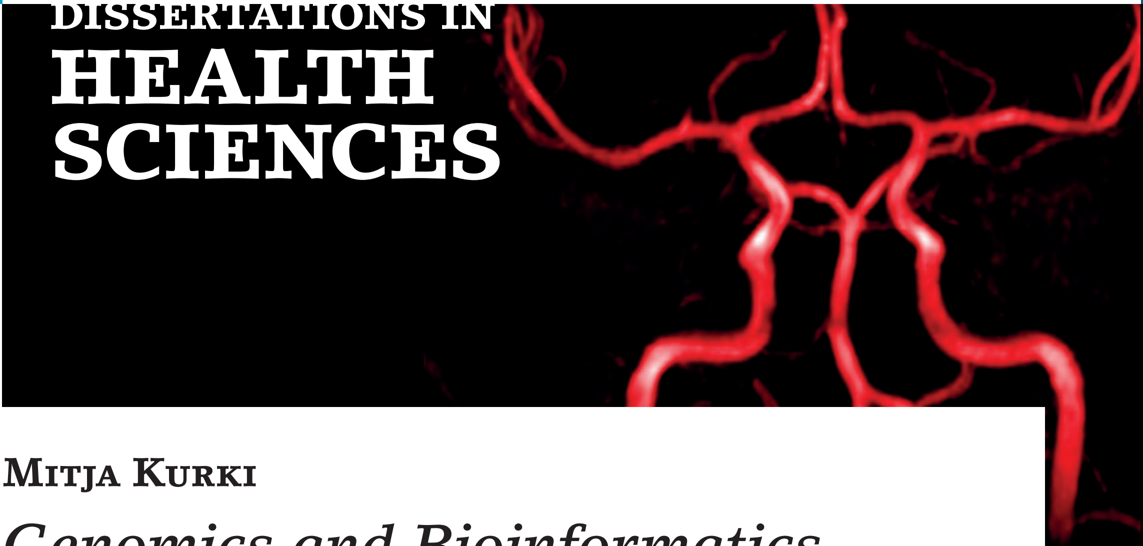
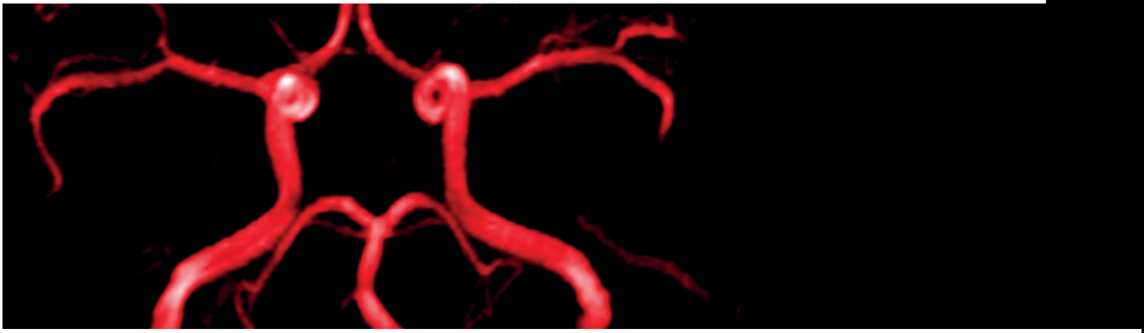


DISSERTATIONS IN
HEALTH
SCIENCES

Mitja Kurki

Genomics and Bioinformatics Approaches in Search of Molecular Pathomechanisms of Saccular Intracranial Aneurysm, A Complex Disease



Publications of the University of Eastern Finland
Dissertations in Health Sciences

UNIVERSITY OF
EASTERN FINLAND

MITJA KURKI

Genomics and Bioinformatics Approaches in Search of Molecular Pathomechanisms of Saccular Intracranial Aneurysm, A Complex Disease

To be presented by permission of the Faculty of Health Sciences, University of Eastern Finland for public examination in Auditorium 1, Kuopio University Hospital, Kuopio, on Saturday, 30.11, at 13

Publications of the University of Eastern Finland
Dissertations in Health Sciences
202

Neurosurgery, Institute of Clinical Medicine, School of Medicine,
Department of Neurobiology, A.I. Virtanen Institute, Faculty of Health Sciences, University of Eastern Finland
Kuopio
2013

Kopijyvä Oy
Kuopio, 2013
Finland

Series Editors:
Professor Veli-Matti Kosma, M.D., Ph.D.
Institute of Clinical Medicine, Pathology
Faculty of Health Sciences

Professor Hannele Turunen, Ph.D.
Department of Nursing Science
Faculty of Health Sciences

Professor Olli Gröhn, Ph.D.
A.I. Virtanen Institute for Molecular Sciences
Faculty of Health Sciences

Professor Kai Kaarniranta, M.D., Ph.D.
Institute of Clinical Medicine, Ophthalmology
Faculty of Health Sciences

Lecturer Veli-Pekka Ranta, Ph.D. (pharmacy)
School of Pharmacy
Faculty of Health Sciences

Distributor:
University of Eastern Finland
Kuopio Campus Library
P.O.Box 1627
FI-70211 Kuopio, Finland
http://www.uef.fi/kirjasto

ISBN (print): 978-952-61-1310-4
ISBN (pdf): 978-952-61-1311-1
ISSN (print): 1798-5706
ISSN (pdf): 1798-5714
ISSN-L: 1798-5706

Author’s address: Neurosurgery of NeuroCenter
Kuopio University Hospital
KUOPIO
FINLAND

Supervisors: Research Director Garry Wong, Ph.D.
Department of Neurobiology
Faculty of Health Sciences
A.I. Virtanen Institute for Molecular Sciences
University of Eastern Finland
KUOPIO
FINLAND

Professor Aarno Palotie, M.D., Ph.D.
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki
HELSINKI
FINLAND

Professor Juha E. Jääskeläinen, M.D., Ph.D.
Neurosurgery, University of Eastern Finland
Neurosurgery, NeuroCenter, Kuopio University Hospital
KUOPIO
FINLAND

Reviewers: Docent Iiris Hovatta, Ph.D.
Department of Biosciences
University of Helsinki
HELSINKI
FINLAND

Docent Tero Aittokallio, Ph.D.
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki
HELSINKI
FINLAND

Opponent: Professor Miikka Vikkula, M.D., Ph.D.
de Duve Institute and Université Catholique de Louvain
BRUSSELS
BELGIUM

Kurki, Mitja
Genomics and Bioinformatics Approaches in Search of Molecular Pathomechanisms of Saccular Intracranial Aneurysm, A Complex Disease.
University of Eastern Finland, Faculty of Health Sciences, 2013.
Publications of the University of Eastern Finland. Dissertations in Health Sciences Number 202. 2013. 122 p.
ISBN (print): 978-952-61-1310-4
ISBN (pdf): 978-952-61-1311-1
ISSN (print): 1798-5706
ISSN (pdf): 1798-5714
ISSN-L: 1798-5706

ABSTRACT

Despite massive efforts to elucidate the genetic and molecular basis of common complex diseases, definitive answers remain elusive. This thesis focuses on the molecular mechanisms of saccular intracranial aneurysm (sIA), a complex trait. The rupture of sIA is the leading cause of aneurysmal subarachnoid hemorrhage, a devastating form of stroke. High-throughput genomic methods can generate a simultaneous assessment of the activity of the human genome, but a drawback is huge amounts of data that pose great challenges for analysis and interpretation, requiring bioinformatic methods. Three complementary approaches utilizing high-throughput genomics and bioinformatics are applied in this thesis. First, differences in genome-wide gene expression profiles between ruptured and unruptured sIA walls were studied, and a number of potential pathways and genes associated with the sIA rupture were identified. Second, a novel bioinformatics method and software was developed, aiding in the interpretation of biological mechanisms reflected by the differentially expressed genes in the sIA walls. Using our novel method, we generated hypotheses about potential links between transcription factors controlling detrimental processes in the sIA walls. Third, we identified four novel sIA loci in Finnish and Dutch sIA samples using genome-wide association analysis. In summary, we identified candidate genes and pathways, which can serve as a basis for future research aiming towards novel diagnostics, preventions, or therapies of sIA disease. Additionally, the developed novel bioinformatic method and software can also be used to study other complex phenotypes.

National Library of Medical Classification: QU 460; QU 26.5; WL 355 ; QU 460; QU 465; QU 550.5.G

Medical Subject Headings: Bioinformatics; Genetics; Genomics; Intracranial Aneurysm; Subarachnoid Hemorrhage; Gene Expression Profiling; Genome-Wide Association Study

Kurki, Mitja
Genomisia ja bioinformatiivisiä menetelmiä monitekijäisen sakkulaarisen intrakraniaalisen aivovaltimoaneurysmataudin molekulaaristen mekanismien tutkimuksessa.
Itä-Suomen yliopisto, terveystieteiden tiedekunta, 2013
Publications of the University of Eastern Finland. Dissertations in Health Sciences Numero 202. 2013. 122 s.

ISBN (print): 978-952-61-1310-4
ISBN (pdf): 978-952-61-1311-1
ISSN (print): 1798-5706
ISSN (pdf): 1798-5714
ISSN-L: 1798-5706

TIIVISTELMÄ

Yleisten monitekijäisten tautien molekulaariset mekanismit eivät ole täysin tunnettuja, vaikka niiden tutkimukseen on maailmanlaajuisesti panostettu huomattavasti. Tässä väitöskirjassa keskitytään sakkulaarisen intrakraniaalisen aivovaltimoaneurysman (sIA) molekulaaristen mekanismien tutkimukseen. sIA on monitekijäinen tauti, jossa muodostuneen aneurysman puhkeaminen aiheuttaa hengenvaarallisen lukinkalvonalaisen verenvuodon. Modernit korkean kapasiteetin genomiset menetelmät mahdollistavat mm. kaikkien ihmisen geenien aktiivisuuden mittaamisen samanaikaisesti. Nämä genomiset menetelmät tuottavat valtavasti mittaustietoa, jonka analysointi ja tulkinta on usein haasteellista. Bioinformatiikka tieteenalana soveltaa ja kehittää genomisen tiedon analysointi- ja tulkintamenetelmiä. Tässä väitöskirjatyössä käytetiin kolmea toisiaan täydentävää lähestymistapaa, joiden avulla pyrittiin ymmärtämään sIA:n muodostumiseen ja puhkeamiseen vaikuttavia molekulaarisia mekanismeja. Aluksi vertailtiin vuotaneiden ja vuotamattomien aneurysmien geenien ilmentymisprofiileja genominlaajuisesti ja tunnistettiin useita vuotoon assosioituneita signalointireittejä ja geenejä. Seuraavaksi kehitettiin bioinformatiivinen menetelmä ja ohjelmisto, joita soveltamalla ensimmäisen osatyön mittaustietoon perustuen luotiin uusia hypoteeseja sIA:n puhkeamiseen liittyvistä mekanismeista. Lopuksi, käyttäen genominlaajuista assosiaatioanalyysiä suomalaisiin ja hollantilaisiin näytteisiin, tunnistettiin neljä uutta geneettistä muunnosta, jotka assosioituvat sIA:n muodostumisriskiin. Yhteenvetona voidaan todeta, että väitöskirjatyössä tunnistettiin useita kandidaattigeenejä ja signalointireittejä, jotka voivat toimia pohjana tutkimuksille, joiden tähtäimessä on uusia sIA-taudin diagnostiikka-, esto- ja hoitomenetelmiä. Kehitetty bioinformatiivinen menetelmä soveltuu lisäksi myös muiden kompleksien fenotyyppien tutkimiseen.

Luokitus: QU 460; QU 26.5; WL 355 ; QU 460; QU 465; QU 550.5.G
Yleinen Suomalainen asiasanasto: Bioinformatiikka; Genomiikka; Perinnöllisyystiede; Aneurysma; Tiedonlouhinta; aivoverenkiertohäiriöt ; populaatiogenetiikka

“Suppose you succeed in breaking the wall with your head.
And what, then, will you do in the next cell?”
Stanisław Jerzy Lec

"Only two things are infinite, the universe and human stupidity, and I'm not sure about the former."
Albert Einstein

Acknowledgements

What just happened? I was just a guy happily writing software for a living, without a clue about anyting biological. Now I am becoming a Ph.D. in molecular medicine/bioinformatics. I did not plan this!

The work presented in this thesis began in the A.I Virtanen Institute and in the Graduate School of Molecular Medicine, the University of Kuopio / University of Eastern Finland and the final part was carried out in Neurosurgery, NeuroCenter in Kuopio University Hospital.

I am honoured that Professor Miikka Vikkula organized time in his busy schedule and accepted the invitation to act as the opponent in my thesis.

I want to thank adjunct professors Iiris Hovatta and Tero Aittokallio for pre-reviewing my thesis. Your comments and suggestions really helped to improve the scientific content as well as the readability of this thesis.

My deepest gratitudes go to my supervisors Professor Juha Jääskeläinen, Professor Aarno Palotie and Research Director Garry Wong. Your collective extensive experience in diverse fields of expertise created a stimulating and educational support network. Garry first inspired and mentored me in the field of bioinformatics and basic science in general. The serendipituous meeting with Juha geared my work and interest towards more direct medical application of genomics and bioinformatics in general and in saccular intracranial aneurysm disease in particular. There is so much more in the details than just the devil. When my work took a genetic twist, I was fortunate to be able to visit the Sanger Institute in Cambridge UK and meet Aarno there. Aarno took me on a fast lane of learning in human complex disease genetics.

I wish to express my heartfelt thanks to all of the co-authors in the manuscripts in this thesis. Especially noteworthy were the contributions of Ph.D. Sanna-Kaisa Häkkinen for all the excellent lab-work for the transcriptomic data utilized in two of the studies in this thesis and Professor Seppo Ylä-Herttuala for making the transcriptomic analyses possible in his lab as well as being extremely supportive and encouraging.

This work would not have been possibly without the utmost important clinical sample and cohort collections: aneurysmal tissue collection in Helsinki Neurosurgery; aneurysm patient databases and DNA collections in Neurogurgery, Kuopio University Hospital; aneurysm cohort in Utrech Medical Center; and DNA collections of Finnish population cohorts Helsinki Birth Cohort Study, Cardiovascular Risk in Young Finns Study and the Health2000 cohort. My sincere thanks goes to all of the participants in these cohorts and the researchers and clinicians involved in the collection of these cohorts.

I wish to thank all of the people I have worked with or who have created an amazing environment to carry out my research in Garry Wong's lab in the A.I.Virtanen Institute, University of Eastern Finland; NeuroCenter of Kuopio University Hospital (KUH); Neurosurgery of Helsinki University Central Hospital (HUCH), and Finnish Institute of Molecular Medicine (FIMM).

Special thanks go to secretaries Seija Kasanen at KUH and Sari Kivikko at FIMM for all of your assistance during these years. The IT personnel, especially Tomi Simonen, at FIMM are thanked for excellent tech-support for the FIMM computer cluster.

High fives to Benewayn koulupojat Harri, Riku, Pasi, Marko, Antti, Mikko, Juha for the support and the possibility to work part-time in the early days of my deviation towards bioinformatics. Those years were tough at times but memorable and educational. Tullaan Kiinaan kiroilemaan!

Wohoo to the family of molecular biology for giving a break at the spring!

I gratefully acknowledge the Academy of Finland, Sydän Sykkii Savolle-fund, and the Graduate School of Molecular medicine, University of Eastern Finland for financial support, which made this work possible.

The nature vs. nurture debate is not an issue when talking about family as both components largely come from the same source. Thank you for those to mom and dad. My sister Sina was quite annoying at first (you laughed at me when you first saw me!!) but I thank you for the friendship we have developed since then.

A big paw goes to Rufus “The Big Purr” and Rudolf “Pakettimarsu” for lowering my blood pressure in critical times during the process. Finally, the greatest support that anybody could hope for has come from Tytteli. I want to thank you for putting up with so many of ”one of my finest qualities”, for being the voice of reason at the most tough times and for your unconditional love. You are the sunshine of my life!

Kuopio, October 2013

Mitja

List of the original publications

This dissertation is based on the following original publications:

I Kurki MI*, Häkkinen SK*, Frösen J, Tulamo R, von und zu Fraunberg M, Wong G, Tromp G, Niemelä M, Hernesniemi J, Jääskeläinen JE, Ylä-Herttuala S. Upregulated signaling pathways in ruptured human saccular intracranial aneurysm wall: an emerging regulative role of Toll-like receptor signaling and nuclear factor-κB, hypoxia-inducible factor-1A, and ETS transcription factors. Neurosurgery 68:1667-75, 2011

II Kurki MI, Paananen J, Storvik M, Ylä-Herttuala S, Jääskeläinen JE, von und zu Fraunberg M, Wong G, Pehkonen P. TAFFEL: Independent Enrichment Analysis of Gene Sets. BMC Bioinformatics 12:171, 2011

III Kurki MI, Gaál E, Kettunen J, Anttila V, van't Hof FNG, von und zu Fraunberg M, Helisalmi S, Hiltunen M, Lehto H, Kivisaari R, Koivisto T, Ronkainen A, Rinne J, Kiemeney LALM, Vermeulen SH, Eriksson JG, Aromaa A, Gunel M, Lehtimäki T, Raitakari OT, Salomaa V, Ruigrok YM, Rinkel GJE, Niemelä M, Hernesniemi J, de Bakker PIW, Ripatti S, Perola M, Palotie A, Jääskeläinen JE. High-risk population isolate reveals low frequency variants predisposing to saccular intracranial aneurysms. Submitted.

The publications were adapted with the permission of the copyright owners. This thesis also contains unpublished data.

* Equal first authorship

Contents

1 INTRODUCTION 1

2 REVIEW OF THE LITERATURE 3

2.1 THE HUMAN GENOME 3

2.1.1 Variation in human genome 3

2.1.2 Linkage disequilibrium 4

2.1.3 Genome variation catalogues 6

2.1.4 Studying human genome variation 9

2.1.4.1 Linkage studies 9

2.1.4.2 Candidate gene association studies 9

2.1.4.3 Genome-wide association studies 10

2.1.4.4 Missing heritability 13

2.1.4.5 Sequencing approaches in genomic studies 13

2.1.4.6 Genotype imputation 14

2.2 THE HUMAN TRANSCRIPTOME 16

2.2.1 Regulation of gene expression 16

2.2.2 Studying the transcriptome 19

2.3 COMPLEX DISEASES 21

2.4 SACCULAR INTRACRANIAL ANEURYSM (SIA) DISEASE 22

2.4.1 Epidemiology 23

2.4.2 Risk factors 23

2.4.3 Molecular biology of intracranial extracerebral arteries and sIA wall 24

2.4.4 Genetics of the sIA disease 25

2.4.5 Whole genome expression profiling of human sIA tissue 29

2.5 CHALLENGES AND BIOINFORMATIC SOLUTIONS FOR INTERPRETATION OF RESULTS FROM HIGH THROUGHPUT ANALYSES 31

2.5.1 Biological body of knowledge and digital representation 31

2.5.2 Pathway analysis methods 36

2.5.2.1 Overrepresentation analyses 36

2.5.2.2 Enrichment analyses 37

2.5.2.3 Topology based pathway analysis 38

2.5.3 Challenges and limitations of pathway analysis 39

2.5.4 Pathway analysis in GWA studies 39

2.5.5 Inferring regulation of differentially expressed genes 40

2.5.6 Clustering in high-throughput genomic data mining 40
2.5.7 Multiple testing burden 41

3 AIMS 43

4 UPREGULATED SIGNALING PATHWAYS IN RUPTURED HUMAN SACCULAR INTRACRANIAL ANEURYSM WALL: AN EMERGING REGULATIVE ROLE OF TOLL-LIKE RECEPTOR SIGNALING AND NUCLEAR FACTOR-KB, HYPOXIA-INDUCIBLE FACTOR-1A, AND ETS TRANSCRIPTION FACTORS 45

4.1 INTRODUCTION 45
4.1.1 Materials and methods 46
4.1.2 Patients and Tissue Samples 46
4.1.3 Isolation of mRNA and Microarray Hybridization 47
4.1.4 Microarray Data Analysis 47
4.1.5 Functional Analysis of Differentially Expressed Genes 48
4.1.6 *In Silico* Transcription Factor Analysis 48
4.1.7 Validation of Microarray Data 49
4.2 RESULTS 49
4.3 DISCUSSION 54
4.4 CONCLUSIONS 58

5 TAFFEL: INDEPENDENT ENRICHMENT ANALYSIS OF GENE SETS 59

5.1 INTRODUCTION 59
5.2 RESULTS 62
5.2.1 Description of the method and tool 62
5.2.2 Availability and running the program 63
5.2.3 Typical analysis flow 64
5.2.4 Analysis of forskolin effect on HEK293T cells 64
5.2.5 IEA suggests novel hypotheses of signaling mechanisms active in the ruptured sIA wall 64
5.2.6 Comparison of TAFFEL to other methods 68
5.3 DISCUSSION 68
5.4 CONCLUSIONS 69
5.5 MATERIALS AND METHODS 70
5.5.1 Annotation data sources 70
5.5.2 Gene clustering method 70

5.5.3 Selection of number of clusters ... 71

5.5.4 Cluster statistics ... 71

5.5.5 Processing of demonstration microarray data sets ... 72

6 HIGH RISK POPULATION ISOLATE REVEALS LOW FREQUENCY VARIANTS PREDISPOSING TO INTRACRANIAL ANEURYSMS ... 73

6.1 INTRODUCTION ... 73

6.2 RESULTS ... 75

6.2.1 Case vs. control analysis in Finnish and Dutch samples ... 75

6.2.2 Association of variants to the number of sIAs ... 79

6.2.3 Analysis of 2q33.1 locus ... 79

6.2.4 Regulatory elements at identified loci ... 80

6.2.5 eQTL analysis ... 81

6.3 DISCUSSION ... 83

6.3.1 Four novel sIA loci ... 83

6.3.2 2q23.3 locus ... 83

6.3.3 5q31.1 locus ... 84

6.3.4 6q24.2 locus ... 84

6.3.5 7p22.1 locus and the number of sIAs ... 85

6.3.6 Previously identified 9p21.3 locus ... 85

6.3.7 2q33.1 locus with previously inconclusive evidence ... 85

6.4 MATERIALS AND METHODS ... 86

6.4.1 Study samples ... 86

6.4.2 Replication strategy ... 88

6.4.3 Genotyping ... 88

6.4.4 Imputation ... 89

6.4.5 eQTL analysis ... 89

6.4.6 Regional association plots ... 89

6.4.7 Search of regulatory elements at identified variants ... 89

6.4.8 Statistical analysis ... 90

6.4.9 Heritability analysis ... 90

7 GENERAL DISCUSSION ... 91

7.1 HIGH-THROUGHPUT METHODS IN THE STUDY OF COMPLEX DISEASES ... 91

7.1.1 Strengths and weaknessess of applied high-throughput approaches in study of complex disease ... 93

7.1.2 Interpretation challenges of genome-wide studies ... 94

7.2 QUEST FOR MOLECULAR PATHOMECHANISMS OF SACCULAR INTRACRANIAL ANEURYSM DISEASE 96

7.3 COMMON FINDINGS IN DIFFFERENT APPROACHES IN THE CURRENT THESIS 99

7.4 NEAR FUTURE AVENUES FOR sIA RESEARCH 99

8 CONCLUSIONS 103

9 REFERENCES 105

Abbreviations

ADPKD	Autosomal dominant polycystic kidney disease
AIC	Akaike Information Criterion
AWS	Amazon Web Services
CAD	Coronary artery disease
CAF	Control allele frequency
cAMP	cyclic adenylate monophosphate
cDNA	Complementary deoxyribonucleic acid
CEU	HAPMAP project population. Utah residents with ancestry from northern and western Europe
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CI	Confidence Interval
CNV	Copy number variation
CREB	Cyclic AMP Response Element Binding
CT	Computed tomography
DAVID	for annotation, visualization, and integrated discovery
DE	differentially expressed
DEA	Dependent enrichment analysis
DNA	Deoxyribonucleic Acid
EA	Enrichment Analysis
EEL	External elastic layer
eQTL	Expression quantitative trait locus
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-wide association study
H2000	The Health 2000 Study
HBCS	Helsinki Birth Cohort Study
HGP	Human Genome Project
HIF1A	Hypoxia inducible factor-1A
IBD	Identity by descent
IDA	Inferred from direct assay
IEA	Independent Enrichment Analysis
IEL	Internal elastic layer
INDEL	Insertion or deletion
IT	Information technology
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage Disequilibrium

LPH	Lactase phlorizin hydrolase
MAF	Minor allele frequency
MCA	Middle cerebral artery
miRNA	Micro ribonucleic acid
mRNA	Messenger ribonucleic acid
NADPH	Nicotinamide adenine dinucleotide phosphate
NF-kB	Nuclear factor-kB
NGS	Next-generation sequencing
NMF	Non-negative matrix factorization
OBO	The Open Biological and Biomedical Ontologies
OR	Odds ratio
ORA	Overrepresentation analysis
QC	Quality control
QQ plot	Quantile-quantile plot
RCA	Reviewed computational analysis
RMA	Robust Multi-array Average
RNA	Ribonucleic acid
RNA-Seq	Ribonucleic acid sequencing
RR	Rate ratio
rRNA	ribosomal ribonucleic acid
SAH	Subarachnoid hemorrhage
sIA	saccular intracranial aneurysm
sIA-SAH	Subarachnoid hemorrhage from saccular intracranial aneurysm rupture
SNP	Single nucleotide polymorphism
SPIA	Signalling pathway impact analysis
SV	Structural variant
TF	Transcription factor
TFBS	Transcription factor binding site
TLR	Toll-like receptor
TP	Topology based
tRNA	Transfer ribonucleicacid
TSS	Transcription start site

1 Introduction

Common complex diseases such as coronary artery disease, diabetes mellitus type 2, and stroke place a heavy burden on public health care systems in developed countries. Despite the massive efforts of the scientific community to elucidate the genetic and molecular basis of common complex diseases, the fundamental basis remains elusive. This knowledge is needed for the development of novel improved preventative measures as well as diagnostic and therapeutic approaches. Complex and largely uncharted combinations of genomic and epigenomic, environmental, and lifestyle factors likely cause complex diseases (Lupski et al. 2011; Marian 2012). The completion of the human genome project in the beginning of the 21st century revealed the 3.2 billion long nucleotide sequence of the 23 human chromosomes, which contain most of the hereditary material in human cells. The human genome and epigenome are the blueprints that amazingly guide the development of a fertilized egg to an adult human being, consisting of a multitude of different cell and tissue types and organs with specific functions. These functions must be maintained throughout the individual's life and protected against internal and external insults, e.g., related to unhealthy lifestyle choices or viral infections. Failures to maintain these functions in the extremely complex interplay of various tissues and organs may lead to the development of various complex diseases.

This thesis focuses on the molecular pathomechanisms of saccular intracranial aneurysm (sIA) disease, a complex trait. The wall of the intracranial artery is a thin but highly complex structure that must withstand various stressors such as blood pressure and shear stress or atherosclerosis throughout the individual's life. For largely unknown reasons, in 2-3% of the population, protective mechanisms fail, leading to the formation of sIAs in the branching sites of intracranial extracerebral arteries. Most sIAs are too small to cause neurological symptoms and go unnoticed during life. In some sIA walls the mechanisms of maintenance and protection fail and the aneurysm ruptures, causing a devastating subarachnoid hemorrhage (sIA-SAH), flow of high-pressure arterial blood to the subarachnoid space on the surface of the brain, and possibly also to the brain tissue and brain ventricles. One third of sIA-SAH patients admitted to neurosurgical care die of complications within one year. Although sIA-SAH accounts for only about 5% of all forms of stroke, it mainly affects the working age population, causing a higher loss of productive years than the two more common forms of stroke, brain infarction and intracerebral hemorrhage. For unknown reasons, sIA-SAH is more common in Finland (about 1000 cases annually) and Japan than elsewhere in the world.

Modern high-throughput genomic methods are an attractive way to unravel the molecular pathomechanisms behind complex diseases, allowing, e.g., studies of whole genomes of patients or interrogation of transcriptomes of diseased tissues. These methods are increasingly applied to understand complex diseases. However, huge amounts of data are generated, and the greatest challenge is meaningful bioinformatic analysis and interpretation of the data in terms of molecular mechanisms, optimal therapies, and likely outcome of a given disease in a given individual.

In this thesis, the aim was to elucidate the signalling pathways of sIA rupture and the genetic background of sIA formation. This knowledge is needed for further understanding of the molecular mechanisms of sIA formation and rupture which could pave the way for the design of novel methods for non-invasive diagnosis, prevention, and occlusive therapy of sIAs.

Three complementary approaches were applied. First, genome-wide differences in gene expression between ruptured and unruptured sIA walls were studied. Second, a novel bioinformatics method and software was developed to aid interpretation and generation of hypotheses from differentially expressed gene sets. Third, a genome-wide association analysis was conducted to identify novel genes contributing to sIA disease susceptibility in Finnish and Dutch samples.

In the review part of this thesis, the landscape of human genetics as well as gene expression and regulation are reviewed. Next, complex diseases are defined, and a literature review of the complex sIA disease is presented. Finally, current computational pathway analysis methods aimed at the interpretation of differential gene expression data sets are reviewed.

After presenting the three studies that form the basis of this thesis, the common themes and differences identified in the complementary approaches are discussed. Finally, some avenues of reseach are envisioned for 1) advancing knowledge of the molecular basis of the sIA disease, and 2) developing bioinformatics methods applicable for the analysis of data generated by high-throughput methods about complex diseases.

2 Review of the literature

2.1 THE HUMAN GENOME

The first draft in 2001 (Lander et al. 2001; Venter et al. 2001) and the final completion of the Human Genome Project (HGP) in 2004 (International Human Genome Sequencing Consortium 2004) laid the foundation for modern high throughput genetic, genomic, and functional genomic studies of human diseases. The preceeding historical milestones were Darwin's theory of evolution and inheritance in 1859 (Darwin 1859), Mendel's observation of regular inheritance patterns in peas in 1866 (Mendel 1866), the identification of the first mendelian disease alkaptonuria in 1902 (Garrod 1902), and the identification of the structure of DNA in 1953 (Watson and Crick, 1953).

The HGP compiled a near complete single consensus basepair sequence of the human genome by analyzing the DNA sequence of four invididuals: two males and two females. The project, for the first time, provided estimates about the composition of the human genome. In 2004, it was estimated that the genome contains 20,000-25,000 protein coding genes, which are contained in less than 2% of the total genome sequence of some three billion basepairs (International Human Genome Sequencing Consortium 2004). The human reference sequence is constantly being improved upon in accuracy and annotation. Selected statistics of the genome build at the time of writing this thesis are shown in Table 1.

All work presented in this thesis is critically dependent on the HGP and subsequent genome annotations. The knowledge of human genome consensus sequence and variation enables the design of technology to study the transcriptome (Study I) and genome (Study III) in a genome-wide manner, and it is also critical for the computational methods predicting the regulation of gene sets obtained by whole-genome transcriptomic studies (Studies I and II).

2.1.1 Variation in human genome

The human reference sequence is based only on a few individuals. However, many different types of variations exist between individuals (Table 2). The sequence between two human individuals is estimated to differ by 0.1% due to single nucleotide polymorphism (SNP) variation (The International HapMap Consortium 2005), but other types of variation can account for a total of 0.5% of sequence difference (Levy et al. 2007).

A variant is a change in the nucleotide sequence of a genome. Variants in the human genome arise from uncorrected errors in DNA replication in dividing cells at an estimated rate of one in every billionth base (McCulloch & Kunkel 2008). In addition, Variants may occur due to environmental exposures such as radiation. If a variant occurs in germ-line cells and proves to be beneficial, leading to increased survival and reproduction, the variant will become more common with every generation in a population – the basis of evolution. In contrast, variants that are harmful in terms of survival and reproduction are under negative selection and will remain rare or vanish in subsequent generations. Variants that do not affect reproduction success, being functionally neutral or affecting probability of diseases at older age, can be carried and distributed randomly in the population.

Variants that occur in less than 1% of a population are called mutations, but if a variant reaches a frequency of 1% or more it is called a polymorphism. Variants with a frequency over 5% in a population are called **common variants** (The 1000 Genomes Project Consortium 2010). Variants occurring at a frequency between 0.5% and 5% are called **low frequency variants,** and those in under 0.5% of population are called **rare variants**. Extremely rare variants seen, e.g., in a single family are called **private variants**. The evolutionary pressure on variations are exemplified by the fact that the frequency of common variants are reduced near genes and regulatory elements (Cai et al. 2009; The 1000 Genomes Project Consortium 2012).

2.1.2 Linkage disequilibrium

Variants that are close to each other and/or not separated by recombination hotspots are likely to be inherited together. This phenomenon is called linkage. Linkage disequilibrium (LD) is the occurrence of a pair of variant alleles in a population more or less often than expected by chance. The term LD can be misleading since variations can be in association with each other without being linked. A simple example is when two populations with different frequencies of variations in two loci are mixed together; the combined population can display disequilibrium between variations without those variations being linked. Gametic phase disequilibrium has been used as an alternative to LD (Neale et al. 2007). In this thesis, like in most scientific articles on genetic association studies, the term LD is used.

When a new mutation occurs in a germline cell chromosome (e.g., *de novo* mutations at a rate of approximately 1.1 X 10^{-8} / base / generation (Conrad et al. 2011)), it is accompanied by all other genetic variants on that chromosome. Recombinations during meiosis slowly erode this LD at a rate of approximately 1 crossover per 100 megabases per generation (Altshuler et al. 2008). The significance of LD to genetic studies of human diseases is that instead of an expensive determination of the full sequence of individuals, only carefully selected variants tagging the regions of interest need to be determined.

Let us consider two biallelic loci (A and B) and their allele frequencies pA1, pA2, pB1 and pB2. If the two loci are independent of each other (i.e., in equilibrium) then the probability of observing any pair of alleles together in the same chromosome is the product of the individual allele frequencies (Equation 1).

$$p_{A1B1} = p_{A1}p_{B1} \qquad (1)$$

Common metrics of LD are D, D′ and r2. The simplest of the measures is covariance D (Equation 2). (Neale et al. 2007).

$$D_{AB} = p_{A1}p_{B1} - p_{A1}p_{A1} \tag{2}$$

A problem with this measure is that it is dependent on allele frequencies and thus it is not useful in comparing different pairs of loci with deviating allele frequencies. The absolute maximum value of D is the smaller of $p_{A1}p_{B1}$ and $p_{A1}p_{B2}$ when D is positive and the smaller of $p_{A1}p_{B1}$ and $p_{A1}p_{B2}$ when D is negative (Neale et al. 2007). A one proposed approach to make D more independent of allele frequencies is to scale the value of D by the maximum value (Equation 3).

$$D' = \frac{D_{AB}}{MAX(D_{AB})} \tag{3}$$

D′ ranges from -1 to 1, or alternatively absolute value, ǀD′ǀ, ranging from 0 to 1 can be used. D′ is more suitable for comparing pairs of loci with different allele frequencies although D′ also is somewhat dependent on the allele frequencies (D′ tends to be inflated when one of the alleles is rare).

Perhaps the most useful measure of LD in association studies is the squared correlation between presence and absence of alleles in a pair of biallelic loci (Equation 4) (Neale et al. 2007).

$$r^2 = \frac{D^2}{P_{A1}P_{A2}P_{B1}P_{B2}} \tag{4}$$

The measure r^2 varies between zero and one, one meaning the variants convey the same information and zero meaning that the variants are independent of each other. In association studies, there is a direct relationship between r^2 of the genotyped marker and causative variant and statistical power to detect association. Sample size would have to be increased by a factor of ~ $1/r^2$ to achieve the same statistical power as if the causative variant had been genotyped directly (Pritchard & Przeworski 2001).

Some distinctive features between D′ and r^2 can be identified. The ǀD′ǀ equals 1 if the frequency of any of the four pairs of genotypes is 0. If D′ = 1, there is not a 100% correlation between the allele at one locus and the allele at the second locus. If r^2=1 then two haplotypes are not observed and there is a 100% correlation between the alleles at two locus. (Neale et al. 2007). As a simplified rule of thumb: D′ might be favored when information on historic recombination events is of interest and r2 provides information on the correlation of the two loci. The measurement of LD in different contexts is a large topic and will not be further discussed here. The interested reader may want to read the publication (Pritchard & Przeworski 2001) and further references therein for an expansion of this topic.

2.1.3 Genome variation catalogues

In the wake of the first draft of the human genome, the International HapMap Project was launched with a mission to provide a catalogue of common human SNP variation and a haplotype map showing how those variants inherit together in different populations. Haplotype is defined as a particular combination of variant alleles along the same chromosome.

The first phase of the HapMap project constructed haplotype maps by analyzing approximately one common SNP in every 5kb across the genome (1,007,329 SNPS) of 269 individuals from four different populations. The samples were 90 individuals (30 parents-offspring trios) from Yoruba, Nigeria, 90 individuals with European ancestry (30 parents-offspring trios) from Utah, USA, 45 Han Chinese from Beijing China and 44 individuals from Tokyo, Japan (The International HapMap Consortium 2005).

In the second phase, the same individuals were used, but the SNP count was increased to over 3.1 million, which was estimated to capture untyped common variation with the maximum r2 ranging from 0.9 to 0.6 (Frazer et al. 2007). In the third phase, the sample size was increased to 1,184 from 11 populations. Rare and low frequency SNPs were also genotyped. They were hypothesized to explain a more substantial fraction of heritability of disease risks than common variants (Altshuler et al. 2010), which have been the focus of most genome-wide association studies to date (See chapter 2.1.4). The HapMap project estimated recombination frequency in different parts of the genome and a block like structure in the LD patterns was observed, where the block boundaries are often separated by recombination hotspots (Figure 1).

The HapMap projects created massive amounts of data on human SNP variation and how they inherit together in different populations. This knowledge was crucial for the design of genome-wide SNP arrays, the tools that started the genome-wide association studies of complex diseases. Study III in this thesis was possible only with the advancements in knowledge brought on by the HapMap projects.

