Laurent F. Thomas

# Effects of Single-nucleotide Polymorphisms on microRNA-Based Gene Regulation and Their Association With Disease

Thesis for the degree of Philosophiae Doctor

Trondheim, November 2012

Norwegian University of Science and Technology
Faculty of Medicine
Department of Cancer Research and Molecular Medicine

NTNU – Trondheim
Norwegian University of
Science and Technology

INTERAGON

**NTNU**
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Medicine
Department of Cancer Research and Molecular Medicine

ISBN 978-82-471-3932-5 (printed ver.)
ISBN 978-82-471-3933-2 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2012:307

Printed by NTNU-trykk

NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET
DET MEDISINSKE FAKULTET

# Sammendrag

## Effekter av enkeltnukleotidpolymorfismer på mikroRNA-basert genregulering og deres assosiasjon med sykdom

DNA-et inneholder variasjoner mellom individer, og DNA-varianter slik som enkeltnukleotidpolymorfismer (SNP) kan påvirke genfunksjon, men også fenotyper. I løpet av de siste årene, har helgenom assosiasjonsstudier (GWAS) forsøkt å identifisere vanlige SNP-er som er assosiert med vanlige sykdommer, slik som kreft. Mange assosierte SNP-er har blitt funnet utenfor protein-kodende regioner og har vært vanskelig å tolke ettersom de ikke endrer proteinstrukturer og funksjoner, men er antatt å ligge i eller i nærheten av genregulatoriske regioner.

For å bedre forstå mekanismene bak slike uforklarte sykdomsassosierte varianter, har vi studert SNP-er involvert i dysregulering av gener, og spesielt de som påvirker genregulering via microRNA (miRNA).

Først identifiserte vi SNP-er som potensielt forstyrrer eller skaper miRNA bindingsseter (miRSNP), og kvantifiserte disse miRSNP-enes effekt på genregulering. Dessuten utviklet vi en metode for å koble miRSNP-ene til sykdomsassosierte SNP-er fra GWAS, for å identifisere sykdom-disposisjon eller kausale miRSNP-er. Ved hjelp av denne metoden, identifiserte vi en miRSNP (rs1434536) som påvirker reguleringen av miRNA mir-125b på genet Bone Morphogenetic Protein Receptor type-1B (*BMPR1b*). Denne SNP-en har vært assosiert til brystkreft, og dens effekt på *BMPR1b* uttrykksnivå ble verifisert eksperimentelt, noe som tyder på at denne SNP-en resulterer i økt disposisjon for brystkreft ved å påvirke miRNA-basert regulering.

Dernest studerte vi regulatoriske varianter (SNP-er) som kan forkorte messenger RNA (mRNA) gjennom alternativ polyadenylering (APA), og spesielt de SNP-ene som kan danne slike APA-signaler. Forkorting kan resultere i tap av regulatoriske regioner som miRNA bindingsseter og dermed påvirke genuttrykk. Vi identifiserte potensielle APA-SNP-er og testet vår hypotese om at APA-SNP-er kan oppregulere genuttrykk gjennom forkorting av mRNA og tap av miRNA bindingsseter.

***Navn kandidat: Laurent F. Thomas***
***Institutt: Institutt for kreftforskning og molekylærmedisin***
***Veiledere: Pål Sætrom (hovedveileder), Finn Drabløs (medveileder)***
***Finansieringskilder: Interagon AS og Nærings-ph.d. fra Norges Forskningsråd.***

*Ovennevnte avhandling er funnet verdig til å forsvares offentlig*
*for graden PhD i medisinsk teknologi.*
*Disputas finner sted i Auditoriet, Blåhø i Øya Helsehus*
*onsdag 21. november 2012, kl. 12.15.*

NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY
FACULTY OF MEDICINE

# Abstract

## Effects of single-nucleotide polymorphisms on microRNA-based gene regulation and their association with disease

The DNA contains variations between individuals, and DNA variants such as single nucleotide polymorphisms (SNPs) may affect gene functions, but also phenotypes. Over the past few years, genome-wide association studies (GWAS) tried to identify common SNPs that are associated with common diseases, such as cancer. However, many associated SNPs were found outside protein-coding regions and have been difficult to interpret as they do not change protein structures and functions, but are thought to lie in or near gene regulatory regions.

To better understand the mechanisms behind unexplained disease-associated variants, we studied SNPs involved in gene dysregulation, and particularly those affecting gene regulation by microRNAs (miRNAs).

First, we identified SNPs potentially disrupting or creating miRNA binding sites (miRSNPs), and tried to quantify miRSNP effects on gene regulation. Furthermore, we described a method to relate miRSNPs to disease-associated SNPs from GWAS, to help identify disease-susceptibility or causal miRSNPs. Using this method, we identified a miRSNP (rs1434536) that affects the regulation of the miRNA miR-125b on the gene Bone Morphogenetic Protein Receptor type 1B (*BMPR1b*). This SNP has been associated with breast cancer and its effect on *BMPR1b* expression level has been verified experimentally, suggesting that this SNP results in increased breast cancer susceptibility by affecting miRNA-based regulation.

Second, we were interested in another type of regulatory variants; SNPs that can shorten messenger RNAs (mRNAs), through alternative polyadenylation (APA), and particularly SNPs that create APA signals. The shortening can result in loss of regulatory regions such as those where miRNAs bind, and thereby affect gene expression. We identified potential APA-SNPs and tested our hypothesis that APA-SNPs can upregulate gene expression through shortening of mRNAs and loss of miRNA binding sites.

UNIVERSITÉ NORVÉGIENNE DE SCIENCES ET DE TECHNOLOGIE, FACULTÉ DE MÉDECINE

# Résumé

## Effets des polymorphismes nucléotidiques simples sur la régulation des gènes par microARNs et leur association aux maladies

L'ADN diffère entre les individus. Ces variations génétiques, et en particuliers les polymorphismes nucléotidiques simples (SNP), peuvent affecter les fonctions des gènes, mais aussi les phénotypes. Au cours des dernières années, des études d'association pangénomique (GWAS) ont tenté d'identifier parmi les SNPs relativement fréquents, ceux qui sont associés à des maladies génétiques multifactorielles telles que le cancer. Cependant, de nombreux SNPs associés à ces pathologies se trouvent à l'extérieur des régions codant pour des protéines et ont été difficiles à interpréter car ils ne changent pas la structure et la fonction des protéines, mais on pense qu'ils se situent dans, ou à proximité, de régions régulatrices des gènes.

Pour mieux comprendre les mécanismes derrière ces prédispositions encore incomprises, nous avons étudié les SNPs impliqués dans la dérégulation des gènes, en particulier ceux qui affectent la régulation des gènes par des transcrits tels que les microARNs (miARN).

Tout d'abord, nous avons identifié des SNPs qui potentiellement perturbent ou créent des sites où se fixent les miARNs (miRSNPs), et nous avons essayé de quantifier leur effets sur la régulation des gènes. En outre, nous avons décrit une méthode pour relater les miRSNPs aux SNPs déjà associés à des pathologies lors de GWAS, afin d'aider à identifier les miRSNPs qui prédisposent ou causent des maladies. En utilisant cette méthode, nous avons identifié un miRSNP (rs1434536) qui influe sur la régulation du gène du récepteur type IB de la protéine morphogénétique osseuse (*BMPR1b*) par le biais du miARN miR-125b. Ce SNP a été associée au cancer du sein et son effet sur le niveau d'expression de *BMPR1b* a été vérifié expérimentalement, ce qui suggère que ce SNP prédispose au cancer du sein en affectant les miARNs.

Deuxièmement, nous nous sommes intéressés à un autre type de variantes régulatrices : les SNPs qui peuvent raccourcir les ARN messagers (ARNm), par le biais d'une polyadénylation alternative (APA), et en particulier les SNPs qui créent des signaux alternatifs de polyadenylation. Ce raccourcissement peut entraîner la perte de régions régulatrices telles que celles où les miARNs se fixent, et donc affecter l'expression des gènes. Nous avons identifié des APA-SNPs potentiellement fonctionnels et testé notre hypothèse où l'APA-SNP peut croître l'expression des gènes par le raccourcissement des ARNm et la perte des sites de fixation des miARNs.

# Acknowledgments

I would like to thank my main supervisor Pål Sætrom. Writing this thesis would not have been possible without his good advices and support. I also would like to thank my co-supervisor Finn Drabløs, as well as articles' co-authors, my colleagues at the medical faculty, my friends and family. I would also like to thank Interagon AS and the Norwegian Research Council for funding this work, as well as the Norwegian University of Science and Technology.

# Contents

# List of papers

**Paper I** **Laurent F. Thomas**, Takaya Saito, and Pål Sætrom. **Inferring causative variants in microRNA target sites**. *Nucleic Acids Research*, 39(16), SEP 2011.

**Paper II** Pål Sætrom, Jacob Biesinger, Sierra M. Li, David Smith, **Laurent F. Thomas**, Karim Majzoub, Guillermo E. Rivas, Jessica Alluin, John J. Rossi, Theodore G. Krontiris, Jeffrey Weitzel, Mary B. Daly, Al B. Benson, John M. Kirkwood, Peter J. O'Dwyer, Rebecca Sutphen, James A. Stewart, David Johnson, and Garrett P. Larson. **A Risk Variant in an miR-125b Binding Site in BMPR1B Is Associated with Breast Cancer Pathogenesis**. *Cancer Research*, 69(18):7459–7465, SEP 15 2009.

**Paper III** **Laurent F. Thomas** and Pål Sætrom. **Single Nucleotide Polymorphisms Can Create Alternative Polyadenylation Signals and Affect Gene Expression through Loss of MicroRNA-Regulation**. Manuscript accepted in *PLoS Computational Biology*, 2012, [In Press].

# List of Figures

# Abbreviations

| | |
|---|---|
| Ago | Argonaute protein, |
| APA | Alternative polyadenylation, |
| CDCV | Common-disease common-variant hypothesis, |
| cDNA | Complementary DNA, |
| CDRV | Common-disease rare-variant hypothesis, |
| CDS | Coding sequence, |
| CGAS | Candidate gene association study, |
| cM | Centimorgan, |
| CNV | Copy number variant, |
| DNA | Deoxyribonucleic acid, |
| DSE | Downstream sequence element, |
| EM | Expectation maximization, |
| eQTL | Expression quantitative trait locus, |
| GEO | Gene expression omnibus, |
| GWAS | Genome-wide association study, |
| HMM | Hidden markov model, |
| HWE | Hardy-Weinberg equilibrium, |
| IBD | Identical by descent, |
| IBS | Identical by state, |
| Indel | Insertion-deletion, |
| LD | Linkage disequilibrium, |
| lncRNA | Long non-coding RNA, |
| LOD | Logarithm of the odds, |

MAF  Minor allele frequency,

MFE  Minimum free energy,

miRNA  MicroRNA,

miRSNP  MicroRNA polymorphism,

mRNA  Messenger RNA,

mRNP  mRNA ribonucleoprotein complex,

ncRNA  Non-coding RNA,

NGS  Next-generation sequencing,

NMD  Nonsense-mediated decay,

NPC  Nuclear pore complex,

nsSNP  Non-synonymous SNP,

PAS  Polyadenylation signal,

piRNA  Piwi-interacting RNA,

pre-miRNA  Precursor microRNA,

pre-mRNA  Precursor mRNA,

pri-miRNA  Primary microRNA,

PTC  Premature termination codon,

qPCR  Quantitative polymerase chain reaction,

QTL  Quantitative trait loci,

RISC  RNA-induced silencing complex,

RNA  Ribonucleic acid,

RNABP  RNA-binding protein,

RNAP  RNA polymerase,

SBE  Single-base extension,

SNP  Single-nucleotide polymorphism,

sSNP  Synonymous SNP,

SSR  Simple sequence repeat,

STR  Short tandem repeat,

TREX  Transcription and export complex,

tRNA Transfer RNA,

USE Upstream sequence element,

UTR Untranslated region,

WGAS Whole-genome association study,

XPO5 Exportin-5 protein,

# Chapter 1

# Introduction

The study of DNA variants and mutations aims at understanding genetic mechanisms involved in diseases. Variants can play a role in pathogenesis, in prognosis, or in treatment response and efficiency. Those that affect protein sequences and therefore gene functions are supposed to be less likely viable and more common among rare diseases. However, since regulatory variants affect gene expression instead of gene function, the proteins produced by the cell are viable, but deregulated. That is why regulatory variants are thought to play an important role in common diseases. Furthermore, a small change in gene expression can have important phenotypical consequences.

In this thesis, I looked at two kinds of regulatory variants. The first directly affects a type of regulatory elements where non-coding RNAs can bind, resulting in gene dysregulation and potentially in disease. The second affects the length of the messenger RNA sequence, which carry the information needed to build its corresponding protein. Shorter messenger RNAs can lose regulatory elements often found at the end of their sequence. This mechanism can also result in gene dysregulation and disease.

First, I will describe the biological background of my thesis, focusing on protein-coding and non-coding genes, as well as DNA variants in general and particularly the regulatory variants. Second, I will describe common technologies involved in genetics and the strategies in disease-association. Third, I will describe relevant algorithms that integrate the data produced by these technologies and these association strategies. Finally, I will detail the aim of my project, sum up my results, and mention some future perspectives.

# Chapter 2

# Biology

Genes are important parts of cells and therefore of living organisms. Some of those genes, termed coding genes, are the recipes specifying how to make proteins, through the formation of important intermediate molecules called messenger RNAs (mRNAs). Other types of genes do not produce proteins (non-coding genes) and may have less obvious functions than coding ones. However, one type of non-coding genes called microRNAs is now quite well understood, as it plays a role in the regulation of mRNAs. Finally, polymorphisms are DNA variants occurring in the DNA sequence, that can affect gene sequences and their resulting protein, as well as gene expression, in which case they are called regulatory variants. I shall focus in this chapter on how coding genes lead to proteins, the regulatory role of non-coding genes like miRNAs, and the characteristics of DNA polymorphisms, particularly those of single nucleotide polymorphisms and their effects on the two gene types above.

## 2.1 Coding genes

Coding genes are genes that code for proteins. Those genes are first transcribed into a molecule called messenger RNA (mRNA), through several processes known as mRNA biogenesis. Depending on many factors, this biogenesis can occur in different ways resulting in different mRNAs (mRNA isoforms). Finally, mRNAs are translated into proteins.

### 2.1.1 Definition of Messenger RNA

Messenger RNA is a ribonucleic acid (RNA) molecule that is built inside the nucleus and transported into the cytoplasm to be translated into protein. Here, I briefly describe the general role of mRNAs and how they are structured.

#### 2.1.1.1 Role

The genetic information is safely stored inside the nucleus as deoxyribonucleic acid (DNA), and can be used by the cell to build proteins. However, protein synthesis occurs in the cytoplasm. The role of mRNA is to work as an intermediate between DNA and proteins, by carrying into the cytoplasm the information needed to build its corresponding protein.

#### 2.1.1.2 Structure

Coding genes consists of successive regions called exons and introns, where only exons are kept in the mature mRNA. Depending on the circumstances, some exons can be either kept or not in the mature mRNA, and are called pseudoexons [1]. Annotations of human mRNAs such as exon positions are available in the RefSeq database [2].

To be able to carry information through its structure, mature mRNAs seem to have evolved into a molecule consisting of five main parts: a modified base at its 5' end called the 5' cap, followed by a noncoding region called the 5' untranslated region (UTR), the coding sequence (CDS) containing the information required to build the protein and delimited by a start codon and a stop codon, then another noncoding region called the 3'UTR harbouring regulatory sequence elements, and finally a sequence of adenine bases at its 3' end called the polyA-tail [3].

The mRNA structure contains several level of information: first, the primary structure is the nucleotide sequence, which defines critical sequence elements such as the coding sequence and therefore the protein to build, and also some regulatory sequence elements where other molecules can bind. Second, the secondary structure is a two dimensional structure of the folded RNA, after pairing of neighbouring nucleotides, creating hairpins and stem-loops. Third, in a similar way, the tertiary structures can be defined as pairing of more distantly separated nucleotides of the RNA creating a three-dimensional structure [4]. The main function of RNA secondary and tertiary structures is the accessibility of sequence elements where proteins and other RNAs can bind, which can have an impact on both gene expression and function [4].

### 2.1.2 Biogenesis

Mature mRNAs are generated from DNA through several processing events that the molecule must go through to become functional and stable: transcription, 5'capping, splicing, polyadenylation, and export to the cytoplasm [5].

