From the Department of Medicine, Solna
Karolinska Institutet, Stockholm, Sweden

FUNCTIONAL ANALYSIS OF GENETIC RISK MARKERS

Jesper Robert Gådin

Karolinska Institutet

Stockholm 2017

Cover Illustration: “A Long Time Ago in a Chromosome Far, Far Away…. ”, Andreas Gådin

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by E-PRINT AB

ISBN: 978-91-7676-870-9

Functional Analysis of Genetic Risk Markers

THESIS FOR DOCTORAL DEGREE (Ph.D.)

Publicly defended at Karolinska Institutet

CMM Lecture Hall (L8:00), Karolinska University Hospital, Solna

Friday, November 24th, 2017, 9:00 AM

By

Jesper Robert Gådin

Principal Supervisor:
Dr. Lasse Folkersen
Karolinska Institutet
Department of Medicine, Solna

Technical University of Denmark
Department of Bio and Health Informatics

Co-supervisors:
Prof. Per Eriksson
Karolinska Institutet
Department of Medicine, Solna

Dr. Ferdinand Van'T Hooft
Karolinska Institutet
Department of Medicine, Solna

Opponent:
Dr. Tuuli Lappalainen
Columbia University
Department of Systems Biology

Examination Board:
Dr. Carsten Daub
Karolinska Institutet
Department of Biosciences and Nutrition

Prof. Richard Sandberg
Karolinska Institutet
Department of Cell and Molecular Biology

Dr. Lars Feuk
Uppsala University
Department of Immunology, Genetics and Pathology

To my dear white ballerina shoes

2008-2015

ABSTRACT

Regulatory variants are the main factors responsible for genetic predisposition to how e.g. humans react differently to the environment. Therefore, it is important to locate and measure their effects, which can result in pre-disease intervention, new drugs, or as part in the personal medicine era, where selection and dose of a drug is based on a person's genetic profile.

In this thesis we have investigated the potential to link genetic markers to transcription using allele specific expression (ASE), which can avoid influence of both population stratification bias and trans-factors, increasing the statistical power compared to using total RNA based linkage methods. To quantify expression levels, we have used RNA-sequencing, which automatically makes it possible to measure ASE, provided that there is a heterozygous variant within the transcribed fragment, which in turn makes it possible to discern the expression between the two alleles. RNA sequencing data tend to be complex and requires to be summarized into count measures before further analyzed for ASE. To facilitate this process and provide additional analytical support, we developed the software AllelicImbalance, which now is freely accessible within bioconductor, a bioinformatics repository for code and data. Using this software we investigated ASE behavior on the individual level of a single transcribed variant, within a gene, and for connections between an ASE event and known risk markers, previously established from Genome Wide Association Studies (GWAS).

We showed in a dataset of 10 individuals that by measuring a consistent ASE over consecutive exons withing the same gene that an ASE signature is robust against dissimilarities in sequence. Further, because we showed that ASE stability covered several SNPs we established that short read sequencing is not a fundamental obstacle to the implementation of this technique. However, more individuals were needed to better assess a link to genetic variants. We continued our analysis in a larger dataset, in which one of the sequenced tissues had a representation of 680 individuals. This was enough to measure ASE as a regression of allelic fraction by genotype (aeQTL), conceptually similar to the regression of expression by genotype commonly used in eQTL studies. In this data we were able to explain novel risk SNPs using the aeQTL method, and showed that any bias for the reference allele had no significant effect on the regression. We moved on to test if aeQTL could pick up unique signals for 205 individuals in a tissue previously investigated for eQTL using a large cohort of more than 5000 individuals. Indeed, we detected 15 novel aeQTLs, which probably were masked by trans-regulation in the previous investigation. In addition, we describe the software ClusterSignificance, which tests for separation of groups in data with reduced dimensionality. The algorithm sets statistical rigor to a task previously done by visual inspection.

This thesis gives an overview of progress of us and others in ASE investigations, which is becoming more than being just a compliment to eQTL. The future signals a more dominant role as more sequencing data becomes readily available, accessing the closest active link to cis-regulation.

POPULÄRVETENSKAPLIG SAMMANFATTNING

Denna avhandling handlar om hur man kan använda modern teknik till att mer precist kunna bestämma kopplingen mellan kända mutations-varianter och gener. Från en individs vävnadsprov kan man med hjälp av avläsningsmaskiner digitalt bestämma exakt mängd av genavskrifter från alla aktiva gener, och dessutom avgöra om genavskrifterna härstammar från mammas eller pappas kromosomala kopia.

Människor har olika genetiska förutsättningar för hur kroppen reagerar på t.ex., rökning, medicin eller andra miljöfaktorer. Det betyder att om alla livsomständigheter skulle vara lika för två personer så är risken att drabbas av sjukdom individuellt betingat på genetik. Även om det kan tyckas vara orättvist att vissa människor från födseln har ökad risk för någon specifik sjukdom, så är skillnader i arvsmassa det som gör det möjligt för genetiker att kartlägga hur samspelet mellan arvsmassa och de funktionella enheterna i celler formar dessa förutsättningar. Mer specifikt är det en kartläggning av vilka vanliga mutationer, även kallat varianter, som är kopplade till sjukdom och hur dessa påverkar regleringen av celler. Denna information kan sedan användas till att dels karaktärisera en sjukdom bättre, men också resultera i mer direkta åtgärder såsom rekommendation av dosering och typ av medicin som är bäst lämpad för en person. Detta kallas på engelska populärt för “personalized medicine”.

Vi har i detta arbete utnyttjat det finurliga med att varje människa har två relativt oberoende aktiva uppsättningar av varje kromosompar, de som härstammar från mamma och pappa. I de fall en person har två olika varianter, kan man mäta deras relativa inflytande på produktionen av genavskrifter. Detta medför att man i princip kan använda endast en person för att finna relationen mellan varianter och genavskrift, men också att varje person blir sin egen statistiska kontroll eftersom miljöfaktorer påverkar personens båda kromosomer lika. Det låter bra i teorin, men kopplingsprocessen är fortfarande kantad av en del tekniska omständigheter, t.ex. att detektionsutrustningen kan behandla genavskrifter olika, om de inte ser likadana ut. I vissa fall finns även faktorer i cellen som påverkar genavskrifterna efter att de har producerats. Men vi har visat att om ett flertal individer studeras samtidigt, så kan man balansera ut deras negativa påverkan, genom en speciell fördelning i olika testgrupper. Vi utvecklade också en metod att statistiskt avgöra ifall två eller fler tidigare kända grupper går att separera med hjälp av mindre delar av den totala datamängden. Syftet med metoden är att undersöka om någon del är tillräcklig för att ensam signalera en skillnad, vilket för avskriftsdata från gener innebär att man kan upptäcka grupper av gener som ensamma signalerar sjuk eller frisk. Skillnaden mot traditionella metoder är att den här metoden inte är bunden till några antaganden kring datapunkternas fördelning.

I arbetet med denna avhandling har vi utvecklat programvara för hantering och analys, vilken kan laddas ner och användas fritt. Vi har visat ett flertal nya kopplingar mellan varianter och gener för hjärtsjukdomar, schizofreni och reumatism. För forskare är kopplingarna användbara för att få en djupare förståelse för sjukdomarnas uppkomst, eller som inkörsport till en ny era av personalized medicine.

LIST OF SCIENTIFIC PAPERS

I. Gådin JR, van't Hooft FM, Eriksson P, Folkersen L, **AllelicImbalance: an R/bioconductor package for detecting, managing, and visualizing allele expression imbalance data from RNA sequencing,** *BMC Bioinformatics* 2015, 12;16:194.

II. Gådin JR, Buil A, Colantuoni C, Nielsen J, Jaffe AE, Shin JH, Hyde TM, Kleinman JE, The BrainSeq Consortium, Plath N, Eriksson P, Brunak S, Didriksen M, Weinberger DR, Folkersen L, **Allelic Imbalance Method Discovers Novel Genes Linked To Schizophrenia**, Manuscript.

III. Gådin JR, Eriksson P, Berg L, Padyukov L, Folkersen L, **Using aeQTL to Characterize Trans-Effects in Peripheral Mononuclear Blood Cells of Rheumatoid Arthritis Patients**, Manuscript.

IV. Serviss JT†, Gådin JR†, Eriksson P, Folkersen L and Grandér D, **ClusterSignificance: a bioconductor package facilitating statistical analysis of class cluster separations in dimensionality reduced data**, *Bioinformatics* 2017, 1;33(19):3126-3128.

† Authors contributed equally

PUBLICATIONS NOT INCLUDED IN THE THESIS

Sennblad B, Basu S, Mazur J, Suchon P, Martinez-Perez A, van Hylckama Vlieg A, Truong V, Li Y, **Gådin JR**, Tang W, Grossman V, de Haan HG, Handin N, Silveira A, Souto JC, Franco-Cereceda A, Morange PE, Gagnon F, Soria JM, Eriksson P, Hamsten A, Maegdefessel L, Rosendaal FR, Wild P, Folsom AR, Trégouët DA, Sabater-Lleal M, **Genome-wide association study with additional genetic and post-transcriptional analyses reveals novel regulators of plasma factor XI levels**, *Hum Mol Genet* 2017, 26(3):637-649.

Drevinge C, Dalen KT, Mannila MN, Täng MS, Ståhlman M, Klevstig M, Lundqvist A, Mardani I, Haugen F, Fogelstrand P, Adiels M, Asin-Cayuela J, Ekestam C, **Gådin JR**, Lee YK, Nebb H, Svedlund S, Johansson BR, Hultén LM, Romeo S, Redfors B, Omerovic E, Levin M, Gan LM, Eriksson P, Andersson L, Ehrenborg E, Kimmel AR, Borén J, Levin MC, **Perilipin 5 is protective in the ischemic heart**, *Int J Cardiol* 2016, 15;219:446-454

McLeod O, Silveira A, Valdes-Marquez E, Björkbacka H, Almgren P, Gertow K, **Gådin JR**, Bäcklund A, Sennblad B, Baldassarre D, Veglia F, Humphries SE, Tremoli E, de Faire U, Nilsson J, Melander O, Hopewell JC, Clarke R, Björck HM, Hamsten A, Öhrvik J, Strawbridge RJ; IMPROVE Study Group, **Genetic loci on chromosome 5 are associated with circulating levels of interleukin-5 and eosinophil count in a European population with high risk for cardiovascular disease**, *Cytokine* 2016, 81:1-9.

