From Department of Medicine
Karolinska Institutet, Stockholm, Sweden

PHENOTYPE AND GENOTYPE EFFECTS ON THE TRANSCRIPTOME IN CARDIOVASCULAR DISEASE - TOOLS TO IDENTIFY CANDIDATE GENES

Lasse Folkersen

Karolinska Institutet

Stockholm 2011

Supervisors:
Dr. Anders Gabrielsen
Experimental Cardiovascular Research
Department of Medicine
Karolinska Institutet, Stockholm

Prof. Per Eriksson
Cardiovascular Genetics and Genomics
Department of Medicine
Karolinska Institutet, Stockholm

Dr. Gabrielle Paulsson-Berne
Experimental Cardiovascular Research
Department of Medicine
Karolinska Institutet, Stockholm

Faculty opponent:
Professor Gerard Pasterkamp
Division Heart and Lungs
University medical center, Utrecht

Thesis committee:
Dr. Ingrid Kockum
Department of Clinical Neuroscience
Karolinska Institutet, Stockholm

Dr. Rickard Sandberg
Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm

Dr. Isabel Gonçalves
Experimental Cardiovascular Research Unit
Lunds Universitet, Malmö

Published by Karolinska Institutet. Printed by Larserics Digital Print AB, Landsvägen 65, 17265 Sundbyberg

Cover image: 'An eQTL', R-script and oil colour on canvas
(R-script available at www.folkersen.com/dnaplotting)

ISBN 978-91-7457-584-2

SHORT INTRODUCTION

The overarching purpose of this thesis is to investigate the expression of human genes and how they relate to cardiovascular disease. Consequently, the ultimate goal is to benefit the patients who are suffering from cardiovascular disease. The only way to improve our treatment of disease is by advancing our knowledge about it. Already, this knowledge is vast and expanding and the work presented herein is therefore only a small piece of a very large context of research, all aimed at this same goal.

The context is that of the medical and biological science of 2011, at a time when a genomic era of medicine has already been declared several times, but where the actual impact of genomics on medicine is lacking. It is a time when headlines routinely report the identification of the genes for one disease or another, but where it is harder to suggest direct practical uses for these new findings. One suggestion that is hard to dispute, however, is that the findings increase our knowledge of disease and that we should strive to translate them into clinical application. Translational medical research is the catchphrase applied to this suggestion.

Since the discovery of the genomic code in the 1960s, it has been known that biological information flows from the genome into practical form and function as proteins. Gene expression is the intermediate of this path and any genomic concept is likely to be mediated through transcription in one way or another. And so, taking cue from the cell itself, the proper subject for translating genomic era findings to the clinical application is to study the products of the genes.

The pieces of the puzzle provided by this thesis all concern gene expression and they all concern translation into medical application. Each of the five papers included investigate different aspects of this focus point. One paper investigates the technology for extracting gene expression information (I). Two look from the genome towards the gene expression in order to interpret genomics better (II and III). Another paper looks from disease towards gene expression to seek clues on the mechanism of disease (IV). And finally the last paper observes the gene expression, irrespective of its biological meaning, and asks how much it can tell about the future of a patient (V).

LIST OF PUBLICATIONS

This thesis is based on the following papers, which are referred to by their Roman numbers in the text

I Folkersen L, Diez D, Wheelock CE, Haeggström JZ, Goto S, Eriksson P, Gabrielsen A. (2009) GeneRegionScan: a Bioconductor package for probe-level analysis of specific, small regions of the genome. Bioinformatics. Aug 1;25(15):1978-9. Epub 2009 Apr 27.

II Folkersen L*, Kyriakou T*, Goel A, Peden J, Mälarstig A, Paulsson-Berne G, Hamsten A, Hugh Watkins, Franco-Cereceda A, Gabrielsen A, Eriksson P; (2009) Relationship between CAD risk genotype in the chromosome 9p21 locus and gene expression. Identification of eight new ANRIL splice variants. PLoS One. Nov 2;4(11):e7677.

III Folkersen L, van't Hooft F, Chernogubova E, Agardh HE, Hansson GK, Hedin U, Liska J, Syvänen AC, Paulsson-Berne G, Franco-Cereceda A, Hamsten A, Gabrielsen A, Eriksson P; (2010) Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. Circ Cardiovasc Genet. Aug;3(4):365-73. Epub 2010 Jun 19.

IV Folkersen L, Wågsäter D, Paloschi V, Jackson V, Petrini J, Kurtovic S, Maleki S, Eriksson MJ, Caidahl K, Hamsten A, Michel JB, Liska J, Gabrielsen A, Franco-Cereceda A, Eriksson P. (2011) Unraveling the divergent gene expression profiles in bicuspid and tricuspid aortic valve patients with thoracic aortic dilatation - the ASAP study. Mol Med. Sep 27. doi: 10.2119/molmed.2011.00286. [Epub ahead of print]

V Lasse Folkersen, Jonas Persson, Johan Ekstrand, Hanna E Agardh, Göran K Hansson, Anders Gabrielsen, Ulf Hedin, and Gabrielle Paulsson-Berne. Prediction of ischemic events based on transcriptomic and genomic profiling in patients undergoing carotid endarterectomy. Manuscript.

*equal contribution

OTHER RELATED PAPERS

These papers were co-authored during the PhD. They have been referred to in the main text as appropriate.

Ueland T, Otterdal K, Lekva T, Halvorsen B, Gabrielsen A, Sandberg WJ, Paulsson-Berne G, Pedersen TM, Folkersen L, Gullestad L, Oie E, Hansson GK, Aukrust P. (2009) Dickkopf-1 enhances inflammatory interaction between platelets and endothelial cells and shows increased expression in atherosclerosis. Arterioscler Thromb Vasc Biol. Aug;29(8):1228-34. Epub 2009 Jun 4.

Folkersen L, Kurtovic S, Razuvaev A, Agardh HE, Gabrielsen A, Paulsson-Berne G. (2009) Endogenous control genes in complex vascular tissue samples. BMC Genomics. Nov 10;10:516.

Breland UM, Michelsen AE, Skjelland M, Folkersen L, Krohg-Sørensen K, Russell D, Ueland T, Yndestad A, Paulsson-Berne G, Damås JK, Oie E, Hansson GK, Halvorsen B, Aukrust P. (2010) Raised MCP-4 levels in symptomatic carotid atherosclerosis: an inflammatory link between platelet and monocyte activation. Cardiovasc Res. May 1;86(2):265-73. Epub 2010 Feb 5.

Gabrielsen A, Qiu H, Bäck M, Hamberg M, Hemdahl AL, Agardh H, Folkersen L, Swedenborg J, Hedin U, Paulsson-Berne G, Haeggström JZ, Hansson GK. (2010) Thromboxane synthase expression and thromboxane A2 production in the atherosclerotic lesion. J Mol Med (Berl). Aug;88(8):795-806. Epub 2010 Apr 12.

Ekstrand J, Razuvaev A, Folkersen L, Roy J, Hedin U. (2010) Tissue factor pathway inhibitor-2 is induced by fluid shear stress in vascular smooth muscle cells and affects cell proliferation and survival. J Vasc Surg. 2010 Jul;52(1):167-75.

Gretarsdottir S, Baas AF, Thorleifsson G, Holm H, den Heijer M, de Vries JP, Kranendonk SE, Zeebregts CJ, van Sterkenburg SM, Geelkerken RH, van Rij AM, Williams MJ, Boll AP, Kostic JP, Jonasdottir A, Jonasdottir A, Walters GB, Masson G, Sulem P, Saemundsdottir J, Mouy M, Magnusson KP, Tromp G, Elmore JR, Sakalihasan N, Limet R, Defraigne JO, Ferrell RE, Ronkainen A, Ruigrok YM, Wijmenga C, Grobbee DE, Shah SH, Granger CB, Quyyumi AA, Vaccarino V, Patel RS, Zafari AM, Levey AI, Austin H, Girelli D, Pignatti PF, Olivieri O, Martinelli N, Malerba G, Trabetti E, Becker LC, Becker DM, Reilly MP, Rader DJ, Mueller T, Dieplinger B, Haltmayer M, Urbonavicius S, Lindblad B, Gottsäter A, Gaetani E, Pola R, Wells P, Rodger M, Forgie M, Langlois N, Corral J, Vicente V, Fontcuberta J, España F, Grarup N, Jørgensen T, Witte DR, Hansen T, Pedersen O, Aben KK, de Graaf J, Holewijn S, Folkersen L, Franco-Cereceda A, Eriksson P, Collier DA, Stefansson H, Steinthorsdottir V, Rafnar T, Valdimarsson EM, Magnadottir HB, Sveinbjornsdottir S, Olafsson I, Magnusson MK, Palmason R, Haraldsdottir V, Andersen K, Onundarson PT, Thorgeirsson G, Kiemeney LA, Powell JT, Carey DJ, Kuivaniemi H, Lindholt JS, Jones GT, Kong A, Blankensteijn JD, Matthiasson SE, Thorsteinsdottir U, Stefansson K. (2010) Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. Nat Genet. Aug;42(8):692-7. Epub 2010 Jul 11.

Agardh HE, Folkersen L, Ekstrand J, Marcus D, Swedenborg J, Hedin U, Gabrielsen A, Paulsson-Berne G. (2011) Expression of fatty acid-binding protein 4/aP2 is correlated with plaque instability in carotid atherosclerosis. J Intern Med. Feb;269(2):200-10. doi: 10.1111/j.1365-2796.2010.02304.x. Epub 2010 Nov 14.

Paloschi V, Kurtovic S, Folkersen L, Gomez D, Wågsäter D, Roy J, Petrini J, Eriksson MJ, Caidahl K, Hamsten A, Liska J, Michel JB, Franco-Cereceda A, Eriksson P. (2011) Impaired splicing of fibronectin is associated with thoracic aortic aneurysm formation in patients with bicuspid aortic valve. Arterioscler Thromb Vasc Biol. Mar;31(3):691-7. Epub 2010 Dec 9.

Juel HB, Kaestel C, Folkersen L, Faber C, Heegaard NH, Borup R, Nissen MH. (2011) Retinal pigment epithelial cells upregulate expression of complement factors after co-culture with activated T cells. Exp Eye Res. Mar;92(3):180-8. Epub 2011 Jan 19.

Gertow K, Nobili E, Folkersen L, Newman JW, Pedersen TL, Ekstrand J, Swedenborg J, Kühn H, Wheelock CE, Hansson GK, Hedin U, Haeggström JZ, Gabrielsen A. (2011) 12- and 15-lipoxygenases in human carotid atherosclerotic lesions: associations with cerebrovascular symptoms. Atherosclerosis. Apr;215(2):411-6. Epub 2011 Jan 21.

Peden JF, Hopewell JC, Saleheen D, Chambers JC, Hager J, Soranzo N, Collins R, Danesh J, Elliott P, Farrall M, Stirrups K, Zhang W, Hamsten A, Parish S, Lathrop M, Watkins H, Clarke R, Deloukas P, Kooner JS, Goel A, Ongen H, Strawbridge RJ, Heath S, Mälarstig A, Helgadottir A, Öhrvik J, Murtaza M, Potter S, Hunt SE, Delepine M, Jalilzadeh S, Axelsson T, Syvanen AC, Gwilliam R, Bumpstead S, Gray E, Edkins S, Folkersen L, Kyriakou T, Franco-Cereceda A, Gabrielsen A, Seedorf U, MuTHER Consortium, Eriksson P, Offer A, Bowman L, Sleight P, Armitage J, Peto R, Abecasis G, Ahmed N, Caulfield M, Donnelly P, Froguel P, Kooner AS, McCarthy MI, Samani NJ, Scott J, Sehmi J, Silveira A, Hellénius ML, van't Hooft FM, Olsson G, Rust S, Assman G, Barlera S, Tognoni G, Franzosi MG, Linksted P, Green FR, Rasheed A, Zaidi M, Shah N, Samuel M, Mallick NH, Azhar M, Zaman KS, Samad A, Ishaq M, Gardezi AR, Fazal-ur-Rehman M, Frossard PM, Spector T, Peltonen L, Nieminen MS, Sinisalo J, Salomaa V, Ripatti S, Bennett D, Leander K, Gigante B, de Faire U, Pietri S, Gori F, Marchioli R, Sivapalaratnam S, Kastelein JJ, Trip MD, Theodoraki EV, Dedoussis GV, Engert JC, Yusuf S, Anand SS. (2011) A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat Genet. Mar 6;43(4):339-44.

Kurtovic S, Paloschi V, Folkersen L, Gottfries J, Franco-Cereceda A, Eriksson P. (2011) Diverging alternative splicing fingerprints in the transforming growth factor-β signaling pathway identified in thoracic aortic aneurysms. Mol Med. ;17(7-8):665-75. doi: 10.2119/molmed.2011.00018. Epub 2011 Mar 24.

Razuvaev A, Ekstrand J, Folkersen L, Agardh H, Markus D, Swedenborg J, Hansson GK, Gabrielsen A, Paulsson-Berne G, Roy J, Hedin U. (2011) Correlations Between Clinical Variables and Gene-expression Profiles in Carotid Plaque Instability. Eur J Vasc Endovasc Surg. Jul 6. [Epub ahead of print]

Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, Petrie JR, Travers ME, Bouatia-Naji N, Dimas AS, Nica A, Wheeler E, Chen H, Voight BF, Taneera J, Kanoni S, Peden JF, Turrini F, Gustafsson S, Zabena C, Almgren P, Barker DJ, Barnes D, Dennison EM, Eriksson JG, Eriksson P, Eury E, Folkersen L, Fox CS, Frayling TM, Goel A, Gu HF, Horikoshi M, Isomaa B, Jackson AU, Jameson KA, Kajantie E, Kerr-Conte J, Kuulasmaa T, Kuusisto J, Loos RJ, Luan J, Makrilakis K, Manning AK, Martínez-Larrad MT, Narisu N, Nastase Mannila M, Ohrvik J, Osmond C, Pascoe L, Payne F, Sayer AA, Sennblad B, Silveira A, Stancáková A, Stirrups K, Swift AJ, Syvänen AC, Tuomi T, van 't Hooft FM, Walker M, Weedon MN, Xie W, Zethelius B; the DIAGRAM Consortium; the GIANT Consortium; the MuTHER Consortium; the CARDIoGRAM Consortium; the C4D Consortium, Ongen H, Mälarstig A, Hopewell JC, Saleheen D, Chambers J, Parish S, Danesh J, Kooner J, Ostenson CG, Lind L, Cooper CC, Serrano-Ríos M, Ferrannini E, Forsen TJ, Clarke R, Franzosi MG, Seedorf U, Watkins H, Froguel P, Johnson P, Deloukas P, Collins FS, Laakso M, Dermitzakis ET, Boehnke M, McCarthy MI, Wareham NJ, Groop L, Pattou F, Gloyn AL, Dedoussis GV, Lyssenko V, Meigs JB, Barroso I, Watanabe RM, Ingelsson E, Langenberg C, Hamsten A, Florez JC. (2011) Genome-Wide Association Identifies Nine Common Variants Associated With Fasting Proinsulin Levels and Provides New Insights Into the Pathophysiology of Type 2 Diabetes. Diabetes. Oct;60(10):2624-2634. Epub 2011 Aug 26.

Jackson V, Olsson T, Kurtovic S, Folkersen L, Paloschi V, Wågsäter D, Franco-Cereceda A, Eriksson P. (2011) Matrix metalloproteinase 14 and 19 expression is associated with thoracic aortic aneurysms. J Thorac Cardiovasc Surg. Sep 26. [Epub ahead of print]

Bown MJ, Jones GT, Harrison SC, Wright BJ, Bumpstead S, Baas AF, Gretarsdottir S, Badger SA, Bradley DT, Burnand K, Child AH, Clough RE, Cockerill G, Hafez H, Julian A Scott D, Futers S, Johnson A, Sohrabi S, Smith A, Thompson MM, van Bockxmeer FM, Waltham M, Matthiasson SE, Thorleifsson G, Thorsteinsdottir U, Blankensteijn JD, Teijink JA, Wijmenga C, de Graaf J, Kiemeney LA, Assimes TL, McPherson R; CARDIoGRAM Consortium; Global BPgen Consortium; DIAGRAM Consortium; VRCNZ Consortium, Folkersen L, Franco-Cereceda A, Palmen J, Smith AJ, Sylvius N, Wild JB, Refstrup M, Edkins S, Gwilliam R, Hunt SE, Potter S, Lindholt JS, Frikke-Schmidt R, Tybjærg-Hansen A, Hughes AE, Golledge J, Norman PE, van Rij A, Powell JT, Eriksson P, Stefansson K, Thompson JR, Humphries SE, Sayers RD, Deloukas P, Samani NJ. (2011) Abdominal Aortic Aneurysm Is Associated with a Variant in Low-Density Lipoprotein Receptor-Related Protein 1. Am J Hum Genet. 2011 Nov 3. [Epub ahead of print].