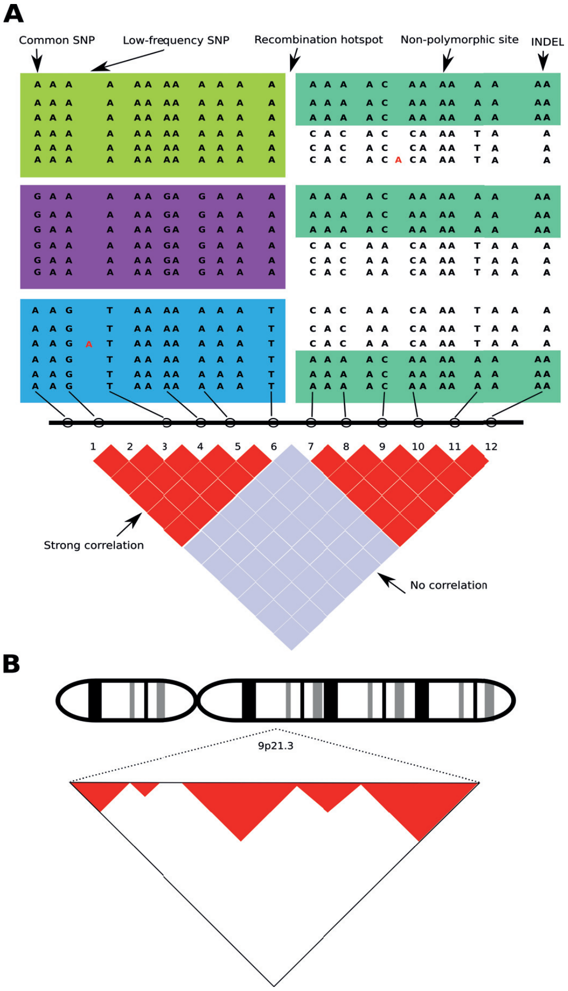


Figure 1. Illustration of the block-like structure of the human genome caused by linkage disequilibrium. A) Typical variation in 18 individuals. Six correlating variants on the left of the recombination hotspot form three different haplotypes (out of 64 possible haplotypes) as indicated by different colors. Recombination breaks the correlation, and the next six correlating variants form two different haplotypes. B) Regions depicted in A are scattered throughout the chromosomes forming LD blocks of different sizes. The correlations shown in the chromosome are fictitious (adapted from (Altshuler et al. 2008).

1000 Genomes Project

The latest major effort to characterize human genetic variation has been the 1000 Genomes project of more than 75 universities and companies worldwide (The 1000 Genomes Project Consortium 2010). The aim is to identify and create accurate haplotype maps of all sorts of human DNA polymorphisms of frequency > 1% and down to a frequency of 0.1% in the coding regions. The first phase of the project performed a low coverage whole-genome sequencing of 179 individuals from 4 populations, high-coverage sequencing of two parents-child trios, and exon sequencing of 697 individuals from 7 populations. The first phase reported 15 million SNPs, 1 million short insertions or deletions, and 20,000 structural variants. One striking finding was that each person's genome contains putative loss-of-function variants in as many as 250-300 genes worldwide (The 1000 Genomes Project Consortium 2010).

The second phase, aimed to comprehensively characterize low frequency (≤5%) variants, sequenced 1092 individuals from 12 populations, including 93 Finns (The 1000 Genomes Project Consortium 2012). Overall, 38 million SNPs, 1.4 million short insertions and deletions, and over 14,000 larger deletions were identified. The project showed that while common variants are typically found across populations and continents, low frequency and rare variants are more restricted to major ancestral groups (e.g., Europeans) or single populations. Interestingly, low-frequency variants (0.5-5%) are enriched among Finns as compared to other European populations.

In study III of this thesis, the interim 1000 Genomes haplotype panel (The 1000 Genomes Project Consortium 2012) was used to impute also low frequency variants, insertions or deletions and structural variants in a Finnish sample of sIA cases and controls. Our hypothesis was that some low-frequency variants, increased in frequency in Finland, would contribute to increased incidence of sIA-SAH in Finland (see Chapter 2.4).

The 1000 Genomes Project is an on-going effort, currently aiming to sequence 2,500 individuals from 27 populations. The data generated will be crucial for the design of new genotyping tools and imputation of a wider range of variants since the focus of whole genome association studies will likely shift towards low frequency variants.

Mankind's colossal endeavour of determining, let alone understanding, the function and consequences of human genome variation is an ongoing effort with an estimated one million individuals having their full genome sequenced by 2015 (Aarno Palotie personal communication; Andreas Sundquist (Harris 2012)) .

Table 1. Current Human genome statistics according to Ensembl database version 72.37, April 2013.

Category	***Count***
Length (basepairs)	*3,323,950,079*
Coding genes	*20,774*
Non coding genes	*22,493*
Pseudogenes	*14,145*
Gene transcripts	*194,846*
SNPs, indels, somatic mutations	*54,965,377*
Structural variants	*10,266,123*

Table 2. Types of variations in the human genome

Type	***Description***
Single nucleotide polymorphism (SNP)	*A change of one nucleotide in DNA.*
Chromosomal abnormality	*Insertion, deletion, translocation or reversal of a large (> 3Mb) parts of a chromosome.*
Copy number variation (CNV)	*A repetition of large (1kb - 3Mb) part of genome zero (deletion) or more times.*
Interspersed element	*A repeating sequence of DNA, a remnant from viral genome. Some repeats contain sequence encoding viral machinery proteins required to copy and move the sequence.*
Simple repeat	*A variable number repetition of a few nucleotide sequence.*
Insertion or deletion (INDEL)	*A change of addition or removal of one or more nucleotides.*
Structural variant (SV)	*1 kb and larger in size and can include inversions and balanced translocations or large insertions or deletions*

2.1.4 Studying human genome variation

Heritability is a measure for percent of variation in a phenotype due to genetic factors (Wray 2008). The classical way of assessing the amount of genetic contribution to human diseases or phenotypes is twin studies. As monozygotic and dizygotic twins share 100% and 50% of the genomes, respectively, and if both share similar environments, then the genetic load on the phenotype can be estimated by studying the phenotype difference between the twins. The concordance of complex diseases in monozygotic twins is often far from 100%, suggesting that, in addition to environmental effects, also epigenomic differences may contribute to phenotype development or complex disease susceptibility (Bell & Spector 2011). The twin studies offer only knowledge about the heritability of the phenotype and methods of linkage and association studies can be used to map the genomic loci harbouring variants affecting the phenotypes.

2.1.4.1 Linkage studies

In linkage studies, segregation of genetic markers scattered throughout the genomes is correlated to the segregation of disease in pedigrees. One shortcoming is that linkage analysis in complex traits can only map genetic loci broadly and the linkage intervals can contain hundreds of genes, thus the underlying mutations are not necessarily easily identifiable (Teare & Barrett 2005).

2.1.4.2 Candidate gene association studies

In association analyses, selected genetic markers are correlated to case-control status in a sample of cases and controls or to quantitative phenotype in a population cohort. The controls should be as similar to the cases as possible, differing optimally only by the phenotype or disease status. The cohorts should usually be unrelated individuals, but pedigree based analysis is also possible (Neale et al. 2007).

Traditionally, the candidate gene approach was used where one or a few genes were selected, based on positional evidence from linkage studies or functional hypotheses about putative susceptibility genes. Genetic markers were then selected among previously known variable genetic markers (Hirschhorn et al. 2002). The obvious shortcoming is the need of insight or prior knowledge about the pathways and genes likely associated to the phenotype studied, and, consequently, many novel associations are likely to be missed.

2.1.4.3 Genome-wide association studies

In recent years, the genome-wide association study (GWAS) approach has been the method of choice in human genetic studies. In GWASs, as in the candidate gene approach, genetic markers are correlated to phenotype in cases (affected) and controls (unaffected) or in population cohorts with phenotypes of interest. Instead of having educated guesses about candidate genes and pathways, genetic markers all around the genome are used. Typical GWAS arrays contain from a few hundred thousand to a few million SNPs and structural variants, and each of these is correlated to the phenotype of interest. The GWAS genotyping chips are commercially available from, e.g., Illumina and Affymetrix whose chips differ in how the SNPs are selected. Affymetrix selects SNPs scattered evenly around the genome whereas Illumina uses LD information of the HapMap CEU population to achieve satisfactory tagging coverage of common variations. One major limitation of the majority of GWASs has been that most commercial genotyping platforms can reliably detect only common SNPs at > 5% frequency. Also, the selection of SNPs may not tag well even the common variation in a given population. More recently, the frequency spectrum has been expanded by genotyping arrays focused on the exomes (Huyghe et al. 2013) or on the loci selected for metabolic (Metabochip) (Voight et al. 2012) or immune related (Immunochip) (Eyre et al. 2012) traits.

The number of published GWASs has steadily increased for the last 5 years, totalling almost 1500 studies of over 200 different phenotypes (Hindorff et al.) (*Figure 2*). The general outcome of the studies is that the associated SNPs only moderately increase the risk of the disease studied, and the identified SNPs explain only a proportion of the heritability estimated from twin studies.

For example, in human height, with an estimated heritability of 80%-90%, the 180 associated GWAS loci explain only about 10% of the heritability (Allen et al. 2011). However, it has been estimated that as much as 45% percent of the phenotypic variance of height can be captured by using all of the variants in the genotyping array and not just the variants reaching stringent genome-wide significance (Yang et al. 2011). These types of multimarker predictive panels need to be evaluated carefully as they typically are not able to explain as much of the phenotypic variances in independent replication studies (Makowsky et al. 2011). It nevertheless seems that many true risk variants (with small risk effect) exists that have not reached the stringent statistical thresholds in GWASs.

Similar GWAS findings apply to less heritable complex human diseases such as ischemic stroke (Bevan et al. 2012), obesity, and mental disorders (Schizophrenia, bipolar disorder) (Visscher et al. 2012). Age-related macular degeneration is a striking exception in which about 50% of the heritability is explained by only five common SNPs (Maller et al. 2006). Another success story is the identification of 163 loci associated to the inflammatory bowel disease, more than to any other complex disease to date, accounting for 13.6% of disease variance in Crohn's disease (one of the two major subtypes of inflammatory bowel disease) (Jostins et al. 2012). However, even in the case of reliably replicated associations with high odds ratios, the predictive ability of the identified variants in common complex diseases is low (Jakobsdottir et al. 2009).

Although the GWAS approach has been successful in identifying numerous consistently replicable SNPs associated to complex human diseases or traits, converting these findings to a molecular understanding of various disease mechanisms has been challenging. In many GWASs the nearest gene to the variants are nominated as susceptibility genes. Only about 5% of the trait associated variants are non-synonymous coding variants, and most of them are intergenic (43%) or intronic (45%) (Hindorff et al. 2009). The problem in nominating the physically closest gene as the susceptibility gene is exemplified by the well known SNP causing adult-type hypolactasia (Enattah et al. 2002). The nearest associated SNP resides in an intron of a gene unrelated to the phenotype, almost 14kb away from the lactase-phlorizin hydrolaze (*LPH*) gene. The enzyme product of *LPH* breaks down lactose, and silencing of LPH causes hypolactasia. Despite being distant to *LPH*, the associated SNP above is strongly associated to the transcriptional regulation of the *LPH* expression (Enattah et al. 2002).

In summary, even though GWASs in many complex human diseases have produced interesting results, suggesting biological pathways involved in the pathophysiology of various complex diseases, follow-up studies are needed to translate these initial findings into mechanistic biological knowledge that paves the way for novel tools for the prediction, diagnosis, and therapy directed at the disease.

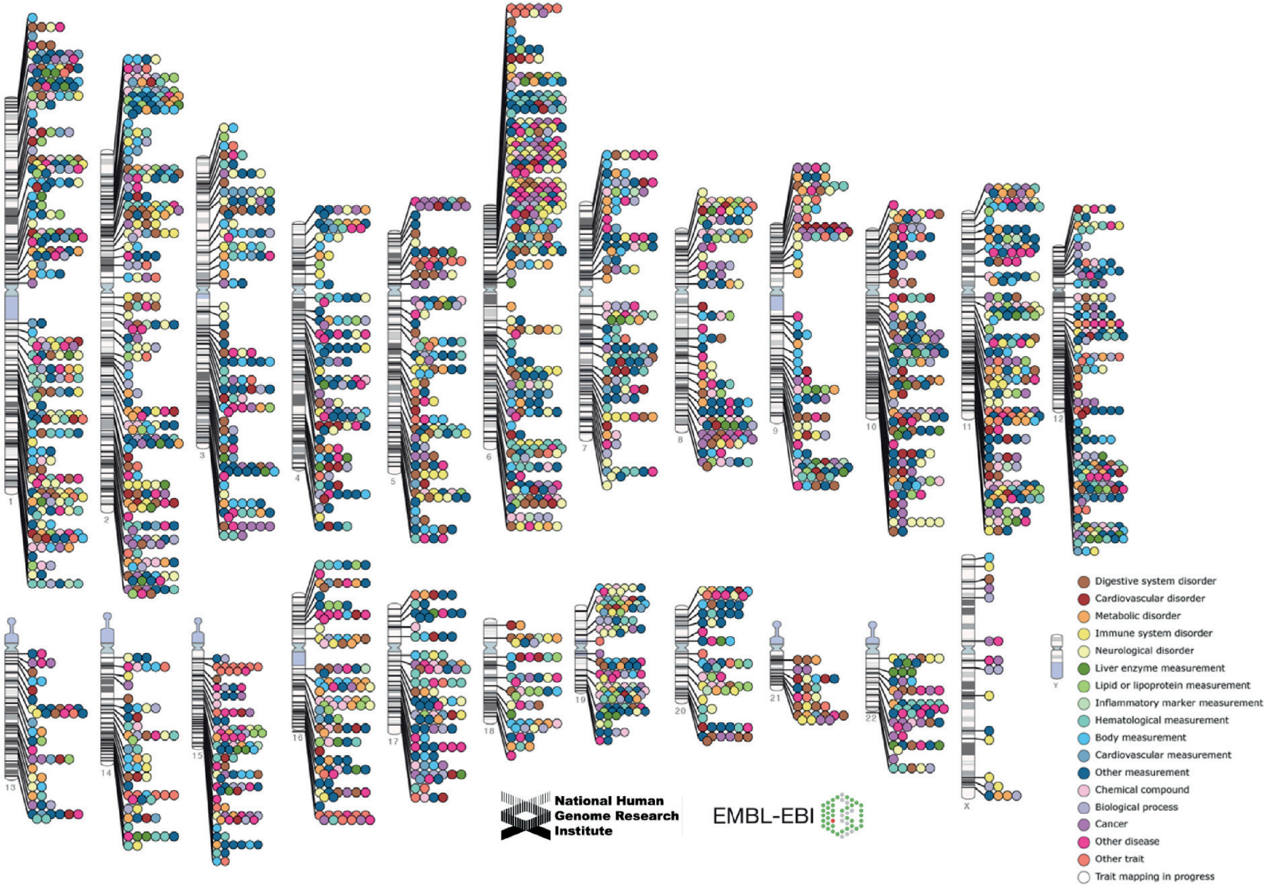


Figure 2. Significant associations ($p<5 \times 10^{-8}$) identified in genome-wide association studies in 18 categories of human traits and diseases (Modified from www.genome.gov/GWAStudies).

2.1.4.4 Missing heritability

The fact that GWAS identified variants explain only a fraction of the heritability has been termed "missing heritability", a topic of much debate in the scientific community (Maher 2008; Manolio et al. 2010). The proposed causes include the possibility that untyped copy number variants could be the variants carrying missing heritability (Maher 2008; Manolio et al. 2010). Future exome or full-genome sequencing projects are likely to shed light on this. Another explanation proposed is that rare variants, inefficiently tagged by SNPs included in the current GWAS genotyping arrays, would explain the missing heritability (Maher 2008; Manolio et al. 2010). Sequencing studies and projects such as the 1000 Genomes Project will elucidate this possibility. The lower frequency variants detected in the 1000 Genomes Project are one of the cornerstones of Study III in this thesis. Further explanations suggest that there are many undiscovered variants with smaller effect sizes than can be detected by current GWAS studies or the effects are modified by other loci and/or environment, reducing the power to detect associations (Maher 2008; Manolio et al. 2010). The fact that the loci surpassing the stringent genome-wide significance explain much less of the variability than considering all the variants together in the array suggests that a great number of false negatives in GWAS studies explain part of the missing heritability (Yang et al. 2011). The identification of these loci would require much larger sample sizes. It has also been suggested that population isolates, potentially enriched in unique variants, are of value in uncovering some of the hidden heritability. This approach has already been succesful in uncovering loci that associate to metabolic traits in the Finnish population isolate (Sabatti et al. 2009). This possibility is utilized in Study III of the present thesis.

2.1.4.5 Sequencing approaches in genomic studies

The latest technological advancement in human genomic studies is next-generation sequencing (NGS) (Metzker 2010). The DNA sequencing costs have plummeted from around $1000 per megabase in 2005 to just about ten cents per megabase, and these techniques have become widely available also outside of large genome centers (Wetterstrand 2013). Sequencing of the exome or the full genome allows interrogation of a complete set of variants (including rare coding variants with possibly large effects) and not just the set of variants selected for, e.g., a GWAS chip. Exome sequencing has accelerated the identification of mutations causing Mendelian diseases (Bamshad et al. 2011), and has been applied to the study of *de novo* mutations possibly causing neurological diseases such as autism (O'Roak et al. 2011). There is also interest to apply exome sequencing in the study of complex diseases (Kiezun et al. 2012). The association of rare variants identified by sequencing with complex diseases and traits necessitates large sample sizes of around 10,000 participants to achieve sufficient statistical power (Kiezun et al. 2012). Full genome sequencing has been mainly applied in somatic mutation detection in cancer studies

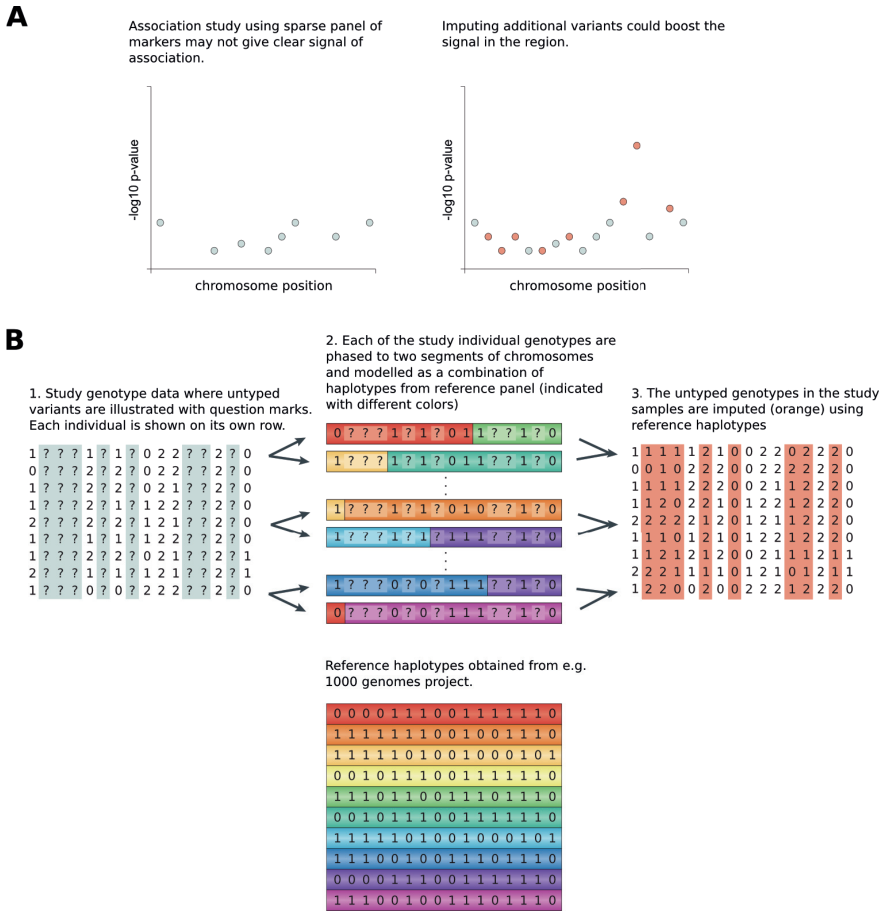
(Kilpivaara & Aaltonen 2013). One of the challenges of applying NGS to study human genetic diseases is to distinguish the disease causing variants from the potentially harmful 250-300 protein product changing variants found in every genome (The 1000 Genomes Project Consortium 2010), and more dauntingly, to identify and to interpret the roles of intergenic non-coding variants.

NGS methods can also be applied to genome-wide studies of epigenomic markers (DNA-methylation, histone modifications) (Ku et al. 2011). The epigenome-wide studies are scarce outside of cancer research and, consequently, the contribution of epigenomic factors to complex human diseases is only beginning to be unravelled (Rakyan et al. 2011).

2.1.4.6 Genotype imputation

Genotyping arrays utilized in GWASs interrogate only a small subset of currently known variants. Utilizing LD between variants, it is possible to impute (i.e. predict) untyped variants based on the smaller set of genotyped ones and larger panel of reference genotypes.

Genotype imputation has several benefits. It can increase the statistical power of the study, and provides a means to fine-map association regions to pinpoint causal variants (Figure 3 A). Imputation also allows prediction of other than SNP variations such as INDELs if suitable reference panels such as the 1000 Genomes Project are used. Finally, genotype imputation facilitates meta-analyses of different studies if they did not include the same set of variants (Marchini & Howie 2010).


A
Association study using sparse panel of markers may not give clear signal of association.
-log10 p-value
chromosome position
Imputing additional variants could boost the signal in the region.
-log10 p-value
chromosome position
B
1. Study genotype data where untyped variants are illustrated with question marks. Each individual is shown on its own row.
2. Each of the study individual genotypes are phased to two segments of chromosomes and modelled as a combination of haplotypes from reference panel (indicated with different colors)
3. The untyped genotypes in the study samples are imputed (orange) using reference haplotypes
Reference haplotypes obtained from e.g. 1000 genomes project.

```
1 ? ? ? 1 ? 1 ? 0 2 2 ? ? 2 ? 0
0 ? ? ? 2 ? 2 ? 0 2 2 ? ? 2 ? 0
1 ? ? ? 2 ? 2 ? 0 2 1 ? ? 2 ? 0
1 ? ? ? 2 ? 1 ? 1 2 2 ? ? 2 ? 0
2 ? ? ? 2 ? 2 ? 1 2 1 ? ? 2 ? 0
1 ? ? ? 1 ? 1 ? 1 2 2 ? ? 2 ? 0
1 ? ? ? 2 ? 2 ? 0 2 1 ? ? 2 ? 1
2 ? ? ? 1 ? 1 ? 1 2 1 ? ? 2 ? 1
1 ? ? ? 0 ? 0 ? 2 2 2 ? ? 2 ? 0
```

```
0 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
                .
                .
1 ? ? ? 1 ? 1 ? 0 1 0 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 1 1 1 ? ? 1 ? 0
                .
                .
1 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
0 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
```

```
1 1 1 1 1 2 1 0 0 2 2 0 2 2 2 0
0 0 1 0 2 2 2 0 0 2 2 2 2 2 2 0
1 1 1 1 2 2 2 0 0 2 1 1 2 2 2 0
1 1 2 0 2 2 1 0 1 2 2 1 2 2 2 0
2 2 2 2 2 1 2 0 1 2 1 1 2 2 2 0
1 1 1 0 1 2 1 0 1 2 2 1 2 2 2 0
1 1 2 1 2 1 2 0 0 2 1 1 1 2 1 1
2 2 2 1 1 1 1 0 1 2 1 0 1 2 1 1
1 2 2 0 0 2 0 0 2 2 2 1 2 2 2 0
```

```
0 0 0 0 1 1 1 0 0 1 1 1 1 1 1 0
1 1 1 1 1 1 1 0 0 1 0 0 1 1 1 0
1 1 1 1 1 0 1 0 0 1 0 0 0 1 0 1
0 0 1 0 1 1 1 0 0 1 1 1 1 1 1 0
1 1 1 0 1 1 0 0 1 1 1 0 1 1 1 0
0 0 1 0 1 1 1 0 0 1 1 1 1 1 1 0
1 1 1 1 1 0 1 0 0 1 0 0 0 1 0 1
1 1 1 0 0 1 0 0 1 1 1 0 1 1 1 0
0 0 0 0 1 1 1 0 0 1 1 1 1 1 1 0
1 1 1 0 0 1 0 0 1 1 1 0 1 1 1 0
```

Figure 3. Genotype imputation. A) Illustration of the gain in power and fine-mapping provided by imputation. B) Schematic illustration of the genotype imputation process modified from the following article (Marchini & Howie 2010). See text for explanation.

The basic process of genotype imputation is illustrated in Figure 3 B. A cohort of individuals has been genotyped with a given GWAS chip, leaving many known (and unknown) variants untyped (1). The genotyping technology is not able to determine which variants reside together on the same chromosome, and this information must be probabilistically inferred, producing two local stretches of variants likely to reside on the same chromosome for each individual (i.e. haplotypes) (2). These short haplotypes will be related to each other in a sample of unrelated individuals and individuals in a reference panel (3) by being identical by descent. Subsequences of the study haplotypes are matched to those of different reference panel individuals (3), resulting in the haplotypes of study individual's being modelled by a mosaic of reference haplotypes (indicated by different colors in Figure 3 B). The missing genotypes are then inferred from the matching haplotypes (4).

The imputed genotypes are probabilistic (contrary to the simple schematic in Figure 3), so imputation methods produce a probability distribution of all possible genotypes for each individual. This distribution can be modelled using Hidden Markov Models in which the transition state space and probabilities of variant-to-variant transitions are modelled from haplotypes of a reference panel as in Impute 2 (Howie et al. 2009).

Using the best guess genotype after imputation in subsequent association analyses can lead to false positives or reduced power and therefore should not be used, rather the uncertainty should be taken into account in the association analyses (Marchini & Howie 2010).

2.2 THE HUMAN TRANSCRIPTOME

The function of the human genome in the cells, tissues, organs and the whole body is ultimately channeled through the transcription of the protein coding parts of genes to mRNA molecules, which are in turn translated to proteins on ribosomes. Other types of functional RNA molecules are also transcribed from the genome such as ribosomal RNA (rRNA), transfer RNA (tRNA) and many other small RNA molecules with diverse functions as enzymes or with gene regulatory functions (Guil & Esteller 2012). The collection of all of the different types of transcribed RNA molecules is called the transcriptome.

A protein-coding gene consists of exons and introns. All introns and exons are first transcribed to mRNA molecules by RNA polymerase II enzyme. The intron parts of full-length mRNA molecules are almost always removed, and combinations of exons are assembled by the spliceosome complex in a process called alternative splicing. After splicing, the mRNA molecules are exported from the cell nucleus to the ribosomes for translation of the mRNA code to the specific sequence of amino acids, i.e., into a protein. Finally, the protein may undergo chemical modifications called post-translation modifications, and the protein folds into its functional conformation.

The approximately 21,000 protein coding genes are coded by only 3 percent of the human genome (Dunham et al. 2012). However, a huge variety of different proteins are produced from those genes via alternative splicing (on average ~4 per gene) (Dunham et al. 2012) and many different post-translational modifications (Choudhary & Mann 2010).

2.2.1 Regulation of gene expression

The activity of protein coding genes need to be precisely controlled spatially and temporally in order to produce proteins in correct time and amount as needed by a cell, tissue, organ or the whole body (Noonan & Mccallion 2010; Davidson & Erwin 2006). The regulation of gene expression can occur at many points, including transcription initiation, elongation (Guenther et al. 2007), or mRNA stability as well as translation and protein degradation (Vogel & Marcotte 2012; Wu & Brewer 2012). Control of transcription initiation is most relevant to the present thesis (Studies I and II), and it is discussed in detail, although a significant amount of control occurs in subsequent steps after mRNA transcription (Vogel & Marcotte 2012).

For a gene to become active, the basic transcriptional machinery needs to have access to the DNA sequence near the transcription start site (TSS), in an area called the promoter of the gene. Transcription factors (TF) are proteins that can bind to the DNA strand in the promoter regions of a gene and guide the transcriptional machinery to initiate transcription. There are two types of transcription factors binding to the promoter region: 1) TFs of the general transcriptional machinery bind to the core promoter in the immediate vicinity of TSS; and 2) more general TFs bind to the proximal promoter. TFs of general transcriptional machinery can usually only drive low levels of expression, whereas proximal TFs can have a greater impact as transcriptional activators or repressors. In addition to promoter sequence elements, more distant elements can also have effect on gene expression. Enhancers and silencers are positive or negative regulatory elements that can reside even several hundred kilobases upstream of a given TSS, downstream of a TSS in an intron, or even beyond the end of the gene (Figure 4) (Maston et al. 2006).

TFs bind to their characteristic short (6-21 bp) DNA sequences called transcription factor binding sites (TFBS). These binding sites are not exact, but are characterized by a consensus sequence where the probability of each nucleotide in each position varies and some positions are more fixed while other positions are more tolerant to different nucleotides (Spitz & Furlong 2012; Bryne et al. 2008; Matys 2003).

Mutations in TFBS can alter the binding affinity and subsequently the transcriptional activity of a gene, and can cause human diseases. A mutation in a TFBS 43 bp upstream from the low-density lipoprotein receptor affects it's expression and cause familial hypercholesterolemia (Koivisto et al. 1994). Favorable lactase persistence trait, the ability to break down lactose in adulthood, is caused by a mutation in a probable distant silencer (Enattah et al. 2002). The lactase gene is typically silenced in mammals after weaning, but an SNP 14kb upstream disrupts the silencer activity, allowing continued expression of the lactase gene and lactase enzyme activity (Rasinperä et al. 2005).

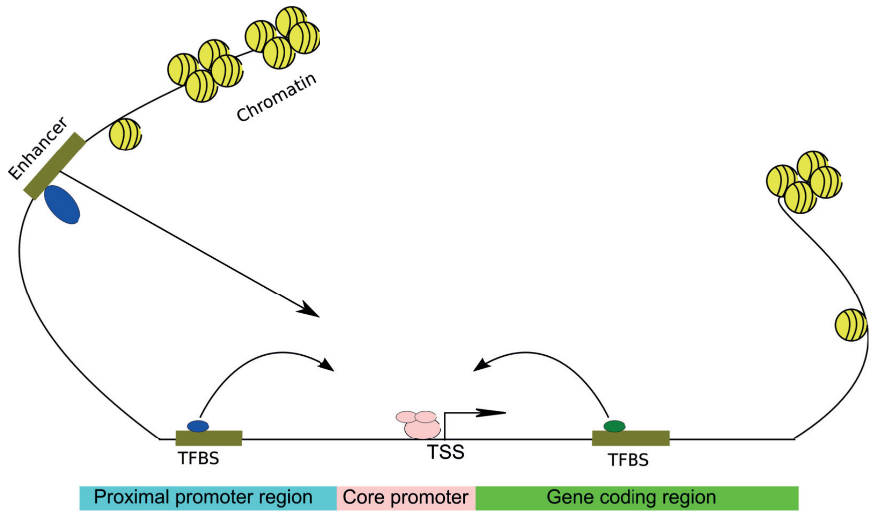


Figure 4. Illustration of a gene promoter and regulation. Chromatin consists of DNA packaged into histones. This packaging must be unwound for the transcriptional machinery to have access to the transcription start site (TSS) in the core promoter. Transcription factors bind to their specific transcription factor binding sites (TFBS) near the TSS (proximal region) or further away (enhancer) and can influence the transcriptional efficiency. Adapted from (Maston et al. 2006; Lenhard et al. 2012).

Another regulatory mechanism controlling transcription is epigenomics. Epigenomic modifications alter the structure and packaging of DNA while the DNA sequence itself remains the same. These epigenetic modifications can be inherited or acquired. Here we adopt the definition of epigenetics by Adrian Bird: ”the structural adaptation of chromosomal regions so as to register, signal, or perpetuate altered activity states” (Bird 2007).

The cytosine residues in DNA strand can be methylated to 5-methylcytosine at CG dinucleotides. CG dinucleotides can be repeated as a so called CpG islands, which often occur in promoter regions and their hypermethylation is associated with silenced gene expression (Hsieh 1994). DNA methylation is also involved in X chromosome inactivation (Panning & Jaenisch 1996), repression of retroviral elements (Walsh et al. 1998) and genomic imprinting (Li et al. 1993), the differential modification distinguishing maternally or paternally inherited chromosomes in the offspring. The amino acids at the tails of the histone proteins are another target for epigenetic modifications. These include methylation, ubiquitylation, phosphorylation, and acetylation. These modifications are considered to modify gene expression by affecting chromating packaging or by serving as signalling factors to other proteins. The chromatin code is very complex, of which we have only a very limited understanding (Berger 2007).

Just recently, the ENCODE project reported a wide catalog of regulatory elements in human cells, including accessible chromatin (DNAse hypersensitivity) in 25 cell types, DNA methylation patterns in 82 cell lines, 12 histone modifications in 46 cell types, and

DNA binding of 119 different transcription factors in 72 different cell types (Dunham et al. 2012). Some key findings were: approximately 80% of the human genome is involved in at least one RNA expression or chromatin associated event in at least one cell type studied; Most TFs show enriched binding signals in a narrow DNA region near TSS; TF binding and chromatin modification in a narrow promoter region around a transcription start site can explain a large fraction of variation in transcription from a specific TSS but not the variation of mRNA transripts; SNPs associated with different diseases in genome-wide association analyses are enriched within identified non-coding functional elements. In study III the functional elements identified in ENCODE project are used to predict putative functional consequences of identified variants.

2.2.2 Studying the transcriptome

Studying the transcriptome has been widely used to gain insight into molecular mechanisms underlying human diseases. The key assumption is that a change in transcriptional activity of genes is an indicator of change in the function of diseased tissues, presumably contributing to disease mechanisms. Identifying signaling pathways and genes involved could lead to a better understanding of the disease process, and lead to development of novel diagnostic, prognostic, and therapeutic tools. (Margulies et al. 2009; Kwong et al. 2012)

Applying transcriptome profiling in human diseases typically involve comparison of two or more types of tissues such as the comparison of the diseased tissue to the corresponding healthy tissue, and identification of transcripts that are differentially expressed between the tissues. The typical result of transcriptomic profiling is then a set of differentially expressed genes. This set of genes might be extremely useful in identifying genes that appear to play a role in the etiology, development, and progression of the studied disease. Often, however, the list contains hundreds of genes and may fail, at least in the light of present knowledge, to provide insight into the molecular mechanism of the disease process. In chapter 2.5, some bioinformatic methods for the functional analysis of differentially expressed gene sets are reviewed. In Study II of the present thesis, a novel computational method and software tool was developed to assist researchers in gaining additional insight and in generating data driven hypotheses from such gene sets.

Another use for transriptome profiling is to study how variation in the genome may lead to differential gene expression, providing mechanistic explanations between, e.g., GWAS findings and phenotypes. These variants, called expression quantitative trait loci (eQTL), are searched for in Study III. In fact, SNPs associated to human traits in GWAS studies are more likely to be eQTLs (Nicolae et al. 2010).

Traditionally, the transcriptional activity of genes has been measured by quantitative PCR, one gene at a time. This approach is limited similarly as the candidate gene association analyses, requiring educated guesses for the selection of candidate genes. This approach may be sensible, however, in functional studies of the loci identified by GWAS. In complex diseases, with limited knowledge of molecular mechanisms, more holistic approaches are required. The advent of microarray technology in the mid 1990's (Schena et al. 1995) revolutionized expression studies by allowing simultaneous measurement of mRNA levels of most human genes.

Microarray technology is dependent on the knowledge of the studied organism's genome sequence and annotations of transcribed regions. Microarrays contain sets of oligonucleotide probe sequences, and each sequence is complementary to a subsequence of each target mRNA. Total mRNA is isolated from tissues or cell cultures, and it is reverse transcribed to complementary DNA (cDNA) and finally labeled complementary RNA (cRNA) is transcribed from cDNA. The total labeled cRNA is then let to hybridize to the microarray chip. Dedicated scanners are used to detect the intensities of signals from labeled cRNA hybridized to specific locations on the microarray chip. The microrrays are commercially available from several vendors, each with their own array designs. The most popular arrays are produced by Affymetrix (www.affymetrix.com) and Illumina (www.illumina.com), which are also the main manufacturers of genotyping arrays. After computational steps of pre-processing and normalization these intensity signals can be used to semi-quantitatively assess the expression level of each transcript. Semi-quantitative refers to the fact that microarray intensity levels cannot be interpreted as specific amounts of mRNA. Instead, the intensities can be used to assess the relative expression levels between experimental groups.

Lately, next-generation sequencing for transcriptome profiling (RNA-Seq) has become available (Wang et al. 2009). In RNA-seq, the total RNA is extracted and reverse transcribed to cDNA, which is then processed in to a library of short sequences, and massively parallel sequencing is finally used to read the sequences. The short DNA sequences can then be either mapped to reference genome or to reference transcripts or alternatively the overlap in short sequences can guide *de novo* assembly of transcripts. The number of reads of each trancript is used as a measure of abundance of the transcript. RNA-seq methods have some advantages over the microarray approach. First, RNA-seq does not necessarily require any *a priori* knowledge about the transcripts of the studied tissue. RNA-seq is suitable for the detection of novel transcripts or alternative mRNA splicing. Another advantage is the possiblity to detect differences in RNA sequences allowing the identification of mutations or mRNA editing events. The data from RNA-seq are expected to be less noisy, not affected by cross-hybridization of similar sequences. Finally, RNA-seq has a broader dynamic range than microarrays (Wang et al. 2009).

The first enthusiastic reports of RNA-seq's technical reproducibility were based on a small sample sizes, and recently pitfalls in RNA-seq approach have been reported. RNA-seq protocols include steps that can induce biases in the data, and the sequence content of the mRNA molecules studied can affect the results (Roberts et al. 2011; Li et al. 2010). These biases require sophisticated normalization procedures to mitigate their impact, similarly as in microarray studies (Hansen et al. 2012). These normalization techniques are mature in microarray technology, but in RNA-seq the biasing problems are just being recognized and appropriate normalization methods are being developed. Despite these differences, the data from the both methods do correlate with each other, but the lower sensitivity of microarrays may reduce the statistical power to detect as many differentially expressed genes (Fu et al. 2009; Bradford et al. 2010; Malone & Oliver 2011).

One shortcoming of the transcriptome profiling in studying complex diseases, with either microarrays or RNA-seq, is that only about 40% of the variation in protein levels might be explained by the mRNA concentrations, and the rest of protein level variation could be explained by post-transcriptional mechanisms. The abundance of mRNA seems, however, to be an excellent indicator of the presence of a given protein (Vogel & Marcotte 2012).

