant. Additionally, while a large proportion of the sentinel variants were intronic, the PFV annotations showed a redistribution to intergenic and coding annotations (Supplemental Figure 3.3A). Thus, approximately half of the pQTLs could be explained by variants previously experimentally shown to influence the associated protein's level.

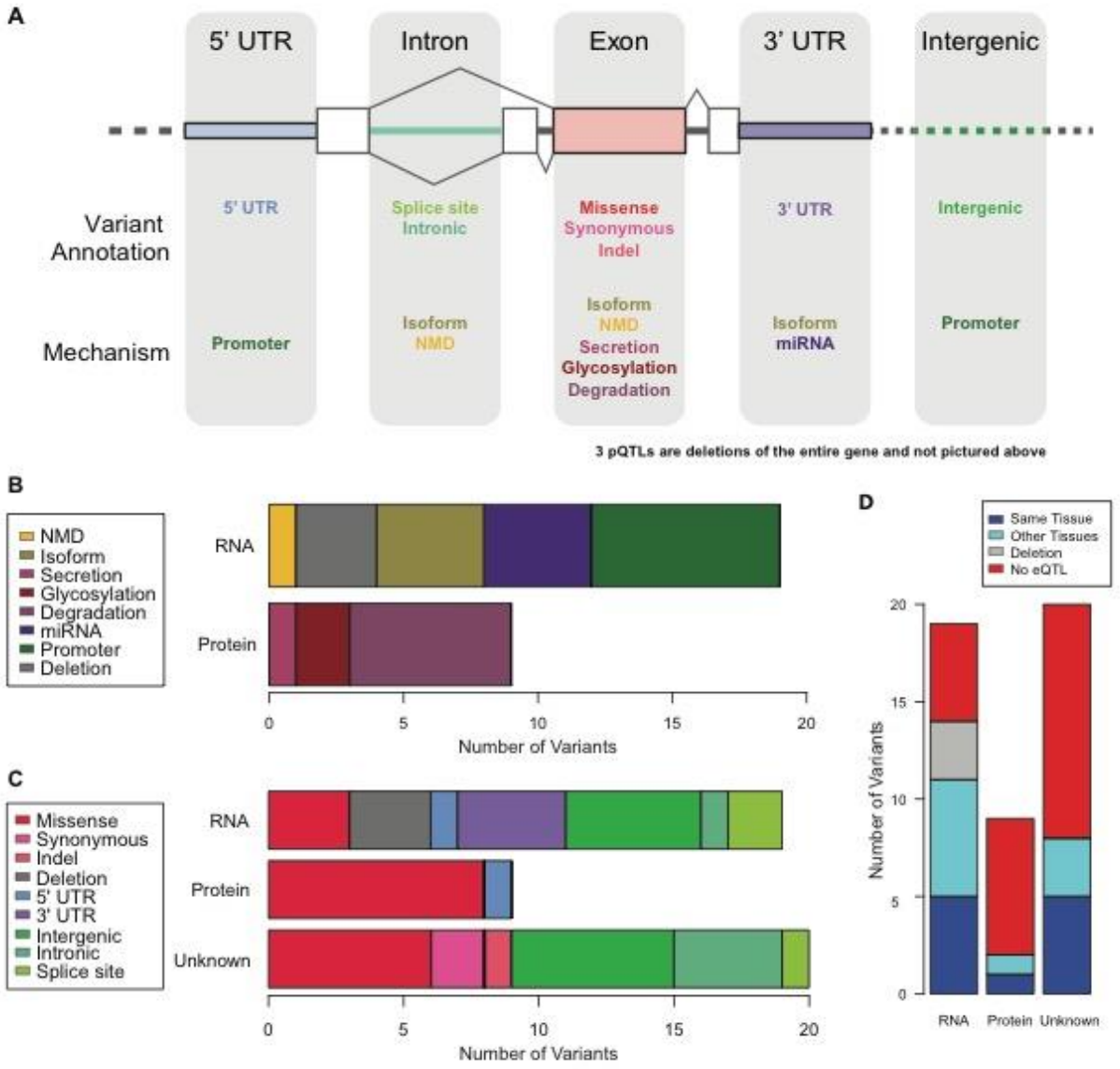


Figure 3.4 Putative functional variant analyses
(A) Cartoon illustrating the genomic locations of variants with particular annotations and mechanisms, relative to the gene body of the pQTL. For example, Indel annotated variants were only located within gene exons, but variants that have an underlying mechanism of "isoform" could be found in introns, exons, or the 3' UTR. The three pQTLs where the PFV was a large genic deletion arae not illustrated. (B) Stacked barplot of the number of PFVs associated with each mechanism, subset by whether the mechanism affects the RNA molecule, or the protein directly. (C) Stacked barplot of the number of PFVs with each SnpEff annotation, subset by whether the PFVs' mechanism affects the RNA molecule, the protein directly, or is unknown. (D) Stacked barplot of the number of PFVs that were eQTLs in GETx, subset by whether the PFVs' mechanism affects the RNA molecule, the protein directly, or is unknown.

Chapter 3.4.7 Examining the Functionality of PFVs

To examine the relative role of different stages of protein level regulation – from gene expression to post-translational modifications – we further classified the PFVs by their molecular mechanism of action using the mechanism linked to the variant during PFV identification. We found the 28 PFVs with suggestive or better evidence to affect a wide range of processes, including 19 (68%) involved in RNA metabolism (7 affected the promoter, 4 affected isoform expression, 1 created a transcript that underwent nonsense-mediated decay, 3 resulted in gene deletions, and 4 affected miRNA processing), and 9 (32%) involved in protein metabolism (6 associated with protein degradation, 2 altered glycosylation, and 1 affected secretion) (Figure 3.4B; Supplemental Figure 3.3B). We next examined if the functional annotation of the variant was correlated with whether the mechanism influenced RNA or protein levels. We observed that PFVs associated with protein levels directly were more often missense variants, whereas PFVs that affected RNA levels were primarily located in non-coding regions (Figure 3.4C). The PFVs that did not have an established mechanism (unknown) were annotated as both missense and noncoding variants, suggesting that some of the unknown PFVs affect protein levels directly, whereas others affect RNA. As variants associated with RNA metabolism would also be expected to show association as an expression QTL (eQTL), we also examined whether these were more often identified in GTEx. We observed that PFVs which affected RNA levels, and were not deletions, were more likely to have been identified as an eQTL (69%, 11/16) than protein PFVs (22%, 2/9) (Figure 3.4D). The unknown PFVs were identified as eQTLs at an intermediate level (40%, 8/20), consistent with this group affecting both RNA and protein levels. These results suggest that, while variants that

affect protein levels often work through mechanisms associated with RNA, and therefore can be detected through eQTL analyses, many variants affect protein levels without affecting RNA levels, and act through molecular mechanisms that are more challenging to measure with current high throughput methods.

Chapter 3.5 Discussion

In this study, we leveraged TMT mass-spectrometry and deep whole exome sequencing data to identify 60 pQTLs (48 common *cis*, 3 rare *cis*, and 9 rare *trans*) associated with 96 unique peptides and 30 proteins across the genome (Supplemental Table 3.6). We then utilized published papers and public databases to examine established molecular mechanisms underlying these pQTLs, and examine how often the mechanisms affected RNA or protein metabolism. We showed that, while the majority of pQTLs exert their effects by modulating the gene's RNA, many affect the protein directly through processes such as degradation, glycosylation, and translation. Our work thus not only shows the importance of identifying functional variation by directly assaying protein levels, but also highlights how identifying the functional variant in pQTL studies can lead to insights into the molecular steps by which the protein is regulated. Based on the types of protein mechanisms that have been described, these results suggest that improved high throughput methods to assess variants that affect protein translation, modification, and degradation are needed.

It is currently unclear how often high throughput protein assays have technical artifacts resulting from genetic variants that affect the ability to correctly quantify peptide

levels due to alterations of coding sequence through missense changes, isoform usage, or cleavage patterns. By integrating the individuals' genotypes within coding sequences with standard TMT mass-spectrometry quantification techniques, we were able to identify pQTLs that were driven by genotype induced technical artifacts and exclude them. We observed the largest impact at the level of peptide-only associations, with the majority of independent associations (67%) being driven by technical artifacts. The majority of independent associations at the protein level (88% of both pQTLs and protein-only pQTLs), however, were unaffected. These findings illustrate the importance of filtering variants that affect peptide quantification, and using quantification techniques that measure proteins at multiple locations and are therefore are more resilient to peptide based quantification artifacts.

Rare variation is likely to be an important contributor to variation in protein levels. By focusing on the proteins that we measured, we identified trans associations between rare variation in FCN3 and the complement cascade. An individual who was homozygous for a rare frameshift variant in FCN3 has been reported to have a deficiency in complement activation (Munthe-Fog et al., 2009); however, this variant was not reported in our study. Our finding thus provides additional evidence that rare variation in FCN3 is associated with variation in levels of the complement pathway proteins in the general population. Additionally, we identified five protein loci with rare variation associated with levels of alpha-1 antitrypsin. Four of the proteins have been characterized as being involved in platelet degranulation, while the fifth, HYOU1, has been shown to act as an oxygen-inducible chaperone for proteins in the endoplasmic reticulum of macrophages (Ozawa et al., 2001). Alpha-1 antitrypsin deficiency is a well-established genetic condition

that predisposes an individual to chronic obstructive pulmonary disease, liver cirrhosis, and hepatocellular carcinoma (Stoller & Aboussouan, 2012). While over 120 alleles of the SERPINA1 gene have been implicated in alpha-1 antitrypsin deficiency, variation in genes other than SERPINA1 have not yet been described (Stoller & Aboussouan, 2012). While the individuals in this study have not been found to have alpha-1 antitrypsin deficiency, the finding that rare variation in many genes can contribute to alpha-1 antitrypsin plasma levels could have implications for the genetic architecture of the disorder.

Due to the fact that our analyses are based on high throughput data, the novel associations that we identified should be further validated by replication in an independent data set. As many of our findings were consistent with previous work, we expect that the majority of the novel associations will be replicated in future studies. Additionally, the annotation of PFVs may have been biased for finding missense variants, as we relied on published literature and databases, and past protein research may have focused on studying missense variation. However, as the majority of the PFVs that we identified were regulatory in nature, and the class of unknown variants showed annotations consistent with them affecting both RNA and protein metabolism, we believe that PFV annotations were likely not strongly biased for previously characterized missense variants.

Chapter 3.6 Acknowledgements

The authors would like to thank Margaret K. R. Donovan for assistance with figure generation.

Chapter 3, in full, is currently being prepared for submission for publication. Terry Solomon, John Lapek, Søren Beck Jensen, Hiroko Matsui, Kristian Hindberg, William Greenwald, Nadezhda Latysheva, Sigrid Braekkan, David Gonzalez, Kelly A. Frazer, Erin Smith, John-Bjarne Hansen. The dissertation author was the primary investigator and author of this paper.

Chapter 3.7 Funding Sources

This work was supported by an independent grant from Stiftelsen Kristian Gerhard Jebsen in Norway. T. Solomon was supported by an institutional award to the UCSD Genetics Training Program from the National Institute for General Medical Sciences, T32 GM008666. This work was supported by the Ray Thomas Edwards Foundation and the University of California Office of the President (D.J.G.). J.D.L. is an IRACDA fellow supported by NIGMS/NIH (K12GM068524).

Chapter 3.8 Supplemental Tables

(see downloadable file for Chapter 3 Supplemental Tables)

Supplemental Table 3.1 Single-site association of cis genetic variants with peptides and/or proteins.

Protein Name corresponds to the gene name. Ensembl ID is the GRCh37 Ensembl ID for the transcript tested. Peptide Sequence is the amino acid sequence of the peptide tested. This is N/A if the association was at the protein level. Sentinel Variant is the rsID for the most significantly associated genetic variant. P-value is the p-value of that variant's association with the tested peptide or protein. Beta is the effect on the normalized peptide or protein level that corresponds to each additional copy of the effect allele for the most significant variant. R2 is the amount of variance in the peptide or protein level that is explained by the most significant variant. Number of Samples lists the number of individuals that the association was calculated in. pQTL Type is whether the association was for a peptide or protein. Number of Peptides Collapsed into Measurement for proteins lists the number of peptides that were summed together to get the protein level. This is N/A for peptides. Conditional Association Result is primary for the first significant variant identified in a locus and secondary if there was an additional variant associated when the first was taken into account. Integrated pQTL Class is whether the association is a

peptide-only pQTL, protein-only pQTL, or both pQTL. LD Between Peptide pQTL and Protein pQTL is whether the sentinel variant was identified or was in LD (R2) with another variant that was identified in this study as a pQTL for that Protein Name. As multiple peptides or proteins can correspond to the same variant, the independent pQTL column is 1 for each independent pQTL and 0 for each repeat. The Artifact column is 1 if the pQTL was deemed an artifact and 0 if not. The Artifact Type lists the evidence supporting whether the variant was deemed an artifact. The In Final Analysis column gives a 1 if the pQTL was an independent pQTL and not an artifact and was thus kept for further analysis. Previously Reported cites the pQTL paper that has previously published this association. [sort table by Protein Name, Integrated pQTL Class, In Final Analysis for clarity]

Supplemental Table 3.2 Grouped association for cis genetic variants with peptides and/or proteins.
Protein Name corresponds to the gene name. Ensembl ID is the GRCh37 Ensembl ID for the transcript tested. Peptide Sequence is the amino acid sequence of the peptide tested. This is N/A if the association was at the protein level. Rare variant criteria is whether the association was found using the MAF < 5%, Deleterious, or CADD-score grouping criteria. P-value column shows the p-value of that region's association with the tested peptide or protein. Number of Samples lists the number of individuals that the association was calculated in. Integrated pQTL Class is whether the association is a peptide-only pQTL, protein-only pQTL, or both pQTL. Rare Variants in Peptide determines whether one of the rare variants tested fell within the peptide sequence and is thus likely an artifact. The Artifact column is 1 if the pQTL was deemed an artifact and 0 if not. Significant in Cis Single-site Analysis column is a 1 if an association was already seen for that peptide or protein in the single-site analysis. The In Final Analysis column gives a 1 if the pQTL was not significant in the cis single-site analysis and not an artifact and was thus kept for further analysis. Previously Reported Citation lists the PubMed ID of any papers that describes an association between one of the underlying rare variants and levels of the protein.