Paloschi V, **Gådin JR**, Khan S, Björck HM, Du L, Maleki S, Roy J, Lindeman JH, Mohamed SA, Tsuda T, Franco-Cereceda A, Eriksson P, **Aneurysm development in patients with a bicuspid aortic valve is not associated with transforming growth factor-β activation**, *Arterioscler Thromb Vasc Biol* 2015, 35(4):973-980

Locke AE†, Kahali B†, Berndt SI†, Justice AE†, Pers TH†, Day FR, Powell C, Vedantam S, Buchkovich ML, Yang J, Croteau-Chonka DC, Esko T, Fall T, Ferreira T, Gustafsson S, Kutalik Z, Luan J, Mägi R, Randall JC, Winkler TW, Wood AR, Workalemahu T, Faul JD, Smith JA, Zhao JH, Zhao W, Chen J, Fehrmann R, Hedman ÅK, Karjalainen J, Schmidt EM, Absher D, Amin N, Anderson D, Beekman M, Bolton JL, Bragg-Gresham JL, Buyske S, Demirkan A, Deng G, Ehret GB, Feenstra B, Feitosa MF, Fischer K, Goel A, Gong J, Jackson AU, Kanoni S, Kleber ME, Kristiansson K, Lim U, Lotay V, Mangino M, Leach IM, Medina-Gomez C, Medland SE, Nalls MA, Palmer CD, Pasko D, Pechlivanis S, Peters MJ, Prokopenko I, Shungin D, Stančáková A, Strawbridge RJ, Sung YJ, Tanaka T, Teumer A, Trompet S, van der Laan SW, van Setten J, Van Vliet-Ostaptchouk JV, Wang Z, Yengo L, Zhang W, Isaacs A, Albrecht E, Ärnlöv J, Arscott GM, Attwood AP, Bandinelli S, Barrett A, Bas IN, Bellis C, Bennett AJ, Berne C, Blagieva R, Blüher M, Böhringer S, Bonnycastle LL, Böttcher Y, Boyd HA, Bruinenberg M, Caspersen IH, Chen YI, Clarke R, Daw EW, de Craen AJM, Delgado G, Dimitriou M, Doney ASF, Eklund N, Estrada K, Eury E, Folkersen L, Fraser RM, Garcia ME, Geller F, Giedraitis V, Gigante B, Go AS, Golay A, Goodall AH, Gordon SD, Gorski M, Grabe HJ, Grallert H, Grammer TB, Gräßler J, Grönberg H, Groves

CJ, Gusto G, Haessler J, Hall P, Haller T, Hallmans G, Hartman CA, Hassinen M, Hayward C, Heard-Costa NL, Helmer Q, Hengstenberg C, Holmen O, Hottenga JJ, James AL, Jeff JM, Johansson Å, Jolley J, Juliusdottir T, Kinnunen L, Koenig W, Koskenvuo M, Kratzer W, Laitinen J, Lamina C, Leander K, Lee NR, Lichtner P, Lind L, Lindström J, Lo KS, Lobbens S, Lorbeer R, Lu Y, Mach F, Magnusson PKE, Mahajan A, McArdle WL, McLachlan S, Menni C, Merger S, Mihailov E, Milani L, Moayyeri A, Monda KL, Morken MA, Mulas A, Müller G, Müller-Nurasyid M, Musk AW, Nagaraja R, Nöthen MM, Nolte IM, Pilz S, Rayner NW, Renstrom F, Rettig R, Ried JS, Ripke S, Robertson NR, Rose LM, Sanna S, Scharnagl H, Scholtens S, Schumacher FR, Scott WR, Seufferlein T, Shi J, Smith AV, Smolonska J, Stanton AV, Steinthorsdottir V, Stirrups K, Stringham HM, Sundström J, Swertz MA, Swift AJ, Syvänen AC, Tan ST, Tayo BO, Thorand B, Thorleifsson G, Tyrer JP, Uh HW, Vandenput L, Verhulst FC, Vermeulen SH, Verweij N, Vonk JM, Waite LL, Warren HR, Waterworth D, Weedon MN, Wilkens LR, Willenborg C, Wilsgaard T, Wojczynski MK, Wong A, Wright AF, Zhang Q; LifeLines Cohort Study, Brennan EP, Choi M, Dastani Z, Drong AW, Eriksson P, Franco-Cereceda A, **Gådin JR**, Gharavi AG, Goddard ME, Handsaker RE, Huang J, Karpe F, Kathiresan S, Keildson S, Kiryluk K, Kubo M, Lee JY, Liang L, Lifton RP, Ma B, McCarroll SA, McKnight AJ, Min JL, Moffatt MF, Montgomery GW, Murabito JM, Nicholson G, Nyholt DR, Okada Y, Perry JRB, Dorajoo R, Reinmaa E, Salem RM, Sandholm N, Scott RA, Stolk L, Takahashi A, Tanaka T, van 't Hooft FM, Vinkhuyzen AAE, Westra HJ, Zheng W, Zondervan KT; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; International Endogene Consortium, Heath AC, Arveiler D, Bakker SJL, Beilby J, Bergman RN, Blangero J, Bovet P, Campbell H, Caulfield MJ, Cesana G, Chakravarti A, Chasman DI, Chines PS, Collins FS, Crawford DC, Cupples LA, Cusi D, Danesh J, de Faire U, den Ruijter HM, Dominiczak AF, Erbel R, Erdmann J, Eriksson JG, Farrall M, Felix SB, Ferrannini E, Ferrières J, Ford I, Forouhi NG, Forrester T, Franco OH, Gansevoort RT, Gejman PV, Gieger C, Gottesman O, Gudnason V, Gyllensten U, Hall AS, Harris TB, Hattersley AT, Hicks AA, Hindorff LA, Hingorani AD, Hofman A, Homuth G, Hovingh GK, Humphries SE, Hunt SC, Hyppönen E, Illig T, Jacobs KB, Jarvelin MR, Jöckel KH, Johansen B, Jousilahti P, Jukema JW, Jula AM, Kaprio J, Kastelein JJP, Keinanen-Kiukaanniemi SM, Kiemeney LA, Knekt P, Kooner JS, Kooperberg C, Kovacs P, Kraja AT, Kumari M, Kuusisto J, Lakka TA, Langenberg C, Marchand LL, Lehtimäki T, Lyssenko V, Männistö S, Marette A, Matise TC, McKenzie CA, McKnight B, Moll FL, Morris AD, Morris AP, Murray JC, Nelis M, Ohlsson C, Oldehinkel AJ, Ong KK, Madden PAF, Pasterkamp G, Peden JF, Peters A, Postma DS, Pramstaller PP, Price JF, Qi L, Raitakari OT, Rankinen T, Rao DC, Rice TK, Ridker PM, Rioux JD, Ritchie MD, Rudan I, Salomaa V, Samani NJ, Saramies J, Sarzynski MA, Schunkert H, Schwarz PEH, Sever P, Shuldiner AR, Sinisalo J, Stolk RP, Strauch K, Tönjes A, Trégouët DA, Tremblay A, Tremoli E, Virtamo J, Vohl MC, Völker U, Waeber G, Willemsen G, Witteman JC, Zillikens MC, Adair LS, Amouyel P, Asselbergs FW, Assimes TL, Bochud M, Boehm BO, Boerwinkle E, Bornstein SR, Bottinger EP, Bouchard C, Cauchi S, Chambers JC,

Chanock SJ, Cooper RS, de Bakker PIW, Dedoussis G, Ferrucci L, Franks PW, Froguel P, Groop LC, Haiman CA, Hamsten A, Hui J, Hunter DJ, Hveem K, Kaplan RC, Kivimaki M, Kuh D, Laakso M, Liu Y, Martin NG, März W, Melbye M, Metspalu A, Moebus S, Munroe PB, Njølstad I, Oostra BA, Palmer CNA, Pedersen NL, Perola M, Pérusse L, Peters U, Power C, Quertermous T, Rauramaa R, Rivadeneira F, Saaristo TE, Saleheen D, Sattar N, Schadt EE, Schlessinger D, Slagboom PE, Snieder H, Spector TD, Thorsteinsdottir U, Stumvoll M, Tuomilehto J, Uitterlinden AG, Uusitupa M, van der Harst P, Walker M, Wallaschofski H, Wareham NJ, Watkins H, Weir DR, Wichmann HE, Wilson JF, Zanen P, Borecki IB, Deloukas P, Fox CS, Heid IM, O'Connell JR, Strachan DP, Stefansson K, van Duijn CM, Abecasis GR, Franke L, Frayling TM, McCarthy MI, Visscher PM, Scherag A, Willer CJ, Boehnke M, Mohlke KL, Lindgren CM, Beckmann JS, Barroso I, North KE, Ingelsson E, Hirschhorn JN, Loos RJF, Speliotes EK, **Genetic studies of body mass index yield new insights for obesity biology**, *Nature* 2015, 518(7538):197-206

CONTENTS

LIST OF ABBREVIATIONS

ASE	Allele Specific Expression
GWAS	Genome Wide Associations Study
eQTL	Expression Quantitative Trait Loci
aeQTL	Allele Specific Expression Quantitative Trait Loci
SNP	Single Nucleotide Polymorphism
txSNP	Transcribed SNP
PCA	Principal Component Analysis

LIST OF SPECIFIC TERMS

In cis	Description of an action directly from and on the same chromosome.
In trans	Description of an action from a chromosome, and then through an intermediate RNA or protein, which acts on the same or another chromosome.
Variant	A more general definition of a genetic variant than SNP.

1 INTRODUCTION

This section starts with a brief historical connection to the discovery of diploid inheritance, evolutionary propulsion, and how natural selection is hijacked by medical applications in order to reduce human suffering. Then follows an introduction to eQTL and aeQTL studies, the global ASE landscape, some observations of ASE in twin studies, a brief overview of ASE-software, and an assumption free method to test class separations.

1.1 The Purpose of Genetics in Evolution and Human Intervention

Since Gregor Mendel presented examples of how a carriers information for a trait always comes in a pair of distinct units[1], the field has taken a relatively huge jump to present-day methods where it is technologically and economically feasible to measure millions of markers in thousands of individuals for these distinct units now called alleles. The advancement has made it possible to address complex traits and diseases from multiple alleles and how they interact.

The driving force behind Mendel's experiments was to gain a deeper insight into breeding of sheep, as it was the economical backbone for the region he lived in. Mendel used his education in plant science and insight in pollination to conduct the cross experiments that would give rise to the important insights about segregation and individual assortment of gametes[2]. Humans are like the pea-plants of Mendel, a diploid organism, and we transfer our traits between generations in a similar fashion. Besides the benefits of recombination and segregation[3], another benefit of having a diploid is its increased power to locate and characterize the impact of genetic markers on gene expression. In this thesis we will present a method using RNA-sequencing and genotype arrays to establish these links. The link is important to understand disease etiology, but to set this in the overarching context of disease prevention, let us first consider harmless variation in the genome, its visual presence and the evolutionary reason for its existence.