Wang J, Razuvaev A, Folkersen L, Hedin E, Roy J, Brismar K, Hedin U. (2011). The expression of IGFs and IGF binding proteins in human carotid atherosclerosis, and the possible role of IGF binding protein-1 in the regulation of smooth muscle cell proliferation. Atherosclerosis [in press].

CONTENTS

LIST OF ABBREVIATIONS

ASAP	Advanced Study of Aortic Pathology
AUC	Area Under the Curve
BAV	Bicuspid Aortic Valve
BiKE	Biobank of Karolinska Endarterectomies
CRP	C-reactive protein
eQTL	Expression Quantitative Trait Locus (see 2.4.1)
FDR	False Discovery Rate
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-Wide Association Study
HDL	High-Density Lipoprotein
LDL	Low-Density Lipoprotein
TAA	Thoracic aortic aneurysm
TAV	Tricuspid Aortic Valve
PBMC	Peripheral Blood Mononuclear Cell
SNP	Single-Nucleotide Polymorphism

1 BACKGROUND

1.1 GENOMICS

DNA is the means by which the design for an organism is stored. It is the means by which an organism conveys itself to future generations. Because of the impact of this design on the future generations, it is the fabric on which the evolution of life is taking place.

A sequence of 3.2 billion DNA nucleotides constitutes the genome of humans. However, the nucleotide composition is not exactly the same for all humans – there are variations such as the single-nucleotide polymorphisms (SNPs) as well as larger segments of re-arrangements of the DNA sequence. These sequence variations are observed even between siblings. The similarities become fewer and the differences increase as comparison is made between more distantly related humans. For example, the Neanderthals, our now extinct relatives, show similarities with some modern human ethnicities of up to 99.99% identity (Green et al. 2010). In chimpanzees this number is 99.0% (Mikkelsen et al. 2005). This pattern is continued, with steadily decreasing similarity between humans and other organisms; e.g. mice (~40%) (Waterston et al. 2002) and dogs (~32%) (Lindblad-Toh et al. 2005). It continues throughout the tree of life, to plants and bacteria where little sequence is identical.

Clearly, DNA sequence variations have the ability to effect large alterations in form and function of any organism. Yet, only the more extreme cases of genetic variation have appreciable and plainly observed effects on the medical futures of individual humans. The idea that more medically relevant information can be extracted from an individual's DNA motivates modern genetics research which seeks to explain the less plainly observed effects of genetic variation.

Recent milestones towards this goal includes the completion of the human genome sequence (Lander et al. 2001), the identification of common human sequence variation (HapMap 2003), and the development of efficient methods to measure this variation (Schena et al. 1995). Taken together, these advances allowed the genome wide association study (GWAS) (Klein et al. 2005). In this first study 96 individuals with age-related macular degeneration were compared with 50 healthy control subjects and two SNPs that had a significant association with disease were reported.

Today, thousands of GWAS papers have been published and even more SNPs have been associated with a vast range of diseases. In spite of all the statistical complexity found in these GWAS papers, the basic methodology is unaltered from 2005 and distinctly simple; a group of individuals with a trait of interest is compared to a group of people without this trait. If there are one or more SNPs that have significantly different frequencies between the two groups they are said to be associated. For example it is an association when 2000 individuals with a cardiovascular disease have a SNP with C-allele frequency of 47.4%, but its C-allele frequency is 55.4% in 3000 healthy individuals. This example was in fact the case for the first SNP reported for cardiovascular disease, e.g. (The Wellcome Trust Case Control Consortium 2007), which have since been widely replicated.

Since the earliest GWAS, there have been two major trends in study setups. One has been towards larger sample sizes (C4D 2011; Schunkert et al. 2011) and one has been towards a wider selection of intermediate phenotypes, e.g. (Kathiresan et al. 2009a; Strawbridge et al. 2011). Larger sample sizes allow the discovery of smaller effects in high-variation data affected by many known and unknown co-variates. Intermediate phenotypes, allow the decrease of this variation because of focus on discrete pathophysiological pathways. It is outside the scope of this thesis to report all cardiovascular GWAS findings, but more detailed accounts are provided elsewhere, e.g. (Malarstig et al. 2010).

Even with thousands of discovered associated variants, their application towards directly improving human health is sparse. It is often stated that the variability in disease patterns explained by known risk-SNPs is too small to have substantial effect on clinical decision. This point, however, is fiercely debated. An important notion that sidesteps the debate is that GWAS results also provide access to greater understanding of the pathophysiology itself. The reason is the strong ability to show causality in humans. Usually, association does not guarantee causation. However, in GWAS setups the far most plausible explanation for association is that the altered SNP-frequency causes the altered disease-risk. The opposite direction would make little biological sense and it is difficult to envision non-genetic confounding effects. Hence, a large part of the popularity of GWAS comes from the fact that they allow the discovery of causative links to disease in humans.

1.2 TRANSCRIPTOMICS

Upon transcription of DNA, RNA is created. Often the future fate of an RNA molecule is translation into protein – other times it has direct RNA-based functions. In all cases, however, RNA is the first step in realizing the information content of the genome. For this reason it is of much interest to characterize and quantify the flow of mRNA, and this field is called transcriptomics.

As with the DNA that it reflects, the sequence of mRNA can vary between individuals (Li et al. 2011). In addition, however, there are much more important variations in the quantity and splicing of mRNA. Unlike DNA, these variations are different between different cells of the body and between different times and states of these cells. Perhaps as a result of this complexity, the analysis and characterization of RNA is much less clear-cut than that of DNA. On the other hand, one can argue that this makes RNA and gene expression analysis more relevant and well-timed for any given medical question.

For this reason, a multitude of studies have investigated transcriptomics as a means to unravel the molecular background of disease. Typically, expression profiles of circulating cells was compared between patients and healthy individuals (Chon et al. 2004), or tissue biopsies was used to compare different forms of a disease (Barth et al. 2005; Barth et al. 2006; Majumdar et al. 2007; Phillippi et al. 2009). This type of comparative expression profiling studies has highlighted several potential target genes, but unlike the GWAS approach there is little assessment of the differences between

cause and effect. A gene with up-regulated expression in disease could just as well be a response to disease as it could be a cause of disease.

Perhaps because of this, many studies are being reconsidered in ways that more clearly separate causes and effects. As discussed later in the Paper III section, studies of the link between risk-SNPs and expression are one way of strengthening the argument for triggering roles in disease. Another attempt to separate cause and effect is central in the discussion of Paper IV.

1.3 DISEASE

The medical scope of this thesis is cardiovascular disease from several angles. Paper II is based on the results of GWAS investigating myocardial infarction, but Paper III includes both this endpoint and several other risk factors – from lipid levels to blood pressure measurements. Paper IV focuses specifically on thoracic aortic aneurysm, and is thus the most disease specific work. Paper V is based on predicting future ischemic strokes and myocardial infarctions.

1.3.1 Myocardial infarctions, ischemic strokes and their causes

Two common and sometimes fatal medical events are myocardial infarction and ischemic stroke. Both are results of blockage, or occlusion, of important vessels. In the case of myocardial infarction, it is occlusion of the coronary arteries of the heart, and in the case of ischemic stroke, it is the vessels of the brain. In the vast majority of cases, this occlusion is a result of the rupture of a vulnerable atherosclerotic plaque, creating a detached mass, an embolus, which eventually will cause the occlusion. Therefore atherosclerosis is a main cause of both myocardial infarction and ischemic stroke.

Unfortunately, the straightforward description of causality ends at this point. The causes and risk-factors for atherosclerotic plaques are many, and instability and tendency to form emboli, rather than just existence and size of the plaque, are thought to have profound influence on the risk of future adverse events.

Low-density lipoprotein (LDL) is well established to play a role in the atherosclerosis (Goldstein et al. 1974; Ridker et al. 2008; Goldstein et al. 2009). Treatment with statins, which lowers LDL, prevents cardiovascular events in survivors of a first-time myocardial infarction (Pedersen et al. 1994) and in subjects with high cardiovascular risk (Shepherd et al. 1995). Likewise, it is well established that conditions such as high blood pressure and diabetes are strongly associated with increased risk of myocardial infarction and ischemic strokes. Other commonly used biomarkers such as serum C-reactive protein (CRP) and serum high-density lipoprotein (HDL), are correlated to atherosclerosis (inversely in the case of HDL), but their actual causal roles are more controversial (Danesh et al. 2009; Nordestgaard et al. 2011).

Taken together, and adding a multitude of other real and suspected risk factors, it is indeed fair to characterize myocardial infarction and ischemic stroke as a complex multi-factorial diseases. For the purposes of this thesis a key point is as follows: in addition to the risk factors we know are causal and the ones we think might be causal,

there are plenty of potentially completely unknown factors. Knowledge of these could both be immensely helpful in diagnosis as well as in treatment of complex cardiovascular disease.

1.3.2 Thoracic aortic aneurysm and bicuspid aortic valve

Paper IV focuses on thoracic aortic aneurysm (TAA). As with myocardial infarction and ischemic stroke, TAA is potentially fatal. Fatality from TAA happens when the aorta is dilated to the point where the vessel wall is broken, either as a rupture or as a dissection of the aorta. One important clinical observation is that amongst TAA patients there is an increased prevalence of patients with bicuspid aortic valve (BAV). This observation is still unexplained, and underlies the main question asked in Paper IV. The aortic valve is the heart valve that controls flow out of the left ventricle and normally it has three cusps (tricuspid aortic valve, or TAV). BAV is a common congenital cardiovascular malformation with prevalence of 1-2%, in which two cusps of the aortic valve are fused together (Figure 1). The aneurysms of patients with BAV grow faster and their TAA develop at a younger age than patients with TAV (Nkomo et al. 2003; El-Hamamsy et al. 2009; Jackson et al. 2011b).

Figure 1. *Illustration of a normal tricuspid aortic valve (left) and a bicuspid aortic valve, with fusion of the right- and left coronary cusp (right).*

TAA and BAV are thought to have strong heritable components (Cripe et al. 2004; Siu et al. 2010). However, no specific causative mutations have been identified. Suggested candidate genes include the NOTCH1, ACTA2, FBN1 and the genes encoding the collagens, elastin, matrix-metalloproteinases and TGF-β (El-Hamamsy et al. 2009; Siu et al. 2010). In addition, it was recently observed in a GWAS that SNPs in the FBN1 gene were associated with TAA and aortic dissection (Lemaire et al. 2011). In further support of these candidate genes, it is of interest to mention the few rare syndromic forms of TAA because they involve known causal genes. These are the Marfan syndrome involving FBN1 (Ramirez et al. 2007) and the Loeys-Dietz syndrome involving the TGF-β receptors (Loeys et al. 2006).

An important point of discussion on the pathophysiology of BAV is the question of hemodynamic changes caused by the malformed aortic valve. Studies have shown that BAV patients have perturbed aortic flow and it is hypothesized that this disturbance could result in the disease development (Hope et al. 2010). An alternative hypothesis, however, is that the same genetic factors which cause BAV also lead to increased TAA risk. This is supported by the fact that the TAA risk in BAV patients is unaffected after aortic valve replacement (Yasuda et al. 2003). As further discussed in the Paper IV section, this controversy is still subject to discussion.

2 METHODS

2.1 GENETICS

Common methods to characterize DNA variation range all the way from the Sanger dideoxy-method, which was the next generation sequencing of 1977 (Sanger et al. 1977), to the latest high-throughput technologies based on massively parallel sequencing. A fundamental point of methodology is to distinguish the aim of determining all sequence and its variation in a given DNA sample (sequencing), with the aim of observing specific known variations in a DNA sample (genotyping). The latter aim can be accomplished with genotype specific PCR amplification for individual SNPs or genotyping microarrays which covers all common SNPs in the human genome. Because GWAS almost exclusively are based on data from genotyping microarrays, the work in this thesis is designed around this method.

2.1.1 Quality metrics

The vast majority of genotyping measurements referred to in this thesis were performed using Illumina Human 610W-Quad Beadarrays at the SNP technology platform at Uppsala University. Part of the quality control was done at the core centre, where SNP-calling and SNP exclusion based on standard Illumina metrics were performed. In addition to core-center methodology, we performed three quality control tests. Firstly, duplicate samples were submitted for three patients. This showed a genotyping concordance rate of 99.90%. Secondly, 8 SNPs had previously been genotyped with a taqman PCR-based method in 89 samples from the BiKE cohort. These showed only 1

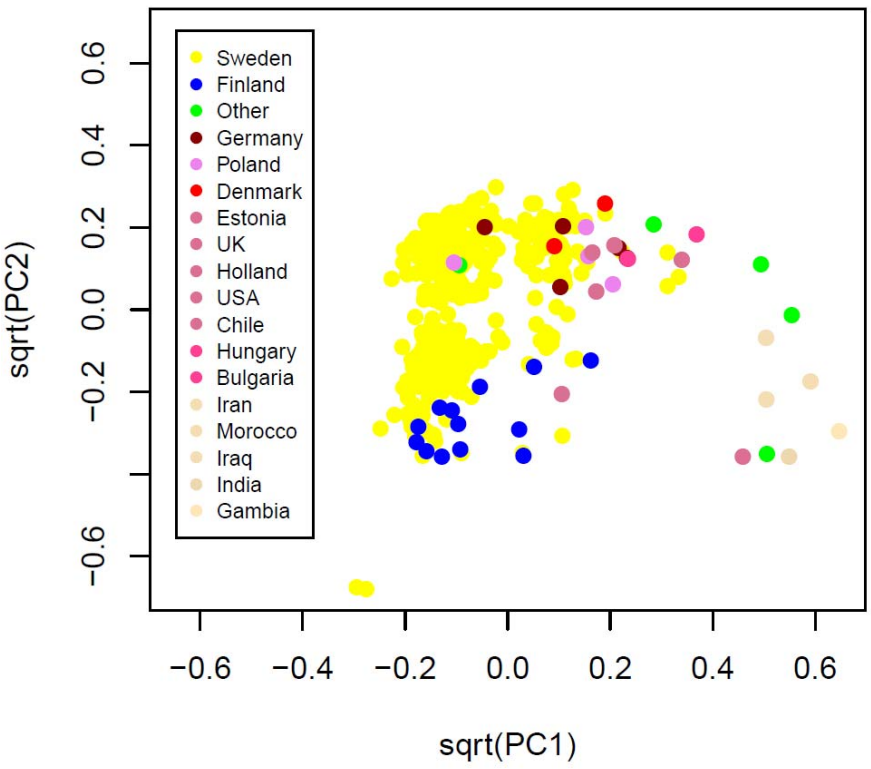

Figure 2. *Principal components analysis (PCA) of ASAP genotype data based on self-reported place of birth. The principal components on each axis do not have any direct biological corollary, but the purpose of the PCA here is to cluster samples with similar genotypes together. Both principal components have been square-root transformed for a better view of relevant clustering.*

discrepant genotype, corresponding to a concordance rate of 99.86%. Thirdly, we investigated sample switching frequency by comparing gender and place-of-birth information with genotypes. All samples with Y-chromosome genotype calls were registered as male and vice versa for females. Principal components analysis of all genotypes combined with labelling by birth place (Figure 2) showed strong clustering of nationalities as expected (Novembre et al. 2008).