Supplemental Table 3.3 Grouped association for trans genetic variants with peptides.
Protein Name corresponds to the gene name. Ensembl ID is the GRCh37 Ensembl ID for the transcript tested. Peptide Sequence is the amino acid sequence of the peptide tested. Locus is the genetic locus that was associated with the peptide. Rare variant criteria is whether the association was found using the MAF < 5%, Deleterious, or CADD-score grouping criteria. P-value column shows the p-value of that region's association with the tested peptide or protein. Number of Samples lists the number of individuals that the association was calculated in.

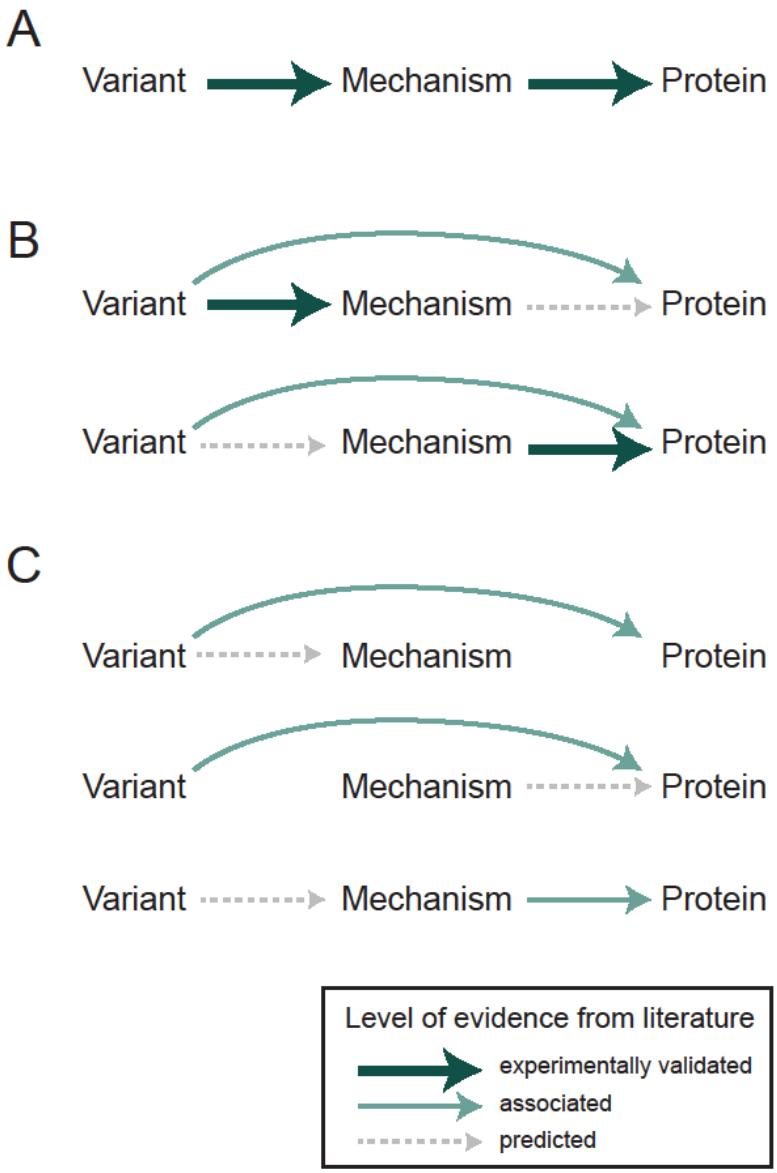
Supplemental Table 3.4 Association of variants in FCN3 with proteins in the Lectin complement pathway.
Protein Name corresponds to the gene name. Ensembl ID is the GRCh37 Ensembl ID for the transcript tested. Peptide Sequence is the amino acid sequence of the peptide tested. Rare variant criteria is whether the association was found using the MAF < 5%, Deleterious, or CADD-score grouping criteria. P-value column shows the p-value of that region's association with the tested peptide.

Supplemental Table 3.5 Annotation of all pQTLs.
Protein Name corresponds to the gene name. Putative Functional Variant (PFV) is the rsID of the proposed functional variant. Sentinel Variant is the rsID of the pQTL. Level is whether the association is a peptide-only pQTL, protein-only pQTL, or both pQTL. LD between PFV and Sentinel lists the r2 between the two variants from HaploReg. Sentinel P-Value lists the lowest p-value of a peptide or protein found with the sentinel variant. PFV P-value lists the corresponding p-value between the PFV and the same peptide or protein measured for the Sentinel P-value. PFV in OMIM lists the OMIM ID if the PFV was reported in OMIM. Sentinel SNPEff Annotation

lists the functional annotation from SNPEff for the sentinel variant. PFV HaploReg Annotation lists the functional annotation and gene locus from HaploReg for the PFV. PFV SNPEff Annotation lists the functional annotation from SNPEff for the PFV. Tissue is the tissue with the highest expression of the protein according to GTEx. PFV eQTL Tissue lists whether the PFV was an eQTL in GTEx in the same tissue as the Tissue column (“Same”), a different tissue (“Other”), or was not an eQTL (“none”). Strength is whether the evidence supporting the mechanism was classified as “known”, “likely”, “suggestive”, or “unknown”. Mechanism lists the category of the mechanism of action. Affected Molecule lists whether the PFV affects RNA levels, Protein levels, or unknown. PMIDs lists the papers with evidence supporting the mechanism.

Supplemental Table 3.6 All pQTLs identified in this study.
Type of Association is which analysis the pQTL was identified in. Protein Name corresponds to the gene name. Ensembl ID is the GRCh37 Ensembl ID for the transcript tested. Associated Protein is the ensemble ID if the pQTL was identified at the protein level. Associated Peptides lists the amino acid sequences of all peptides that were associated with that variant. Sentinel Variant is the rsID of the pQTL or the name of the gene locus if this was a rare pQTL. Integrated pQTL Class is whether the association is a peptide-only pQTL, protein-only pQTL, or both pQTL.

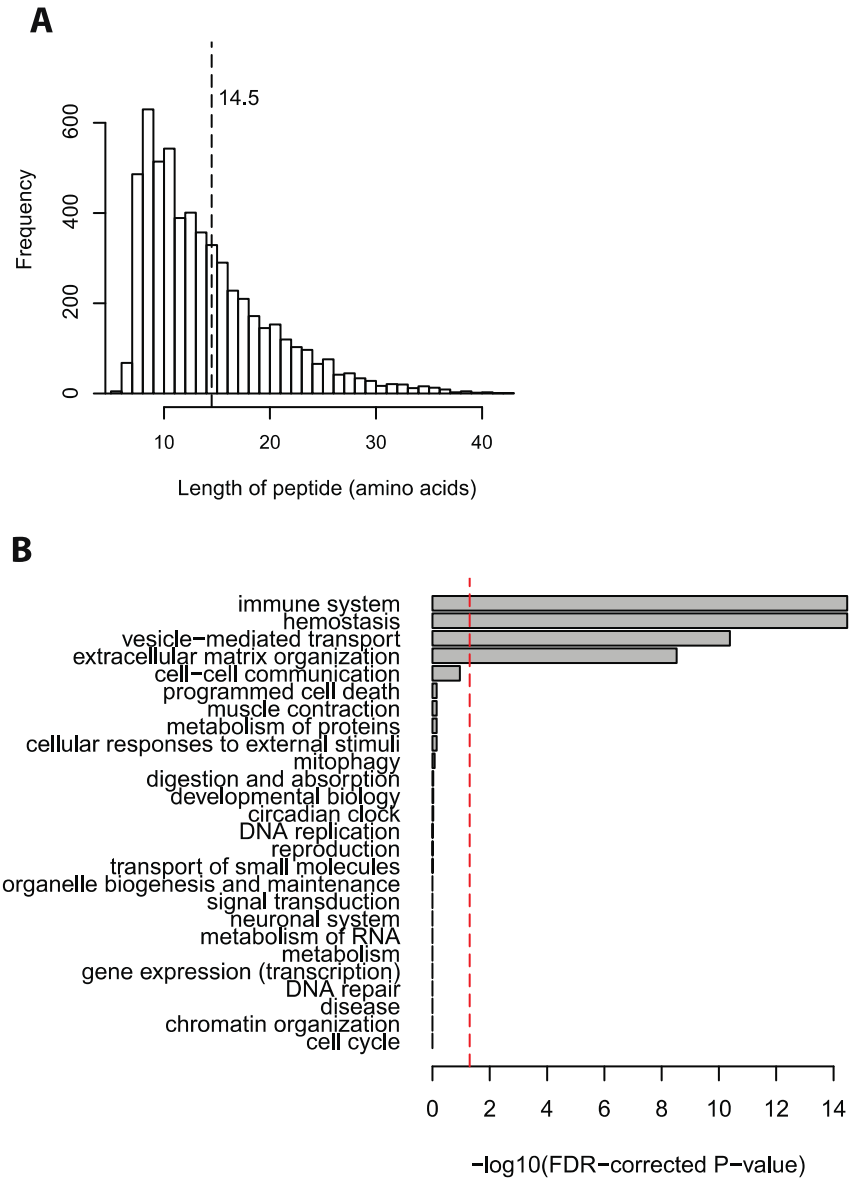
Chapter 3.7 Supplemental Figures



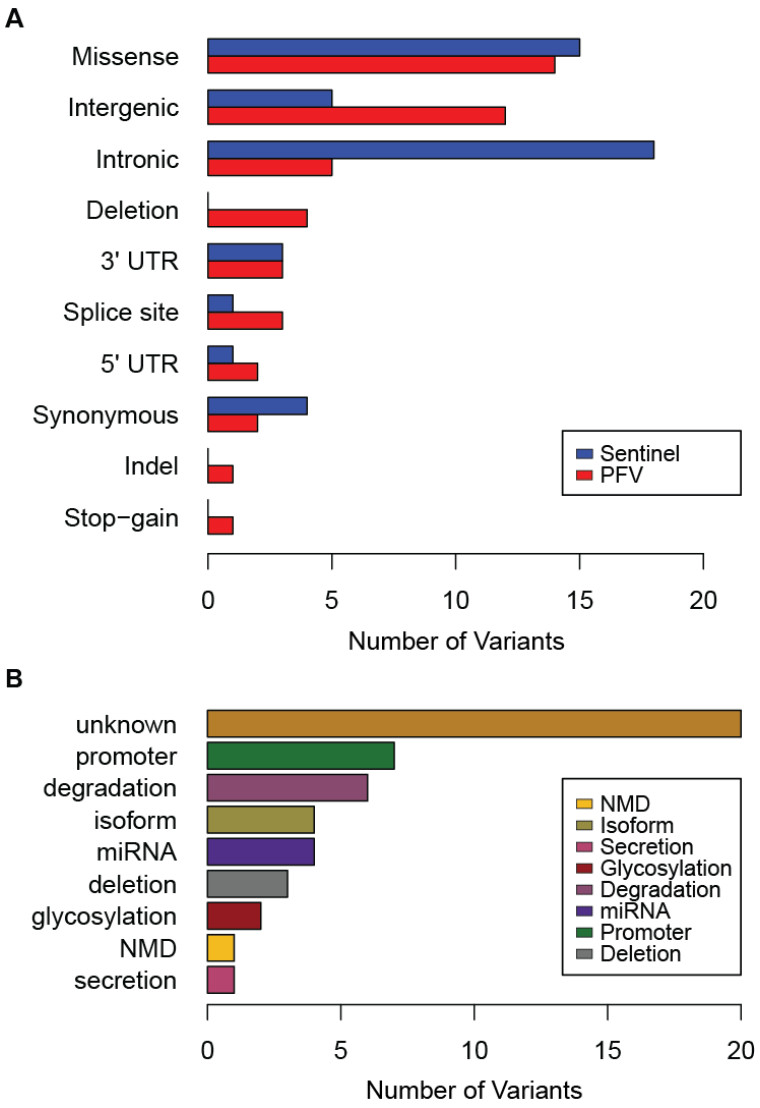
Supplemental Figure 3.1 Schematic showing the categories of strength of supporting evidence for PFV classification

(A) Known: the PFV had been experimentally shown to affect the protein's level through a particular molecular mechanism. (B) Likely: either the PFV had been experimentally validated to have a molecular mechanism that was predicted to affect the protein's level, and the variant was previously associated with the protein's level (top), or the PFV was predicted to have a mechanism that had been experimentally shown to affect protein's level, and the variant was previously associated with the protein's level (bottom). (C) Suggestive: either the PFV was predicted to act through a specific molecular mechanism, and the PFV was previously associated with the protein's level (top), the PFV was associated with the protein's level and a molecular mechanism had been previously predicted to affect the protein's level (middle), or the PFV was predicted to

act through a molecular mechanism, and the proposed mechanism had been associated with the protein's level (bottom).



Supplemental Figure 3.2 Data characterization of the plasma peptides and proteins
(A) Histogram of the distribution of the number of amino acids within each of the 5,608 quantified peptides, with the mean of 14.5 amino acids indicated by the dotted black line. (B) Bar plot of the top 25 pathways from Reactome pathway analysis. Dotted line indicates significance at FDR q < 0.05.



Supplemental Figure 3.2 Characterization of putative functional variants
(A) Barplot showing the number of sentinel variants (blue) or PFVs (red) in each SnpEff annotation. (B) Barplot of the number of PFVs with each mechanism.