#### 2.1.2.1 Transcription

The transcription consists in copying a DNA sequence into a complementary RNA sequence, called precursor mRNA (pre-mRNA). This process starts by a step called pre-initiation, which happens on the DNA molecule a few base-pairs upstream of a protein coding gene at its promoter region. Specifically, transcriptional activators bind to the promoter and recruit chromatin-modifying factors to open the chromatin (DNA and its histones) and make the DNA region available for transcription [6]. Activators also recruit an enzyme called RNA polymerase II (RNAP) and some proteins to form the transcription machinery [6]. Then, RNAP creates an initiation bubble (initiation step) and starts the synthesis of the pre-mRNA. The 5' capping step (described below) happens, and RNAP can enter the elongation step where the RNA sequence is synthesised while recruiting factors for splicing and polyadenylation events (also described below) [6]. The last step called termination is the release of the RNA molecule.

#### 2.1.2.2 5' capping

The first event in mRNA processing is 5'-end capping, occurring when the first 25-30 nucleotides of the pre-mRNA have been transcribed [5]. It consists in adding a 5' cap structure at the 5' end of the precursor mRNA. This structure enables nuclear export, translation into protein by the ribosomes, increases splicing efficiency and protects from cleavage by enzymes like exonucleases [5]. After capping, the polymerase can continue the transcription of the rest of the pre-mRNA [5].

#### 2.1.2.3 Splicing

RNA splicing consists in removing introns from the pre-mRNA and in joining exons together to produce a mature mRNA [5]. This mechanism is achieved by the spliceosome, a complex of hundreds of small RNAs and proteins [1]. Introns should be removed at very precise positions to avoid shifting the reading frame of the mRNA which would result in completely different proteins [7]. Therefore introns' boundaries are well defined by sequence elements at 5' and 3' splicing sites, whereas other sequence elements within the intron are also involved in the splicing [8].

#### 2.1.2.4 Polyadenylation

The polyadenylation process consists in cleaving the 3' end of a pre-mRNA and synthesising a sequence of multiple adenosine bases (called the polyA tail) onto the upstream cleavage product [9]. This process occurs for all human mRNAs except replication-dependent histone mRNAs [10]. Polyadenylation cleavage sites are generally indicated by a polyadenylation signal (PAS), usually the canonical RNA sequence AAUAAA [9], but a few other hexamers can be used [11]. Cleavage sites can be found 10 to 30 nucleotides downstream of the signal [12] and a downstream sequence element (DSE) rich in GU nucleotides can be found 20 to 40 nucleotides downstream of the cleavage site [12]. Similarly, an upstream sequence element (USE) upstream of the PAS can contribute to the polyadenylation efficiency, particularly for weak signals [13]. Therefore, the mammalian sequence pattern for polyadenylation can be summarised as USE-AAUAAA-DSE [7]. Furthermore, non-canonical sites do not necessarily need a polyadenylation signal, as the GU-rich region can be sufficient [14].

The role of polyadenylation of mRNA is to enable nuclear export, to increase the mRNA stability in the cytoplasm and the translation efficiency [3]. The polyadenylation machinery involves several protein complexes: the cleavage and polyadenylation specificity factor (CPSF) which recognizes the PAS, and the cleavage stimulatory factor (CstF) which binds to the GU-rich region [9], and also two cleavage factors (CFIm and CFIIm) [5]). After the cleavage of pre-mRNA at the polyA site, the polyA polymerase adds the polyA tail to finalise the mature mRNA [7].

#### 2.1.2.5 Export

The mature mRNA is transported from the nucleus to the cytoplasm during a step called export. This process starts during transcription by assembling onto the pre-mRNA different proteins forming a complex called transcription and export (TREX) complex [15]. Together with the mRNA, the TREX results in an mRNA ribonucleoprotein (mRNP) complex and is transported to the cytoplasm through protein complexes which can cross the nuclear envelope: the nuclear pore complexes (NPCs) [16].

### 2.1.3 Alternative processing: mRNA isoforms

Messenger RNA variants encoded by the same gene are called mRNA isoforms. They can arise from several kinds of alternative processing: alternative transcription initiation, alternative splicing, alternative polyadenylation, RNA editing, and post-transcriptional modification [8]. In human, there are about ten times more mRNA isoforms than genes [17], which suggests that this is a way of generating complexity among RNAs [8]. But those mRNA isoforms can cause disease [18], for instance by influencing mRNA transport, localization or stability [3]. I will focus here on alternative splicing and alternative polyadenylation.

#### 2.1.3.1 Alternative splicing

Alternative splicing consists in splicing a gene in different ways to produce different mature mRNAs from the same DNA sequence [1]. More than 90% of human genes encounter alternative splicing events [19, 20]. It can happen by selecting different combinations of exons, but also different splice sites which changes exon lengths [1]. Some introns can also be kept in the mature mRNA: those are known as pseudoexons [1]. This mechanism can produce many different mRNA and protein isoforms with different functions [1]. Several factors can affect the choice of a particular splicing site: the site strength depends on *cis*-acting elements, where the splicing machinery binds [1], chromatin and histone modifications [21] and the transcription rate [22].

Abnormal splicing creating coding frameshifts can happen and may create a too early stop codon, also known as premature termination codon (PTC) [8], which results in truncated proteins and in triggering of quality control processes such as nonsense-mediated decay (NMD) pathways to avoid erroneous proteins [23]. Abnormal splicing can cause diseases [1] and affect drug response [24], and can be used as cancer biomarkers [24]. Since abnormal splicing can generate dangerous protein variants, the cell uses quality processes to avoid export of those mRNAs into the cytoplasm [5]. Furthermore, several methods have been investigated to correct wrongly spliced transcripts in disease: the use of antisense oligonucleotides complementary to a particular splicing element enables to skip an unwanted exon [25], and trans-splicing, which is splicing between two pre-mRNA transcripts, can be used to correct one mutated exon by a normal one [26].

#### 2.1.3.2 Alternative polyadenylation

Similarly to alternative splice sites, pre-mRNAs can have several polyadenylation signals and therefore cleavage sites [3]. This concept of multiple sites is known as alternative polyadenylation (APA), resulting in mature mRNAs with different 3'UTRs [3] and occurring in about 54% of human protein-coding genes [11]. Also, APA can affect stability, localization, transport and translation of the mRNA. For instance, it plays a regulatory role, since mRNAs with shorter 3'UTRs might lack regulatory elements often found along the 3'UTR, resulting in differentially regulated transcripts and proteins [9].

The choice of polyA site is tissue- and development-specific: some tissues are more likely to use proximal polyA sites, while others use the distal ones [9]. Specifically, proliferating cells, cancer cells and less differentiated cells preferably use proximal polyA sites [27, 28, 29], while non-proliferative tissues might use distal sites [8]. Interestingly, strong canonical polyA sites are often distally located, while weaker polyA sites are often more proximal [9]. Generally, the site selection depends on it strength (signal, USE and DSE) and on physiological conditions such as concentration of polyA factors [9], and is thought to be also regulated by epigenetic marks

[9]. Furthermore, APA is found deregulated in an increasing number of diseases [30], therefore a high coordination between polyadenylation and splicing is necessary to avoid selection of intronic PAS, which could result in truncated proteins [7].

### 2.1.4 Translation to protein

The translation into proteins consists in translating the nucleotide sequence of an mRNA into a chain of amino acids called protein. It happens in three steps: initiation, elongation and termination [31].

**Translational initiation** takes place at the 5' cap: the small ribosomal subunit (40S) and initiation factors form a complex and bind to the 5' cap [31]. One protein of this complex, the polyA-binding protein (PABP), binds to the polyA, putting the mRNA in a circle shape to maintain its stability during translation [32]. The complex then scans the mRNA, to look for the translation start codon (usually AUG) [31]. Some initiation factors are then released and the large ribosomal subunit (60S) is recruited to form the ribosome complex (80S), which is the main actor of the translation [31].

**Elongation** starts when the ribosome (80S) reads the mRNA sequence by triplets called codons and uses Transfer RNAs (tRNA), which are molecules that associate a codon to an amino acid, as described by the genetic code (64 possible codons map to 20 amino acids), to produce an amino acid sequence until it reads the stop codon, indicating the end of the protein sequence.

**Termination** consists in releasing the new protein and the mRNA from the ribosome [31].

## 2.2 Non-coding genes

Genes do not necessarily encode for a protein, but are also functional as RNA transcript. Those are called non-coding genes. Several types of non-coding genes exist, like for instance microRNAs, piRNAs and lncRNAs. Those three non-coding RNA classes have in common to guide RNA-binding proteins (RNABPs) to a specific target nucleotidic molecule, to achieve a specific function.

MicroRNAs (miRNAs) are small non-coding RNAs of about 22 nucleotides that bind to mRNAs to inhibit translation. Piwi-interacting RNAs (piRNAs) are small non-coding RNAs of 24-32 nucleotides which are thought to play a role in germline development and gene regulation, particularly silencing of transposons [33]. Long non-coding RNAs (lncRNAs) are non-protein coding RNAs longer than 200 nucleotides, which play a role in epigenetic, splicing, transcription, translation apoptosis, cell cycle, imprinting and differentiation [34]. In this section, I shall focus on miRNAs, their biogenesis, their binding to mRNAs, and their involvement in diseases.

### 2.2.1 MicroRNA

MicroRNAs (miRNAs) are abundant small endogenous single-stranded non-coding RNAs (ncRNA) of about 22 nucleotides, which inhibit gene expression mostly by binding to 3' UTR of target mRNAs [35]. It is estimated that at least 60% of coding genes are repressed by miRNAs in humans [36]. Since they are expressed differently in each tissue, resulting in a tissue-specific gene regulation [37], they regulate developmental and physiological processes such as differentiation, growth, and apoptosis [38]. In arrested cells, miRNAs are thought to activate translation [39]. Furthermore, they also have been reported to stimulate translation by binding to 5'UTR during amino acid starvation [40].

### 2.2.2 Biogenesis

MiRNAs can be generated through several types of biogenesis: a canonical one, and alternative biogeneses.

#### 2.2.2.1 Canonical biogenesis

Similarly to mRNAs, the biogenesis of miRNAs consists of several processing events resulting in a mature miRNA: transcription and formation of the hairpin structure, different kinds of cleavage of that structure, export to the cytoplasm, and loading into the silencing complex. All known miRNA genes, their hairpin structure and their mature form are annotated in the MirBase database [41].

**Transcription:** In the nucleus, at a miRNA gene locus, RNA Polymerase II synthesises a primary miRNA (pri-miRNA) transcript containing a 5' cap, splicing events and a polyA tail like mRNAs [42]. Pri-miRNA can also be transcribed by RNA Polymerase III [43]. The pri-miRNA typically contains one or several hairpin structures, each of them having a hairpin stem, a terminal loop and single-stranded regions up- and down-stream of the hairpin [44]. Several mature miRNAs can cluster on the same miRNA gene and therefore share similar expression patterns [45].

**Cleavage by microprocessor:** The pri-miRNA is cleaved by a complex called microprocessor (a dimer of the Drosha enzyme and the DGCR8 protein) at the base of the pri-miRNA hairpin and results in a precursor-miRNA (pre-miRNA) of about 60 nucleotides, only consisting of the hairpin stem and loop [46]. The Drosha enzyme accomplishes the cleavage [47], while the DGCR8 protein recognises the single-stranded RNA/double-stranded RNA junction of the hairpin to have the correct cleavage site [48].

**Cleavage by Dicer:** The pre-miRNA is exported from the nucleus to the cytoplasm by the Exportin-5 (XPO5) protein [44]. In the cytoplasm, the enzyme Dicer cleaves the pre-mRNA's terminal loop to result in an imperfect miRNA:miRNA* duplex of ∼22-nucleotide length [44].

**Loading into Argonaute:** The miRNA:miRNA* duplex is split into two separated strands after the cleavage by Dicer: the functional one is loaded into an Argonaute (Ago) protein (Ago1-4) from a complex performing gene silencing, called the RNA-induced silencing complex (RISC), while the non-functional strand (often denoted miRNA*) is degraded [44]. In the cytoplasm, the Ago protein protects the mature miRNA from degradation [44].

#### 2.2.2.2 Alternative biogenesis

MicroRNAs can be generated by alternative processing, independent of either Drosha or Dicer [49]. One example of Drosha-independent biogenesis is the Mirtron pathway, where miRNAs are hosted in introns of mRNAs [50]. During splicing of the pre-mRNA, introns are cleaved and some of them resemble pre-miRNA hairpins and can be processed by Dicer [51]. Splicing is thought to replace cleavage by microprocessor [51], and the expression of miRNAs from introns is often correlated with host gene expression level [52]. One example of Dicer-independent biogenesis is the slicing of the pri-miRNA by Ago2, which is the unique Ago protein with slicing abilities [49]. It results in a functional mature miRNA loaded into Ago.

Similarly to mRNAs, miRNAs have isoforms as well, which are called isomiR. An isomiR is a variation of a mature miRNA and comes from the same pre-miRNA, but has either 5'-, 3'-trimming or nucleotide substitutions or additions [53]. IsomiRs are generally less expressed than their corresponding canonical mature sequence.

### 2.2.3 Targeting

After the miRNA has been processed, it becomes functional. The role of the mature miRNA is to recognise the target mRNA and to guide the RISC to it, to achieve gene silencing [44]. The miRNA possesses at its 5'end a region called seed sequence, which can detect its target mRNA by binding to a partially complementary sequence known as seed site, generally located in the 3' UTR of the mRNA [54].

#### 2.2.3.1 Silencing

MicroRNAs bring RISC to the target mRNA to inhibit protein synthesis in several ways [31]. The first is translation repression: RISC inhibits the process of translation, either at the translation initiation step, by interfering with the cap recognition process and the ribosomal subunits, or at the elongation step by inhibiting ribosome elongation and inducing ribosome drop-off [31]. The second way of inhibiting protein level is transcript decay: RISC recruits a deadenylase complex to induce shortening of the polyA tail (deadenylation) or a decapping complex to remove the 5' cap, both leading to degradation of the transcript [31]. Another miRNA-mediated transcript decay is cleavage, which is rare in animals and more common in plants [55]. Cleavage requires perfect complementarity between the miRNA and mRNA, but animal miRNAs often have mismatches and bulges preventing cleavage [31].

#### 2.2.3.2 Seed sites

The recognition of the target to silence is made by Watson-Crick base-pairing of the 2nd to 7th first nucleotides of 5' end of the miRNA (called seed sequence), which is generally perfectly complementary to the mRNA [54].

Several stringent seed types exist; from the most efficient to the less efficient, these are 8mer, 7mer-m8, 7mer-A1 and 6mer [56]. The 8mer has perfect nucleotide match between the 2nd and 8th nucleotides of the 5' end of the miRNA and has an adenosine base at position 1 of the target site on the mRNA. The 7mer-m8 is like the 8mer but without the adenosine at position 1. The 7mer-A1 is like an 8mer but without the base match position 8 and the 6mer is like a 7mer-A1 without the adenosine base at position 1 [56]. Furthermore, an adenosine at position 1, and a uracil or an adenosine at position 9 can enhance target recognition without particular base-pairs [31]. There are also several moderately stringent seed types, defined by one G:U pairing, or one bulged nucleotide or one loop [57].

Pairing of 3' end of miRNA can improve the targeting recognition [56] but this pairing is estimated to happen for less than 10% [54]. Stringent seed sites that show 3 or 4 base pairings at positions 13-16 are called 3'-supplementary sites, and moderately stringent sites with 4 or 5 base pairings at positions 13 to 19 are called 3'-compensatory sites as they compensate their weak seed with an additional pairing [54]. Finally, centred sites are miRNA target sites which do not have canonical seed sites and 3' compensatory pairing, but have 11 to 12 contiguous base-pairing from position 4 to 15 [58].

#### 2.2.3.3 Target sites

The seed sequences are used by Ago proteins to find the target sites on target mRNAs. Functional target sites downregulating gene expression are mostly found in 3'UTRs and less frequently in 5'UTRs and coding regions [31]. It has been shown that multiple target sites at optimal distances between each other can act in synergy [59], and that a lot of target sites are conserved among species, particularly at the region matching the miRNA seed [36]. Target sites in the coding region are not as efficient as those in the 3'UTR, unless they are preceded by rare codons upstream which slow down the translation rate and enable miRNAs to bind in a more efficient way [60].