In common complex diseases, transcriptome profiling has indicated novel candidate genes and pathways, increased the understanding of the molecular mechanisms of those diseases, and has led to novel diagnostics approved also for clinical use. For example, microarray profiling of liver, visceral fat, skeletal muscle, atherosclerotic vs. unaffected arterial wall isolated from coronary artery disease (CAD) patients, has identified a group of genes in atherosclerotic vascular and visceral fat tissues that associates to the extent of coronary stenosis (Hägg et al. 2009). The association of this gene group to the coronary stenosis was confirmed by expression profiling of an independent CAD cohort. The gene group was linked to the leukocyte transendothelial migration signalling pathway, and the *LBD2* transcription factor binding site was identified in the promoter regions of all these genes, suggesting *LBD2* as a key regulator in coronary stenosis. In another study, the transcriptomic profiling of breast cancer tissue by microarrays led to the identification of a panel of 70 genes, which proved be a better predictor of the 5-year probability of distant metastases than any other clinical predictor used (Veer et al. 2002). The FDA has approved this panel, called Mammaprint, after validation in further studies.

2.3 COMPLEX DISEASES

Human diseases can be roughly divided into two classes based on the impact of inheritance in their etiology. A disease is termed Mendelian if it segregates in families following the Mendelian laws of segregation. In Mendelian diseases, there is typically a single gene or a locus with a causal effect but additional genes may modulate the appearance of the disease (Chial 2008). Mutations that cause severe consequences remain rare in the population since they diminish the probability of having offspring.

The identification of a causative mutation can be a relatively straightforward when a sufficient number of members in families with the genetic condition are identified. There are currently over 3,200 Mendelian diseases tied to a specific gene but causative genes in as many rare diseases or phenotypes have not been identified (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore).

With no obvious pattern of inheritance, a disease is termed complex or multifactorial such as type 2 diabetes, hypertension, schizophrenia, or multiple sclerosis. The development of complex diseases is determined in various proportions by multiple genomic, epigenomic and environmental risk factors and their interactions. Unraveling the mechanisms and impact of these factors in complex disease development is much more difficult than in Mendelian diseases because the development these diseases are controlled by multiple genetic, epigenetic and environmental factors and their complex interactions (Craig 2008).

2.4 SACCULAR INTRACRANIAL ANEURYSM (SIA) DISEASE

Intracranial extracerebral aneurysms are classified into: 1) saccular intracranial aneurysms (sIAs), pouches that develop during life in the forks of extracerebral arteries (Figure 5); and to 2) fusiform intracranial aneurysms, spindle-like dilatations of extracerebral artery trunks. Fusiformic aneurysms are the dominant type in the aorta, but for unknown reasons, saccular aneurysms account for about 96% of the intracranial extracerebral artery aneurysms. The current thesis, similar to most previous research, focuses on sIA disease.

Unruptured sIAs are almost always too small to cause neurological symptoms, and most remain undetected during life (Brisman et al. 2006). Large unruptured sIAs can cause neural symptoms by compressing, e.g., the optic nerve (visual deficits) or the oculomotor nerve (diplopia and proptosis), or cerebral ischemia due to emboli from inside the aneurysm (Friedman et al. 2001). The rupture of the sIA pouch, however, causes subarachnoid hemorrhage (sIA-SAH), pouring of high-pressure arterial blood into the subarachnoid space on the surface of the brain and possibly into the brain tissue or the brain ventricles. Aneurysmal SAH is a devastating form of stroke that affects the working age population (van Gijn et al. 2007).

The case fatality of SAH is variable but has been estimated to have been almost around 50%, however recently improving to under 40% (Stegmayr et al. 2004; Nieuwkamp et al. 2009) and about a third of the survivors are moderately or severely disabled (Hop et al. 1997). Neurosurgery at Kuopio University Hospital (KUH) solely serves a defined Eastern Finnish catchment population (Huttunen et al. 2010), and 26% of the sIA-SAH patients admitted alive to KUH died at 12 months due to complications of the bleeding (Karamanakos et al. 2012).

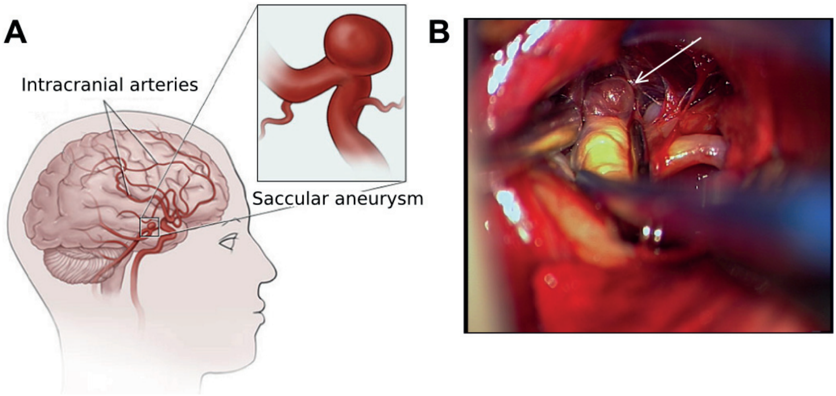


Figure 5. A) Drawing of a saccular intracranial aneurysm at a typical location in the Circle of Willis at the base of the brain (modified from www.nebraskamed.org). B) Photograph of a human saccular intracranial aneurysm taken during surgery.

2.4.1 Epidemiology

The prevalence of sIA has been estimated to be 3% in a population with a median age of 50 years and consisting of 50% males (Vlak et al. 2011). Most sIAs do not rupture during life as the incidence of sIA-SAH is only 9.1 (95% CI 8.8-9.5) per 100,000 life-years in Europe and North America, about half of that observed in South and Central America 4.2 (3.1 to 5.7) and strikingly about twice in Japan 22.7 (21.9 to 23.5) and Finland 19.7 (18.1 to 21.3) (de Rooij et al. 2007). The reasons for the higher incidence in Finland and Japan are not known.

The overall annual rupture rate of unruptured sIAs is estimated to be around 1% (Wermer et al. 2007; Morita et al. 2012; Müller et al. 2013). The risk of rupture can vary by aneurysm size, location, shape, patient gender and ethnicity (Wermer et al. 2007; Morita et al. 2012). Posterior aneurysms are more likely to rupture than MCA aneurysms and also larger aneurysms and aneurysms with secondary pouches are more likely to rupture (Morita et al. 2012). In a recent meta-analysis, the incidence of sIAs seems comparable in Japan and Finland to that of the rest of the world, hinting to the possibility of higher rupture risk explaining the higher incidence of SAH in those countries (Vlak et al. 2011). The possible genetic basis behind the higher rupture or incidence rate in Finland is investigated in study III.

2.4.2 Risk factors

sIA disease is acquired during life as suggested by the increasing incidence of sIAs by age and very low incidence in the first two decades of life (Rinkel et al. 1998). The average age of aSAH has been in the sixth decade of life but is increasing and currently estimated to be 62 years (Nieuwkamp et al. 2009).

The incidence of SAH becomes dominant in females in the sixth decade of life (de Rooij et al. 2007). The higher incidence of SAH in women after menopause suggest that hormonal changes may play a role (Mhurchu et al. 2001).

Smoking, hypertension, and excess alcohol consumption are the strongest modifiable risk factors for SAH (V. L. Feigin et al. 2005; V. Feigin et al. 2005). Hypertension increases the risk of SAH approximately 2.5 times and the risk effect of hypertension has been reported to be 30% higher in females (V. L. Feigin et al. 2005). Smoking at some point in life increases the risk of SAH by 2.2- to 3.1-fold when compared to never smoking. Similarly, previous smoking also increases the risk by 1.5 to 2.3-fold as compared to never smoking. Current smokers hav the highest risk of 2.2-3.1-fold as compared to never and previous smokers combined (V. L. Feigin et al. 2005). Excess alcohol intake (> 150g/week) approximately doubles the risk of SAH (V. L. Feigin et al. 2005). These three acquired risk factors accounted for SAH as follows: smoking 20%; alcohol ≥ 300 g/wk 21% and 100 to 299 g/wk 11%; and hypertension 17% (Ruigrok et al. 2001).

Over 10 % of aSAH patients have first or second degree relatives with SAH or unruptured aneurysms (Ruigrok et al. 2005; Huttunen et al. 2010). The relative risk of aSAH of first degree relatives of aSAH patients has been estimated to range from 4.1 to 6.6 with a population attributable risk of 11% (Ruigrok et al. 2001). Familial patients also more often seem to have multiple aneurysms (Mackey 2012; Ruigrok et al. 2004; Huttunen et al. 2010), suggesting that genetic factors may increase the risk of developing aneurysms. The association of smoking, female gender, and hypertension to the number of sIAs remains to be verified in large enough cohorts (Juvela 2000; Ellamushi et al. 2001; Qureshi et al. 1998). The association of SNPs, INDELs and structural variants to the number of sIAs Finland is investigated in Study III of the present thesis.

2.4.3 Molecular biology of intracranial extracerebral arteries and sIA wall

The phenotypical tissue of the sIA disease is the branching site of the intracranial extracerebral arteries. The arterial wall consists of three histologically separate layers: the intima consisting of endothelial cells; the media containing mainly smooth muscle cells; and the adventitia containing mostly fibroblasts. In extracranial arteries, the media is separated from the intima and the adventitia by the internal elastic layer (IEL) and the external elastic layer (EEL), respectively. Instead, the intracranial extracerebral arteries lack the EEL (Figure 6) (Ruigrok et al. 2005).

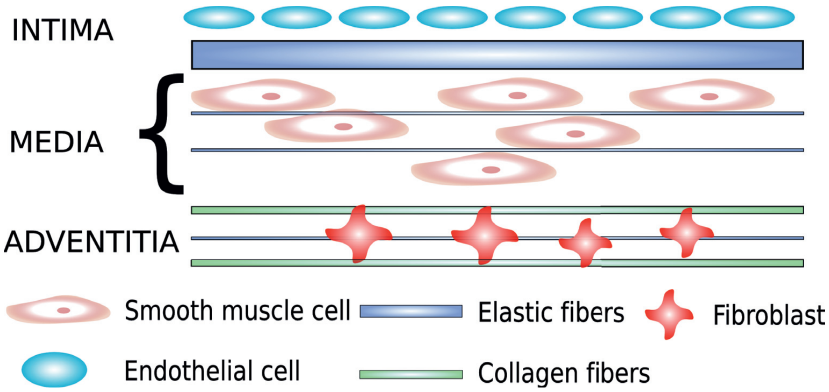


Figure 6. Schematic diagram of intracranial extracerebral artery wall structure. Adapted from (Ruigrok et al. 2005)

The histological studies of ruptured and unruptured sIA walls show degenerative and inflammatory changes associated to the rupture. Unruptured sIA walls already manifest damaged endothelial layer, absence of IEL, and disorganized media. The amount of smooth muscle cells in medial layer is also decreased (Scanarini et al. 1978; Frösen et al. 2004).

Ruptured sIA walls expectedly differ from unruptured ones. Some differences are likely due to the rupture event, but inflammatory and wall remodeling processes likely also precede the rupture (Frösen et al. 2004; Laaksamo et al. 2008; Kataoka et al. 1999). Specifically, immunohistochemichal studies indicate loss of endothelium and subsequent intraluminal thrombus formation, loss of mural cells and collagen matrix degradation (Frösen et al. 2004; Kataoka et al. 1999).

2.4.4 Genetics of the sIA disease

The tendency of the sIA disease to cluster in families (Ronkainen et al. 1997; Ruigrok et al. 2005) and the fact that familial background predisposes to the rupture (Ruigrok et al. 2001) at an earlier age (Huttunen et al. 2010; Bromberg et al. 1995) suggests that genetic factors might affect the risk of sIA formation and sIA rupture. On the other hand, the familial sIA clustering could also be affected by clustering of acquired socioenvironmental risk factors. Based on a large Nordic twin study, the concordance of subarachnoid hemorrhage was low in monozygotic twins which suggests low heritability of SAH in general (Korja et al. 2010).

A few Mendelian diseases are associated with sIA disease. Autosomal dominant polycystic kidney disease (ADPKD) caused by mutations in PKD1 (85% of the cases) and/or PKD2 genes is clearly and specifically associated with the saccular form of IAs (Torres et al. 2007). Some 8% to 12% of ADPKD patients develop sIAs (Pirson et al. 2002; Xu et al. 2011). A recent meta-analysis reported a prevalence ratio of unruptured IAs of 6.9 (95% CI 3.5–14.0) in ADPKD and of 3.4 (1.9–5.9) in familial sIA disease as compared to the general population (Vlak et al. 2011). ADPKD however accounts for only 0.3% of aSAH cases

(Ruigrok et al. 2001). Autosomal dominant fibromuscular dysplasia patients also seem to have increased prevalence of sIAs (7.3%) (Cloft et al. 1998). Many candidate gene studies have been conducted since the late 90's, motivated mainly by Mendelian diseases suggested to be associated to the SIA disease, but these studies did not identify consistently replicating risk variants (Nahed et al. 2007). Several linkage studies aimed to identify chromosomal regions segregating in sIA families have identified many chromosomal loci (Nahed et al. 2007; Ruigrok & Rinkel 2008) but only a few loci (1p34.3–p36.13, 7q11, 19q13.3, and Xp22) have been replicated in an independent studies in other populations. The identified regions are large, containing many genes, and the significance of these loci is unknown.

The first genome-wide association study of sIA disease was published in 2008 (Bilguvar et al. 2008). The study employed a two-stage design: genome-wide discovery (289,271 SNPs) in Finnish (920 cases, 985 controls) and Dutch (781 cases, 6424 controls) samples followed by replication of the most promising associations ($p < 5 \times 10^{-7}$) in a Japanese sample (495 cases, 676 controls). The study identified and replicated three SNPs in three loci (2q33.1, 8q11.23–q12.1 and 9p21.3). In addition, the 8q11 locus contained a second independent genome-wide significant SNP in the European sample, which failed to replicate in the Japanese sample.

In a follow-up study, two samples were added to the discovery phase, a German sample (789 cases and 2228 controls) and a combined European sample from Germany, Great Britain, Hungary, The Netherlands, Switzerland, and Spain (475 cases and 1940 controls) (Yasuno et al. 2010). The replication phase was augmented by adding a new Japanese sample (2,282 cases and 905 controls) and by increasing the cases and controls (a total of 829 cases and 761 controls) in the previous Japanese replication sample. The number of studied SNPs was increased to ~832,000 by imputation based on the HapMap phase 2 CEU reference panel. The study identified three new risk loci in 18q11.2, 13q13.1 and 10q24.32 and confirmed two of the previous loci in 8q11.23–q12.1 and 9p21.3. The previously reported locus 2q33.1 was not carried forward to replication as the probability of association was lower than the probability of no association in the applied Bayesian association analysis. In the 8q11 locus, the replicated SNP was the one that failed replication in the first study, and the SNP previously replicated, failed to reach genome-wide significance in the discovery phase and also failed to replicate in the Japanese sample.

In a second followup study using the same samples, the 14 loci with weaker associations in the discovery phase (probability of association 0.1 - 0.5) were taken to replication (Yasuno et al. 2011). Three loci (4q31.23, 12q22 and 20p12.1) replicated in the Japanese sample, but only one of those on 4q31.23 reached genome-wide significance when the samples were combined. It is noteworthy that the 2q33.1 locus discovered in the first study and not replicated in the second study was one of the 14 loci carried to the replication, and it did replicate in terms of statistical significance but the risk allele was different (i.e., the risk allele in the discovery phase was a protective allele in the replication phase). The authors suggested that this controversy might be caused by allelic heterogeneity between European and Japanese populations, and stated that more studies are needed. This locus is further studied in Study III of the current thesis.

The three GWA studies have identified 6 definitive and 1 probable loci, and they are estimated to account for 6.1%, 4.4%, and 4.1% of the four-fold sibling recurrence risk in Finnish, other European, and Japanese populations, respectively (Yasuno et al. 2011). Although most of the associated variants are located within or in LD with adjacent genes, suggesting those genes as IA candidate genes, the functional significance of the loci are not known.

The first functional clues of the pathophysiological mechanisms of sIA GWAS loci was revealed in a cross-phenotype study which investigated the association of the sIA loci with hypertension, a strong acquired risk factor of sIA disease (Gaál et al. 2012). Among the six established and 13 suggestive loci (Yasuno et al. 2010; Yasuno et al. 2011), a suggestive variant on 5q23.2 was significantly associated with increased systolic blood pressure in 9893 Finnish individuals. The association was replicated in a cohort of 200,000 individuals of European descent. Summary of replicated genomic loci affecting affecting sIA disease risk are shown in Table 3.

Table 3. Replicated genetic variants/loci linked to sIA disease

Locus	***Gene***	***RAF***	***Study type***	***Effect size***	***Reference***
4q31.23	*EDNRA*	*85%**	*GWAS*	*OR 1.22 (95%-CI 1.14–1.31)*	*(Yasuno et al. 2011)*
8q12.1	*SOX17*	*83%**	*GWAS*	*1.28 (1.20–1.38)*	(Yasuno et al. 2010)
9p21.3	*CDKN2A/ CDKN2B*	*58%**	*GWAS*	*1.32 (1.25–1.39)*	(Yasuno et al. 2010)
10q24.32	*CNNM2*	*91%**	*GWAS*	*1.29 (1.19–1.40)*	(Yasuno et al. 2010)
13q13.1	*KL, STARD13*	*23%**	*GWAS*	*1.20 (1.13–1.28)*	(Yasuno et al. 2010)
18q11.2	*RBBP8*	*48%**	*GWAS*	*1.22 (1.15–1.28)*	(Yasuno et al. 2010)
1p34.3-1p.36.13	*Many. Potential candidate HSPG2 (Ruigrok et al. 2006)*	*Likely rare*	*LNK*	-	(Nahed et al. 2005; Ruigrok et al. 2008)
7q11	*Many. Potential candidate ELN* (Onda et al. 2001; Akagawa et al. 2006)	*Likely rare*	*LNK*	-	(Onda et al. 2001; Farnham et al. 2004)
19q13.3	*Many.*	*Likely rare*	*LNK*	-	(Olson et al. 2002; Yamada et al. 2004; van der Voet et al. 2004; Mineharu et al. 2007)
Xp22	*Many.*	*Likely rare*	*LNK*	-	(Yamada et al. 2004; Ruigrok et al. 2008)
16p13.3	*PKD1*	*Rare*	*MEND*	*Prevalence of sIA is 6.9 higher in ADPKD patients than in general population.*	*(Vlak et al. 2011)*
4q22.1	*PKD2*	*Rare*	*MEND*	*Prevalence of sIA is 6.9 higher in ADPKD patients than in general population.*	*(Vlak et al. 2011)*

RAF: Risk allele frequency; LNK: linkage; MEND: Mendelian disease associated with sIA disease; ADPKD: Autosomal Dominant Polycystic Kidney Disease; * European populations in 1000 Genomes Project

2.4.5 Whole genome expression profiling of human sIA tissue

Five whole genome gene expression studies of human sIA tissues have been published (Table 4). Krischek et al. (Krischek et al. 2008) and Shi et al. (Shi et al. 2009) compared combined ruptured and unruptured intracranial aneurysm (IA) wall tissues with control artery walls and reported, for example, major histocompatibility complex class II overexpression, inflammatory response, and apoptosis as characteristic processes of aneurysm wall tissue. Li et al. (Li et al. 2009) compared unruptured IAs only with control arteries and, in contradiction, reported downregulation of several immune-related genes. Pera et al. reported cell adhesion and muscle system related genes to be downregulated and immune system and inflammatory response related genes to be upregulated when comparing ruptured and unruptured aneurysms together to control arteries (Pera et al. 2010). Marchese et al. compared ruptured sIAs to control arteries and reported structural proteins of the extracellular matrix, members of matrix metalloproteinase (MMP) family (which resulted as being overexpressed) and genes involved in apoptotic phenomena to be differentially expressed (Marchese et al. 2010).

Transcriptomic comparisons of human ruptured and unruptured sIA walls have been performed in four studies (Table 4). Very few differentially expressed genes were identified in those studies and consequently statistically significant pathway analyses were not possible. Pera et al (Pera et al. 2010) reported only 1 upregulated gene in ruptured IA walls compared with unruptured walls and stated, in slight contradiction to previous immunohistochemical studies, that some inflammatory genes were more highly expressed in unruptured IAs.

In these previously published whole-genome expression studies there were some seemingly contradicting results expecially whether the inflammatory reactions are upregulated in ruptured or unruptured sIAs or whether inflammatory genes are downregulated in IA tissues as compared to controls. Many differences could lead to these results including: different case vs. control selection, different statistical methods used, small sample sizes in many studies, suboptimal annotation files (Sandberg & Larsson 2007) used in Affymetrix analysis by Marchese et al and Li et al. In study II of the current thesis whole-genome expression profile between ruptured and unruptured sIAs is investigated to shed light on the role of inflammation in aneurysm rupture and to identify more specific pathways potentially contributing to sIA rupture.

Table 4. Whole-genome gene expression profiling studies of human saccular intracranial aneurysm tissue.

Cases	***Controls***	***Genes ↑***	***Genes ↓***	***Array***	***Ref.***
6 RAs and 4 UAs	*4 AVM feeder arteries*	*263*	*258*	*Agilent Human 1A(v2)*	(Krischek et al. 2008)
6 RA	*4 UAs*	*2**	*0*	*Agilent Human 1A(v2)*	(Krischek et al. 2008)
3 RAs and 3 UAs	*6 STAs from the same patients as cases*	*172*	*154*	*Illumina Human WG6-v2*	(Shi et al. 2009b)
3 RAs	*3 UAs*	*0*	*0*	*Illumina Human WG6-v2*	(Shi et al. 2009b)
3 UAs	*3 STAs*	*164*	*996*	*Affymetrix HU133 Plus 2.0*	(Li et al. 2009)
8 RAs and 6 UAs	*5 MMA*	*8*	*123*	*Affymetrix GeneChip Human Gene ST 1.0*	(Pera et al. 2010)
8 RAs	*6 UAs*	*1*	*31*	*Affymetrix GeneChip Human Gene ST 1.0*	(Pera et al. 2010)
12 RAs	*10 UAs*	*10*	*4*	*Affymetrix U133A*	(Marchese et al. 2010)
12 RAs	*5 STA/MMA from the same patients as cases*	*22*	*8*	*Affymetrix U133A*	(Marchese et al. 2010)

RA: Ruptured saccular intracranial aneurysm; UA: Unruptured saccular intracranial aneurysm; STA: Superficial temporal artery; AVM Arteriovenous malformation; Middle meningeal Artery; * Did not specify if the genes were upregulated or downregulated.

2.5 CHALLENGES AND BIOINFORMATIC SOLUTIONS FOR INTERPRETATION OF RESULTS FROM HIGH THROUGHPUT ANALYSES

High-throughput methods have revolutionized the research of complex diseases by enabling simultaneous genome-wide measurement of genetic variants, gene expression, and protein levels. Current technologies are able to produce massive amounts of data in a short time and the bottleneck in the research process has shifted from producing the data to intelligent bioinformatic analyses. Each technology platform has its own idiosyncrasies often requiring sophisticated statistical methods to obtain reliable and comparable results. A researcher obviously must be aware of these and to apply appropriate statistical methods, not to mention proper experimental design. These questions are not discussed further as they are out of the scope of this review but the interested reader is referred to reviews on these topics (Leek et al. 2010; Trapnell et al. 2012; Roberts et al. 2011; Allison et al. 2006).

After initial data-analysis, the researcher often ends up with a set of genes that are differentially expressed between some conditions in gene expression profiling or with genome-wide risk estimates and their associated statistical significance values in GWAS experiments. These lists in isolation, especially in gene expression studies, fail to convey much insight into the molecular mechanisms underlying complex diseases. In this chapter, bioinformatic approaches for analyzing gene sets from high-throughput methods are reviewed focusing mainly on methods for analyzing gene sets obtained from e.g. transcriptomic or proteomic studies, while similar methods in GWAS studies are only briefly treated.

2.5.1 Biological body of knowledge and digital representation

The first essential building block of computational analysis of gene sets from high-throughput analyses is computationally accessible biological knowledge about genes such as functions, regulation of expression, localization in cell or organism, interactions with other genes and environment. Obtaining such knowledge for a larger set of genes from traditional scientific literature is time consuming manually, and error-prone computationally, as the data about gene functions is disseminated in individual research articles, described in natural languages. A common strategy to represent scientific knowledge is the creation of controlled vocabularies and using the vocabulary to describe the knowledge gained from original data. This standardization is important so that discoveries on gene functions by individual researchers can be unambiguosly defined, made possible to search computationally and faster to comprehend manually (Bard & Rhee 2004).

Arguably the most widely used and accepted source of information on individual gene product functions is the Gene Ontology (GO) project and associated databases (Ashburner et al. 2000). Gene Ontology is a standardized controlled vocabulary of terms used to annotate different aspects of gene functions. The GO defines three categories of annotations: biological process, molecular function, and cellular component. The Biological process category describes end points of biological activity that a gene somehow contributes to e.g. ″positive regulation of cell death″. The molecular function annotations describe biochemical activity such as ″receptor binding″, irrespective of when or where that binding might occur. The cellular component annnotations describe a location within a cell where the gene product is active. The terms are organized as a directed acyclic graph where the root nodes are the three different categories and each child node describes the aspect more specifically than the parent node (Figure 7). This has the benefit that two genes can be inferred to share a same more general function even though they have been described at a different level of abstraction. All of the gene product annotations have evidence code associations. Some of the annotations are based on experimental evidence and manually checked by the curators such as the evidence code "Inferred from direct assay (IDA)" while some annotations are reviewed by curators but are based on computational predictions such as inferred from Reviewed Computational Analysis (RCA). One of the largest annotation category is comprised of e.g. automatic transfer of annotations from other databases or computational predictions without manual curation: "Inferred from electronic annotation (IEA)". The Gene Ontology consortium provides annotations for over 20 species, including human and model organisms such as mouse, dog, and the worm Caenorhabditis elegans. Each species' annotations are governed by a single member organisation of the consortium, each with their own procedures for annotation. Human annotations are governed by the European Bioinformatics Institute (www.ebi.ac.uk/GOA), providing GO annotations for proteins in UniProt protein resource (http://www.uniprot.org). For the full list of consortium members, included species, annotation codes, and their explanations refer to the consortium's web site (www.geneontology.org)

While Gene Ontology is somewhat comprehensive in human annotations (45,555 protein products and 200,047 curated annotations as of 13.8.2013) and widely used, it allows annotation of genes and gene products with a limited attribute set. For example it does not contain attributes for disease involvement because e.g. oncogenesis is not a normal function of any gene. For the same reasons the GO does not contain attributes for in what tissues or in what developmental stage the gene might be expressed. There are other gene ontologies being developed for these types of more specific use cases although the actual gene annotations are often lacking. The Open Biological and Biomedical Ontologies (OBO) is a community of different ontology developers aiming to develop shared principles for ontology development. OBO Foundry is an extension to the original OBO, establishing a set of principles for ontology creation with the aim to create a set of interoperable and non-reduntant ontologies in the biomedical domain (Smith et al. 2007). The OBO foundry site lists the foundry ontologies as well as foundry candidate ontologies (www.obofoundry.org). Some emerging ontologies likely to to be relevant to complex human disease studies are ontologies like human disease ontology (Osborne et al. 2009) and human phenotype ontology (Robinson & Mundlos 2010).

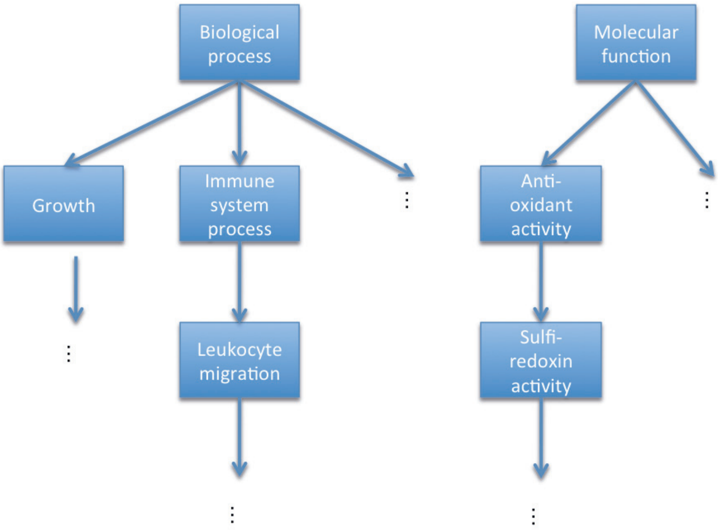


Figure 7. Illustration of a Gene Ontology tree. Annotations can be depicted as directed acyclic graph where directed edges (arrows) go from broader term to more specific term.

One of the limitations of using Gene Ontology annotations for gaining understanding to disease mechanisms from a gene set obtained from high-throughput studies is that the biological process annotations do not necessarily describe a single coherent pathway, but just annotates the gene product's functions that aim to a similar biological endpoint or function in the same cellular location. In that sense, Gene Ontology is not strictly a pathway resource. The other class of pathway resources aims to define more focused signalling pathways (often called canonical pathways). In addition to classifying genes as performing a certain biological process, a pathway topology is desribed, defining interactions between gene products (Figure 8). One of the earliest and most widely used such resources is the Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY database (Kanehisa & Goto 2000; Kanehisa et al. 2012). The KEGG PATHWAY database contains manually compiled signalling pathways based on literature. KEGG provides pathway maps aiming to capture, from some aspect, metabolic pathways (e.g. steroid hormone biosynthesis), signalling cascades related to different human diseases (e.g Alzheimers disease), specific signalling pathways (e.g. Wnt signalling pathway), different organismal systems (e.g. Cardiac muscle contraction), and cellular processess (e.g. Apoptosis). Another similar noteworthy resource gaining popularity and providing manually compiled and curated pathways is Reactome (Croft et al. 2011). Reactome pathways are created by experts in collaboration with Reactome staff but are also subjected to peer-review, which is expected to increase the accuracy of the provided pathways. Currently, Reactome contains ~1,300 human pathways containing ~6,300 proteins as well as pathways for 19 other species. Several other pathway resources are constantly being developed, each with their own annotation processes and focus areas. The Pathguide online resource (www.pathguide.org) lists and updates pathway databases and currently lists over 60 signalling pathway resources.

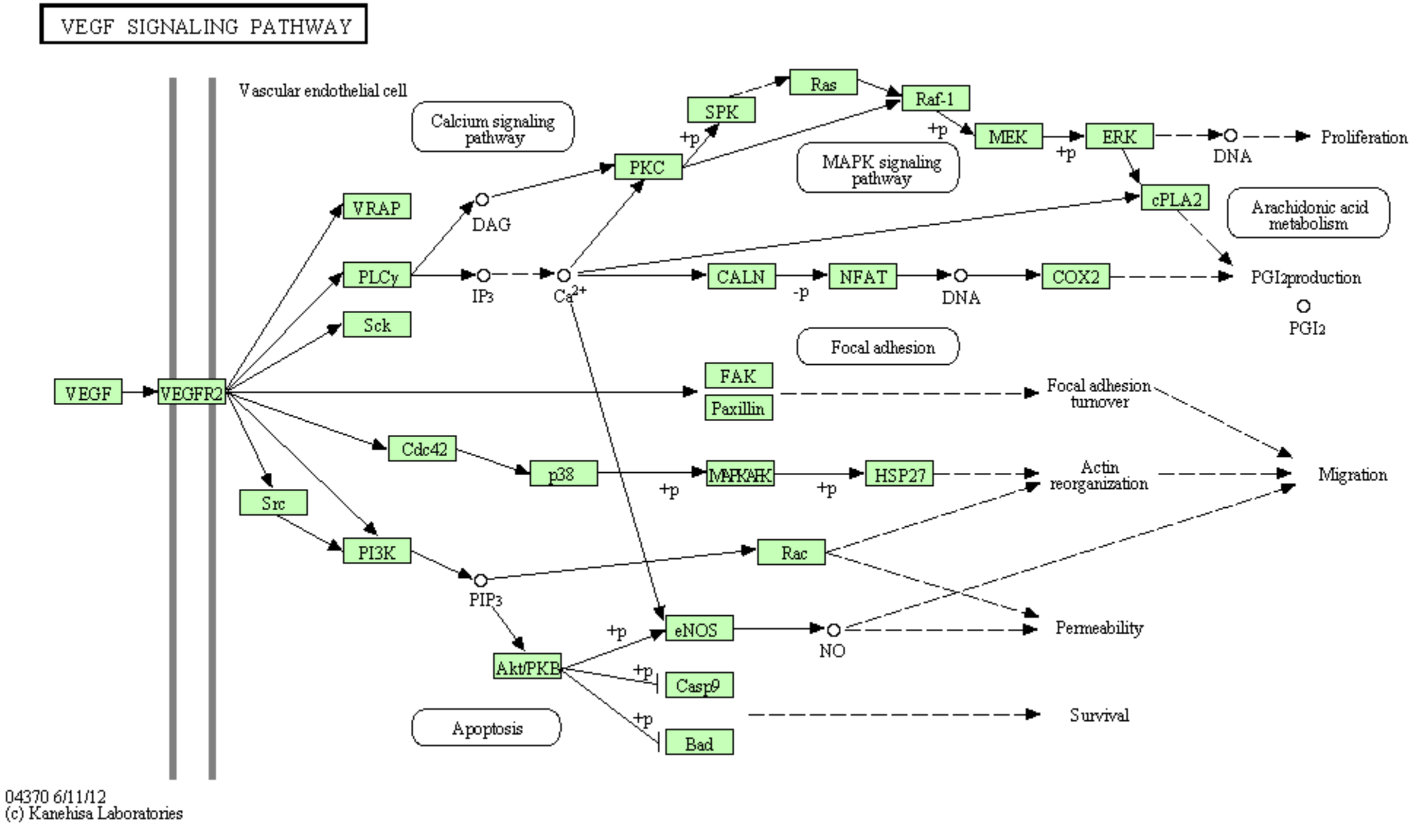


Figure 8. Example of a more elaborate pathway including pathway topology. (http://www.genome.jp/kegg/)

2.5.2 Pathway analysis methods

Once pathway resources are selected, the next step in pathway analysis of gene sets is to computationally relate the differentially expressed genes to the pathways and statistically assess the degree of belief that this relation would not have occurred just by chance. During the last decade, this has been an active area of research and several classes of analytical approaches have emerged. There are far too many individual methods and software developed for this purpose and hence individual methods will not be evaluated or discussed, rather the different types of approaches are presented and the interested reader is referred to a comparison article of 68 different pathway analysis methods (Huang et al. 2009). The classification of pathway tools and discussions are inspired and based on a review by Khatri et al. (Khatri et al. 2012)

2.5.2.1 Overrepresentation analyses

The first methods developed assess the overlap between the differentially expressed gene set (e.g. genes significantly differentially expressed between healthy vs. diseased samples) and gene sets derived from chosen pathway resources. Each of the gene sets in a pathway resource in turn is tested for overlap, and statistical significance of the overlap is tested using e.g. conventional Chi-square or Fisher's exact test. As there are typically hundreds of gene sets tested, the likelihood of getting significant results just by chance increases and multiple testing correction must be applied to control against false positive findings.

The result of overrepresentation analysis is then a ranked list of statistically significant pathways and the differentially expressed genes grouped into the respective pathways. This aids the researcher in focusing on enriched biological processess instead of just individual genes.

Despite the usefulness and widespread usage of overepresentation analysis, some limitations can be identified. First, this type of analysis ignores the magnitude of differential expression and the strength of statistical evidence and treats every gene in the gene set equally. In some circumstances it could be appropriate to weight the genes e.g. with higher fold change more, which might lead to identifying significant pathways. However, also the genes that exhibit only subtle differential expression can be equally, or more important in given context and inappropriately downweighting such genes might then actually lead to not being able to identify the important pathways. A second type of potential limitation is that in this approach, the researcher must decide a strict significance cut-off when obtaining the differentially expressed gene set whereafter the rest of the genes are ignored. Marginally less significant genes (say, two genes with p-values 0.049 and 0.051) are practically just as significant but the one is considered to be fully significant and the other one is discarded, potentially losing important information. The third limitation is that these methods treat each gene and pathway independently of each other (other than assigning genes belonging to the same pathway). In reality the biological processess in cells and tissues are highly complex interactions between genes and different pathways, and novel insights might be gained if the interaction information were appropriately utilized. A final shortcoming is that these methods totally rely on previous knowledge about gene functions and on a limited and narrow focused knowledge on the complex interactions between genes in forming biological pathways. The curated pathways are often based on experimental conditions in cell lines or tissues and conditions that might differ in unknown ways from the context the researcher is facing. Study II presents one bioinformatic approach that can be used for hypothesizing potentially novel gene functions and biological processes in studied condition.

2.5.2.2 Enrichment analyses

In order to overcome the need to define strict cut-offs for selecting differentially expressed gene sets and to utilize more of the measurement information (fold-change or statistical significance), gene set enrichment analysis (EA) methods were developed. Many methods have been developed differing in individual aspects but most follow a similar general framework (Ackermann & Strimmer 2009).

These approaches first compute a gene-level statistic from molecular measurement of gene expression, such as t-statistic or Z-test statistics, comparing e.g. healthy vs. diseased tissue. The choice of the gene-level statistic seems not to be critical for detection of differentially expressed gene sets, however untransformed statistics can fail to detect pathways with up- and down-regulated genes. Transforming the statistic e.g. to absolute value can be useful in such situations. (Ackermann & Strimmer 2009). Next the individual gene-level statistics are aggregated to a single statistic per pathway. Statistics suggested can be multi-variate such as Hotelling T (Kong et al. 2006) , thus taking into account the interdependence between gene measurements , or univariate like average of t-test statistics (Tian et al. 2005), effectively ignoring the correlation between genes. Perhaps unexpectedly, although multi-variate statistics does lead to higher statistical power in the presence of high correlation in a simulation study, univariate statistics showed higher or at worst equal power in analysis of a real dataset, suggesting that some aspects of complex gene expression dynamics cannot be efficiently simulated (Glazko & Emmert-Streib 2009).

Statistical significance of pathway score is typically achieved by generating empirical null distribution of pathway-level test-statistics by permuting either the phenotypes of samples (e.g. diseased / healthy) or by permuting genes in the gene sets. Gene permutation tests if the observed gene set can be distinguished from randomly chosen gene set of the same size. On the other hand, phenotype permutation tests whether the association of gene set with phenotype is distinguishable from a random correlation of shuffled phenotypes (Tian et al. 2005). Notably, permuting sample labels preserves the complex correlation structure between genes and is often favored. (Dinu et al. 2009; Nam & Kim 2008). One of the first and most widely used methods (33,000 registered users and > 3,100 citations) Gene Set Enrichment analysis (GSEA), ranks genes based on differential expression statistic (e.g. fold change) and calculates a running sum statistic for the top or bottom of the list

(Weighted Kolmogorov-Smirnov type statistic) (Subramanian et al. 2005; Mootha et al. 2003). Also simpler and computationally faster parametric approaches such as z-test have been proposed (Irizarry et al. 2009).