2.1.2 Imputation

In addition to the directly genotyped SNPs, several cases required knowledge of SNPs which were not directly measured. This data was obtained through imputation. Imputation is the process of estimating the genotypes of SNPs that are not measured, based on information from proximal SNPs that have been measured and information on linkage disequilibrium as obtained from either the HapMap consortia (HapMap 2003) or the 1000 genomes project (Durbin et al. 2010). Essentially the process allows guessing an additional 4-5 million SNPs from the measurement of only half a million. Imputations in this thesis have been performed either with the Plink algorithm (Purcell et al. 2003) or the Mach 1.0 algorithm (Li et al. 2010), based on HapMap reference data or 1000 genomes data, respectively. All imputation calculations were performed on the UPPMAX computing cluster.

2.2 TRANSCRIPTOMICS

Several methods exist to quantify the amount of mRNA in a sample – starting from the very first northern blots, over real-time PCR, gene-expression microarrays and into high throughput RNA sequencing. All these methods have differences, strengths and shortcomings, but the general trend has been to move towards higher throughput (more genes measured at the same time), and higher resolution (more complete knowledge of RNA transcript architecture). While there can be no doubt that throughput and resolution is improving, previous methods still have their proponents in terms of precision, accuracy and price-considerations and most published gene expression studies are required to undergo validation using real-time PCR.

2.2.1 Quality metrics

A later section in the statistics section is dedicated to a discussion of terms in validation, but for a general introduction to gene expression measurement it is of interest to establish if the different methods show the same results. Expression microarrays and real-time PCR were compared in a study by the MicroArray Quality Control consortium, which tested a broad range of platforms and different test sites in a small collection of standardized samples (Canales et al. 2006; Shi et al. 2006). The type of expression microarray also used in the BiKE cohort, was reported to give a correlation of $R = 0.92$ over ~450 genes when comparing fold-changes between two samples. Of interest, a later study found a correlation of $R = 0.95$, between real-time PCR and high-throughput sequencing in the same samples (Wang et al. 2008).

A more relevant quality metric to the works presented here would be how individual gene expression levels compare across samples, when using different measurement methods. The reason is that virtually no comparisons are made between the expression levels of different genes. Most comparisons are of individual genes and their expression level between different sample types, e.g. in different genotypes or in different patient groups. Of interest to this type of question is a study, comparing 53 cell line samples for which gene expression was available both from a high-throughput RNA-sequencing platform and from expression microarrays (Pickrell et al. 2010). Correlations in the range 0.60 to 0.78 was observed across genes, but across samples this dropped to an average of 0.3 (range -0.4 – 0.9), for all genes at the medium-high expression level. Correlation was worse at more extreme expression levels. Cross-sample studies between expression microarray and real-time PCR data in the BiKE database showed an average correlation of 0.53 (range 0.17 – 0.76) in 15 genes over 88 samples (Folkersen et al. 2009b). Compared to correlations across genes, these across-sample correlations are disappointing, and the high level of variance is worth keeping in mind when interpreting transcriptomics data here and in general.

2.2.2 Array mechanism and design

Several different companies manufacture expression microarrays. Common for all is a mechanism based on hybridisation of a sample of unknown labelled RNA sequences (the sample), to a set of known short RNA sequences located in an array with defined positions (the probes). After hybridization, a scan of the array will reveal the amount of labelled unknown RNA at each position, which can then be translated to a quantitative parameter for the RNA at that particular position. Before hybridization, a number of reverse transcriptions, amplification, or fragmentation steps might be performed, but this depends on the brand and type of expression microarray. Likewise there are differences in the organization of probes, where some companies use one or two longer probes for each gene (Agilent, Illumina) and others group several shorter probes for the same gene or gene region into so-called probe sets (Affymetrix).

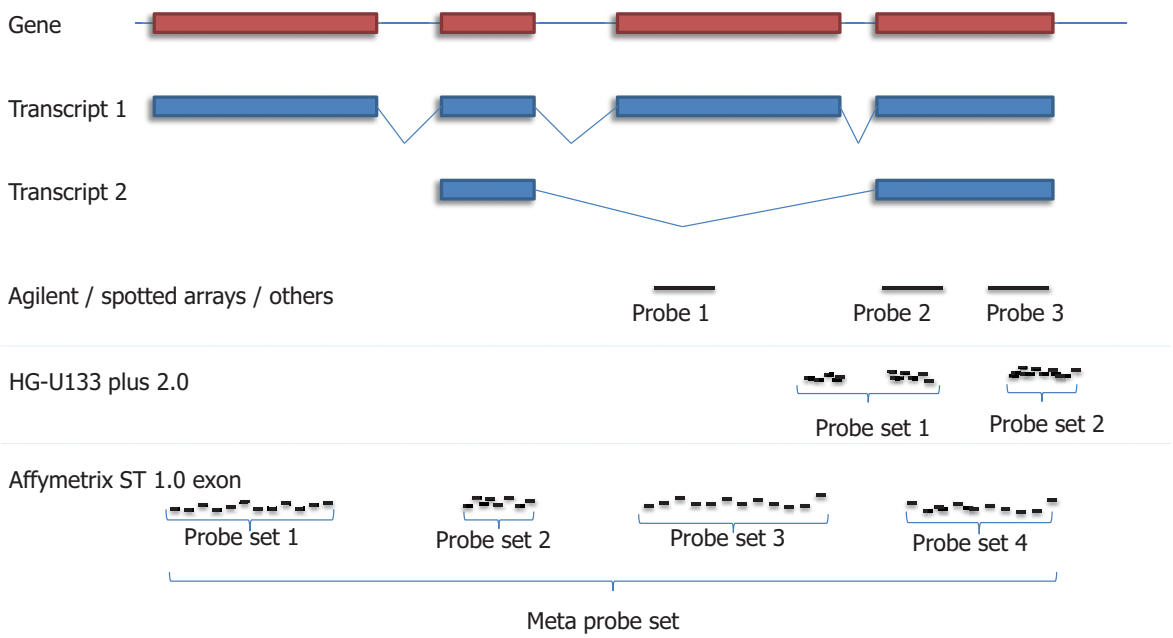

Figure 3. *Overview of concepts in the organization of Affymetrix expression arrays.*

Particularly for Paper I and II, it is important to explain the organization of these probe sets on Affymetrix expression microarrays in detail. Unfortunately some of the terms can be confusing as their meaning change with the type of Affymetrix microarray. In all cases the term *probe* is defined as meaning a 25 nucleotide long sequence, designed to match in one, and only one, region of a gene. Groups of 4-13 of these probes are termed as *probe sets* and the probes from a set always match genomic locations in close proximity to each other. In the older generation of expression microarrays, of which the HG-U133 plus 2.0 is part, these probe sets are focused at the 3' part of the gene, and their hybridization intensity is meant to be taken as a value for the quantity of the entire gene. In the newer generation of expression microarrays, of which the ST 1.0 exon array is part, these probe sets are distributed over the gene, with approximately one probe set per exon. In this array type, the *probe set* value is meant to be taken as the quantity of an individual exon. The value of an entire gene, is then retrieved from grouping all probes, from all gene-specific probe sets into one so-called *meta probe set*, which is meant to be taken as the quantity of the entire gene. This organization is illustrated in Figure 3. The main point to keep in mind is that different array types organize their probes differently – either they all can be summarized to yield one value for one gene, or else one can keep track of individual probes in order to reveal more detailed information.

2.2.3 Normalization and pre-processing

The last point of microarray analysis is how to arrive at one value per gene and per sample which is comparable across the study. Because of the fragmentation, amplification and hybridization steps, there will be systematic hybridization intensity variation between any two arrays. A normalization step is therefore required. Also, because of the organization of probes described above, a step of summarization must be included. Often some kind of background estimation and subtraction is included as well. These matters of pre-processing can be the source of much discussion and many competing algorithms have been developed. These have been fully reviewed elsewhere (Calza et al. 2010). Today, PLIER and RMA seem to be generally accepted norms for pre-processing, even though both include some rather drastic assumptions which might not hold when comparing samples with large differences. In support of RMA and PLIER, however, it should be mentioned that in our hands they produce the best correlation with real-time PCR data (in comparison with MAS5.0, gcRMA, and dChip algorithms). Because of this, it was chosen to use the RMA algorithm (Irizarry et al. 2003) with log2 transformation for all pre-processing of expression microarray data in this thesis.

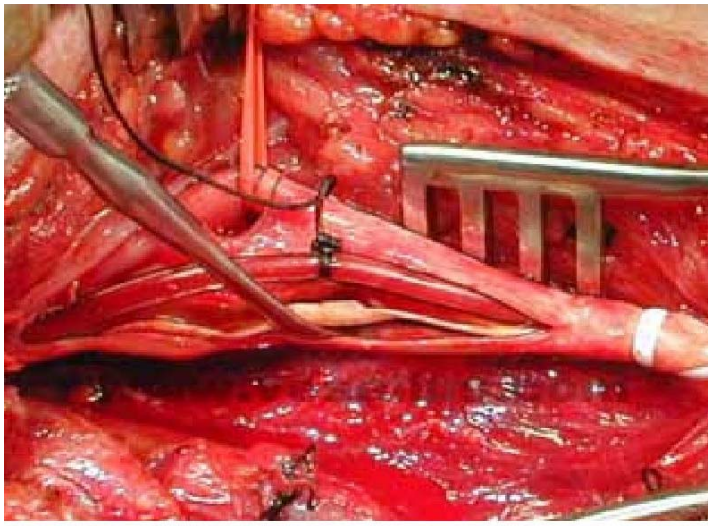

Figure 4. *Ongoing carotid endarterectomy at the Karolinska Hospital. Photo provided by professor Jesper Swedenborg.*

2.3 BIOBANKS

The studies described herein are based on the *Biobank of Karolinska Endarterectomies* (BiKE) and the *Advanced Study of Aortic Pathology* (ASAP). It can hardly be understated how important these two biobanks have been to this thesis. A biobank is a collection of several biological samples together with clinical information about the samples. How many samples and how much information vary (Shalhoub et al. 2011). In the BiKE and ASAP databases, information is available on clinical records, biochemical measurements, ultrasound imaging, follow up, genotype profiling and expression profiling in several different tissues. All these measurements taken together provide one of the highest amounts of information per patient in any existing cardiovascular biobank study. The trade-off to this is that expansion of sample size becomes very costly. Both in monetary material-cost but also from the fact that each of these biopsies are taken during surgery at times when a multitude of other clinical priorities are in effect.

The BiKE database consists of plaque samples removed during carotid endarterectomy on consecutive patients treated at the Karolinska Hospital (Figure 4). Carotid endarterectomy is indicated in cases of substantial narrowing of the carotid artery, as detected with duplex ultrasound. Typically, a patient presents with symptoms of transient ischemic attacks, stroke, or amaurosis fugax. Endarterectomy is also indicated in men with asymptomatic narrowing of the carotid artery. The efficacy of the surgery was established in the NASCET trial (Ferguson et al. 1999) and the ECST trial (ECST 1998). The NASCET criteria were used for selection, and no further exclusion criteria were pre-defined. For some BiKE-investigations ad-hoc criteria (e.g. atrial fibrillation exclusion) have been applied subsequently, as described in relevant publications. Other biobanks of carotid endarterectomies have been described in (Hurks et al. 2009; Goncalves et al. 2010; Saksi et al. 2011).

The ASAP database consists of biopsies taken during consecutive elective surgeries for aortic aneurysm and aortic valve repair. Inclusion criteria was age above 18 and aortic

valve disease (aortic stenosis or regurgitation) or ascending aorta dilatation (aneurysm or ectasia). Exclusion criteria included coronary artery disease as defined by coronary angiogram. More clinical details can be found in (Jackson et al. 2011b). Patients were classified as having either BAV or TAV and as having either dilated or non-dilated thoracic aorta. The criteria for the latter were >45 mm (dilated) and < 40 mm (non-dilated), respectively. These thresholds are based on clinical guidelines which indicate surgery at 40 mm dilation in BAV patients and 45 mm dilation TAV patients.

Description of sample size of the biobanks is complicated by the fact that some measurements only have been performed in some of the samples. There are 489 BiKE individuals and 500 ASAP individuals, but this number is not very informative, because analysis effectively is performed only on the expression microarrays. This fact does leave open possibilities of validation in independent patients, however. Figure 5 describes current sample composition in more detail. Note, however, that both biobanks are continuously being expanded and that sample sizes in earlier papers are somewhat less than indicated here.

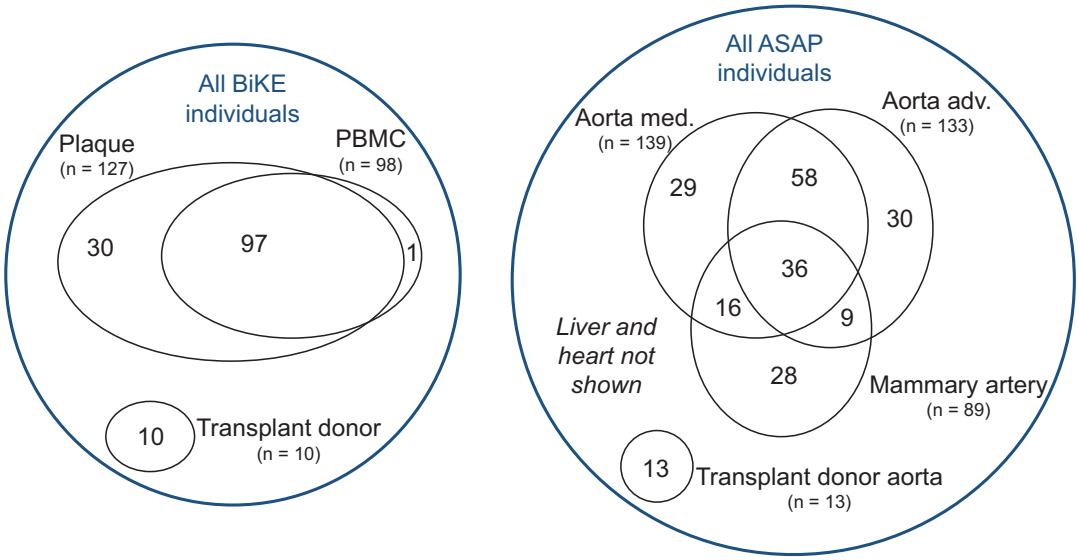

Figure 5. *Overview of expression microarray measurements in BiKE and ASAP, as of October 2011. For ASAP there are 309 individuals with at least one array and a total of 700 arrays. For BiKE there are 138 individuals with at least one array and a total of 235 arrays. Genotyping microarray measurements overlap with expression microarray measurements for all but 2 BiKE patients and all but 1 ASAP patient. In addition to the samples shown for ASAP, there are expression array measurements from 127 heart biopsies and 212 liver biopsies. A majority of these overlap with the vascular samples.*

Both the BiKE and the ASAP biobanks are large collaborative efforts, and the papers in this thesis merely represents a small part of the output. For further reading on other aspects of both biobanks see Table 1 and Table 2.