### 2.2.4 Disease

MicroRNAs do not only regulate physiological processes, they are involved in diseases such as for instance cancer [61] and neurodegenerative diseases [62]. This can happen by dysregulation of mature miRNAs, generally through the disruption of miRNA biogenesis, but miRNAs can also be involved in disease if their target site is disrupted, for instance by a mutation.

**Dysregulation:** MicroRNAs are generally downregulated in tumour tissues compared to normal [37]. Consequently, miRNA profile can be used as a biomarker to subclassify tumour tissues [37], and miRNA signatures are important information for disease diagnosis, progression, prognosis, and treatment response [63]. Furthermore, miRNA signature correction is a potential therapeutic approach [64], but still raises challenges for delivering miRNAs.

**MicroRNA biogenesis disruption:** Each step of the miRNA biogenesis can be disrupted and associated with disease. For instance, Drosha can be upregulated in some cancers, resulting in processing more pri-miRNAs, and a global over-expression of miRNAs [65]. Drosha has also been shown to malfunction in primary tumours, resulting in accumulation of unprocessed pri-miRNAs [66]. In several cancer cell lines, an important number of pre-miRNAs are kept inside the nucleus, suggesting dysfunction in the export process [67] and could explain the global downregulation of mature miRNAs.

**MicroRNA targeting disruption:** Target sites can also be altered by changes in accessibility or by polymorphisms and mutations. Target genes become then dysregulated and can be implicated in tumourigenesis if they are oncogenes or tumour suppressor genes [68]. The following section will describe polymorphisms in detail.

## 2.3 Polymorphisms

Polymorphisms are DNA sequence variations that encompass several types, such as single-nucleotide polymorphism (SNP) and structural variants. A SNP is a variant of one nucleotide change, commonly used to analyse heritable DNA diseases. Other variants are called structural variants and encompass insertions/deletions (indels), block substitutions, inversions or copy number variants (CNV) and are estimated to account for 20% of human polymorphisms and 1% of the genome bases [69]. Furthermore, a CNV is defined as a sequence that is repeated several times, the number of copies changing from one individual to another [69]. Other polymorphisms exist such as microsatellite, also known as simple sequence repeat (SSR) or short tandem repeat (STR), which is a DNA sequence that consists of a motif long of two to six nucleotides and is repeated several times [70]. Similarly to SNPs, microsatellites are also used in disease analyses, but here I shall focus in SNPs and their effects on diseases.

### 2.3.1 Single nucleotide polymorphisms

A single-nucleotide polymorphism (SNP) is a change of one single nucleotide in the DNA sequence [69]. Each form the variant can take is called allele and the majority of SNPs are diallelic, which means they have two possible alleles in a population [71]. SNPs are the most common kind of DNA variants in humans [69] and millions of them are annotated in the dbSNP database [72].

An important characteristic of a SNP is the allele frequency in a given population; *i.e.* how frequent each allele occurs. The major and minor alleles are respectively defined as the most and least common alleles [73]. However, people generally only refer to the minor allele frequency (MAF), since the major allele frequency can be deducted from the MAF. For most diallelic variants, the MAF is used to distinguish between SNPs and rarer variants: diallelic variants require a MAF greater than 1% in a population to be termed as SNPs [73]. The ones with lower MAF are termed rare variants.

Other characteristics of SNPs are the combination of parental alleles (genotype), the combination of neighbouring alleles along the chromosome (haplotype), and the allele correlation between SNPs (linkage disequilibrium).

#### 2.3.1.1 Genotype

Humans are diploid, which means their cells contain two versions of each autosomal chromosome, and there is therefore a combination of two alleles at each SNP: one from the father and one from the mother. This combination of alleles is called genotype and each diallelic SNP harbours one of the three possible genotypes: either homozygous for the major allele (twice the major allele), heterozygous (both major and minor alleles) or homozygous for the minor allele (twice the minor allele) [74].

**Hardy-Weinberg equilibrium** (HWE) is a principle stating that the genotype frequency of an autosomal variant stays constant from one generation to the next one, assuming random mating [71]. The two alleles $A$ and $a$ of a variant with respective frequencies $p$ and $q = 1 - p$ in a population, are expected to result by random mating in the genotypes $AA$, $Aa$ and $aa$, with the respective frequencies $p^2$, $2pq$ and $q^2$ in the next generation [71]. Departure from HWE can happen by inbreeding, mutation, and natural or artificial selection and can be tested using Pearson's $\chi^2$-test by comparing the observed genotype counts and the expected ones under HWE based on allele counts [71].

**Dominance and recessiveness:** A genotype can be responsible for a simple trait. A trait is defined as dominant if the trait allele is stronger than the other allele, which means that each person having the trait allele (heterozygous or homozygous for the trait allele) harbours the trait. In contrast, a trait is recessive if a person needs both trait alleles to harbour the trait (only homozygous for the trait allele).

#### 2.3.1.2 Haplotype

In contrast to genotype which is a combination of two alleles at one position on the chromosome pair, a haplotype is a combination of alleles occurring at different positions on the same chromosome [73]. For each chromosome pair, each human inherit one haplotype from the father and one from the mother. During the formation of gametes (meiosis), chromosome pairs can cross over, resulting in new combinations (recombination) of alleles along each chromosome, and therefore new haplotypes [74].

Recombination events between two markers are studied within family pedigrees by analysing the recombination fraction ($\theta = \frac{r}{n}$; for $r$ recombinant among $n$ offspring, by computing directly $\theta$ in case of phased markers, or by estimating it). It enables to quantify the genetic linkage (loci inherited together; linked) as the logarithm of the odds (LOD) score, which is the log ratio between the likelihood of linkage for a given recombination fraction ($\theta < 0.5$) and the likelihood of no linkage ($\theta = 0.5$): $\text{LOD}(\theta) = \log_{10}\left(\frac{L(\theta)}{L(\theta=0.5)}\right)$ [75]. The most likely genetic distance between the markers is given by the maximum likelihood estimate of $\theta$; the $\theta$ that gives the highest LOD score [75]. Scores greater than 3 are generally seen as evidence for linkage, while scores lower than $-2$ are evidence for independence of the markers [75].

Some DNA regions harbour a high density of recombination events and are known as recombination hotspots, while regions with low density of recombination are inherited as blocks (Figure 2.1) through generations where variants inside the blocks are linked together [73]. Furthermore, the block structure of the genome has also been shaped by historical events reducing the population size, such as migration of a subpopulation, and high mortality rate.

#### 2.3.1.3 Linkage disequilibrium

Alleles of SNPs that are closely located are often correlated to each other, because the closer they are, the less likely a recombination event can happen between them [69]. This non-random association of alleles in a haplotype is called linkage disequilibrium (LD) [69]. While linkage equilibrium describes a situation where alleles of two variants occur in an independent way, LD describes the dependence between the alleles, *i.e.* some haplotypes occur more often in a population than expected by chance. Interestingly, LD decreases with space (genomic distance) and time (number of generations) because of recombination events between loci [74].

This correlation between variants can be measured in several ways: the most important ones are $D'$ and $r^2$ [69]. For example, two loci with the alleles A/a and B/b respectively, have $p_A$, $p_B$ and $p_{AB}$ the probabilities of allele A, allele B and haplotype AB. Then $D = p_{AB} - p_A p_B$ is the difference between the actual haplotype frequency and the expected one for independent loci [76]. $D'$ is the normalised measure of $D$: $D$ is divided by its theoretical maximum for the observed allele frequencies [77], and ranges between 0 and 1, where 0 means no LD and 1 means LD.

Another measure of LD is $r^2$, which is the square of the correlation coefficient between allele frequencies, or the percentage of variance at one SNP that can be explained by the other one. $r^2$ is given by $r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$ and ranges between 0 and 1 [78]. The main difference between $r^2$ and $D'$ is that $r^2$ contains allele fre-

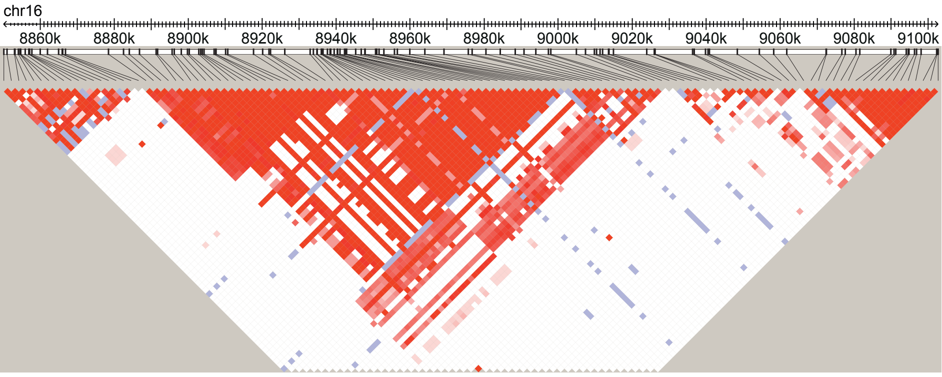


Figure 2.1: The block structure of a DNA region. The SNPs are shown on the horizontal axis, and the colours show linkage disequilibrium (LD) between pairs of SNPs: red is high LD, and white is low LD. Blocks are regions with low recombination rate, and appear as red triangles.

quency information and therefore rapidly decreases with low MAF [76]. In high LD regions, such as haplotype blocks, LD can be used to predict the allele at one SNP knowing the allele at another SNP [73]. The Haplotype Map (Hapmap) database provides haplotype and LD data for 3.1 million SNPs from different populations (European, African, and Asian populations) [79].

### 2.3.2 Effects of DNA variants

DNA variants, such as SNPs, are called functional when they perturb functional elements within the cell. I will first describe the effects of variants within coding regions, and then those within non-coding regions.

#### 2.3.2.1 Coding variants

Coding variants are variants within the coding region of mRNAs, and there are two different types: the ones that change the amino-acid (non-synonymous) and those that do not (synonymous).

**Synonymous variants** are variants in mRNA coding regions that do not change the amino acid in the resulting protein [80]. Specifically, synonymous variants of one nucleotide change are called synonymous SNPs (sSNPs). For a long time, sSNPs have been thought to be silent, but it is now known that they can affect protein expression, structure and function and that they can play a role in disease [80]. Synonymous SNPs can cause disease through several ways. By disrupting or creating exonic sequence elements involved in mRNA splicing, they can result in aberrant splicing and disease [81, 1]. Similarly, by disrupting miRNA target sites in coding regions, they can result in change of mRNA stability, protein expression and possibly disease [82]. Furthermore, by changing mRNA structure and protein folding, they can affect transcript stability [80]. Finally, by switching between a rare and a frequent codon, and affecting translation rate and space between translating ribosomes, they can result in protein misfolding, or ribosome blockage and translation abortion [80].

**Non-synonymous variants** (nsSNP) are SNPs in a coding region that change the amino acid sequence, resulting in protein isoforms (missense mutations) or truncated proteins when they change into a stop codon (nonsense mutations) [80]. Changes in proteins due to nsSNPs can affect protein function and cause disease, particularly single-gene disorders [73].

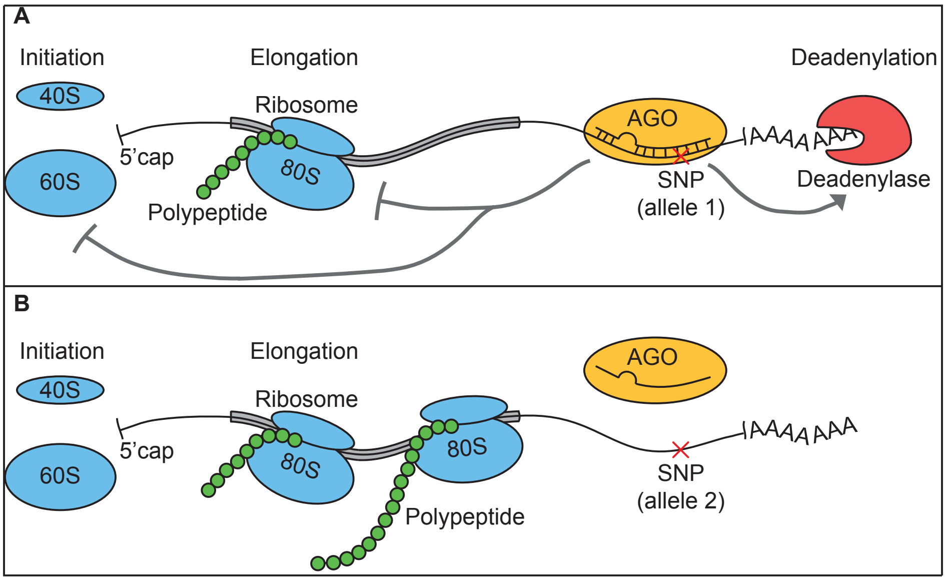


Figure 2.2: A SNP in a miRNA target site affects gene expression of an mRNA and its protein. (A) Allele 1 of the SNP makes the target site complementary to the miRNA, which can bind to it with the silencing machinery illustrated by the Argonaute (AGO) protein. The silencing machinery inhibits protein translation by either disrupting the translation initiation of the ribosome, the elongation of the protein sequence, or affects mRNA expression by deadenylation. (B) Allele 2 disrupts the target site. The miRNA and the silencing machinery cannot bind to downregulate gene expression.

#### 2.3.2.2 Non-coding variants

Variants do not necessarily occur in coding regions, but also in non-coding regions such as non-coding RNAs, but also 5' UTRs, 3' UTRs, introns and promoter regions. Several variants in those regions have been associated with diseases, however the mechanisms that those polymorphisms affect is less clear than for nsSNPs. Nevertheless, polymorphisms in 5' and 3' UTRs can alter the mRNA structure and have been associated with diseases [83].

**Variants affecting microRNAs** (called miRSNPs), can impact gene expression, disease risk, treatment, and prognosis in several ways. First, SNPs within miRNA target sites (Figure 2.2) can increase or decrease affinity between miRNAs and their targets, disrupting or creating new sites on the target mRNAs, possibly affecting mRNA transcript and protein expression levels [84]. This kind of variant has been associated with risk for several different diseases, such as cancer and Parkinson's disease [85]. Seed-complementary sequence regions of target sites conserved between species harbour a lower polymorphism density possibly due to negative selection [86]. Second, miRNA genes have a low density of polymorphisms and particularly the seed region [87], because variation in those sequences would have a strong impact on miRNA expression and function, gene expression, and phenotype [84]. However, there are a few SNPs in miRNA genes, and they can disrupt miRNA function by affecting the pre-miRNA hairpin pairing, resulting in a different mature miRNA, or by affecting the binding to target mRNAs, and have been associated with increased cancer risk [84]. Third, several variants in the miRNA machinery proteins, such as Drosha, DGCR8, XPO5, Dicer and AGO, affecting either the function or expression of those proteins, have been identified as non-synonymous, or potentially affecting splicing, or causing frameshift [88], and some have been associated with diseases [89, 90].

**Variant affecting polyadenylation** are variants in 3'UTR that can deregulate polyadenylation by affecting sequence elements where the polyadenylation machinery can bind [30]. By creating new sequence elements, those variants can result in polyadenylation stimulation upstream of the normal site, or disrupt normal sequence elements and postpone polyadenylation further downstream. Those mutation-caused APAs may result in miRNA dysregulation and be associated with diseases [30]. One particular sequence element that can be subject to mutation is the polyA signal (Figure 2.3). A variant can change a canonical signal into a non-canonical signal, which is weaker, and which often requires USE and DSE to compensate the signal weakness [14]. Also mutations in GU-rich DSE can affect the polyadenylation process [91].

**Variant in lncRNAs** are also important because lncRNAs have sequence elements that can bind to DNA, RNA or protein, and affect primary and secondary structures, function and expression level. Those variants have been strongly associated with diseases such as cancer and neurodegenerative diseases [34].