When all nucleotides in the genome are the same, as for monozygotic twins their phenotypical appearance will be similar. Apart from environmental variation, the reason people look different is because of nucleotides varying at certain areas in the genome, regulating gene expression responsible for phenotypic traits. This indicates that genetics is a major contributor in shaping the development of the human being. Today, the dbSNP database[4] contains more than 135 million variants discovered throughout the genome. Luckily, the majority of genetic variants don't confer any remarkable or dangerous differences. Instead, the purpose of their presence is to act as an evolutionary catalyst in adaptations to changes in the environment. However, variants can have a serious influence on complications that come late in life, as the level of fitness steadily loses importance by the presence of offspring. Speculatively, in an evolutionary perspective, early death could actually act have an importance in the balance of available resources to the generation of offspring for each generation? Nevertheless, in the context of the modern society, a favorable evolutionary step would be to treat diseases caused by the environment, instead of a large

pool of variation. As a long-term goal, a world free from disease would reduce the economical and emotional burden of the suffering involved with disease and death. To reach there, it is important to map which genetic variants that contribute to changes in expression, and from that knowledge we can take measures in how we want to intervene, e.g., using the temporary effects of drugs, or through gene therapy.

1.2 The Role of Allelic Imbalance in Functional Analysis of Genetic Risk Markers

Several names, e.g. Allelic Imbalance (AI)[5], Allele Specific Expression (ASE)[6], Differential Allele Expression (DAE)[7], have been suggested to explain the phenomena of unequal expression of a gene caused by diploid chromosomal heterogeneity. We started out using AI, but as the term competes with the popular concept of Artificial Intelligence (also AI), I have in this thesis chosen to use ASE. The underlying assumption in using ASE for genetic mapping, is founded on the idea that if the region around a gene is similar on both the maternal and paternal chromosome for a gene, then all factors responsible for transcription should act equally on both alleles, resulting in an equal level of transcriptional expression.

Mapping of a genetic variants influence on the transcriptional process becomes possible first when we have a collection of individuals with at least two different genotype groups, e.g. one group containing the heterozygous case AB, and another group containing the homozygous case AA. The process of mapping the influence of genotypes on total RNA is called expression quantitative trait loci (eQTL) analysis, where significant association can be attributed as eQTLs. As more high-throughput expression data are made accessible from microarray or RNA-sequencing, eQTL studies are being increasingly common[8,9], and are suitable either as independent studies or complimentary to other investigations on expression.

In RNA sequencing experiments, ASE information can without further experiments directly be used as a compliment to eQTL analyses adding power in that it e.g. avoids population stratification bias, as each individual is its own control. In theory, the presence of ASE is itself a proof of existing regulatory control over the transcription. However, for a better characterization of the genetic impact, it is of interest to detect which variants are responsible for ASE, which can be done in a similar fashion to eQTL studies, with the pre-requisite that the phase has to be determined, i.e., we need to know which regulatory allele is on the same phase as which transcribed allele. Such an analysis constitutes the backbone of this thesis, and because of its similarities to eQTL, we chose to denote the method as allele specific quantitative trait loci (aeQTL).

1.3 ASE is Measured if Heterozygous.

With existing methods of microarrays and RNA-sequencing it is only possible to differentiate ASE if it contains a marker, in this case a heterozygous variant. Even though the genome is rich in variation, exons from which much of the transcriptional content origins from, have a sparser set, and even for exons harboring common variation, it is not certain the variant is a

heterozygote for a specific individual. For this reason, eQTL will still be the only way to assess associations for transcripts of less heterozygosity.

1.4 ASE is Necessary for Functional Analysis of Genetic Risk Markers.

A reduction in population stratification is not the only benefit using aeQTL, but the internal case-control condition between alleles also applies to transcriptional factors, which are regulated by other parts of the genome in adaptation to a cells spatial experience, i.e., any transcriptional factor will have the same probability of initiating transcription of a gene, if the genetic region it binds to is exactly the same. The large benefit is that total levels of RNA, which can fluctuate during the day won't have an impact on ASE, which only uses the fraction of expression, which is relative and stays the same. However, in an eQTL analysis the total RNA levels will have an impact, and force large sample sizes to detect a signal through all inter-individual variation. This opens up the opportunity for aeQTL to detect cis QTLs, using smaller sample sizes, provided that they have the required heterozygosity.

1.5 Knowledge of ASE from Twin Studies

As a proof of concept for the ASE phenomenon and to investigate its heredity, Cheung et al.[10], conducted a study where they investigated the co-expression pattern of ASE in monozygotic twins (MZ) using genotype arrays. After filtering SNPs to have at least five heterozygote twin pairs, they were left with 211 SNPs. For 63 SNPs there were a significant similarity in ASE expression. Their results suggest that MZ not only show coherent ASE signals, but as well share levels of expression. Additionally, the high intra correlations for ASE fractions suggest a strong heredity of the levels of imbalance.

In a more recent study, Buil et al.[11], sequenced fat, skin, blood and lymphoblastoid cell lines(LCL) for ~400 female twin pairs, to investigate the proportions of genetic and environmental effects on ASE. They found that environmental effects do have an effect on ASE (11% LCL, 16% skin, 21% fat and 35% blood), and affects the magnitude of ASE. However they clearly state that, initially the difference in transcription of the alleles must come from a difference in sequence. About 80% of the genes with a reported ASE also had a significant eQTL, i.e., the identification of an ASE event is likely to also be an eQTL.

1.6 Correcting for Mapbias Underlies All Further Analysis

The post-processing step, which can bias reference mapping has been seen to affect several studies and different methods have been developed to adjust for it: Measuring the bias by producing artificial reads of equal quantity[12]. Allow more mismatches in the alignment step, but increasing the risk of false mappings[13]. Map the reads to personalized phased genomes, or transcriptomes[14,15]). But the most straightforward and accessible way to handle mapping bias, here denoted as mapbias, is to mask all SNPs within the reference genome with the generic nucleotide indicator N. However, this will make it harder to map, and the loss of reads can be huge if the amount of SNPs in a region is large. The remaining source of mapbias after mapping to a masked genome would be in the regions with presence of a novel SNP, and

solely not masked[16]. Compensating mapbias using the map ratio for sequenced gDNA has been popular, and should ideally give 1:1 ratio in absence of technical variation[17,18,19,20], but has been shown less effective compared to masking SNPs[21]. Another option to reduce bias is to produce longer reads[22], as longer reads in general increase the confidence in mapping.

It has also been a concern that the problems with mapbias seen for ASE and aeQTL could be equally problematic for eQTLs, potentially disqualifying previous discoveries. This concern is correct in theory, but has been shown to have limited effect in practice[23]. Additionally, eQTL results have normally been validated in a separate experiment, increasing the reliability of the findings. When eQTLs are performed in RNA sequencing data, extending the analysis to detect aeQTLs is an additional cost-effective validation.

1.7 Investigations of ASE in Tissues or Other Sub-compartments

A global analysis is important to describe the ASE in certain regions, among tissues or to quantify the amount of technical ASE bias among all or a subset of SNPs. For the biological and technical understanding it can be of interest to study the variation of ASE in different regions e.g. like start, middle and end of transcripts or for different types of transcripts e.g. like lncRNA, miRNA or protein coding genes. Mapbias on a global level can be measured by averaging the allele fraction from all heterozygotes with the same allele as the reference genome. The more deviation of the average fraction from 1:1 ratio, the more mapbias is present.

ASE quantification of only one SNP in one individual can unfortunately not by its own say anything of the presence of bias, but subsetting SNPs into smaller regions, i.e. investigating all SNPs located in a gene or exon of interest can reveal regional stability of ASE measurements[24].

Comparing ASE globally between tissues have in a study by Kukurba et al[25] shown that ASE is co-expressed in tissues more often for tissues with the same embryonic origin. Comparing ASE between tissues are robust in the sense of avoiding bias from regional differences in the DNA, which is the concern but also necessity for aeQTL, i.e., with no genotypic difference we cannot distinguish the maternal and paternal allele.

1.8 An R, and Bioconductor Based Software for Allele Imbalance

Numerous softwares and pipelines have recently been developed to explore and detect global ASE for different experimental conditions. The AlleleSeq[26] software constructs paternal and maternal reference genome based on variant phasing from family-trios. MMSEQ[27] uses the polyHap software to phase according to the hapmap project. The software Allim[28] for F1 crosses handles the phasing in the alignment step with the read aligning software GSNAP[29]. Other pipelines or softwares are e.g. Allele Workbench[30], WASP[31], MBASED[32], EMASE[33], ALEA[34], ASARP[35], QuASAR[36], GeneiASE[37], MAMBA[38] and RASQUAL[39]. Two softwares for ASE analysi in the R and Bioconductor environment are iASeq[40] and AllelicImbalance[24].

1.9 Assumption Free Method to Test Separation of Clusters with Known Label

It is common that gene expression data is analyzed using dimensionality-reducing methods that reduce a high-dimensional space to something more comprehensible, which makes it possible to detect the formation of clusters, which in turn would suggest relatedness of the points in each cluster. Often these are assessed by visual inspection, lacking statistical objectivity, but can be ok if a separation is clear, i.e., there is no overlap. In the situations when the formations of clusters are not clear, because of a substantial overlap between classes, more sophisticated ways are required to assess if there is a true separation of clusters. The basic idea of our algorithm is to use a computer approximation of how humans comprehend patterns, by using principle curves[41], which makes it possible to further reduce the data down to one dimension. For the many dimensionality reduction methods available a general test requires to assume no particular distribution, and the algorithm we have developed has taken that into consideration.

2 AIMS

The general aim has been to investigate the potential of ASE information within RNA sequencing data to find the link between regulatory genetic markers and gene expression.

PAPER I To develop a software to handle ASE data from RNA sequencing and investigate the potential of using ASE in the ASAP cohort; can independent ASE signals be used to reliably assess links to nearby regulatory variants?

PAPER II What is the potential of using ASE in a larger cohort, which allows for establishing an association in a regression over genotypes; can we detect associations for GWAS risk variants in schizophrenia that eQTL cannot?

PAPER III Can we discover similar results as in PAPER II in a smaller cohort of less sequencing depth; is it still possible to detect associations by ASE, which haven't previously been found by large eQTL efforts within the same tissue?

PAPER IV To develop a software to test separations in clusters resulting from dimensionality reduction methods.

3 METHODOLOGICAL CONSIDERATIONS

This section describes how to we assured enough heterozygosity and read depth in our analyses, followed by a description of the material and experimental techniques used in each study. Lastly, a presentation of the filters and statistics applied in the studies.

3.1 Examinations of Heterozygous Probability and Read Depth

In order to assess the amount of individuals required to increase the chances to make a statement about ASE in any selected gene, we needed to characterize the heterozygous landscape in humans. To calculate the probability for a gene to contain at least one heterozygote site, we used dbSNP135[42] and UCSC-known-genes-hg19 (2012-06-27)[43].