Chapter 3.8 References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., McVean, G. A., & Consortium, G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi:10.1038/nature11632

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Braekkan, S. K., Mathiesen, E. B., Njølstad, I., Wilsgaard, T., Størmer, J., & Hansen, J. B. (2008). Family history of myocardial infarction is an independent risk factor for venous thromboembolism: the Tromsø study. *J Thromb Haemost, 6*(11), 1851-1857. doi:10.1111/j.1538-7836.2008.03102.x

Brantly, M., Courtney, M., & Crystal, R. G. (1988). Repair of the secretion defect in the Z form of alpha 1-antitrypsin by addition of a second mutation. *Science, 242*(4886), 1700-1702.

Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet, 98*(1), 116-126. doi:10.1016/j.ajhg.2015.11.020

Burgess, S., Timpson, N. J., Ebrahim, S., & Davey Smith, G. (2015). Mendelian randomization: where are we now and where are we going? *Int J Epidemiol, 44*(2), 379-388. doi:10.1093/ije/dyv108

Carson, A. R., Smith, E. N., Matsui, H., Brækkan, S. K., Jepsen, K., Hansen, J. B., & Frazer, K. A. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics, 15*, 125. doi:10.1186/1471-2105-15-125

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin), 6*(2), 80-92. doi:10.4161/fly.19695

Cohen, J. C., Boerwinkle, E., Mosley, T. H., & Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med, 354*(12), 1264-1272. doi:10.1056/NEJMoa054013

Consortium, G. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet, 45*(6), 580-585. doi:10.1038/ng.2653

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Group, G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods, 4*(3), 207-214. doi:10.1038/nmeth1019

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom, 5*(11), 976-989. doi:10.1016/1044-0305(94)80016-2

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., & D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Res, 46*(D1), D649-D655. doi:10.1093/nar/gkx1132

Folkersen, L., Fauman, E., Sabater-Lleal, M., Strawbridge, R. J., Franberg, M., Sennblad, B., Baldassarre, D., Veglia, F., Humphries, S. E., Rauramaa, R., de Faire, U., Smit, A. J., Giral, P., Kurl, S., Mannarino, E., Enroth, S., Johansson, A., Enroth, S. B., Gustafsson, S., Lind, L., Lindgren, C., Morris, A. P., Giedraitis, V., Silveira, A., Franco-Cereceda, A., Tremoli, E., group, I. s., Gyllensten, U., Ingelsson, E., Brunak, S., Eriksson, P., Ziemek, D., Hamsten, A., & Malarstig, A. (2017). Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet, 13*(4), e1006706. doi:10.1371/journal.pgen.1006706

Garred, P., Honore, C., Ma, Y. J., Munthe-Fog, L., & Hummelshoj, T. (2009). MBL2, FCN1, FCN2 and FCN3-The genes behind the initiation of the lectin pathway of complement. *Mol Immunol, 46*(14), 2737-2744. doi:10.1016/j.molimm.2009.05.005

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res, 33*(Database issue), D514-517. doi:10.1093/nar/gki033

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van

Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R., & Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res, 22*(9), 1760-1774. doi:10.1101/gr.135350.111

Jacobs, J. M., Adkins, J. N., Qian, W. J., Liu, T., Shen, Y., Camp, D. G., 2nd, & Smith, R. D. (2005). Utilizing human blood plasma for proteomic biomarker discovery. *J Proteome Res, 4*(4), 1073-1085. doi:10.1021/pr0500657

Jacobsen, B. K., Eggen, A. E., Mathiesen, E. B., Wilsgaard, T., & Njolstad, I. (2012). Cohort profile: the Tromso Study. *Int J Epidemiol, 41*(4), 961-967. doi:10.1093/ije/dyr049

Jensen, S. B., Hindberg, K., Solomon, T., Smith, E.N., Lapek, J.D., Gonzalez, D.J., Latysheva, N., Frazer, K.A., Brækkan, S.K., Hansen, J.B. Discovery of novel plasma biomarkers for future incident venous thromboembolism by untargeted synchronous precursor selection mass spectrometry proteomics. *In progress*.

Johansson, Å., Enroth, S., Palmblad, M., Deelder, A. M., Bergquist, J., & Gyllensten, U. (2013). Identification of genetic variants influencing the human plasma proteome. *Proc Natl Acad Sci U S A, 110*(12), 4673-4678. doi:10.1073/pnas.1217238110

Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods, 4*(11), 923-925. doi:10.1038/nmeth1113

Kang, H. M. (2014). EPACTS (Version 3.2.5): University of Michigan Center for Statistical Genetics. Retrieved from http://www.sph.umich.edu/csg/kang/epacts/

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet, 42*(4), 348-354. doi:10.1038/ng.548

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res, 12*(6), 996-1006. doi:10.1101/gr.229102

Kim, S., Swaminathan, S., Inlow, M., Risacher, S. L., Nho, K., Shen, L., Foroud, T. M., Petersen, R. C., Aisen, P. S., Soares, H., Toledo, J. B., Shaw, L. M., Trojanowski, J. Q., Weiner, M. W., McDonald, B. C., Farlow, M. R., Ghetti, B., Saykin, A. J., & , A. s. D. N. I. (2013). Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. *PLoS One, 8*(7), e70269. doi:10.1371/journal.pone.0070269

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. doi:10.1038/ng.2892

Lapek, J. D., Jr., Lewinski, M. K., Wozniak, J. M., Guatelli, J., & Gonzalez, D. J. (2017). Quantitative Temporal Viromics of an Inducible HIV-1 Model Yields Insight to Global Host Targets and Phospho-Dynamics Associated with Protein Vpr. *Mol Cell Proteomics, 16*(8), 1447-1461. doi:10.1074/mcp.M116.066019

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics, 13*(4), 762-775. doi:10.1093/biostatistics/kxs014

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Liu, Y., Buil, A., Collins, B. C., Gillet, L. C., Blum, L. C., Cheng, L. Y., Vitek, O., Mouritsen, J., Lachance, G., Spector, T. D., Dermitzakis, E. T., & Aebersold, R. (2015). Quantitative variability of 342 plasma proteins in a human twin population. *Mol Syst Biol, 11*, 786.

Lourdusamy, A., Newhouse, S., Lunnon, K., Proitsi, P., Powell, J., Hodges, A., Nelson, S. K., Stewart, A., Williams, S., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M., Vellas, B., Lovestone, S., Dobson, R., Consortium, A., & Initiative, A. s. D. N. (2012). Identification of cis-regulatory variation

influencing protein abundance levels in human plasma. *Hum Mol Genet, 21*(16), 3719-3726. doi:10.1093/hmg/dds186

MacKeigan, J. P., Clements, C. M., Lich, J. D., Pope, R. M., Hod, Y., & Ting, J. P. (2003). Proteomic profiling drug-induced apoptosis in non-small cell lung carcinoma: identification of RS/DJ-1 and RhoGDIalpha. *Cancer Res, 63*(20), 6928-6934.

Melzer, D., Perry, J. R., Hernandez, D., Corsi, A. M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J. R., Paolisso, G., Rafiq, S., Simon-Sanchez, J., Lango, H., Scholz, S., Weedon, M. N., Arepalli, S., Rice, N., Washecka, N., Hurst, A., Britton, A., Henley, W., van de Leemput, J., Li, R., Newman, A. B., Tranah, G., Harris, T., Panicker, V., Dayan, C., Bennett, A., McCarthy, M. I., Ruokonen, A., Jarvelin, M. R., Guralnik, J., Bandinelli, S., Frayling, T. M., Singleton, A., & Ferrucci, L. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet, 4*(5), e1000072. doi:10.1371/journal.pgen.1000072

Munthe-Fog, L., Hummelshoj, T., Honore, C., Madsen, H. O., Permin, H., & Garred, P. (2009). Immunodeficiency associated with FCN3 mutation and ficolin-3 deficiency. *N Engl J Med, 360*(25), 2637-2644. doi:10.1056/NEJMoa0900381

Ong, S. E., Schenone, M., Margolin, A. A., Li, X., Do, K., Doud, M. K., Mani, D. R., Kuai, L., Wang, X., Wood, J. L., Tolliday, N. J., Koehler, A. N., Marcaurelle, L. A., Golub, T. R., Gould, R. J., Schreiber, S. L., & Carr, S. A. (2009). Identifying the proteins to which small-molecule probes and drugs bind in cells. *Proc Natl Acad Sci U S A, 106*(12), 4617-4622. doi:10.1073/pnas.0900191106

Ozawa, K., Kondo, T., Hori, O., Kitao, Y., Stern, D. M., Eisenmenger, W., Ogawa, S., & Ohshima, T. (2001). Expression of the oxygen-regulated protein ORP150 accelerates wound healing by modulating intracellular VEGF transport. *J Clin Invest, 108*(1), 41-50. doi:10.1172/JCI11772

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P., Sullivan, P. F., Bergen, S., Moran, J. L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Purcell, S. M., Haas, D. W., Liang, L., Sunyaev, S., Patterson, N., de Bakker, P. I., Reich, D., & Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet, 44*(6), 631-635. doi:10.1038/ng.2283

Reitsma, P. H., Versteeg, H. H., & Middeldorp, S. (2012). Mechanistic view of risk factors for venous thromboembolism. *Arterioscler Thromb Vasc Biol, 32*(3), 563-568. doi:10.1161/ATVBAHA.111.242818

Schofield, C. J., & Ratcliffe, P. J. (2004). Oxygen sensing by HIF hydroxylases. *Nat Rev Mol Cell Biol, 5*(5), 343-354. doi:10.1038/nrm1366

Solomon, T., Smith, E. N., Matsui, H., Braekkan, S. K., Wilsgaard, T., Njølstad, I., Mathiesen, E. B., Hansen, J.-B., Frazer, K. A., & Consortium, I. (2016). Associations Between Common and Rare Exonic Genetic Variants and Serum Levels of Twenty Cardiovascular-Related Proteins: The Tromsø Study. *Circulation: Cardiovascular Genetics*, CIRCGENETICS. 115.001327.

Spivak, M., Weston, J., Bottou, L., Kall, L., & Noble, W. S. (2009). Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. *J Proteome Res, 8*(7), 3737-3745. doi:10.1021/pr801109k

Stengaard-Pedersen, K., Thiel, S., Gadjeva, M., Moller-Kristensen, M., Sorensen, R., Jensen, L. T., Sjoholm, A. G., Fugger, L., & Jensenius, J. C. (2003). Inherited deficiency of mannan-binding lectin-associated serine protease 2. *N Engl J Med, 349*(6), 554-560. doi:10.1056/NEJMoa022836

Stoller, J. K., & Aboussouan, L. S. (2012). A review of alpha1-antitrypsin deficiency. *Am J Respir Crit Care Med, 185*(3), 246-259. doi:10.1164/rccm.201108-1428CI

Suhre, K., Arnold, M., Bhagwat, A. M., Cotton, R. J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., DeLisle, R. K., Gold, L., Pezer, M., Lauc, G., El-Din Selim, M. A., Mook-Kanamori, D. O., Al-Dous, E. K., Mohamoud, Y. A., Malek, J., Strauch, K., Grallert, H., Peters, A., Kastenmuller, G., Gieger, C.,

& Graumann, J. (2017). Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun, 8*, 14357. doi:10.1038/ncomms14357

Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., Oliver-Williams, C., Kamat, M. A., Prins, B. P., Wilcox, S. K., Zimmerman, E. S., Chi, A., Bansal, N., Spain, S. L., Wood, A. M., Morrell, N. W., Bradley, J. R., Janjic, N., Roberts, D. J., Ouwehand, W. H., Todd, J. A., Soranzo, N., Suhre, K., Paul, D. S., Fox, C. S., Plenge, R. M., Danesh, J., Runz, H., & Butterworth, A. S. (2017). Consequences Of Natural Perturbations In The Human Plasma Proteome. *bioRxiv*. doi:10.1101/134551

The UniProt, C. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res, 45*(D1), D158-D169. doi:10.1093/nar/gkw1099

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics, 11*(1110), 11.10.11-11.10.33. doi:10.1002/0471250953.bi1110s43

Wang, X., & Zhang, B. (2013). customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics, 29*(24), 3235-3237. doi:10.1093/bioinformatics/btt543

Ward, L. D., & Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res, 40*(Database issue), D930-934. doi:10.1093/nar/gkr917

Chapter 4: Discovery of novel plasma biomarkers for future incident venous thromboembolism by untargeted SPS-MS^3 proteomics.

Chapter 4.1 Abstract

Objective: Prophylactic anticoagulant treatment may substantially reduce the incidence of venous thromboembolism (VTE) but entails considerable risk of severe bleeding. Identification of individuals at high risk of VTE through the use of predictive biomarkers is desirable in order to achieve a favorable benefit-to-harm ratio. Therefore, we aimed to identify predictive protein biomarker candidates of VTE.

Approach and Results: We performed a case-control study of 200 individuals that participated in the Tromsø Study, a population-based cohort, where blood samples were collected before the VTE events occurred. Untargeted TMT-SPS-MS3-based (tandem mass tag-synchronous precursor selection-mass spectrometry) proteomic profiling was used to study the plasma proteomes of each individual. Of the 501 proteins detected in a sufficient number of samples to allow multivariate analysis, 46 proteins were associated with VTE case-control status with p-values below the 0.05 significance threshold. The strongest predictive biomarker candidates, assessed by statistical significance, were transthyretin, vitamin K-dependent protein Z, and protein/nucleic acid deglycase DJ-1.

Conclusions: Our untargeted approach of plasma proteome profiling revealed novel predictive biomarker candidates of VTE and confirmed previously reported candidates, thereby providing conceptual support to the validity of the study. A larger nested case-control study will be conducted to validate our findings.

Chapter 4.2 Introduction

Venous thromboembolism (VTE), a collective term for deep vein thrombosis and pulmonary embolism, has an annual incidence rate of 1-2 per 1000 persons (Heit, 2015). The health burden caused by VTE is immense, and it is expected to grow with the aging of the population and the increasing prevalence of major risk factors for VTE such as obesity and cancer (Afshin et al., 2017; Ferlay J, 2012; W. Huang, Goldberg, Anderson, Kiefe, & Spencer, 2014; "Thrombosis: a major contributor to the global disease burden," 2014). Prophylactic anticoagulant treatment in situations of high VTE risk provides an effective strategy for VTE prevention but entails a substantial risk of severe bleeding (Cohen et al., 2008; Mayer, Streiff, Hobson, Halpert, & Berenholtz, 2011). Thromboprophylaxis with anticoagulants should therefore be targeted towards individuals with the highest VTE risk in order to achieve a favorable benefit-to-harm ratio.