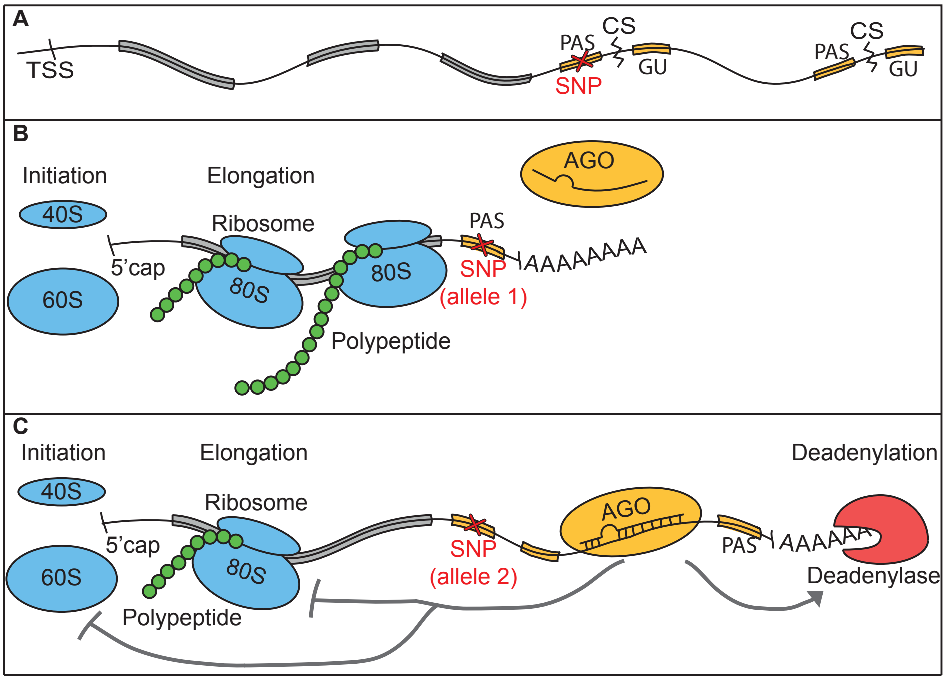


Figure 2.3: A SNP creates an alternative polyadenylation signal and affects gene expression. (A) A DNA region harbours a gene, here illustrated by its transcription start site (TSS), its coding region in grey, and two polyadenylation sites shown by polyadenylation signals (PAS), cleavage sites (CS), and GU-rich regions. A SNP lies in the first PAS. (B) Allele 1 makes the first polyA site functional, resulting in a short 3'UTR. (C) Allele 2 makes the first PAS non-functional, resulting in cleavage at the second polyA site and a long 3'UTR, which contains a miRNA target site. The miRNA machinery binds to the mRNA and downregulates gene expression, through translation inhibition or deadenylation.

# Chapter 3

# Technologies and strategies

To analyse polymorphisms and their effects on RNA sequences and expression levels, several technologies, such as microarray and sequencing technologies, have been developed. Furthermore, those technologies can be involved in several strategies to analyse the effects of polymorphisms on a trait such as a genetic disorder: linkage studies, genome-wide association studies and exome sequencing studies. Here, I shall first describe the technologies' principles, advantages and limitations, before explaining the different strategies.

## 3.1 Technologies in genetics

Genetics involves several important processes, such as identifying the genotype of an individual at a polymorphic site (genotyping), quantifying the expression levels of mRNAs and non-coding RNAs in a sample (expression quantification) and identifying the DNA/RNA sequences in a sample (sequencing). Several technologies have been developed to achieve these processes. Here, I am going to talk about microarrays and RNA-seq technologies.

### 3.1.1 Microarrays

A microarray contains target-specific DNA probes and uses hybridisation to those probes to catch single stranded DNA fragments of interest that are complementary to those probes [92]. By using different DNA probes, they can genotype SNPs and quantify transcript expression levels. One limitation is that they cannot measure unknown targets: they require prior knowledge of the target to design its probe [92]. Also, hybridisation errors can occur when probes bind to molecules that are similar to their target [92]. Nevertheless, quality control standards have been developed to reduce biases [93]. In this section, I shall describe how microarrays can be used for SNP genotyping and RNA expression level quantification.

#### 3.1.1.1 SNP genotyping

Microarrays can be used to genotype SNPs anywhere in the DNA as long as the flanking sequences are known. Microarrays, such as Illumina's Infinium Beadchips, Affymetrix GeneChip Human Mapping arrays, Invader or Perlegen, are commonly used to genotype SNPs genome-wide and can analyse from 10 thousand to 2 million SNP assays in parallel with high accuracy [94].

**Principle:** Each technology either hybridises a single stranded DNA sequence consisting of the target SNP and its flanking regions to allele-specific probes, or hybridises the 5' flanking region to a primer and extends the primer with the nucleotide that is complementary to the allele (single-base extension or SBE). Those probes can be coupled with fluorescent labelling specific to each allele, whose intensities can then be measured [94], but other labelling methods exist. Then, genotypes are statistically estimated based on those signal intensities. Furthermore, microarray genotyping technologies can be customised, generally for replication and validation studies, to analyse a smaller amount of SNPs in a high number of samples with high accuracy [94].

**Limitations:** SNP arrays enable to genotype only known SNPs and genome-wide arrays provide limited customisation [94].

#### 3.1.1.2 RNA expression

Transcript expression microarrays were the first technology to enable transcriptome-wide expression analyses in many different cell types, differentiation states and diseases [95]. Many of these experiments generated expression results that are archived within the Gene Expression Omnibus (GEO) database [96].

**Principle:** Similarly to SNP arrays that use probes based on sequences that flank the SNP of interest, mRNA expression arrays use probes based on gene complementary sequences [95]. Probes are based on exonic sequences from mRNAs, or mature sequences from small RNAs, such as miRNAs. Fluorescent labelled RNAs hybridise to their respective probes and light intensities are measured and correspond to gene expressions [95]. Expression microarrays can be used to measure expression of mRNA isoforms such as alternatively spliced mRNAs, by designing probes that target exon junctions, to measure the expression of that particular exon combination [97].

**Limitations:** Expression microarrays can quantify transcript expression of annotated genes, but cannot detect unknown expressed transcripts, unknown exon junctions (alternative splicing) or unknown poly(A) sites (alternative polyadenylation). Furthermore, isoforms disrupting the matching to the probe [3] and noise from hybridisation signal (cross-hybridization) [8] may affect the resulting expression data. Also, microarrays cannot provide good sensitivity for low and high gene expression levels when looking at differential expression [98]. Finally, allele-specific expression quantification is possible but quite limited compared to RNA-seq approaches.

### 3.1.2 RNA-seq

RNA-seq is a next-generation sequencing (NGS) technology that aims at sequencing the whole transcriptome profile and at addressing microarray limitations [8], such as measuring unknown transcripts. This high-throughput technology is also known as whole transcriptome shotgun sequencing (sequencing of small fragments). Like microarrays, it can be used for quantifying RNA expression level, SNP genotyping and identifying any exon junctions, but also allele-specific expression levels and polymorphism detection (variant calling). In this section, I shall describe RNA sequencing, SNP genotyping, and quantification of RNA expression levels.

#### 3.1.2.1 Sequencing

Sequencing consists in extracting the RNA, and breaking it in small fragments that are then sequenced. Once sequenced, those fragments, called reads, are then mapped to reference genomes (Figure 3.1A,B) such as the human reference genome to identify expressed regions [99].

**Principle:** RNAs are digested into small fragments (RNA fragmentation) which are converted into complementary DNA (cDNA) fragments and repeatedly sequenced in a massively parallel way and in a short time. Alternatively, fragmentation can occur after cDNA synthesis (cDNA fragmentation). Sequencing in itself consists in reading a lot of single stranded DNA fragments simultaneously by generating their complement strand one nucleotide after the other by a DNA polymerase (sequencing by synthesis; *i.e.* Illumina Genome Analyzer), or all consecutive identical nucleotides at a time (sequencing by synthesis pyrosequencing; *i.e.* Roche 454 Life Science), or with oligonucleotide fragments one after the other by a DNA ligase (sequencing by ligation; *i.e.* Applied Biosystems SOLiD), each method having fluorescent labelling [100]. Those are the current high-throughput sequencing methods, but new ones are emerging as well. The sequence of fluorescent intensities enables to identify the RNA sequence fragment, called the read, and to estimate uncertainty of each base (probability of wrong base) [100]. That information is stored in a FASTQ file [101],

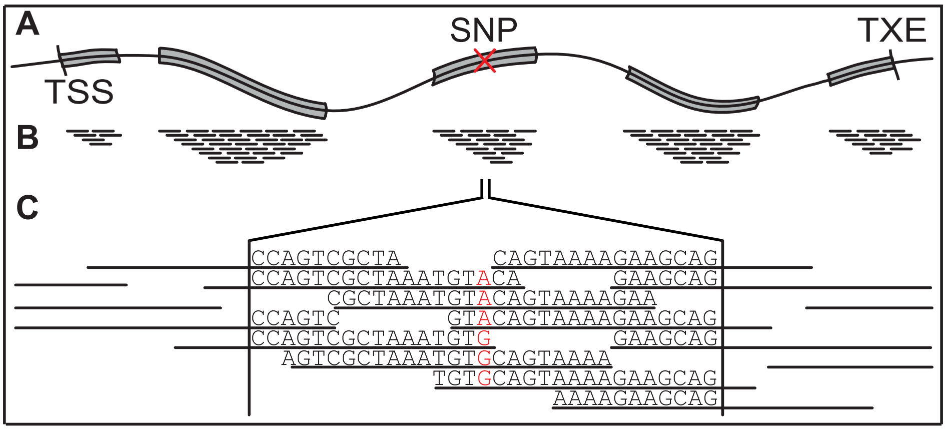


Figure 3.1: RNA-sequencing. (A) A gene is shown on a DNA strand, from its transcription start site (TSS) to its transcription end (TXE); the exons are shown in grey and one of them contains a SNP. (B) Sequenced reads are aligned against the gene, showing where exons are expressed as they come from RNA. The reads can be used to estimate transcript expressions. (C) A zoom at the SNP locus shows the mapped reads, and the nucleotides at the SNP position are A and G with equal proportions, suggesting that this SNP is heterozygous.

where the uncertainty is converted into a quality scores $Q_{phred} = -10 \log_{10} P(error)$ and mapped to an ASCII character, resulting in a quality sequence.

**Reads** are sequence fragments and each experiment generates millions of them. The read length ranges from 30 to 400 nucleotides according to the sequencing method used [98]. Sequencing from one end of the fragment (respectively both ends) generates single-end (respectively paired-end) reads [98]. Therefore paired-end reads correspond to sequences of both 5' and 3' ends of the DNA fragment, which may be separated by an unsequenced gap [98]. Depending on the fragment size, those read pairs are located more or less distantly on the original RNA molecule. The short reads resulting from sequencing can then be aligned to the human reference genome [100].

**Read mapping** consists in aligning millions of reads against the reference genome, to know from which transcripts those fragments came from. The mapping process can take into account polymorphisms or sequencing errors, by using base quality scores and allowing a few mismatches, insertions, or deletions between the read and the reference sequence [92]. However, since the mapping process can make mistakes as reads may map ambiguously to different loci, discarding those reads is a way to reduce mapping errors. Furthermore, paired-end reads contain more information than single reads, and can therefore increase alignment accuracy [100].

**Advantages:** RNA-seq can identify new transcripts and isoforms in a high-throughput way with a single base resolution, while requiring a low amount of RNA. Such high resolution maps have improved the annotation of gene boundaries (reads showing transition between UTRs and polyA/polyT tails), exon junctions (reads containing splice site motif and mapping the two flanking sequences to different exons), introns (showing low expression compared to exons), and RNA editing events [98, 95].

**Limitations:** RNA-seq can have some limitations in sequencing. For example, the pyrosequencing method adds one type of nucleotide at a time (either A, C, G or T) and measures the signal intensity to identify the number of consecutive identical nucleotides that have been added by the DNA polymerase. This method can have problems to estimate the precise number of consecutive identical nucleotides the higher it gets, particularly with homopolymeric sequences, resulting in false positive insertion or deletion and mapping problems [92]. Also, methods that read one base at a time can address that issue, but they usually provide lower quality at the 3' end of the read [92], because of asynchrony in the sequencing cycles [100]. Also it is not always easy to map reads back to the genome, because some reads may come from several potential locations. Finally, RNA-seq produces much more data than microarrays, which raises storage and computer processing problems [98].

#### 3.1.2.2 SNP genotyping

Once the reads have been aligned to the reference genome, RNA-seq can be used to genotype SNPs in exons [98] and to compute allele-specific expression .

**Principle:** Genotype calling consists in estimating the genotype of an individual at one known polymorphic site [100]. With RNA-seq data, this is based on read counts of the alleles (Figure 3.1C) and their base quality scores, by counting only high quality bases (base accuracy $\geq 0.99$) and then determining the proportions of the two alleles [100]. A simple threshold rule can be used to infer genotypes: for instance, both allelic proportions greater than 0.15 classifies the site as heterozygous, and otherwise homozygous to the allele with highest proportion.

**Limitations:** This genotyping procedure on RNA-seq data can only work on exonic variants from expressed genes. It can be affected by several kinds of errors, such as sequencing and mapping errors [100]. Also low read depth can result in only one chromosome sequenced from the chromosome pair and increase the number of heterozygotes wrongly classified as homozygous [100]. Therefore high coverage sequencing as well as focusing on highly expressed loci can reduce uncertainty. Alternatively, computing genotype likelihood (as described in Chapter 4) can reduce and quantify uncertainty, based on sequencing and mapping errors, allele or genotype frequencies and LD [100]. The genotype with the highest likelihood is chosen and this value provides a measure of confidence that can be used in downstream analyses such as association tests [100].

**Allelic expression:** Genotyping a SNP with RNA-seq data involves computing allele-specific expression [95]. At heterozygous loci, the expression of both alleles is in general thought to be the same for autosomal chromosomes, and any deviation from that equilibrium (allelic imbalance) can be of interest, because it can for instance mean that the two gene copies are regulated differentially. Allelic imbalance can be measured by either the proportion of alleles, their ratio or their log ratio, and similarly to the genotyping method, it can be affected by sequencing and mapping errors, but also by bias towards the allele in the reference genome (reference allele) if the mapping method was carried out without the SNP information.

#### 3.1.2.3 RNA expression

In a similar way to allelic expression, RNA-seq data can be used to estimate expression of mRNAs and non-coding RNAs, by counting reads mapped to the gene region (Figure 3.1B).

**Principle:** The read distribution across a gene shows the different exons. Using RNA fragmentation gives a more uniform expression distribution in the coding region, but less coverage at both ends, therefore each exon expression level can be estimated by the number of reads mapped divided by the exon length [98]. In contrast, using cDNA fragmentation tends to give a biased distribution towards the 3' end, therefore the expression level is estimated by counting reads in a window near the 3' end [98].

**Advantages:** RNA-seq can clearly identify gene boundaries and exon inclusion and junction and therefore quantify mRNA isoforms without any prior knowledge about the existence of any particular isoform, in contrast to microarrays [95]. Also, sequencing can discover miRNA isoforms, new miRNAs and classes of non-coding RNAs [92]. Furthermore, mapping reads to previously unannotated regions may suggest the existence of unknown genes, in contrast to exon tiling microarrays which require annotation. Transcript expression is more precise with RNA-seq than microarrays and correlates with traditional quantification methods like quantitative polymerase chain reaction (qPCR) [98]. Finally, RNA-seq has a lower noise, no upper limit of expression level and high levels of biological and technical reproducibility [98].

**Limitations:** Bias in expression levels can arise from several sources: the fragmentation bias which produces non-uniform read distribution over a gene and can affect the final expression level [98], and the sequencing bias which gives to RNA fragments a non-uniform chance to get sequenced according to their motifs [92]. Like microarrays, RNA-seq is limited for the quantification of rare transcripts [95].

## 3.2 Trait-locus association strategies

The preferred strategy to identify an association between a trait with one or several genetic causes depends on many factors, such as the expected frequencies of the genetic variants, the probability of having the trait given the trait genotype (penetrance). Here, I first define special types of traits that are genetic disorders, particularly complex diseases, and their aetiology, before detailing several strategies to identify causal variants.

### 3.2.1 Traits and aetiology

Traits may be any phenotypical features, but here I shall focus on genetic disorders and their aetiology.

#### 3.2.1.1 Genetic disorders

There are two main types of genetic disorders: Mendelian diseases and complex diseases.

**Mendelian diseases** are single gene disorders that follow Mendel's law of inheritance. They are caused by one single variant and often cluster in families [74]. More than 1500 genes involved in rare Mendelian disorders have been identified [94].

**Complex diseases** are disorders that do not follow Mendelian inheritance but that can combine multiple genetic and nongenetic causes with small contributions each [74]. This term encompasses most of the heritable diseases.