Enrichment analysis methods typically overcome the three issues of over-representation analyses. First, they circumvent the need to define a strict threshold between significant and not significant differential expression of genes. Second, EA methods utilize the molecular measurement of expression in order to detect coordinated changes in the expression of genes in the same pathway. Third, the utilization of gene expression changes allows taking into account the correlation of genes in the pathway. Enrichment analyses share the shortcoming of over-representation analyses that they consider each pathway in isolation and ignore the topology of the signalling network of genes.

2.5.2.3 Topology based pathway analysis

EA and ORA approaches consider only the number of genes in a pathway or gene coexpressions to detect differentially expressed pathways and totally ignore the topology of the interaction provided by some signalling pathway resources, such as KEGG and Reactome. A third type of pathway analyses, topology based (TP) analyses, aims to take into account the topological ordering of genes in a pathway.

TP methods are typically similar to the three-step process of EA methods except that they use pathway topology information in computing the gene-level statistic. (Rahnenführer et al. 2004; Tarca et al. 2009; Draghici et al. 2007). For example, Impact Factor analysis (IF) and its enhancement SPIA combine the evidence of overrepresentation analysis and topology based analysis (Draghici et al. 2007; Tarca et al. 2009. IF defines a gene-level statistic (termed Perturbation Factor, PF) as the sum of fold change of the gene and PF of all the genes preceding the gene in the pathway. Pathway-level statistic (termed Impact Factor, IF) is then defined as a sum of all PFs of the genes in the pathway. Statistical significance of pathway-level statistic is calculated by repeatedly assigning as many random genes from the list of differentially expressed genes as there were differentially expressed genes in the pathway to replace a random gene in the pathway. IF is calculated for all of these repeated randomizations and the p-value is calculated by observing the number of times the randomizations achieved higher IF than the actual data (Draghici et al. 2007; Tarca et al. 2009).

TP methods are more recently developed and despite their intuitively appealing advantages, no method has gained significant popularity like GSEA, presumably at least due to lack of comprehensive comparison between methods. One limitation of PT analyses is that the topology of signalling is dependent on cell-specific gene expression patterns and conditions being studied. This knowledge is often not readily available (Bauer-Mehren et al. 2009).

2.5.3 Challenges and limitations of pathway analysis

Pathway analyses are an essential tool for interpreting large differentially expressed gene sets from experimental condititions but are mainly limited by incomplete and inaccurate gene function annotations. For instance, 95% of human Gene Ontology annotations are inferred from electronic annotations without manual curation. These annotations are likely to contain more false positives, although this has not been shown conclusively (Rhee et al. 2008). Furthermore, most of the pathways in pathway databases are curated by combining evidence from many different cell types and conditions (Bauer-Mehren et al. 2009). Therefore the available pathways may not correctly reflect the signalling pathway under the studied condition. Methods using literature mining of pubmed abstracts has been proposed to alleviate the problem of incomplete annotations in pathway analysis (Jelier et al. 2011).

Finally, all pathway analysis tools typically map the input genes to a single unique gene identifier. However, over 90% of the human genes are estimated to be alternatively spliced and these isoforms of the same gene may have both related and opposing functions (Wang et al. 2008). This mapping to a single gene identifier can be advantageous in that the tool is not dependent on the technology that generated the gene measurements but naturally leads to a loss of potentially valuable information. Study II introduces one approach for gene set analysis, where exact knowledge of all of the gene functions in studied condition might not be available.

2.5.4 Pathway analysis in GWA studies

Due to the high multiple testing burden in GWA studies, genes or loci may be genuinely associated with disease status but might not reach a stringent genome-wide significance threshold in a given GWA study. However, related genes showing coordinated moderate association to disease might pinpoint true positives among less significant genes. To study this possibility, a similar type of pathway analyses have been developed for genome-wide association studies as for genome-wide expression studies studies (Wang et al. 2010; Zhong et al. 2010; Ramanan et al. 2012). One key difference is that instead of having expression measures of each gene in pathway analysis, the involvement of a gene is assigned by a SNPs proximity to a gene. A SNP could be, for example, assigned to a gene if the SNP resides less than 1000 nucleotides away from a gene. The general framework of GWAS pathway analyses is to calculate a pathway score by taking a logarithm of the lowest p-value of each gene in the pathway and sum the scores of individual genes. This pathway score is then tested if it is larger than would be expected by chance. As pathways containing larger genes are more likely to get higher pathway scores (a larger gene has more SNPs and by chance would get lower p-values), many methods use phenotype permutation to correct for this bias. The GWAS pathway analysis methods mainly differ by the way gene-assignment is done and how the significance is tested. Similar issues are present in gene expression pathway analysis, but no clear consensus exists about the best approaches, as these methods are relatively recently developed (Wang et al. 2010).

2.5.5 Inferring regulation of differentially expressed genes

Once a differentially expressed gene set has been identified, the next logical question might be to determine the cause of differential expression. When studying complex human diseases and target tissues in a typical one time-point study, this question might be next to impossible to answer. Clues to the differential expression regulation can however be gained by analyzing the regulatory transcription factor binding sites in the vicinity of differentially expressed genes (see chapter 2.2.1). As the specific regulation of most of the genes by transcription factors in different cell types and conditions are not known, bioinformatic predictions can be a useful way to gain some insight into the observed differential expression.

Known transcription factor binding site motifs from resources such as Jaspar (Bryne et al. 2008) or Transfac (Wingender et al. 1996) can be used to search promoter sequences of differentially expressed genes for potentially functional binding sites (Wasserman & Sandelin 2004). A noteworthy problem with such an approach is that a mammalian organism can contain hundreds of times the number of binding motifs than are actually bound by a transcription factor in a given cell type (Zhang et al. 2009). Conservation of sequence around regulatory elements between closely related species is one common strategy to reduce identification of false positive binding sites (Huber & Bulyk 2006; Robertson et al. 2006), but may miss lineage specific regulatory elements (Hardison & Taylor 2012). The relative merits of different identification methods are not reviewed here but the interested reader is referred to a recent excellent review and references therein (Hardison & Taylor 2012).

Once putative regulatory elements are identified, overrepresentation or an enrichment type of analyses (see chapter 2.5.2) can be used to assess if some regulatory elements occur in the promoter regions of differentially expressed genes more often than would be expected by chance. Identification of such elements among promoters of differentially expressed genes between healthy and disease tissue could pinpoint common regulatory mechanisms putatively contributing to pathogenesis of human diseases (Nischan et al. 2009).

2.5.6 Clustering in high-throughput genomic data mining

In high-throughput data analysis such as microarray data mining, it is often desirable to group similar items (e.g., genes) together to facilitate interpretation. Clustering methods identify subgroupings of observations so that members in a group are more similar to each other (with some chosen metric) than to members of other groups. One clustering method is applied in Study II to identify functionally similar gene groups from differentially expressed genes. The result of clustering is dependent on the chosen clustering method and parameter values. For a general technical treatment of clustering methods see e.g. (Hastie et al. 2008) and specifically in application to microarray gene expression data see e.g. (Kerr et al. 2008).

2.5.7 Multiple testing burden

The possibility to make a large number of simultaneous measurements, either by performing genome-wide association or transcription experiments, or testing hundreds of pathways for enrichment is a double-edged sword. The positive side is that links between genetic marker and phenotype, gene-expression and phenotype or pathway and phenotype will be tested (if they exist), but the flip-side is that by conducting hundreds to millions of statistical tests, the probability of false positive findings increases. This occurs in frequentist statistics because in a single test we typically might reject a null hypothesis if the probability of being wrong is less than 1/20 (i.e. p-value < 0.05). If we make 20 statistical tests, the probabilities of being wrong add up and one of those tests is expected to reject the null hypothesis, even if the null hypothesis were true. Appropriate multiple testing correction methods must therefore be applied in the analysis of high-throughput data and in bioinformatic datamining.

If the statistical tests are uncorrelated and moderate in number, traditional Bonferroni correction is appropriate (divide the critical p-value by the number of tests performed). However when the tests are correlated (e.g. SNPs in LD or correlated genes) the Bonferroni correction becomes overly conservative (i.e. increased false negative rate), hence decreases the statistical power. In such situations computationally demanding permutation methods can be an option (Subramanian et al. 2005; Churchill & Doerge 1994).

Multiple testing correction methods that control the probability of making one or more false positives among all the statistical tests are called family-wise error rate (FWER) methods. For example Bonferroni correction is a FWER method. False discovery rate (FDR) methods offer less stringent multiple testing correction. FDR methods control the expected number of false positives among rejected null hypotheses. FDR methods such as Benjamini and Hochberg False Discovery Rate (Benjamini & Hochberg 1995) are often favored to increase statistical power e.g. in whole-genome gene expression studies where the number of tests far exceed the number of samples (e.g. Study I).

3 Aims

The overarching aim of the current study was to elucidate: (a) genetic predisposition to sIA disease and (b) signaling processes leading to sIA wall rupture, using genomewide methods and related bioinformatics. This knowledge is needed for a better understanding of the molecular mechanisms of sIA pouch formation and sIA wall rupture which could pave the way for the design of novel methods for non-invasive diagnosis, prevention, and occlusion of sIAs. The main approaches were signalling pathway analysis by whole-genome expression study, development of a novel bioinformatic method to be applied to the expression data, and a genome-wide association analysis utilizing the high-risk population of Finland.

The specific aims of the study were:

1) **To identify signalling pathways active in ruptured human sIA walls as compared to unruptured ones using whole genome transcriptome profiling and related bioinformatics.**

- We aimed to identify differentially expressed genes and signalling pathways and their putative control by transcription factors by comparing ruptured and unruptured human sIA walls obtained and snap frozen during microsurgery. The acquired knowledge would increase the understanding of mechanisms leading to sIA wall rupture. Identified candidate genes and pathways might be used as a target for future therapeutic development.

2) **To develop a novel method and software for interpreting differentially expressed gene sets to gain further insight into the signaling processes orchestrated by the differentially expressed gene set in Study I (Aim 1).**

- We hypothesized that there are a multitude of signalling processes driven by subsets of the differentially expressed gene set in Study I. Such specific processes might not be identified using the existing enrichment and pathway analysis software.
- We aimed to develop a method that could identify functionally coherent gene subsets from differentially expressed genes and would provide clues of transcriptional regulation of those subsets.
- We hypothesized that such a method could identify additional candidate genes and biological processes that could serve as a basis for novel hypothesis generation concerning the mechanisms leading to sIA rupture.

- We aimed to make the method publicly available as an easy to use software for other researchers and to be used as a basis for further method development.

3) To identify novel genetic factors affecting susceptibility to sIA disease by genome-wide association analysis and followup replication in the high-risk population of Finland.

- We aimed to identify novel genetic risk loci predisposing to sIA disease by performing genome-wide association analysis augmented by genotype imputation of low frequency variants in a high risk population isolate of Finland.
- We aimed to identify novel genetic risk loci predisposing to aneurysm formation using the number of aneurysms as a phenotype.
- We sought to confirm previously published association loci with inconclusive positive evidence in European populations, including Finland, but no replication in Japan.
- We aimed to gain evidence to support our hypothesis that the higher incidence of sIA in Finland is at least in part due to genetic factors.

4 *Upregulated signaling pathways in ruptured human saccular intracranial aneurysm wall: an emerging regulative role of Toll-like receptor signaling and nuclear factor-κB, hypoxia-inducible factor-1A, and ETS transcription factors*[1]

4.1 INTRODUCTION

The mechanisms of the initiation, progression, and rupture of the sIA pouch need to be elucidated for the design of novel methods for noninvasive diagnosis, prevention, or occlusion of sIAs. In previous microarray studies, Krischek et al. (Krischek et al. 2008) and Shi et al. (Shi et al. 2009) compared combined ruptured and unruptured intracranial aneurysm (IA) wall tissues with control artery walls and reported, for example, major histocompatibility complex class II overexpression, inflammatory response, and apoptosis as characteristic processes of aneurysm wall tissue. Significant changes were not found between unruptured and ruptured IA walls. Li et al (Li et al. 2009) compared unruptured IAs only with control arteries and, in contradiction, reported downregulation of several immune-related genes. Pera et al (Pera et al. 2010) reported only 1 upregulated gene in ruptured IA walls compared with unruptured walls and stated, in contradiction, that inflammatory genes were more highly expressed in unruptured IAs. Marchese et al (Marchese et al. 2010) reported 10 upregulated and 4 downregulated genes in ruptured IAs.

The mechanisms in sIA wall rupture are poorly understood. Because most sIAs do not rupture, it is possible that ruptured sIAs have a pathobiology distinct from unruptured sIAs. So far, transcriptome profiling of ruptured and unruptured sIA walls has not revealed specific pathways related to sIA wall rupture. In the present study, we compared the transcriptomes of 11 ruptured and 8 unruptured human sIA walls to identify pathways that are askksociated with the rupture and to computationally predict transcriptional control of those pathways.

4.1.1 Materials and methods

4.1.2 Patients and Tissue Samples

Fundi of 18 ruptured and 11 unruptured sIAs were resected after microsurgical clipping of the neck (*Table 1*) as previously described (Frösen et al. 2004; Frösen et al. 2006; Tulamo et al. 2006; Laaksamo et al. 2008). All patients were of Finnish ethnicity. The samples were immediately snap-frozen in liquid nitrogen and stored in the Helsinki Neurosurgery sIA Tissue Bank. The medical records of the 29 sIA patients were reviewed (Table 5). The Ethical Committee of Neurology, Ophthalmology, Otorhinolaryngology, and Neurosurgery of the Helsinki University Central Hospital approved the study.

Table 5. Patients, Saccular Intracranial Aneurysm Samples, and Methods

Sample No	Sex	Age Years	Location of sIA	Rupture of sIA	Time from Rupture (h)	Microarray	qRT-PCR
1.	F	60	MCA	-		+	-
2.	F	64	ICA	-		+	-
3.	M	47	MCA	-		+	-
4.	F	37	MCA	-		+	-
5.	M	42	MCA	-		+	+
6.	F	62	MCA	-		+	+
7.	M	56	PCoA	-		+	+
8.	F	65	MCA	-		+	+
9.	F	56	MCA	-		-	+
10.	M	42	MCA	-		-	+
11.	F	59	MCA	-		-	+
12.	F	54	MCA	+	16	+	-
13.	F	46	ACoA	+	96	+	-
14.	M	58	MCA	+	24	+	+
15.	F	71	ACoA	+	216	+	+
16.	F	52	ICA	+	168	+	+
17.	F	32	MCA	+	3	+	+
18.	F	83	MCA	+	2.6	+	+
19	M	73	MCA	+	3.6	+	+

20.	F	69	MCA	+	6.7	+	+
21.	F	53	MCA	+	N.A.	+	+
22.	M	70	MCA	+	14	+	+
23.	F	57	PCoA	+	6.4	-	+
24.	F	58	MCA	+	12	-	+
25.	F	44	MCA	+	11	-	+
26.	F	53	MCA	+	24	-	+
27.	F	47	ACoA	+	5.2	-	+
28.	M	41	ACoA	+	9.1	-	+
29.	F	62	PCoA	+	72	-	+

F = female; M = male; MCA = middle cerebral artery; PCoA = posterior communicating artery; ACoA = anterior communicating artery; ICA = internal carotid artery; N.A. = not available.

4.1.3 Isolation of mRNA and Microarray Hybridization

Total RNA was isolated with Trizol Reagent (Invitrogen, Carlsbad, California) according to the manufacturer's instructions and as described previously(Hiltunen et al. 2002). The quantity and quality of RNA were analyzed with NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies Inc, Wilmington, Delaware), and only good-quality RNA with an A260/A280 ratio of > 1.8 and < 2.0 was used. RNA from the whole wall of 11 ruptured and 8 unruptured samples was used to prepare hybridization mixes according to the Affymetrix 2-cycle amplification protocol. Briefly, 100 ng of total RNA was used to synthesize double-stranded cDNA. The cDNA was purified and transcribed to biotin-labeled cRNA. Purified cRNA (20 mg) was hybridized to Affymetrix Human Genome U133 Plus 2.0 GeneChips. The chips were stained, washed (Affymetrix Fluidics Station 400), and scanned (Affymetrix GeneChip Scanner 3000) according to the manufacturer's instructions.

4.1.4 Microarray Data Analysis

Microarray analyses were performed with R statistical software version 2.9.1 (R Development Core Team 2009) and Bioconductor version 2.4.1 (Gentleman et al. 2004). Data import was done with Affy package version 1.2226 with the BrainArray CustomCDF version 12 custom chip description file for probe set matching and gene annotations (Dai et al. 2005; Sandberg & Larsson 2007). There were 17 788 distinct genes defined by the custom chip description file.

The Robust Multichip Average was used to normalize expression values between arrays and to generate a single expression measure for each gene from individual probes (Wu et al. 2004). Nonspecific filtering was used to filter out less informative probe sets not linked to genes and probe sets with small variance across samples (50% of probe sets with the least variation). Linear models for microarray data version 2.18.3 analysis package (Smyth 2004) was used to detect the differentially expressed genes between sample groups, using fitting of linear models and applying empirical Bayes variance smoothing to each probe set. Because of the anticipated large biological gene specific variation between individuals, a robust MM estimator was used (as implemented by rlm method in MASS R package (Venables & Ripley 2002)), as it is less sensitive to outliers than the least-squares estimation. The Benjamini and Hochberg (Benjamini & Hochberg 1995) false discovery rate was used to adjust for multiple testing, and adjusted values of $P < 0.05$ were considered significant.

4.1.5 Functional Analysis of Differentially Expressed Genes

An overrepresentation analysis was performed for the upregulated and downregulated gene lists separately. For this enrichment analysis of gene ontology terms and Kyoto Encyclopedia of Genes and Genomes pathways, GOstats R package version 2.1 (Beissbarth & Speed 2004) and database for annotation, visualization, and integrated discovery (DAVID) (Huang da et al. 2009; Dennis Jr et al. 2003) bioinformatics resource were used. All of the distinct 17 788 genes in the array were used as a background gene set. To avoid reporting redundant ontologies, a conditional gene ontology analysis strategy was used that reports only the most specific gene ontology terms in the hierarchy that are statistically overrepresented in the differentially expressed gene sets (Beissbarth & Speed 2004). The Benjamini and Hochberg (Benjamini & Hochberg 1995) false discovery rate was used to adjust for multiple testing, and adjusted values of $P < 0.05$ were considered significant.

The similarity of differentially expressed gene sets to genes genetically associated with different diseases and disease classes was assessed with DAVID (Huang da et al. 2009) with the Genetic Association Database as the data source used for disease association.

In the ruptured sIA wall samples, the time elapsed from rupture to resection of the sample may affect the gene expression levels. The levels were compared between the early (2.6-16 hours) and delayed (24-216 hours) time groups. The correlation between the elapsed time and the expression level of each gene was calculated. In both tests, the P values were adjusted for multiple testing correction with the Benjanimi and Hochberg (Benjamini & Hochberg 1995) false discovery rate, and corrected values of $P < 0.05$ were considered significant. The Kruskal nonmetric multidimensional scaling method implemented in the MASS R package (Venables & Ripley 2002) was used to arrange each sample according to expression level differences of all genes between samples, and clustering according to the elapsed time was visually assessed.

4.1.6 *In Silico* Transcription Factor Analysis

The enrichment of binding motifs for transcription factors located within 5000 bp upstream of the transcription start site of differentially expressed genes was assessed with Whole Genome rVista (Zambon et al. 2005). Start of the 5′ exon was always used to define transcription start site. The binding motifs and surrounding genomic sequences were required to be conserved between human and mouse to reduce the number of false positives among the detected binding sites. Enrichment values of $P < 10^{-5}$ were considered significant.

4.1.7 Validation of Microarray Data

Expression of 5 genes was studied by quantitative polymerase chain reaction in 17 ruptured and 6 unruptured sIA samples (Table 5). Briefly, 500 ng of total RNA was reverse transcribed into cDNA using random hexamers (Promega, Madison, Wisconsin) and M-MuLV reverse transcriptase (MBI Fermentas, Hanover, Maryland). Quantitative measurements of mRNA levels were per-formed with assays-on-demand gene expression reagents (Applied Biosystems, Foster City, California) with the ABI PRISM 7700 Sequence Detection System (Applied Biosystems) and 1 x gene expression product target (Applied Bio-systems) in a final volume of 23 µL. The assayed genes and assay identifications were CD44 (Hs00153304_m1), TIMP1 (Hs00171558_m1), VEGFA (Hs00900055_m1), TNFR1 (Hs01042313_m1), and TNFR2 (Hs00153550_m1). These genes were chosen arbitrarily in order to validate the microarray measurements of genes showing larger and smaller fold changes. Measurements were performed in duplicates. Amplification of 18S ribosomal RNA was used as an endogenous control to standardize the amount of total RNA in each sample. The differential expression was tested by the Welch t-test, and values of $P < 0.05$ were considered significant.

4.2 RESULTS

We compared the transcription profiles of 11 ruptured and 8 unruptured sIA walls resected during microsurgery from Finnish patients at median ages of 58 and 56 years, respectively. We were able to screen the expression of 17 788 distinct genes. In the ruptured sIA walls, 686 genes were significantly upregulated and 740 were downregulated compared with the unruptured sIA walls (See supplemental Table 1 of the original article). Five upregulated genes (CD44, TIMP1, VEGFA, TNFRS1A, TNFRS1B) were studied by quantitative polymerase chain reaction (Figure 9). All 5 were consistently upregulated but TNFRS1A not significantly.

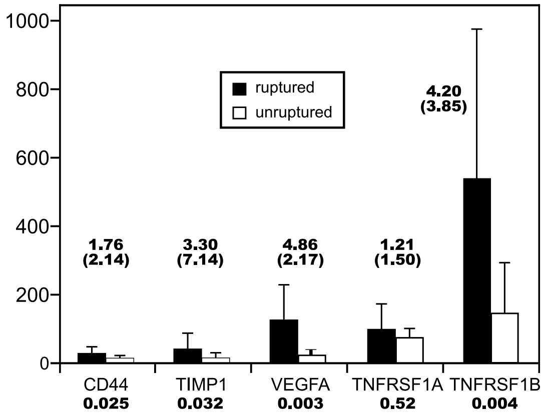


Figure 9. Comparison of expression of 5 selected genes in 16 ruptured and 7 unruptured saccular intracranial aneurysm (sIA) wall samples by quantitative real-time polymerase chain reaction (RT-PCR). Gene expression ratios in the RT-PCR study and in the microarray study (in parentheses; 11 ruptured and 8 unruptured sIA wall samples) are presented. The y-axis indicates in arbitrary units the expression level of each gene normalized by ribosomal RNA expression levels. SDs are shown on top of each bar. P values are indicated below the gene names.

In the ruptured sIA wall group, the time elapsed from rupture to resection of the sample did not seem to affect the gene expression levels. There were no statistically significant differences in the gene expression levels between the early and delayed sample groups. The elapsed time correlated significantly only to the expression of GREM1, not in the differentially expressed gene set, and the samples did not cluster according to the elapsed time in multidimensional scaling (Figure 10).

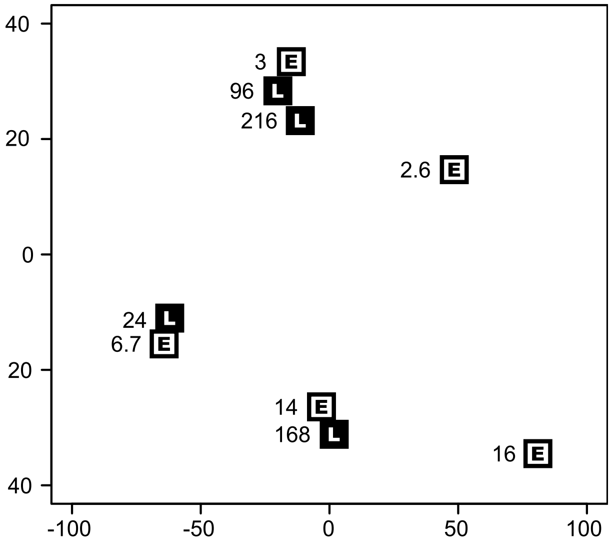


Figure 10. Ten ruptured saccular intracranial aneurysm (sIA) wall samples arranged according to the expression levels of all 17 788 genes studied (see Materials and Methods). The 5 samples resected < 24 hours after the rupture are denoted by E (early), and the 5 samples resected >= 24 hours are denoted by L (late). The time elapsed in hours from rupture to the resection is shown in front of the rectangles. The axes are in arbitrary units.

Among the upregulated genes in the ruptured sIA walls, the significantly enriched pathways were cytokine-receptor interaction, toll-like receptor (TLR) signaling, hematopoietic cell lineage, and leukocyte transendothelial migration (Table 6). The most interesting (based on previous literature and partially conflicting previous microarray studies) and significantly enriched ontologies were related to the immune system, and to the chemotaxis of cells, specifically including neutrophil chemotaxis. Of the cellular compartment ontologies, the Arp2/3 protein complex and the NADPH oxidase complex were enriched. The upregulated gene set of the ruptured sIA wall was significantly associated with the following disease classes in Genetic Association Database: immune; infection; cardiovascular, including atherosclerosis; and renal (see Table 3 of the original article).

Among the genes significantly downregulated in the ruptured sIA walls, there were no significantly enriched pathways after multiple testing correction. However, there were significantly enriched gene ontologies, revealing strong enrichment of zinc finger proteins of transcription factor activity (data not shown) and genes of tight junction and adherens junction (Table 6).

We performed in silico prediction of conserved transcription factor binding sites in the promoter regions of the differentially expressed gene sets. There were 6 and 58 enriched transcription factor binding sites in the upregulated (Table 7) and downregulated (see Supplemental Table 2 of the original publication) gene sets, respectively. The transcription factors enriched among the upregulated genes consisted of several members of the ETS family of transcription factors, nuclear factor-kB (NF-kB) p65 subunit, and hypoxia inducible factor-1A (HIF1A). The factors enriched in the downregulated gene set were diverse but contained many transcription factors of the SOX family.

Table 6. Biological processes differentially expressed in ruptured vs. unruptured sIA wall samples.

UPREGULATED GENES

Gene Ontology (GO) biological processes	**GO ID***	***P* value****	**FDR *****	**OR**	**Count *******	**Size*********
chemotaxis	GO:0006935	2.60E-14	5.66E-11	7	30	125
immune response	GO:0006955	5.70E-14	6.26E-11	4	49	333
response to external stimulus	GO:0009605	1.00E-13	7.41E-11	3	74	651
inflammatory response	GO:0006954	1.40E-13	7.52E-11	4.2	46	296
locomotory behavior	GO:0007626	1.50E-11	6.46E-09	4.5	35	208
response to stress	GO:0006950	3.60E-09	1.31E-06	2.1	99	1224
response to other organism	GO:0051707	4.70E-08	1.49E-05	5.7	18	87
positive regulation of tumor necrosis factor production	GO:0032760	5.80E-08	1.60E-05	127	6	7
locomotion	GO:0040011	1.30E-07	3.18E-05	4.7	20	114
cytokine production	GO:0001816	2.50E-05	5.49E-03	3.8	16	108
phosphate metabolic process	GO:0006796	5.30E-05	1.06E-02	1.8	66	893
positive regulation of interleukin-6 production	GO:0032755	5.90E-05	1.08E-02	42	4	6
regulation of cell proliferation	GO:0042127	1.80E-04	3.10E-02	1.9	44	550
intracellular lipid transport	GO:0032365	2.60E-04	3.55E-02	21	4	8
neutrophil chemotaxis	GO:0030593	2.70E-04	3.55E-02	12	5	14
regulated secretory pathway	GO:0045055	2.70E-04	3.55E-02	12	5	14
Protein amino acid phosphorylation	GO:0006468	2.70E-04	3.55E-02	1.8	47	613
regulation of cytokine biosynthetic process	GO:0042035	3.90E-04	4.55E-02	4.2	10	61

Purine ribonucleoside monophosphate biosynthetic process	GO:0009168	3.90E-04	4.55E-02	11	5	15
KEGG biological process	**KEGG ID**	**P value**	**FDR**	**OR**	**Count**	**Size**
Cytokine-cytokine receptor interaction	4060	8.50E-06	7.82E-04	2.7	31	245
toll-like receptor signaling pathway	4620	1.10E-05	7.82E-04	4	17	94
hematopoietic cell lineage	4640	1.70E-05	8.10E-04	4.3	15	78
epithelial cell signaling in Helicobacter pylori infection	5120	5.80E-05	2.03E-03	4.3	13	67
fructose and mannose metabolism	51	3.30E-04	8.81E-03	5.6	8	33
leukocyte transendothelial migration	4670	3.80E-04	8.81E-03	3	16	112
Gene Ontology cellular compartment	**GO ID**	**P value**	**FDR**	**OR**	**Count**	**Size**
membrane	GO:0016020	2.50E-07	5.30E-05	1.5	325	6028
vacuole	GO:0005773	3.30E-07	5.30E-05	3.5	26	193
Arp2/3 protein complex	GO:0005885	9.30E-07	9.80E-05	110	5	6
cytoplasm	GO:0005737	5.10E-05	3.80E-03	1.4	329	6428
integral to plasma membrane	GO:0005887	6.00E-05	3.80E-03	1.7	72	1053
NADPH oxidase complex	GO:0043020	2.20E-04	1.18E-02	22	4	8
extracellular space	GO:0005615	3.20E-04	1.43E-02	1.9	39	497
membrane raft	GO:0045121	5.60E-04	2.23E-02	3.9	10	66
lysosome	GO:0005764	8.10E-04	2.85E-02	3.1	13	108
cytosol	GO:0005829	1.30E-03	4.09E-02	1.6	55	827
Proton-transporting V-type ATPase, V0 domain	GO:0033179	1.50E-03	4.40E-02	22	3	6

DOWNREGULATED GENES						
Gene Ontology						
cellular compartment	**GO ID**	**P value**	**FDR**	**OR**	**Count**	**Size**
nucleus	GO:0005634	1.40E-05	4.81E-03	1.5	235	4487
intracellular part	GO:0044424	5.90E-05	1.04E-02	1.8	79	1487
costamere	GO:0043034	9.60E-05	1.11E-02	31	4	7
adherens junction	GO:0005912	5.90E-04	4.36E-02	3.6	11	82
tight junction	GO:0005923	6.30E-04	4.36E-02	5.6	7	36

* identification code; ** p-value uncorrected for multiple testing; *** false discovery rate, p-value after multiple testing correction (see Materials and Methods); **** number of differentially expressed genes in each biological category; ***** total number of genes assayed in the present study in each category.

Table 7. Enriched transcription factor binding sites in the -5 kb promoter regions of the 686 upregulated genes.

Transcription factor					
TRANSFAC® *	**HGNC ****	**Number of binding sites**	**Number of binding sites in human genome**	**P value*****	**Family**
ELF1	ELF1	411	11549	1.66E-08	ETS
PEA3	ETV4	714	21887	1.66E-07	ETS
ETS2	ETS2	413	11968	3.64E-07	ETS
ETS1	ETS1	728	22501	3.87E-07	ETS
HIF1	HIF1A	304	8663	3.13E-06	-
NFKAPPAB65	RELA	96	2232	7.99E-06	Rel/ankyrin

Transcription factor name at *TRANSFAC® (www.gene-regulation.com) and ** HUGO Gene Nomenclature Committee gene symbol; *** nominal p-value.

4.3 DISCUSSION

In the present study, we compared the transcriptomes of 11 ruptured and 8 unruptured human sIA walls to identify pathways that are associated with the rupture and to computationally predict transcriptional control of those pathways. The processes identified in the ruptured sIA walls were response to turbulent blood flow, chemotaxis, leukocyte migration, oxidative stress, vascular remodeling, and extracellular matrix degradation (See *Table 4* and discussion below). Signaling pathway and transcription factor analyses suggested that TLR signaling and regulation by NF-kB, HIF1A, and ETS transcription factors have key roles in the processes active in the ruptured sIA walls. Our results may provide clues to the molecular mechanism in sIA wall rupture and insight for novel therapeutic strategies to prevent rupture.

Previous whole-genome gene expression profiling studies comparing ruptured and unruptured sIAs have only identified low number of differentially expressed genes (see chapter 2.4.5 for a review). In contrast, we identified 686 significantly upregulated and 740 significantly downregulated genes in the ruptured sIA walls. The larger number of differentially expressed genes is most likely due to increased sample size combined with different statistical analyses and up-to-date custom annotations for microarray oligonucleotide probes.

In our study, some of the differential gene expressions may be caused by reaction of the sIA wall to the rupture. However, different times from the rupture to the resection of the sIA wall samples did not seem to have a significant effect on the differential gene expression. This is in line with the comparison of 44 ruptured and 27 unruptured walls by Kataoka et al (Kataoka et al. 1999) in which the time from rupture to resection of the aneurysm samples did not correlate to the scores of histological inflammation and aneurysm wall fragility. Frösen et al (Frösen et al. 2004) studied 42 ruptured and 24 unruptured sIA walls, and their comparison of leukocyte density and time from rupture to sample resection suggested that leukocytes could be present in the sIA wall already before the rupture. Another limitation in our differential gene expression profiling is that the sIA wall samples contain a mixture of cell types, including endothelial cells, smooth muscle cells, fibroblasts, and leukocytes. Consequently, it is difficult to pinpoint which cell populations are responsible for the overall differential profile.

Turbulent flow and low shear stress may cause inflammation, leukocyte migration, and oxidative stress at arterial bifurcations (Chiu et al. 2009), the site of sIAs as well. In our study, many genes and processes upregulated in turbulent and low-blood-flow conditions were enhanced, including TLR signaling (Dunzendorfer et al. 2004), CTSL1 (Platt et al. 2006), VEGF, NADPH complex, NF-kB signaling, IL8, CXCR4, PTX3, TNFRSF21, PHLDA1, and ICAM1. (Chiu et al. 2009) Dai et al. (Dai et al. 2004) detected 72 upregulated genes in endothelial cells under turbulent blood flow compared with laminar flow, and 13 of those genes were significantly upregulated in our ruptured sIA walls. The probability of finding as many or more overlapping genes just by chance is 1.38E-5 as assessed by Fisher's exact test.

Inflammation is associated with atherosclerosis and many cardiovascular diseases (Sprague & Khalil 2009), as well as experimental cerebral aneurysm formation (Aoki, Kataoka, et al. 2009). Our transcriptome data suggest that inflammatory processes are strongly associated with the rupture of the human sIA wall. Increased expression was observed in inflammatory pathways such as the cytokine-cytokine receptor interaction, TLR signaling, leukocyte transendothelial migration (Table 6), NF-kB signaling (Table 7) and many other inflammation-related gene ontology categories (Table 6).

Increased expression of the macrophage marker CD163 and the neutrophil marker FCGR3B (Bux 2008) was observed, as well as enrichment of the neutrophil chemotaxis gene ontology category. Here, we provide the first indication that neutrophils may have a role in the rupture of human sIA walls. Prolonged neutrophil presence can damage healthy tissues (McGettrick & O'Neill 2007) (Figure 11). FCGR3B is possibly involved in small vessel vasculitis associated with antineutrophil cytoplasmic antibodies (Kettritz 2012), it is required for neutrophil mediated tissue damage in bullous pemphigoid disease (Yu et al. 2010), and neutrophil depletion decreased the progression of experimental abdominal aortic aneurysms with a mechanism that so far is unknown (Eliason et al. 2005). Importantly, FCGR3B mediated reactive oxidant production can be selectively suppressed (Fossati et al. 2002), suggesting a potential future approach to prevent sIA wall rupture.

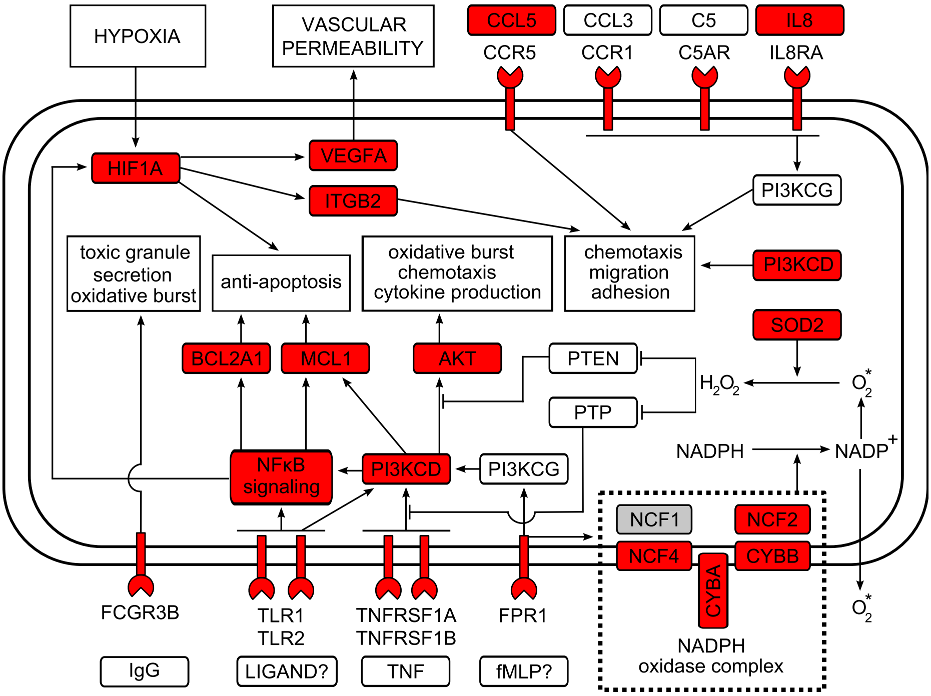


Figure 11. Upregulated genes in selected cellular signaling pathways in a neutrophil granulocyte, hypothesized here to infiltrate a ruptured saccular intracranial aneurysm wall. Neutrophil signaling pathways are adapted from the following references (Reichel et al. 2006; Nizet & Johnson 2009; Rommel et al. 2007; Tanaka et al. 2003). Red boxes indicate proteins of the upregulated genes; white boxes indicate proteins of the nondifferentially expressed genes. NCF1 gene in the grey box was not included in the array. Many of the signaling pathways are also used by macrophages. NADP+, nicotinamide adenine dinucleotide phosphate; NADPH, reduced NADP+.