VTE is a complex disease that occurs as a result of interactions between inherited and acquired factors (Rosendaal, 1999). Several genetic variants and the levels of numerous plasma proteins, mostly with roles in coagulation or fibrinolysis, have been shown to be associated with VTE (Bruzelius et al., 2016; Christiansen et al., 2006; Fashanu et al., 2017; Germain et al., 2015; Heit, 2015; Karasu, Baglin, Luddington, Baglin, & van Hylckama Vlieg, 2016; Meltzer et al., 2010; Norgaard, Nielsen, & Nordestgaard, 2016; Puurunen et al., 2016; Reitsma & Rosendaal, 2004; Ridker, Cushman, Stampfer, Tracy, & Hennekens, 1997; Tsai et al., 2002; van

Hylckama Vlieg et al., 2015; van Montfoort et al., 2013). However, few prospective studies have successfully shown associations between protein biomarker levels at baseline and risk of future incident VTE (Christiansen et al., 2006; Fashanu et al., 2017; Puurunen et al., 2016; Ridker et al., 1997; Tsai et al., 2002). The discovery of novel biomarkers for risk prediction of incident VTE in the general population is therefore warranted. Furthermore, the identification of individuals at high risk of VTE is challenging, as it requires integration of both clinical risk factors and biomarkers. Current risk prediction models for VTE are often restricted to patient subgroups and they have shown limited predictive power, particularly in validation studies (Bruzelius et al., 2015; de Haan et al., 2012; Folsom et al., 2016; Greene et al., 2016; Mahan, Burnett, Fletcher, & Spyropoulos, 2017; Park et al., 2017; Pepin et al., 2016; Puurunen et al., 2016; van Es et al., 2017).

The proteomic profile of blood plasma is influenced by both genetic and environmental factors that may affect the risk of VTE. Combined with the minimal invasiveness and cost of blood sampling, blood plasma is a clinically attractive and relevant specimen for the discovery of novel biomarkers for VTE. Recent advances in mass spectrometry technology have increased the feasibility of mass spectrometry (MS)-based biomarker discovery studies. Improved accuracy in relative protein quantification combined with the development of sample multiplexing protocols have made MS an attractive technology for plasma biomarker discovery (Cominetti et al., 2016; Dayon, Nunez Galindo, Corthesy, Cominetti, & Kussmann, 2014; McAlister et al., 2014; Ting, Rad, Gygi, & Haas, 2011).

This study was designed to identify novel plasma protein biomarkers for future incident VTE. We combined Tandem-Mass-Tag (TMT)10-plexing with synchronous precursor selection (SPS)-MS (MS3) to generate untargeted proteomic profiles (McAlister et al., 2014). Our study included 100 individuals who developed VTE and 100 age and sex-matched control individuals selected from a population-based cohort where plasma samples were collected before the VTE events occurred. To our knowledge, this study is the first to employ untargeted plasma proteomic profiling with the objective to discover predictive biomarkers for incident VTE, and is the first to take advantage of the improved accuracy of MS3 in a larger plasma proteomic study. We identified a panel of 46 biomarker candidates worthy of further investigation and validation.

Chapter 4.3 Materials and Methods

Chapter 4.3.1 Source Population

Participants were recruited from the fourth survey of the Tromsø Study conducted in 1994-95, where all inhabitants of Tromsø (Norway) older than 24 years of age were invited to participate in a prospective health survey (Jacobsen, Eggen, Mathiesen, Wilsgaard, & Njolstad, 2012). The participation rate was 77% with 27,158 individuals attending the first visit. Additionally, a subset of the participants was invited for a more extensive examination, and 7,965 individuals participated in the second visit. Those who did not consent to medical research (n=300), who were not officially registered as inhabitants of the municipality of Tromsø at baseline

(n=43), and those with a known pre-baseline history of VTE (n=47) were excluded from the study. The remaining participants (n=26,768) were followed from the date of enrollment until September 1, 2007. All first lifetime events of VTE occurring among the participants during follow-up were identified from the discharge diagnosis registry, the autopsy registry, and the radiology procedure registry at the University Hospital of North Norway, which is the sole hospital in the Tromsø region. Trained personnel adjudicated each VTE by extensive medical records review. A VTE was adjudicated if the presence of signs and symptoms of deep vein thrombosis or pulmonary embolism were combined with objective confirmation by radiological procedures, which resulted in treatment initiation (unless contraindications were specified) as previously described (Braekkan et al., 2010). In total, 462 VTE events occurred in the follow-up period.

Chapter 4.3.2 The Study Population

From the source population, we established a case-control study of 100 VTE cases and 100 controls. For each VTE case, an age- and sex-matched control was randomly sampled from the source cohort. Cases were prioritized according to the shortest time from blood sampling to VTE, and the first 100 case-control pairs where both plasma samples passed quality control (as described below) were included to form our case-control study.

Chapter 4.3.3 Ethics Approval

All participants provided informed written consent to participate in accordance with the declaration of Helsinki. The study was approved by the Regional Committee of Medical and Health Research Ethics.

Chapter 4.3.4 Plasma Collection and Base Line Characteristics

Baseline characteristics including age, sex, and anthropometrics were collected by physical examination at study enrollment. Height and weight were measured with subjects wearing light clothing and no shoes. BMI (Body mass index) was calculated as the weight in kilograms divided by the square of height in meters (kg/m2). Non-fasting blood samples were drawn from an antecubital vein into 5 mL vacutainer tubes containing EDTA (Ethylenediaminetetraacetic acid) as an anticoagulant (K3-EDTA 40 µL, 0.37 mol/L per tube). Blood samples were processed within 1 hour by centrifugation at 3000 g for 10 min at 22°C, and plasma was collected and frozen in 1 mL aliquots. The plasma samples were stored at -70°C until analysis.

Chapter 4.3.5 Quality Control

The plasma samples obtained from the Tromsø Study were inspected visually for signs of hemolysis and the protein content was determined by Bradford assay (Biorad, Hercules, CA, USA). Signs of sample protein degradation were

assessed by Coomassie Blue visualization of 10 µg of protein from each sample separated by sodium dodecyl sulfate polyacrylamide gel electrophoresis on a 4-20% Criterion, gradient gel (Biorad, Hercules, CA, USA). The first 100 sample-pairs where both case- and control samples passed quality control as assessed by hemolysis, protein concentration (mean ± 2 standard deviations), and sodium dodecyl sulfate polyacrylamide gel electrophoresis band pattern were included in the study. After albumin and IgG depletion, 17 samples were picked randomly for quality control on sodium dodecyl sulfate polyacrylamide gel electrophoresis as described above and passed quality control (Supplementary Figure 4.1).

Chapter 4.3.6 Sample Preparation, Digestion, Labeling, and Multiplexing

Plasma samples were depleted for albumin and IgG on Q-proteome spin columns (Qiagen, Hilden, Germany, Cat#: 37521) following the manufacturer's instruction replacing the kit buffer with 100 mM triethylammonium bicarbonate. From each depleted plasma sample, 60 µg of protein was brought to 0.1% sodium dodecyl sulfate in 100 mM triethylammonium bicarbonate, reduced (tris(2-carboxyethyl)phosphine), alkylated (iodoacetamide) and digested with trypsin. TMT-labeling was performed by mixing the tryptic peptides with their relevant TMT-10plex reagent in accordance with a labeling plan that randomized each sample pair to 25 multiplexed experimental samples. Labeling reactions were terminated with hydroxylamine before each of the 25 multiplexed samples were combined from their respective labeled peptides according to the labeling plan. Multiplexed samples

were diluted to an acetonitrile concentration below 5% before solid phase purification on Oasis HLB-cartridges (Waters, Saint-Quentin, France) was performed and samples were eluted according to manufacturer's instructions.

Chapter 4.3.7 Sample Fractionation, Liquid Chromatography and Mass Spectrometry

Strong cation exchange chromatography on a polySULFOETHYL-A column (PolyLC, Columbia, MD, USA) on a high-performance liquid chromatography system from Waters Alliance (2695) (Saint-Quentin, France) was used to separate 300 µg of peptide from each experimental sample into eight fractions. The fractions were desalted on Oasis HLB cartridges and dried (Waters, Saint-Quentin, France). Each peptide fraction was re-suspended and analyzed in duplicate by liquid chromatography-MS3 using an EASY-nLC 1000 system coupled to an Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). Re-suspended peptides were loaded onto a nanoViper C18 Acclaim PepMap 100 pre-column (Thermo Fischer Scientific) and resolved using an increasing gradient of 0.1% Formic acid in acetonitril through a 50 cm PepMap RSLC analytical column (Thermo Fisher Scientific, Waltham, MA, USA) at a 200 nL/min flow rate. Peptide mass spectra were acquired throughout the chromatographic run of 180 min using a top 10 high-energy collision induced dissociation method for Fourier Transform-MS2 scans following each Fourier Transform-MS scan. SPS of several MS2 fragment ions followed by higher energy collisional dissociation fragmentation

released the reporter ions, which were detected in the Orbitrap at a resolving power of 30000 at 400 m/z (McAlister et al., 2012). Proteomic Sciences (Cobham, United Kingdom) performed the plasma sample quality control and subsequent steps for the generation of 400 raw-data files.

Chapter 4.3.8 Mass Spectrometry Data Analysis

Proteome Discoverer v2.1 (Thermo Scientific) was used as a data processing interface for all raw files, which were processed together to yield an accurate false discovery rate (Savitski, Wilhelm, Hahne, Kuster, & Bantscheff, 2015). The false discovery rate was set to 1% for both peptide and protein levels using a reverse database strategy (Elias & Gygi, 2007).We used spectrum selector default settings and SequestHT to identify peptides mapping to the Genecode human proteins sequence database (Gencode 19) (Eng, McCormack, & Yates, 1994). Oxidized methionine was included as a variable modification. Carbamidomethylation of cysteine, and 10-plex TMT-labels on peptide amino-termini and lysines were included as fixed modifications. Trypsin was selected as proteolytic enzyme and a maximum of three potential missed cleavages was allowed. Reporter ion signal-to-noise ratios were extracted with the reporter ions quantifier node in Proteome Discoverer were exported for relative quantification.

Chapter 4.3.9 Data Processing and Analysis

Peptide level filtering excluded peptides with isolation interference greater than 25% or average reporter ion signal-to-noise ratios below 10. Peptide level signal-to-noise ratios were summed to estimate protein abundances enforcing the principle of parsimony. Values from technical duplicates were averaged if both values were available, otherwise non-missing values were used. Data was normalized in a two-step process as previously described (Lapek, Lewinski, Wozniak, Guatelli, & Gonzalez, 2017).

Chapter 4.3.10 Post Normalization Data Quality Control

Proteins with measurements in all samples were used in unsupervised hierarchical clustering of Spearman's correlations between individual samples, and the heatmap.2 function in R was used to visually identify batch effects from TMT-label, experimental sample number, or Tromsø survey visit number.

Chapter 4.3.11 Statistical Analysis

Univariate and multivariate linear regression adjusting for age, sex, and BMI were performed to identify VTE-biomarker candidates with significantly different protein expression levels between cases and controls. To stabilize estimates in the multivariate linear regression, 10 measurements were required per explanatory variable, i.e. only proteins with valid measurements in at least 40 samples were

analyzed. Regression coefficients were standardized according to the standard deviation of the control group. We used a significance threshold of $p < 0.05$. All analysis were performed in R (version 3.3.3) using standard packages.

Chapter 4.4 Results

We established a case-control study of 100 VTE cases and 100 controls matched for age and sex with plasma samples available from the Tromsø Study that passed quality control procedures (Figure 4.1) (baseline characteristics in Supplementary Table 4.1). TMT10-multiplexing and liquid chromography-MS3 was used to generate plasma proteomic profiles of each individual sample in 25 multiplexed mass spectrometry experiments. We identified and performed relative quantification of 6,117 peptides mapping to 681 proteins in 200 human plasma samples (Figure 4.2A). Of the 681 proteins identified, 287 proteins (42%) were measured in all samples and 431 proteins (63%) were measured in more than half of the samples (Figure 4.2B). Of the 681 proteins, 488 proteins (71%) were identified by more than one peptide and a median of three peptides per protein were used for identification (Figure 4.2C). A two-step normalization was performed to account for slight differences in pipetting and TMT-labeling efficiency, and to allow comparison of relative protein levels across all samples in the study (Supplementary Figure 4.2A and 2B). A heatmap of Spearman's correlations revealed two clusters of highly correlated samples. These clusters contained almost exclusively samples collected at the second visit of the Tromsø survey, and only a single sample collected at the

second visit was not found in these two clusters (Supplementary Figure 4.2C). Therefore, the 24 samples collected at the second visit were removed from the analysis. Additionally, eight samples obtained from participants with active cancer at the time of blood sampling were removed (i.e. individuals diagnosed with cancer within 5 years before to 1 year after blood sampling). Baseline characteristics of the study participants after the removal of these 32 samples are summarized in Table 4.1. Data normalization and clustering analysis were re-performed. Clustering analysis revealed no batch effects of MS experiment number or TMT-label and indicated appropriate data normalization (Figure 4.2D). The normalized protein estimates from two technical replicates showed high correlations (range [0.80-0.98], median 0.91).