#### 3.2.1.2 Aetiology of complex diseases

Studying the cause of complex diseases consists in identifying the causes of complex phenotypic disorders as well as their mechanisms to improve diagnostics and treatments, but also in cataloguing their risk factors for prevention purpose [74]. Risks factors affecting such phenotypes can be genetic and environmental, and the phenotypic variance depends on the genetic and environmental variances and covariance. Estimating heritability of a complex disease consists in separating the phenotypic variance into the genetic and environmental components. Once the two components have been separated, candidate risk factors can be analysed to try to identify those that can partly explain each component variance.

**Genetic component** , also called heritability, is the proportion of phenotypic variance that the genetic variance can explain [74]. Estimating heritability is usually done by studying monozygotic (identical) twin pairs and comparing their phenotypic concordance with non-identical sibling pairs or closely related pairs. It is because diseases with a genetic component are more likely to co-occur in a group of related people than in a group of unrelated ones [74]. Heritability includes all the genetic risk variants, ranging from rare to common variants with high or low penetrance: multiple risk variants can affect the phenotype independently of each other, or epistatically (synergistic or antagonistic epistasis) [69].

**Environmental component:** Studying adopted children enable the estimation of the environmental component proportion, which includes environmental factors that the patients have been exposed to. Their measurements are generally less accurate than genetic factors, because they are often based on patients' recollection [74].

### 3.2.2 Strategies

The different strategies to identify DNA variants that contribute to common complex disease susceptibility are based on assumptions regarding allele frequencies and disease penetrance. Two main hypotheses have emerged: the common-disease common-variant (CDCV) hypothesis focuses on multiple common variants with low penetrance while the common-disease rare-variant (CDRV) hypothesis focuses on multiple rare variants with higher penetrance [102].

**The CDCV hypothesis** states that multiple common variants with small effects result in susceptibility to common complex diseases [69]. Common variants are generally defined as having a MAF greater than 5% in the studied population.

**The CDRV hypothesis** states that multiple rare variants with high penetrance result in susceptibility to common complex diseases [69]. Rare variants are generally defined as having a MAF from 0.1% or 1% and up to 5% in the studied population. The idea behind the CDRV hypothesis is that several individuals can have different variants affecting the same DNA region, resulting in the same disease (allelic heterogeneity).

The first step of genetic disorder analyses has been for a long time to identify broad genomic regions through linkage studies, with a follow-up inside those regions by candidate gene association studies. With the formulation of the CDCV hypothesis, genome-wide association study has become the strategy that has mainly been used to identify common variants associated with common complex diseases. Then, since the CDCV hypothesis could not explain an important part of heritability, the CDRV has been formulated and exome sequencing has now become a promising strategy to identify multiple rare variants in common complex diseases.

#### 3.2.2.1 Linkage analysis

Linkage analysis consists in analysing a population whose relatedness is known (such as a family pedigree) and which contains many cases of a particular genetic disease, to identify the DNA region responsible for that trait. It can be used for Mendelian diseases (model-based analysis) or complex disease (allele-sharing analysis) [75], and their family-based approach has been able to identify rare mutations with high penetrance [103]. However, gathering genotypes and pedigrees from many affected families takes time and is not easy [74].

**Model-based linkage analysis** consists in analysing recombination events within a pedigree, to identify genetic regions that are associated with a trait or a disease [75]. The analysis is based on a heredity model of the trait (dominant/recessive, autosomal/sex-linked, and penetrance) and a set of genetic markers through the whole genome [75]. Given a trait model, a pedigree with affected individuals and their genotypes, a two-point mapping is carried out between each marker and the disease, by computing LOD scores. The region that shows the less recombination with the disease locus can then be further analysed in detail by a multipoint mapping. This mapping is based on a set of markers and a linkage map, which shows the recombination between the markers (in terms of centimorgan (cM)). The multipoint mapping computes LOD scores based on multiple markers simultaneously by calculating the likelihood of the pedigree given the disease variant is lying within an interval of the linkage map. This method requires the knowledge of the true model of inheritance, which is possible only for simple Mendelian diseases. Also multiple models of inheritance of one unique disease (model heterogeneity) reduce the power of that method. Generally, for complex diseases, this method cannot be used because it is not possible to identify the model [75].

**Allele sharing analysis** is a model-free linkage analysis that can be used for complex diseases. It consists in analysing the proportions of allele-sharing between related individuals, such as affected sibling pairs. Allele-sharing is generally defined as Identical By Descent (IBD), which means the alleles of two relatives are the same and inherited from a common ancestor [104], in contrast with Identical By State (IBS) where the alleles are the same but not necessarily inherited from the same ancestor. At a locus the number of alleles that are IBD between an affected sibling pair can be 0, 1 or 2. Based on many sib-pairs, it is possible to estimate the proportions of 0 IBD allele, 1 IBD allele, and 2 IBD alleles. Those proportions can be compared to the expected IBD proportions under the null hypothesis that there is no linkage between the tested locus and the disease locus. A significant deviation from the expected proportions would be a sign of linkage between the locus and the disease locus, while not assuming any inheritance model of the disease [75]. Several statistics can be computed to test the significance, such as the goodness of fit or the mean number of IBD alleles, but those require known IBD status [105]. Otherwise, IBD can be estimated by maximum likelihood methods to compute a LOD score [75]. In case of analysing a continuous trait, quantitative trait loci (QTL) analysis can be achieved by linear regression of IBD on the trait value (or on the trait difference between relatives). Since the inheritance models of complex diseases are unknown, those model-free methods only have enough power to detect large regions of linkage, and can therefore be a first step in the genetic analysis of one disease [75].

#### 3.2.2.2 Candidate gene analysis

The candidate gene association study (CGAS) consists in analysing a candidate gene for association with a disease. It requires choosing a gene, often one that lies within a DNA region formerly identified by linkage studies [73]. Within the selected gene, a set of independent markers is chosen through SNP tagging methods, so that the markers are not in LD with each other, to avoid redundancy, which would reduce the power of detecting association [73]. Parts of the gene like the coding region or the promoter region can be prioritised in the selection of markers [74], because in case of association, their effect would be easier to interpret than in non-coding regions or introns. The selected set of SNPs is then genotyped among patients and healthy people, and their genotypes or alleles are tested for association with a disease. In case of a case-control study design, the frequencies of each genotype (or allele) are compared between the case and the control groups, using for example the $\chi^2$ test, to identify the genotypes (or alleles) that are significantly found more often among cases [73]. It results in a set of SNPs associated with the disease. Those SNPs can be disease-causative, partly contributing to the disease, or in LD with causative variants, in which case they are proxies for the real causative variants [73]. Haplotypes may also be tested when assuming the trait comes from a combination of variants [73].

The advantage of association studies over linkage studies is that they can identify smaller regions of association, and that they provide more power, particularly for common variants with low-penetrance (small effects) [74]. However, they are based on choosing a candidate gene [73], and a large sample size is needed to detect small effect variants [74]. Furthermore, CGASs could identify many variants associated with diseases, but most of them lack reproducibility [106], possibly because of heterogeneous populations (population stratification) that contain several subpopulations which have different allele frequencies.

#### 3.2.2.3 Genome-wide association study

A genome-wide association study (GWAS) or whole-genome association study (WGAS) is a type of association study based on the CDCV hypothesis. The CDCV-based strategies like GWASs have been the main focus of genetic epidemiology during the last few years for identifying the heritability of complex diseases, by analysing common SNPs for association with common diseases and identifying and replicating significant associations [69].

**Principles:** GWAS consists in genotyping hundreds of thousands of common markers that cover most of the genome for hundreds to thousands of cases and controls, which has become possible with high-throughput genotyping platforms. As for CGAS, genotypes or alleles of each marker are tested among the case and control groups, to identify associated markers that are correlated (in LD) with the susceptibility one [74]. Because many SNPs are tested in a GWAS, the chance that they appear significant just by chance is high. Therefore, to avoid many false-positive associations, p-values need to be corrected for multiple testing such as the Bonferroni approach [74] and associations need to be replicated in independent studies [107].

**Markers** (or tagSNPs) are variants selected to capture the association signals at particular loci. As all the SNPs cannot be tested, because they are too many, and because they are not independent, markers are tested as proxies for all the variants that are in LD with them. Those tagSNPs are selected to cover most of the common variants genome-wide with the optimal and smallest subset [103]. Methods to select tagSNPs are either based on haplotype data (a SNP that identify a common haplotype) or based on LD statistics (LD block identification); the latter one giving better results for complex haplotype structures [73]. To know which SNP to genotype in a study, the tagSNP selection requires haplotype and LD data of common SNPs that covers the whole genome. This is provided by the Hapmap database for several different populations [79]. Other SNPs of interest can be included as marker: synonymous and non-synonymous SNPs (nsSNPs), miRSNPs and CNVs. For instance, Illumina's Infinium Beadchips provide genotyping of tagSNPs, genic nsSNP and CNVs with genome coverage at LD $r^2 = 0.8$ [94].

**Applications:** The aim of GWAS is to identify common risk variants with low effect, for a better understanding of complex disease and also to identify risk populations for prevention purposes. Many SNPs have been identified as associated with traits and highly significant ones have been gathered in a catalogue which contains 1260 publications and about 6400 SNPs in May 2012 [108]. Most of disease-associations of common SNPs with low effects identified by GWASs link to non-coding regions rather than coding ones, suggesting that an important part of the genetic heritability of complex diseases may be related to changes in gene regulation [85].

**Advantages:** GWAS is hypothesis-free in that there is no prior assumption about which gene, variant or pathway to test, except that it has to involve common variants. Particularly it can identify new loci or pathways that were not previously related to the disease, and improve knowledge about disease aetiologies. Another important aspect of GWAS is that it provides high coverage of the genome with a minimum set of SNPs (tagSNPs). Also GWAS has a high power for common alleles, and free controls from many databases can be used [109]. Furthermore, in contrast to linkage studies that focus on families, GWAS uses unrelated individuals, whose recombination events are much older than in families, providing a high mapping resolution [110].

**Limitations:** Once a variant has been associated with a trait through GWAS, it is difficult to identify the actual functional variant that can explain the mechanism behind the association signal measured within the LD block, particularly when the associated markers lie in non-coding regions or intergenic regions [69]. Furthermore, association results from GWAS can be difficult to replicate in independent populations because of the difference of LD patterns [69]. Also, undetected population stratification can result in false positive associations, as the statistics remain unadjusted. The biggest issues with GWASs are that they currently can explain only a small fraction (less than 10%) of genetic heritability of complex traits and that the effect sizes of associated SNPs are much smaller than expected (odd ratios typically around 1.2) [69]. This suggests that the CDCV hypothesis has reached its limits and that the missing heritability should be sought through alternative hypotheses, such as rare variants, epigenetics, CNV, and gene-gene interactions [111]. Recently, most research has been focused on multiple rare variants (MAF between 1 and 5%), since GWASs and linkage studies have low power to detect them [69].

#### 3.2.2.4 Exome sequencing

As we saw, GWAS focuses on common variants genome-wide and provides a lot of intergenic and intronic association signals that are difficult to interpret. In contrast, exome sequencing focuses on DNA regions, containing exonic variants, that can affect mRNA processing and regulation, protein translation and protein sequence and structure, but also DNA regions containing regulatory variants [112]. By focusing on coding regions, the exome sequencing strategy enables to analyse rarer variants than the ones from whole-genome analyses like whole-genome sequencing or GWAS. Furthermore exome sequencing tries to answer the missing heritability issue from GWASs, by shifting from the CDCV to the CDRV hypothesis: allelic heterogeneity of rare variants with moderate to large effect sizes.

**Principles:** Exome sequencing consists in analysing a DNA sample, by using probes to select exonic DNA regions, and then sequencing those exonic fragments using high-throughput sequencing methods, resulting in hundreds of millions short DNA reads, in a similar way as RNA-seq [112]. Once the protein-coding genes have been sequenced, variant calling and genotype calling are achieved, to identify the genotypes of new or known coding variants. Then, since rare variant associations are difficult to detect, those affecting the same locus can be aggregated together to test them in a case-control setting [69]. Similar to the Hapmap database [79] which provided to GWASs a control resource of common variants, their haplotypes and LD data, the exome sequencing approach can use the rare variants catalogued by the 1000 Genomes Project [113].

**Advantages:** By focusing on coding variants, exon sequencing makes the identification of functional variants such as nsSNPs easier, and is cheaper compared to whole-genome sequencing [109]. It does not need multiple affected relatives like linkage studies to identify rare disease-causing variants, but can aggregate rare variants to compare unrelated affected individuals with controls from the 1000 genomes project. Furthermore, sequencing-based genotyping enables to compute genotyping uncertainty and to integrate it in association tests to limit false positive associations due to genotyping errors [100].

**Limitations:** Detecting rare variant association with exome sequencing is limited to those with large effects and that lie in or near coding regions [109]. Also risk prediction of rare variants is much less precise than of common variants [109]. Furthermore, not all the regions of interest are selected for deep-sequencing yet (incomplete coverage) and it is difficult to identify CNV by exome sequencing [112]. Also, since rarer variants are more population-specific than common ones, replication studies in other populations may be even more difficult than for GWASs. Finally, rare variants may require a larger sample size than GWAS depending on the effect size [107].

# Chapter 4

# Algorithms and software

As we saw in Chapter 3, SNPs can be genotyped through hybridisation reactions of DNA or cDNA fragments. However during an experiment, not all the known variants are genotyped, generally for cost reasons, and because common SNPs close to each other are not independent, providing redundant information. Furthermore, during an analysis, the DNA or RNA materials are not necessarily available for experimental typing of SNPs that miss genotype information (referred as missing genotypes). For those reasons, it is important to be able to estimate genotypes with the available information, as we shall see in the first section.

Once genotypes are known, they can be used to analyse the effects of SNPs, such as those affecting functional parts of mRNAs and non-coding RNAs, as described in Chapter 1. One functional part described earlier is miRNA target site; the mRNA region where a miRNA binds to its target mRNA. In the second section, I shall describe the existing databases and software that analyse SNPs in miRNA target sites.

## 4.1 Genotype imputation

In a study, missing genotypes of SNPs can be imputed through several ways, depending on the type of data available. If reference haplotypes and study genotypes from neighbouring SNPs are available, it is possible to estimate missing genotypes through linkage disequilibrium, and more precisely through haplotype phasing of the neighbouring SNPs. Also, if sequencing data are available and mapped to a SNP that misses genotype, it can be estimated by analysing the reads mapped to the locus.

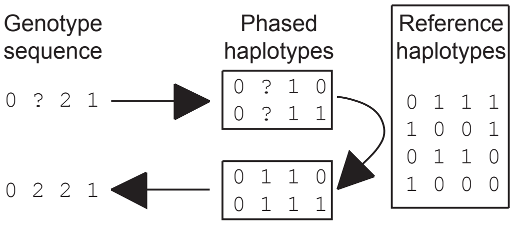


Figure 4.1: Basic example of genotype imputation using Clark's algorithm. Among four neighbouring SNPs, two of them are homozygous (0 and 2), one is heterozygous (1) and one misses genotype (?). As there is only one heterozygous SNP, the genotype sequence can be phased unambiguously into two haplotypes (0-?-1-0 and 0-?-1-1), where 0 and 1 represent alleles of each SNP. A set of reference haplotypes can be compared to our two phased haplotypes to infer the two missing alleles. Finally, combining the two haplotypes into a diplotype gives the resulting genotype sequence (0-2-2-1), where the missing SNP has been imputed as homozygous (2).

### 4.1.1 Genotype estimation from linkage disequilibrium

Genotypes of SNPs that are located within high LD regions can be estimated through genotypes of neighbouring SNPs. However, genotype imputation depends on haplotype phasing: given the genotypes of some SNPs of studied individuals and reference haplotypes from a similar population, the known genotypes must be phased to relate them to reference haplotypes and infer missing genotypes. I shall quickly describe here the naive phasing algorithm of Clark, the Expectation Maximization algorithm for phasing, and more complex methods using Hidden Markov models, such as the Impute and FastPhase tools.

**Clark's algorithm** [114] can achieve simple haplotype phasing, by first identifying multilocus genotypes that do not have more than one heterozygous site, because they can be phased unambiguously (Figure 4.1). Those new haplotypes are added to the set of known haplotypes, which then enables to phase other remaining multilocus genotypes unambiguously. After several iterations, the algorithm stops when all the haplotypes have been phased or when no more haplotypes can be resolved. Then from the phased haplotypes, missing genotypes can be inferred thanks to reference haplotypes that include the missing SNPs. Problems with this phasing method is that it may leave unresolved diplotypes, the results depend on processing order, and the method is limited to a small amount of SNPs.