A two-sided binomial test for which one allele has 39 reads and the other has 61 reads is enough to detect significant ASE signals (p=0.0352), assuming no bias or sequencing errors. An experiment is however not free from bias or technical variation, and sadly not equally distributed among all genes, i.e., one gene might be affected more by e.g. bias or sequencing errors than another. Similar to other statistical tests experiments, the more samples included (in this case read coverage), allows us to detect smaller effect size differences and reduce the impact of spurious technical artifacts. Yet, mapbias continues to pose a problem for regions hard to align, and its impact is problematic for single txSNPs.

To decide on a cost-effective sequencing depth, we knew from previous microarray experiments, that only about 50% of the genes in a measured tissue had a meaningful level of intensity[44]. However, in RNA sequencing, we are not limited by cross-hybridization problems of low level expression, but each tissue can have a quite dramatic difference in expression, and more common genes or remains of eg. rRNA will consume the majority of nucleotides used in the sequencing, leaving genes with low expression a smaller pool of nucleotides to share. Early RNA sequencing experiments outlined the potential for RNA-sequencing using around 100 million paired-end 36-42mer reads to be enough for a gene specific depth coverage of at least 10 reads for about 50% of ~36 000 EnsEMBL genes[45]. We designed a pilot experiment of 100 million pair-end 100mer reads, to ensure an acceptable depth for the majority of genes to be investigated for ASE.

3.2 RNA Sequencing of Vascular Tissue from the ASAP Cohort

In PAPER I, we both described our developed software for managing ASE-data, and some applications on RNA sequenced Aorta and Liver tissue from the ASAP study[46]. In total 26 sample were sequenced at a depth of ~100 million pair-end reads each, 8 individuals for each tissue, plus 2 additional individuals for aorta, to accompany another ongoing experiment in the lab. The library preparation protocol used RiboZero and strand-specific sequencing[47], randomly distributed in the sequencing lanes to reduce batch effects. The mean fragment size generated was 100bp long, which had been showing good QC measures by the sequencing facility. However, to our disappointment, it many times resulted in the same regions being read twice, losing many of the advantages of paired-end sequencing.

The mean fragment size of 100 yielded many shorter fragments, which all of them contained adapter sequences, additionally reduced the reason for using a read length of 100 in the sequencing. All SNPs investigated in PAPER I were determined by DNA genotype arrays for 500 000 variants. However, for the development of the AllelicImbalance software, which took place previous to our prototype RNA sequencing experiment, we used public data from Montgomery et. al.[48]

3.3 RNA Sequenced Brain Tissue from Consortium

In PAPER II, RNA sequenced brain tissue from Dorso Lateral Prefrontal Cortex (DLPFC)(n=680) and Hippocampus (n=421) in 782 individuals, was used in collaboration with the Lieber Consortium. Paired-end read sequencing with two samples per lane with a mean depth of 129.8 million for each sample. The library preparation protocol contained a mixed setup of poly-A capture or RiboZero, no retained strand-specific information and a mean fragment size of ~300bp, a size exploiting the full use of pair-end sequencing. The use of poly-A capture achieves a maximization of read coverage for mRNA specific fragments, by not spending reads on non-coding RNAs, or degradation products. For genotyping, the Illumina Human1M-Duo v3.0, illumina 650K and omniX+ microarrays were used. In the purpose of establishing a phase estimate as link between risk SNPs and txSNPs, and to increase the amount of investigable variants, we used SHAPEIT2[49], and IMPUTE2[50], using the 1000 Genomes phase III haplotype panel[51].

3.4 RNA Sequenced PBMC from the Combine Study

In PAPER III, we applied our developed method in PAPER II in a mixed cohort of rheumatology patients (n=150) and healthy subjects (n=55), where 137 individuals have complete data for two visits, three months apart. For RNA-sequencing, the library was prepared by a polyA+ extraction of all fragments containing a poly-A tail. Two samples per lane were used producing an average of 11 million paired-end reads per sample. For genotyping, the Illumina OmniExpress-platform was used, while phasing and imputation was done in the same manner using SHAPEIT2[49] and IMPUTE2[50] as in PAPER II.

3.5 Public Microarray Data to Exemplify a ClusterSignificance Use Case

In PAPER IV, the data used to demonstrate the algorithm came from public microarray data of 2096 patient samples and representing 6 different hematological malignancies as well as non-leukemic and healthy patients. These were subset to only include lncRNA expression profiles, resulting in 4283 unique lncRNA genes, subsequently dimensionally reduce by t-SNE[52], which is characterized by a preservation of neighbourhoods better than a standard PCA.

3.6 Generation of an N-masked Reference Genome

In order to reduce mapbias, in Paper I, II and III we based our alignment procedures on a N-masked reference genome, inspired by Degner et al.[16], who had reported a decrease in mapbias on a global level, which after masking was within the boundary of the expected

binomial distribution. However, they also noted that several independent loci still showed an imbalance for simulated reads, caused by a reads ability to align equally or better at another loci in the genome[53]. In this regard, we only used unique alignments in PAPER I.

The N-masked reference genome was generated through functionality in the AllelicImbalance software, which is fed coordinates of SNP evidence (dbSNP138[54]). Two popular splice aware aligners main aligners were considered: TopHat[55], which had gain a lot of user-feedback and mileage due to early popularity and STAR[56], which at release outperformed other aligners by a factor >50 in run-time, as well as increasing mapping sensitivity and precision[56]. In paper I, we used TopHat as example to demonstrate its ability to mapbias, but due to the all-around better performance of STAR, it was the only aligner used for Paper I, II and III.

3.7 Gene Annotation and Filtering

To reduce the risk of technical artifacts showing up as ASE, for paper II and III, we set several filtering cutoffs: A threshold of at least 10 reads from each allele for a txSNP. And to hinder sequencing errors creating false allelic imbalance, we also required that at least 10% of the total expression came from the less expressed allele. To perform a linear regression at least 5 usable individuals in at least two groups were required. These ad-hoc thresholds were determined using trial and error on early data, and proved to prevent many systematic misinterpretations. Further, txSNP-regions were defined according to the UCSC RefSeq gene annotation database (version 12-11-2015[57] in PAPER II, and version 01-16-2017[58] in PAPER III) in a way that each txSNP-region corresponded to one gene. The only exception was overlapping genes, which were included in the same txSNP-region. Using the phase information we then summarized all txSNPs in the region to get a single value for each sample. To reduce the multiple testing burden, the association tests were constructed in pairs of genes and risk variants that were within a distance of 200kb in PAPER II and 500kb in PAPER III, a distance proven sufficient in previous studies[59].

3.8 Using a Binomial Test to Assess ASE

In PAPER I, we investigated the presence of individual ASE:s using the binomial test, which statistically is similar to any random sampling of two random entities from the same set of data. When testing for ASE, the two random entities are the two types of allele-fragments randomly sampled from within the same tissue. Like many other statistical methods, the more samples (here reads) available, the smaller differences can be given statistical evidence.

3.9 Using a Regression Analysis For eQTL and aeQTL

The regression method used in PAPER II and III, is especially appealing in its resemblance of the three-genotype group eQTL analysis, which facilitates a direct comparison of the result, see figure 1. The method was inspired by a similar method by Almlöf et al.[60], who instead of RNA sequencing used RNA in the form of cDNA on a DNA microarray

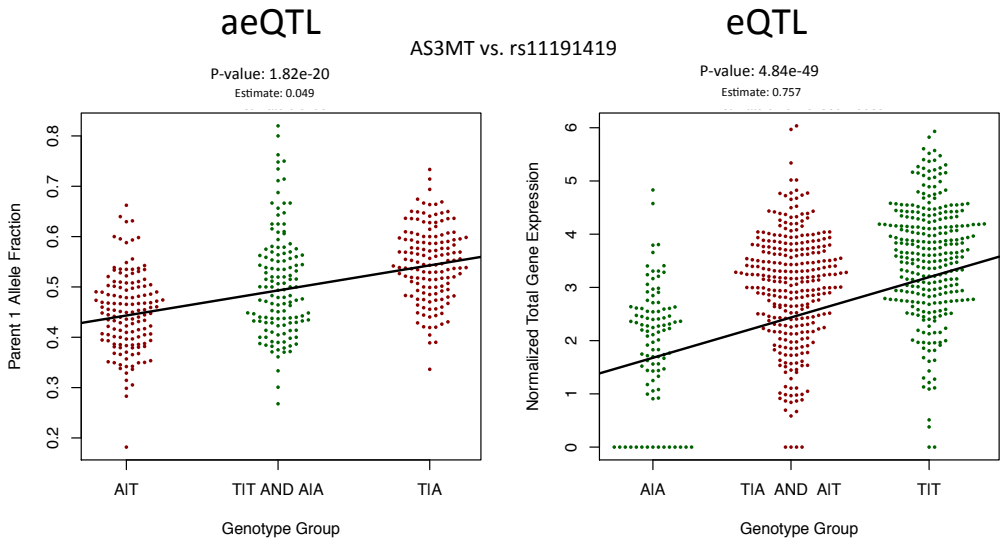
platform. As for allelic imbalance in general, we rely on the presence of heterozygote txSNPs to construct a fraction, and as all genotypes were computationally phased, the relevant allelic imbalance fractions were always calculated in the parent-1-allele direction. Taken together this prompts the $fraction_{P1}$ variable, defined as the amount of parent-1 reads divided by reads from both chromosome copies. Since the non-transcribed risk-SNP is also phased, the association between risk-SNP and the allelic imbalance it creates can be investigated using the linear regression model:

$Fraction_{P1} \sim$ genotype-group + covariates

Even if aeQTL is then theoretically set up to ignore trans-effects caused by confounders, to reduce the risk of their influence, we included several covariates, age, sex, RIN, the first three principal components (PCs) from a PCA on the genotype profile, and the ten first PCs from a PCA on normalized whole gene expression estimates. In order to compare the aeQTLs ability in detecting QTLs, we performed an eQTL analysis, which used the same covariates, but instead of txSNP expression values, we used normalized whole gene read counts. Additionally in PAPER III, we assessed eQTLs using total RNA from each txSNP respectively; with purpose of creating groups comparable enough to characterize the relative influence of trans-regulation, using the difference in effect size between aeQTL and eQTL. In PAPER III, we facilitated the use of mixed models to take the extra information of both visits into account:

fractionP1 ~ genotype-group + covariates | patient-ID

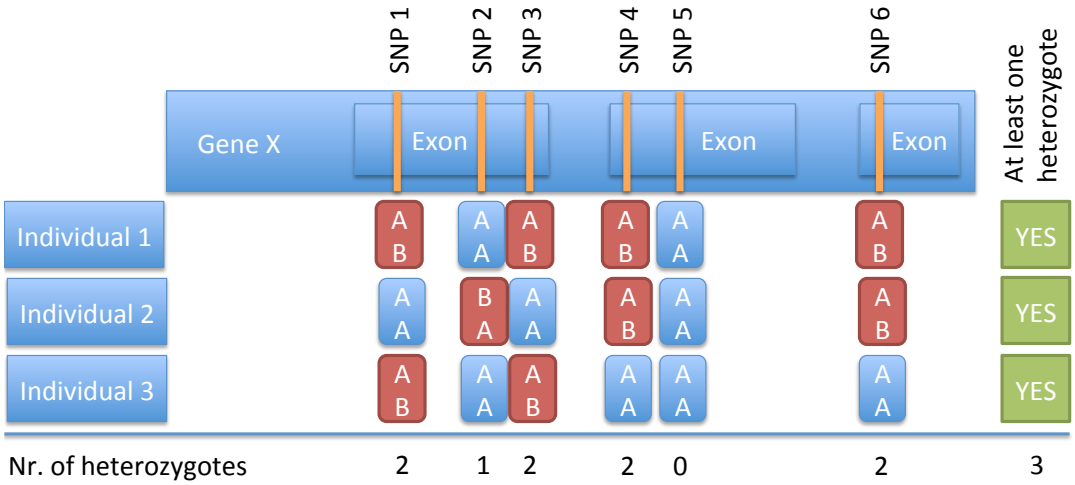
In the performed aeQTL the genotype-group value is enumerated as 0, 1 or 2; being either heterozygote of one phase, e.g. A|G , homozygote, e.g. A|A and G|G, or heterozygote in opposing phase, e.g. G|A. The symbol | indicates that the genotype is phased, and on the left side is the allele variant from parent-1.