	Main finding	Role of BiKE
(Agardh et al. 2011)	Finds FABP4 as the most differentially expressed transcript in unstable plaques	Central role in unbiased identification of FABP4
(Breland et al. 2010)	MCP-4 levels are increased in patients with symptomatic carotid atherosclerosis	*In vivo* plaque mRNA measurements supporting main data
(C4D 2011)	GWAS that identifies five new risk-SNPs for myocardial infarction	Finding expression quantitative trait loci (eQTL) effects of novel GWAS risk-SNPs
(Diez et al. 2010)	Theoretical overview of network analysis methods	Methodology exemplified using BiKE expression data
(Folkersen et al. 2009b)	Investigation of real-time PCR normalization methods	Central role as source of gene expression measurement
(Folkersen et al. 2009c)	Paper II	Essential
(Folkersen et al. 2010)	Paper III	Essential
(Gabrielsen et al. 2010)	TBXAS1 is increased in murine atherosclerosis and correlates with macrophage markers	Major role on macrophage marker correlation
(Gertow et al. 2011)	Investigation of leukotriene genes in plaque highlights ALOX15B	Central role as source of all gene expression measurement
(Malarstig et al. 2008)	IRF5 is expressed in the atherosclerotic plaque and is affected by proximal SNPs	Major role as source of eQTL measurements to match with association studies
(Olofsson et al. 2009)	Tnfsf4 expression was associated with increased atherosclerosis in mice.	A role in human gene expression together with other major biobanks
(Qiu et al. 2006)	Leukotriene component LTA4H is correlated with time from last symptom in	Major role as human i*n vivo* validation of animal model data

	plaque	
(Razuvaev et al. 2011)	Investigation of a selected set of 317 known atherosclerosis genes	Central role as source of all expression measurements on the gene set
(Tran et al. 2007)	HSPG2 is reduced in human atherosclerotic lesions	Central role as the source of all immunostaining and gene expression material
(Ueland et al. 2009)	Broad mechanistic and clinical studies on DKK-1 levels in atherosclerosis	*In vivo* plaque mRNA measurements supporting main data
(Wang et al. 2011)	Expression of IGFs and IGF-binding proteins in human carotid atherosclerosis	Central role as the source of all gene expression material

Table 1. *Overview of all publications that uses data from BiKE.*

	Main finding	Role of ASAP
(Bown et al. 2011a)	GWAS that identifies one new risk-SNP for abdominal aortic aneurysm	Finding eQTL effects of LRP1 gene.
(C4D 2011)	GWAS that identifies five new risk-SNPs for myocardial infarction	Finding eQTL effects of proximal genes.
(Folkersen et al. 2009c)	Paper II	Essential
(Folkersen et al. 2010)	Paper III	Essential
(Folkersen et al. 2011)	Paper IV	Essential
(Gretarsdottir et al. 2010)	GWAS that identifies one new risk-SNP for abdominal aortic aneurysm	Finding eQTL effects of DAB2IP gene.
(Jackson et al. 2011a)	Investigations of MMP gene expression in TAA patients finds MMP14 and 19 to be differentially expressed	Central role as source of MMP expression measurements
(Jackson et al. 2011b)	Studies on cusp morphology implying intrinsic mechanisms as source of BAV-related TAA	Central role as source of all morphology measurements
(Kurtovic et al. 2011)	New method to investigate alternative splicing and identification of TGF-beta gene splice variants	Central role as source of exon expression measurements
(Paloschi et al. 2011)	Known splice isoforms of fibronectin differs with with cuspidity of TAA patients	Central role as source of fibronectin gene expression measurements
(Strawbridge et al. 2011)	GWAS that identifies nine new risk-SNPs for pro-insulin levels	Expression level and eQTL effects of proximal genes.

Table 2. *Overview of all publications that uses data from ASAP.*

2.4 STATISTICS

All calculations in this thesis have been performed in the R/Bioconductor programming language (Gentleman et al. 2004; R Development Core Team 2011). Choice of statistical tests has been prioritized towards the simplest test-type that would suffice for the question at hand. In addition to a large number of Student's T-tests and Pearson correlations, this includes a few more advanced concepts which will be described here. A feature of exclusively scripting all statistics is that all calculations are saved for later review – this is useful both in cases of auditing suspected errors and whenever previous methods are required in similar but not identical context. Altogether the calculations performed during the creation of this thesis consists of 168 846 lines of R-code, collected in 805 script-files. A web-interface connecting a graphical user interface to the most frequently used scripts is under development. This project can be found under the name of ExpressionWebExpress, at code.google.com.

2.4.1 Additive linear models for eQTLs

The analyses in Paper I, II, and III makes extensive use of additive linear models for detecting association between SNPs and gene expression levels. It is therefore important to thoroughly understand the mechanism of this test. In a more general sense, linear modelling or linear regression is a widely used method for describing any response variable as a function of several covariant variables. In the simplest case of two numerical variables, it will simply give the slope of the best fit of a regression line describing variable Y as a function of variable X. In the specific context of eQTL studies this is very often used as a so-called additive linear model, where the expression level of a gene is modelled as a function of genotype, where genotype is encoded numerically as e.g. 0 for AA, 1 for AG, and 2 for GG. When expression and

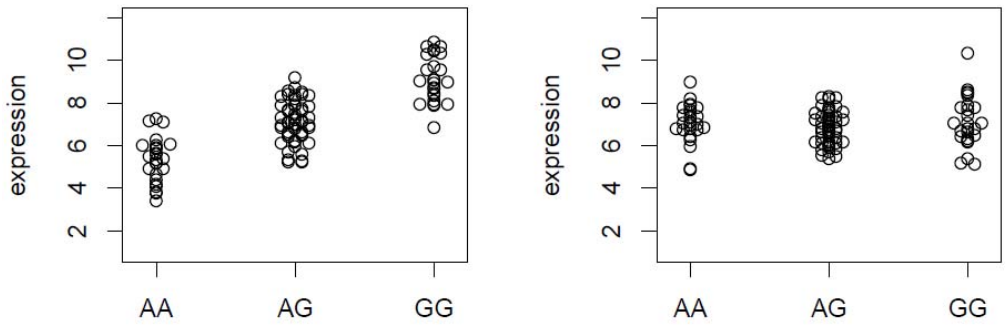

Figure 6. *Each dot shows a gene expression measurement in a patient with a given genotype from an arbitrary SNP, which can be either AA, AG or GG. In the example to the left there is a clear association between genotype and expression and the P-value of a linear-additive model is much below 0.05. In the example to the right, there is no association and the P-value of a linear additive model is above 0.05.*

numerically encoded genotype are used as input in a linear model function, an estimate of the slope and a P-value is obtained. The magnitude of this estimate can be interpreted as the per-allele effect of the genotype on the expression level. Because of the log2-scale expression data used, a value of 1 therefore equals a doubling of gene expression. As usual the P-value is the probability that the null hypothesis is true, where the null hypothesis in this case is when the genotype has no effect on the gene expression level. An example of both a significant association and a non-significant association is given in Figure 6. Alternatives to the additive linear model for eQTL studies, is the use of the Spearman rank correlation. This non-parametric statistic is preferred in non-normally distributed data sets, which is rarely the case in this thesis.

2.4.2 Cox proportional hazards regression model

The survival analysis in Paper V revolves around the use of the Cox proportional hazards model, or Cox-model. This model is essentially an extension of the linear model with time as the main covariate. However, the model has the important addition that data points either can be exact ("an event happened precisely at this time") or boundaries ("no event happens before this time", also known as censored points). An example of this is given in Figure 7, where four patients are recruited into a survival study. Patients 1 and 2 suffer from an adverse advent in the beginning of 2011. Although patient 1 has the event before patient 2, he still is registered as having a longer follow up time because he was recruited into the study at an earlier time. Patients 3 and 4 do not have any adverse events. Nonetheless we have to treat them differently in our calculations because patient 4 has been followed for a much shorter time, and could have suffered from an adverse event at some time in late 2010 after the end of his follow-up time. The particular reasons why patients 3 and 4 are not in the study anymore are not essential, as long it is not an event of the type under study. In the case of patient 3 for example the study simply reached its end-date. For patient 4 it could have been either a death due to unrelated factors (e.g. an automobile accident) or because of loss of follow-up, such as moving to another continent.

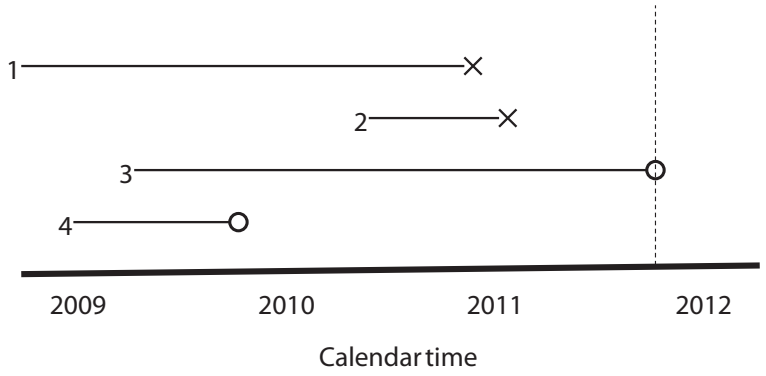

Figure 7. *Example of four patients in a survival study.*

The purpose of the Cox-model is to estimate and quantify how likely it is that a clinical variable is associated with chance of event-free survival. This is extensively used in Paper V, where clinical variables with known effect on survival, such as age, gender and smoking are compared with new variables based on genetics and transcriptomics.

In evaluating prognostic properties of any type it is common to provide receiver operating characteristic (ROC) curves. The ROC curve describes how capable a measurement is at correctly categorizing samples in two groups. The main advantage of a ROC curve is that it shows how capable the measurement is at all thresholds values. The measurement threshold where a sample is categorized as positive does therefore not have to be pre-specified.

A plot of the ROC curve is necessary to fully interpret the range of all possible thresholds. However, a common derivate statistic to summarise the same information is to calculate the area under the curve (AUC). The AUC is connected to the probability that a true positive will have higher prediction score than a true negative. An AUC of 0.5 therefore indicates that true positive or true negative samples are equally likely to have higher prediction scores, i.e. a useless prediction based on pure chance. An AUC of 1.0 on the other hand indicates that prediction scores will always be higher for true positive samples, i.e. a perfect prediction that is always correct.

2.4.3 Multiple testing correction

Throughout this thesis, multiple testing issues play a large role and it is therefore important with a careful explanation of the causes and consequences of multiple testing. A P-value indicates the probability that an observation, such as a change in gene expression, is not correct. By convention it is accepted that this probability has to be 5% or lower before a researcher can present anything as significant. However, because the 5% probability of non-difference exists at every single test, an example study with 20 Student's T-tests performed on completely randomly generated numbers would have a 64% probability of having at least one P-value below 5% by pure chance. This is the reason for multiple testing correction.

Several approaches exist to correct for multiple testing. The simplest is known as Bonferroni correction. In this method the P-values obtained are multiplied by the number of tests, and then required to be below 5%. A common criticism against the Bonferroni correction is that it is overly conservative with inflated false-negative rate. A handful of other multiple testing correction methods exist. The tests vary in the amount of correction, but they all lower the P-value threshold from the non-corrected 5%-level to a value that is increasingly lower with the number of tests performed. One such test that is used in Paper III and IV is the false discovery rate (FDR). This test reports the proportion of false positives in a set of significant results, rather than the probability of the individual tests being false positive. Beneath the semantic difference, however, it is still a method to lower the P-value threshold in order to take multiple testing into account.

Multiple testing issues are in fact so central to any data intensive analysis that it will be explained once more using simpler terminology. For average coins, heads-up and tails-up happens half of the time. If we were to find a coin that landed heads up 9 out of 10

times, it would be special. This is essentially the same kind of special that we search for in genes and SNPs. However imagine that a special coin had been identified from a group of 99 other coins, all of which had been flipped 10 times each. Intuitively this makes the special coin less special. A single coin out of hundreds would probably just be found by chance. Theoretically it is in fact not special. The calculated probability that one coin will be heads 9 of 10 times is 1.07%. The calculated probability of this happening in at least one coin out of a hundred is 66%.

2.4.4 Cross-validation

Multiple testing is just one issue on a long list of biases and problems that can contribute to false positive results. For this reason it should be demanded that results are validated. The type and level of validation can vary. Doubts about measurement-accuracy require technical validation, using independent methods. Doubts about general reliability of results require biological validation, using independent samples. More extraordinary claims require more extensive validations. A typical experimental setup, as for example used in GWAS, is to use a discovery cohort and a validation cohort. Because methods and techniques can differ between these cohorts, this can even have the advantage of overcoming many hidden biases. On the other hand the technical differences might cause perfectly true results to be rejected on false grounds, so it is advisable to change as few things as possible in each validation.

Sometimes validation cohorts are not available. There can be several reasons for this, but ultimately it usually relates to time, costs and sample-restraints. If the cohort being studied is large this does not need to be a problem. Obviously one could just randomly split the total cohort into two halves and name one discovery and another validation. This would work, and would yield correctly validated results – albeit only with the power of half the samples.

The split would also work if a larger fraction of the total cohort was reserved as discovery-cohort, and this would increase the power to detect correspondingly. The disadvantage is that with a smaller validation cohort the validation becomes less precise. Importantly, however, it does not become less accurate. The difference between precision and accuracy is important here. The more precise estimate will more often give the same value when repeated. The more accurate estimate will on average be closer to the correct value. Thus, if we imagine a very small validation-cohort of, say, two samples, we would have almost all the power and accuracy of the remaining total cohort, but our estimate would be extremely imprecise. It would not be advisable to rely on such imprecise estimates.

Building on the above argument, cross-validation is introduced. In cross validation the data set is split as described above, the discovery calculation is performed, the validation calculation is performed, and then the entire process is repeated again – this time with a new and different split of the data sets. For each of these repetitions an estimate is recorded. This estimate has high accuracy, but low precision. All the part-estimates taken together, however, will give a more precise and accurate estimate of the magnitude of the effect (Simon et al. 2003). Cross-validation is an established part of prediction studies based on expression profiles in cancer, e.g. (Tibshirani et al. 2002; Borup et al. 2010).

3 RESULTS AND DISCUSSION

3.1 PAPER I

The purpose of Paper I (Folkersen et al. 2009a) was to present a software package that could visualize expression microarray data at the most detailed level possible. The motivation for creating this software package was the incomplete annotation of the chromosome 9p21 region, which recently had come into focus as a major myocardial infarction risk locus and which would be the focus for Paper II.

A standard analysis of Affymetrix expression arrays rely on the use of Affymetrix-provided annotation files. Principally one file is needed for mapping between gene names and probe set ID-numbers and another is needed for mapping between probe set ID-number and the physical array location of each probes in the probe set. Practically, there are quite a few differences in the setup depending on the array-type. For example, to get from physical probe-location to gene name in the ST 1.0 class of arrays, four mapping files are required; the pgf-file, the clf-file, the mps-file and the transcript.csv file. More detail is provided in Figure 3 in the methods section. For the purposes of this paper, however, the important point is that the measured expression of a gene is understood to depend on the hybridization to a set of defined probes – but that this set definition is given by Affymetrix-provided annotation files.

As described in the quality metrics section of methods, this setup works adequately for most cases of gene expression analysis. The expression of a gene with well-defined exon-boundaries and no transcript variants should theoretically be completely described by the hybridization-intensity of probes matching to its sequence. In this case, any measurement error can be attributed to unavoidable technical noise. In cases where the annotation is incomplete, however, the theoretical soundness fails. If for example, a poorly characterized gene is defined by 20 probes, but only 10 of these are actually in transcribed regions – then the measured expression value will be half of the true expression value. This is particularly a problem in genes with imprecise exon-boundaries.

The solution to this problem was to develop a software package that could extract the hybridization intensity and the nucleotide sequence of individual probes in any gene from any set of raw data files. This was combined with functions that allowed matching of extracted probe sequence with downloaded gene sequence and easy plotting of expression as function of gene-location. All these different data components were readily available in the raw data and downloadable annotation files, but even with knowledge of where to look it took time to find and match them all. With a software package, this time was reduced from hours of manual work to minutes of processing time, as well as a simplification of the amount of technical background required to understand. The software package was named GeneRegionScan.

Fast and easy matching of individual probe sequence with gene sequence provided by the user can resolve the problem of incomplete annotation. If the Affymetrix-provided annotation is problematic, the user can easily provide new annotation in the form of simple gene sequences. This can be either sequence with the latest updated information or custom information based on results from ongoing studies.

With the help of the Bioinformatics Center at Kyoto the script was adapted to R-package format and published in the Bioconductor repository. Since publication in august 2009, it has been downloaded a little more than 100 times per month – approximately equal to the median amount of downloads of all software packages in the Bioconductor repository.

3.2 PAPER II

The purpose of Paper II (Folkersen et al. 2009c) was to investigate the functional consequences of the chromosome 9 genetic variation that had been robustly associated with risk of cardiovascular disease. As already referred to in the introduction section, this variant was the first SNP to be associated with cardiovascular disease through the use of GWAS (Helgadottir et al. 2007; McPherson et al. 2007; Samani et al. 2007; The Wellcome Trust Case Control Consortium 2007). One common variation with larger effect size has subsequently been identified (Clarke et al. 2009), but the 9p21 risk allele is still today the most significant association between a SNP and cardiovascular disease (C4D 2011; Schunkert et al. 2011).