**Likelihood-based Expectation Maximization algorithm** [115] computes through the expectation step the diplotype probabilities of an individual given the studied genotypes and haplotype frequencies assuming HWE. During the maximisation step, it updates the haplotype probabilities based on the diplotype probabilities from the expectation step of all individuals. Those two steps iterate until convergence and missing genotypes are inferred as for Clark's algorithm. This algorithm is more robust than Clark's algorithm, but is still limited in the amount of SNPs and assumes HWE.

**IMPUTE** tool [116] is a more computationally intensive tool, as it uses hidden Markov models (HMM). For a given individual, the model is based on an observed sequence (sequence of genotypes) and a set of hidden states (known haplotype pairs, for example from Hapmap). For $N$ haplotypes, there are $N^2$ ordered pairs, which are all the possible states. Therefore, the sequence of hidden states represents the phased diplotype. In the model, the initial state probability is uniform $\left(\frac{1}{N^2}\right)$, and the transition probabilities between states (probabilities of changing from a reference haplotype pair to another between two loci) depends on the genetic distance between the current SNP and the previous one: a small genetic distance gives a high probability that it is the same state, whereas a large one gives a low probability. Furthermore, the output probabilities of the model are the probabilities of observing the genotype given the current state (the current haplotype pair) and include mutation rate. Finally, the probability distribution of the genotype at a locus is estimated by the forward-backward algorithm.

An improved version of IMPUTE [117] divides the SNPs into two groups: those that are typed (T) in both the reference haplotype data and in the study genotype data, and those that are only typed (U) in the reference haplotype data. Then it estimates the phase of haplotypes consisting of SNPs from the T group in the study population based on all data except data from the individual being phased, by using the previous HMM model of diploid states. After phasing, it uses an HMM model of haplotype states to impute alleles at SNPs from the U group, to then estimate genotypes. Separating the phasing from the imputation enables to reduce processing time, as the diplotype-based phasing is quadratic on a reduced number of haplotypes (in T) and the haplotype-based imputation is linear on the number of all haplotypes, while the first version is quadratic on the number of all haplotypes.

**FastPhase** [118] uses a reduced amount of states compared to IMPUTE, by clustering similar haplotypes to improve computational efficiency. The clustering is based on different parameters that are estimated, as well as the recombination rate between each marker pair, by the Expectation Maximization (EM) algorithm (maximum likelihood estimates that give the known genotypes). Since the likelihood surface can have local maximums depending on the initialisation, parameters are estimated several times with several starting points. Again, for a given SNP, the genotype distribution is computed based on an HMM model for each set of parameters using the forward backward algorithm. For each possible genotype the mean probability of several sets of parameters (several starting points in the EM) gives the estimate of the genotype probability, whose maximum estimates the genotype. This method has a reduced amount of states, which makes it more efficient, but it requires parameter estimations.

### 4.1.2 Genotype estimation from sequencing data

In low LD regions or for rarer variants, LD imputation is more difficult. However, as briefly seen, in Chapter 3, genotypes can be estimated from sequencing data. Different methods such as threshold and probabilistic methods are based on counts of reads of each allele.

**Threshold-based genotyping** uses a threshold on allelic proportion to distinguish between the genotypes of a biallelic SNP: heterozygous when both allelic proportions are greater than the threshold, otherwise homozygous for the allele with higher proportion. An empirical study suggests that the threshold should belong to the interval $[0.12; 0.22]$ depending on the coverage depth [119]. The threshold approach is a simple method that enables to genotype biallelic SNPs from sequencing data. However, it does not provide any uncertainty of genotype estimates, which could be used for downstream analyses like association testing.

**Binomial distribution** can be used to compute the likelihood of each genotype given the observed allelic counts. For a biallelic A/B SNP, sequenced with a base-call error $p$, the number $X$ of B alleles among $N$ reads follows a binomial distribution: $B(N,p)$ for AA homozygous, $B(N,\frac{1}{2})$ for AB heterozygous and $B(N,1-p)$ for BB homozygous. Assuming equal prior probabilities of genotype, as they may be unknown, the genotype estimate is the one with the highest likelihood:

$$\begin{cases} P\left(X|G=BB,N,p\right)=\binom{N}{X}(1-p)^X p^{N-X} \\ P\left(X|G=AB,N,p\right)=\binom{N}{X}\left(\frac{1}{2}\right)^N \\ P\left(X|G=AA,N,p\right)=\binom{N}{X}p^X(1-p)^{N-X} \end{cases}$$

The probabilistic approach gives better estimates than the threshold one. However, since genotypes cannot have the same prior probability under HWE, it can result in overestimation of rarer genotypes.

**Bayes' theorem** can be used to classify genotypes as previously, but with prior probabilities of genotypes (the $p_{BB}$, $p_{AB}$, $p_{AA}$ frequencies). The highest joint probability of allele counts and genotypes determines the genotype estimate:

$$\begin{cases} P\left(X\wedge(G=BB)|N,p\right)=p_{BB}\binom{N}{X}(1-p)^X p^{N-X} \\ P\left(X\wedge(G=AB)|N,p\right)=p_{AB}\binom{N}{X}\left(\frac{1}{2}\right)^N \\ P\left(X\wedge(G=AA)|N,p\right)=p_{AA}\binom{N}{X}p^X(1-p)^{N-X} \end{cases}$$

Including prior knowledge of genotype frequency gives a higher accuracy [100], but this information is not necessarily known.

**Parameter estimation** may be needed in case of unknown base-call error probability and unknown genotype frequencies. Those parameters can be estimated by maximising their likelihood with the EM algorithm, as implemented in the SeqEM tool [120]. Alternatively the base error can be estimated by the quality scores from the sequencing and mapping processes.

## 4.2 Prediction of SNP effects in miRNA target sites

Once genotypes are known or imputed, it may be interesting to predict their effects, if they lie in regulatory regions such as miRNA target sites. The identification of SNPs that affect miRNA target sites is mostly based on the identification of functional miRNA target sites. Lists of validated miRNA target sites (such as the TarBase database [121]) are available, but few of them overlap with SNPs. Therefore, most of the methods that try to identify SNPs in miRNA target sites are based on target site prediction tools, such as TargetScan [122] or miRanda [123]. Details on the many different target prediction tools and the features they are based on are described elsewhere [124]. The main features generally used for target predictions are perfect matching at the seed region, the site accessibility for miRNAs, and site conservation between species. Several databases have tried to gather miRSNPs and their effect on gene expression.

**PolymiRTS** database [125] provides 3'UTR SNPs that create or disrupt seed regions of predicted miRNA target sites from TargetScan [122]. The association of those SNPs with host gene expression, also known as cis-acting expression quantitative trait locus (eQTL), have been computed in mice and humans and the high score SNPs were mapped to QTL of physiological and behavioural traits in mice, to try to identify the miRSNPs that could be responsible for the traits. However, this approach does not take expression levels of miRNA into accounts. Furthermore, phenotypes studied are only mice physiological and behavioural traits and the database lacks analyses of human traits, and particularly human diseases. This issue has been considered in the new version PolymiRTS 2.0 [126], where human SNPs from the GWAS catalogue [108] have been mapped to their nearby gene, if containing miRNA SNPs. However, this approach does not take LD into account to assume a link between the GWAS SNPs and the miRNA SNP. Furthermore, SNPs affecting experimentally validated miRNA target sites, or that lie in miRNA seed sequence have been added.

**Patrocles** database [88] provides SNPs affecting regulatory regions identified by Xie et al. [127] and sites from TargetScan predictions [122]. As for PolymiRTS, they computed mRNA eQTL from microarray data, but also included miRNA expression from sequencing in order to provide coexpression between mRNA targets and miRNAs. Furthermore, SNPs in miRNA genes and miRNA machinery were also provided. However, those miRSNPs have not been analysed for phenotype association except on SNP in sheep.

**MicroSNiPer** [128] is a web-tool that can identify miRSNPs on the fly, given a sequence or a gene. It can consider haplotypes of maximum six SNPs and one gene at a time. However, it is based on sequence search only; *i.e.* it looks for sequences complementary to miRNA seed region, resulting in probably many false positive target sites. Furthermore, it does not quantify SNP effect, and its flexible approach does not enable eQTL analysis.

The above databases can provide many SNPs and miRNAs for one gene search and may require additional filtering before testing candidate miRSNPs. Expression QTL and miRNA expression can be a way of filtering, but those expression data are not necessarily available for a tissue of interest. Filtering can also be done through LD mapping of significant SNPs from GWAS. In any case, without prior knowledge of gene expression- or phenotype-associated variants, filtering can be done after quantifying SNP effects on miRNA regulation. None of those database provides this type of quantification. However, Nicoloso *et al.* [129] also predicted miRNA target sites for each allele of 3'UTR SNPs with the miRanda target prediction tool, and calculated minimum free energy (MFE) for each allele. They used the difference of MFE to quantify SNP effects, which can be used for example for rank filtering, and tested experimentally the effects of miRSNPs that overlap known breast cancer associated SNPs and genes. However, this approach may miss many interesting results as it does not take linkage disequilibrium (LD) into account to map miRSNPs to phenotype-associated SNPs. In general, miRSNP databases do not provide SNP effect quantification and LD mapping to GWASs, as a way to analyse GWAS results. But this will be covered in Chapter 5.

# Chapter 5

# Project

The project is meant to follow up on the results generated by GWASs, to try to identify SNPs that are the cause or susceptibility for diseases, by affecting gene regulation. I shall describe its aim more in detail, the three publications it resulted in, and the potential directions it can evolve into in the near future.

## 5.1 Aim of the study

The increased use of GWASs has identified many disease-associated variants that are generally in linkage disequilibrium with the susceptibility variants. Associated variants were generally found outside coding regions, suggesting that they may affect gene regulation rather than protein structures. The aim of this project was to study DNA variants affecting gene regulation, particularly those lying within regions associated with genetic disorders from GWASs, to try to understand unexplained association signals. The study focused on SNPs affecting gene regulation by microRNAs (miRNAs) through two types of mechanisms. First, SNPs disrupting or creating miRNA target sites may affect the stability of the target mRNAs and change gene expression. Second, SNPs in polyadenylation signals may shorten 3' end of mRNAs, possibly removing miRNA target sites and making the mRNAs more stable and therefore upregulated.

## 5.2 Summary of results

**Paper I: Inferring causative variants in microRNA target sites** [130]. This paper describes a method to identify SNPs that may affect mRNA regulation by miRNAs. Based on miRNA target prediction tools, the paper identifies and analyses SNPs lying in mRNA regions complementary to miRNA seeds (miRSNPs), to try to quantify their effects on mRNA expression levels. Predicted effects were compared to mRNA allele-specific expressions from sequencing, and the two values correlated well when using the SVM target prediction tool, while predicted effects based on TargetScan scores or minimum free energy gave lower or no correlations. Furthermore, the paper describes a way to map interesting miRSNPs to disease-associated SNPs from GWASs, and shows examples of analyses on several published GWAS data. Specifically, the paper shows that SNPs in miRNA target sites that are in linkage disequilibrium with top-ranking SNPs from GWASs have a higher predicted effect, suggesting that those miRSNPs may explain some of the association signals from GWASs. Finally, the paper provides a database of miRSNPs and their predicted effects on mRNA expression levels.

**Paper II: A Risk Variant in an miR-125b Binding Site in BMPR1B Is Associated with Breast Cancer Pathogenesis** [131]. This paper is a practical use of the mapping method described in paper I. The study was based on genes dysregulated in estrogen receptor-stratified breast tumours, particularly the genes that contains SNPs affecting predicted miRNA target sites. Those miRSNPs were then mapped to top-ranking SNPs from a breast cancer GWAS study, using the method from paper I. One miRSNP (rs1434536) affecting the miR-125b miRNA regulation of the Bone Morphogenetic Receptor type 1B (*BMPR1b*) gene has been identified as being in strong linkage disequilibrium (LD) with two SNPs (rs1970801 and rs11097457) from the 100 top-ranking markers in the GWAS. The disease-association of that miRSNP was independently validated and it was shown that the two alleles of the miRSNP differently regulate the expression level of *BMPR1b*, suggesting that the miRSNP could be responsible for the disease-increased risk. Furthermore, after our study, this miRSNP has been associated with prostate cancer in Chinese men [132]. This association in another population and disease strengthens confidence about the causative role of this variant.

**Paper III: Single Nucleotide Polymorphisms Can Create Alternative Polyadenylation Signals and Affect Gene Expression through Loss of MicroRNA-Regulation** [133]. This manuscript presents how SNPs may upregulate mRNA expression levels by triggering alternative polyadenylation (APA), which results in shortening of 3' UTRs and loss of miRNA target sites. It is known that somatic mutations may trigger this mechanism and result in diseases. This manuscript shows that SNPs can also result in increased disease risk through that mechanism. The identification of candidate SNPs in APA elements such as polyA signals enabled us to show with EST and RNA-seq that such SNPs can shorten 3'UTRs, and with RNA-seq and microarray data that they can upregulate mRNA expression, particularly mRNAs losing miRNA target sites through alternative polyadenylation. Finally, through linkage disequilibrium, alleles giving APA were associated with risk alleles from GWASs.

## 5.3 Future perspectives

Those three publications focused on SNPs that affect microRNA-based regulation, through two different mechanisms: SNPs creating or disrupting miRNA target sites and SNPs creating alternative polyadenylation signal, which shortens the UTR and suppresses miRNA target sites downstream.

It would be interesting to analyse SNPs disrupting polyadenylation sites, making 3' UTRs longer and destabilising mRNAs by miRNA targeting, which would results in decreased gene expression. RNA-seq will provide precious data to try to estimate 3' end of longer transcripts. However, longer transcripts may be challenging to validate *in vitro*.

Also, integrating miRSNPs, APA-SNPs and alternative polyadenylation through a haplotype-based analysis that involves miRNA target prediction and quantification of haplotype effects on gene expression could be a way to follow up on article I and III. Furthermore, data from the 1000 genomes project will be very helpful for looking at rarer SNPs.

Finally, other regulatory elements than miRNA target sites can be affected by SNPs. Particularly, regulatory regions involved in mRNA transcription such as transcription factor binding sites may be strongly affected by SNPs. Analyses of these SNPs will be mostly based on emerging sequencing methods such as ChIP-seq, that provides better transcription factor binding site predictions than the previously used position weight matrix algorithms.

# Bibliography

[1] Douglas AGL, Wood MJA (2011) RNA splicing: disease and therapy. Brief Funct Genomics 10: 151-164.

[2] Pruitt K, Tatusova T, Maglott D (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501-D504.

[3] Lutz CS (2008) Alternative Polyadenylation: A Twist on mRNA 3 ‘ End Formation. ACS Chem Biol 3: 609-617.

[4] Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY (2011) Understanding the transcriptome through RNA structure. Nat Rev Genet 12: 641-655.

[5] Hocine S, Singer RH, Grunwald D (2010) RNA Processing and Export. Cold Spring Harbor Perspect Biol 2.

[6] Lee T, Young R (2000) Transcription of eukaryotic protein-coding genes. Annu Rev Genet 34: 77-137.

[7] Proudfoot NJ (2011) Ending the message: poly(A) signals then and now. Genes Dev 25: 1770-1782.

[8] Licatalosi DD, Darnell RB (2010) RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 11: 75–87.

[9] Di Giammartino DC, Nishida K, Manley JL (2011) Mechanisms and Consequences of Alternative Polyadenylation. Mol Cell 43: 853-866.

[10] Lopez MD, Samuelsson T (2008) Early evolution of histone mRNA 39 end processing. RNA-Publ RNA Soc 14: 1-10.

[11] Tian B, Hu J, Zhang H, Lutz C (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res 33: 201-212.

[12] Colgan D, Manley J (1997) Mechanism and regulation of mRNA polyadenylation. Genes Dev 11: 2755-2766.

[13] Danckwardt S, Kaufmann I, Gentzel M, Foerstner KU, Gantzert AS, et al. (2007) Splicing factors stimulate polyadenylation via USEs at non-canonical 3 ‘ end formation signals. Embo J 26: 2658-2669.