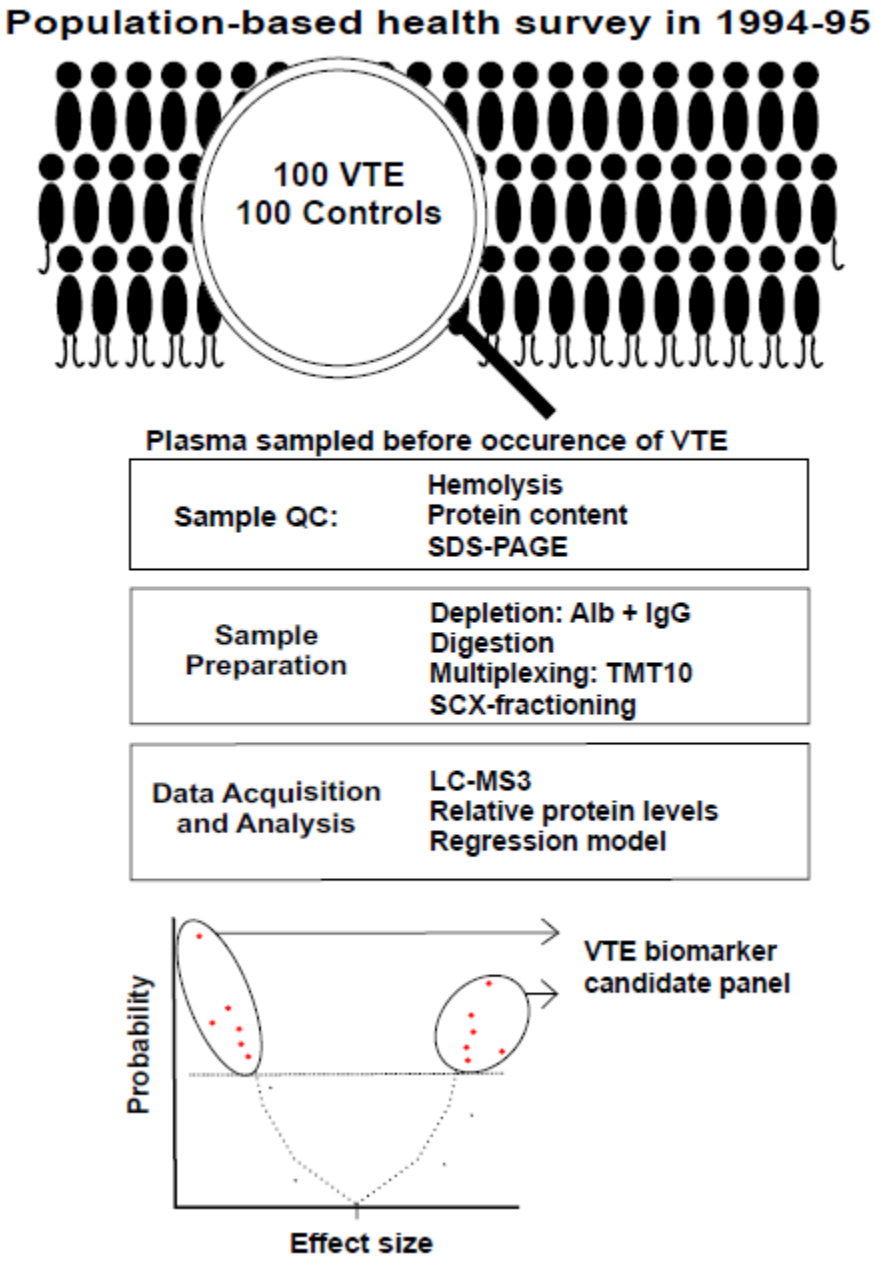


Figure 4.1 Study Overview

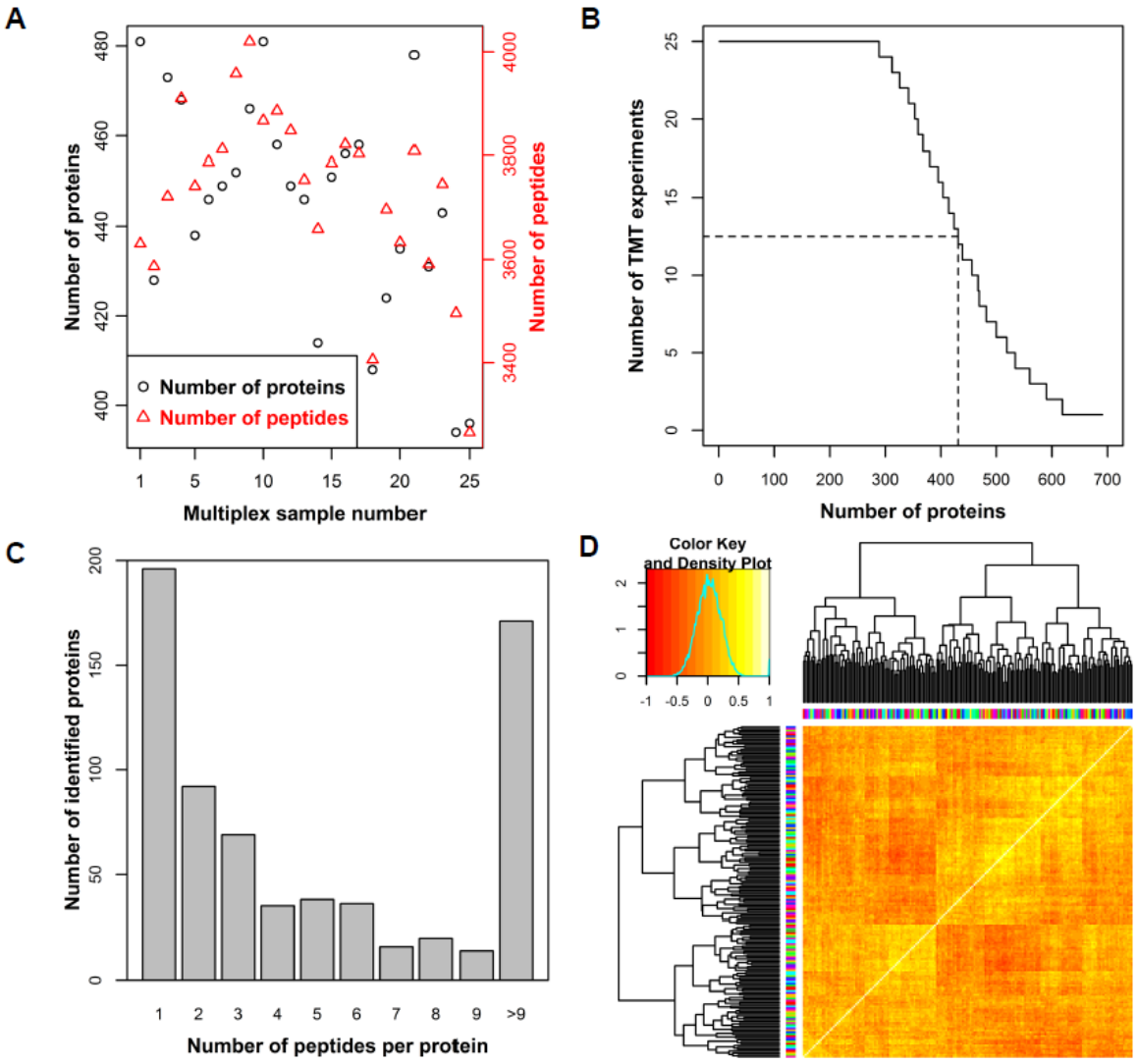


Figure 4.2 Characterization of proteins identified
Number of peptides (red) and proteins (black) identified in each multiplex sample. The sum of identifications in two technical replicates is shown (A). The number of proteins identified in a given number of multiplexed experiments. The dashed lines indicate identification in half of the TMT reactions (B). The number of proteins identified by a given number of peptides. For each protein, the sum of peptides across the dataset is provided (C). Heatmap of Spearman's correlation clustering for the study summarized in Table 4.1. Colors on axis indicate TMT-label (vertical) and multiplex sample number (horizontal) (D).

Table 4.1 Baseline characteristics of the study after removal of Tromsø Study second visit samples and participants with active cancer.
Abbreviations: Deep vein thrombosis (DVT), pulmonary embolism (PE).

	Cases	Controls
Participants	80	86
Median age, y [range]	65 [28-83]	65 [28-83]
Sex, male	32 (40%)	39 (45%)
BMI, kg/m^2 (mean ± SD)	27.0±4.1	24.7±3.5
Years to VTE, mean [range]	3.82 [0.09-6.85]	
DVT	55 (69%)	
PE	25 (31%)	
Cancer (at event)	17 (31%)	
Unprovoked	34 (43%)	

The normalized relative protein levels were regressed on age, sex, BMI, and VTE status in a multivariate linear model. To yield stable estimates we required a minimum of 40 measurements for a protein to be considered. The obtained p-value for the association with VTE status was used to evaluate the biomarker potential for each protein. Out of the 501 proteins tested in the multivariate analysis, 46 proteins had p-values below the 0.05 significance threshold (Figure 4.3 and Supplementary Table 4.2). For the proteins that were identified in too few samples to be considered in multivariate analysis, univariate statistics are provided in Supplementary Table 4.2.

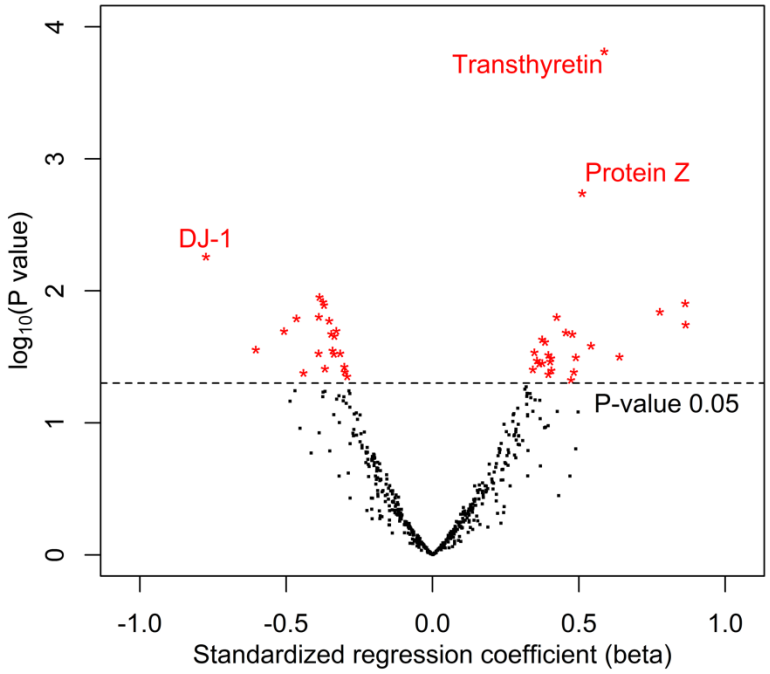


Figure 4.3 Volcano plot of plasma proteins identified in 40 or more samples.
For each protein, the standardized regression coefficient for VTE-status is plotted against -log (p-value). The multivariate model 10 included age, sex, and BMI as covariates. The black dashed line indicates a p-value of 0.05. The three candidates with lowest pvalues are indicated by their protein name. ProZ: Vitamine K-dependent protein Z, DJ-1: Protein/nucleic acid deglycase DJ-1

Based on statistical probability, the strongest biomarker candidate identified in this study was transthyretin with a nominal p-value of 0.00015 (Figure 4.3). We also found vitamin K-dependent protein Z (ProZ) to be overexpressed in cases although with a less extreme p-value of 0.0018 (Figure 4.3). Interestingly, the third lowest p-value was obtained for protein/nucleic acid deglycase DJ-1 (DJ-1) (p = 0.0055), which is also the candidate with the largest effect size (Figure 4.3). Figure

4.4 shows the relative protein estimates for cases and controls for each of the three aforementioned biomarker candidates.

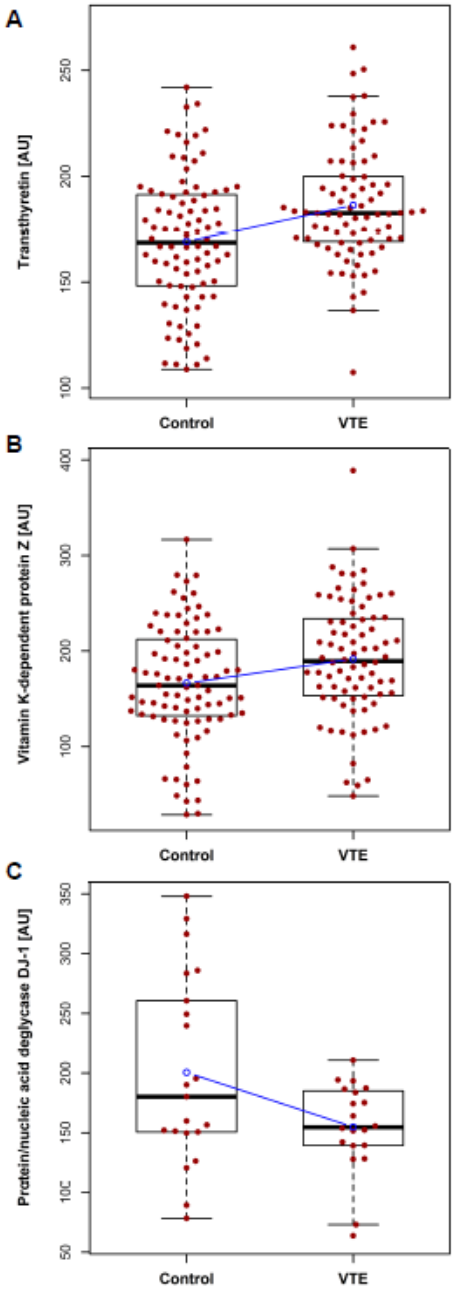


Figure 4.4 Boxplot of the relative plasma protein levels of transthyretin (A), DJ-1 (B), and ProZ (C) in cases and controls.
The regression line for VTE status is shown in blue. AU = arbitrary units.

We found a significant inverse correlation (Pearson's R = -0.41, p-value = 0.0046) between the plasma levels of transthyretin and DJ-1 (Figure 4.5). Sequence analysis revealed that the proposed optimal target sequence for DJ-1-mediated proteolysis is closely resembled by the 34-37th amino acids in transthyretin and may suggest that DJ-1-mediates cleavage of transthyretin after position V36 (Mitsugi et al., 2013).