Subunits of NADPH oxidase complex (NCF1, NCF2, CYBB, CYBA; Table 6), producers of reactive oxygen species, and a number of oxidative stress response genes (ERCC1, PTGS1, CCL5, SOD2, SRXN1, HMOX1, UCP2, PNKP (Ashburner et al. 2000)) were upregulated in the ruptured sIA walls. Aoki et al (Aoki, Nishimura, Kataoka, et al. 2009) found that NCF1 was upregulated in sIA walls and that NCF1 knockout decreased cerebral aneurysm formation in a mouse neurysm model. Increased oxidative stress and upregulated NADPH complex play a role in, for example, coronary artery disease and atherosclerosis, and reactive oxygen species production is increased at coronary artery branching sites (Chiu et al. 2009; Guzik et al. 2006). Tight junction and adherens junction genes were downregulated in the ruptured sIA walls (Table 6), suggesting loosening of contact between endothelial cells and smooth muscle cells. Elastin- and collagen-degrading enzymes (cathepsins A, L1, S, B, C), matrix metalloproteinases (MMP9, MMP19), heparan sulfate proteoglycan degrading enzyme heparanase (HPSE), and plasminogen activating receptor (PLAUR) were highly upregulated, whereas 3 collagen genes (COL4A5, COL21A1, COL14A1) were strongly down-regulated (see Supplemental table 1 of the original publication). These data suggest that ECM degradation predisposes or follows the sIA wall rupture or both. Extracellular matrix degradation is central in many arterial wall diseases (Lutgens et al. 2007; Raffetto & Khalil 2008) and matrix metalloproteinase mediated vascular remodeling likely promotes intracranial aneurysm formation in rats (Aoki et al. 2007). Overexpression of HPSE in mouse endothelium decreased aortic stiffness and strength and increased the incidence of spontaneous aneurysms (Baker et al. 2009).

The canonical TLR pathway ends in the activation of NF-kB. In our study, TLR signaling was significantly enriched, and genes responsive to TLR to NF-kB signaling (IL6, MMP9, CCL5) (Lee et al. 2009; O'Neill et al. 2009) were upregulated in the ruptured sIA walls. In our in silico prediction, the binding site for the p65 transcription factor, a subunit of NF-kB, was significantly enriched among the promoter regions of the upregulated gene set. TLR activation contributes to the development and progression of atherosclerosis, cardiac dysfunction in sepsis, and congestive heart failure (Frantz et al. 2007). TLR4 and TLR10 are upregulated during the formation of cerebral aneurysms in rats, and TLR4 was also detected in human cerebral aneurysms (Aoki, Nishimura, Ishibashi, et al. 2009).

We observed the enrichment of 6 conserved transcription factor binding sites among the upregulated genes: 4 from the ETS family (ELF1, ETV4, ETS2, ETS1), HIF1A, and an NF-kB subunit (RELA) (Table 7). ETS transcription factors are involved in vascular inflammation and remodeling (Ni et al. 2007; Oettgen 2006; Zhan et al. 2005). HIF1A transcriptional activity is increased in hypoxic conditions, and the activation of HIF1A has been reported in several inflammatory conditions also in normoxic conditions e.g. in atherosclerosis (Vink et al. 2007). NF-kB is a key player in the induction of IAs in rats (Tomohiro Aoki et al. 2007), and the induction of abdominal aortic aneurysms in rabbits was regressed by dual inhibition of NF-kB and ETS1 (Miyake et al. 2007).

4.4 CONCLUSIONS

Transcriptome comparison of 11 ruptured and 8 unruptured human sIA walls indicated that response to turbulent blood flow, chemotaxis, leukocyte migration, oxidative stress, vascular remodeling, and extracellular matrix degradation were active in the ruptured sIA walls. Signaling pathway analysis and computational transcription factor analysis suggested that TLR signaling and regulation by NF-kB, HIF1A, and ETS transcription factors have a key role in processes active in the ruptured sIA walls. Further analyses are required to distinguish between inflammatory reactions that predispose the sIA wall to rupture and the immediate responses to the rupture and subsequent closure of the ruptured site by an acute thrombus and to confirm the mRNA level gene expression patterns at the protein level.

5 *TAFFEL: Independent Enrichment Analysis of gene sets*[1]

5.1 INTRODUCTION

Gene expression studies often compare samples from two or more experimental conditions, the most typical outcome being a set of genes that differ in expression between the conditions. Several databases, computational methods and software programs have been recently published for analysis of such differentially expressed (DE) gene sets. Usually these tools are aimed at finding out associated (differentially active) biological mechanisms by searching for associations of DE genes to various biological functions, processes and pathways reported in the biological databases such as Gene Ontology (GO) (Ashburner et al. 2000). The output of these tools is usually a list of biological terms (functions, processes, pathways etc.) that are more frequently associated to the gene set than expected by chance. Therefore, this analysis is often referred to as overrepresentation analysis (ORA) (for a review see chapter 2.5.2).

Standard EA has some shortcomings that should be taken into account, especially in the case of DE genes. DE genes are often, dependent on the studied condition, likely to be associated to multiple distinct biological pathways rather than one or a few, consequently the possible large list of genes might prevent the identification of more subtly perturbed processes driven by fewer genes. This problem has been addressed by applying various clustering methods for finding gene subgroups with homogeneous functional annotations (Pehkonen et al. 2005; Dennis Jr et al. 2003; Martin et al. 2004) and combining similar functional annotations together (Martin et al. 2004). Clustering can reveal interesting gene subgroups, but so far, there are no definitive methods available to verify them or obtain further interpretation about their biological significance in the studied cases, other than calculating the internal homogeneity of clusters.

Here we present a novel method Independent Enrichment Analysis (IEA) and its implementation in a software tool called TAFFEL. The principal idea of IEA and TAFFEL is to facilitate the discovery of relevant biological phenomena from subsets of a set of differentially expressed genes and potential mechanisms of the regulation of those processes. The developed application allows quick and easy explorative analysis of data by performing three main steps (Figure 12). First, TAFFEL uses functional annotations from Gene Ontology (Ashburner et al. 2000) to separate differentially expressed genes into functionally homogenous gene groups. This facilitates the discovery of multiple biological phenomena associated to DE genes. Secondly, TAFFEL discovers groups of genes with similar *cis*-regulatory transcription factor binding sites (TFBSs) in their regulatory regions, using annotations of TFBS to specific transcription factors (TF) from the cisRED database (Robertson et al. 2006). This enables the identification of putatively co-regulated genes from the gene list and identification of their regulators. At this point, the analyst has several groups of genes that are homogenous in either GO or TF annotations. Therefore, as a third step TAFFEL includes a novel method referred to as Independent Enrichment Analysis (IEA) which evaluates the enrichment of TFs in gene clusters homogeneous in GO terms, and vice versa, enrichment of GO terms in gene clusters homogeneous in TF annotations. IEA provides clues to the regulatory control of genes sharing common functions. Simultaneously, it serves as an extrinsic biological validation of the obtained gene groups that can be used to point out the most interesting gene clusters among several. A detailed description of typical analysis flow with TAFFEL is provided in Methods and illustrated in Figure 12.

In order to demonstrate the utility of our method and the associated software, we applied TAFFEL to two datasets. Firstly, we analyzed differentially expressed genes in human HEK293T cell culture after treatment with forskolin, a cyclic AMP (cAMP) pathway inducer. A researcher, not involved in the method or software development, independently performed this analysis. This analysis served two main purposes: 1) as a positive control to see if our methodology is able to identify the involvement of the known cyclic AMP response element (CREB) in regulation of cAMP responsice genes; and 2) as a way to assess if a researcher knowledgeable in the biology of a studied phenomena (cAMP signalling) but without extensive bioinformatic knowledge is able to use the software and produce new insights.

Secondly, we re-analyzed differentially expressed genes between human ruptured and unruptured saccular intracranial artery aneurysm (sIA) walls of study I. This dataset was re-analyzed using TAFFEL in order to demonstrate the capability of TAFFEL to find novel phenomena overlooked in standard analysis and to identify factors that might be causing the reported phenomena. The results suggest hypotheses of novel molecular mechanisms in ruptured sIA walls and demonstrate the usefulness of TAFFEL in typical snapshot type research settings and in diseases of poorly characterized molecular pathogenesis.

We compared TAFFEL gene clustering results against results from five other methods or tools used for enrichment analysis: standard list of GO-terms sorted according to Fisher's Exact test p-values, a sorted list of GO-terms and transcription factors resulting from FatiGO+ tool (Al-Shahrour et al. 2004), annotation sets resulting from the Functional Annotation Clustering tool available in DAVID (Dennis Jr et al. 2003), co-occurring sets of GO-terms and transcription factors resulting from a priori association rule discovery algorithm implemented in GeneCodis (Carmona-Saez et al. 2007) and results from GSEA (Subramanian et al. 2005). The comparison shows that TAFFEL can discover important individual themes and relations between transcription factors and biological processes that are not reported at all by other methods.

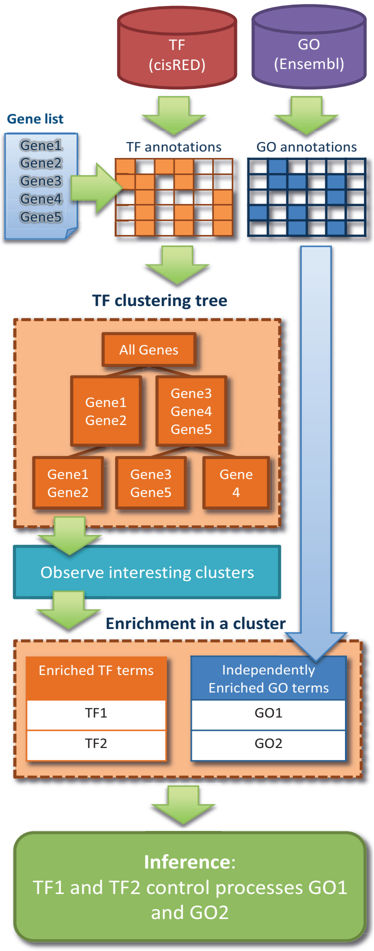


Figure 12. The flow diagram of TAFFEL analysis. From the top: the list of genes given by the user is annotated by GO and TF information from Ensembl (20) and cisRED (12) databases. The genes are clustered separately in parallel, based on GO and TF annotations (for simplicity only the TF clustering tree is shown). In each resulting cluster, the enrichment of both GO and TF annotations is determined, providing a basis for suggesting implications between the biological processes and their regulator molecules.

5.2 RESULTS

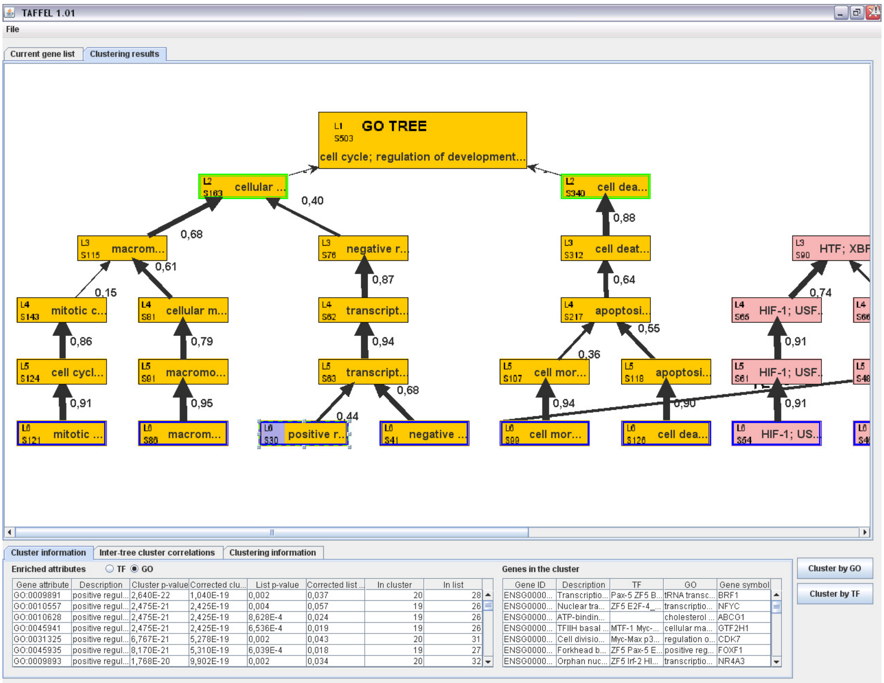
5.2.1 Description of the method and tool

TAFFEL uses a non-nested hierarchical clustering scheme (Pehkonen et al. 2005) for finding gene subgroups that are homogeneous in GO terms or TF annotations. The gene subgroups are a partition of the whole gene set i.e. they are disjoint sets that cover the whole gene set.

The clustering of genes is performed using only GO or TF data and no gene expression data is needed. The method creates multiple clustering solutions with different numbers of clusters and combines them into a single visualization. Each clustering solution is visualized as a set of horizontally ordered rectangles, each rectangle representing a single cluster (Figure 13). Different clustering solutions are ordered vertically according to the number of clusters. Thus, the visualization contains several levels, the top level representing the whole gene list as a single cluster, the second level representing clustering of genes into two clusters, the third level representing a solution with three clusters etc. The best correlating clusters between adjacent levels are combined with edges, creating a tree-like structure. Unlike regular hierarchical clustering, the different tree levels are independent of each other. This visualization can be used to track coherent clusters that stay similar in different solutions despite the changing number of clusters and initialization for clustering, and to observe the hierarchical relationships in the data. In addition, the tool performs automatic evaluation of clustering solutions with a different number of clusters using a statistical model selection (see Selection of number of clusters). The best scoring levels are highlighted in the visualization. The tree that is obtained using GO terms as data for clustering is referred to as a "GO tree" and the tree obtained using TF annotations is referred to as a "TF tree".

For each gene cluster, TAFFEL reports both the enriched GO terms and TF annotations, regardless of what information (GO or TF) was used for clustering. For the first level of the tree, representing the whole analyzed gene list, the enrichment is measured in the list versus the genome. This is analogous to the traditional enrichment analysis and can be used for observing the most interesting themes in general. This enrichment is also reported for the annotations in the clusters of subsequent tree levels (column "List p-value" in the software) as additional evidence of their biological significance. However, as a principal description for each cluster in the subsequent tree levels, TAFFEL reports annotations that are enriched in each cluster versus the whole gene list (column "Cluster p-value" in the software). This gives the user a compact overview of the different biological phenomena present in the analyzed list of genes.

In order to gain more evidence about the biological meaningfulness of resulting clusters, TAFFEL performs two types of extrinsic evaluation steps. Firstly, in the IEA evaluation, each functionally homogeneous gene cluster is evaluated in terms of enrichment of TFs, and each gene cluster homogeneous in TFs is evaluated in terms of enrichment in GO terms. Secondly, TAFFEL allows measuring correlations of gene memberships between all possible cluster pairs where one cluster comes from the GO tree and another from the TF tree. This measure, referred to as *inter-correlation*, can be used to identify the gene clusters that share the same genes regardless of using TF's or GO terms as a basis for clustering. Both the IEA and inter-correlation can be used for validating the biological significance of gene subgroups, and to interpret relations between transcription regulators and processes they regulate.



Enriched attributes ○ TF ● GO

Gene attribute	Description	Cluster p-value	Corrected clu...	List p-value	Corrected list ...	In cluster	In list
GO:0009891	positive regul...	2,640E-22	1,040E-19	0,002	0,037	20	28
GO:0010557	positive regul...	2,475E-21	2,425E-19	0,004	0,057	19	26
GO:0010628	positive regul...	2,475E-21	2,425E-19	8,628E-4	0,024	19	26
GO:0045941	positive regul...	2,475E-21	2,425E-19	6,536E-4	0,019	19	26
GO:0031325	positive regul...	6,767E-21	5,278E-19	0,002	0,043	20	31
GO:0045935	positive regul...	8,170E-21	5,310E-19	6,039E-4	0,018	19	27
GO:0009893	positive regul...	1,768E-20	9,902E-19	0,002	0,034	20	32

Genes in the cluster

Gene ID	Description	TF	GO	Gene symbol
ENSG0000...	Transcriptio...	Pax-5 ZF5 B...	tRNA transc...	BRF1
ENSG0000...	Nuclear tra...	ZF5 E2F-4_...	transcriptio...	NFYC
ENSG0000...	ATP-bindin...		cholesterol ...	ABCG1
ENSG0000...	TFIIH basal ...	MTF-1 Myc-...	cellular ma...	GTF2H1
ENSG0000...	Cell divisio...	Myc-Max p3...	regulation o...	CDK7
ENSG0000...	Forkhead b...	ZF5 Pax-5 E...	positive reg...	FOXF1
ENSG0000...	Orphan nuc...	ZF5 Irf-2 Hl...	transcriptio...	NR4A3

Figure 13. TAFFEL user interface. The clustering trees represents the clustering result for the DE genes after 4 hours of forskolin treatment in HEK293T cells. The genes have been clustered by the GO terms (left) and TFs (right). The topmost box represents the whole gene set without clustering. Below that, each level represents clustering to two, three, or more clusters. The green outline indicates the cluster number selection by AIC and blue by dAIC. The clusters obtained from the IEA analysis with FDR p < 0.1 are highlighted with the light blue background on the right side of cluster box. The best intercorrelating clusters (cell morphogenesis cluster in the GO tree and COUP cluster in the TF tree) between the trees are connected with the bold line. Information at bottom shows enriched annotations (left list) and cluster genes (right list). Positive regulation of biosynthetic process related cluster is selected in the picture.

5.2.2 Availability and running the program

TAFFEL is a Java Web Start application written using Java Standard Edition 6 with NetBeans integrated development environment (www.netbeans.org). MySQL database (www.mysql.com) is used to store all the persistent data. Running TAFFEL requires Java Runtime Environment version 6. TAFFEL program, help-pages, and example data sets are freely available under LGPL license from http://kokki.uef.fi/bioinformatics/taffel/.

5.2.3 Typical analysis flow

A typical analysis flow with TAFFEL is shown in Figure 12. Firstly, the gene list is imported to TAFFEL and clustered using GO terms and TF annotations as data. Secondly, the root levels of the GO and TF trees are observed to study the themes associated to the whole gene list in general. Thirdly, the clusters at the tree levels with the smallest dAIC scores in both the GO and TF trees are observed in order to find which separate themes are associated to the analyzed gene list and which respective gene subgroups constitute it. Fourthly, the coherency of these clusters is evaluated by observing their conservation throughout the tree. Finally, special focus is set on the clusters in the selected levels by using IEA and inter-correlation methods for cluster evaluation. The independently enriched themes in each cluster can be used to infer the TFs that drive a particular biological process or function in the analyzed condition.

The resulting clusters can be further analyzed by multiple ways such as highlighting the clusters including particular GO terms or TFs, to find correlations between clusters in different trees, and to show the list of genes associated to specific GO terms and/or TF annotations in each cluster. The results can be exported from the program in text form, and all results can be saved in one XML file.

5.2.4 Analysis of forskolin effect on HEK293T cells

The application of TAFFEL and IEA to forskolin treated human HEK293T cells by cAMP signaling by an external domaine expert researcher indicated that the method and software are usable and useful also for non-bioinformatics researchers. CREB involvement was correctly identified but also other results were identified, unexpected to the analyst at first, but highly plausible backed by evidence from the literature. As the author of this thesis is not expert on cAMP signaling and it is out of the scope of the thesis, the interested reader is referred to Publication II for a more detailed report of the external analysts results.

5.2.5 IEA suggests novel hypotheses of signaling mechanisms active in the ruptured sIA wall

The 498 overexpressed and 491 underexpressed genes in the ruptured sIA wall group were input to TAFFEL and clustered separately first by GO and then by TFBS. Automatic cluster number selection identified 9 and 11 clusters in GO and TFBS clustering respectively. We focused then on the clusters with significant (corrected $p<0.05$) independent enrichment (IEA) or significant *inter-correlation*. Three statistically significant IEA clusters and one significant *inter-correlation* cluster were identified (Table 8).

The first cluster was identified among the upregulated gene set. Metal ion transport and protein kinase signaling were the main functions in the cluster and MTF-1 and ATF-1 transcription factor binding sites were independently enriched. MTF-1 is a pleiotropic metal, oxidative stress and hypoxia activated transcription factor controlling e.g. zinctransferring metallothioneins (MT) and other genes (Günther et al. 2012). The cluster is enriched in ion-transferring proteins and contains MT2A (a metallothionein), a primary target of MTF-1 (Saydam et al. 2002). Metallothioneins can control intracellular zinc availability (Kang 2006; Giacconi et al. 2008; Maret & Krężel 2007). MT activation and reduced zinc bioavailability has been observed to increase with aging and cardiovascular diseases in the elderly (Giacconi et al. 2008). Although MTF-1 is mainly vascular protective, chronic low grade inflammation can maintain long-term elevation of MTs, which in turn may lead to a pro-inflammatory response possibly due to decreased zinc bioavailability (Giacconi et al. 2008; Conway et al. 2010). This cluster contains also many other potassium and calcium transporters. Ion channels play a role in vascular tone regulation and many inflammatory conditions (Eisenhut & Wallace 2011; Hu & Zhang 2012; Das et al. 2010). One such gene in the cluster is *P2RX4*, a gene having a role in regulating large arterial tone in response to shear stress. A loss-of-function mutation in P2RX4 is associated with increased pulse pressure (Stokes et al. 2011).

Such possible coordinated regulation of ion transporters important for cardiovascular function suggests that the role of these transporters, metallothioneins, and metal ions in aneurysm rupture should be investigated in more focused studies to address the question of whether these changes are just a benign defence mechanisms or detrimental to sIA wall structure.

Another cluster found in IEA from the analysis of under expressed genes was related to oxidation-reduction and independently enriched the NF-1 (nuclear factor 1 C, NF1C) transcription factor (FDR corrected p=0.037). NF1C activation capability is repressed by oxidative stress and NFIC knockout decreases the activity of Cytochrome p450 family gene CYP1A1(Barouki & Morel 2001). The cluster contains 2 CYP-family genes, and many other lipid and amino acid metabolizing genes as well as genes protecting against or controlling oxidative stress (*NXN*, *OXR1*). NF1C activity is repressed by oxidative stress (Barouki & Morel 2001) and thus the down-regulation of the genes in this cluster might be caused by inactivation of NF1C by oxidative stress likely present in the ruptured aneurysms (Laaksamo et al. 2013).

The third cluster found in IEA was identified among the down-regulated genes. The cluster was enriched with cell development related GO terms, *cell motion*, *cell projection and organization,* and more specifically *blood vessel morphogenesis* and independently enriched Tal-1 transcription factor (FDR corrected p=0.031). Tal-1 protein is known to drive endothelial cell migration and morphogenesis in angiogenesis (Chetty et al. 1997; Lazrak et al. 2004). Tal-1 regulates VE-cadherin expression in endothelial cells. VE-cadherin concentrates on cell-to-cell adherens junctions and maintains cell adhesion, controls vascular permeability and relays signals necessary for vascular stabilization. VE-cadherin is a positive controller of TGF-β signalling and deletion of various components of this signalling pathway leads to several vascular manifestations, often including hemorrhages (Rudini et al. 2008).

In order to find out whether the clustering by GO terms and TF annotations would yield any clusters with common genes, the TAFFEL *inter-correlation* method was applied. The link between apoptosis and TF MEF2A and Lhx3a was strongly observed (FDR corrected p=7.5E-6). MEF2A is a myocyte enhancer factor, which controls many muscle-specific genes. A deletion of MEF2A causes autosomal dominant coronary artery disease (Wang et al. 2003) and knock-down of MEF2A enhances smooth muscle cell proliferation (Zhao et al. 2012). A low number of smooth muscle cells with disorganized architecture has been associated with aneurysm rupture (Frösen et al. 2004). Our results suggest that MEF2A might be involved in smooth muscle cell apoptosis or phenotype in the ruptured sIA walls.

Table 8. Statistically significant clusters in up-regulated (sIA↑) and down-regulated (sIA↓) genes in ruptured intracranial aneurysms. CLUSTER column indicates the clustered dataset, annotations used for clustering (either GO or TF) and the size of the cluster, respectively. ANNOTATION column indicates enriched GO terms and TF annotations from TRANSFAC in each cluster. P and P LIST columns indicate Benjamini-Hochberg FDR corrected Fisher's exact test p-values for the enrichment of the annotation in the cluster and in the gene list, respectively. N and N LIST columns show the number of genes associated with the annotation in the cluster and in the gene list.

CLUSTER			ANNOTATION	P	P LIST	N	N LIST
sIA ↑	GO	58	cation transport	3.8E-08	1.8E-01	17	23
			ion transport	3.8E-08	3.7E-01	18	26
			metal ion transport	1.2E-05	4.2E-01	12	16
			…				
			protein amino acid phosphorylation	4.5E-4	9.6E-03	16	35
			regulation of protein kinase activity	1.0E-02	1.5E-01	7	12
			G-protein signaling, coupled to IP3 second messenger (phospholipase C activating)	2.9E-02	1.6E-01	4	5
			regulation of Ras GTPase activity	7E-02	1.5E-01	4	6
			MTF-1	4.8E-02	4.7E-01	13	30
			ATF-1	4.8E-02	6.5E-01	9	17
sIA ↓	GO	22	nervous system development	1.3E-11	1.3E-02	17	32
			generation of neurons	2.1E-06	2.8E-01	8	10
			cell development	2.1E-06	2.8E-01	10	17
			…				
			blood vessel morphogenesis	3.5E-04	2.8E-01	5	6
			cell migration	4.3E-04	7.2E-01	4	4
			Tal-1	3.1E-02	1.0E+00	4	5
			AR	9.4E-02	9.8E-01	6	17
sIA ↓	GO	49	organic acid metabolic process	1.3E-08	2.8E-01	13	13
			carboxylic acid metabolic process	1.3E-08	2.8E-01	13	13
			oxidation reduction	1.2E-07	2.8E-01	14	16

...				
lipid metabolic process	5.3E-07	4.4E-01	13	16
carbohydrate metabolic process	2.8E-3	5.5E-01	7	9
amino acid metabolic process	3.2E-03	5.2E-01	5	5
NF-1	3.7E-02	9.8E-01	14	30

5.2.6 Comparison of TAFFEL to other methods

Several different approaches for analyzing differentially expressed gene sets exists, such as GENERATOR (Pehkonen et al. 2005), DAVID (Dennis Jr et al. 2003), FatiGO (Al-Shahrour et al. 2004), GOToolBox (Martin et al. 2004), GenMAPP (Dahlquist et al. 2002), GoMiner (Zeeberg et al. 2003), OntoTools (Draghici et al. 2003), and GSEA (Subramanian et al. 2005), which can report the enriched terms e.g. the functional annotations, or TF information but no relation between these concepts. GeneCodis (Carmona-Saez et al. 2007) is aimed at addressing partly the same concerns as TAFFEL by seeking relations between different annotation systems within a set of genes but without considering subsets of the genes (i.e. clustering).

The main advancement of TAFFEL is that the developed IEA method, which allows statistically interpretable evaluations for the found clusters, helps to draw attention to the most interesting gene clusters among many, and provides information about the control of regulator proteins in functionally homogeneous gene subgroups.

We compared the IEA method to several other pathway analysis methods (See publication II for details) and showed that our method can identify additional phenomena from differentially expressed gene sets, which are not reported by other methods. IEA is not, however, a replacement for more standard overrepresentation analyses but a complementary way to aid in gaining additional insight.

5.3 DISCUSSION

We presented a novel method for the analysis of differentially expressed (DE) genes for the discovery of co-functional and co-regulated subsets of genes, and for further analysis of such clusters with functional annotations and regulatory protein information. As information about gene regulatory elements, we have used TF predictions and annotations from cisRED database where putative binding sites are validated in terms of evolutionary conservation (Robertson et al. 2006). Such validation has shown to be advantageous as it can significantly reduce the amount of false positives in predictions (Kankainen et al. 2006;

Ho Sui et al. 2005). Moreover, our clustering of genes by TF/GO and validation of discovered clusters using functional annotations not used in the clustering process should reveal relevant patterns from the data and reduce the amount of noise.

A major limitation in our and many other methods employing GO and TF data is that the knowledge on gene functions (the GO annotations) (Rhee et al. 2008) and regulation (TFs) is incomplete. Furthermore, the GO annotations are biased towards well-studied biological phenomena and the predicted TF binding sites (cisRED) often contain a large number of false positives (Hannenhalli 2008). Still the clustering method alleviates this problem in the sense that the clustering is not driven by randomly distributed annotations (false positives or negatives) but by stable annotations shared by many genes. The constantly improving quality of the annotations is also likely to improve the results obtained using our method. It should also be noted that gene expression is not necessarily functional in the sense that co-expressed or similarly expressed genes do not necessarily share any GO annotations. Thus our clustering approach does not necessarily produce clusters of co-expressed genes, which likely results in fewer significant IEA clusters. Also the used AIC method for cluster number selection is not necessarily optimal, but rather it strikes a good balance between accuracy and number of parameters. The cluster number selection is a very general problem and usually there is no single best solution for every dataset (see for example reviews (Halkidi et al. 2002a; Halkidi et al. 2002b). In our method we use cluster number selection as a guide for the analyst to focus on some particular clustering level to start the analysis with.

The result of TAFFEL analysis for the DE genes after forskolin treatment of human HEK293T cells in culture showed the expected results at the first level of the clustering tree, e.g., the enrichment of cAMP related GO terms and CREB TF. The independent non-bioinformatician domain expert could additionally identify a piece of a complex MAPK-AP1-AhR related transcription network, related to proliferation and regulation of metabolism. As CREB signalling is out of focus of this thesis and outside the area of expertise of the author, the reader interested in these discussions is referred to the original Publication II.

In the analysis of over and under-expressed genes in the ruptured saccular intracranial aneurysm (sIA) walls TAFFEL identified several interesting clusters. TAFFEL suggested signalling relating *TAL1* to cell development and blood vessel morphogenesis, *MTF1* to metal ion transport activity, *NF1C* to oxidation reduction and lipid metabolism and MEF2A to apoptosis. These processes might be related to processess detrimental to the sIA wall integrity but they could also be vascular wall reactions to rupture. Further more focused studies are needed to address this issue.

5.4 CONCLUSIONS

In conclusion, we have demonstrated that the developed method and TAFFEL tool is usable by a non-bioinformatician and can give new insight into the analysis of differentially expressed genes. Our comparison to other popular methods showed that the IEA method implemented in TAFFEL can generate novel hypotheses of biological phenomena, which are not reported by other methods at all. The downside of the method, as is typical of many computational methods, is that the generated hypotheses should be validated in further functional studies .IEA is however not meant to be a replacement for or claimed superior to more standard overrepresentation analyses but a complementary way to aid in gaining additional insight into the studied phenomenon.

Firstly, the analysis of forskolin-treated HEK293T cells indicates that TAFFEL will identify well-known and expected phenomena such as differential expression of CREB regulated genes, but can also lead to new hypotheses. Secondly, the results with the sIA wall rupture related data give confidence to the usefulness of TAFFEL in the analysis of complex and poorly characterized clinical conditions, affected by inherited and acquired risk factors. These findings suggest that TAFFEL is an efficient method to generate new hypotheses to be further tested in more focused studies.

5.5 MATERIALS AND METHODS

5.5.1 Annotation data sources

For the functional grouping of genes, TAFFEL uses Gene Ontology (Ashburner et al. 2000) annotations (December 2008 release used in this study) from Ensembl database (Flicek et al. 2008) (version 53 used in this study). The included species are human, mouse, rat and *C. elegans*. The current version of TAFFEL can use biological process and molecular function ontologies from GO, either separately or in parallel.

Secondly, TAFFEL uses information about predicted TFBSs available in the public cisRED database (Robertson et al. 2006), containing genome wide collections of sequence motifs conserved in gene regulatory regions. The motifs have been annotated by transcription factors (TFs) found in TRANSFAC (Matys 2003) and JASPAR (Bryne et al. 2008) databases. In TAFFEL, we have included all TF annotations from both of these databases that have similarity p-value < 0.001 with the found sequence motif. We have included data for human (version 9), mouse (version 4) and *C. elegans* (version 4).

5.5.2 Gene clustering method

In order to perform gene clustering, associations between genes and annotations (GO terms and TFs) are represented as a binary matrix. Each row in the matrix represents a gene and each column represents an annotation. In the matrix, the cell value one indicates association and zero indicates no association between the row (gene) and the column (annotation). For clustering, we apply a Non-negative matrix factorization (NMF) (Lee & Seung 1999) based approach. This approach has been advantageous in clustering of sparse binary data and finding clusters that are defined in a (possibly small) subset of all data attributes (Seppänen et al. 2003). Both of these features are important in our cases described here. Firstly, the data are sparse by nature. Secondly, one set of genes often associates to numerous biological attributes (TFs and GO terms), many of which may not be relevant (Pehkonen et al. 2005).

5.5.3 Selection of number of clusters

In order to choose a clustering solution with a suitable balance between goodness of fit in the data and complexity, TAFFEL uses Akaike Information Criterion (AIC) (Akaike 1974) for statistical model selection. AIC is calculated by taking the number of parameters of the statistical model representing the evaluated clustering solution and subtracting them from the maximized log-likelihood of the data for the same model. Due to simplicity and robustness of the method, it has been widely used in similar clustering applications (see e.g. (Chen & Murphy 2005; Liu et al. 2009; Huang et al. 2003)).

As the abundance of dimensions (GO terms or TFs) in the gene annotation data are distributed randomly in resulting clusters, the clusters tend to exist in a relatively small subset of all dimensions (Pehkonen et al. 2005). Besides being problematic for clustering, this behaviour is also problematic for model selection. The model selection tends to be overwhelmed by such dimension and systematically favour a result with only one or a few clusters with different data sets. Thus, we also calculated a modified AIC, referred to as dAIC, for which we used only the dimensions that are distributed in a non-random fashion in at least one of the clustering solutions with >2 clusters in the whole TAFFEL tree. This was tested by comparing the AIC score of the dimension in the whole gene list versus the AIC score in each clustering solution. If the AIC score is better (smaller) in any of the clustering solutions, then the evaluated dimension was included in the calculation of dAIC. The same set of dimensions was then used for calculating dAIC for different clustering solutions including the whole gene list as one cluster. This feature selection filtered out at least 50% of the GO terms in our forskolin and sIA datasets (see *Results* section for detailed description of the datasets). When the remaining dimensions were used for calculating AIC score, the number of selected clusters was systematically higher than when using all dimensions.

5.5.4 Cluster statistics

The statistical testing of enrichment in TAFFEL is calculated using Fisher's exact test. Only annotations with occurrences in a cluster are used in the testing. The resulting p-values are corrected for multiple testing using Benjamini-Hochberg False Discovery Rate (FDR) (Benjamini & Hochberg 1995).

The interpretation of p-values reported by TAFFEL warrants a special note. In each cluster, enrichment is analyzed for the annotations of the same (Dependent Enrichment Analysis, DEA) and different (IEA) annotation system that was used in clustering. The p-values resulting from IEA have reasonable statistical interpretation as they test null hypotheses such as: "*TF x is not dependent of the gene group y homogeneous in GO terms*". Due to statistical independence between variables *x* and *y*, these p-values can be used reasonably to detect their biological significance and dependence of each cluster. As an opposite, the p-values from DEA would test null hypotheses such as: "*GO term x is not dependent of gene group y homogeneous in GO terms*". Here, variable *y* is statistically dependent on *x* and thus treating the resulting values as standard p-values for statistical decision-making would lead to circular argumentation. Still, these values from DEA are suitable as relative enrichment scores representing the most characteristic annotations in each cluster.

The inter-correlation measurements are also calculated using Fisher's exact test with Benjamini-Hochberg correction. As dependencies exist among the clusters between different inter-correlation comparisons, the correction tends to be highly conservative for this situation and should be interpreted with care.

The correlation between each cluster pair between the adjacent clustering solutions in the same clustering tree is calculated using standard correlation between two binomial distributions representing the gene memberships in the clusters.

5.5.5 Processing of demonstration microarray data sets

Gene expression microarray data (GSE2060 Affymetrix Human Genome U133A Array) concerning the effect of forskolin in human HEK293T culture was downloaded from Gene Expression Omnibus (GEO) and normalized using the RMA method. Forskolin-treated and control HEK293T cells (both in duplicates) in culture were compared at 4 hours to find out differentially expressed genes. Welch's t-test with Benjamini-Hochberg correction was used. Due to a low number of replicates, the fold change was used as an additional measure for filtering. P-value < 0.05 and fold change > 1.25 resulted in 691 differentially expressed genes.

Whole genome expression data of 11 ruptured and 8 unruptured sIA wall samples resected after microsurgical clipping of the sIA neck were compared using Affymetrix HG-U133 Plus 2.0 microarrays (see publication 1). The data was RMA normalized and compared using Welch's t-test with Benjamini-Hochberg correction for p-values. Genes with p-value < 0.05 were regarded as differentially expressed genes. This resulted in 498 overexpressed and 491 underexpressed genes in the ruptured sIA wall group.

6 *High risk population isolate reveals low frequency variants predisposing to intracranial aneurysms*

6.1 INTRODUCTION

About 3% of the population develops saccular intracranial aneurysms (sIAs) during life (Vlak et al. 2011; Ronkainen et al. 1998). Some 95% of subarachnoid hemorrhages are caused by ruptured sIA (sIA-SAH), a devastating form of stroke affecting individuals mainly in the sixth decade of life (van Gijn et al. 2007). The annual incidence of SAH is 4-9 per 100 000 worldwide (Feigin et al. 2009) but over twice as high in Finland and in Japan (de Rooij et al. 2007). The sIA disease is a complex trait, the risk of which is affected by age, sex, smoking, hypertension, excess drinking (V. L. Feigin et al. 2005), and in over 10% of the cases family history of sIA disease (Ronkainen et al. 1997; Huttunen et al. 2010; Ruigrok et al. 2001).

To date, genome wide association (GWA) studies have identified six definite and one probable (replicated in Japan but not strictly genome-wide significant) loci with common variants associated to sIA: 4q31.23 (OR 1.22) (Yasuno et al. 2011; Low et al. 2012); 8q11.23–q12.1 (OR 1.28); 9p21.3 (OR 1.31); 10q24.32 (OR 1.29); 12q22 (OR 1.16) (Yasuno et al. 2011); 13q13.1 (OR 1.20); 18q11.2 (OR 1.22)(Yasuno et al. 2010); and 9p21.3 (Helgadottir et al. 2008; Foroud et al. 2012) (see Supplementary Table 5 of the publication of study III). These seven loci were estimated to explain 6.1%, 4.4%, and 4.1% of the four-fold sibling recurrence risk in Finland, Europe, and Japan, respectively (Yasuno et al. 2011). In these previous GWA studies, results on 2q33.1 locus were inconsistent: the locus was significant in the first GWAS (Bilguvar et al. 2008), not significant in the enlarged follow-up GWAS (Yasuno et al. 2010), and in the third GWAS the risk allele was reversed in the Japanese replication sample (Yasuno et al. 2011).