[14] Nunes NM, Li W, Tian B, Furger A (2010) A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. Embo J 29: 1523-1536.

[15] Abruzzi K, Lacadie S, Rosbash M (2004) Biochemical analysis of TREX complex recruitment to intronless and intron-containing yeast genes. Embo J 23: 2620-2631.

[16] Kelly SM, Corbett AH (2009) Messenger RNA Export from the Nucleus: A Series of Molecular Wardrobe Changes. Traffic 10: 1199-1208.

[17] Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559-1563.

[18] Cooper TA, Wan L, Dreyfuss G (2009) RNA and Disease. Cell 136: 777-793.

[19] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40: 1413–1415.

[20] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470-476.

[21] Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T (2011) Epigenetics in Alternative Pre-mRNA Splicing. Cell 144: 16-26.

[22] de la Mata M, Alonso C, Kadener S, Fededa J, Blaustein M, et al. (2003) A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell 12: 525-532.

[23] Kalsotra A, Cooper TA (2011) Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet 12: 715–729.

[24] Skotheim RI, Nees M (2007) Alternative splicing in cancer: Noise, functional, or systematic? Int J Biochem Cell Biol 39: 1432-1449.

[25] Aartsma-Rus A, Van Ommen GJB (2007) Antisense-mediated exon skipping: A versatile tool with therapeutic and research applications. RNA-Publ RNA Soc 13: 1609-1624.

[26] Mansfield S, Chao H, Walsh C (2004) RNA repair using spliceosome-mediated RNA trans-splicing. Trends Mol Med 10: 263-268.

[27] Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB (2008) Proliferating cells express mRNAs with shortened 3 ‘ untranslated regions and fewer microRNA target sites. Science 320: 1643-1647.

[28] Mayr C, Bartel DP (2009) Widespread Shortening of 3 ‘ UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. Cell 138: 673-684.

[29] Ji Z, Lee JY, Pan Z, Jiang B, Tian B (2009) Progressive lengthening of 3 ' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A 106: 7028-7033.

[30] Danckwardt S, Hentze MW, Kulozik AE (2008) 3 ' end mRNA processing: molecular mechanisms and implications for health and disease. Embo J 27: 482-498.

[31] Fabian MR, Sonenberg N, Filipowicz W (2010) Regulation of mRNA Translation and Stability by microRNAs. In: Annual Review of Biochemistry, volume 79 of *Annual Review of Biochemistry*. pp. 351-379. doi:{10.1146/annurev-biochem-060308-103103}.

[32] Kahvejian A, Svitkin Y, Sukarieh R, M'Boutchou M, Sonenberg N (2005) Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms. Genes Dev 19: 104-113.

[33] Thomson T, Lin H (2009) The Biogenesis and Function of PIWI Proteins and piRNAs: Progress and Prospect. Annu Rev Cell Dev Biol 25: 355-376.

[34] Wapinski O, Chang HY (2011) Long noncoding RNAs and human disease. Trends Cell Biol 21: 354-361.

[35] Bartel D (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116: 281-297.

[36] Friedman RC, Farh KKH, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19: 92-105.

[37] Lu J, Getz G, Miska E, Alvarez-Saavedra E, Lamb J, et al. (2005) MicroRNA expression profiles classify human cancers. Nature 435: 834-838.

[38] Flynt AS, Lai EC (2008) Biological principles of microRNA-mediated regulation: shared themes amid diversity. Nat Rev Genet 9: 831-842.

[39] Vasudevan S, Tong Y, Steitz JA (2007) Switching from repression to activation: MicroRNAs can up-regulate translation. Science 318: 1931-1934.

[40] Orom UA, Nielsen FC, Lund AH (2008) MicroRNA-10a binds the 5 ' UTR of ribosomal protein mRNAs and enhances their translation. Mol Cell 30: 460-471.

[41] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34: D140-D144.

[42] Lee Y, Kim M, Han J, Yeom K, Lee S, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. Embo J 23: 4051-4060.

[43] Borchert GM, Lanier W, Davidson BL (2006) RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol 13: 1097-1101.

[44] Winter J, Jung S, Keller S, Gregory RI, Diederichs S (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat Cell Biol 11: 228-234.

[45] Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, et al. (2005) Clustering and conservation patterns of human microRNAs. Nucleic Acids Res 33: 2697-2706.

[46] Denli A, Tops B, Plasterk R, Ketting R, Hannon G (2004) Processing of primary microRNAs by the Microprocessor complex. Nature 432: 231-235.

[47] Lee Y, Ahn C, Han J, Choi H, Kim J, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. Nature 425: 415-419.

[48] Han J, Lee Y, Yeom K, Nam J, Heo I, et al. (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell 125: 887-901.

[49] Yang JS, Lai EC (2011) Alternative miRNA Biogenesis Pathways and the Interpretation of Core miRNA Pathway Mutants. Mol Cell 43: 892-903.

[50] Shomron N, Levy C (2009) MicroRNA-Biogenesis and Pre-mRNA Splicing Crosstalk. J Biomed Biotechnol .

[51] Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell 130: 89-100.

[52] Baskerville S, Bartel D (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA-Publ RNA Soc 11: 241-247.

[53] Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, et al. (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18: 610-621.

[54] Bartel DP (2009) MicroRNAs: Target Recognition and Regulatory Functions. Cell 136: 215-233.

[55] Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466: 835-U66.

[56] Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, et al. (2007) MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. Mol Cell 27: 91-105.

[57] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics 8.

[58] Shin C, Nam JW, Farh KKH, Chiang HR, Shkumatava A, et al. (2010) Expanding the MicroRNA Targeting Code: Functional Sites with Centered Pairing. Mol Cell 38: 789-802.

[59] Saetrom P, Heale BSE, Snove O Jr, Aagaard L, Alluin J, et al. (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res 35: 2333-2342.

[60] Gu S, Jin L, Zhang F, Sarnow P, Kay MA (2009) Biological basis for restriction of microRNA targets to the 3 ‘ untranslated region in mammalian mRNAs. Nat Struct Mol Biol 16: 144-150.

[61] Garzon R, Fabbri M, Cimmino A, Calin GA, Croce CM (2006) MicroRNA expression and function in cancer. Trends Mol Med 12: 580-587.

[62] Hebert SS, De Strooper B (2009) Alterations of the microRNA network cause neurodegenerative disease. Trends Neurosci 32: 199-206.

[63] Calin GA, Croce CM (2006) MicroRNA signatures in human cancers. Nat Rev Cancer 6: 857-866.

[64] Witkos TM, Koscianska E, Krzyzosiak WJ (2011) Practical Aspects of microRNA Target Prediction. Curr Mol Med 11: 93-109.

[65] Muralidhar B, Goldstein LD, Ng G, Winder DM, Palmer RD, et al. (2007) Global microRNA profiles in cervical squamous cell carcinoma depend on Drosha expression levels. J Pathol 212: 368-377.

[66] Thomson JM, Newman M, Parker JS, Morin-Kensicki EM, Wright T, et al. (2006) Extensive post-transcriptional regulation of microRNAs and its implications for cancer. Genes Dev 20: 2202-2207.

[67] Lee EJ, Baek M, Gusev Y, Brackett DJ, Nuovo GJ, et al. (2008) Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors. RNA-Publ RNA Soc 14: 35-42.

[68] van Kouwenhove M, Kedde M, Agami R (2011) MicroRNA regulation by RNA-binding proteins and its implications for cancer. Nat Rev Cancer 11: 644-656.

[69] Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241-251.

[70] Guichoux E, Lagache L, Wagner S, Chaumeil P, Leger P, et al. (2011) Current trends in microsatellite genotyping. Mol Ecol Resour 11: 591-611.

[71] Mayo O (2008) A century of Hardy-Weinberg equilibrium. Twin Res Hum Genet 11: 249-256.

[72] Sherry S, Ward M, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.

[73] Crawford D, Nickerson D (2005) Definition and clinical importance of haplotypes. Annu Rev Med 56: 303+.

[74] Williams MA, Carson R, Passmore P, Silvestri G, Craig D (2011) Introduction to genetic epidemiology. Optometry 82: 83-91.

[75] Dawn Teare M, Barrett JH (2005) Genetic linkage studies. "Lancet" 366: 1036 - 1044.

[76] Sved JA (2009) Linkage Disequilibrium and Its Expectation in Human Populations. Twin Res Hum Genet 12: 35-43.

[77] Lewontin R (1964) Interaction of Selection + Linkage .I. General Considerations - Heterotic Models. Genetics 49: 49-&.

[78] Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38: 226–231.

[79] Altshuler D, Brooks L, Chakravarti A, Collins F, Daly M, et al. (2005) A haplotype map of the human genome. Nature 437: 1299-1320.

[80] Sauna ZE, Kimchi-Sarfaty C (2011) Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12: 683-691.

[81] Cartegni L, Chew S, Krainer A (2002) Listening to silence and understanding nonsense: Exonic mutations that affect splicing. Nat Rev Genet 3: 285-298.

[82] Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, et al. (2011) A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nature Genet 43: 242-U24.

[83] Halvorsen M, Martin JS, Broadaway S, Laederach A (2010) Disease-Associated Mutations That Alter the RNA Structural Ensemble. PLoS Genet 6.

[84] Ryan BM, Robles AI, Harris CC (2010) Genetic variation in microRNA networks: the implications for cancer research. Nat Rev Cancer 10: 389-402.

[85] Sethupathy P, Collins FS (2008) MicroRNA target site polymorphisms and human disease. Trends Genet 24: 489-497.

[86] Chen K, Rajewsky N (2006) Natural selection on human microRNA binding sites inferred from SNP data. Nature Genet 38: 1452-1456.

[87] Saunders MA, Liang H, Li WH (2007) Human polymorphism at microRNAs and microRNA target sites. Proc Natl Acad Sci U S A 104: 3300-3305.

[88] Hiard S, Charlier C, Coppieters W, Georges M, Baurain D (2010) Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. Nucleic Acids Res 38: D640-D651.

[89] Clague J, Lippman SM, Yang H, Hildebrandt MAT, Ye Y, et al. (2010) Genetic Variation in MicroRNA Genes and Risk of Oral Premalignant Lesions. Mol Carcinog 49: 183-189.

[90] Melo SA, Ropero S, Moutinho C, Aaltonen LA, Yamamoto H, et al. (2009) A TARBP2 mutation in human cancer impairs microRNA processing and DICER1 function. Nature Genet 41: 365-370.

[91] Uitte De Willige S, Rietveld IM, De Visser MCH, Vos HL, Bertina RM (2007) Polymorphism 10034c>t is located in a region regulating polyadenylation of fgg transcripts and influences the fibrinogen $\gamma'/\gamma a$ mrna ratio. J Thromb Haemost 5: 1243–1249.

[92] Pais H, Moxon S, Dalmay T, Moulton V (2011) Small RNA Discovery and Characterisation in Eukaryotes Using High-Throughput Approaches. In: Collins, LJ, editor, RNA Infrastructure and Networks, volume 722 of *Advances in Experimental Medicine and Biology*. pp. 239-254.

[93] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24: 1151-1161.

[94] Ragoussis J (2009) Genotyping Technologies for Genetic Research. Annu Rev Genomics Hum Genet 10: 117-133.

[95] Malone JH, Oliver B (2011) Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 9.

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

[97] Johnson J, Castle J, Garrett-Engele P, Kan Z, Loerch P, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302: 2141-2144.

[98] Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

[99] Lander E, Linton L, Birren B, Nusbaum C, Zody M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

[100] Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12: 443-451.

[101] Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38: 1767-1771.

[102] Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19: 212-219.

[103] Hindorff LA, Gillanders EM, Manolio TA (2011) Genetic architecture of cancer and other complex diseases: lessons learned and future directions. Carcinogenesis 32: 945-954.

[104] Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nat Rev Genet 12: 703-714.

[105] Neale BM, AR FM, Medland SE, Posthuma D (2008) Statistical genetics: gene mapping through linkage and association. Taylor & Francis Group. URL `http://books.google.fr/books?id=Tf5EAQAAIAAJ`.

[106] Hirschhorn J, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4: 45-61.

[107] Chung CC, Chanock SJ (2011) Current status of genome-wide association studies in cancer. Hum Genet 130: 59-78.

[108] Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed May 2012.

[109] Carvajal-Carmona LG (2010) Challenges in the identification and use of rare disease-associated predisposition variants. Curr Opin Genet Dev 20: 277-281.

[110] Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, et al. (2009) Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. Plant Cell 21: 2194-2202.

[111] Danchin E, Charmantier A, Champagne FA, Mesoudi A, Pujol B, et al. (2011) Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. Nat Rev Genet 12: 475-486.

[112] Singleton AB (2011) Exome sequencing: a transformative technology. Lancet Neurol 10: 942-946.

[113] Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

[114] Clark A (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7: 111-122.

[115] Excoffier L, Slatkin M (1995) Maximum-likelihood-estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921-927.

[116] Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genet 39: 906-913.

[117] Howie BN, Donnelly P, Marchini J (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5.

[118] Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629-644.

[119] Hedges D, Burges D, Powell E, Almonte C, Huang J, et al. (2009) Exome Sequencing of a Multigenerational Human Pedigree. PLoS One 4.

[120] Martin ER, Kinnamon DD, Schmidt MA, Powell EH, Zuchner S, et al. (2010) SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. Bioinformatics 26: 2803-2810.

[121] Sethupathy P, Corda B, Hatzigeorgiou A (2006) TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA-Publ RNA Soc 12: 192-197.

[122] Lewis B, Burge C, Bartel D (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15-20.

[123] Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, et al. (2006) A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. Cell 126: 1203-1217.

[124] Saito T, Saetrom P (2010) MicroRNAs - targeting and target prediction. New Biotech 27: 243-249.

[125] Bao L, Zhou M, Wu L, Lu L, Goldowitz D, et al. (2007) PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits. Nucleic Acids Res 35: D51-D54.

[126] Ziebarth JD, Bhattacharya A, Chen A, Cui Y (2012) PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. Nucleic Acids Res 40: D216-D221.

[127] Xie X, Lu J, Kulbokas E, Golub T, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3 ‘ UTRs by comparison of several mammals. Nature 434: 338-345.

[128] Barenboim M, Zoltick BJ, Guo Y, Weinberger DR (2010) MicroSNiPer: A Web Tool for Prediction of SNP Effects on Putative microRNA Targets. Hum Mutat 31: 1223-1232.

[129] Nicoloso MS, Sun H, Spizzo R, Kim H, Wickramasinghe P, et al. (2010) Single-Nucleotide Polymorphisms Inside MicroRNA Target Sites Influence Tumor Susceptibility. Cancer Res 70: 2789-2798.

[130] Thomas LF, Saito T, Saetrom P (2011) Inferring causative variants in microRNA target sites. Nucleic Acids Res 39.

[131] Saetrom P, Biesinger J, Li SM, Smith D, Thomas LF, et al. (2009) A Risk Variant in an miR-125b Binding Site in BMPR1B Is Associated with Breast Cancer Pathogenesis. Cancer Res 69: 7459-7465.

[132] Feng N, Xu B, Tao J, Li P, Cheng G, et al. (2012) A miR-125b binding site polymorphism in bone morphogenetic protein membrane receptor type IB gene and prostate cancer risk in China. Mol Biol Rep 39: 369-373.

[133] Thomas LF, Saetrom P (2012) Single Nucleotide Polymorphisms Can Create Alternative Polyadenylation Signals and Affect Gene Expression through Loss of MicroRNA-Regulation. Manuscript accepted in *PLoS Comput. Biol.*, [In Press].