The rs2891168 SNP investigated in Paper II was part of an LD block consisting of 14 SNPs in high linkage disequilibrium with each other (Broadbent et al. 2007). Because of this, GWAS studies have often reported on the different SNPs in this region as equivalents. The SNPs in the haploblock were found to span a sparsely described gene, by the gene symbol of CDKN2BAS. A previous study from the field of cancer had linked the deletion of this gene to cancer in a French family and the gene was named Antisense Noncoding RNA in INK4/ARF Locus (ANRIL) (Pasmant et al. 2007). Based on this study, ANRIL was annotated as having a long 19-exon form and a short 13-exon form. Work by our collaborators at the Oxford Wellcome Trust Centre for Human Genetics showed that the alternative splicing of ANRIL was more complicated than what was shown in the annotation data base (Figure 8, with UCSC genome browser data) and previous sequencing work (Pasmant et al. 2007). The annotated short and long isoforms could not be readily detected, but 8 novel exon-combinations of the ANRIL gene were found based on exon 1, 14 and 20 PCR primer amplification followed by capillary sequencing.

At the time of publication of Paper II, two other functional real-time PCR based studies of the 9p21 risk locus had recently been published: one study showed that the 9p21 risk allele was associated with *decreased* expression of the neighbouring genes CDKN2A, CDKN2B and ANRIL in peripheral blood T-cells from 170 healthy individuals, measured at an unreported exonic location (Liu et al. 2009). Another study showed an association between the 9p21 risk allele and *increased* expression levels of the "short" transcript of ANRIL in whole blood cells from 124 healthy individuals (Jarinova et al. 2009). The same study reported that the "long" transcript isoform of ANRIL had *decreased* expression with the risk allele.

With at least 10 different transcript isoforms of ANRIL detected, it was clear that a single real-time PCR based measurement point on the gene was insufficient. It could even be speculated that there was no inconsistency between the studies by Liu and Jarinova, but that they merely were measuring different splice variants. We therefore

turned to expression microarrays and the newly developed GeneRegionScan package from Paper I, which had in fact been developed towards this use.

The data sets at our disposal at the time was the BiKE database of carotid plaque samples, the ASAP database in an early version that consisted of a subset of the aorta media and mammary artery media samples, and two publically available expression microarray databases from Gene Expression omnibus (Edgar et al. 2002). The BiKE and ASAP databases at that time had genotype specific PCR-genotyping performed for the rs2891168 SNP. The two publically available data sources were the CEU HapMap dataset (HapMap 2003), which provided genotype data, and an expression microarray database (Kwan et al. 2008; Zhang et al. 2008), which provided gene expression from lymphoblastoid cell lines from the same HapMap individuals.

Consequently, the hypothesis being tested was if any particular exon-region of ANRIL or any other proximal gene was associated with the 9p21 risk allele. The somewhat disappointing result of testing this hypothesis was that no clear region of association could be observed. In the publication of these findings our aim was to highlight the inconsistencies involved in reporting gene expression measured in arbitrary exons of a gene with known alternative splicing.

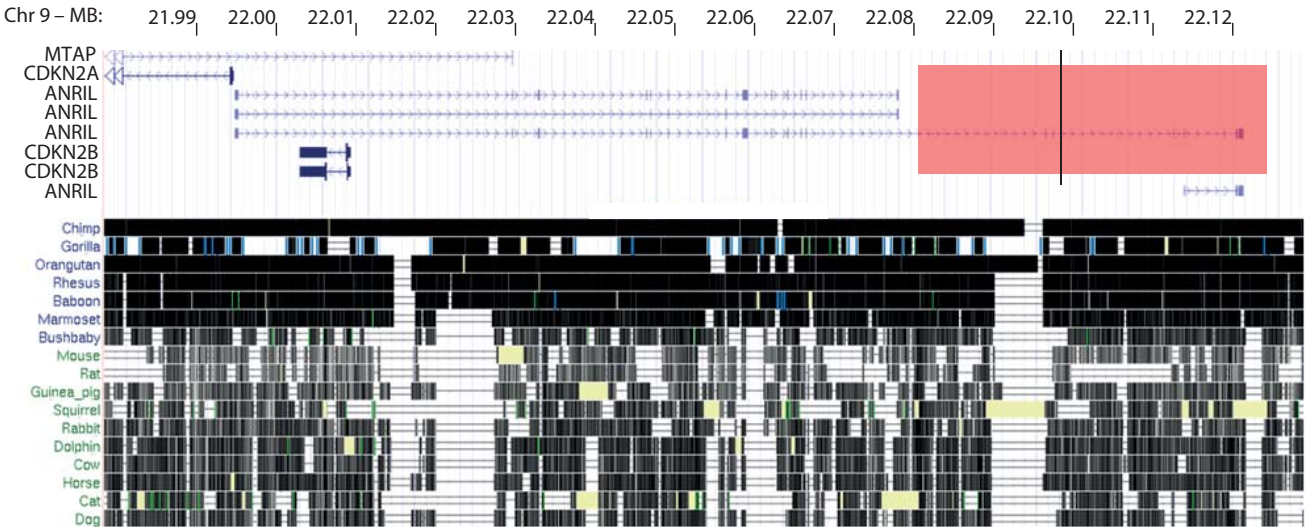

Figure 8. *Overview of the 9p21 region around ANRIL. Gene annotation and alignment to other animal species downloaded from UCSC genome browser (Kent et al. 2002). Note that the UCSC genome browser only shows a few of the ANRIL transcript variants that are known today. Red box indicates the span of the risk haploblock (Broadbent et al. 2007) and the thin black vertical line within indicates the position of rs2891168.*

Shortly after the release of Paper II, it was published that ANRIL had *increased* expression with the 9p21 risk allele in peripheral blood mononuclear cells from 1098 healthy individuals (Holdt et al. 2010) and that ANRIL had *decreased* expression with the 9p21 risk allele in the leukocytes of 487 healthy individuals (Cunnington et al. 2010). In total that brought the number of studies of ANRIL expression and 9p21 risk genotype up to 5, with serious disagreements in observed direction (Table 3). This disagreement has not been clearly resolved, but a number of speculative reasons should be mentioned. Tissue type could be one possible reason. All effects have been observed

in circulating cells and it is possible that this effect is specific to the immune system. This cannot be a full explanation since inconsistent directions are reported for the same cell types. Secondly, alternative splicing could be another reason. As asserted in Paper II, variations in exonic location of measurement points could result in different transcript isoforms being measured. Thirdly, with a high-profile finding like the 9p21 risk locus, there is an increased risk of the so-called winner's curse – which is simply to say that if many different research groups initiated investigations and only the groups with significant results reported on them, there would be a large risk of selecting for false positive findings. Finally, a very possible reason is that half the groups have mistaken the risk locus for the non-risk locus. Unfortunately, this is not entirely unlikely, and would underscore how ambiguous the technical identification of strands is in major genotyping platforms and databases. An allele from any human SNP can currently be reported in at least three strand definitions (dbSNPs forward/reverse, reference genome positive/negative, Illumina TOP/BOT). Various genotyping methods uses various definitions. To safeguard against this, a check of allele frequencies is usually performed during the analysis. In paper III, for example, all indications of direction are accompanied by an indication of frequency. However, because the 9p21 locus has a frequency of close to 50%, a flip of strand would never be revealed to researchers.

	Tissue	Sample size	ANRIL direction with risk locus
(Liu et al. 2009)	Peripheral blood T-cells	170	Decrease
(Jarinova et al. 2009)	Whole blood cells	124	Increase and decrease, depending on transcript
(Folkersen et al. 2009c)	Various, vessel	5 x ~100	No change
(Holdt et al. 2010)	Peripheral blood mononuclear cells	1098	Increase
(Cunnington et al. 2010)	Leukocytes	487	Decrease

Table 3. *Overview of 2009-2010 studies of ANRIL expression and 9p21 risk genotype.*

Naturally other methods have been brought to bear in the functional investigation of the 9p21 SNPs. Long range interaction studies have implicated STAT1 as a downstream target (Harismendy et al. 2011). STAT1 is involved in the response to interferons, providing a putative link to the immune system and a support of circulating cells as study material for the 9p21 locus. Yet other methods have shown that ANRIL is important in epigenetic silencing (Yap et al. 2010) and some groups have even sought to investigate knock-out mouse models of the 9p21 region (Visel et al. 2010), even

though the genetic similarity in this region is so low that it is unlikely that rodents will be very informative (Figure 8).

Finally, the question of alternative splicing was recently revisited in a very interesting study of ANRIL using RNA sequencing and rapid amplification of cDNA ends (Burd et al. 2010). Several new transcript variants were discovered, association with 9p21 risk genotype was found for some of them, and as, perhaps the most fascinating part, it was shown that several of the transcript variants were round rather than linear.

In a broader perspective, the interest in the function of the 9p21 risk genotype mechanism is more than just of academic interest to molecular biology. One stated potential of the GWAS was to provide better medical diagnostics methods. However, the variation in common disease patterns explained by GWAS risk SNPs is now known to be a relative small. A compelling alternative use of the GWAS is therefore to provide knowledge of new pathways and putative drug targets. Such a target could for example be the mechanism that dictates that an individual with a 9p21 risk-allele has a higher risk of early myocardial infarction than an individual without. However, in spite of much research, it is fair to say that we still do not know this mechanism today.

3.3 PAPER III

The purpose of Paper III (Folkersen et al. 2010) was to investigate whether any established risk SNPs were associated to the expression levels of nearby genes. Inspired by the inquiry into the 9p21-risk allele mechanism and its effect on gene expression, we hypothesised that a similar investigation of all other known risk SNPs would reveal more clear cases of eQTLs. At the time of publication, there were 17 SNPs with established association to myocardial infarction (Samani et al. 2007; Erdmann et al. 2009; Kathiresan et al. 2009b). In addition we made a selection of SNPs relevant to risk-factors, intermediate traits and other cardiovascular diseases. These included lipoproteins, triglyceride, hypertension, atherosclerosis, abdominal-aortic aneurysm and waist circumference, and gave a total of 168 established risk SNPs.

The method of Paper III was straightforward: for each of these 168 risk SNPs we identified all genes within 200 kb distance. For each of these genes we compared the expression levels with the genotype of the SNP in question. This was done in each of the tissues available in the BiKE and ASAP biobanks: liver, mammary artery, aorta media, aorta adventitia and carotid plaque. If the association between gene expression and genotype was significant at $P < 0.005$, it was reported as a risk-SNP eQTL.

Using this method we found 47 SNPs that had at least one proximal, significantly associated gene, in at least one of the investigated tissue types. By far the most likely interpretation of this finding is that the risk associated with the SNP is mediated through the gene associated with the SNP. As noted in the discussion, one reason of the popularity of the GWAS is their delineation of cause and effect. The strength of the eQTL approach is that it extends this cause and effect delineation to actual genes. One can assume that the SNP must be the cause of the altered disease risk, and that the SNP must be the cause of the altered gene expression level – if this assumption is true, the altered gene expression level should be the cause of the altered disease risk. And the assumption can indeed be confidently made – a disease or a gene product that affects

the sequence of the DNA of an entire organism is highly improbable. The eQTL method therefore allows the extension of GWAS results from risk-SNPs to risk-genes.

In addition to the presentation of these 47 novel gene targets, the paper contains three considerations of more conceptual nature. The first point is that risk-SNPs are more likely to have eQTL effects than non-risk SNPs. This was shown by comparing the 168 known risk-SNPs with random samplings of non-risk SNPs, each set having similar size and frequency. In 10 samplings and complete re-calculations, there was found no SNP set with stronger eQTL associations than the 168 known risk-SNPs. We therefore made the conclusion that risk-SNPs mechanism is often likely to be mediated through gene expression change.

Secondly, we observed that the SNP was not always found directly in or next to the associated gene. Of the 47 significant eQTLs, 22 were associated with genes at distances of more than 35 kb. Moreover, the most proximal gene only occasionally showed the most significant eQTL effect. This might seem like a trivial observation. However, it goes firmly against the GWAS convention of labelling risk-SNPs with the name of the closest gene or gene-clusters. Based on these results, we conclude that the labelling risk-SNPs by the closest gene-name will often be misleading. Additionally it is worth noting, that we only could conclude on effects within the ±200 kb window. Even more distal eQTL effects would not have been detected. This distance limit was required to control the multiple testing issues. It was based on reports that a majority of general eQTL effects were observed at this distance (Dimas et al. 2009) as well as on the fact that none of these particular risk-SNPs were found in haploblocks of larger size.

Thirdly, we found that the risk-SNP phenotype was often determining the tissue type in which the eQTL was observed. For example the risk-SNPs originally found to be associated with lipid-levels often showed association with gene expression only in liver. Conversely, there were a higher proportion of vessel-wall eQTL effects from risk-SNPs that were originally determined to be associated with myocardial infarction independently of lipid levels. This specificity was even observed when gene expression level where similar or higher in non-eQTL tissues. This observation has implications towards the choice of sample material in future eQTL studies. Because of sampling accessibility, most published eQTL studies are based on samples from circulating blood. This is advantageous when studying risk-SNPs associated with phenotypes of the cells of the circulating blood. When studying risk-SNPs associated with other diseases it is recommendable to study eQTLs in tissue types linked to these diseases.

Taken together, the findings in Paper III illustrate one way to pursue GWAS results further towards clinical application. As with the specific 9p21 functional analysis, it should be beneficial towards the identification of putative drug targets. It has already been beneficial for the GWAS papers in which we have collaborated based on this method (Gretarsdottir et al. 2010; Bown et al. 2011a; C4D 2011; Strawbridge et al. 2011).

3.4 PAPER IV

The purpose of Paper IV (Folkersen et al. 2011) was to analyze the overall gene expression profile of the patients in the ASAP database. The ASAP database was designed with the intention of answering a specific clinical question about thoracic aortic aneurysm (TAA). As described in the introduction section, TAA is a common complication in patients with a bicuspid aortic valve (BAV). It is today not known why TAA consistently occurs at younger age and with a higher frequency in BAV patients than in patients with a tricuspid aortic valve (TAV). The purpose of Paper IV was to investigate this increased TAA incidence in BAV patients.

The fundamental analytical approach of the study was to ask what gene expression changes was observed when comparing the dilated aorta with the non-dilated aorta. This question was asked both for the TAV and the BAV patients. Because these two patient groups both end up with aortic dilation, we hypothesised that observed expression changes that are common to the two groups more likely would be connected to the final common stages of dilation, whereas the observed expression changes unique to either group would be connected to earlier steps in the pathophysiology. When compared to the eQTL-methods described in Paper II and III, this approach certainly has less capacity to determine causality. However, the setup of the question is still an attempt towards delineating cause and effect. Doing this is an advance in comparison to other earlier expression microarray studies which simply reports differentially expressed genes as targets, notwithstanding that they could just as well be up-regulation of repair responses. The obvious strength of the Paper IV over the genetics-methods of Paper II and III is of course that there is more to pathophysiology than genetics.

After establishing the main analytical approach, we were faced with a number of statistical choices. Many different statistical methods have been developed for the analysis of expression microarrays, and many arguments have been made as to why any given test is superior. The one advantage of the Student's T-test is that it is universally recognised and understood. In a translational study that fact is not trivial. The second method we based our studies on, was gene set enrichment analysis (GSEA) (Luo et al. 2009). Although less widespread, this test is attractive for its simple premise: in any defined group of genes, one asks if the genes are more differentially expressed than would have been expected by chance. The definition of gene groups is done separately from the analysis, but is usually taken as those defined by the Gene Ontology, which is a bioinformatics initiative aimed at categorizing all genes (Ashburner et al. 2000). Thus an output from a GSEA might report that in a given gene group, e.g. "vasodilation-related genes", there are many more differentially expressed genes than would have been expected for a similarly sized random group of genes. With these two tests, the Student's T-test and the GSEA, we set out to compare aortic dilation in BAV and TAV patients.