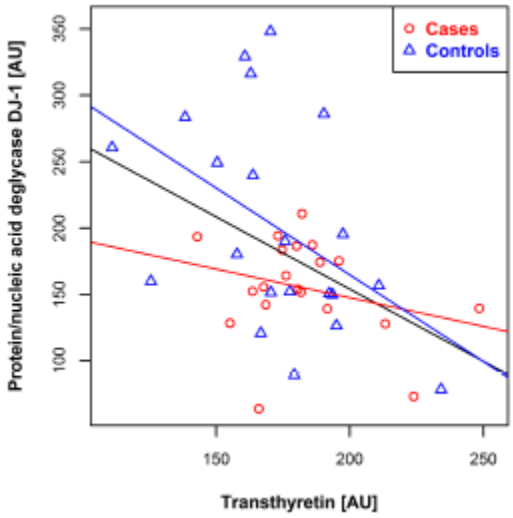


Figure 4.5 Scatter plot of relative transthyretin levels versus DJ-1 levels
Controls are shown in blue and cases in red, with corresponding regression lines. The black regression line is created with respect to all samples. AU = arbitrary units.

In our panel of predictive biomarker candidates, we found coagulation factor IX, galectin-3-binding protein, and both subunits of the heterodimeric S100A8/9 (correlation between subunits $R^2 = 0.96$) to be differentially expressed in cases and

controls (Supplementary Figure 4.3A-C and Supplementary Table 4.2). These biomarker candidates have previously been linked to VTE in retrospective case-control studies (Heikal et al., 2013; van Hylckama Vlieg, van der Linden, Bertina, & Rosendaal, 2000) or in animal models of VTE (DeRoo et al., 2015; Wang et al., 2017). Moreover, our candidate list included proteins related to the complement system and the ProZ-dependent protease inhibitor. The previously described predictive VTE biomarker von Willebrand factor showed differences in expression levels in the expected direction (i.e. overexpressed in VTE cases) without reaching statistical significance (p-value = 0.16) (Smith et al., 2010; Tsai et al., 2002). (Supplementary Figure 4.3D and Supplementary Table 4.2).

Chapter 4.5 Discussion

In this study, we present a large-scale MS3-based plasma proteomic profiling with the objective to discover novel biomarker candidates with the potential to predict incident VTE in the general population. We identified a panel of 46 biomarker candidates that included transthyretin, ProZ, and DJ-1 as the most promising candidates. Moreover, we revealed a negative correlation between transthyretin and DJ-1 that might suggest a mechanistic implication of these biomarkers in the pathogenesis of VTE. Finally, we support the concept that the proteins galectin-3-binding protein and S100A8/S100A9, previously reported to be involved in VTE pathogenesis using mouse models, are predictive biomarker candidates in humans. Moreover, the identification of galectin-3-binding protein and S100A8/9 as biomarker candidates, and the expected direction of difference in von Willebrand factor expression, lend conceptual support to the validity of this study.

Of the 681 proteins identified, 501 proteins were detected in a sufficient number of samples to allow multivariate analysis. We chose to present all proteins with p-values below 0.05 as biomarker candidates. This resulted in a panel of 46 proteins. When 501 statistical tests are conducted at a 0.05 significance threshold, 25 type I errors are expected. In a discovery study, the aim is to identify as many promising candidates as possible. Therefore, we omit control of the study-wide type I error rate since limitation hereof will increases the chance of type II error. Inflation of the type II error will erode the objective of a discovery study when followed up by

a validation study. Therefore, we promoted all candidates with p-values below 0.05 to our future validation study.

As we identified a high number of candidate plasma proteins associated with VTE, it is possible that many proteins act together to increase risk. The associations of VTE with elevated thrombin potential and hypofibrinolytic capacity support this notion (Karasu et al., 2016; Meltzer et al., 2010; van Hylckama Vlieg et al., 2015). Indeed, knowledge about non-linear interactions between single risk factors, such as the non-additive effects of prothrombin mutation 20210A and factor V Leiden (Simone et al., 2013), will be of pivotal importance to meet the challenge of VTE prediction and suggests a need for the development of panels of cooperating biomarkers (Demler, Pencina, & D'Agostino, 2013).

The strongest plasma biomarker candidate that we identified, transthyretin, forms a homotetramer that has two binding sites for thyroxine (Pettersson, Carlstrom, & Jornvall, 1987). Transthyretin misfolding can lead to amyloidosis, which affect as much as 25% of the elderly population, and may be linked to VTE through low-grade inflammation (Saghazadeh & Rezaei, 2016; Tanskanen et al., 2008). Interestingly, the inverse correlation between transthyretin and DJ-1 identified in this study is consistent with a previously reported proteolytic role for DJ-1 towards transthyretin reported in a study that also found an association between transthyretin amyloidosis and the secretion of an inactive form of DJ-1 (Koide-Yoshida et al., 2007). An alternative mechanistic explanation to DJ-1-mediated protection against VTE could be a reduction of advanced glycation end-products

that may contribute to VTE development (Richarme et al., 2015; Wautier & Wautier, 2013).

This study showed an upregulation of ProZ in subjects who later developed VTE, which might be surprising given its regulatory role in coagulation. Deficiency in ProZ has previously been associated with increased risk of VTE in retrospective studies (Al-Shanqeeti, van Hylckama Vlieg, Berntorp, Rosendaal, & Broze, 2005; Bafunno, Santacroce, & Margaglione, 2011; Sofi et al., 2010). However, in these studies blood was sampled after the occurrence of VTE entailing the risk of reverse causation. We note that plasma levels of ProZ are known to be affected by warfarin treatment and oral contraceptive use, and that a more controversial inverse relationship with interleukin-6 levels has been described (Al-Shanqeeti et al., 2005; Bafunno et al., 2011; Miletich & Broze, 1987). In plasma, ProZ is bound to a stoichiometric excess of protein Z-dependent protease inhibitor and promotes its inhibition of coagulation factor Xa (Han, Fiehler, & Broze, 1998). However, ProZ also impairs antithrombin mediated inhibition of coagulation factor Xa, which in combination with the vulnerability of protein Z-dependent protease inhibitor function to lipid oxidation may result in a procoagulant effect of ProZ in microenvironments with high levels of oxidative stress (Han et al., 1998; X. Huang et al., 2017). Our study is the first prospective study to assess the association between ProZ plasma levels and risk of future incident VTE.

The strength of our study lies in the combination of an epidemiological study design with the hypothesis free discovery approach offered by MS3-based

proteomics. The source cohort is recruited form a single-centered survey of the general population with a 77% participation rate that limits selection bias. Important to the discovery of predictive biomarkers, blood samples were drawn years before the VTE events occurred, and the VTE-events were well validated without knowledge on the proteome status. Additionally, we exploited the improved quantitative accuracy of MS3 and obtained individual untargeted plasma proteomic profiles (Ting et al., 2011).

In general, an important limitation of the data-dependent MS approach is the decreasing detectability of proteins with their decreasing abundancy. For example, we detected candidates like DJ-1, structural maintenance of chromosomes protein 5, and complement component C1q receptor in just enough samples to allow multivariate assessment. The statistical significance of these candidates was driven by large effect size, which may suggest these candidates to be the stronger predictive biomarkers for VTE. We offered univariate statistics for proteins that were detected in too few samples to yield stable estimates.

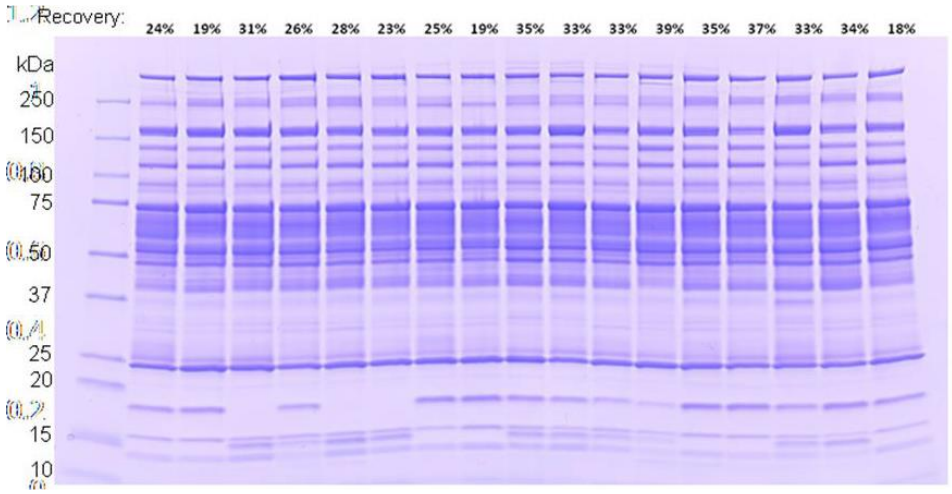
In conclusion, we present a large-scale MS3-based plasma proteomic profiling study designed to discover biomarker candidates with the potential to predict incident VTE in the general population. In a prospective case-control design with a discovery approach, we identified a panel of 46 biomarker candidates including transthyretin, ProZ and DJ-1. The biomarker candidates will be further validated in a larger, nested case-control study.

Chapter 4.6 Acknowledgements

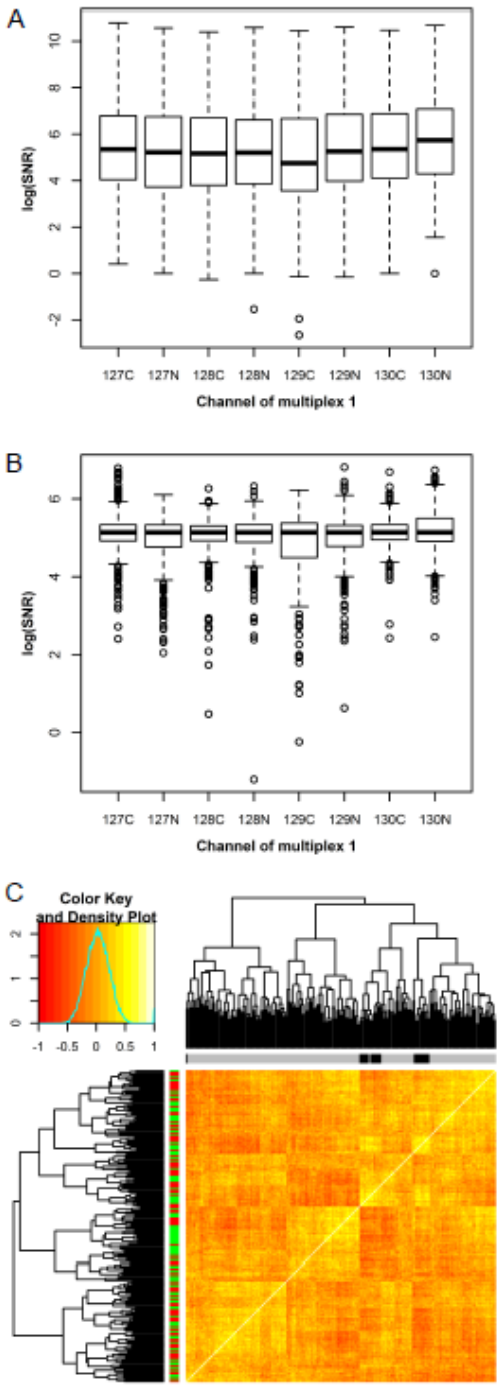
Chapter 4, in part, is currently being prepared for submission for publication. Søren Beck Jensen, Kristian Hindberg, Terry Solomon, Erin Smith, John Lapek, David Gonzalez, Nadezhda Latysheva, Kelly A. Frazer, Sigrid Braekkan, John-Bjarne Hansen. The dissertation author worked on the proteomics dataset and was a co-author of this paper.

The K. G. Jebsen Thrombosis Research and Expertise Center (TREC) is supported by an independent grant from Stiftelsen Kristian Gerhard Jebsen. J.D.L is an IRACDA fellow supported by NIGMS/NIH (K12GM068524).