The population of Finland is one of the most thoroughly characterized genetic isolates. Due to the small size of the founder population, subsequent bottleneck effects and genetic drift, the Finnish population is enriched for low frequency variants that are almost absent in other European populations and some variants rare elsewhere are increased in frequency (The 1000 Genomes Project Consortium 2012). This is best illustrated by the increased prevalence of 36 rare Mendelian, mostly recessive, disorders in Finland (www.findis.org); the so called Finnish disease heritage (FDH) (Peltonen et al. 1999). We hypothesized that some of the enriched rare or low frequency variants could contribute to the increased sIA-SAH susceptibility in Finland.

In this GWA study, we combined the power of the 1000 Genomes imputation, the special benefits of a population isolate and enrichment of familial cases in the discovery cohort. Familial sIA patients more often carry multiple sIAs as compared to sporadic sIA patients, which may confer an additional genetic burden to sIA formation (Ruigrok et al. 2004; Mackey 2012; Huttunen et al. 2010). Therefore, in addition to the case vs. control analysis, we also analyzed the number of sIAs per individual as an intermediate phenotype. We conducted an association analysis in a discovery sample of 760 Finnish sIA cases and 2,513 matched controls followed by replication in an additional sample of 858 Finnish sIA cases and 4048 controls. The successfully replicated loci in Finland were further studied in a Dutch cohort of 717 sIA cases and 3004 controls to assess the extent to which the allele frequencies and risk effect sizes match between the isolate of Finland and a continental European population (Figure 14). In addition, we hypothesized that a previously inconclusive locus on 2q33.1 (Bilguvar et al. 2008; Akiyama et al. 2010; Yasuno et al. 2011) is a true sIA risk locus at least in Finland and aimed to replicate the best discovery associations in the locus in this study in the Finnish and in the Dutch samples.

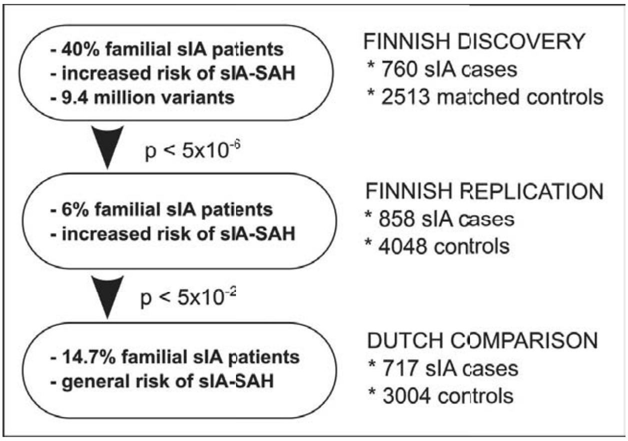


Figure 14. Study design. The Finnish discovery and replication cohorts represent a population with an over two-fold increased risk of subarachnoid hemorrhage from ruptured saccular intracranial aneurysm (sIA-SAH). The Finnish discovery cohort was intentionally enriched with familial sIA patients, and 9.4M genotyped and imputed variants were studied. The loci with p < 5E-6 were replicated in an independent and unselected Finnish sIA sample. The allele frequencies and effect sizes of the replicated variants in Finland were finally compared to a continental European population using a Dutch sample. The sIA-SAH risk is not increased in the Netherlands ('general risk' in the figure).

6.2 RESULTS

6.2.1 Case vs. control analysis in Finnish and Dutch samples

To increase the potential genetic load in the study sample, our discovery sample consisted of 760 cases from the isolated, high-risk Finnish population, purposefully enriched for familial sIA (40%) patients and 2513 genetically matched Finnish controls. The imputation of the 304,399 previously genotyped variants (Yasuno et al. 2010) against the 1000 Genomes Project reference panel (v3, March 2012 release) increased the number of common and low frequency variants available for the association analysis to 9,359,231. Quantile-quantile (QQ) plots of association p-values and genomic inflation factor ($\lambda = 1.04$) did not indicate substantial population stratification (see Supplementary Figure 1 of the original publication). The discovery association analysis revealed one locus at 12p11.1 driven by rs653464 at genome-wide significance ($p < 5 \times 10^{-8}$) and 14 other loci at $p < 5 \times 10^{-6}$ (see Supplementary Table 1 of the original publication).

We chose 17 SNPs representing the 15 promising loci (p < 5 x 10-6) above for replication in an independent sample of 858 Finnish sIA cases and 4,048 controls (Table *11*). Four SNPs and one deletion were associated at p<0.05 with sIA disease (see Supplementary Table 1 of the original publication), two of them in the previously reported sIA loci 9p21.3 (rs1333042;

OR 1.3, p = 6.3 x 10-7) and 13q13.1 (rs113124623; OR 0.88, p = 0.01). The genome-wide significant 12p11.1 locus in the discovery sample did not replicate (p = 0.29). In the meta-analysis of the two Finnish samples, four SNPs reached genome-wide significance at p < 5x10-8 (Table 9). Three were novel: 2q23.3 (rs74972714; OR 2.1, 95% CI 1.68 - 2.63, p = 7.4 x 10-11, control allele frequency or CAF 2.35%), 5q31.3 (rs113816216; OR 1.92, CI 1.53 – 2.40, p = 1.74 x 10-8, CAF 2.09%) and 6q24.2 (rs75018213; OR 1.97, CI 1.6- 2.43, p = 2.25 x 10-10, CAF 2.53%). One was previously reported at 9p21.3 (rs1333042; OR 1.31, CI 1.21 – 1.42, p = 1.8 x 10-11, CAF 42.3%) (Table 9). To assess how the allele frequencies and effect sizes of variants identified in the Finnish population compare to other European populations, we studied those variants in a Dutch sample consisting of 717 sIA cases and 3,004 controls (Table *11*). All three variants tagging the novel loci at 2q23.3, 5q31.3 and 6q24.2 had a similar low minor allele frequency (1.6-3.9%) in Finland and the Netherlands (Table 9). Two of them had similar effect sizes and were also significantly replicated: 5q31.3 (rs113816216; OR 1.3, CI 0.98 - 1.75, p = 0.045, CAF 3.87%) and 6q24.2 (rs75018213; OR 1.5, CI 0.98 – 2.3 p = 0.034, CAF 2.92%). The previously reported 9p21.3 locus also replicated in the Dutch sample (rs1333042; OR 1.32, CI 1.17 – 1.49, p = 3.42 x 10-6, CAF 47.86%). In the meta-analysis of the Finnish and Dutch samples, all three novel loci 2q23.3 (rs74972714; OR 1.89, p = 1.42 x 10-9), 5q31.3 (rs113816216; OR 1.66, p = 3.17 x 10-8) and 6q24.2 (rs75018213; 1.87, p = 7.1 x 10-11) were significantly associated to the sIA disease at genome-wide significance (Table 9).

Table 9. Five loci with genome-wide significant association to saccular intracranial aneurysm (sIA) disease in the Finnish and Dutch samples.

Case vs. control analysis		**Finnish discovery**				**Finnish replication**				**Finnish meta-analysis**				**Dutch replication**				**All meta-analysis**	
SNP*	**Gene**	**Case MAF**	**Ctrl MAF**	**OR**	**P**	**Case MAF**	**Ctrl MAF**	**OR**	**P**	**Case MAF**	**Ctrl MAF**	**OR**	**P**	**Case MAF**	**Ctrl MAF**	**OR**	**P**	**OR**	**P**
rs74972714 (C/A) 2q23.3	LYPD6 (40kb)**	3.4%	1.7%	2.73	3.43E-06	4.9%	2.8%	1.88	4.11E-06	4.2%	2.4%	2.10	7.41E-11	1.7%	1.6%	1.04	4.37E-01	1.89	1.42E-09
rs113816216 (G/C) 5q31.1	FSTL4***	4.5%	2.1%	2.31	8.26E-07	3.2%	2.1%	1.60	2.57E-03	3.8	2.1%	1.92	1.74E-08	4.5%	3.9%	1.30	4.53E-02	1.66	3.17E-08
rs75018213 (A/G) 6q24.2	EPM2A***	5.1%	2.7%	2.11	3.44E-06	4.2%	2.4%	1.85	2.85E-05	4.6%	2.5%	1.97	2.25E-10	2.9%	2.3%	1.50	3.39E-02	1.87	7.14E-11
rs1333042 (G/A) † 9p21.3	CDKN2B-AS1	50%	43.2%	1.32	3.01E-06	48.1%	41.7%	1.30	6.30E-07	49%	42.3%	1.31	1.81E-11	54.3%	47.9%	1.32	3.42E-06	1.31	6.71E-16
rs919433 (A/G) ‡ 2q33.3	ANKRD44 ***	48%	42.8%	1.25	2.53E-04	48.6%	44.6%	1.18	1.01E-03	48.3%	44%	1.21	2.15E-06	41.8%	33.2%	1.43	9.77E-09	1.27	2.20E-12
rs12472355 (A/C) ‡ 2q33.3	ANKRD44 (30kb)**	47.8%	42.7%	1.24	2.89E-04	48.8%	44.3%	1.21	2.23E-04	48.3%	43.7%	1.23	4.84E-07	39.1%	31%	1.39	1.05E-07	1.27	1.87E-12

* For each variant major allele / minor allele and locus are given.; ** The variant's distance (kb) to the nearest gene is given.; *** Located in the intron of the given gene.; † The previously reported 9p21.3 locus (Helgadottir et al. 2008; Yasuno et al. 2010).; ‡ The previously studied 2q33.3 locus with inconclusive evidence (see Materials and Methods).

Table 10. The locus with a genome-wide significant association to the number of saccular intracranial aneurysms (sIA) per individual in the Finnish samples.

Association to sIA count		**Finnish discovery**				**Finnish replication**				**Finnish meta-analysis**				**Dutch replication**				**All meta-analysis**	
SNP*	**Gene**	**Case MAF**	**Ctrl MAF**	**RR**	**P**	**Case MAF**	**Ctrl MAF**	**RR**	**P**	**Case MAF**	**Ctrl MAF**	**RR**	**P**	**Case MAF**	**Ctrl MAF**	**RR**	**P**	**RR**	**P**
rs150927513 (T/A) 7p22.1 (4894744 bp)	RADIL*	6.0%	3.6%	1.95	8.86E-08	7.0%	5.2%	1.39	8.36E-4	6.5%	4.6%	1.60	4.92E-09	0.3%	0.3%	0.97	4.82E-1	1.59	6.08E-09

* For each variant major allele / minor allele, locus and base pair position are given. ; * Located in the intron of the given gene

6.2.2 Association of variants to the number of sIAs

Some 20-30% of the sIA patients carry multiple sIAs, a phenomenon more commonly seen in familial sIA disease (Huttunen et al. 2010; Mackey 2012; Ruigrok et al. 2004). We hypothesized that an increased number of sIAs (> 2) in a given patient would reflect a higher underlying genetic load, motivating us to use aneurysm count as an intermediate phenotype to increase statistical power. The number of sIAs was used as a count data using the negative binomial regression analysis in the discovery sample of 760 Finnish sIA cases (1-8 sIAs per patient) and 2,513 controls. The QQ plot (see Supplementary Figure 2 of the original publication) and the genomic inflation factor (1.05) did not indicate substantial population stratification.

Nine loci had variants at p < 5E-6 (Supplementary Table 2). The most significant variant of each locus was selected for replication in the new Finnish sample of 858 sIA cases (1-6 sIAs per patient) and 4,048 controls. Two loci were replicated at p < 0.05: 7p22.1 (rs150927513; RR 1.39, $p = 8.36 \times 10^{-4}$, CAF 5.24%) and 16p13.3 (rs144159053; RR 1.66, $p = 4.4 \times 10^{-3}$, CAF 1.27%) (see Supplementary Table 2 of the original publication). In the meta-analysis of the Finnish samples, 7p22.1 was genome-wide significant (rs150927513; RR 1.6, CI 1.37 – 1.88, p = 4.92 x 10-9, CAF 4.61%); Table 10).

To compare the allele frequency and effect size of rs150927513 identified in the Finnish population to those of a continental European population we studied the variant also in the Dutch, but the imputation quality (Impute info 0.38) and estimated allele frequency (0.29%) were too low to obtain reliable estimates (RR 0.97; 95% CI 0.17 - 4.03, p = 0.97).

6.2.3 Analysis of 2q33.1 locus

Previously published results on the 2q33.1 locus are inconsistent, being significant in the first GWAS (Bilguvar et al. 2008), not significant in the enlarged follow-up GWAS (Yasuno et al. 2010), and uncertain in the third GWAS (Yasuno et al. 2011). We aimed to study if the 2q33.1 would replicate in Finland, even though no variant in this region reached p < 5E-6 in the discovery sample. We chose two of the most significant SNPs (in this study) at 2q33.1 for replication in the new Finnish replication sample, which was not used in the previous studies (rs12472355; OR 1.21, $p = 2.23 \times 10^{-4}$, CAF 43.7%, and rs919433; OR 1.18, p = 1.01 x 10-3, CAF 43.9%). They are in LD with the three previously investigated SNPs (rs787994, rs1429412, rs700651; LD r2 0.75-0.96). The variants rs12472355 (OR 1.23, CI 1.13 – 1.33, $p = 4.83 \times 10^{-7}$) and rs919433 (OR 1.23, CI 1.13 – 1.33, $p = 2.15 \times 10^{-6}$) did not reach genome-wide significance in the combined Finnish samples (). They were highly significant in the Dutch sample (rs12472355; OR, 1.39, CI 1.23 - 1.57, p = 1.05 x 10-7 and rs919433; OR 1.43, CI 1.26 – 1.61, p 9.77 x 10-9), and in the meta-analysis of all three samples they reached genome-wide significance (Table 9). The allele frequencies were notably higher in the Finnish samples (44% and 43.7%) than in the Dutch samples (33.2% and 31%).

Table 11. The Finnish and Dutch study samples used in the association analysis of saccular intracranial aneurysm (sIA) disease.

	Finnish discovery		Finnish replication		Dutch replication	
	Cases	Controls	Cases	Controls	Cases	Controls
N	760	2,513	858	4,048	717	3,004
Women	443 (58%)	1,454 (58%)	532 (62%)	2,182 (54%)	492 (67%)	1,135 (38%)
Familial sIA	300 (40%)	-	51 (6%)	-	100 (15%)*	-
sIA-SAH	561 (74%)	-	587 (68%)	-	658 (92%)	-
Mean age (yrs)	50	52	52	40	54	68
Number of sIAs						
mean	1.54 (1-8)	-	1.46 (1-6)	-	1.26 (1-7)	-
> 2	242 (32%)	-	257 (30%)	-	127 (18%)**	-

* Unknown familial sIA status for 35 patients.: ** Number of sIAs not known for 16 patients

6.2.4 Regulatory elements at identified loci

The UCSC Genome Browser and HaploReg version 2 (Ward & Kellis 2012) were used to search for ENCODE regulatory elements at the five genome-wide significant variants.

rs74972714 at 2q23.3 and rs150927513 at 7p22.1 reside within a DNAse hypersensitivity peak. The rs75018213 at 6q24.2 resides on an ENCODE GATA2 transcription factor binding site peak (see Supplementary Table 4 of the original publication).

Using genome-wide ChIP-seq analysis, Ernst et al. constructed a predicted cell-type specific regulatory region map of nine chromatin marks in nine cell lines (Ernst et al. 2011). rs113816216 at 5q31.3 resides on a predicted erythroleukemia cell specific (K562) strong enhancer and rs75018213 at 6q24.2 on a predicted lymphoblastoid cell (GM12878) weak enhancer (see Supplementary Table 4 of the original publication).

We searched for putative transcription factor binding sites affected by the four variants, based on position weight matrices from Transfac, Jaspar and ENCODE (top 3 enriched motifs for each transcription factor, identified in transcription factor ChIP-seq peaks (Ward & Kellis 2012)). rs74972714 at 2q23.3 affects putative binding sites for EBF1 (ENCODE), HDAC2 (ENCODE), RXRA:PPARG complex (Transfac), ZNF423 (Jaspar) and ZIC3 (Jaspar). rs113816216 at 5q31.3 affects the putative binding sites for RFX1 (Transfac), SREBP1 (ENCODE), STAT3 (Transfac) and IKZF3 (Transfac). rs150927513 at 7p22.1 affects putative binding sites of T (brachyury) (Transfac), CEBPB (Transfac) and P300 (ENCODE).

rs75018213 at 6q24.2 is not directly on any putative transcription factor binding site (see Supplementary Table 4 of the original publication).

At the 2q33.1 locus neither of the studied variants (rs919433, rs12472355) were on ENCODE DNAse hypersensitivity or transcription factor binding site peaks. However, rs919433 is on a predicted lymphoblastoid (GM12878) cell enhancer whereas rs12472355 is not directly on any regulatory region. rs919433 disrupts a putative transcription factor binding sites for RUNX2 (OSF2,Transfac) and the MYC:MAX complex (Transfac).

6.2.5 eQTL analysis

To study the potential effects of the variants in the five significant loci on the transcripts of nearby genes, we correlated the variants to expression levels of exons of nearby genes (expression quantitative trait locus (eQTL) analysis) obtained using RNA-sequencing in lymphoblasts of genotyped European individuals from the 1000 Genomes Project (Finnish, British, Toscani and CEPH populations, n=373; www.geuvadis.org) (Lappalainen et al. 2013). Each variant was correlated to transcripts residing within 1MB. There were 55 genes in 586 exons available for analysis (see Materials and Methods) and in total 748 tests were performed corresponding to Bonferroni corrected significance threshold of 8.7 x 10-5. Strongest association for each variant are reported below and all eQTL results in Supplementary Table 6 of the original publication.

The most significant eQTL associations were observed at the 2q33.1 locus: rs12472355 associated significantly to the closest gene ANKRD44 (FC 0.94, p = 1.83 x 10-5) and also to HSPD1 (FC 0.94, p = 1.6 x 10-4), whereas rs919433 was associated to the same genes but in different order of significance; HSPD1 (FC 0.94, p = 3.8 x 10-5) and ANKRD44 (FC 0.95, p = 1.4 x 10-4). Among the novel low-frequency variants, only rs150927513 at 7p22.1 was significantly associated to TNRC18 (allelic fold change (FC) 1.23, p = 5.1 x 10-5). Nominal associations were observed for two other novel low frequency variants: rs113816216 at 5q31.3 to VDAC1 (FC 2.12, p 4.6E-4); rs74972714 at 2q23.3 to EPC2 (FC 0.75, p = 3.9 x 10-2). rs75018213 at 6q24.2 did not have any association even at nominal p < 0.05 (see Supplementary Table 6 of the original publication).

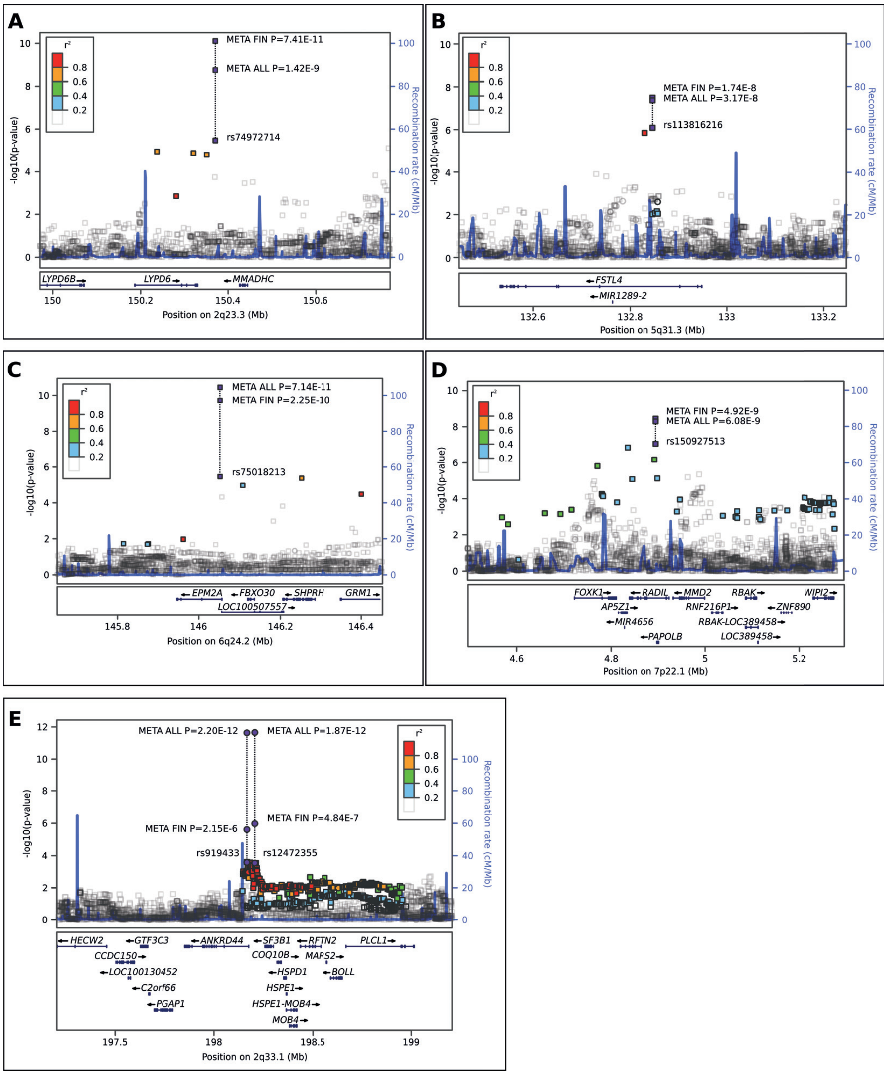


Figure 15. Regional association plots of the five identified saccular intracranial aneurysm (sIA) loci in the combined Finnish samples and the Dutch sample. Association p-values (−log10 scale, y-axis) of variants are shown according to their chromosomal positions (x-axis). Blue lines indicate the genetic recombination rate (cM/Mb). Figures A-C present the loci identified in the case vs. control analysis at 2q23.3, 5q31.3, and 6q24.2, respectively. Figure D presents the 7p22.1 locus associated to the sIA count per patient. Figure E presents the 2q33.1 locus with inconclusive previous evidence. Purple rectangles indicate the most significant variant in a) the Finnish discovery sample and, along the dashed line, its p-values in b) the combined Finnish samples (META FIN) and in c) all samples (META ALL). Adjacent variants in linkage disequilibrium (r^2; EUR populations, 1000 Genomes March 2012) to the index variant are shown in colours indicating their r^2 levels (r^2 box in each figure).

6.3 DISCUSSION

In this study, we used three approaches to improve the power to identify new loci associated to the sIA disease. First, we focused on the Finnish population isolate with increased risk for subarachnoid haemorrhage from ruptured sIAs (sIA-SAH) (de Rooij et al. 2007). Second, we enriched the proportion of familial sIA patients in the discovery sample, thus possibly increasing the prevalence of risk alleles. Third, we increased genome-wide coverage through imputing ungenotyped variants based on the 1000 Genomes Project data. Using this combination of strategies, we were able to identify three new loci associated with sIA disease, and one locus associated with the number of aneurysms. Additionally, we replicated a locus where the evidence so far was inconclusive. Together these five loci account for 2.1% of the heritability in the Finnish samples. In comparison, the six previously identified SNPs explain 2.5% of the heritability in the discovery sample of the current study. Our results likely reflect the varying genetic background of complex traits, such as sIA, in different populations.

6.3.1 Four novel sIA loci

The lead SNPs in the four novel loci all have a low frequency (< 5%) in the general population and could not have been identified without imputing the genotype data against the 1000 Genomes reference. One of the variants, rs150927513 at 7p22.1 that was associated with the number of sIAs, indicates a strong bottleneck effect, for it was 15 times more frequent in the controls of combined Finnish samples (4.6%) than in the Dutch sample (0.3%), and it is virtually non-existent in other populations (1000 Genomes). The three other loci had similar frequencies in Finland and other European populations (1000 Genomes). These four novel loci explain 1.7% of the heritability in the Finnish samples.

The four sIA loci had higher effect sizes (point estimates ranging 1.59-1.88) than the lead SNPs identified by previous GWA studies. We cannot yet conclude whether relatively high ORs of low frequency risk alleles are a typical feature of sIA disease. Similar, odds ratios for low frequency and rare variants have been reported in isolates for other traits (Sulem et al. 2011; Jonsson et al. 2013). It is likely that this first wave of low frequency and rare susceptibility variants represent "low hanging" fruits that do not allow general conclusions about the susceptibility landscape of sIA or other complex traits.

6.3.2 2q23.3 locus

The variant rs74972714 at 2q23.3 has a frequency of about 2% in European populations, including Finns. It was significantly associated to sIA in the Finnish samples but did not show a trend for being associated in the Dutch sample despite having a similar allele frequency. Further studies are required to find out whether this variant tags a risk allele specific to Finnish sIA patients. The variant is located 40kb downstream of LYPD6 and 55kb upstream of MMADHC (Figure 15 A). LYPD6 has recently been characterized as a member of the Ly-6 protein superfamily (Zhang et al. 2010). LYPD6 is ubiquitously expressed with highest levels in heart and brain. GPI-anchored Ly-6 proteins such as PLAUR function, e.g., in cellular adhesion (Zhang et al. 2010). LYPD6 overexpression can inhibit transcriptional activity of the AP1 transcription factor complex (Zhang et al. 2010), a key inflammation mediator activated, e.g., in endothelial cells in atherogenic disturbed blood flow conditions, leading in turn to upregulation of pro-inflammatory molecules (Nigro et al. 2011). Similar transcriptional changes were observed in the ruptured human sIA wall in Study I of this thesis. MMADHC is an intracellular vitamin B12 trafficking gene. Mutations in this gene can cause methylmalonic aciduria or homocystinuria, or both (Lerner-ellis et al. 2008). rs74972714 was only nominally associated to exon expression level of EPC2 (FC 0.75, p = 3.9 x 10^{-2}).

6.3.3 5q31.1 locus

The variant rs113816216 at 5q31.3 has a frequency of 1-3% in Finland and most other European populations, except in Spain (7%). It was significantly associated to the sIA disease in the Finnish samples and was also significant in the Dutch sample but had a somewhat lower OR (Table 9).

The meta-analysis of all combined samples exceeded the genome wide significance threshold. The variant is located in the intron of FSTL4 (Figure 15 B), a poorly characterized gene. FSTL1, a paralog of FSTL4, codes a protein inducing innate immunity as TLR4 agonist (Murakami et al. 2012). Increased tissue levels of FSTL1 were associated to the severity of heart failure (Lara-Pezzi et al. 2008) and to the coronary artery aneurysm formation in Kawasaki disease (Gorelik et al. 2012). Variants in FSTL4 were modestly associated to human ischemic stroke (Luke et al. 2009), and a variant 70 kb from FSTL4 nominally to hypertension (Guo et al. 2012). The strongest eQTL of rs113816216 was suggestive association with an exon of VDAC1 (FC 2.12, p 4.64 x 10^{-4}).

6.3.4 6q24.2 locus

The variant rs75018213 at 6q24.2 has similar frequencies (2%) in European populations, including Finns. It was significantly associated to the sIA disease in the Finnish samples and was also significant in the Dutch sample but had a somewhat lower OR (Table 9)

It is located in the intron of EPM2A. The LD spans over 300 kb downstream covering FBXO30, LOC100507557, SHPRH, and GRM1 (Figure 15 C). In the ENCODE data, rs75018213 is located in a GATA2 transcription factor binding site ChIP-seq peak. Homozygous deletions in the EPM2A gene result in progressive myoclonus epilepsy (PME) with Lafora bodies (OMIM 254780) (Minassian et al. 1998). No vascular anomalies have been reported in EPM2 deletion patients with a PME phenotype or their heterozygote parents. EPM2A encodes a phosphatase, which dephosphorylates glycogen, but it is likely that EPM2A has broader functions in regulation of glycogen biosynthesis, endoplasmic reticulum stress, autophagy, and possibly also cell cycle (Gentry et al. 2013).

6.3.5 7p22.1 locus and the number of sIAs

The variant rs150927513 at 7p22.1 was significantly associated to sIA count per individual in the Finnish population (Table 10). Its frequency was 4.6% in the Finnish samples but only 0.3%, in the Dutch sample, in line with most European populations. The variant is located in the intron of RADIL (Figure 15 D), a rap GTPase interactor, an essential effector of RAP1 in activation of integrins in cell-adhesive signalling by G protein-coupled receptors (Ahmed et al. 2010). RADIL has also been shown to control, together with RAP1, neutrophil adhesion and chemotaxis (Liu et al. 2012). Neutrophils seem to have a role in the formation and rupture of intracranial and abdominal aortic aneurysm (Frösen et al. 2012; Kurki et al. 2011; Eliason et al. 2005). The strongest eQTL association was to an exon of TNRC18, (FC 1.23, p 5.1 x 10-5). TNRC18 has not been functionally characterized.

As we analysed the number of sIAs as a count variable from 0-8, the inherent assumption was that the same variant would increase the risk of the first and the subsequent sIA formation. Thus, any variant associated to the number of sIAs will to some extent be associated in the case vs. control analysis. Indeed, in the analysis of combined Finnish cohorts rs150927513 was associated in the case-control analysis (OR 1.54, p = 6.5 x 10^{-7}) and consistently also in the analysis of multiple vs. single sIA patients (OR 1.65, p = 8.4 x 10^{-4}). The association of this variant, should be interpreted as reflecting the tendency of sIA formation, rather than considering multiple sIAs as a completely different dichotomous end point.

6.3.6 Previously identified 9p21.3 locus

The 9p21.3 locus has been robustly associated to the sIA disease (Yasuno et al. 2010) as well as to cardiovascular, metabolic, and cancer traits (Helgadottir et al. 2008; Wellcome et al. 2007), and it has been extensively studied by others (Johnson et al. 2013). The allele frequency and effect size in the current study, although with a different lead SNP (r2 = 0.7 to previous lead SNP rs1333040), are in strong agreement with the previous study (Yasuno et al. 2010). This locus is not therefore discussed further here.

6.3.7 2q33.1 locus with previously inconclusive evidence

Two common variants, rs12472355 and rs919433 at 2q33.1 were significantly associated to the sIA disease in the Finnish and Dutch samples (Table 9). rs919433 is located intronic and rs12472355 upstream 30 kb from ANKRD44 (Figure 15 E). The allele frequencies were somewhat higher in the Finnish samples (rs919433, 44%; rs12472355 43.7%) than in the Dutch samples (33.2%; 31%) or in the Japanese population according to 1000 Genomes Project (28.1%; 27.5%). In this locus, the risk allele was reversed in the Japanese cohort of the previous sIA GWA study (Yasuno et al. 2011). *ANKRD44* is likely a subunit of protein phosphatase 6 (Stefansson et al. 2008) that functions, e.g., in cell cycle control (Stefansson & Brautigan 2007) and in inhibition of NF-κB activation (Stefansson & Brautigan 2006). NF-κB is a significant mediator in experimental sIA formation in rats, highly expressed in human sIA wall (Tomohiro Aoki et al. 2007), and it was associated to human sIA wall rupture in Study I of this thesis. In eQTL analysis, rs12472355 was significantly associated to *ANKRD44* (FC 0.94, $p = 1.83 \times 10^{-5}$) and rs919433 to HSPD1 (FC 0.94, $p = 3.8 \times 10^{-5}$)

In conclusion, we identified four novel loci associated to sIA disease and confirmed one additional locus with previously inconclusive evidence, together explaining 2.1% of the sIA heritability in Finland. Our data illustrates the utility of high-risk population isolates, familial disease history, and dense genotype imputation in search of low-frequency variants associated to complex human diseases. The identification of the four novel low frequency variants would likely have required much larger sample sizes in more mixed populations. Further studies of the identified five loci are needed to explain their functional mechanisms in the pathogenesis of sIA disease.

6.4 MATERIALS AND METHODS

6.4.1 Study samples

A. Finnish discovery sample

The initial discovery GWAS data consisted of previously Illumina genotyped 974 Finnish intracranial aneurysm patients and 740 controls (Yasuno et al. 2010). The patients were collected from the registries of Neurosurgery, Kuopio University Hospital, and Neurosurgery, Helsinki University Hospital, solely serving their catchment populations in Eastern and Southern Finland, respectively. The sIAs were angiographically verified and the cases of subarachnoid hemorrhage from ruptured sIA (sIA-SAH) with computed tomography (CT). Patients with at least 1 first-degree relative carrying sIA disease were considered familial (Huttunen et al. 2010).

The Helsinki Birth Cohort Study (HBCS) includes 8,760 individuals born in the Helsinki Central Hospital between 1934 and 1944 (Barker et al. 2005). A subset of 1676 Illumina genotyped individuals were available for the present study. The Health 2000 Cohort (H2000) includes 2 402 Finns, and of those 2138 Illumina genotyped individuals were available for the present study (Aromaa & Koskinen 2004; THL - National Institute for Health and Welfare. 2000).

The following 210 cases and 119 controls were removed from the discovery sample: fusiform aneurysm carriers (n=5); duplicated cases (n=9) and controls (n=10); blind duplicate cases (n=15) and controls (n=5); genotyping rate <97% (29 cases, 31 controls);

individuals with higher missingness from cryptically related pairs (Identity by descent (IBD) >0.1875, similarity halfway between 2nd and 3rd degree relatives: 69 cases, 55 controls); genetic distance to 5 nearest neighbours > 4 standard deviations longer than the average distance (2 cases, 18 controls); patients not traceable from the database or with traumatic SAH (n=81); or polycystic kidney disease (n=4).

The following SNPs were removed: missing genotypes > 5%; minor allele frequency < 1%; Hardy-Weinberg disequilibrium p-value in controls < $1*10^{-6}$; symmetric SNPs (A/T, C/G); and SNPs not on all the genotyping platforms.

To minimize false positives, each sIA case was matched to three controls by gender and genetic distance from control individuals. First, a sliding window approach was used to thin the set of SNPs to be approximately independent of each other. A sliding window of 1500 SNPs was shifted by 150 SNPs at a time along chromosomes, and in each step SNPs were filtered if any pairwise r2 was > 0.2, resulting in 79596 independent SNPs. Pairwise IBS distances of these SNPs were used in multidimensional scaling and four first dimensions were used in matching. Plink v. 1.07 (Purcell et al. 2007) was used for thinning and MDS analysis. R package optmatch was used to pair each case to three controls. After 1:3 matching, additionally all Eastern Finnish controls from the previous sIA study were included (Purcell et al. 2007).

The final discovery sample consisted of 760 sIA cases and 2,513 controls (Table *11*). After SNP filtering, there were 304,399 genotyped SNPs and 9,046,433 imputed SNPs and indels (see imputation paragraph for imputation QC) for the discovery sample.

B. Finnish replication sample

The replication sample consisted of 858 independent sIA patients from the registry of Neurosurgery, Kuopio University Hospital. There were 1,605 independent controls, 453 from Eastern Finland and 1152 from the FINRISK study, both genotyped using the Sequenom iPLEX technique. Additionally, 2,443 whole genome genotyped controls from The Cardiovascular Risk in Young Finns Study were acquired and replication SNPs were extracted after imputation (Table 11).

The Cardiovascular Risk in Young Finns Study is a follow-up study of cardiovascular risk factors from childhood to adulthood (Raitakari et al. 2008; Anon 2008). The participants were randomly chosen from the Finnish Population Registry and recruited from five university cities in Finland. The baseline study launched in 1980 and included 3,596 individuals. Follow-ups have taken place at every three to six years with the last one in 2007 at 27 years of age.

The FINRISK cohort is a national survey on risk factors of chronic and non-communicable diseases in Finland (Vartiainen et al. 2010). The survey has been conducted every five years since 1972 in randomly selected, representative population samples from different parts of Finland.

C. Dutch replication sample

The Dutch sample consisted of previously GWAS genotyped 786 Dutch sIA cases (Yasuno 2010), and the 3,110 controls were recruited as part of the Nijmegen Biomedical Study (n=1,832) and the Nijmegen Bladder Cancer Study (n=1,278) (Wetzels et al. 2007; Kiemeney et al. 2008). The relevant medical ethical committees approved all studies and all participants provided written informed consent. The patients were admitted to the Utrecht University Medical Center between 1997 and 2007. The sIA-SAH cases were verified with CT scan and sIAs by angiography. Unruptured sIAs were identified by angiography in the absence of clinical or radiological signs of SAH (Yasuno et al. 2010). Patients reporting at least 1 first-degree relative carrying sIA disease were considered familial.

The Nijmegen Biomedical Study is a population based cross-sectional study conducted by the Radboud University Nijmegen Medical Centre (Wetzels et al. 2007; Kiemeney et al. 2008). Age and sex stratified, randomly selected adults (> 18 years) of Nijmegen (n=22,452) received an invitation to fill out a postal questionnaire on lifestyle and medical history.

The following cases and controls were excluded: missingness >= 0.05 (n=10); IBD >= 0.2 (n=102); heterozygosity >/< 3 standad deviations from the mean (n=46); and principal component analysis outliers (n=43). The intersection of SNPs in different platforms was first extracted and symmetric SNPs were removed (A/T, C/G). SNPs prior to the imputation were filtered by the following QC criteria: genotype missingness > 0.05; MAF < 0.01; HWE p < 0.001; and differential missingness between cases and controls p < 1E-5. The final Dutch replication sample consisted of 717 cases and 3,004 controls (Table 11).

6.4.2 Replication strategy

From both of the analyses (the case vs. controls and the number of sIAs) the best independent SNPs were taken to replication if p < 5E-6. Additional significant independent SNPs in a locus were tested by analyzing each SNP within 1 MB from the top SNP while adding the top SNP as a covariate. Additionally the most significant SNP in the current study in 2q33.1 region with uncertain evidence in previous sIA GWASs was taken to replication. A variant was considered replicated if it reached one-tailed significance of p<0.05 and was consistent in terms of the risk allele. In all of the results, one-tailed p-values are given for the Finnish replication and in the Dutch results.