Figur 1 A side-by-side comparison between the data point distribution in respect to the two models of eQTL and aeQTL. Heterozygote points indicate homozygote individuals and red points indicate heterozygote individuals. There are slightly fewer points in the aeQTL analysis, which can be explained by the methods dependence on a presence of heterozygote txSNPs. The established link between AS3MT and rs11191419 is one of the strongest associations in respect to p-value in PAPER II.

3.10 Combining Gene-wide ASE Measures to Increase Statistical Power

Multiple variants can be present within an exon, and for which of the variants an individual is heterozygote for varies. Summarizing the fraction measures over a gene-region thereby increases the amount of genes amendable for analysis by raising the total number of contributing individuals compared to specific txSNP measures, see example figure 1.



Figur 2 An example of how a united measure for a gene region can result in more individuals available for an aeATL test. Heterozygote SNPs are here illustrated as red rectangles, and in the presence of at least one heterozygote SNP for an individual, it is indicated as a green box. In this example all three individuals will be able to contribute to test for aeQTL on the gene level, but maximum two for individual txSNPs.

3.11 Test Cluster Separations in Data of Known Label

In its most basic form the separation test performed in ClusterSignificance uses a permutation on ranked data to avoid any assumptions on distribution or assessment of effects size. The aim was to set statistical rigor to a field in science where separations otherwise was assessed by pure visual inspection. To become a general test appropriate for all types of distributions produced by the plethora of dimension-reduced algorithms available[61], no assumptions of the underlying distribution were acceptable, and to mimic the interpretational assessment of human vision, we applied principal curves to make the final simplification down to one dimension, which moreover is ignorant of assumptions. The simplification to one-dimensional space also makes it feasible to produce a ROC curve from all possible linear discriminants, resulting in a score calculated using a formula provided by Song et al.[62], which e.g. makes it possible to compare groups of unequal size. To test against a certain null case, we advocate an initialization of the permutation so that it wraps the whole pipeline. In this way, any introduced artificial separations by the original dimensionality reduction algorithm are included in the null case. Another reason to perform randomization of the data before the initial dimensionality reduction can be in a setup where e.g. one wants to assess if a set of genes produce a better separation than picking a random set of genes.

4 RESULTS

Here the work is presented in chronological order to best describe how certain results or opportunities guided the projects within this thesis forward. It starts with an early explorative analysis leading to PAPER I, and thereafter a section describing how we planned to deal with the many uncertainties of ASE measures. Subsequently, how the opportunity using larger cohorts (in PAPER II and III) allowed us to try a more robust variant of ASE analysis, partly ignorant to some of the problems individual ASE measures are subjected to. Finally, a description of ClusterSignificance, which is a method and software to statistically test cluster separations.

4.1 Preliminary Results and Considerations

Before the time of any pilot experiments we wanted to calculate the total amount of individuals needed to get a reasonable chance to have at least one individual with a heterozygous SNP for any given gene. From the Hardy-Weinberg equilibrium equation, we know that frequency of heterozygotes is the frequency of allele q multiplied by allele f. The probability p for a gene to have at least one heterozygote SNP is then p=1-(1-(f * q))^n, where n equals number of individuals. Applying this equation on coding exonic frequencies demonstrated that more than 80 percent of all gene transcriptions larger than 2000bp would be available for investigation using 8 individuals, figure 3A.

Our RNA-sequenced samples matched well previous described gene-read-depth relationship with 16500 genes with 10 reads or less, figure 3B.

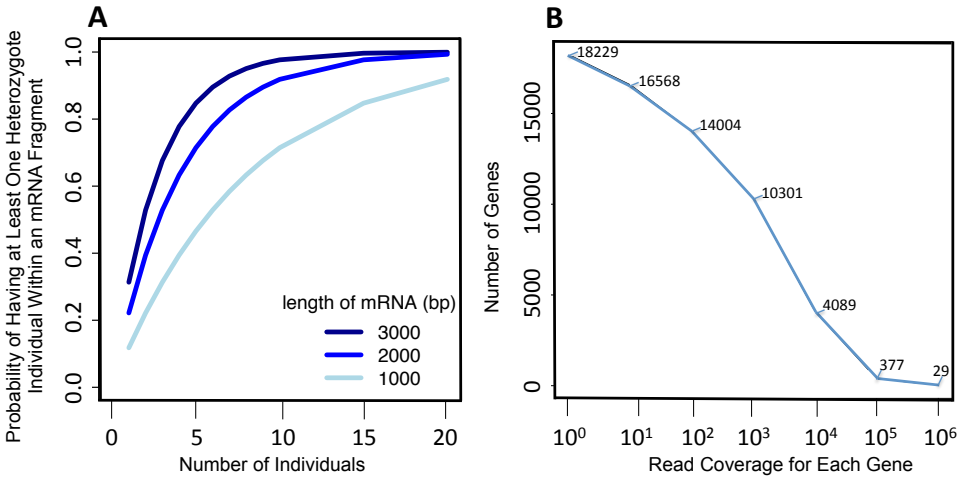


Figure 1 Simulated Heterozygote Distribution and Read Distribution Over Genes. A) Probability of having at minimum one individual with a heterozygote SNP, which is the requirement to assess ASE. B) Describes the number of genes available in respect to different read coverage in Aorta.

4.2 Paper I

Towards our aim of using ASE to detect gene-expression links in experiments of small sample sizes, our first objective was to develop a simple framework in how to organize and

process allelic count data from RNA-sequencing. The result was a software AllelicImbalance[24], and alongside extensive documentation and an introductory vignette, it was accepted and published in the Bioconductor project[63], a bioinformatics code repository written in the R programming language[64]. For development and as examples in the vignette we used public RNA-sequencing data from Montgomery et al.[48], including 16.9 ± 5.9 million poly-A purified 37bp sequencing reads. As second objective, and in order to reveal factors potentially influencing the early steps of library assembly and sequencing, we decided to test the performance of the software in our own pilot experiment, using a read depth of 100 million pair-end, RiboZero and strand-specific sequencing. The in-house aorta and liver samples sequenced were used with the AllelicImbalance package to demonstrate methodological insights: 1) Even in a small sample size of only 8 individuals there is enough heterozygosity to calculate ASE in consecutive exons spanning the same gene region. 2) By measuring txSNPs in consecutive exons it is possible to self-validate an ASE in a gene by confirming a consistent pattern. 3) When using N-masked data, it is possible to achieve a near mapbias-free read count for a gene. 4) Strand-specific ASE can effectively be visualized in a "dual barplot", which represents one strand as upward bars and the other as downward bars. The difference between sense and anti-sense transcription tend to be large, and there can be significant p-values also for the opposing strand. 5) Equal expression of alleles tends to be rather common, and does not show a significant difference even when the read count for each allele is over 10 000, Fig2-a (PAPER I).

4.3 Follow up on RNA-sequenced Samples, Characteristics, Mapbias and Filtering

From our RNA sequenced aorta samples we could demonstrate that about 14000 genes had a read coverage larger than 100 reads, figure 3B. The AllelicImbalance software proved useful to investigate occurrence of ASE, and shone light on a fundamental impact of mapbias for single ASE measures. Even if we offered ways to reduce mapbias influence e.g. N-masking the reference genome, map-bias remains a critical source of doubt for all genes not harboring heterozygotes in multiple exons in the same individual. To increase the reliability, we began to develop a suite of tests to measure the robustness of ASE measures in respect to different applied filters. Although, promising results, extensive filtering also reduce the amount of SNPs amendable for analysis. It started to become clear that more individuals were needed to provide additional support for the detected AIs.

4.4 Paper II

Due to a cost-wise uncertainty if an investment in more RNA sequencing was a justified prioritization, and as RNA sequencing started to become a more common procedure among researchers, we looked for collaborative opportunities to investigate the potential of aeQTL. More samples would allow us to move from assessing the impact of mapbias on a single individuals ASE measures, to a more traditional inference technique using a linear regression, which circumvents mapbias, under the criterion the reference allele is spread equally among the genotypes tested for association. In a collaboration with the Lieber Consortium, we had

access to 782 individuals in two tissues, DLPFC(n=680) and hippocampus(n=421), with an average depth of ~100 million paired-end reads per sample. Complimenting an ongoing eQTL analysis, we designed a study linking Schizophrenia GWAS findings with aeQTL and compared our results to previous eQTLs and eQTLs in the same cohort. We could demonstrate that 1) 4 riskSNPs could only be linked by using aeQTL. 2) For all situations with both an eQTLs and an aeQTLs, their direction of correlation was the same. 3) 340 of the 493 risk-SNP proximate genes were amendable to aeQTL analysis in respect to heterozygosity and read depth 4) RIN value was not a main covariate in our cohort. 5) Using covariates influence eQTL results more than aeQTL. 6) The closest gene to a risk SNP is not necessarily the one under control. 7) In a down-sampling experiment we showed that more aeQTLs and eQTLs could be attributed DLPFC than hippocampus.