Chapter 4.7 Supplemental Figures

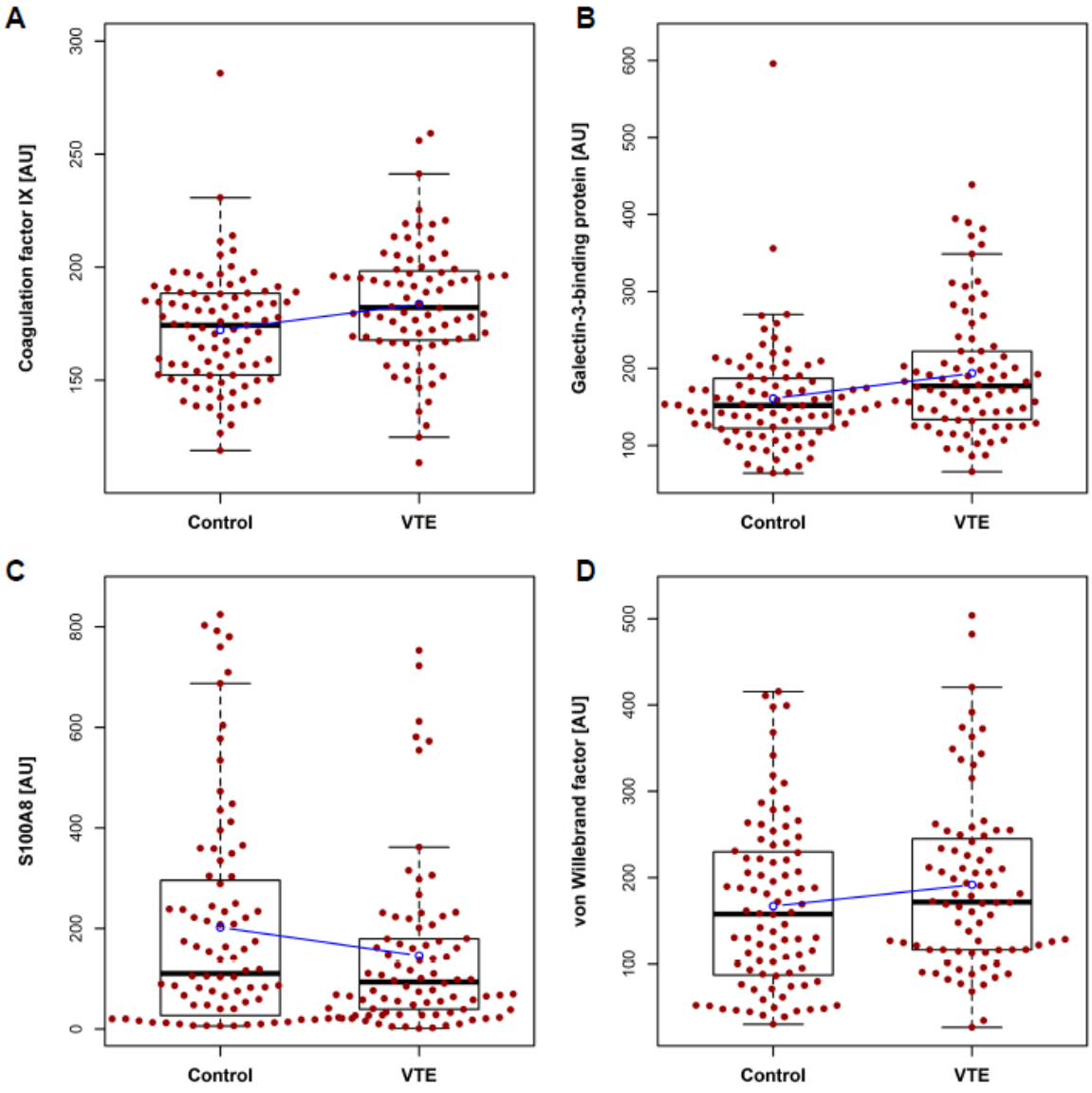


Supplemental Figure 4.1 Comassie Blue stain of SDS-PAGE-separated proteins from 17 randomly picked samples.
After sample depletion, 10 µg of protein was analyzed in each lane. The fraction of protein recovered after depletion is given as a percentage of the initial protein content above each lane.



Supplemental Figure 4.2 Boxplots of raw (A) and median normalized (B) relative protein estimates from each TMT label in experimental sample 1 replicate 1

(used for illustration). Heatmap of unsupervised clustering of Spearman's correlations of all samples. Tromsø Study visit number is indicated above the heatmap (1st visit in grey, 2nd visit in black). VTE-status is indicated to the left of the heatmap (cases in red, controls in green) (C). SNR = signal-to-noise ratio.



Supplemental Figure 4.3 Boxplot of the relative plasma protein levels of coagulation factor IX (A), galectin-3-binding protein (B), S100A8 (C), and von Willebrand factor (D) in cases and controls.

The regression line for VTE status is shown in blue. AU = arbitrary units.

Chapter 4.8: Supplemental Tables

Supplemental Table 4.1 Baseline characteristics of full case-control sample set.
DVT = deep vein thrombosis, PE = pulmonary embolism

	Cases	**Controls**
Participants	100	100
Median age [range]	65 [28-83]	65 [28-83]
Sex (male)	43	43
BMI (mean±SD)	27.0±4.1	24.8±3.6
Years to VTE (mean & [range])	3.82 [0.09-6.85]	
DVT	70	
PE	30	
Cancer (at event)	23	
Unprovoked	40	

(see downloadable file for Supplemental Table 4.2)

Supplemental Table 4.2 Transcript identifier for all identified proteins
Transcript identifier for all identified proteins is given with the number of peptides used for identification. For each protein, the number of detections in case- and control samples is provided. The standardized regression coefficients and p-values for VTE-status are provided for multivariable linear regression with adjustment for age, sex, and BMI and for univariate linear regression. Corresponding Uniprot protein or Ensemble gene descriptions are provided.

Chapter 4.9: References

Afshin, A., Forouzanfar, M. H., Reitsma, M. B., Sur, P., Estep, K., Lee, A., Marczak, L., Mokdad, A. H., Moradi-Lakeh, M., Naghavi, M., Salama, J. S., Vos, T., Abate, K. H., Abbafati, C., Ahmed, M. B., Al-Aly, Z., Alkerwi, A., Al-Raddadi, R., Amare, A. T., Amberbir, A., Amegah, A. K., Amini, E., Amrock, S. M., Anjana, R. M., Arnlov, J., Asayesh, H., Banerjee, A., Barac, A., Baye, E., Bennett, D. A., Beyene, A. S., Biadgilign, S., Biryukov, S., Bjertness, E., Boneya, D. J., Campos-Nonato, I., Carrero, J. J., Cecilio, P., Cercy, K., Ciobanu, L. G., Cornaby, L., Damtew, S. A., Dandona, L., Dandona, R., Dharmaratne, S. D., Duncan, B. B., Eshrati, B., Esteghamati, A., Feigin, V. L., Fernandes, J. C., Furst, T., Gebrehiwot, T. T., Gold, A., Gona, P. N., Goto, A., Habtewold, T. D., Hadush, K. T., Hafezi-Nejad, N., Hay, S. I., Horino, M., Islami, F., Kamal, R., Kasaeian, A., Katikireddi, S. V., Kengne, A. P., Kesavachandran, C. N., Khader, Y. S., Khang, Y. H., Khubchandani, J., Kim, D., Kim, Y. J., Kinfu, Y., Kosen, S., Ku, T., Defo, B. K., Kumar, G. A., Larson, H. J., Leinsalu, M., Liang, X., Lim, S. S., Liu, P., Lopez, A. D., Lozano, R., Majeed, A., Malekzadeh, R., Malta, D. C., Mazidi, M., McAlinden, C., McGarvey, S. T., Mengistu, D. T., Mensah, G. A., Mensink, G. B. M., Mezgebe, H. B., Mirrakhimov, E. M., Mueller, U. O., Noubiap, J. J., Obermeyer, C. M., Ogbo, F. A., Owolabi, M. O., Patton, G. C., Pourmalek, F., Qorbani, M., Rafay, A., Rai, R. K., Ranabhat, C. L., Reinig, N., Safiri, S., Salomon, J. A., Sanabria, J. R., Santos, I. S., Sartorius, B., Sawhney, M., Schmidhuber, J., Schutte, A. E., Schmidt, M. I., Sepanlou, S. G., Shamsizadeh, M., Sheikhbahaei, S., Shin, M. J., Shiri, R., Shiue, I., Roba, H. S., Silva, D. A. S., Silverberg, J. I., Singh, J. A., Stranges, S., Swaminathan, S., Tabares-Seisdedos, R., Tadese, F., Tedla, B. A., Tegegne, B. S., Terkawi, A. S., Thakur, J. S., Tonelli, M., Topor-Madry, R., Tyrovolas, S., Ukwaja, K. N., Uthman, O. A., Vaezghasemi, M., Vasankari, T., Vlassov, V. V., Vollset, S. E., Weiderpass, E., Werdecker, A., Wesana, J., Westerman, R., Yano, Y., Yonemoto, N., Yonga, G., Zaidi, Z., Zenebe, Z. M., Zipkin, B., & Murray, C. J. L. (2017). Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *N Engl J Med, 377*(1), 13-27. doi:10.1056/NEJMoa1614362

Al-Shanqeeti, A., van Hylckama Vlieg, A., Berntorp, E., Rosendaal, F. R., & Broze, G. J., Jr. (2005). Protein Z and protein Z-dependent protease inhibitor. Determinants of levels and risk of venous thrombosis. *Thromb Haemost, 93*(3), 411-413. doi:10.1160/th04-11-0715

Bafunno, V., Santacroce, R., & Margaglione, M. (2011). The risk of occurrence of venous thrombosis: focus on protein Z. *Thromb Res, 128*(6), 508-515. doi:10.1016/j.thromres.2011.08.007

Braekkan, S. K., Mathiesen, E. B., Njolstad, I., Wilsgaard, T., Stormer, J., & Hansen, J. B. (2010). Mean platelet volume is a risk factor for venous thromboembolism: the Tromso Study, Tromso, Norway. *J Thromb Haemost, 8*(1), 157-162. doi:10.1111/j.1538-7836.2009.03498.x

Bruzelius, M., Bottai, M., Sabater-Lleal, M., Strawbridge, R. J., Bergendal, A., Silveira, A., Sundstrom, A., Kieler, H., Hamsten, A., & Odeberg, J. (2015). Predicting venous thrombosis in women using a combination of genetic markers and clinical risk factors. *J Thromb Haemost, 13*(2), 219-227. doi:10.1111/jth.12808

Bruzelius, M., Iglesias, M. J., Hong, M. G., Sanchez-Rivera, L., Gyorgy, B., Souto, J. C., Franberg, M., Fredolini, C., Strawbridge, R. J., Holmstrom, M., Hamsten, A., Uhlen, M., Silveira, A., Soria, J. M., Smadja, D. M., Butler, L. M., Schwenk, J. M., Morange, P. E., Tregouet, D. A., & Odeberg, J. (2016). PDGFB, a new candidate plasma biomarker for venous thromboembolism: results from the VEREMA affinity proteomics study. *Blood, 128*(23), e59-e66. doi:10.1182/blood-2016-05-711846

Christiansen, S. C., Naess, I. A., Cannegieter, S. C., Hammerstrom, J., Rosendaal, F. R., & Reitsma, P. H. (2006). Inflammatory cytokines as risk factors for a first venous thrombosis: a prospective population-based study. *PLoS Med, 3*(8), e334. doi:10.1371/journal.pmed.0030334

Cohen, A. T., Tapson, V. F., Bergmann, J. F., Goldhaber, S. Z., Kakkar, A. K., Deslandes, B., Huang, W., Zayaruzny, M., Emery, L., Anderson, F. A., Jr., & Investigators, E. (2008). Venous thromboembolism risk and prophylaxis in the acute hospital care setting (ENDORSE study): a multinational cross-sectional study. *Lancet, 371*(9610), 387-394. doi:10.1016/S0140-6736(08)60202-0

Cominetti, O., Nunez Galindo, A., Corthesy, J., Oller Moreno, S., Irincheeva, I., Valsesia, A., Astrup, A., Saris, W. H., Hager, J., Kussmann, M., & Dayon, L. (2016). Proteomic Biomarker Discovery in 1000 Human Plasma Samples with Mass Spectrometry. *J Proteome Res, 15*(2), 389-399. doi:10.1021/acs.jproteome.5b00901

Dayon, L., Nunez Galindo, A., Corthesy, J., Cominetti, O., & Kussmann, M. (2014). Comprehensive and Scalable Highly Automated MS-Based Proteomic Workflow for Clinical Biomarker Discovery in Human Plasma. *J Proteome Res*. doi:10.1021/pr500635f

de Haan, H. G., Bezemer, I. D., Doggen, C. J., Le Cessie, S., Reitsma, P. H., Arellano, A. R., Tong, C. H., Devlin, J. J., Bare, L. A., Rosendaal, F. R., & Vossen, C. Y. (2012). Multiple SNP testing improves risk prediction of first venous thrombosis. *Blood, 120*(3), 656-663. doi:10.1182/blood-2011-12-397752

Demler, O. V., Pencina, M. J., & D'Agostino, R. B., Sr. (2013). Impact of correlation on predictive ability of biomarkers. *Stat Med, 32*(24), 4196-4210. doi:10.1002/sim.5824

DeRoo, E. P., Wrobleski, S. K., Shea, E. M., Al-Khalil, R. K., Hawley, A. E., Henke, P. K., Myers, D. D., Jr., Wakefield, T. W., & Diaz, J. A. (2015). The role of galectin-3 and galectin-3-binding protein in venous thrombosis. *Blood, 125*(11), 1813-1821. doi:10.1182/blood-2014-04-569939

Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods, 4*(3), 207-214. doi:10.1038/nmeth1019

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom, 5*(11), 976-989. doi:10.1016/1044-0305(94)80016-2

Fashanu, O. E., Heckbert, S. R., Aguilar, D., Jensen, P. N., Ballantyne, C. M., Basu, S., Hoogeveen, R. C., deFilippi, C., Cushman, M., & Folsom, A. R. (2017). Galectin-3 and Venous Thromboembolism Incidence: the Atherosclerosis Risk in Communities (ARIC) Study. *Res Pract Thromb Haemost, 1*(2), 223-230. doi:10.1002/rth2.12038

Ferlay J, S. I., Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. (2012). GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. *Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr, accessed on 15/12/2017*.

Folsom, A. R., Tang, W., Weng, L. C., Roetker, N. S., Cushman, M., Basu, S., & Pankow, J. S. (2016). Replication of a genetic risk score for venous thromboembolism in whites but not in African Americans. *J Thromb Haemost, 14*(1), 83-88. doi:10.1111/jth.13193

Germain, M., Chasman, D. I., de Haan, H., Tang, W., Lindstrom, S., Weng, L. C., de Andrade, M., de Visser, M. C., Wiggins, K. L., Suchon, P., Saut, N., Smadja, D. M., Le Gal, G., van Hylckama Vlieg, A., Di Narzo, A., Hao, K., Nelson, C. P., Rocanin-Arjo, A., Folkersen, L., Monajemi, R., Rose, L. M., Brody, J. A., Slagboom, E., Aissi, D., Gagnon, F., Deleuze, J. F., Deloukas, P., Tzourio, C., Dartigues, J. F., Berr, C., Taylor, K. D., Civelek, M., Eriksson, P., Psaty, B. M., Houwing-Duitermaat, J., Goodall, A. H., Cambien, F., Kraft, P., Amouyel, P., Samani, N. J., Basu, S., Ridker, P. M., Rosendaal, F. R., Kabrhel, C., Folsom, A. R., Heit, J., Reitsma, P. H., Tregouet, D. A., Smith, N. L., & Morange, P. E. (2015). Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet, 96*(4), 532-542. doi:10.1016/j.ajhg.2015.01.019

Greene, M. T., Spyropoulos, A. C., Chopra, V., Grant, P. J., Kaatz, S., Bernstein, S. J., & Flanders, S. A. (2016). Validation of Risk Assessment Models of Venous Thromboembolism in Hospitalized Medical Patients. *Am J Med, 129*(9), 1001.e1009-1001.e1018. doi:10.1016/j.amjmed.2016.03.031

Han, X., Fiehler, R., & Broze, G. J., Jr. (1998). Isolation of a protein Z-dependent plasma protease inhibitor. *Proc Natl Acad Sci U S A, 95*(16), 9250-9255.