6.4.3 Genotyping

Genomic DNA was extracted from peripheral blood and genotyped by Illumina arrays: the Finnish discovery sample and the Dutch replication cases by CNV370k DUO chip; the HBCS and YFS controls by Illumina Human670K customBeadChip; and the H2000 controls by Illumina Infinium HDHuman610-Quad BeadChip. In the Finnish replication sample, DNA was genotyped using Sequenom MassARRAY system and iPLEX Gold assays (Sequenom Inc., San Diego, USA). The data was collected using the MassARRAY Compact System (Sequenom) and the genotypes were called using TyperAnalyzer software (Sequenom). Genotyping quality was examined by a detailed QC procedure consisting of success rate checks, duplicates, water controls and Hardy-Weinberg Equilibrium (HWE) testing.

6.4.4 Imputation

For imputation of additional genotypes in the discovery sample, the Young Finns replication cohort and in the 2nd Dutch replication sample the genotypes were first pre-phased (Howie et al. 2012) using the Shape-IT (Delaneau et al. 2012) phasing software and the pre-phased haplotypes were subjected to imputation. The Impute version 2.2.2 software (Howie et al. 2009) with 1000 Genomes Phase I integrated variant set release (v3) reference panel (05 Mar 2012 release, http://mathgen.stats.ox.ac.uk/impute/) was used. Imputed genotypes were filtered if the Impute info measure was < 0.5 or minor allele frequency < 0.01.

6.4.5 eQTL analysis

We analyzed whether the identified genome-wide significant SNPs might affect gene expression by using the European samples of the Geuvadis RNA-sequencing data set, with mRNA sequencing data from LCLs of 373 samples from the FIN, CEU, GBR, and TSI populations of 1000 Genomes project (Lappalainen et al. 2013).

We did eQTL analysis for each of the associating variants and all the genes within a 1MB window that were expressed in >50% of the individuals. We used exon quantifications based on individual read counts per exon, after correction by the total number of mapped reads per sample and PEER normalization to remove technical variation. For each exon, we calculated linear regression between these expression values and genotype dosage of the associating variants in the 1000 Genomes data.

6.4.6 Regional association plots

Regional association plots were generated using LocusZoom with LD data from European populations of the 1000 Genomes project (Hg19/ March 2012) (Pruim et al. 2010).

6.4.7 Search of regulatory elements at identified variants

The UCSC Genome Browser and HaploReg version 2 (Ward & Kellis 2012) were used to search for ENCODE regulatory element regions located at the five genome-wide significant variants. HaploReg database also annotates if SNP resides on a putative transcription-factor binding site (TFBS) according to Transfac or Jaspar TFBS profiles and also 10 most enriched

TFBS profiles identified in ENCODE TF ChIP-eq peaks. We used all the Jaspar and Transfac annotations and the three most enriched ENCODE based TFBS annotations for each TF.

6.4.8 Statistical analysis

GWA was performed against two complementary phenotypes: the case vs. control status and the number of sIAs.

Case vs. control analysis

SNPTEST v2.3.0 was used for the association analysis, assuming additive effect. Genotype uncertainty in the imputed SNPs was taken in to account by treating them as continuous expected genotype dosages. The gender was used as a covariate.

Aneurysm count analysis

The Vuong test (Vuong 1989) showed that the negative binomial model was a significantly better fit to the sIA count per individual when compared to the Poisson model. The zero-inflated negative binomial model was not significantly better, so the simpler negative binomial model (glm.nb function in MASS R package) was used. When assessing the model fits, the gender was used as a predictor. Imputation uncertainty was taken into account by treating the imputed SNPs as continuous expected genotype dosages, and the gender was used as a covariate.

Meta-analysis

The association evidence from the discovery and replication samples were combined by inverse variance-weighted fixed-effects meta-analysis, using Plink v. 1.07 (Purcell et al. 2007).

6.4.9 Heritability analysis

The fraction of additive genetic variance explained by the five identified loci was estimated using the liability threshold model (So et al. 2011). The model assumes an additive effect at each locus, which shifts the mean of a normally distributed distribution of disease liability for each genotype. The combined genetic variance explained by the five SNPs (rs74972714, rs113816216, rs7501821, rs1509275133, rs12472355) in the five loci was assumed to be the sum of variances explained by each SNP. Risk allele frequencies in controls and OR's from combined Finnish samples was used and population prevalence of 3% of the sIA disease was assumed (Vlak et al. 2011). Heritability of the six previously identified lead SNPs (rs9298506, rs1333040, rs12413409, rs9315204, rs11661542, rs6841581) was estimated using the allele frequencies and effect sizes from the discovery cohort of the current study.

7 General discussion

Despite efforts to elucidate the genetic and molecular basis of common complex diseases, the answers still remain elusive. The intracranial aneurysm disease is no exception; rather it has received much less attention than more common complex diseases like Alzheimer's disease or type 2 diabetes.

Common complex diseases are a cause of much humanitarian suffering and also place a huge economical burden on societies in treatment costs and loss of productive life years. In the United States, common diseases were estimated to cost 277 billion dollars in treatment costs and 1,047 billion in loss of productive life years annually in 2003 (Ross & Armen 2007). Although stroke represents only 0.8% of the cases, the costs are still 36 billion annually. Thus, even for only economical reasons, the search for the molecular basis of complex diseases will and should continue.

7.1 HIGH-THROUGHPUT METHODS IN THE STUDY OF COMPLEX DISEASES

The application of high-throughput genomic methods in the study of molecular mechanism and genetics of complex diseases produces massive amounts of data, which is further typically complicated by noise caused by measurement technology or external factors related to study design. A brief summary of most typical methods is presented in Table 12 (see chapters 2.1.4 and 2.2.2 for more detailed discussion). Especially the application of next-generation or now-generation sequencing is expected to grow in popularity as the costs are rapidly plummeting. For example, sequencing of the human genome currently costs approximately a few thousands euros, the cost of which was around 100 million just a decade ago. As the cost decreases and the technology matures, producing the raw data and basic analyses are becoming more trivial but the greatest challenges are the bioinformatic analysis, interpretation of the produced data, information technology (IT) related issues, and importantly functional followup studies to conclusively prove the proposen mechanistic links between findings and phenotypes. The IT requirements for imputation and storing genotypes of > 3,200 individuals of the discovery cohort in our study III required about a week of computation in a cluster computing environment of 81 nodes with 24 CPUs each and 2 TB of disk space. In comparison, storing the full genome sequencing results of a single individual can take around 300 Gb (30x coverage) and take over a day to just call the variants (16 CPU environment) (Sboner et al. 2011). Making sense of the genome sequence and linking variants to phenotypes can take even months from a team of bioinformaticians, statisticians, geneticists, and physicians (Sboner et al. 2011).

A traditional way of handling the computational and storage requirements is to build and administer in-house cluster computing environments. This approach requires investments in hardware and maintenance personnel and the hardware capacity must be fitted to the heaviest use case even if that use would be infrequent. Cloud computing offers a flexible alternative to efficiently allocate virtual computational and storage resources via the Internet. Public Cloud computing providers such as Amazon Web Services (AWS) (http://aws.amazon.com/) enable a pay-as-you-go type of computational infrastructure where you can dynamically allocate computing capacity and pay only for the used resources. Transfer of large amounts of data to the public cloud can still be a bottleneck and it is not uncommon to ship a hard-drive to the Cloud Service provider to be transferred to the cloud. AWS has a free public repository of some large datasets that can be used in own virtual computing instances from AWS, such as Ensembl Human genome database (size 310GB) and 1000 Genomes project data (size 200 TB). For a review of cloud computing in bioinformatics see (Dai et al. 2012).

Table 12. Summary of common high-throughput methods in genomic studies of complex human diseases.

Technology	***Purpose***	***Description***	***Advantages***	***Disadvantages***
Oligonucleotide Microarrays	*mRNA, miRNA, profiling of disease tissue.*	*Oligonucleotide baits to capture mRNA printed on chips designed based on known transcripts in human reference genome.*	*Relatively cheap, mature bioinformatics methods for quality control and data-analysis.*	*Can only detect known transcripts, biological interpretation of results can be challenging.*
RNA-seq	*mRNA, miRNA, profiling of disease tissue.*	*High throughput sequencing of transcripts.*	*Higher dynamic range than microarrays, can detect novel transcripts, identification of sequence variants in transcripts.*	*More expensive than microarrays, less mature statistical methods for quality control and analysis, biological interpretation of results can be challenging.*
SNP microarrays	*Genetic studies to identify genetic risk variants / loci.*	*Assayed variants designed based on known variations and LD in human genome (e.g. HapMap).*	*Very good genotyping quality, relatively cheap, mature statistical methodology*	*Detection of known variants only, most arrays identify only common variants, translating findings into pathophysiological*

			and software.	*understanding is often challenging.*
Next generation exome sequencing.	*Genetic studies to identify (causative) genetic risk variants / loci.*	*Protein coding parts of genome are captured using capture kit, which is designed based on known transcripts. Next generation sequencing of captured DNA.*	*Affordable high coverage sequencing of protein coding parts of genome, robust single nucleotide variant and small INDEL detection methodology, moderate IT requirements, coding variants can be easier to interpret biologically.*	*Misses non-coding variants, can detect coding variants of only known transcripts.*
Next generation whole-genome sequencing	*Genetic studies to identify (causative) genetic risk variants / loci*	*Next generation sequencing of full genomic DNA.*	*Identification of variants everywhere in the genome instead of pre-selected parts, robust single nucleotide variant and small INDEL detection methodology*	*Especially high-coverage sequencing is expensive, larger structural variant detection still unreliable, very high data storage and analysis requirements, interpretation of non-coding parts of genome is very challenging*

7.1.1 Strengths and weaknessess of applied high-throughput approaches in study of complex disease

In trying to understand the molecular mechanisms of complex human diseases, it would seem important to study human samples of the phenotypical tissue in a holistic way, like transcriptome profiling of the sIA wall in the current thesis. This approach has at least one serious limitation; the molecular phenomena observed in the diseased tissue might not reflect the causative processes, but just the end stage status of the disease. Observations from such studies can, however, be a useful way to generate data driven hypotheses of genes and pathways that otherwise might not have been considered relevant. These hypotheses must then be proven in more focused studies in further research. Animal models can be a useful way to test such hypotheses as the researcher can control the experminental conditions and coincide the sampling time of tissues with various stages of the diasese. Unfortunately, there are no naturally occurring sIAs in animal models but the sIAs have to be induced by e.g. hypertension and elastin degradation (Nuki et al. 2009).

Studying the genetic susceptibility to complex diseases is not as susceptible to confounders as studies of phenotypical tissues. The genotype of an individual is not dependent on environmental effects and therefore identifying variants predisposing to complex diseases therefore are either in a directly causative path or predisposes to concomitant disease or phenotype, that in turn increases the risk of the disease of interest. The translation of association to knowledge can be often challenging but has had success in highlighting important pathways and genes e.g. in IL23-R pathway in Crohn's disease, factor H in age related macular degeneration (Visscher et al. 2012), and regulation of *SORT1* in controlling plasma LDL levels likely contributing to myocardial infarction (Musunuru et al. 2010).

7.1.2 Interpretation challenges of genome-wide studies

The challenges of development for advancing the interpretation of high-throughput methods can be broadly divided into two: annotation challenges and methodological challenges. First, the functional understanding and annotation of genes, regulatory elements, signaling molecules and their interplay in pathways need improvement. These annotations should become more precise e.g. in terms of different isoforms or different tissues and conditions and they should be made computationally accessible in a systematic manner. Second, the bioinformatics analysis methods need to keep pace with the evolving annotation systems to maximally benefit from the higher resolution knowledgebases. The results from current pathway analysis methods (either positive or negative) cannot often be just taken at face value but combined knowledge of domain experts in the studied disease, cellular molecular mechanisms, bioinformatics, and used technology platforms is often needed. A short summary of common pathway analysis methods of high-throughput studies is given in Table 13 (see chapter 2.5 for detailed discussion).

In the analysis of differentially expressed gene sets in study I, we also utilized the most popular gene set enrichment method, GSEA (Subramanian et al. 2005) and one topology based analysis method SPIA (Tarca et al. 2009) (data not shown). The SPIA method identified the same KEGG pathways as simpler over-representation analysis, only the ordering was different. The practical utility of a topology-based method in our case was therefore limited, although utilizing the topology in signaling pathway analysis feels intuitively appealing. It is expected that topology based analyses suffer from lack of precise knowledge of signaling pathways in different tissues and will greatly benefit from the evolving knowledge and resolution of signaling pathway databases in the future. Similar results were also obtained from GSEA analysis of GO categories. Thus, the GSEA method also did not have practical utility in our case. The advantage of GSEA is however that no strict statistical cut-off has to be specified. One shortcoming of GSEA could be that seeking enrichment in the top or the bottom of the gene list, which is ordered by strength of correlation to phenotype of interest might not reveal all relevant changes. Functionally similar genes might not be necessarily similarly differentially regulated. Small differential expression in functionally similar genes could also convey larger effects, and the method of Study II was developed to study this type of hypotheses, which are likely missed by the existing methods.

Table 13. Short summary of typical pathway analysis methods of high-throughput genomic studies.

Method	*Description*	*Pros*	*Cons*
Over-representation analysis	*Interpretation of differentially expressed gene sets. Assess if some pre-specified gene annotations (e.g. biological functions) occur in the analyzed gene set more often than would be expected just by chance.*	*Aids interpretation of gene sets by highlighting higher-level themes, computationally easy, huge variety of software available.*	*Depends on the accuracy of annotations, existing gene annotations may not be relevant in the studied condition/tissue, the differentially expressed gene set is based on arbitrary statistical cut-off, huge variety of software available.*
Enrichment analysis	*Interpretation of differentially expressed gene sets. Assess if pre-specified annotations (e.g pathways) occur in top or bottom of ranked list of genes more often than would be expected just by chance.*	*Aids interpretation of gene sets by highlighting higher level themes, is not dependent on arbitrary statistical cut-off for differentially expressed gene set, significance calculation by permutation is computationally heavy*	*Depends on the accuracy of annotations, existing gene annotations may not be relevant in the studied condition/tissue, ignores interaction of genes in canonical pathways.*
Topology based pathway analysis	*Interpretation of differentially expressed gene sets. Takes into account the topological ordering of interacting genes in a pathway when*	*Aids interpretation of gene sets by highlighting biological pathways, utilizes the topological information of*	*Depends on the accuracy of pathways, existing pathways may not be relevant in the studied condition/tissue*

	calculating pathway significance	*existing pathway resources*	
GWAS pathway analysis	*Identification of possibly weak but coordinated statistical signal among genes belonging to a common pathway.*	*Helps to put the genetic findings in biological context, aid in discovery of genomic loci with weaker risk effects.*	*Permutation based analyses computationally demanding, relatively recently developed and no clear consensus exist about best methodology.*

7.2 QUEST FOR MOLECULAR PATHOMECHANISMS OF SACCULAR INTRACRANIAL ANEURYSM DISEASE

In this thesis, we aimed to elucidate signalling pathways and genetic background of saccular intracranial aneurysm formation and rupture. This knowledge is needed for understanding the molecular mechanisms of aneurysm formation and rupture, which could lead to the design of novel methods for non-invasive diagnosis, prevention, or occlusion of sIAs.

In study I, we identified pathways and transcription factors potentially contributing to the process of sIA wall rupture, which might serve as a target for novel non-invasive therapies to stabilize the wall of sIAs. Some of the identified signaling molecules have already shown promise in animal models of sIA. We identified overrepresentation of NFKB and the ETS-family of transcription factors in the promoter region of upregulated genes in ruptured sIA wall. Consistent with our results, a recent study showed that a dual inhibition of Nuclear Factor-kB and Ets-1 dimished the size and thickened the wall of existing IAs in rats (Aoki et al. 2012). Our results also add to the growing body of scientific evidence that inflammatory changes could precede and predispose to sIA rupture (Chalouhi et al. 2012) and we additionally pinpointed more specific putative targets for future studies and novel therapy development.

The limitations of study I are that it is possible that some of the observed signaling pathways are mere reactions to rupture, although we did not find evidence to support such a view by analyzing gene expression levels in relation to time from rupture. Additionally, we did not study protein expression or localization in sIA tissues. Thus, we are not able to pinpoint, which of the cell populations exhibit differential expression and if the mRNA level changes are reflected to protein level expression. Further more focused studies are therefore warranted. One such study has already been performed where oxidative stress activity was observed in polymorphonuclear cells in the luminal thrombus (in perfect agreement with our hypothesis in Publication I) and oxidative stress response genes

HMOX1 identified in the Publication I was shown to be upregulated in the ruptured sIA walls but was also associated to wall degeneration in unruptured ones (Laaksamo et al. 2013). Laaksamo et al. also conclude that the observed changes are likely not just reactions to rupture.

In study II, we developed a bioinformatics method, Independent Enrichment Analysis (IEA), for data-mining the differentially expressed genes of study I. Specifically, we hypothesized that interesting biological phenomena are driven by a subset of the differentially expressed genes, which are not revealed by existing pathway analysis methods: a key novelty of IEA is to identify functional subgroups from large gene sets and provide clues about their regulatory control. As similar approach could potentially be useful in research of many other conditions, we also sought to develop publicly available, easy to use software TAFFEL. TAFFEL was designed to enable agile evaluation and use of our method by other researchers without bioinformatics skills.

Using the developed method and software, we generated novel data driven hypotheses of signaling pathways active in the ruptured sIAs, which were not identified by other methods. One such hypothesis relates *TAL1* transcription factor to controlling downregulation of cell developmental processes, and more specifically blood vessel development. As TAL1->VE-Cadherin->TGF-beta pathway maintain vascular stability (Rudini et al. 2008), the downregulation of these genes could potentially be involved in the weakening of the vessel wall. Another plausible hypothesis generated was the link between *MEF2A* transcription factor and apoptosis. Low vascular smooth muscle cell count with disorganized architecture in sIAs is associated with aneurysm rupture (Frösen et al. 2004), and our results suggest that the role of MEF2A in these processes should be investigated in further studies.

In our IEA analysis, also many other unreported clusters also contained many interesting links between function and regulation, even though IEA did not reach formal statistical significance after multiple testing corrections. This is actually expected, as functional knowledge of many genes is limited and therefore not annotated with the functions the genes might actually be involved in. Also some of the gene functions inferred are from cell culture studies or from different tissues/conditions and thus might not hold in other tissues or conditions (here sIA wall). These IEA clusters could anyway serve as a basis for generating new hypotheses about differentially expressed genes functions in studied conditions even if all of the genes are not (yet) annotated with hypothesized function and gain insight on transcriptional regulation of those functions.

A main limitation of the study is that we did not provide additional evidence to support the generated hypotheses. This aspect touches also on one key hindrance in signalling pathway analysis method development. The ground truth (i.e. what pathways are really differentially active) is typically not known so it is difficult to systematically compare the methods in terms of e.g. sensitivity to identify correct biological processess. In specific cases, like in our study, the role of identified processes could potentially be studied in e.g. knockout studies in arterial wall cells or in animal models of sIA disease.

Many interesting studies on human transcriptional regulation have been published after the publication of study II. One of the most interesting ones is the ENCODE projects report of DNA binding of 119 different transcription factors in 72 different cell types (Dunham et al. 2012). Using these binding sites in the IEA method instead of computationally inferred ones would likely lead to improvements in the method. Also pathway annotations have been evolving since the study and we aim to extend our methodology and software in the future also to include ENCODE TF data and recent pathway databases.

In study III, we hypothesized that novel sIA susceptibility loci could be identified in a high-risk population of Finland especially among low frequency variants. Identification of such variants could potentially highlight genes and pathways relevant to molecular pathogenesis of sIA disease and to be of general interest to the genetics research community. We identified low frequency variants in four novel loci associated to either sIA status or number of aneurysms, a phenotype hypothesized to reflect genetic load of sIA disease. We also provided evidence that a previously controversial locus in 2q33.1 with inconclusive evidence is associated to sIA disease at least in Finland and Europe.

As is typical for genome-wide association studies, it is not yet possible to conclusively associate genes and functions through, which the identified variants convey the sIA risk effect. Two of the loci had functionally hypothesizable genes whose effects converge with some of the biological processes identified in the Study I and in literature on the sIA disease. The variant identified on 7p22.1 predisposing to multiple aneurysm formation is located in an intron of *RADIL* gene. The variant in this locus also shows a strong bottleneck effect: it is observed only in Finns (MAF 4%) and Italians (MAF 0.5%) in 1000 genomes project data. *RADIL* controls cell matrix adhesion and has been shown to control neutrophil adhesion and chemotaxis (Liu et al. 2012). Neutrophil signaling was associated to aneurysm rupture in study I and neutrophils have also been observed to be trapped in luminal thrombus of human sIAs (Frösen et al. 2012) and to be a source of potentially wall damaging oxidative stress as suggested in Publication I and by Laaksamo et al. (Laaksamo et al. 2013). As the identified variant is on a DNAse hypersensitivity peak of two fibroblast cell lines in the ENCODE data, this gives rise to another hypothesis that the variant could affect fibroblast cell adhesion properties in the adventitia, which has essential structural and functional roles in vascular wall (Stenmark et al. 2013).

The other locus with increased frequency in Finland was the previously inconclusively reported 2q33.1. The most strongly associated variant was more frequent in Finland (37.6%) than in other Europan populations (20%-30%) in 1000 genomes data. Notably in the Japanese population, where the replication in previous studies failed, the MAF is only

27.5%. The variants are located in the intron and upstream of *ANKRD44*, a subunit of Protein phosphatase 6, which has been linked to regulating NfKB activity, which in turn likely plays a crucial role in aneurysm development (Tomohiro Aoki et al. 2007) and rupture (Study I). These results indicate that the role of *ANKRD44* in aneurysm disease should be further investigated.

7.3 COMMON FINDINGS IN DIFFFERENT APPROACHES IN THE CURRENT THESIS

Inflammation related genes and pathways were common between the expression study and genetic study of the current thesis. The involvement of neutrophils specifically was suggested by the gene expression study and the genetic study (*RADIL* gene). Neutrophils had not been previously linked to sIA disease, which opens up new directions for sIA research. Different components of NfKB, a master regulator of inflammatory signalling, were also suggested to be involved in sIA formation and rupture. Computational predictions by developed TAFFEL software suggested more vessel wall structural phenomena to be associated with sIA rupture rather than inflammatory signalling. This would be expected *a priori,* since the wall would be expected to weaken for sIA to form or rupture.

7.4 NEAR FUTURE AVENUES FOR SIA RESEARCH

As both the gene-expression (Study I) profiles and genetic variants (Study III) suggest neutrophil involvement in sIA formation and rupture this should be a primary focus for subsequent functional studies. The potential role of identified variant in *RADIL* gene on neutrophil phenotypes could be studied by e.g. extracting leukocytes from patients and controls with and without the variant and performing comparative adhesion assays between wild type and variant cells. For following up the hypotheses generated by IEA method (Study II) the identified transcription factors and the genes they were predicted to regulate should be further studied. The localization and expression levels of the proteins in human IA tissues could be studied by immunohistochemistry. These same genes could also be knocked out or over-expressed in mouse models and the phenotypical consequences on vascular tissue observed.

The largest GWAS studies have likely identified most common variants (MAF ≥ 10%) with modest effect sizes (genotype relative risk > 1.25) (Yasuno et al. 2010; Yasuno et al. 2011). First, a natural continuation to study the genetics of the sIA disease is to investigate lower frequency variants and/or variants with even lower effect sizes. The primary way to find variants with smaller effect sizes is to organize GWAS studies of very large cohorts of sIA patients and controls. Low-frequency (MAF 0.5%-5%) or rare (MAF < 0.5%) variants have received much of the focus in the discussions where the hidden heritability may lie. The genotyping arrays used in the majority of the GWAS studies do not capture well these lower frequency variants. Futhermore, the effect sizes are likely not high enough to be captured by linkage in families (Manolio et al. 2010). Imputation based on whole-genome-sequenced individuals (Study III), use of exome targeted (Huyghe et al. 2013), or custom GWAS chips (Voight et al. 2012), and exome or full-genome sequencing are the tools available for studying the association of lower frequency variants to sIA and other complex diseases. Study III is the first study reporting low-frequency variants associated to sIA disease in genome-wide association analysis. The identified variants have relatively high odds ratios. The sIA disease commonly affects individuals past their prime reproductive age, and consequently it is possible that the negative selection pressure against variants with higher risks has not been strong. Future studies of low-frequency variants are needed to address the question whether high-risk low-frequency variants are a feature of sIA disease. Studying very rare penetrant Mendelian variants in families using exome or full-genome sequencing is another valuable approach in that they may highlight disease associated functions and pathways. This approach has identified *SMAD3* mutations causing Thoracic aortic aneurysms and dissection and many different forms of arterial aneurysms, including intracranial aneurysms (Regalado et al. 2011). We have identified an Eastern Finnish family where parents, all six children and two out of six of the father's siblings are affected with sIA disease. We hypothesize that a highly penetrant very rare coding variant is causing the disease in this family and we will apply exome sequencing to identify the putative mutations.

The role of epigenetics, the alteration of phenotypic expression of genomic information without changes in DNA sequence, in sIA disease has not been studied so far. Genome-wide epigenetic modification can be studied using immunoprecipitation of DNA sequences associated with the chromatin modification of interest followed by sequence detection by arrays or increasingly by next-generation sequencing (Bock et al. 2010; Ku et al. 2011). Epigenetic changes have been associated to many cardiovascular functions in development and disease such as angiogenesis, flow-dependent regulation of gene-expression, smooth muscle cell proliferation, and vascular inflammation (Lorenzen et al. 2012; Schnabel et al. 2012). Most of the evidence is coming from basic science, however, and epidemiological and clinical data is lacking (Schnabel et al. 2012). One clinical example is the observation of DNA hypomethylation in promoter regions of upregulated genes in end-stage failing human hearts. These observations, however, can not discriminate cause from consequence (Movassagh et al. 2011).

Differential expression of microRNAs (miRNA), a class of non-protein coding RNA molecules, which post-transcriptionally regulate mRNA expression, has been associated to many cardiovascular functions and diseases (Quiat & Olson 2013) including abdominal aortic aneurysms (Boon & Dimmeler 2011). Only one study in experimental intracranial aneurysms in rats have studied the genome-wide differential expression of miRNAs (Lee et al. 2013), consequently the role of miRNAs in sIA disease development is almost completely unknown.

Many studies relating genetic, epigenetic, transcriptomic, proteomic, lipidomic, and environmental variables to complex disease have studied these phenomena in isolation. Studies presented in this thesis belong to this category. This simplistic view is unlikely to fully reflect the complex interactions occurring in the development of complex human diseases, including sIA disease. Modern systems biology aims to integrate disparate data sources and analyze them together as a network of interacting participants (Medina 2013). Although very challenging, due to the explosion of combinatorial possibilities alone, development of systems biology promises to provide a more holistic view of the processes leading to complex diseases, hopefully leading to novel methods for disease prevention, diagnosis, and personalized treatment.

8 Conclusions

The overarching aim of the current study was to elucidate signaling processes leading to sIA rupture and genetic predisposition to formation and rupture of the sIA pouch using genomewide methods and related bioinformatics. This knowledge is needed for understanding the molecular mechanisms of aneurysm formation and rupture, which could be used as a basis for novel methods for diagnosis, prevention, or occlusion of sIAs. The primary approaches used in this thesis were signalling pathway analysis by whole-genome microarrays, novel bioinformatic method development, and the methods application to our microarray data and genome-wide association analysis in the Finnish population.

By using three complementary approaches, we were able to identify several genes and biological processes associated with sIA formation and rupture. In addition to shedding light on molecular pathomechanism of sIA disease, these findings may also serve as a basis for more focused studies aiming to find druggable targets for development of novel methods to identify rupture prone patients and to prevent sIA development or rupture.

In study I, we identified biological processes associated to sIA rupture and transcription factors putatively controlling those processes using genome-wide transcriptome analysis. Some of the identified genes and processes have lately been shown to be associated to sIA disease in independent studies in human sIA tissue and animal models.

In study II, we developed a novel bioinformatic method and software to gain additional insight in to differentially expressed gene set of study I. The principal idea of the method development was to be able to identify biological processes and transcription factors putatively controlling those processes from subsets of the whole set of differentially expressed genes.

We also developed easy to use, publicly available software for other researchers without requirement for advanced computing skills. Using the developed method and software, we were able to identify additional biological processes and transcription factors associated to sIA rupture, which were not reported by other popular pathway analysis methods.

In study III, we identified low frequency variants in three novel loci associated to sIA disease and one variant associated to the number of aneurysms in the high-risk population of Finland. Our study highlights the utility of population isolates, imputation based on whole genome sequencing projects, and use of alternative phenotypes in identification of susceptibility variants to complex diseases. Two of the identified variants were also replicated in a Dutch case/control cohort whereas two of the variants are putatively specific to Finland. The putative Finnish specific variants may be related to, but not explain, the higher than average SAH incidence in Finland. The potential mechanisms how the identified variants are related to sIA disease are not definitely known and the role of implicated genes in sIA disease should be a focus of further studies.

It should finally be noted that the identification of these different genes and pathways were basically based on statistical associations to various aspects of the sIA disease. Mere associations can never prove causality and therefore the identified genes and pathways should be considered as data driven hypotheses and serve as a basis for more focused studies in the future. But as Albert Einstein once said: "I think that only daring speculation can lead us further and not accumulation of facts".

9 References

Ackermann, M. & Strimmer, K., 2009. A general modular framework for gene set enrichment analysis. BMC bioinformatics, 10, p.47.

Ahmed, S.M. et al., 2010. G protein betagamma subunits regulate cell adhesion through Rap1a and its effector Radil. The Journal of Biological Chemistry, 285(9), pp.6538–51.

Akaike, H., 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6), pp.716–723.

Akiyama, K. et al., 2010. Genome-wide association study to identify genetic variants present in Japanese patients harboring intracranial aneurysms. Journal of human genetics, 55(10), pp.656–61.

Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J., 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics (Oxford, England), 20(4), pp.578–580.

Allen, H.L. et al., 2011. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature, 467(7317), pp.832–838.

Allison, D.B. et al., 2006. Microarray data analysis: from disarray to consolidation and consensus. Nature reviews.Genetics, 7(1), pp.55–65.

Altshuler, D., Daly, M.J. & Lander, E.S., 2008. Genetic mapping in human disease. Science (New York, N.Y.), 322(5903), pp.881–8.

Altshuler, D.M. et al., 2010. Integrating common and rare genetic variation in diverse human populations. Nature, 467(7311), pp.52–8.

Anon, 2008. The Cardiovascular Risk in Young Finns Study. Available at: http://vanha.med.utu.fi/cardio/youngfinnsstudy/index.html [Accessed January 22, 2013].

Aoki, T., Kataoka, H., et al., 2009. Impact of monocyte chemoattractant protein-1 deficiency on cerebral aneurysm formation. Stroke; a journal of cerebral circulation, 40(3), pp.942–951.

Aoki, T. et al., 2007. Macrophage-derived matrix metalloproteinase-2 and -9 promote the progression of cerebral aneurysms in rats. Stroke, 38(1), pp.162–9.

Aoki, T., Nishimura, M., Kataoka, H., et al., 2009. Reactive oxygen species modulate growth of cerebral aneurysms: a study using the free radical scavenger edaravone and p47phox(-/-) mice. Laboratory investigation; a journal of technical methods and pathology, 89(7), pp.730–741.

Aoki, T. et al., 2012. Regression of intracranial aneurysms by simultaneous inhibition of nuclear factor-κB and Ets with chimeric decoy oligodeoxynucleotide treatment. Neurosurgery, 70(6), pp.1534–43; discussion 1543.

Aoki, T., Nishimura, M., Ishibashi, R., et al., 2009. Toll-like receptor 4 expression during cerebral aneurysm formation. Journal of neurosurgery.

Aromaa, A. & Koskinen, S. eds., 2004. HEALTH AND FUNCTIONAL CAPACITY IN FINLAND. Baseline Results of the Health 2000 Health Examination Survey, Publications of the National Public Health Institute.

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics, 25(1), pp.25–29.

Baker, A.B. et al., 2009. Heparanase alters arterial structure, mechanics, and repair following endovascular stenting in mice. Circulation research, 104(3), pp.380–387.

Bamshad, M.J. et al., 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nature reviews. Genetics, 12(11), pp.745–55.

Bard, J.B.L. & Rhee, S.Y., 2004. Ontologies in biology: design, applications and future challenges. Nature reviews. Genetics, 5(3), pp.213–22.

Barker, D.J.P. et al., 2005. Trajectories of growth among children who have coronary events as adults. The New England Journal of Medicine, 353(17), pp.1802–9.

Barouki, R. & Morel, Y., 2001. Repression of cytochrome P450 1A1 gene expression by oxidative stress: mechanisms and biological implications. Biochemical pharmacology, 61(5), pp.511–516.

Bauer-Mehren, A., Furlong, L.I. & Sanz, F., 2009. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Molecular systems biology, 5(290), p.290.

Beissbarth, T. & Speed, T.P., 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics (Oxford, England), 20(9), pp.1464–1465.

Bell, J.T. & Spector, T.D., 2011. A twin approach to unraveling epigenetics. Trends in genetics◎: TIG, 27(3), pp.116–25.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society.Series B (Methodological), 57(1), pp.289–300.

Berger, S.L., 2007. The complex language of chromatin regulation during transcription. Nature, 447(7143), pp.407–12.

Bevan, S. et al., 2012. Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. Stroke; a journal of cerebral circulation, 43(12), pp.3161–7.

Bilguvar, K. et al., 2008. Susceptibility loci for intracranial aneurysm in European and Japanese populations. Nature Genetics, 40(12), pp.1472–7.

Bird, A., 2007. Perceptions of epigenetics. Nature, 447(7143), pp.396–8.

Bock, C. et al., 2010. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nature biotechnology, 28(10), pp.1106–14.

Boon, R. a & Dimmeler, S., 2011. MicroRNAs and aneurysm formation. Trends in cardiovascular medicine, 21(6), pp.172–7.

Bradford, J.R. et al., 2010. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. BMC genomics, 11, p.282.

Brisman, J.L., Song, J.K. & Newell, D.W., 2006. Cerebral aneurysms. The New England journal of medicine, 355(9), pp.928–39.

Bromberg, J.E. et al., 1995. Familial subarachnoid hemorrhage: distinctive features and patterns of inheritance. Annals of neurology, 38(6), pp.929–34.

Bryne, J.C. et al., 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic acids research, 36(Database issue), pp.D102–6.

Bux, J., 2008. Human neutrophil alloantigens. Vox sanguinis, 94(4), pp.277–285.

Cai, J.J. et al., 2009. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS genetics, 5(1), p.e1000336.

Carmona-Saez, P. et al., 2007. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome biology, 8(1), p.R3.

Chalouhi, N. et al., 2012. Biology of intracranial aneurysms: role of inflammation. Journal of Cerebral Blood Flow and Metabolism, 32(9), pp.1659–76.

Chen, X. & Murphy, R.F., 2005. Objective clustering of proteins based on subcellular location patterns. Journal of biomedicine & biotechnology, 2005(2), pp.87–95.

Chetty, R. et al., 1997. TAL-1 protein expression in vascular lesions. The Journal of pathology, 181(3), pp.311–315.

Chial, H., 2008. Mendelian genetics: Patterns of inheritance and single-gene disorders. Nature education, 1(1).

Chiu, J.J., Usami, S. & Chien, S., 2009. Vascular endothelial responses to altered shear stress: pathologic implications for atherosclerosis. Annals of Medicine, 41(1), pp.19–28.

Choudhary, C. & Mann, M., 2010. Decoding signalling networks by mass spectrometry-based proteomics. Nature reviews. Molecular cell biology, 11(6), pp.427–39.

Churchill, G. a & Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics, 138(3), pp.963–71.

Cloft, H.J. et al., 1998. Prevalence of cerebral aneurysms in patients with fibromuscular dysplasia: a reassessment. Journal of neurosurgery, 88(3), pp.436–40.

Conrad, D.F. et al., 2011. Variation in genome-wide mutation rates within and between human families. Nature genetics, 43(7), pp.712–714.

Conway, D.E. et al., 2010. Endothelial metallothionein expression and intracellular free zinc levels are regulated by shear stress. American Journal of Physiology. Cell physiology, 299(6), pp.C1461–7.

Craig, J., 2008. Complex diseases: Research and applications. Nature education, 1(1).

Croft, D. et al., 2011. Reactome: a database of reactions, pathways and biological processes. Nucleic acids research, 39(Database issue), pp.D691–7.

Dahlquist, K.D. et al., 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nature genetics, 31(1), pp.19–20.

Dai, G. et al., 2004. Distinct endothelial phenotypes evoked by arterial waveforms derived from atherosclerosis-susceptible and -resistant regions of human vasculature. Proceedings of the National Academy of Sciences of the United States of America, 101(41), pp.14871–14876.

Dai, L. et al., 2012. Bioinformatics clouds for big data manipulation. Biology direct, 7, p.43; discussion 43.

Dai, M. et al., 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic acids research, 33(20), p.e175.

Darwin, C., 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, London: John Murray.

Das, R., Pluskota, E. & Plow, E.F., 2010. Plasminogen and its receptors as regulators of cardiovascular inflammatory responses. Trends in Cardiovascular Medicine, 20(4), pp.120–4.

Davidson, E.H. & Erwin, D.H., 2006. Gene regulatory networks and the evolution of animal body plans. Science (New York, N.Y.), 311(5762), pp.796–800.

Delaneau, O., Marchini, J. & Zagury, J.-F., 2012. A linear complexity phasing method for thousands of genomes. Nature Methods, 9(2), pp.179–81.

Dennis Jr, G. et al., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology, 4(5), p.P3.

Dinu, I. et al., 2009. Gene-set analysis and reduction. Briefings in bioinformatics, 10(1), pp.24–34.

Draghici, S. et al., 2007. A systems biology approach for pathway level analysis. Genome research, 17(10), pp.1537–45.

Draghici, S. et al., 2003. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucleic acids research, 31(13), pp.3775–3781.

Dunham, I. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), pp.57–74.

Dunzendorfer, S., Lee, H.K. & Tobias, P.S., 2004. Flow-dependent regulation of endothelial Toll-like receptor 2 expression through inhibition of SP1 activity. Circulation research, 95(7), pp.684–691.

Eisenhut, M. & Wallace, H., 2011. Ion channels in inflammation. Pflügers Archiv◎: European journal of physiology, 461(4), pp.401–21.

Eliason, J.L. et al., 2005. Neutrophil depletion inhibits experimental abdominal aortic aneurysm formation. Circulation, 112(2), pp.232–240.