4.5 Paper III

In response to the success of applying aeQTL to detect links not possible using standard eQTL in brain tissue, we wanted to further test the method in in-house RNA-sequencing data originating from blood of 150 RA-patients and 55 healthy controls. Multiple eQTLs had previously been detected in a large-scale analysis based on riskSNPs from a GWAS on rheumatology, however far from all risk SNPs were given a functional link. Except the direct application of the method developed in Paper II, we provided a theoretical reasoning on the source of trans-regulatory footprints detectable by comparing aeQTL and eQTL. The main results in paper III were 15 novel QTLs, only detectable by aeQTL, suggesting a strong influence of trans regulatory mechanisms, since Okada et al were not able to establish these links in their eQTL analysis of over 5000 individuals. To test the potential of a more detailed genome-relative investigation of trans-regulatory control within a tissue, we used the same individuals for both eQTL and aeQTL for the same expression readings over each txSNP. However, even for significant gene level eQTLs, no txSNP specific eQTLs were significant, prompting for larger cohorts to allow direct comparisons to txSNP-eQTLs. In summary, paper III demonstrates 1) aeQTLs are detectable in a cohort of both lower sample size as well as sequencing depth compared to the work on SZ. 2) aeQTLs reveal associations not possible using eQTL in thousands of samples. 3) Trans regulation is likely the main source to inter-individual variation of expression as variation from e.g. Brownian motion is stabilized through gene bursts.

4.6 Paper IV

This thesis also describes a methodological work, not obviously linked to previous work on ASE, instead budding from a statistical need in the characterization of gene expression in respect to different diseases. The aim was to provide an assumption free statistical method to assess class cluster separations, which is a common procedure in gene expression analysis to detect e.g. if only long-coding RNA alone can reveal separations of two classes of data. Previously, a separation was established only by visual inspection, a difficult task when classes have a considerate overlap. Instead our method justifies a separation by the use of principal curves to catch the fidelity of the data similar to the human eye, but puts it in a

permutation framework in order to establish a p-value. To facilitate usage of the method, we constructed a software application, which was released within the Bioconductor environment. Subsequently, we published an article describing a use case of selected long non-coding RNAs from public microarray data. Altogether, we used the software to demonstrate that long non-coding RNA represented on microarray to alone be enough to separate all except one pair of cancers types in the data.

5 DISCUSSION

5.1 A "One Hit Wonder"

One of the initial issues in this project was if heterozygosity, a pre-requisite to measure ASE, was widespread enough within exons to measure a sizeable part of the exome. A developed probability model predicted that around 8 individuals would be sufficient to describe ASE for around 80% genes. In theory, one individual could be enough to link regulatory SNPs to expression, provided that only one gene around a risk SNP is subjected to presence of ASE. This we discussed as the one hit wonder possibility. In reality, however, it is rare to fulfill these conditions as e.g. several nearby genes could be subjected to a non-disease risk SNP altering a change from 1:1, and therefore be incorrectly determined to be important factors in disease. Another detrimental factor complicating the discovery of a "one hit wonder" analysis is the difficulty to assess mapbias influence on ASE, making every event unreliable without further validation. Of course, an ASE that is in the favor of the non-reference allele is more likely to represent a true ASE, which can be furthermore supported by comparing the reference/alternative allele composition in the read length for all reads overlapping the examined txSNP. Provided that bias from mapping has no influence on the establishment of ASE, the detection of a large discrepancy from 1:1, also prompts for a strict regulation, by one or more variants in cis. Still provided adequate mapbias handling, for multiple samples it would be possible to find group-specific ASE signals, which would suggest a common regulatory variation responsible to e.g. disease.

5.2 Exploiting Consecutive Heterozygous Exons for Internal Validation

PAPER I, described consistent 1:1 signals over multiple exons in the same gene, and how N-masking corrects deviations due to mapbias. In Fig1 (P1) some txSNPs seem to respond to N-masking, while others don't, suggesting that either the masking was unsuccessful in removing mapbias or there was no mapbias to remove. As the masked reference genome seem to center the expression near 1:1 for all involved txSNPs, speaks for a quite successful reduction using an N-masked genome. Another technical message revealed by studying Fig1(P1) and worth to highlight is that the effect among txSNPs is somewhat different, while the bias seem to occur at almost the same effect size in e.g rs7596677. As the occurrence of additional SNPs around a measured txSNP reduces the mappability of reads coming from that region, one consistent interpretation of our observations is that the region for each txSNP varies more than one region does between individuals, but also a recent shared ancestry among the subjects.

5.3 The Success of Unsuccessful Discrimination Tests

The more reads over a txSNP the smaller deviations from 1:1 can be statistically established, but a small effect size is not necessarily relevant for disease. Interestingly, for some txSNPs covered by thousands of reads not even a tiny difference from 1:1 generated a p-value lower than 0.05, e.g. Fig2:a and b (W1). It supports the idea that alleles truly are equally expressed

when gene-regulation is the same, which is a fundamental assumption in ASE and aeQTL studies. It also confirms RNA sequencing as a technique capable to deliver reliable results. However, these heavily read covered genes might pose an exception in a genomic landscape otherwise full of technically or non-genetic regulator induced ASE. In complex regions of high variation and e.g. enhancer influence a distant variant has a regulatory effect, and it becomes difficult to assure the baseline of equal expression compared to regions of low variation and no enhancer influence.

5.4 Unstable ASE measures Become Robust as aeQTLs

In addition to linking txSNPs to nearby genetic variation, aeQTL gain additional properties by basing the ASE-fraction value on one of the parental phases (haplotypes), which allows for a fraction value based on either the ref or alt-allele. An equal amount of ref-allele in all genotype groups should cancel most of the mapbias influence on the correlation to genotype, i.e. present bias is spread equally among the correlated genotypes. Supplemental S2-2C Paper II can prove an example of this phenomena, as the mean reference fraction is higher than expected 0.5 for heterozygous genotypes tested. Intriguingly, the reference fraction for the homozygous genotype appears to actually be centered on 0.5, possibly explained from a mapbias-causing heterozygote SNP only present when the tested genotype also is heterozygote.

5.5 Library Preparation and Allele Specific mRNA Degradation

Compared to mapbias, library preparation and mRNA degradation are two less investigated sources of allele specific bias. In library preparation several sources of bias have been described:

> *There are three potential sources for technical bias in library preparation: RNA-specific molecular biology (RNA fragmentation, reverse-transcription), RNA selection method (rRNA depletion, polyA selection), and sequencing-specific molecular biology (adapter ligation, library enrichment, bridge PCR).* (Lahens et. al, 2014)[65]

Translated into allele specific terms, in library preparation, restriction endonucleases or primers might favor fragments for one allele. Library preparation protocols using ribosome removal techniques might be more sensitive to allele specific degradation than poly-A capture techniques, i.e., some alleles might be persistent to degradation after tail or cap removal. Another potential source is miRNA directed mRNA degradation, which can act on any part of the mRNA fragment, likely introducing a bias for sites of variation[66,67,68]. However, in the same way as aeQTL inherently reduce mapbias influence on by an equal distribution of bias, we could assume the same to be true for allele bias occurring in any regulatory step that in some way act on a heterozygous fragment.

5.6 Phase Reproducibility, Sequencing and LD

To know if a risk-SNP allele is on the same chromosome as the txSNP requires a phasing procedure. Independent from our RNA sequencing results, we established a phase using the

1000 genomes haplotype panel and genotype data from DNA. Although, it would certainly be an advantage to make use of the phase information within RNA sequencing reads, as the process wouldn't rely on a haplotype panel as reference. For the Shapeit2 software, the process would exclude phase informative heterozygote reads not matching the haplotype panel filling in where phase generation is uncertain. However, we considered that a genotype call from RNA-sequencing comprises a risk to subsequent phasing for cases of wrongly called genotypes, therefore only including genotypes from DNA. Haplotype generation through a reference panel is not perfect, but uncertainty would translate into a random genotype assignment, which would show up as noise rather than a bias. The relation between haplotypes and LD is strong, and as expected we found the reproducibility of haplotypes to be equal or better as a pure LD R^2 estimate, supplemental material S1-5 (PAPER II). In the future, if long-read DNA sequencing costs continues to fall, it would be a more effective way to generate a reliable phase not limited by inference, likely replacing imputation techniques used today.

5.7 Read Depth and Overlap

A substantial read coverage is critical in order to get reliable fraction estimates for aeQTL, which besides heterozygosity represents a stern limitation in respect to todays sequencing costs. In eQTL analysis read-depth is also important, but low levels of total RNA is acceptable, as individuals with a coverage of zero reads for a certain gene still can be useful in comparison to individuals with a coverage over the same gene. After the alignment of reads, there is a quantification step counting the number of overlapping reads over exons or txSNPs. Long reads allow coverage of multiple exons and txSNPs, which directly impact the united fraction values, which we applied to increase sample power for gene-regions (PAPER II and III). Even though there is an underlying complexity of all isoforms originating a region, such as e.g. alternative transcription initiation, alternative splicing, alternative poly-adenylation or alternative translation initiation[69], a merge over all fraction values in a region is still representative to characterize the united gene expression. Using this gene summarization technique, we consider the level of false positives to remain low, as a significant fold change most probably represents a difference regardless of isoform setup. A false negative not due to sample size, would occur when two or more isoforms perfectly cancel eachother. RNA sequencing experiments that use ribosome removal (PAPER I) can catch fragments of premature mRNA, containing intronic regions rich in heterozygosity, which makes it interesting for ASE. However, in competition with mature mRNA fragments the amount of reads available are relatively low, at date making it infeasible for robust ASE measures, but can prove valuable in a future perspective e.g. as a proxy in determination of allele specific degradation rate. Another general complication in eQTL, but where aeQTL performs slightly worse, is for comparing detection rates of QTLs between tissues, which gets inherently difficult as expression levels tend to be tissue-specific, making it statistically possible to detect a QTL in one tissue, whilst the same gene in another tissue sometimes simply haven't the reads required to detect the same effect size, or perform a test at all.

5.8 Fine-mapping aeQTLs in The Era of Personalized Medicine

As a first step, linking GWAS variants to gene expression helps to prioritize among the drug target candidates in the risk variant region. The main limiting factor in an aeQTL analysis is the level of heterozygosity in a gene for a population, which makes it impossible to link genes with no variation. However, after application of quality filters, we were in our SZ-analysis (PAPER II) able to investigate presence of aeQTL in 340 of 493 genes for the 200kb +/- region around the 101 risk loci, and in the RA-analysis (PAPER III) we were able to investigate 911 of 1351 genes for the 500kb +/- region around the 100 risk loci. A gene under regulatory control from a disease-associated polymorphism e.g. found by GWAS, might as well be under control by other polymorphisms, which can prove particularly valuable to detect in personal medicine applications. The potential of aeQTL for fine-mapping has also been in the focus of others offering various methodological approaches[70,71,39]. Like pure ASE measures, fine-mapping will be limited to the same pool of heterozygous genes amendable for analysis, but as tested variants can be homozygous for inclusion in the association model, any additional decrease in sample size only occurs when no more than two groups are available. Similar to eQTL, fine-mapping additional parameters are valuable to take into consideration e.g. picking individuals from different ethnic groups to reduce LD masked polymorphisms. As we slowly enter the era of personalized medicine, aeQTL finemapping might prove useful to map the landscape of genetic influence for existing or new drug target genes.