The first major observation was that the difference between the BAV and the TAV patients. Of the genes found to be differentially expressed with dilation, only few

(<4%) were differentially expressed in both BAV and TAV patients. This underscores a point that has only recently been appreciated, namely that these two patient groups exhibit vastly different molecular profiles (Siu et al. 2010). In contrast to this, we observed very little difference when comparing the non-dilated TAV with the non-dilated BAV samples. This observation has impact on the open debate on the role of BAV-dependent hemodynamic changes, because similar pre-dilation profiles in both patient groups corresponds poorly with life-long flow-alteration as an explanation of increased TAA in BAV patients. Lists of individual genes that were differentially expressed in BAV, TAV or both are included as supplementary materials of the paper.

The second major observation was from the GSEA, which showed that the *cell adhesion* and *extracellular region* genes had a much larger proportion of differentially expressed genes than expected by chance. This was true both in the BAV and TAV patients. Because aorta dilation is known as a disease that involves the degradation of structural proteins, this finding was not surprising. The other significant finding from GSEA, however, was that *immune response* genes were highly differentially expressed between BAV and TAV patients. The pattern of this regulation was better visualized using plots of the individual genes, such as figures 4 and 5 of Paper IV. Here it was seen that the observed GSEA scores were a result of a general up-regulation of the *immune response* in dilated aorta media from TAV patients, but not in BAV patients. This finding is clearly novel, and satisfies our initial hypothesis on differences and similarities between the BAV and TAV patients (Figure 9). In other words, we consider this to indicate that the immune response is involved at an earlier point in the aorta dilation pathogenesis. We consider this TAV-specific observation to be an important difference between BAV and TAV patients.

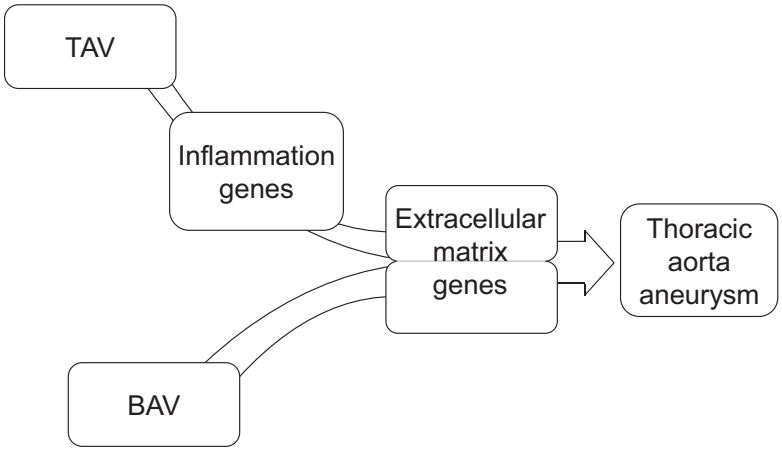

Figure 9. *Common and unique pathways activated in TAA pathogenesis suggest a causal order of events. In this conceptual framework several other established facts have been omitted for the sake of clarity. Examples are the monogenic forms acting through TGF-β (Loeys-Dietz, Marfan) and COL3A1 (Ehlers-Danlos), which illustrate other "starting-points" of disease that could be included here.*

We observed no BAV-specific trends from the GSEA methodology. However, several individual genes were specifically up-regulated only in the dilated aorta of BAV-patients, and these do satisfy our initial hypothesis. It is of interest to note some of these (*PLXNA1*, *EP400* and *BAT2*) have previously been discussed in connection with BAV and TAA (Majumdar et al. 2007; Wooten et al. 2010). In addition, the TGF-beta binding proteins *LTBP3* and *LTBP4* were highly specific for dilation in BAV patients only. TGF-beta has previously been noted for its association to aortic dilation (Loeys et al. 2006; Paloschi et al. 2011). In the scope of this paper, however, we did not perform any further experimental analysis on any of these genes. In addition, it is of interest that one of the individually significant differentially expressed genes is the LRP1 gene, which recently was identified in a GWAS for abdominal aortic aneurysm (Bown et al. 2011b).

Further work on this project is expected to revolve around these gene targets. Because of the recent publication of this paper, however, these projects are only in early planning stages.

3.5 PAPER V

The purpose of Paper V, which is not yet a published article, is to investigate what genomics and transcriptomics can reveal about the medical future of patients with established atherosclerosis. Predicting the medical future of individuals is inherent to any clinical setting in which a prognosis is given. In cardiovascular disease, guidelines typically include consideration of age, gender, smoking, cholesterol, diabetes, e.g. (Roques et al. 2003). Several other emerging biomarkers have been suggested and shown to have predictive properties, e.g. (Hellings et al. 2010; Peeters et al. 2010). In Paper V we ask if an omics-based approach to the question can be advantageous.

Genomics-based approaches to prediction are the essence of the GWAS as already described. To our knowledge, there have not been any GWAS investigating the risk of ischemic events in patients with established atherosclerosis. However, because of the observation that risk-SNPs can be associated with several related phenotypes (Helgadottir et al. 2008), an inviting hypothesis is that risk SNPs for myocardial infarction will have predictive value in similar settings, such as in our patient cohort. This has for example been shown for the 9p21 myocardial infarction risk-SNP in patients undergoing coronary artery bypass surgery (Muehlschlegel et al. 2010). We therefore based the genomics-part of our prediction study on 25 SNPs with firmly established association to myocardial infarction (Kathiresan et al. 2009b; C4D 2011; Schunkert et al. 2011).

The transcriptomics-based part could not benefit from already established risk genes, as no such studies had been performed. The closest such thing was a study of myocardial biopsies, which found 46 genes that were differentially expressed between patients with less than 2 year survival and patients with more than 5 year survival (Heidecker et al. 2008). We therefore decided to pursue a *de novo* discovery of genes with predictive properties. This was done both in the set of 126 carotid plaques and in the overlapping set of 98 peripheral blood mononuclear cell samples (PBMC), for which we had expression microarray data available. To obtain a realistic estimate of the predictive

properties, this was done using a leave-one-out cross-validation setup as previously recommended (Simon et al. 2003).

Taken together, this provided us with risk scores based on the genome of the patients, the transcriptome of the PBMC, and the transcriptome of plaque tissues. For each of these the prognostic strength was calculated using a Cox regression model. The gene expression risk scores from carotid plaque had the best predictive power, at an AUC of 0.72. Gene expression risk scores based on PBMC and scores from risk-SNP genotype had more moderate scores of AUC 0.59 and 0.55, respectively. We then compared these scores with classical risk scores. Classical risk scores were here defined as age, gender, LDL and smoking status, but other combinations of established risk factors were attempted with similar results. The combination of the classical risk scores, with the new transcriptomic and genomic risk scores improved the classical risk score from AUC 0.66 to AUC 0.79. Recall that an AUC of 0.5 corresponds to prediction by pure chance whereas 1.0 corresponds to flawless prediction. So this improvement is respectable.

As an illustration of the prognostic power of these results we give an example of a test based on a set threshold for event prediction. If this threshold is set where it will identify 77% of the patients with future ischemic events, then the number of false positives will be 35% of the remaining patients. However, if gene expression profiling is not performed only 58% of the patients with ischemic events will be detected at the same false positive rate. This can also be stated as a need for testing 6 patients, for every extra correct prediction made with gene expression profiling. We believe that these findings may pave the way for better identification and treatment of patients with increased cardiovascular risk.

One limitation of the study was the necessity of the leave-one-out cross-validation. Optimally results should be validated in an independent cohort. A suitable validation cohort would be a large biobank of carotid plaque samples having both global expression profiling data and clinical follow up information (preferably with a similar time-period and expression profiling method). At the moment, however such studies are left for future collaborations.

In addition, it is important to note the choice of focus. Paper V is focused on the magnitude of improvement of prediction methods using high-throughput methods. We expect these findings to have impact in their own right, but there is another aspect that consequently is not in focus: the specific genes which contain the predictive properties. A main reason for this choice of focus is that we found that no genes were predictive after multiple testing correction. Nonetheless, the data contain several genes which have very significant levels of prediction. Without further validation it is not advisable to proceed with these. However, when that has been resolved it could conceivably overcome several other limitations.

First of all the current sample analysis necessarily restricts the method to patients undergoing carotid endarterectomy. If individual predictive genes could be identified it could provide opportunities for much less invasive measurements of the products of these genes. Positron emission tomography imaging techniques have been developed towards the goal of detecting plaque properties, namely inflammation measured as the degree of fluorodeoxyglucose uptake (Tawakol et al. 2006; Pedersen et al. 2011).

Similarly one could speculate that imaging techniques directed towards a biomarker in the carotid plaque could be informative (Gustafsson et al. 2006). Of course a whole host of technical issues must be overcome for this to be possible.

Secondly it is of course of interest to understand the mechanical basis of why the identified genes are predictive. Ultimately, a goal could be to provide drug targets as well as diagnostics tools. However, in such speculations it is extremely important to keep in mind that this prediction from expression levels is not better suited to establish cause and effect than any other comparative expression studies. The gene that is clearly increased in patients at increased risk of future ischemic attack, might just as well represent an essential healing function. The scope of this study is therefore solely diagnostic.

4 CONCLUSIONS

The overarching purpose of this thesis was to investigate the expression of human genes and how they relate to cardiovascular disease. With this, it was hoped that a contribution could be made to the larger goal of improved treatment of cardiovascular disease. It is probably impossible to quantify how much improvement was in fact contributed, if anything at all. However, that can be argued with most basic research and it is nonetheless basic research that drives the discoveries that ultimately improve treatment. We therefore focus on the immediate basic research objectives that were accomplished:

- We provided the first study of the transcript isoforms of the genes in the 9p21 cardiovascular risk SNP region.
- We identified 47 additional novel target genes using the eQTL methods on other risk SNPs. Additionally we contributed to a concept of systematically checking newly discovered GWAS results for eQTL effects, in order to move from target-SNPs to target-genes.
- We provided the first analysis of the gene expression differences between aorta dilation of BAV and TAV patients. This highlighted a possible fundamental difference in pathophysiology of these two forms of aortic aneurysm.
- We showed as proof-of-principle that gene expression signatures from carotid plaque samples are able to measurably improve prediction of future myocardial infarctions and ischemic strokes.

The underlying theme throughout the thesis has been that biobanks with human samples are crucial in answering fundamental medical questions. The overall conclusion therefore is that biobanks are important in translational research, and that future medical research to an even higher degree would benefit from investment in biobanks.

5 ACKNOWLEDGEMENTS

The first and the warmest acknowledgements will go to my supervisors.

Thanks to **Anders Gabrielsen** for inviting me to Stockholm in the first place. It is a great shame that our initial salt and microgravity project still has not been published to a larger audience interested in space biology. Perhaps it really is as you say that people are more interested in a disease that affects millions. Thank you for your ability to always appear out of no-where with a swarm of relevant and creative suggestions.

Thanks to **Per Eriksson** for providing such a great environment for creative thinking. Thank you so much for all the support, encouragement and guidance over the past years.

Thanks to **Gabrielle Paulsson-Berne** for always being there. It is only because of the immense work that you have done in biobanking that this thesis is even possible.

Thanks to my family, my parents **Per** and **Kirsten** in Odense and my sister **Emilie** in Oslo. Perhaps one day we will manage to live in the same country. Until then, I think we are all doing a very good job at keeping in touch in spite of long distances. Thanks also to my in-law family; my "placebo"-mother and father in-law, **Yudi Pawitan** and **Marie Reilly**. And my real mother and father in-law,莫永仙 and韦自国 (which I unfortunately am not allowed to know how to pronounce, because it would be insulting if I were to call them anything but **ma** and **pa**). And of course my sister-in-law, jiějiě, **Tianhong Wei**.

Thanks to all my friends and colleagues: **Valentina Paloschi**, **Anders Franco-Cereceda**, **Anders Hamsten**, **Göran Hansson**, **Sanela Kurtovic**, **Sarah-Jayne Reilly**, **Anders Mälarstig**, **Hanna Agardh**, **Phil Cheong**, **Rona Strawbridge**, **Ulf Hedin**, **Anuj Goel**, **Daniel Johansson**, **Dick Wågsäter**, **Hanna Björk**, **Stefan Gavelin**, **Andreas Bornø**, **Angela Silveira**, **Anton Razuvaev**, **David Eldrup**, **Elena Gabets**, **Ferdinand van't Hooft**, **Janne Schrøder Jensen**, **Jesper Swedenborg**, **Johan Ekstrand**, **John Öhrvik**, **Simone Picelli**, **Stefano Sorrentino**, **Søren Baunø**, **Tatjana Rebesa**, **Theodosios Kyriakou**, **Therese Olsson**, **Torsten Schultz-Larsen**, **Andreas Hougaard**, **Bengt Sennblad**, **Ekaterina Chernogubova**, **Jing Wang**, **Joëlle Magné**, **Karl Gertow**, **Magnus Bäck**, **Maria Nastase Mannila**, **Maria Sabater Lleal**, **Olga Ovchinnikova**, **Shohreh Maleki**, **Alexandra Bäcklund**, **Andre Strodthoff**, **Andreas Hermansson**, **Anna Aminoff**, **Anna Deleskog**, **Anton Gisterå**, **Björn Gustafsson**, **Brandon Hurrie**, **Brendan Ivey**, **Cao Jia**, **Cecilia Österholm Corbascio**, **Cheryl Xiaoying Zhang**, **Clara Ibel Chamorro**, **Daniel Figueiredo**, **Daniel Ketelhuth**, **Daniela Strodthoff**, **Edit nagy**, **Emina Vorkapic**, **Ewa Ehrenborg**, **Florian Meisgen**, **Hanane M'Hamdi**, **Hovsep Mahdessian**, **Joanna Chmielewska**, **John Andersson**, **Jonas Persson**, **Karin Danell-Toverud**, **Leif Söderström**, **Lin Cao**, **Lollo Sjöholm**, **Louisa Cheung**, **Magnus Kjærgaard**, **Magnus Mossfeldt**, **Malin Larsson**, **Marcelo Petri**, **Mari Männik**, **Maria Gonzalez Diez**, **Maria Iglesias**, **Maria Klement**, **Olga Piksasova**, **Olivera Werngren**, **Peder Olofsson**, **Rachel Fisher**, **Rehannah Borup**, **Robert Badeau**, **Sabrina Görgen**, **Said Zeiai**, **Santi Sole**, **Stanley Cheuk**, **Susanne Brauner**, **Tatjana Adamovic**, **Yajuan Wang**, **Zhong-Qun Yan**. Each person on this list is somebody that I am very lucky to have met, and who deserves to be acknowledged. Perhaps I should have written a

special note for each, but with such a long list of fantastic people the acknowledgements section would have extended dozens of pages. Instead I have chosen to present the list sorted by a value calculated as the number of co-authored papers plus the number of co-tagged facebook pictures.

In addition, I'd like to give special thanks to the administrators, **Ami Björkholm** and **Ann Hellström**. Thank you for helping me out with all the exceptions that comes from being registered in two research groups.

Also, a special thanks to the lab-technicians, **Anneli Olsson**, **Ingrid Törnberg**, **Karin Husman**, **Linda Haglund**, and **Therese Olsson**. Without your work there would be no biobanks to analyze.

To the students that I have worked with: **Sofi Lanevik** and our work on western blots; **Jesper Gådin**, **Johan Dahlberg**, **Niklas Malmqvist** for their excellent work on ExpressionWebExpress; **Emma Koopmans** and her work on microarray normalization – everything in this thesis is ultimately affected by that report.

Very many thanks to **Carlsberg** and the **Danish Agency for Science, Technology and Innovation (DASTI)** for generous grants which made it possible for me to move to Sweden and stay here for 5 years.

Finally, huge thanks to my wife **Tianling Wei**. I am not going to use big words here, because I think you already know what I want to say. I'll hint you though; I am looking forward to the spring day soon when we will be walking in the park with '**Neil Armstrong**', our unborn and unnamed baby-boy or girl.

Thanks in advance to you too, Nelly – for not coming out before the dissertation date.