Ellamushi, H.E. et al., 2001. Risk factors for the formation of multiple intracranial aneurysms. Journal of neurosurgery, 94(5), pp.728–32.

Enattah, N.S. et al., 2002. Identification of a variant associated with adult-type hypolactasia. Nature genetics, 30(2), pp.233–7.

Ernst, J. et al., 2011. Systematic analysis of chromatin state dynamics in nine human cell types. Nature, 473(7345), pp.43–49.

Eyre, S. et al., 2012. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nature Genetics, 44(12), pp.1336–40.

Feigin, V. et al., 2005. Smoking and elevated blood pressure are the most important risk factors for subarachnoid hemorrhage in the Asia-Pacific region: an overview of 26 cohorts involving 306,620 participants. Stroke; a journal of cerebral circulation, 36(7), pp.1360–5.

Feigin, V.L. et al., 2005. Risk factors for subarachnoid hemorrhage: an updated systematic review of epidemiological studies. Stroke; a journal of cerebral circulation, 36(12), pp.2773–80.

Feigin, V.L. et al., 2009. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. Lancet neurology, 8(4), pp.355–69.

Flicek, P. et al., 2008. Ensembl 2008. Nucleic acids research, 36(Database issue), pp.D707–14.

Foroud, T. et al., 2012. Genome-Wide Association Study of Intracranial Aneurysms Confirms Role of Anril and SOX17 in Disease Risk. Stroke, 43(11), pp.2846–2852.

Fossati, G. et al., 2002. Differential role of neutrophil Fcgamma receptor IIIB (CD16) in phagocytosis, bacterial killing, and responses to immune complexes. Arthritis and Rheumatism, 46(5), pp.1351–1361.

Frantz, S., Ertl, G. & Bauersachs, J., 2007. Mechanisms of disease: Toll-like receptors in cardiovascular disease. Nature clinical practice.Cardiovascular medicine, 4(8), pp.444–454.

Frazer, K.A. et al., 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature, 449(7164), pp.851–61.

Friedman, J.A. et al., 2001. Small cerebral aneurysms presenting with symptoms other than rupture. Neurology, 57(7), pp.1212–1216.

Frösen, J. et al., 2006. Growth factor receptor expression and remodeling of saccular cerebral artery aneurysm walls: implications for biological therapy preventing rupture. Neurosurgery, 58(3), pp.534–541.

Frösen, J. et al., 2004. Remodeling of saccular cerebral artery aneurysm wall is associated with rupture: histological analysis of 24 unruptured and 42 ruptured cases. Stroke, 35(10), pp.2287–2293.

Frösen, J. et al., 2012. Saccular intracranial aneurysm: pathology and mechanisms. Acta Neuropathologica.

Fu, X. et al., 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC genomics, 10, p.161.

Gaál, E.I. et al., 2012. Intracranial aneurysm risk locus 5q23.2 is associated with elevated systolic blood pressure. PLoS Genetics, 8(3), p.e1002563.

Garrod, A.E., 1902. The incidence of alkaptonuria: A study in chemical individuality. Lancet, 160(4137), pp.1616–1620.

Gentleman, R.C. et al., 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology, 5, p.R80.

Gentry, M.S., Romá-Mateo, C. & Sanz, P., 2013. Laforin, a protein with many faces: glucan phosphatase, adapter protein, et alii. The FEBS Journal, 280(2), pp.525–37.

Giacconi, R. et al., 2008. Pro-inflammatory genetic background and zinc status in old atherosclerotic subjects. Ageing research reviews, 7(4), pp.306–318.

Van Gijn, J., Kerr, R.S. & Rinkel, G.J.E., 2007. Subarachnoid haemorrhage. Lancet, 369(9558), pp.306–18.

Glazko, G. V & Emmert-Streib, F., 2009. Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics (Oxford, England), 25(18), pp.2348–54.

Gorelik, M. et al., 2012. Plasma follistatin-like protein 1 is elevated in Kawasaki disease and may predict coronary artery aneurysm formation. The Journal of Pediatrics, 161(1), pp.116–9.

Guenther, M.G. et al., 2007. A chromatin landmark and transcription initiation at most promoters in human cells. Cell, 130(1), pp.77–88.

Guil, S. & Esteller, M., 2012. Cis-acting noncoding RNAs: friends and foes. Nature structural & molecular biology, 19(11), pp.1068–75.

Günther, V., Lindert, U. & Schaffner, W., 2012. The taste of heavy metals: gene regulation by MTF-1. Biochimica et biophysica acta, 1823(9), pp.1416–25.

Guo, Y. et al., 2012. A genome-wide linkage and association scan reveals novel loci for hypertension and blood pressure traits. PloS One, 7(2), p.e31489.

Guzik, T.J. et al., 2006. Coronary artery superoxide production and nox isoform expression in human coronary artery disease. Arteriosclerosis, Thrombosis, and Vascular Biology, 26(2), pp.333–339.

Hägg, S. et al., 2009. Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: the Stockholm Atherosclerosis Gene Expression (STAGE) study. PLoS genetics, 5(12), p.e1000754.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M., 2002a. Clustering validity checking methods: Part I. ACM Special Interest Group on Management of Data Record, 31(2), pp.40–45.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M., 2002b. Clustering validity checking methods: part II. ACM Special Interest Group on Management of Data Record, 31(3), pp.19–27.

Hannenhalli, S., 2008. Eukaryotic transcription factor binding sites--modeling and integrative search methods. Bioinformatics (Oxford, England), 24(11), pp.1325–1331.

Hansen, K.D., Irizarry, R. a & Wu, Z., 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics (Oxford, England), 13(2), pp.204–16.

Hardison, R.C. & Taylor, J., 2012. Genomic approaches towards finding cis-regulatory modules in animals. Nature Reviews Genetics, 13(7), pp.469–483.

Harris, D., 2012. As genomics data approaches exascale, cloud could save the day. Available at: http://gigaom.com/2012/01/23/as-genomics-pushes-big-data-limits-cloud-could-save-the-day/ [Accessed August 13, 2013].

Hastie, T., Tibshirani, R. & Friedman, J., 2008. The Elements of Statistical Learning. Data Mining, Inference, and Prediction 2nd editio., Springer.

Helgadottir, A. et al., 2008. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nature Genetics, 40(2), pp.217–24.

Hiltunen, M.O. et al., 2002. Changes in gene expression in atherosclerotic plaques analyzed using DNA array. Atherosclerosis, 165(1), pp.23–32.

Hindorff, L. et al., A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies [Accessed June 25, 2013].

Hindorff, L. a et al., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America, 106(23), pp.9362–7.

Hirschhorn, J.N. et al., 2002. A comprehensive review of genetic association studies. Genetics in medicine◎: official journal of the American College of Medical Genetics, 4(2), pp.45–61.

Ho Sui, S.J. et al., 2005. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic acids research, 33(10), pp.3154–3164.

Hop, J.W. et al., 1997. Case-fatality rates and functional outcome after subarachnoid hemorrhage: a systematic review. Stroke; a journal of cerebral circulation, 28(3), pp.660–664.

Howie, B. et al., 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics, 44(8), pp.955–9.

Howie, B.N., Donnelly, P. & Marchini, J., 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics, 5(6), p.e1000529.

Hsieh, C.L., 1994. Dependence of transcriptional repression on CpG methylation density. Molecular and cellular biology, 14(8), pp.5487–94.

Hu, X.-Q. & Zhang, L., 2012. Function and regulation of large conductance Ca(2+)-activated K+ channel in vascular smooth muscle cells. Drug discovery today, 17(17-18), pp.974–87.

Huang, D.W., Sherman, B.T. & Lempicki, R. a, 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37(1), pp.1–13.

Huang da, W., Sherman, B.T. & Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), pp.44–57.

Huang, J., Shimizu, H. & Shioya, S., 2003. Clustering gene expression pattern and extracting relationship in gene network based on artificial neural networks. Journal of bioscience and bioengineering, 96(5), pp.421–428.

Huber, B.R. & Bulyk, M.L., 2006. Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. BMC bioinformatics, 7, p.229.

Huttunen, T. et al., 2010. Saccular intracranial aneurysm disease: distribution of site, size, and age suggests different etiologies for aneurysm formation and rupture in 316 familial and 1454 sporadic eastern Finnish patients. Neurosurgery, 66(4), pp.631–8; discussion 638.

Huyghe, J.R. et al., 2013. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nature genetics, 45(2), pp.197–201.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature, 431(7011), pp.931–45.

Irizarry, R. a et al., 2009. Gene Set Enrichment Analysis Made Simple. Statistical Methods In Medical Research, 18(6), pp.565–575.

Jakobsdottir, J. et al., 2009. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS genetics, 5(2), p.e1000337.

Jelier, R. et al., 2011. Literature-aided interpretation of gene expression data with the weighted global test. Briefings in bioinformatics, 12(5), pp.518–29.

Johnson, A.D. et al., 2013. Resequencing and clinical associations of the 9p21.3 region: a comprehensive investigation in the framingham heart study. Circulation, 127(7), pp.799–810.

Jonsson, T. et al., 2013. Variant of TREM2 associated with the risk of Alzheimer's disease. The New England Journal of Medicine, 368(2), pp.107–16.

Jostins, L. et al., 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature, 491(7422), pp.119–24.

Juvela, S., 2000. Risk factors for multiple intracranial aneurysms. Stroke; a journal of cerebral circulation, 31(2), pp.392–7.

Kanehisa, M. et al., 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research, 40(Database issue), pp.D109–14.

Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), pp.27–30.

Kang, Y.J., 2006. Metallothionein Redox Cycle and Function. Experimental Biology and Medicine, 231, pp.1459–1467.

Kankainen, M. et al., 2006. POXO: a web-enabled tool series to discover transcription factor binding sites. Nucleic acids research, 34(Web Server issue), pp.W534–40.

Karamanakos, P.N. et al., 2012. Risk factors for three phases of 12-month mortality in 1657 patients from a defined population after acute aneurysmal subarachnoid hemorrhage. World Neurosurgery, 78(6), pp.631–9.

Kataoka, K. et al., 1999. Structural fragility and inflammatory response of ruptured cerebral aneurysms. A comparative study between ruptured and unruptured cerebral aneurysms. Stroke; a journal of cerebral circulation, 30(7), pp.1396–1401.

Kerr, G. et al., 2008. Techniques for clustering gene expression data. Computers in biology and medicine, 38(3), pp.283–93.

Kettritz, R., 2012. How anti-neutrophil cytoplasmic autoantibodies activate neutrophils. Clinical and experimental immunology, 169(3), pp.220–8.

Khatri, P., Sirota, M. & Butte, A.J., 2012. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS computational biology, 8(2), p.e1002375.

Kiemeney, L.A. et al., 2008. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nature Genetics, 40(11), pp.1307–12.

Kiezun, A. et al., 2012. Exome sequencing and the genetic basis of complex traits. Nature genetics, 44(6), pp.623–30.

Kilpivaara, O. & Aaltonen, L. a., 2013. Diagnostic Cancer Genome Sequencing and the Contribution of Germline Variants. Science, 339(6127), pp.1559–1562.

Koivisto, U.M. et al., 1994. A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia. Proceedings of the National Academy of Sciences of the United States of America, 91(22), pp.10526–30.

Kong, S.W., Pu, W.T. & Park, P.J., 2006. A Multivariate Approach for Integrating Genome-wide Expression Data and Biological Knowledge. Bioinformatics (Oxford, England), 22(19), pp.2373–2380.

Korja, M. et al., 2010. Genetic epidemiology of spontaneous subarachnoid hemorrhage: Nordic Twin Study. Stroke; a journal of cerebral circulation, 41(11), pp.2458–62.

Krischek, B. et al., 2008. Network-based gene expression analysis of intracranial aneurysm tissue reveals role of antigen presenting cells. Neuroscience, 154(4), pp.1398–1407.

Ku, C.S. et al., 2011. Studying the epigenome using next generation sequencing. Journal of Medical Genetics, 48(11), pp.721–30.

Kurki, M.I. et al., 2011. Upregulated signaling pathways in ruptured human saccular intracranial aneurysm wall: an emerging regulative role of Toll like receptor signaling and NF-κB, HIF1A and ETS transcription factors. Neurosurgery, 68(6), pp.1667–1676.

Kwong, R., Lupton, M.K. & Janitz, M., 2012. Single-cell and regional gene expression analysis in Alzheimer's disease. Cellular and molecular neurobiology, 32(4), pp.477–89.

Laaksamo, E. et al., 2008. Involvement of mitogen-activated protein kinase signaling in growth and rupture of human intracranial aneurysms. Stroke; a journal of cerebral circulation, 39(3), pp.886–892.

Laaksamo, E. et al., 2013. Oxidative stress is associated with cell death, wall degradation, and increased risk of rupture of the intracranial aneurysm wall. Neurosurgery, 72(1), pp.109–17.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. Nature, 409(6822), pp.860–921.

Lappalainen, T. et al., 2013. Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501(7468), pp.506–511.

Lara-Pezzi, E. et al., 2008. Expression of follistatin-related genes is altered in heart failure. Endocrinology, 149(11), pp.5822–7.

Lazrak, M. et al., 2004. The bHLH TAL-1/SCL regulates endothelial cell migration and morphogenesis. Journal of cell science, 117(Pt 7), pp.1161–1171.

Lee, D.D. & Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), pp.788–791.

Lee, Hyung-Jin et al., 2013. Dysregulated Expression Profiles of MicroRNAs of Experimentally Induced Cerebral Aneurysms in Rats. Journal of Korean Neurosurgical Society, 53(2), pp.72–6.

Lee, Y. et al., 2009. Activation of toll-like receptors 2, 3 or 5 induces matrix metalloproteinase-1 and -9 expression with the involvement of MAPKs and NF-kappaB in human epidermal keratinocytes. Experimental dermatology.

Leek, J.T. et al., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature reviews. Genetics, 11(10), pp.733–9.

Lenhard, B., Sandelin, A. & Carninci, P., 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature reviews. Genetics, 13(4), pp.233–245.

Lerner-ellis, J.P. et al., 2008. Gene Identification for the cblD Defect of Vitamin B12 Metabolism. The New England Journal of Medicine, 358(14), pp.1454–1464.

Levy, S. et al., 2007. The diploid genome sequence of an individual human. PLoS biology, 5(10), p.e254.

Li, E., Beard, C. & Jaenisch, R., 1993. Role for DNA methylation in genomic imprinting. Nature, 366(6453), pp.362–5.

Li, J., Jiang, H. & Wong, W.H., 2010. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome biology, 11(5), p.R50.

Li, L. et al., 2009. Transcriptome-wide characterization of gene expression associated with unruptured intracranial aneurysms. European neurology, 62(6), pp.330–337.

Liu, L. et al., 2012. Radil controls neutrophil adhesion and motility through β2-integrin activation. Molecular Biology of the Cell.

Liu, T. et al., 2009. Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments. BMC bioinformatics, 10, p.146.

Lorenzen, J.M., Martino, F. & Thum, T., 2012. Epigenetic modifications in cardiovascular disease. Basic research in cardiology, 107(2), p.245.

Low, S.-K. et al., 2012. Genome-wide association study for intracranial aneurysm in the Japanese population identifies three candidate susceptible loci and a functional genetic variant at EDNRA. Human molecular genetics, 21(9), pp.2102–10.

Luke, M.M. et al., 2009. Gene variants associated with ischemic stroke: the cardiovascular health study. Stroke, 40(2), pp.363–8.

Lupski, J.R. et al., 2011. Clan genomics and the complex architecture of human disease. Cell, 147(1), pp.32–43.

Lutgens, S.P. et al., 2007. Cathepsin cysteine proteases in cardiovascular disease. The FASEB journal®: official publication of the Federation of American Societies for Experimental Biology, 21(12), pp.3029–3041.

Mackey, J., 2012. Unruptured intracranial aneurysms in the Familial Intracranial Aneurysm and International Study of Unruptured Intracranial Aneurysms cohorts: differences in multiplicity and location. Journal of Neurosurgery, 117(1), p.192.

Maher, B., 2008. Personal genomes: The case of the missing heritability. Nature, 456(7218), pp.18–21.

Makowsky, R. et al., 2011. Beyond missing heritability: prediction of complex traits. PLoS genetics, 7(4), p.e1002051.

Maller, J. et al., 2006. Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. Nature genetics, 38(9), pp.1055–9.

Malone, J.H. & Oliver, B., 2011. Microarrays, deep sequencing and the true measure of the transcriptome. BMC biology, 9, p.34.

Manolio, T.A. et al., 2010. Finding the missing heritability of complex diseases. Nature, 461(7265), pp.747–753.

Marchese, E. et al., 2010. Comparative evaluation of genome-wide gene expression profiles in ruptured and unruptured human intracranial aneurysms. Journal of Biological Regulators and Homeostatic Agents, 24(2), pp.185–195.

Marchini, J. & Howie, B., 2010. Genotype imputation for genome-wide association studies. Nature reviews. Genetics, 11(7), pp.499–511.

Maret, W. & Krężel, A., 2007. Cellular Zinc and Redox Buffering Capacity of Metallothionein / Thionein in Health and Disease. Molecular Medicine, 13, pp.371–375.

Margulies, K.B., Bednarik, D.P. & Dries, D.L., 2009. Genomics, transcriptional profiling, and heart failure. Journal of the American College of Cardiology, 53(19), pp.1752–9.

Marian, A.J., 2012. Elements of "missing heritability". Current opinion in cardiology, 27(3), pp.197–201.

Martin, D. et al., 2004. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome biology, 5(12), p.R101.

Maston, G. a, Evans, S.K. & Green, M.R., 2006. Transcriptional regulatory elements in the human genome. Annual review of genomics and human genetics, 7, pp.29–59.

Matys, V., 2003. TRANSFAC(R): transcriptional regulation, from patterns to profiles. Nucleic Acids Research, 31(1), pp.374–378.

McCulloch, S.D. & Kunkel, T. a, 2008. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell research, 18(1), pp.148–61.

McGettrick, A.F. & O'Neill, L.A., 2007. Toll-like receptors: key activators of leucocytes and regulator of haematopoiesis. British Journal of Haematology, 139(2), pp.185–193.

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, M., Online Mendelian Inheritance in Man, OMIM®. Available at: http://omim.org/ [Accessed October 25, 2012].

Medina, M.A., 2013. Systems biology for molecular life sciences and its impact in biomedicine. Cellular and molecular life sciences◎: CMLS, 70(6), pp.1035–53.

Mendel, G., 1866. Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereins Brünn.

Metzker, M.L., 2010. Sequencing technologies - the next generation. Nature reviews. Genetics, 11(1), pp.31–46.

Mhurchu, C.N. et al., 2001. Hormonal Factors and Risk of Aneurysmal Subarachnoid Hemorrhage◎: An International Population-Based, Case-Control Study Editorial Comment◎: The Gender Gap in Aneurysmal Subarachnoid Hemorrhage. Stroke, 32(3), pp.606–612.

Minassian, B.A. et al., 1998. Mutations in a gene encoding a novel protein tyrosine phosphatase cause progressive myoclonus epilepsy. Nature Genetics, 20(2), pp.171–4.

Miyake, T. et al., 2007. Regression of abdominal aortic aneurysms by simultaneous inhibition of nuclear factor kappaB and ets in a rabbit model. Circulation research, 101(11), pp.1175–1184.

Mootha, V.K. et al., 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics, 34(3), pp.267–273.

Morita, A. et al., 2012. The natural course of unruptured cerebral aneurysms in a Japanese cohort. The New England journal of medicine, 366(26), pp.2474–82.

Movassagh, M. et al., 2011. Distinct epigenomic features in end-stage failing human hearts. Circulation, 124(22), pp.2411–22.

Müller, T.B. et al., 2013. Unruptured Intracranial Aneurysms in the Norwegian HUNT-study: Risk of Rupture Calculated from Data in a Population-based Cohort Study. Neurosurgery.

Murakami, K. et al., 2012. Follistatin-related protein/follistatin-like 1 evokes an innate immune response via CD14 and toll-like receptor 4. FEBS Letters, 586(4), pp.319–24.

Musunuru, K. et al., 2010. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature, 466(7307), pp.714–9.

Nahed, B. V et al., 2007. Genetics of intracranial aneurysms. Neurosurgery, 60(2), pp.213–25; discussion 225–6.

Nam, D. & Kim, S.-Y., 2008. Gene-set approach for expression pattern analysis. Briefings in bioinformatics, 9(3), pp.189–97.

Neale, B. et al. eds., 2007. Statistical Genetics: Gene Mapping Through Linkage and Association, Taylor and Francis, London.

Ni, W. et al., 2007. Ets-1 is a critical transcriptional regulator of reactive oxygen species and p47(phox) gene expression in response to angiotensin II. Circulation research, 101(10), pp.985–994.

Nicolae, D.L. et al., 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS genetics, 6(4), p.e1000888.

Nieuwkamp, D.J. et al., 2009. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. Lancet neurology, 8(7), pp.635–42.

Nigro, P., Abe, J.-I. & Berk, B.C., 2011. Flow shear stress and atherosclerosis: a matter of site specificity. Antioxidants & Redox Signaling, 15(5), pp.1405–14.

Nischan, J. et al., 2009. Binding sites for ETS family of transcription factors dominate the promoter regions of differentially expressed genes in abdominal aortic aneurysms. Circulation.Cardiovascular genetics, 2(6), pp.565–572.

Nizet, V. & Johnson, R.S., 2009. Interdependence of hypoxic and innate immune responses. Nature reviews.Immunology, 9(9), pp.609–617.

Noonan, J.P. & Mccallion, A.S., 2010. Genomics of Long-Range Regulatory Elements. Annual review of genomics and human genetics, 11, pp.11–23.

Nuki, Y. et al., 2009. Elastase-induced intracranial aneurysms in hypertensive mice. Hypertension, 54(6), pp.1337–44.

O'Neill, L.A., Bryant, C.E. & Doyle, S.L., 2009. Therapeutic targeting of toll-like receptors for infectious and inflammatory diseases and cancer. Pharmacological reviews, 61(2), pp.177–197.

O'Roak, B.J. et al., 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature genetics, 43(6), pp.585–9.

Oettgen, P., 2006. Regulation of vascular inflammation and remodeling by ETS factors. Circulation research, 99(11), pp.1159–1166.

Osborne, J.D. et al., 2009. Annotating the human genome with Disease Ontology. BMC genomics, 10 Suppl 1, p.S6.

Panning, B. & Jaenisch, R., 1996. DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes & Development, 10(16), pp.1991–2002.

Pehkonen, P., Wong, G. & Toronen, P., 2005. Theme discovery from gene lists for identification and viewing of multiple functional groups. BMC bioinformatics, 6, p.162.

Peltonen, L., Jalanko, a & Varilo, T., 1999. Molecular genetics of the Finnish disease heritage. Human molecular genetics, 8(10), pp.1913–23.

Pera, J. et al., 2010. Gene expression profiles in human ruptured and unruptured intracranial aneurysms: what is the role of inflammation? Stroke; a journal of cerebral circulation, 41(2), pp.224–231.

Pirson, Y., Chauveau, D. & Torres, V., 2002. Management of cerebral aneurysms in autosomal dominant polycystic kidney disease. Journal of the American Society of Nephrology®: JASN, 13(1), pp.269–76.

Platt, M.O., Ankeny, R.F. & Jo, H., 2006. Laminar shear stress inhibits cathepsin L activity in endothelial cells. Arteriosclerosis, Thrombosis, and Vascular Biology, 26(8), pp.1784–1790.

Pritchard, J.K. & Przeworski, M., 2001. Linkage disequilibrium in humans: models and data. American Journal of Human Genetics, 69(1), pp.1–14.

Pruim, R.J. et al., 2010. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics (Oxford, England), 26(18), pp.2336–7.

Purcell, S. et al., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics, 81(3), pp.559–575.

Quiat, D. & Olson, E.N., 2013. MicroRNAs in cardiovascular disease: from pathogenesis to prevention and treatment. The Journal of clinical investigation, 123(1), pp.11–8.

Qureshi, A.I. et al., 1998. Risk Factors for Multiple Intracranial Aneurysms. Neurosurgery, 43(1), pp.22–26.

R Development Core Team, R., 2009. R: A Language and Environment for Statistical Computing R. D. C. Team, ed. R Foundation for Statistical Computing, 1(2.11.1).

Raffetto, J.D. & Khalil, R.A., 2008. Matrix metalloproteinases and their inhibitors in vascular remodeling and vascular disease. Biochemical pharmacology, 75(2), pp.346–359.

Rahnenführer, J. et al., 2004. Calculating the statistical significance of changes in pathway activity from gene expression data. Statistical applications in genetics and molecular biology, 3, p.Article16.

Raitakari, O.T. et al., 2008. Cohort profile: the cardiovascular risk in Young Finns Study. International Journal of Epidemiology, 37(6), pp.1220–6.

Rakyan, V.K. et al., 2011. Epigenome-wide association studies for common human diseases. Nature reviews. Genetics, 12(8), pp.529–41.

Ramanan, V.K. et al., 2012. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends in genetics⊚: TIG, 28(7), pp.323–32.

Rasinperä, H. et al., 2005. Transcriptional downregulation of the lactase (LCT) gene during childhood. Gut, 54(11), pp.1660–1.

Regalado, E.S. et al., 2011. Exome sequencing identifies SMAD3 mutations as a cause of familial thoracic aortic aneurysm and dissection with intracranial and other arterial aneurysms. Circulation research, 109(6), pp.680–6.

Reichel, C.A. et al., 2006. Chemokine receptors Ccr1, Ccr2, and Ccr5 mediate neutrophil migration to postischemic tissue. Journal of leukocyte biology, 79(1), pp.114–122.

Rhee, S.Y. et al., 2008. Use and misuse of the gene ontology annotations. Nature reviews.Genetics, 9(7), pp.509–515.

Rinkel, G.J. et al., 1998. Prevalence and risk of rupture of intracranial aneurysms: a systematic review. Stroke; a journal of cerebral circulation, 29(1), pp.251–256.

Roberts, A. et al., 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome biology, 12(3), p.R22.

Robertson, G. et al., 2006. cisRED: a database system for genome-scale computational discovery of regulatory elements. Nucleic acids research, 34(Database issue), pp.D68–73.

Robinson, P.N. & Mundlos, S., 2010. The human phenotype ontology. Clinical genetics, 77(6), pp.525–34.

Rommel, C., Camps, M. & Ji, H., 2007. PI3Kδ and PI3Kγ: partners in crime in inflammation in rheumatoid arthritis and beyond?7, 473-483 (2006).-->. Nature reviews.Immunology.

Ronkainen, A. et al., 1997. Familial intracranial aneurysms. Lancet, 349(9049), pp.380–4.

Ronkainen, A. et al., 1998. Risk of harboring an unruptured intracranial aneurysm. Stroke; a journal of cerebral circulation, 29(2), pp.359–62.

De Rooij, N.K. et al., 2007. Incidence of subarachnoid haemorrhage: a systematic review with emphasis on region, age, gender and time trends. Journal of Neurology, Neurosurgery, and Psychiatry, 78(12), pp.1365–72.

Ross, D. & Armen, B., 2007. An Unhealthy America: The Economic Burden of Chronic Disease -- Charting a New Course to Save Lives and Increase Productivity and Economic Growth,

Rudini, N. et al., 2008. VE-cadherin is a critical endothelial regulator of TGF-beta signalling. The EMBO journal, 27(7), pp.993–1004.

Ruigrok, Y.M. et al., 2004. Characteristics of intracranial aneurysms in patients with familial subarachnoid hemorrhage. Neurology, 62(6), pp.891–4.

Ruigrok, Y.M., Buskens, E. & Rinkel, G.J., 2001. Attributable risk of common and rare determinants of subarachnoid hemorrhage. Stroke, 32(5), pp.1173–1175.

Ruigrok, Y.M. & Rinkel, G.J., 2008. Genetics of Intracranial Aneurysms. Stroke; a journal of cerebral circulation, 39(3), pp.1049–1055.

Ruigrok, Y.M., Rinkel, G.J. & Wijmenga, C., 2005. Genetics of intracranial aneurysms. Lancet neurology, 4(3), pp.179–189.

Sabatti, C. et al., 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nature genetics, 41(1), pp.35–46.

Sandberg, R. & Larsson, O., 2007. Improved precision and accuracy for microarrays using updated probe set definitions. BMC bioinformatics, 8, p.48.

Saydam, N. et al., 2002. Regulation of metallothionein transcription by the metal-responsive transcription factor MTF-1: identification of signal transduction cascades that control metal-inducible transcription. The Journal of biological chemistry, 277(23), pp.20438–45.

Sboner, A. et al., 2011. The real cost of sequencing: higher than you think! Genome biology, 12(8), p.125.

Scanarini, M. et al., 1978. Histological and ultrastructural study of intracranial saccular aneurysmal wall. Acta neurochirurgica, 43(3-4), pp.171–82.

Schena, M. et al., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (New York, N.Y.), 270(5235), pp.467–470.

Schnabel, R.B. et al., 2012. Next steps in cardiovascular disease genomic research--sequencing, epigenetics, and transcriptomics. Clinical chemistry, 58(1), pp.113–26.

Seppänen, J.K., Bingham, E. & Mannila, H., 2003. A Simple Algorithm for Topic Identification in 0-1 Data. In N. Lavrac et al., eds. Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Springer-Verlag, pp. 423–434.

Shi, C. et al., 2009. Genomics of human intracranial aneurysm wall. Stroke; a journal of cerebral circulation, 40(4), pp.1252–1261.

Smith, B. et al., 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology, 25(11), pp.1251–5.

Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology, 3, p.Article3.

So, H.-C. et al., 2011. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genetic Epidemiology, 35(5), pp.310–7.

Spitz, F. & Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. Nature reviews. Genetics, 13(9), pp.613–26.

Sprague, A.H. & Khalil, R.A., 2009. Inflammatory cytokines in vascular dysfunction and vascular disease. Biochemical pharmacology, 78(6), pp.539–552.

Stefansson, B. et al., 2008. Protein phosphatase 6 regulatory subunits composed of ankyrin repeat domains. Biochemistry, 47(5), pp.1442–51.

Stefansson, B. & Brautigan, D.L., 2006. Protein phosphatase 6 subunit with conserved Sit4-associated protein domain targets IkappaBepsilon. The Journal of Biological Chemistry, 281(32), pp.22624–34.

Stefansson, B. & Brautigan, D.L., 2007. Protein phosphatase PP6 N terminal domain restricts G1 to S phase progression in human cancer cells. Cell Cycle, 6(11), pp.1386–92.

Stegmayr, B., Eriksson, M. & Asplund, K., 2004. Declining mortality from subarachnoid hemorrhage: changes in incidence and case fatality from 1985 through 2000. Stroke; a journal of cerebral circulation, 35(9), pp.2059–63.

Stenmark, K.R. et al., 2013. The adventitia: essential regulator of vascular wall structure and function. Annual review of physiology, 75, pp.23–47.

Stokes, L. et al., 2011. A loss-of-function polymorphism in the human P2X4 receptor is associated with increased pulse pressure. Hypertension, 58(6), pp.1086–92.

Subramanian, A. et al., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), pp.15545–15550.

Sulem, P. et al., 2011. Identification of low-frequency variants associated with gout and serum uric acid levels. Nature Genetics, 43(11), pp.1127–30.

Tanaka, S. et al., 2003. Fc gamma RIIIb allele-sensitive release of alpha-defensins: anti-neutrophil cytoplasmic antibody-induced release of chemotaxins. Journal of immunology (Baltimore, Md.: 1950), 171(11), pp.6090–6096.

Tarca, A.L. et al., 2009. A novel signaling pathway impact analysis. Bioinformatics (Oxford, England), 25(1), pp.75–82.

Teare, D.M. & Barrett, J.H., 2005. Genetic Epidemiology 2◎: Genetic linkage studies. Lancet, 366(9490), pp.1036–44.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature, 467(7319), pp.1061–73.

The 1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes. Nature, 135(V), pp.0–9.

The International HapMap Consortium, 2005. A haplotype map of the human genome. Nature, 437(7063), pp.1299–320.

THL - National Institute for Health and Welfare., 2000. Health (2000). Available at: http://www.terveys2000.fi/indexe.html [Accessed January 22, 2013].

Tian, L. et al., 2005. Discovering statistically significant pathways in expression profiling studies. Proceedings of the National Academy of Sciences of the United States of America, 102(38), pp.13544–9.

Tomohiro Aoki, M.D. et al., 2007. NF-B Is a Key Mediator of Cerebral Aneurysm Formation. Circulation, 116, p.2830.

Torres, V.E., Harris, P.C. & Pirson, Y., 2007. Autosomal dominant polycystic kidney disease. Lancet, 369(9569), pp.1287–301.

Trapnell, C. et al., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols, 7(3), pp.562–78.

Tulamo, R. et al., 2006. Complement activation associates with saccular cerebral artery aneurysm wall degeneration and rupture. Neurosurgery, 59(5), pp.1067–1069.

Vartiainen, E. et al., 2010. Thirty-five-year trends in cardiovascular risk factors in Finland. International Journal of Epidemiology, 39(2), pp.504–18.

Veer, L.J. Van et al., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(345), pp.530–536.

Venables, W.N. & Ripley, B.D., 2002. Modern Applied Statistics with S Fourth., New York: Springer.

Venter, J.C. et al., 2001. The sequence of the human genome. Science (New York, N.Y.), 291(5507), pp.1304–51.

Vink, A. et al., 2007. HIF-1 alpha expression is associated with an atheromatous inflammatory plaque phenotype and upregulated in activated macrophages. Atherosclerosis, 195(2), pp.e69–75.

Visscher, P.M. et al., 2012. Five years of GWAS discovery. American journal of human genetics, 90(1), pp.7–24.

Vlak, M.H. et al., 2011. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. Lancet Neurology, 10(7), pp.626–636.

Vogel, C. & Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature reviews. Genetics, 13(4), pp.227–32.

Voight, B.F. et al., 2012. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS genetics, 8(8), p.e1002793.

Vuong, Q.H., 1989. LIkelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica, 57(2), pp.307–333.

Walsh, C.P., Chaillet, J.R. & Bestor, T.H., 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. , 20(2), pp.116–117.

Wang, E.T. et al., 2008. Alternative Isoform Regulation in Human Tissue Transcriptomes. Nature, 456(7221), pp.470–476.

Wang, K., Li, M. & Hakonarson, H., 2010. Analysing biological pathways in genome-wide association studies. Nature Reviews. Genetics, 11(12), pp.843–54.

Wang, L. et al., 2003. Mutation of MEF2A in an Inherited Disorder with Features of Coronary Artery Disease. Science, 302(5650), pp.1578–1581.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics, 10(1), pp.57–63.

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 Research, 40(Database issue), pp.D930–4.

Wasserman, W.W. & Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. Nature reviews. Genetics, 5(4), pp.276–87.

Wellcome, T., Case, T. & Consortium, C., 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447(7145), pp.661–78.

Wermer, M.J.H. et al., 2007. Risk of rupture of unruptured intracranial aneurysms in relation to patient and aneurysm characteristics: an updated meta-analysis. Stroke; a journal of cerebral circulation, 38(4), pp.1404–10.

Wetterstrand, K., 2013. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcosts [Accessed June 13, 2013].

Wetzels, J.F.M. et al., 2007. Age- and gender-specific reference values of estimated GFR in Caucasians: the Nijmegen Biomedical Study. Kidney International, 72(5), pp.632–7.

Wingender, E. et al., 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic acids research, 24(1), pp.238–241.

Wray, N.R., 2008. Estimating Trait Heritability. Nature education, 1(1).

Wu, X. & Brewer, G., 2012. The regulation of mRNA stability in mammalian cells: 2.0. Gene, 500(1), pp.10–21.

Wu, Z. et al., 2004. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association, 99(9), p.909.

Xu, H.W. et al., 2011. Screening for intracranial aneurysm in 355 patients with autosomal-dominant polycystic kidney disease. Stroke; a journal of cerebral circulation, 42(1), pp.204–6.

Yang, J. et al., 2011. Genome partitioning of genetic variation for complex traits using common SNPs. Nature genetics, 43(6), pp.519–25.

Yasuno, K. et al., 2011. Common variant near the endothelin receptor type A (EDNRA) gene is associated with intracranial aneurysm risk. Proceedings of the National Academy of Sciences of the United States of America.

Yasuno, K. et al., 2010. Genome-wide association study of intracranial aneurysm identifies three new risk loci. Nature Genetics, 42(5), pp.420–5.

Yu, X. et al., 2010. FcγRIIA and FcγRIIIB are required for autoantibody-induced tissue damage in experimental human models of bullous pemphigoid. The Journal of investigative dermatology, 130(12), pp.2841–4.

Zambon, A.C. et al., 2005. Gene expression patterns define key transcriptional events in cell-cycle regulation by cAMP and protein kinase A. Proceedings of the National Academy of Sciences of the United States of America, 102(24), pp.8561–8566.

Zeeberg, B.R. et al., 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. , 4(4), p.R28.

Zhan, Y. et al., 2005. Ets-1 is a critical transcriptional regulator of reactive oxygen species and p47(phox) gene expression in response to angiotensin II. The Journal of clinical investigation, 115(9), pp.2508–2516.

Zhang, Y. et al., 2010. Identification and characterization of human LYPD6, a new member of the Ly-6 superfamily. Molecular Biology Reports, 37(4), pp.2055–62.

Zhang, Y. et al., 2009. Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1. Nucleic acids research, 37(21), pp.7024–38.

Zhao, W., Zhao, S. & Peng, D., 2012. The effects of myocyte enhancer factor 2A gene on the proliferation, migration and phenotype of vascular smooth muscle cells. Cell biochemistry and function, 30(2), pp.108–13.

Zhong, H. et al., 2010. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. American journal of human genetics, 86(4), pp.581–91.

Mitja Kurki

Genomics and Bioinformatics Approaches in Search of Molecular Pathomechanisms of Saccular Intracranial Aneurysm, A Complex Disease

High-throughput genomic methods generate huge amounts data that pose challenges for analysis and interpretation, requiring bioinformatic methods. In this thesis, to study molecular mechanisms of saccular intracranial aneurysm (sIA) disease, two complementary approaches utilizing high-throughput genomics and bioinformatics were applied. Additionally, a bioinformatic method and software was developed. Candidate genes and pathways were identified, which can serve as a basis for future research aiming to novel diagnostics, preventions, or therapies of sIA disease.

UNIVERSITY OF
EASTERN FINLAND

Publications of the University of Eastern Finland
Dissertations in Health Sciences

ISBN 978-952-61-1310-4