Heikal, N. M., Murphy, K. K., Crist, R. A., Wilson, A. R., Rodgers, G. M., & Smock, K. J. (2013). Elevated factor IX activity is associated with an increased odds ratio for both arterial and venous thrombotic events. *Am J Clin Pathol, 140*(5), 680-685. doi:10.1309/ajcpagor4q2iikug

Heit, J. A. (2015). Epidemiology of venous thromboembolism. *Nat Rev Cardiol, 12*(8), 464-474. doi:10.1038/nrcardio.2015.83

Huang, W., Goldberg, R. J., Anderson, F. A., Kiefe, C. I., & Spencer, F. A. (2014). Secular trends in occurrence of acute venous thromboembolism: the Worcester VTE study (1985-2009). *Am J Med, 127*(9), 829-839 e825. doi:10.1016/j.amjmed.2014.03.041

Huang, X., Liu, B., Wei, Y., Beyea, R., Yan, H., & Olson, S. T. (2017). Lipid oxidation inactivates the anticoagulant function of protein Z-dependent protease inhibitor (ZPI). *J Biol Chem, 292*(35), 14625-14635. doi:10.1074/jbc.M117.793901

Jacobsen, B. K., Eggen, A. E., Mathiesen, E. B., Wilsgaard, T., & Njolstad, I. (2012). Cohort profile: the Tromso Study. *Int J Epidemiol, 41*(4), 961-967. doi:10.1093/ije/dyr049

Karasu, A., Baglin, T. P., Luddington, R., Baglin, C. A., & van Hylckama Vlieg, A. (2016). Prolonged clot lysis time increases the risk of a first but not recurrent venous thrombosis. *Br J Haematol, 172*(6), 947-953. doi:10.1111/bjh.13911

Koide-Yoshida, S., Niki, T., Ueda, M., Himeno, S., Taira, T., Iguchi-Ariga, S. M., Ando, Y., & Ariga, H. (2007). DJ-1 degrades transthyretin and an inactive form of DJ-1 is secreted in familial amyloidotic polyneuropathy. *Int J Mol Med, 19*(6), 885-893.

Lapek, J. D., Jr., Lewinski, M. K., Wozniak, J. M., Guatelli, J., & Gonzalez, D. J. (2017). Quantitative Temporal Viromics of an Inducible HIV-1 Model Yields Insight to Global Host Targets and Phospho-Dynamics Associated with Protein Vpr. *Mol Cell Proteomics, 16*(8), 1447-1461. doi:10.1074/mcp.M116.066019

Mahan, C. E., Burnett, A. E., Fletcher, M., & Spyropoulos, A. C. (2017). Extended thromboprophylaxis in the acutely Ill medical patient after hospitalization - a paradigm shift in post-discharge thromboprophylaxis. *Hosp Pract (1995)*. doi:10.1080/21548331.2018.1410053

Mayer, R. S., Streiff, M. B., Hobson, D. B., Halpert, D. E., & Berenholtz, S. M. (2011). Evidence-based venous thromboembolism prophylaxis is associated with a six-fold decrease in numbers of symptomatic venous thromboembolisms in rehabilitation inpatients. *PM R, 3*(12), 1111-1115 e1111. doi:10.1016/j.pmrj.2011.07.022

McAlister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C., Kuhn, K., Pike, I., Grothe, R. A., Blethrow, J. D., & Gygi, S. P. (2012). Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal Chem, 84*(17), 7469-7478. doi:10.1021/ac301572t

McAlister, G. C., Nusinow, D. P., Jedrychowski, M. P., Wuhr, M., Huttlin, E. L., Erickson, B. K., Rad, R., Haas, W., & Gygi, S. P. (2014). MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem, 86*(14), 7150-7158. doi:10.1021/ac502040v

Meltzer, M. E., Lisman, T., de Groot, P. G., Meijers, J. C., le Cessie, S., Doggen, C. J., & Rosendaal, F. R. (2010). Venous thrombosis risk associated with plasma hypofibrinolysis is explained by elevated plasma levels of TAFI and PAI-1. *Blood, 116*(1), 113-121. doi:10.1182/blood-2010-02-267740

Miletich, J. P., & Broze, G. J., Jr. (1987). Human plasma protein Z antigen: range in normal subjects and effect of warfarin therapy. *Blood, 69*(6), 1580-1586.

Mitsugi, H., Niki, T., Takahashi-Niki, K., Tanimura, K., Yoshizawa-Kumagaye, K., Tsunemi, M., Iguchi-Ariga, S. M., & Ariga, H. (2013). Identification of the recognition sequence and target proteins for DJ-1 protease. *FEBS Lett, 587*(16), 2493-2499. doi:10.1016/j.febslet.2013.06.032

Norgaard, I., Nielsen, S. F., & Nordestgaard, B. G. (2016). Complement C3 and High Risk of Venous Thromboembolism: 80 517 Individuals from the Copenhagen General Population Study. *Clin Chem*. doi:10.1373/clinchem.2015.251314

Park, M. S., Spears, G. M., Bailey, K. R., Xue, A., Ferrara, M. J., Headlee, A., Dhillon, S. K., Jenkins, D. H., Zietlow, S. P., Harmsen, W. S., Ashrani, A. A., & Heit, J. A. (2017). Thrombin generation profiles as predictors of symptomatic venous thromboembolism after trauma: A prospective cohort study. *J Trauma Acute Care Surg, 83*(3), 381-387. doi:10.1097/ta.0000000000001466

Pepin, M., Kleinjan, A., Hajage, D., Buller, H. R., Di Nisio, M., Kamphuisen, P. W., Salomon, L., Veyradier, A., Stepanian, A., & Mahe, I. (2016). ADAMTS-13 and von Willebrand factor predict venous thromboembolism in patients with cancer. *J Thromb Haemost, 14*(2), 306-315. doi:10.1111/jth.13205

Pettersson, T., Carlstrom, A., & Jornvall, H. (1987). Different types of microheterogeneity of human thyroxine-binding prealbumin. *Biochemistry, 26*(14), 4572-4583.

Puurunen, M. K., Enserro, D., Xanthakis, V., Larson, M. G., Benjamin, E. J., Tofler, G. H., Wollert, K. C., O'Donnell, C. J., & Vasan, R. S. (2016). Biomarkers for the prediction of venous

thromboembolism in the community. *Thrombosis Research, 145*(Supplement C), 34-39. doi:https://doi.org/10.1016/j.thromres.2016.07.006

Reitsma, P. H., & Rosendaal, F. R. (2004). Activation of innate immunity in patients with venous thrombosis: the Leiden Thrombophilia Study. *J Thromb Haemost, 2*(4), 619-622. doi:10.1111/j.1538-7836.2004.00689.x

Richarme, G., Mihoub, M., Dairou, J., Bui, L. C., Leger, T., & Lamouri, A. (2015). Parkinsonism-associated protein DJ-1/Park7 is a major protein deglycase that repairs methylglyoxal- and glyoxal-glycated cysteine, arginine, and lysine residues. *J Biol Chem, 290*(3), 1885-1897. doi:10.1074/jbc.M114.597815

Ridker, P. M., Cushman, M., Stampfer, M. J., Tracy, R. P., & Hennekens, C. H. (1997). Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *N Engl J Med, 336*(14), 973-979. doi:10.1056/nejm199704033361401

Rosendaal, F. R. (1999). Venous thrombosis: a multicausal disease. *Lancet, 353*(9159), 1167-1173.

Saghazadeh, A., & Rezaei, N. (2016). Inflammation as a cause of venous thromboembolism. *Crit Rev Oncol Hematol, 99*, 272-285. doi:10.1016/j.critrevonc.2016.01.007

Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B., & Bantscheff, M. (2015). A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol Cell Proteomics, 14*(9), 2394-2404. doi:10.1074/mcp.M114.046995

Simone, B., De Stefano, V., Leoncini, E., Zacho, J., Martinelli, I., Emmerich, J., Rossi, E., Folsom, A. R., Almawi, W. Y., Scarabin, P. Y., den Heijer, M., Cushman, M., Penco, S., Vaya, A., Angchaisuksiri, P., Okumus, G., Gemmati, D., Cima, S., Akar, N., Oguzulgen, K. I., Ducros, V., Lichy, C., Fernandez-Miranda, C., Szczeklik, A., Nieto, J. A., Torres, J. D., Le Cam-Duchez, V., Ivanov, P., Cantu-Brito, C., Shmeleva, V. M., Stegnar, M., Ogunyemi, D., Eid, S. S., Nicolotti, N., De Feo, E., Ricciardi, W., & Boccia, S. (2013). Risk of venous thromboembolism associated with single and combined effects of Factor V Leiden, Prothrombin 20210A and Methylenetethraydrofolate reductase C677T: a meta-analysis involving over 11,000 cases and 21,000 controls. *Eur J Epidemiol, 28*(8), 621-647. doi:10.1007/s10654-013-9825-8

Smith, N. L., Chen, M. H., Dehghan, A., Strachan, D. P., Basu, S., Soranzo, N., Hayward, C., Rudan, I., Sabater-Lleal, M., Bis, J. C., de Maat, M. P., Rumley, A., Kong, X., Yang, Q., Williams, F. M., Vitart, V., Campbell, H., Malarstig, A., Wiggins, K. L., Van Duijn, C. M., McArdle, W. L., Pankow, J. S., Johnson, A. D., Silveira, A., McKnight, B., Uitterlinden, A. G., Wellcome Trust Case Control, C., Aleksic, N., Meigs, J. B., Peters, A., Koenig, W., Cushman, M., Kathiresan, S., Rotter, J. I., Bovill, E. G., Hofman, A., Boerwinkle, E., Tofler, G. H., Peden, J. F., Psaty, B. M., Leebeek, F., Folsom, A. R., Larson, M. G., Spector, T. D., Wright, A. F., Wilson, J. F., Hamsten, A., Lumley, T., Witteman, J. C., Tang, W., & O'Donnell, C. J. (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*(12), 1382-1392. doi:10.1161/CIRCULATIONAHA.109.869156

Sofi, F., Cesari, F., Abbate, R., Gensini, G. F., Broze, G., Jr., & Fedi, S. (2010). A meta-analysis of potential risks of low levels of protein Z for diseases related to vascular thrombosis. *Thromb Haemost, 103*(4), 749-756. doi:10.1160/TH09-09-0645

Tanskanen, M., Peuralinna, T., Polvikoski, T., Notkola, I. L., Sulkava, R., Hardy, J., Singleton, A., Kiuru-Enari, S., Paetau, A., Tienari, P. J., & Myllykangas, L. (2008). Senile systemic amyloidosis affects 25% of the very aged and associates with genetic variation in alpha2-macroglobulin

and tau: a population-based autopsy study. *Ann Med, 40*(3), 232-239. doi:10.1080/07853890701842988

Thrombosis: a major contributor to the global disease burden. (2014). *J Thromb Haemost, 12*(10), 1580-1590. doi:10.1111/jth.12698

Ting, L., Rad, R., Gygi, S. P., & Haas, W. (2011). MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods, 8*(11), 937-940. doi:10.1038/nmeth.1714

Tsai, A. W., Cushman, M., Rosamond, W. D., Heckbert, S. R., Tracy, R. P., Aleksic, N., & Folsom, A. R. (2002). Coagulation factors, inflammation markers, and venous thromboembolism: the longitudinal investigation of thromboembolism etiology (LITE). *Am J Med, 113*(8), 636-642.

van Es, N., Di Nisio, M., Cesarman, G., Kleinjan, A., Otten, H. M., Mahe, I., Wilts, I. T., Twint, D. C., Porreca, E., Arrieta, O., Stepanian, A., Smit, K., De Tursi, M., Bleker, S. M., Bossuyt, P. M., Nieuwland, R., Kamphuisen, P. W., & Buller, H. R. (2017). Comparison of risk prediction scores for venous thromboembolism in cancer patients: a prospective cohort study. *Haematologica, 102*(9), 1494-1501. doi:10.3324/haematol.2017.169060

van Hylckama Vlieg, A., Baglin, C. A., Luddington, R., MacDonald, S., Rosendaal, F. R., & Baglin, T. P. (2015). The risk of a first and a recurrent venous thrombosis associated with an elevated D-dimer level and an elevated thrombin potential: results of the THE-VTE study. *J Thromb Haemost, 13*(9), 1642-1652. doi:10.1111/jth.13043

van Hylckama Vlieg, A., van der Linden, I. K., Bertina, R. M., & Rosendaal, F. R. (2000). High levels of factor IX increase the risk of venous thrombosis. *Blood, 95*(12), 3678-3682.

van Montfoort, M. L., Stephan, F., Lauw, M. N., Hutten, B. A., Van Mierlo, G. J., Solati, S., Middeldorp, S., Meijers, J. C., & Zeerleder, S. (2013). Circulating nucleosomes and neutrophil activation as risk factors for deep vein thrombosis. *Arterioscler Thromb Vasc Biol, 33*(1), 147-151. doi:10.1161/atvbaha.112.300498

Wang, Y., Gao, H., Kessinger, C. W., Schmaier, A., Jaffer, F. A., & Simon, D. I. (2017). Myeloid-related protein-14 regulates deep vein thrombosis. *JCI Insight, 2*(11). doi:10.1172/jci.insight.91356

Wautier, J. L., & Wautier, M. P. (2013). Molecular basis of erythrocyte adhesion to endothelial cells in diseases. *Clin Hemorheol Microcirc, 53*(1-2), 11-21. doi:10.3233/ch-2012-1572