6 REFERENCES

Agardh, H. E., L. Folkersen, J. Ekstrand, D. Marcus, J. Swedenborg, U. Hedin, A. Gabrielsen and G. Paulsson-Berne (2011). "Expression of fatty acid-binding protein 4/aP2 is correlated with plaque instability in carotid atherosclerosis." J Intern Med **269**(2): 200-210.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Barth, A. S., R. Kuner, A. Buness, M. Ruschhaupt, S. Merk, L. Zwermann, S. Kääb, E. Kreuzer, G. Steinbeck, U. Mansmann, A. Poustka, M. Nabauer and H. Sültmann (2006). "Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies." J Am Coll Cardiol **48**(8): 1610-1617.

Barth, A. S., S. Merk, E. Arnoldi, L. Zwermann, P. Kloos, M. Gebauer, K. Steinmeyer, M. Bleich, S. KÇÏÇÏb, M. Hinterseer, H. Kartmann, E. Kreuzer, M. Dugas, G. Steinbeck and M. Nabauer (2005). "Reprogramming of the human atrial transcriptome in permanent atrial fibrillation: expression of a ventricular-like genomic signature." Circ Res **96**(9): 1022-1029.

Borup, R., M. Rossing, R. Henao, Y. Yamamoto, A. Krogdahl, C. Godballe, O. Winther, K. Kiss, L. Christensen, E. Hogdall, F. Bennedbaek and F. C. Nielsen (2010). "Molecular signatures of thyroid follicular neoplasia." Endocr Relat Cancer **17**(3): 691-708.

Bown, M. J., G. T. Jones, S. C. Harrison, B. J. Wright, S. Bumpstead, A. F. Baas, S. Gretarsdottir, S. A. Badger, D. T. Bradley, K. Burnand, A. H. Child, R. E. Clough, G. Cockerill, H. Hafez, A. S. D. Julian, S. Futers, et al. (2011a). "Abdominal Aortic Aneurysm Is Associated with a Variant in Low-Density Lipoprotein Receptor-Related Protein 1." Am J Hum Genet.

Bown, M. J., G. T. Jones, S. C. Harrison, B. J. Wright, S. Bumpstead, A. F. Baas, S. Gretarsdottir, S. A. Badger, D. T. Bradley, K. Burnand, A. H. Child, R. E. Clough, G. Cockerill, H. Hafez, D. J. A. Scott, S. Futers, et al. (2011b). "Abdominal aortic aneurysm is associated with a functional variant in the Low density lipoprotein receptor Related Protein 1 (LRP1) gene." Am J Hum Genet **[in press]**.

Breland, U. M., A. E. Michelsen, M. Skjelland, L. Folkersen, K. Krohg-Sorensen, D. Russell, T. Ueland, A. Yndestad, G. Paulsson-Berne, J. K. Damas, E. Oie, G. K. Hansson, B. Halvorsen and P. Aukrust (2010). "Raised MCP-4 levels in symptomatic carotid atherosclerosis: an inflammatory link between platelet and monocyte activation." Cardiovasc Res.

Broadbent, H. M., J. F. Peden, S. Lorkowski, A. Goel, H. Ongen, F. Green, R. Clarke, R. Collins, M. G. Franzosi, G. Tognoni, U. Seedorf, S. Rust, P. Eriksson, A. Hamsten, M. Farrall and H. Watkins (2007). "Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked, SNPs in the ANRIL locus on chromosome 9p." Hum Mol Genet.

Burd, C. E., W. R. Jeck, Y. Liu, H. K. Sanoff, Z. Wang and N. E. Sharpless (2010). "Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk." PLoS Genet **6**(12): e1001233.

C4D (2011). "A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease." Nat Genet **43**(4): 339-344.

Calza, S. and Y. Pawitan (2010). "Normalization of gene-expression microarray data." Methods Mol Biol **673**: 37-52.

Canales, R. D., Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E. H. Peters, K. Poulter, P. L. Ruppel, et al. (2006). "Evaluation of DNA microarray results with quantitative gene expression platforms." Nat Biotechnol **24**(9): 1115-1122.

Chon, H., C. A. Gaillard, B. B. van der Meijden, H. M. Dijstelbloem, R. J. Kraaijenhagen, D. van Leenen, F. C. Holstege, J. A. Joles, H. A. Bluyssen, H. A. Koomans and B. Braam (2004). "Broadly altered gene expression in blood leukocytes in essential hypertension is absent during treatment." Hypertension **43**(5): 947-951.

Clarke, R., J. F. Peden, J. C. Hopewell, T. Kyriakou, A. Goel, S. C. Heath, S. Parish, S. Barlera, M. G. Franzosi, S. Rust, D. Bennett, A. Silveira, A. Malarstig, F. R. Green, M. Lathrop, B. Gigante, et al. (2009). "Genetic variants associated with Lp(a) lipoprotein level and coronary disease." N Engl J Med **361**(26): 2518-2528.

Cripe, L., G. Andelfinger, L. J. Martin, K. Shooner and D. W. Benson (2004). "Bicuspid aortic valve is heritable." J Am Coll Cardiol **44**(1): 138-143.

Cunnington, M. S., M. Santibanez Koref, B. M. Mayosi, J. Burn and B. Keavney (2010). "Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression." PLoS Genet **6**(4): e1000899.

Danesh, J. and M. B. Pepys (2009). "C-reactive protein and coronary disease: is there a causal link?" Circulation **120**(21): 2036-2039.

Diez, D., A. M. Wheelock, S. Goto, J. Z. Haeggstrom, G. Paulsson-Berne, G. K. Hansson, U. Hedin, A. Gabrielsen and C. E. Wheelock (2010). "The use of network analyses for elucidating mechanisms in cardiovascular disease." Mol Biosyst **6**(2): 289-304.

Dimas, A. S., S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, M. Gutierrez Arcelus, M. Sekowska, M. Gagnebin, J. Nisbett, P. Deloukas, E. T. Dermitzakis and S. E. Antonarakis (2009). "Common regulatory variation impacts gene expression in a cell type-dependent manner." Science **325**(5945): 1246-1250.

Durbin, R. M., G. R. Abecasis, D. L. Altshuler, A. Auton, L. D. Brooks, R. A. Gibbs, M. E. Hurles and G. A. McVean (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

ECST (1998). "Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European Carotid Surgery Trial (ECST)." Lancet **351**(9113): 1379-1387.

Edgar, R., M. Domrachev and A. E. Lash (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.

El-Hamamsy, I. and M. H. Yacoub (2009). "Cellular and molecular mechanisms of thoracic aortic aneurysms." Nat Rev Cardiol **6**(12): 771-786.

Erdmann, J., A. Grosshennig, P. S. Braund, I. R. Konig, C. Hengstenberg, A. S. Hall, P. Linsel-Nitschke, S. Kathiresan, B. Wright, D. A. Tregouet, F. Cambien, P. Bruse, Z. Aherrahrou, A. K. Wagner, K. Stark, S. M. Schwartz, et al. (2009). "New susceptibility locus for coronary artery disease on chromosome 3q22.3." Nat Genet **41**(3): 280-282.

Ferguson, G. G., M. Eliasziw, H. W. Barr, G. P. Clagett, R. W. Barnes, M. C. Wallace, D. W. Taylor, R. B. Haynes, J. W. Finan, V. C. Hachinski and H. J. Barnett (1999). "The North American Symptomatic Carotid Endarterectomy Trial : surgical results in 1415 patients." Stroke **30**(9): 1751-1758.

Folkersen, L., D. Diez, C. E. Wheelock, J. Z. Haeggstrom, S. Goto, P. Eriksson and A. Gabrielsen (2009a). "GeneRegionScan: a Bioconductor package for probe-level analysis of specific, small regions of the genome." Bioinformatics **25**(15): 1978-1979.

Folkersen, L., S. Kurtovic, A. Razuvaev, H. E. Agardh, A. Gabrielsen and G. Paulsson-Berne (2009b). "Endogenous control genes in complex vascular tissue samples." BMC Genomics **10**: 516.

Folkersen, L., T. Kyriakou, A. Goel, J. Peden, A. Malarstig, G. Paulsson-Berne, A. Hamsten, W. Hugh, A. Franco-Cereceda, A. Gabrielsen and P. Eriksson (2009c). "Relationship between CAD risk genotype in the chromosome 9p21 locus and gene expression. Identification of eight new ANRIL splice variants." PLoS ONE **4**(11): e7677.

Folkersen, L., D. Wagsater, V. Paloschi, V. Jackson, J. Petrini, S. Kurtovic, S. Maleki, M. J. Eriksson, K. Caidahl, A. Hamsten, J. B. Michel, J. Liska, A. Gabrielsen, A. Franco-Cereceda and P. Eriksson (2011). "Unraveling the divergent gene expression profiles in bicuspid and tricuspid aortic valve patients with thoracic aortic dilatation - the ASAP study." Mol Med.

Folkersen, L., F. van't Hooft, E. Chernogubova, H. E. Agardh, G. K. Hansson, U. Hedin, J. Liska, A. C. Syvanen, G. Paulsson-Berne, A. Franco-Cereceda, A. Hamsten, A. Gabrielsen and P. Eriksson (2010). "Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease." Circ Cardiovasc Genet **3**(4): 365-373.

Gabrielsen, A., H. Qiu, M. Back, M. Hamberg, A. L. Hemdahl, H. Agardh, L. Folkersen, J. Swedenborg, U. Hedin, G. Paulsson-Berne, J. Z. Haeggstrom and G. K. Hansson (2010). "Thromboxane synthase expression and thromboxane A2 production in the atherosclerotic lesion." J Mol Med (Berl) **88**(8): 795-806.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Gertow, K., E. Nobili, L. Folkersen, J. W. Newman, T. L. Pedersen, J. Ekstrand, J. Swedenborg, H. Kuhn, C. E. Wheelock, G. K. Hansson, U. Hedin, J. Z. Haeggstrom and A. Gabrielsen (2011). "12- and 15-lipoxygenases in human carotid atherosclerotic lesions: associations with cerebrovascular symptoms." Atherosclerosis **215**(2): 411-416.

Goldstein, J. L. and M. S. Brown (1974). "Binding and degradation of low density lipoproteins by cultured human fibroblasts. Comparison of cells from a normal subject and from a patient with homozygous familial hypercholesterolemia." J Biol Chem **249**(16): 5153-5162.

Goldstein, J. L. and M. S. Brown (2009). "The LDL receptor." Arterioscler Thromb Vasc Biol **29**(4): 431-438.

Goncalves, I., K. Stenstrom, G. Skog, S. Mattsson, M. Nitulescu and J. Nilsson (2010). "Short communication: Dating components of human atherosclerotic plaques." Circ Res **106**(6): 1174-1177.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, et al. (2010). "A draft sequence of the Neandertal genome." Science **328**(5979): 710-722.

Gretarsdottir, S., A. F. Baas, G. Thorleifsson, H. Holm, M. den Heijer, J. P. de Vries, S. E. Kranendonk, C. J. Zeebregts, S. M. van Sterkenburg, R. H. Geelkerken, A. M. van Rij, M. J. Williams, A. P. Boll, J. P. Kostic, A. Jonasdottir, G. B. Walters, et al. (2010). "Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm." Nat Genet **42**(8): 692-697.

Gustafsson, B., S. Youens and A. Y. Louie (2006). "Development of contrast agents targeted to macrophage scavenger receptors for MRI of vascular inflammation." Bioconjug Chem **17**(2): 538-547.

HapMap (2003). "The International HapMap Project." Nature **426**(6968): 789-796.

Harismendy, O., D. Notani, X. Song, N. G. Rahim, B. Tanasa, N. Heintzman, B. Ren, X. D. Fu, E. J. Topol, M. G. Rosenfeld and K. A. Frazer (2011). "9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response." Nature **470**(7333): 264-268.

Heidecker, B., E. K. Kasper, I. S. Wittstein, H. C. Champion, E. Breton, S. D. Russell, M. M. Kittleson, K. L. Baughman and J. M. Hare (2008). "Transcriptomic biomarkers for individual risk assessment in new-onset heart failure." Circulation **118**(3): 238-246.

Helgadottir, A., G. Thorleifsson, K. P. Magnusson, S. Gretarsdottir, V. Steinthorsdottir, A. Manolescu, G. T. Jones, G. J. Rinkel, J. D. Blankensteijn, A. Ronkainen, J. E. Jaaskelainen, Y. Kyo, G. M. Lenk, N. Sakalihasan, K. Kostulas, A. Gottsater, et al. (2008). "The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm." Nat Genet **40**(2): 217-224.

Helgadottir, A., G. Thorleifsson, A. Manolescu, S. Gretarsdottir, T. Blondal, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, A. Baker, A. Palsson, G. Masson, D. F. Gudbjartsson, K. P. Magnusson, K. Andersen, A. I. Levey, V. M. Backman, et al. (2007). "A common

variant on chromosome 9p21 affects the risk of myocardial infarction." Science **316**(5830): 1491-1493.

Hellings, W. E., W. Peeters, F. L. Moll, S. R. Piers, J. van Setten, P. J. Van der Spek, J. P. de Vries, K. A. Seldenrijk, P. C. De Bruin, A. Vink, E. Velema, D. P. de Kleijn and G. Pasterkamp (2010). "Composition of carotid atherosclerotic plaque is associated with cardiovascular outcome: a prognostic study." Circulation **121**(17): 1941-1950.

Holdt, L. M., F. Beutner, M. Scholz, S. Gielen, G. Gabel, H. Bergert, G. Schuler, J. Thiery and D. Teupser (2010). "ANRIL Expression Is Associated With Atherosclerosis Risk at Chromosome 9p21." Arterioscler Thromb Vasc Biol **30**(3): 620-627.

Hope, M. D., T. A. Hope, A. K. Meadows, K. G. Ordovas, T. H. Urbania, M. T. Alley and C. B. Higgins (2010). "Bicuspid aortic valve: four-dimensional MR evaluation of ascending aortic systolic flow patterns." Radiology **255**(1): 53-61.

Hurks, R., W. Peeters, W. J. Derksen, W. E. Hellings, I. E. Hoefer, F. L. Moll, D. P. de Kleijn and G. Pasterkamp (2009). "Biobanks and the search for predictive biomarkers of local and systemic outcome in atherosclerotic disease." Thromb Haemost **101**(1): 48-54.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.

Jackson, V., T. Olsson, S. Kurtovic, L. Folkersen, V. Paloschi, D. Wagsater, A. Franco-Cereceda and P. Eriksson (2011a). "Matrix metalloproteinase 14 and 19 expression is associated with thoracic aortic aneurysms." J Thorac Cardiovasc Surg.

Jackson, V., J. Petrini, K. Caidahl, M. J. Eriksson, J. Liska, P. Eriksson and A. Franco-Cereceda (2011b). "Bicuspid aortic valve leaflet morphology in relation to aortic root morphology: a study of 300 patients undergoing open-heart surgery." Eur J Cardiothorac Surg **40**(3): e118-124.

Jarinova, O., A. F. Stewart, R. Roberts, G. Wells, P. Lau, T. Naing, C. Buerki, B. W. McLean, R. C. Cook, J. S. Parker and R. McPherson (2009). "Functional Analysis of the Chromosome 9p21.3 Coronary Artery Disease Risk Locus." Arterioscler Thromb Vasc Biol.

Kathiresan, S., C. J. Willer, G. M. Peloso, S. Demissie, K. Musunuru, E. E. Schadt, L. Kaplan, D. Bennett, Y. Li, T. Tanaka, B. F. Voight, L. L. Bonnycastle, A. U. Jackson, G. Crawford, A. Surti, C. Guiducci, et al. (2009a). "Common variants at 30 loci contribute to polygenic dyslipidemia." Nat Genet **41**(1): 56-65.

Kathiresan, S., B. F. Voight, S. Purcell, K. Musunuru, D. Ardissino, P. M. Mannucci, S. Anand, J. C. Engert, N. J. Samani, H. Schunkert, J. Erdmann, M. P. Reilly, D. J. Rader, T. Morgan, J. A. Spertus, M. Stoll, et al. (2009b). "Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants." Nat Genet **41**(3): 334-341.

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

Klein, R. J., C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable and J. Hoh (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **308**(5720): 385-389.

Kurtovic, S., V. Paloschi, L. Folkersen, J. Gottfries, A. Franco-Cereceda and P. Eriksson (2011). "Diverging alternative splicing fingerprints in the transforming growth factor-beta signaling pathway identified in thoracic aortic aneurysms." Mol Med **17**(7-8): 665-675.