5.9 Avoidance and Measurement of Trans Regulation

ASE as a measure avoids influence of trans regulation, defined as regulatory mechanisms from other chromosomes. A special case is when an initiation of transcription relies on the fixed presence of another chromosome, which for one cell then could contribute with a form of cis expression, slightly breaking its definition as the regulatory difference is present on another chromosome than the gene investigated. However, in pooled samples, the chromosome-chromosome interactions are likely random, enforcing ASE:s capability to detect in cis associations.

5.10 ClusterSignificance

Like other tests, the larger the sample size, the smaller the effect size results in a significant p-value. The recommended follow-up on a significant separation is therefore to evaluate if the result is meaningful in terms of effect. However, a problem with dimensionally reduced data is that effect size can be difficult to interpret, as the transformation from high-dimensional to low-dimensional space is non-trivial. Nevertheless, an established difference does not necessarily depend on the effect size to be meaningful, as a difference always origins from a real difference. Another issue is the final required reduction down to one-dimensional space, in that every reduction excludes information. However, to minimize the risk of loosing the separating information in that last reduction, several different dimensionality reduction algorithms can be tested on the data before testing for a separation in ClusterSignificance.

5.11 R software Design Considerations

Both AllelicImbalance and ClusterSignificance are products of optimization in storage and calculative operations, as well as functional extendibility design, which does not interfere with existing users, i.e., new functions do not disrupt existing workflows. AllelicImbalance is a general tools package for ASE operations, and has been used extensively throughout Paper I, II and II. The ASEset class object functions as a container of pre-processed sequencing, genetoype and other sample variables like e.g. covariate measures. The data storage is copy-on-modification, lowering the memory usage and processing time, required to smoothly process millions of SNPs in a single run. The package takes full advantage of the GenomicRanges-suit in Bioconductor, which lowers over-head maintenance costs and facilitates integration and usability within the community. Some updates under consideration are handling of txSNP indels and shrinking the count data using sparse-matrices or other index based implementations. One non-released branch of a sparse-matrices implementation exists, but the development required assistance of many basic arithmetic operations to accommodate the current 3D-array, which is the internal representation. This shines light of a conflict between storage and rapid processing, i.e., it is very unnecessary to use a top of the art storage system if every access requires a transformation back to a memory-consuming structure to be able to use existing arithmetic operations. The development of arithmetic operations for tailored indexing has potential, but risks a large over-head cost. Generally, if computer capacity is not increasing in the same speed as more data become available for analysis, we will be forced to balance more time on optimization than interpretation. ClusterSignificance is compared to AllelicImbalance not a general tools package, and requires less flexibility in storage considerations. Expected input is merely the output of a dimensional reduction algorithm in the format of a 2D-matrix, limited by sample and feature size.

5.12 Future Perspectives

With recent technical innovations it is possible to apply more sample-efficient as well as cell specific ASE examinations.

A more careful design can increase statistical power for genes with low read depth or with few heterozygots

One path to reduce the loss of data from filtering heterozygote sites, is to perform an experiment from e.g. biobank samples previously genotyped and only include samples harboring at least one heterozygote in e.g. genes previously indicated important in disease. Combining this pre-selection with a full multiplex RNA-sequencing experiment[72] focusing on a probe-capture[73] lowly expressed genes can reveal their relatively unexplored landscape of genetic regulation in a high-throughput setting; lowly expressed genes that otherwise only accessible through very costly deep sequencing experiments.

Single cell sequencing can reveal spatial ASE and within each cell type

RNA sequencing for single cells has recently become feasible, and especially important for e.g. lineage tracing during cellular differentiation and organ or embryo development[74]. Using single cell sequencing information to measure ASE is an interesting option but has to take into account e.g. that cells burst allele specific expression[75]. This would reduce the ability to perform accurate ASE measurement, i.e. an individual might shift which allele is the only one expressed. This makes it impossible to get a robust measure of ASE for genetic linking, using just one cell at one time point. Instead, using RNA sequencing on a population of cells will comprise the mean expression of all RNA molecules among the pooled cells, and thus represent a better estimate of ASE. Bioinformatically, there is no difference whether the merge of allele expression takes place before or after the sequencing, and in that respect there is no strong gain in using single cells to address ASE, compared to sequence a whole population of cells. On the contrary, there has been a certain success for single cell eQTLs, using connectivity between cells to reduce the influence of trans-regulatory variability, which prompts a possibility to enhance ASE measures using the same principles. Additionally, single cell sequencing can reveal detailed spatial distribution within tissues[76], in turn revealing potential differences in ASE within the same tissue, triggered by the use of different transcription factors. This reasoning is an extension of the tissue specific results from e.g. Korkuba et al[25], and organ differentiation by Saliba et al[74]. Single-cell sequencing has the potential of narrowing down genetic contributed ASE in smaller spatial tissue space than in pooled tissue samples, but maybe more important is its capability to purify ASE measures for mixed cell population. Additionally, single-cell sequencing has also been proven useful for phase generation, facilitating coherent ASE within a gene[77].

Digital qPCR can be used to validate ASE findings from RNA-sequencing

Another way to validate ASE-events is to use digital qPCR[78], which similar to RNA-sequencing can quantify and discern the alleles, in the existence of heterozygote sites. However, this method is not standard procedure and could likely be exposed to bias in the library construction, whereas any of the alleles may be more or less easy to replicate.

Suggestions to reduce the need for heterozygosity to assess ASE

The perfect aeQTL experiment totally avoids the need for heterozygosity, but adding an artificial marker to only one portion of identical RNA fragments based on ancestry is difficult to achieve. An unelaborate proposal would be to infer allele specific expression based on additional regulatory mechanisms, such as histone markers, chromatin-status coupled with the information depth provided by single cell expression measures. Or more wishfully, a micro-monitoring device, which captures and follows all mRNA processes in living cells.

aeQTL can be a major player in advancing personalized medicine

Personalized medicine is part of the EU agenda and is described as " a fast-growing market and Europe's healthcare industry has the potential to build on its leading position, providing economic growth and jobs." (European Commission[79]). Many drugs act on repressing gene expression or their protein product to lower the effect of a protein involved in a certain disease. The individual response varies, and in treatment the choice of drug is today more a

"trial and error" process than based on scientific evidence. Therefore, to reveal genetic influence on drug-targeted genes and providing the markers affecting the transcriptional machinery is of great public interest. Using aeQTL to connect genetic markers and the expression for e.g. drug-targeted genes, allows the pharmaceutical industry to select candidates less influenced of genetic regulation or to specifically address a smaller portion of the population in the clinical trial.

5.13 Ethical Risks

Genetic analysis is always sensitive, since the donors' genetic information also can be used to infer attributes for their relatives. This thesis is using genotypes from a hundreds of individuals, but which is a small figure compared to the thousands of individuals used in e.g. GWAS studies. With that said, even if the sample size is small in comparison to GWAS studies, using anonymized keys we kept the integrity of our participants to not let their contribution end up where they did not consent.

If we extend the ethical concerns from how to handle the data in a secure way to discuss the general danger of genetic research, can it be so that the bad effects could outweigh the good? There are several moral concerns regarding economy, humanity and the environment. E.g. insurance companies might request genetic information to set up a lower fee for individuals with a genome less likely to experience high-cost diseases. Similar in dating, individuals might be locked out of society by not having a preferable genome. All research can end up being misused with the wrong intentions, but as long as we are aware of the dangers, and can provide sufficient transparence, the risks might be worth taken. The alternative would be to stop science and stay satisfied with how life is today.

6 FINAL REMARKS

ASE analyses are of growing importance in the field of genetic linkage analysis, while it is also the closest genetic active link to explain expression. This will become increasingly important in order characterize specific drug-target genes for the use in personalized medicine. Despite all possible harmful elements that risk biasing an ASE analysis, it seems possible to design experiments so the distribution of test groups allows e.g. measuring a single gene target to be analyzed with the full power of ASE. Even if the global analysis methods are important to screen out the mass QTL effects, the non-symmetrical heterogeneity among SNPs makes it difficult to compare QTLs with each other. Unless the pool of accessible samples is not large enough, p-values and effect size measures will only indicate associations, not their relative importance. To be used in a clinical setting the findings are still required to be validated and independently quality controlled for bias. Gene or region centric quality control tools, like the ones in the AllelicImblance package[24], are thus of importance.

The ClusterSignificance is a statistical tool to set p-values of cluster separations, especially an aid for overlapping separations where visual inspection is not enough. Like the AllelicImbalance package, it is an open-source and free package completely written in the R programming language and can be downloaded from the Bioconductor repository.

7 ACKNOWLEDGEMENTS

Lasse, even though we in 99.9% of all time have been communicating via web resources, it is nevertheless so that you might be the person I have spent most conversational time together with during these years, in fact you never felt especially far away. No matter what time or day, you have mostly replied to my messages within a minute. Thank you for being a realist when I am naive, and motivator when things have gone less well. Regardless of scientific achievements, I have at least for others been an active chat-link to your knowledge. You have many times said "Yes, I did it like that first, but then I changed opinion, you will do the same in due time", and as I secretly start to like Word, it must mean you were right also for that extreme case.

Per, The first time Lasse talked about you; he mentioned that you like good hardware for computers. To my happy surprise I later experienced that as true, if you remember we did even discuss setting up a computer cluster of our own. However, besides the pure scientific skills you have given me, you always transfer sparks of motivational energy to facilitate whatever challenges are ahead. I did like the book Stoner very much.

Ferdinand, thank you for all articles put on my office desk, they have all been very valuable in this work. I do miss our discussion on how to validate ASE signals; a lot of paper and white board pens were consumed.

Dan, even though we were in the middle of a cardio-program at the gym, you were always up for the latest news in the field of ASE science or anything else importance in life. Thanks for initiating our bioinformatic collaboration; I am significantly wiser now when it comes to statistics. The gym won't be the same without you.

Jason, you are the only person I know who would be able to compete with Lasse in straightforwardness, and effectiveness. I am looking forward to your thesis, if to any help, I can recommend two tools I'm personally found of, Knee-wizardry and Ninja editing, works like a charm.

All group members and roommates. **Louisa**, my dear office neighbor, you are both very kind and very smart, thanks for being you. I am sorry if my raising and lowering of the table made your life a mess sometimes. **Lei**, my other office neighbor and bioinformatics soulmate, like a Jante you master the art of concealed appreciation and encouragement. **Hanna**, It feels great that you have returned from absence, and perhaps you are moving back to Kungsholmen as well? **Valentina**, without you my social life would have involved much fewer people and activities. Remember, Partek is a good first step to become a bioinformatician. **Karin**, besides being a fundamental asset in the group, your northern accent steam stability. **Flore-Anne**, I promise we will have time for that mining game after the defense. **Nancy**, thank you for all coffee breaks and the coffee mug, it makes me feel special. **Apostholos**, I propose that we continue our interest in coffee, and practice fika like a pro. **Shohreh**, thanks for sharing your hidden insights into how academia functions, and that you always dare to discuss sensitive

matters. **Maria**, if you would have entered the office a little earlier, we could have discussed much more about genetics, R-scripting, or carrot juice. **Peter**, thank for co-experiencing all the phases of pursuing a PhD.