Kwan, T., D. Benovoy, C. Dias, S. Gurd, C. Provencher, P. Beaulieu, T. J. Hudson, R. Sladek and J. Majewski (2008). "Genome-wide analysis of transcript isoform variation in humans." Nat Genet **40**(2): 225-231.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lemaire, S. A., M. L. McDonald, D. C. Guo, L. Russell, C. C. Miller, 3rd, R. J. Johnson, M. R. Bekheirnia, L. M. Franco, M. Nguyen, R. E. Pyeritz, J. E. Bavaria, R. Devereux, C. Maslen, K. W. Holmes, K. Eagle, S. C. Body, et al. (2011). "Genome-wide association study identifies a susceptibility locus for thoracic aortic aneurysms and aortic dissections spanning FBN1 at 15q21.1." Nat Genet **43**(10): 996-1000.

Li, M., I. X. Wang, Y. Li, A. Bruzel, A. L. Richards, J. M. Toung and V. G. Cheung (2011). "Widespread RNA and DNA sequence differences in the human transcriptome." Science **333**(6038): 53-58.

Li, Y., C. J. Willer, J. Ding, P. Scheet and G. R. Abecasis (2010). "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." Genet Epidemiol **34**(8): 816-834.

Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, M. Kamal, M. Clamp, J. L. Chang, E. J. Kulbokas, 3rd, M. C. Zody, E. Mauceli, X. Xie, M. Breen, R. K. Wayne, E. A. Ostrander, C. P. Ponting, et al. (2005). "Genome sequence, comparative analysis and haplotype structure of the domestic dog." Nature **438**(7069): 803-819.

Liu, Y., H. K. Sanoff, H. Cho, C. E. Burd, C. Torrice, K. L. Mohlke, J. G. Ibrahim, N. E. Thomas and N. E. Sharpless (2009). "INK4/ARF transcript expression is associated with chromosome 9p21 variants linked to atherosclerosis." PLoS ONE **4**(4): e5027.

Loeys, B. L., U. Schwarze, T. Holm, B. L. Callewaert, G. H. Thomas, H. Pannu, J. F. De Backer, G. L. Oswald, S. Symoens, S. Manouvrier, A. E. Roberts, F. Faravelli, M. A. Greco, R. E. Pyeritz, D. M. Milewicz, P. J. Coucke, et al. (2006). "Aneurysm syndromes caused by mutations in the TGF-beta receptor." N Engl J Med **355**(8): 788-798.

Luo, W., M. S. Friedman, K. Shedden, K. D. Hankenson and P. J. Woolf (2009). "GAGE: generally applicable gene set enrichment for pathway analysis." BMC Bioinformatics **10**: 161.

Majumdar, R., D. V. Miller, K. V. Ballman, G. Unnikrishnan, S. H. McKellar, G. Sarkar, R. Sreekumar, M. E. Bolander and T. M. Sundt, 3rd (2007). "Elevated expressions of osteopontin and tenascin C in ascending aortic aneurysms are associated with trileaflet aortic valves as compared with bicuspid aortic valves." Cardiovasc Pathol **16**(3): 144-150.

Malarstig, A. and A. Hamsten (2010). "Genetics of atherothrombosis and thrombophilia." Curr Atheroscler Rep **12**(3): 159-166.

Malarstig, A., S. Sigurdsson, P. Eriksson, G. Paulsson-Berne, U. Hedin, L. Wallentin, A. Siegbahn, A. Hamsten and A. C. Syvanen (2008). "Variants of the interferon regulatory factor 5 gene regulate expression of IRF5 mRNA in atherosclerotic tissue but are not associated with myocardial infarction." Arterioscler Thromb Vasc Biol **28**(5): 975-982.

McPherson, R., A. Pertsemlidis, N. Kavaslar, A. Stewart, R. Roberts, D. R. Cox, D. A. Hinds, L. A. Pennacchio, A. Tybjaerg-Hansen, A. R. Folsom, E. Boerwinkle, H. H. Hobbs and J. C. Cohen (2007). "A common allele on chromosome 9 associated with coronary heart disease." Science **316**(5830): 1488-1491.

Mikkelsen, T. S., L. W. Hillier, E. E. Eichler, M. C. Zody, D. B. Jaffe, S. P. Yang, W. Enard, I. Hellmann, K. Lindblad-Toh, T. K. Altheide, N. Archidiacono, P. Bork, J. Butler, J. L. Chang, Z. Cheng, A. T. Chinwalla, et al. (2005). "Initial sequence of the chimpanzee genome and comparison with the human genome." Nature **437**(7055): 69-87.

Muehlschlegel, J. D., K. Y. Liu, T. E. Perry, A. A. Fox, C. D. Collard, S. K. Shernan and S. C. Body (2010). "Chromosome 9p21 variant predicts mortality after coronary artery bypass graft surgery." Circulation **122**(11 Suppl): S60-65.

Nkomo, V. T., M. Enriquez-Sarano, N. M. Ammash, L. J. Melton, 3rd, K. R. Bailey, V. Desjardins, R. A. Horn and A. J. Tajik (2003). "Bicuspid aortic valve associated with aortic dilatation: a community-based study." Arterioscler Thromb Vasc Biol **23**(2): 351-356.

Nordestgaard, B. G. and A. Tybjaerg-Hansen (2011). "Genetic determinants of LDL, lipoprotein(a), triglyceride-rich lipoproteins and HDL: concordance and discordance with cardiovascular disease risk." Curr Opin Lipidol **22**(2): 113-122.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens and C. D. Bustamante (2008). "Genes mirror geography within Europe." Nature **456**(7218): 98-101.

Olofsson, P. S., L. A. Soderstrom, C. Jern, A. Sirsjo, M. Ria, E. Sundler, U. de Faire, P. G. Wiklund, J. Ohrvik, U. Hedin, G. Paulsson-Berne, A. Hamsten, P. Eriksson and G. K. Hansson (2009). "Genetic variants of TNFSF4 and risk for carotid artery disease and stroke." J Mol Med (Berl) **87**(4): 337-346.

Paloschi, V., S. Kurtovic, L. Folkersen, D. Gomez, D. Wagsater, J. Roy, J. Petrini, M. J. Eriksson, K. Caidahl, A. Hamsten, J. Liska, J. B. Michel, A. Franco-Cereceda and P. Eriksson (2011). "Impaired splicing of fibronectin is associated with thoracic aortic aneurysm formation in patients with bicuspid aortic valve." Arterioscler Thromb Vasc Biol **31**(3): 691-697.

Pasmant, E., I. Laurendeau, D. Heron, M. Vidaud, D. Vidaud and I. Bieche (2007). "Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF." Cancer Res **67**(8): 3963-3969.

Pedersen, S. F., M. Graebe, A. M. Hag, L. Hoejgaard, H. Sillesen and A. Kjaer (2011). "Microvessel Density But Not Neoangiogenesis Is Associated with (18)F-FDG Uptake in Human Atherosclerotic Carotid Plaques." Mol Imaging Biol.

Pedersen, T. R., J. Kjekshus, K. Berg, T. Haghfelt, O. Faergeman, G. Faergeman, K. Pyörälä, T. Miettinen, L. Wilhelmsen, A. Olsson, H. Wedel and S. S. S. S. Group (1994). "Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S)." Lancet **344**(8934): 1383-1389.

Peeters, W., D. P. de Kleijn, A. Vink, S. van de Weg, A. H. Schoneveld, S. K. Sze, P. J. van der Spek, J. P. de Vries, F. L. Moll and G. Pasterkamp (2010). "Adipocyte fatty acid binding protein in atherosclerotic plaques is associated with local vulnerability and is predictive for the occurrence of adverse cardiovascular events." Eur Heart J.

Phillippi, J. A., E. A. Klyachko, J. P. t. Kenny, M. A. Eskay, R. C. Gorman and T. G. Gleason (2009). "Basal and oxidative stress-induced expression of metallothionein is decreased in ascending aortic aneurysms of bicuspid aortic valve patients." Circulation **119**(18): 2498-2506.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J. B. Veyrieras, M. Stephens, Y. Gilad and J. K. Pritchard (2010). "Understanding mechanisms underlying human gene expression variation with RNA sequencing." Nature **464**(7289): 768-772.

Purcell, S., S. S. Cherny and P. C. Sham (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits." Bioinformatics **19**(1): 149-150.

Qiu, H., A. Gabrielsen, H. E. Agardh, M. Wan, A. Wetterholm, C. H. Wong, U. Hedin, J. Swedenborg, G. K. Hansson, B. Samuelsson, G. Paulsson-Berne and J. Z. Haeggstrom (2006). "Expression of 5-lipoxygenase and leukotriene A4 hydrolase in human atherosclerotic lesions correlates with symptoms of plaque instability." Proc Natl Acad Sci U S A **103**(21): 8161-8166.

R Development Core Team. (2011). "R: A Language and Environment for Statistical Computing." 2011, from http://www.R-project.org.

Ramirez, F. and H. C. Dietz (2007). "Marfan syndrome: from molecular pathogenesis to clinical treatment." Curr Opin Genet Dev **17**(3): 252-258.

Razuvaev, A., J. Ekstrand, L. Folkersen, H. Agardh, D. Markus, J. Swedenborg, G. K. Hansson, A. Gabrielsen, G. Paulsson-Berne, J. Roy and U. Hedin (2011). "Correlations Between Clinical Variables and Gene-expression Profiles in Carotid Plaque Instability." Eur J Vasc Endovasc Surg.

Ridker, P. M., E. Danielson, F. A. Fonseca, J. Genest, A. M. Gotto, Jr., J. J. Kastelein, W. Koenig, P. Libby, A. J. Lorenzatti, J. G. MacFadyen, B. G. Nordestgaard, J. Shepherd, J. T. Willerson and R. J. Glynn (2008). "Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein." N Engl J Med **359**(21): 2195-2207.

Roques, F., P. Michel, A. R. Goldstone and S. A. Nashef (2003). "The logistic EuroSCORE." Eur Heart J **24**(9): 881-882.

Saksi, J., P. Ijas, K. Nuotio, R. Sonninen, L. Soinne, O. Salonen, E. Saimanen, J. Tuimala, E. M. Lehtonen-Smeds, M. Kaste, P. T. Kovanen and P. J. Lindsberg (2011). "Gene expression differences between stroke-associated and asymptomatic carotid plaques." J Mol Med (Berl) **89**(10): 1015-1026.

Samani, N. J., J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R. J. Dixon, T. Meitinger, P. Braund, H. E. Wichmann, J. H. Barrett, I. R. Konig, S. E. Stevens, S. Szymczak, D. A. Tregouet, M. M. Iles, et al. (2007). "Genomewide association analysis of coronary artery disease." N Engl J Med **357**(5): 443-453.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Schunkert, H., I. R. Konig, S. Kathiresan, M. P. Reilly, T. L. Assimes, H. Holm, M. Preuss, A. F. Stewart, M. Barbalic, C. Gieger, D. Absher, Z. Aherrahrou, H. Allayee, D. Altshuler, S. S. Anand, K. Andersen, et al. (2011). "Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease." Nat Genet **43**(4): 333-338.

Shalhoub, J., K. J. Davies, N. Hasan, A. Thapar, P. Sharma and A. H. Davies (2011). "The Utility of Collaborative Biobanks for Cardiovascular Research." Angiology.

Shepherd, J., S. M. Cobbe, I. Ford, C. G. Isles, A. R. Lorimer, P. W. MacFarlane, J. H. McKillop and C. J. Packard (1995). "Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. West of Scotland Coronary Prevention Study Group." N Engl J Med **333**(20): 1301-1307.

Shi, L., L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, et al. (2006). "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." Nat Biotechnol **24**(9): 1151-1161.

Simon, R., M. D. Radmacher, K. Dobbin and L. M. McShane (2003). "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification." J Natl Cancer Inst **95**(1): 14-18.

Siu, S. C. and C. K. Silversides (2010). "Bicuspid aortic valve disease." J Am Coll Cardiol **55**(25): 2789-2800.

Strawbridge, R. J., J. Dupuis, I. Prokopenko, A. Barker, E. Ahlqvist, D. Rybin, J. R. Petrie, M. E. Travers, N. Bouatia-Naji, A. S. Dimas, A. Nica, E. Wheeler, H. Chen, B. F. Voight, J. Taneera, S. Kanoni, et al. (2011). "Genome-Wide Association Identifies Nine Common Variants Associated With Fasting Proinsulin Levels and Provides New Insights Into the Pathophysiology of Type 2 Diabetes." Diabetes **60**(10): 2624-2634.

Tawakol, A., R. Q. Migrino, G. G. Bashian, S. Bedri, D. Vermylen, R. C. Cury, D. Yates, G. M. LaMuraglia, K. Furie, S. Houser, H. Gewirtz, J. E. Muller, T. J. Brady and A. J. Fischman (2006). "In vivo 18F-fluorodeoxyglucose positron emission tomography imaging provides a noninvasive measure of carotid plaque inflammation in patients." J Am Coll Cardiol **48**(9): 1818-1824.

The Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.

Tibshirani, R., T. Hastie, B. Narasimhan and G. Chu (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proc Natl Acad Sci U S A **99**(10): 6567-6572.

Tran, P. K., H. E. Agardh, K. Tran-Lundmark, J. Ekstrand, J. Roy, B. Henderson, A. Gabrielsen, G. K. Hansson, J. Swedenborg, G. Paulsson-Berne and U. Hedin (2007). "Reduced perlecan expression and accumulation in human carotid atherosclerotic lesions." Atherosclerosis **190**(2): 264-270.

Ueland, T., K. Otterdal, T. Lekva, B. Halvorsen, A. Gabrielsen, W. J. Sandberg, G. Paulsson-Berne, T. M. Pedersen, L. Folkersen, L. Gullestad, E. Oie, G. K. Hansson and P. Aukrust (2009). "Dickkopf-1 enhances inflammatory interaction between platelets and endothelial cells and shows increased expression in atherosclerosis." Arterioscler Thromb Vasc Biol **29**(8): 1228-1234.

Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge (2008). "Alternative isoform regulation in human tissue transcriptomes." Nature **456**(7221): 470-476.

Wang, J., A. Razuvaev, L. Folkersen, E. Hedin, J. Roy, K. Brismar and U. Hedin (2011). "The expression of IGFs and IGF binding proteins in human carotid atherosclerosis, and the possible role of IGF binding protein-1 in the regulation of smooth muscle cell proliferation." Atherosclerosis **in press**.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-562.

Visel, A., Y. Zhu, D. May, V. Afzal, E. Gong, C. Attanasio, M. J. Blow, J. C. Cohen, E. M. Rubin and L. A. Pennacchio (2010). "Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice." Nature **464**(7287): 409-412.

Wooten, E. C., L. K. Lyer, M. C. Montefusco, A. K. Hedgepeth, D. D. Payne, N. K. Kapur, D. E. Housman, M. E. Mendelsohn and G. S. Huggins (2010). "Application of gene network analysis techniques identifies AXIN1/PDIA2 and endoglin haplotypes associated with bicuspid aortic valve." PLoS ONE **5**(1): e8830.

Yap, K. L., S. Li, A. M. Munoz-Cabello, S. Raguz, L. Zeng, S. Mujtaba, J. Gil, M. J. Walsh and M. M. Zhou (2010). "Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a." Mol Cell **38**(5): 662-674.

Yasuda, H., S. Nakatani, M. Stugaard, Y. Tsujita-Kuroda, K. Bando, J. Kobayashi, M. Yamagishi, M. Kitakaze, S. Kitamura and K. Miyatake (2003). "Failure to prevent progressive dilation of ascending aorta by aortic valve replacement in patients with bicuspid aortic valve: comparison with tricuspid aortic valve." Circulation **108**: 291-294.

Zhang, W., S. Duan, E. O. Kistner, W. K. Bleibel, R. S. Huang, T. A. Clark, T. X. Chen, A. C. Schweitzer, J. E. Blume, N. J. Cox and M. E. Dolan (2008). "Evaluation of genetic variation contributing to differences in gene expression between populations." Am J Hum Genet **82**(3): 631-640.