Anton, if it would not have been for you, the gym would have been a much more boring place, and thanks for introducing me to Dan, it resulted in a pretty cool article. And FYI, this thesis was written using all eccentric powers available. **Ulf**, if it would not have been for your Urkult initiative, my life might well have taken a different direction than pursuing a PhD. **Gabrielle**, thank you for the opportunity to work in your lab during my final project for the masters, and for the opportunity to work with the BiKE database in the EWE project. **Eva P**, thanks for dragging me out of all administrational nightmares.

Nu först förstår jag vad du **Mamma** genomgick när jag var liten, och vad en doktorand verkligen innebär. Tack för allt engagemang du och **Pappa** alltid visat inför de idéer jag kommit upp med genom åren. Men vem vore jag utan min bror **Andreas**, en livskamrat i både små och stora projekt. Detta hade inte varit möjligt utan den resa som varit innan. **Isabelle**, hoppas den senaste tidens anspänning och sena nätter inte varit dig ovän. Tack för all din extra omtanke om mig, men det viktigaste är och har hela tiden varit att du är du.

8 REFERENCES

1. Mendel, G. Experiments in Plant Hybridisation, [translated by Charles T. Druery], with an introductory note by W. Bateson. *J. R. Hortic. Soc.* **26,** 1–32 (1901).

2. Westerlund, J. F. & Fairbanks, D. J. Gregor Mendel's classic paper and the nature of science in genetics courses: Gregor Mendel and the nature of science in genetics courses. *Hereditas* **147,** 293–303 (2010).

3. Jiang, X., Hu, S., Xu, Q., Chang, Y. & Tao, S. Relative effects of segregation and recombination on the evolution of sex in finite diploid populations. *Heredity* **111,** 505 (2013).

4. Sherry, S. T. & Ward, M.-H. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

5. Milani, L. *et al.* Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells. *Nucleic Acids Res.* **35,** e34 (2007).

6. Gaur, U., Li, K., Mei, S. & Liu, G. Research progress in allele-specific expression and its regulatory mechanisms. *J. Appl. Genet.* **54,** 271–283 (2013).

7. Serre, D. *et al.* Differential Allelic Expression in the Human Genome: A Robust Approach To Identify Genetic and Epigenetic Cis-Acting Mechanisms Regulating Gene Expression. *PLoS Genet.* **4,** e1000006 (2008).

8. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B Biol. Sci.* **368,** 20120362–20120362 (2013).

9. Westra, H.-J. & Franke, L. From genome to function by studying eQTLs. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* **1842,** 1896–1902 (2014).

10. Cheung, V. G. *et al.* Monozygotic Twins Reveal Germline Contribution to Allelic Expression Differences. *Am. J. Hum. Genet.* **82,** 1357–1360 (2008).

11. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47,** 88–91 (2014).

12. Stevenson, K. R., Coolon, J. D. & Wittkopp, P. J. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics* **14,** 536 (2013).

13. Quinn, A., Juneja, P. & Jiggins, F. M. Estimates of allele-specific expression in Drosophila with a single genome sequence and RNA-seq data. *Bioinformatics* **30,** 2603–2610 (2014).

14. Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.* **32,** 261–266 (2014).

15. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci.* **111,** 9869–9874 (2014).

16. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25,** 3207–3212 (2009).

17. Ge, B. *et al.* Survey of allelic expression using EST mining. *Genome Res.* **15,** 1584–1591 (2005).

18. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.* **11,** 533–538 (2010).

19. Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* **21,** 1728–1737 (2011).

20. Zhang, X. & Borevitz, J. O. Global Analysis of Allele-Specific Expression in Arabidopsis thaliana. *Genetics* **182,** 943–954 (2009).

21. Liu, Z. *et al.* Comparing Computational Methods for Identification of Allele-Specific Expression based on Next Generation Sequencing Data. *Genet. Epidemiol.* **38,** 591–598 (2014).

22. Cho, H. *et al.* High-Resolution Transcriptome Analysis with Long-Read RNA Sequencing. *PLoS ONE* **9,** e108095 (2014).

23. Panousis, N. I., Gutierrez-Arcelus, M., Dermitzakis, E. T. & Lappalainen, T. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol.* **15,** 467 (2014).

24. Gådin, J. R., van't Hooft, F. M., Eriksson, P. & Folkersen, L. AllelicImbalance: An R/Bioconductor package for detecting, managing, and visualizing allele expression imbalance data from RNA sequencing. *BMC Bioinformatics* **16,** 194 (2015).

25. Kukurba, K. R. *et al.* Allelic Expression of Deleterious Protein-Coding Variants across Human Tissues. *PLoS Genet.* **10,** e1004304 (2014).

26. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522–522 (2014).

27. Turro, E. *et al.* Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* **12,** R13 (2011).

28. Pandey, R. V., Franssen, S. U., Futschik, A. & Schlötterer, C. Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Mol. Ecol. Resour.* **13,** 740–745 (2013).

29. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).

30. Soderlund, C. A., Nelson, W. M. & Goff, S. A. Allele Workbench: Transcriptome Pipeline and Interactive Graphics for Allele-Specific Expression. *PLoS ONE* **9,** e115740 (2014).

31. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12,** 1061–1063 (2015).

32. Mayba, O. *et al.* MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol.* **15,** 405 (2014).

33. Munger, S. C. *et al.* RNA-Seq Alignment to Individualized Genomes Improves Transcript Abundance Estimates in Multiparent Populations. *Genetics* **198,** 59–73 (2014).

34. Younesy, H. *et al.* ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics* **30,** 1172–1174 (2014).

35. Li, G. *et al.* Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.* **40,** e104–e104 (2012).

36. Harvey, C. T. *et al.* QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics* **31,** 1235–1242 (2015).

37. Edsgärd, D. *et al.* GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Sci. Rep.* **6,** (2016).

38. Pirinen, M. *et al.* Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics* **31,** 2497–504 (2015).

39. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48,** 206–213 (2015).

40. Wei, Y., Li, X., Wang, Q. & Ji, H. iASeq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. *BMC Genomics* **13,** 681 (2012).

41. Hastie, T. & Stuetzle, W. Principal Curves. *J. Am. Stat. Assoc.* **84,** 502 (1989).

42. *Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 135).*

43. Carlson, M. & Maintainer, B. *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s). R package version 2.15.0.* (2012).

44. Hackstadt, A. J. & Hess, A. M. Filtering for increased power for microarray data analysis. *BMC Bioinformatics* **10,** 11 (2009).

45. Griffith, M. *et al.* Alternative expression analysis by RNA sequencing. *Nat. Methods* **7,** 843–847 (2010).

46. Folkersen, L. *et al.* Unraveling divergent gene expression profiles in bicuspid and tricuspid aortic valve patients with thoracic aortic dilatation: the ASAP study. *Mol. Med.* **17,** 1365 (2011).

47. Ares, M. Methods for Processing High-Throughput RNA Sequencing Data. *Cold Spring Harb. Protoc.* **2014,** pdb.top083352-top083352 (2014).

48. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).

49. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet.* **10,** e1004234 (2014).

50. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* **5,** e1000529 (2009).

51. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

52. Maaten, L. van der & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).

53. Vijaya Satya, R., Zavaljevski, N. & Reifman, J. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Res.* **40,** e127–e127 (2012).

54. *Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 138).*

55. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

56. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

57. Carlson, M. & Maintainer, B. *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s). R package version 3.2.0.* (2015).

58. Carlson, M. & Maintainer, B. *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s). R package version 3.3.2.* (2017).

59. Folkersen, L. *et al.* Association of Genetic Risk Variants With Expression of Proximal Genes Identifies Novel Susceptibility Genes for Cardiovascular Disease. *Circ. Cardiovasc. Genet.* **3,** 365–373 (2010).

60. Almlöf, J. C. *et al.* Powerful Identification of Cis-regulatory SNPs in Human Primary Monocytes Using Allele-Specific Gene Expression. *PLoS ONE* **7,** e52260 (2012).

61. Van Der Maaten, L., Postma, E. & Van den Herik, J. Dimensionality reduction: a comparative. *J Mach Learn Res* **10,** 66–71 (2009).

62. Song, B., Zhang, G., Zhu, W. & Liang, Z. ROC operating point selection for classification of imbalanced data with application to computer-aided polyp detection in CT colonography. *Int. J. Comput. Assist. Radiol. Surg.* **9,** 79–89 (2014).

63. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12,** 115–121 (2015).

64. R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.* (2017).

65. Lahens, N. F. *et al.* IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.* **15,** R86 (2014).

66. Norgren, N., Hellman, U., Ericzon, B. G., Olsson, M. & Suhr, O. B. Allele Specific Expression of the Transthyretin Gene in Swedish Patients with Hereditary Transthyretin Amyloidosis (ATTR V30M) Is Similar between the Two Alleles. *PLoS ONE* **7,** e49981 (2012).

67. Yousef, G. M. miRSNP-Based Approach Identifies a miRNA That Regulates Prostate-Specific Antigen in an Allele-Specific Manner. *Cancer Discov.* **5,** 351–352 (2015).

68. Võsa, U., Esko, T., Kasela, S. & Annilo, T. Altered Gene Expression Associated with microRNA Binding Site Polymorphisms. *PLOS ONE* **10,** e0141351 (2015).

69. de Klerk, E. & 't Hoen, P. A. C. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet.* **31,** 128–139 (2015).

70. Kang, E. Y. *et al.* Discovering Single Nucleotide Polymorphisms Regulating Human Gene Expression Using Allele Specific Expression from RNA-seq Data. *Genetics* **204,** 1057–1064 (2016).

71. Cheng, H. H. *et al.* Fine mapping of QTL and genomic prediction using allele-specific expression SNPs demonstrates that the complex trait of genetic resistance to Marek's disease is predominantly determined by transcriptional regulation. *BMC Genomics* **16,** 816 (2015).

72. Hou, Z. *et al.* A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Sci. Rep.* **5,** (2015).

73. Mercer, T. R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9,** 989–1009 (2014).

74. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* **42,** 8845–8860 (2014).

75. Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science* **343,** 193–196 (2014).

76. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).

77. Edsgärd, D., Reinius, B. & Sandberg, R. scphaser: haplotype inference using single-cell RNA-seq data. *Bioinformatics* **32,** 3038–3040 (2016).

78. Baker, M. Digital PCR hits its stride. *Nat. Methods* **9,** 541–544 (2012).

79. European Commission, Research & Innovation, Health, Policies, Personalized Medicine, https://ec.europa.eu/research/health/index.cfm?pg=policy&policyname=personalised, Accessed: 19-10-2017.