# Paper I

*Published online 21 June 2011* *Nucleic Acids Research, 2011, Vol. 39, No. 16* ***e109***
*doi:10.1093/nar/gkr414*

# Inferring causative variants in microRNA target sites

**Laurent F. Thomas[1,2,*], Takaya Saito[1] and Pål Sætrom[1,2,3,*]**

[1]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, N-7489 Trondheim, Norway, [2]Interagon AS, Laboratoriesenteret, NO-7006 Trondheim and [3]Department of Computer and Information Science, Norwegian University of Science and Technology, N-7489 Trondheim, Norway

Received December 21, 2010; Revised May 5, 2011; Accepted May 9, 2011

## ABSTRACT

**MicroRNAs (miRNAs) regulate genes post transcription by pairing with messenger RNA (mRNA). Variants such as single nucleotide polymorphisms (SNPs) in miRNA regulatory regions might result in altered protein levels and disease. Genome-wide association studies (GWAS) aim at identifying genomic regions that contain variants associated with disease, but lack tools for finding causative variants. We present a computational tool that can help identifying SNPs associated with diseases, by focusing on SNPs affecting miRNA-regulation of genes. The tool predicts the effects of SNPs in miRNA target sites and uses linkage disequilibrium to map these miRNA-related variants to SNPs of interest in GWAS. We compared our predicted SNP effects in miRNA target sites with measured SNP effects from allelic imbalance sequencing. Our predictions fit measured effects better than effects based on differences in free energy or differences of TargetScan context scores. We also used our tool to analyse data from published breast cancer and Parkinson's disease GWAS and significant trait-associated SNPs from the NHGRI GWAS Catalog. A database of predicted SNP effects is available at http://www.bigr.medisin.ntnu.no/mirsnpscore/. The database is based on haplotype data from the CEU HapMap population and miRNAs from miRBase 16.0.**

## INTRODUCTION

MicroRNAs (miRNAs) are small non-coding single stranded RNAs of about 22 nucleotides length that regulate genes post transcription by partially pairing with 3′-untranslated regions (3′-UTR) of messenger RNA (mRNA) (1). Watson–Crick pairing to nucleotides 2–7 of the 5′-end of microRNAs (seed sites) is known to be important in mRNA targeting. Specifically, miRNAs require almost perfect complementarity at seed sites for binding and reducing the protein levels of targets (2). However, mRNA sites with perfect complementarity to the seed nucleotides are not necessarily functional (3) and those with imperfect seed complementarity can also be functional (2). Consequently, considering seed sites alone gives many false positive miRNA target sites. Predictions can be improved, however, by using information about the target sites' context, such as their position within the 3′-UTR (4) and the distance to neighbouring sites (5), as such context is critical for target site functionality and efficacy.

Genome-wide association studies (GWAS) can identify genomic regions that contain genomic alterations, such as single nucleotide polymorphisms (SNPs), associated with common disease (6). The biological effects of identified alterations are usually not known, however, as few of the functional variants that show association in GWAS change the amino acid sequence. Moreover, a sizeable proportion is thought to reside in regulatory regions, since several associated regions found in GWAS lack known genes (7). Variants in regulatory regions can, for example, result in altered protein levels, so identifying and understanding their effects can improve diagnostics and treatments for diseases (8). Specifically, SNPs in regulatory elements such as miRNA target sites can affect phenotype (9) and have been associated with increased cancer risk (10) and other diseases (11). The increased use of GWAS to study genetic factors in common disease necessitates a tool that can identify and interpret effects of regulatory variants.

Several research groups have tried to look at regulatory variant effects. Bao *et al.* (12) looked for SNPs in putative conserved miRNA target sites [from the target site prediction tool TargetScan (13)], and integrated such SNP sites with phenotype (physiological and behavioural traits

*To whom correspondence should be addressed. Fax: +47 72571463; Email: laurent.thomas@ntnu.no
Correspondence may also be addressed to Pål Sætrom. Tel: +4798203874; Email: pal.satrom@ntnu.no

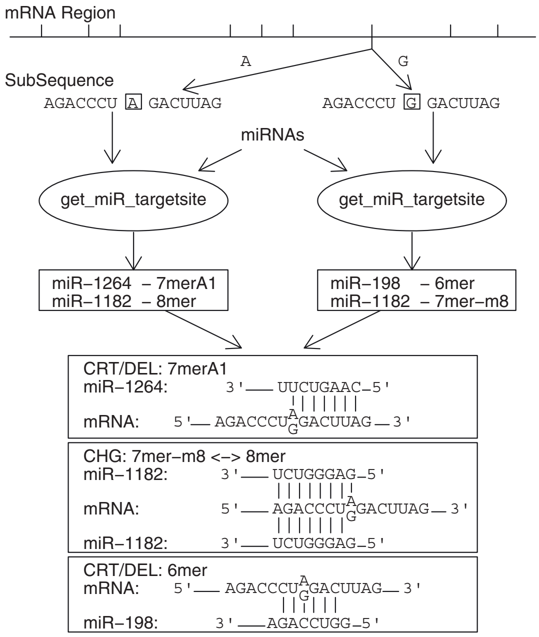


**Figure 1.** Identifying SNPs in miRNA target sites. The illustration shows an mRNA region that contains SNPs represented by small vertical lines. The considered SNP has two alleles: A and G. We make one subsequence for each allele by using the flanking regions of the SNP (7 nucleotides on each side). Given miRNA seed motifs (nucleotides 2–8 from the 5′-end of miRNA sequences), we look for target sites in each allele sequence and then compare results to characterise the effect of the SNP (create/delete (CRT/DEL) target sites, or change (CHG) site type).

of mice as quantitative trait loci) and expression data (of mice and human transcripts) into a database. However, the studied phenotypes only concern physiology of mice instead of human diseases. Georges *et al*. (14) also made a database with SNPs in putative miRNA target sites [regulatory motifs identified in (15) and predicted sites from (13)], but Georges *et al*. (14) did not map their site SNPs to phenotypes, except for one SNP in sheep. Barenboim *et al*. (16) developed an online tool that finds SNPs in microRNA target sites on the fly. The tool takes haplotype into account, but is limited to one single gene and six SNPs per run and does not quantify SNP effects. Nicoloso *et al*. (17) used the miRanda tool (18) to identify breast cancer-associated SNPs that disrupt miRNA target sites. The authors filtered SNPs based on minimum free energy (MFE) and tested the remaining ones in a case-control study.

A basic way of detecting SNPs in microRNA target sites (mirSNPs) in a gene *g*, starts by looking at SNPs lying in a region of interest, such as 3′-UTR, 5′-UTR, coding or promoter region (Figure 1). Here, we will use the 3′-UTR as an example, since SNPs affecting miRNA target sites are more likely to reside in the 3′-UTR (19,20). Let us consider a SNP *s* in this region of interest. The SNP *s* has several alleles, usually two, that we want to evaluate for targeting by a microRNA seed motif *m*. Specifically, for each allele $a_i$, we determine whether there is a microRNA target site in a sequence $als_i$ consisting of the allele $a_i$ and its flanking sequences. Target sites are detected by using any miRNA target site prediction tool based on sequence search. It is convenient to disregard target sites with mismatches in the seed region and only consider 6-mer, 7-mer and 8-mer seed sites. For each allelic sequence $als_i$, we get a list $l_i$ of target sites for microRNA *m*. We can then compare these lists to determine if a target site is created, deleted, or changed between the alleles (Figure 1).

All existing tools use variants of the approach above of evaluating candidate sites individually (Figure 1), but this approach ignores that 3′-UTRs can contain multiple linked SNPs that can affect miRNA targeting by altering site context. Instead, we propose to analyse all the SNPs of the 3′-UTR at the same time, to have a general overview of the SNPs' regulatory effect on the considered mRNA.

In this article, we present a computational tool that can help identifying SNPs causative to diseases, such as cancer. The tool focuses on SNPs that may affect miRNA targeting and thereby cause gene dysregulation. More precisely, the tool predicts the effects of SNPs in miRNA target sites and uses linkage disequilibrium to map those mirSNPs to SNPs of interest in GWAS. We show that the tool's predictions correspond well to the SNP's measured effects on miRNA regulation, and that the predictions correlate better to those effects than do the predictions of other existing tools. We further demonstrate the tool's utility by analysing two published GWAS data sets and specific SNPs reported to affect miRNA targeting.

## MATERIALS AND METHODS

The following sections will present a method that uses context-based miRNA target prediction to quantify the effects of SNPs in miRNA target sites (mirSNPs) and uses linkage disequilibrium to map candidate mirSNPs to disease data from GWAS. The tool allows additional filtering of candidate genes and candidate miRNAs. The tool's mapping method is general and can therefore be applied to SNPs independent of the scoring method used.

### Data

We used the SNP data from the human haplotype map project [HapMap, (21)]; particularly, SNP data from the CEU population (CEPH - Utah residents with ancestry from northern and western Europe), release 22 for haplotype data, and release 27 for linkage disequilibrium data. We used DNA sequences from the human and mouse genome assemblies hg18 and mm9 (22,23). SNPs and Gene annotations (hg18,mm9) came from UCSC Genome browser (24). MicroRNA sequences came from miRBase, release 13.0 and 16.0 (25). GWAS data were from a breast cancer study from Cancer Genetic Markers of Susceptibility (CGEMS) (26), from a Parkinson disease study (*P*-values from tier 1) (27), and from the NHGRI GWAS catalog (28) (http://www.genome.gov/gwastudies).

### MicroRNA regulation score of haplotypes

To analyse all the SNPs of the 3′-UTR at the same time, we use population haplotype data for the 3′-UTR (Figure 2 and Supplementary Figure S1). Specifically, we first use haplotype data to build haplotype sequences $hs_i$; i.e. 3′-UTR sequences containing the combinations of alleles found in the considered population. Second, for a given miRNA $m$, we use a miRNA target prediction tool (29) to score each haplotype sequence $hs_i$. The prediction tool uses a two-step SVM classifier, where one SVM step classifies individual target sites and a subsequent SVM step classifies overall mRNA targeting potential. Features the SVM uses at the first step include seed pairing, 3′ supplementary pairing, the site's AU context and relative position in the 3′-UTR, and distance to neighbouring sites, whereas features at the second step include 3′-UTR length, the number and predicted strength of target sites, and the number of optimally spaced sites in the 3′-UTR (29). As output, the SVM-based prediction tool gives a score such that a high output score indicates that the miRNA $m$ is likely to down-regulate this mRNA. Third, we compare the score-haplotype pairs to find the differences of haplotypes that can explain any differences of SVM scores. From the differences of haplotypes, we can make a list of candidate SNPs and predict their impact on gene regulation.

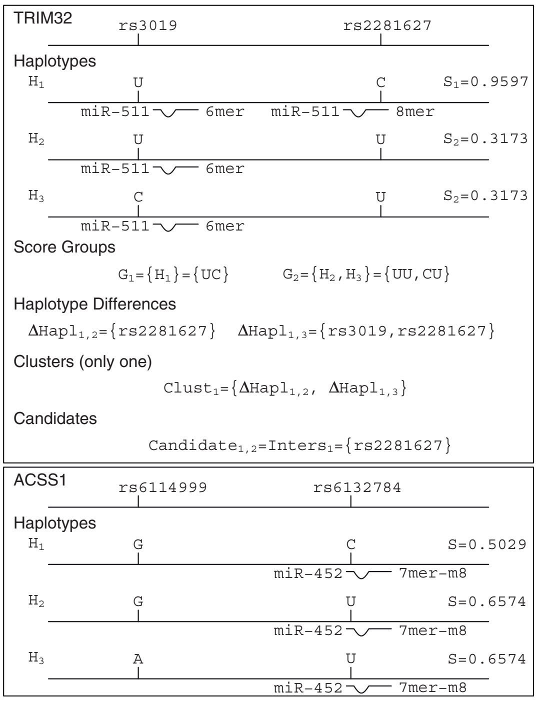


**Figure 2.** Scoring SNPs in miRNA target sites. rs3019 and rs2281627 are SNPs in the 3′-UTR of *TRIM32*. There are 3 different haplotypes in the CEU population: UC/UU/CU. *TRIM32* is targeted by miR-511, but the U allele of rs2281627 disrupts one seed site, which results in a lower score $S_2$ for the UU/CU haplotypes. To identify rs2281627 as the effect SNP, first the 3 haplotypes $H_1$, $H_2$ and $H_3$ are grouped by scores into $G_1$ and $G_2$. Second, we identify the differences between haplotypes from groups $G_1$ and $G_2$; i.e. differences between $H_1$ and $H_2$ and between $H_1$ and $H_3$. Third, we cluster those haplotype differences, so that the intersection within the cluster is not empty; here, there is only one cluster. Finally, we take the intersection of haplotype differences within this cluster, which gives the SNP rs2281627. Similarly, rs6114999 and rs6132784 lie in the 3′-UTR of *ACSS1*. There are 3 haplotypes: GC/GU/AU. Both SNPs lie outside of any seed sites of miR-452, but rs6132784 lies in a 3′-supplementary site and has a small effect on the scores.

The haplotype score comparison works as follows. First we group haplotypes $H_i$ by scores, since we are interested in score differences:

$$G_s = \{H_i \in H \mid Score(H_i) = s\}.$$

Second, we look at the difference of haplotypes between groups, to identify which SNPs differ between two score groups: $\forall(G_m, G_n)$, $m \neq n$, $\forall H_i \in G_m$, $\forall H_j \in G_n$,

$$\Delta Haplo_{ij} = \{snp | H_i(snp) \neq H_j(snp)\}.$$

Third, we cluster the $\Delta Haplo$ SNP sets, to handle particular cases such as two SNPs in one target site (Supplementary Figure S2). Specifically, we cluster $\Delta Haplo$ sets such that in each cluster, the intersection of all the $\Delta Haplo_{ij}$ of the cluster is not empty:

$$Clust_k = \left\{\Delta Haplo_{ij} \middle| \bigcap \Delta Haplo_{ij} \neq \emptyset\right\}.$$

Fourth, we take the intersection of the $\Delta Haplo$ SNP sets in each cluster, to identify which SNP is responsible for the score difference in each cluster:

$$Inters_k = \bigcap Clust_k = \bigcap_{\Delta Haplo_{ij} \in Clust_k} \Delta Haplo_{ij}.$$

Finally, we merge all the clusters to create a list of SNPs responsible for the score difference for the clusters:

$$Candidate_{mn} = \bigcup_k Inters_k.$$

$Candidate_{mn}$ are candidate SNPs that might explain the difference between the scores $m$ and $n$.

### Normalization of target site scores

The miRNA target site prediction tool (29) predicts both the targeting potential of individual candidate sites and the total regulatory potential of candidate 3′-UTRs; i.e. if a gene's 3′-UTR sequence contains one or more candidate miRNA target sites, the tool scores the miRNA's regulatory effect on the target gene. However, the tool does not score mRNAs without target site candidates. Consequently, to score and compare scores for sequences with and without candidate sites, we needed to create a normalized score. The desired distribution should be mainly uniform, because the difference between two transformed scores should reflect a difference in percentiles in the original distribution. Since we only get scores for sequences with target sites, we had to find a way to score sequences that do not have target sites and to compare sequences with and without target sites. Our solution consisted of normalizing the scores in the interval [0, 1]. As there are more sequences without target sites than with target sites, we normalized scores so that the codomain of the normalization has an exponential distribution in [0, 0.01] and a uniform distribution in [0.01, 1], according to the following probability density function:

$$df(y) = \begin{cases} \lambda e^{-\alpha\lambda y} & y \in [0, \tau] \\ \frac{P_{Unif}}{1-\tau} & y \in [\tau, 1]. \end{cases}$$

Here, τ is the threshold that separates the two distributions in the codomain. To jointly score sequences with and without target sites, we considered sequences with only one target site as an intermediate. Since we needed to put the worst target site scores in the exponential part, we used the score distribution of mRNAs that have only one target site, which is a 6-mer. Specifically, we used the fifth percentile of the 6-mer distribution to define the threshold $T$: $P(X_{6m} < T) = 0.05$. This threshold then separated the exponential distribution from the uniform distribution in the domain of the normalization morphism. As a result, the exponential part contained scores for sequences that have no target site (TS) (including those with mismatch target sites) or canonical target sites with a score lower than $T$. The proportion of scores that will be in the uniform part is $P_{Unif} = P[X \geq T]P_{TS}$, where $P_{TS}$ is the probability of having a target site and $P[X \geq T]$ is the proportion of scores greater than $T$. The proportion of scores in the exponential part is $P_{Exp} = 1 - P_{Unif}$. The parameter $\lambda = -\frac{1}{\alpha\tau}\log(1 - \alpha P_{Exp})$ makes the cumulative distribution of the exponential part fit $P_{Exp}$. The parameter $\alpha \in \left]0, \frac{1}{P_{Exp}}\right[$ makes the two distributions continuous in τ and minimizes

$$f(\alpha) = \left( -\frac{1 - \alpha P_{Exp}}{\alpha\tau}\log(1 - \alpha P_{Exp}) - \frac{P_{Unif}}{1 - \tau} \right)^2.$$

We chose τ = 0.01 as a trade-off between τ being so small that all the scores from the exponential part had the same tendency, and being so large that we could find the α that minimized $f(\alpha)$.

## Mapping candidate SNPs to disease

We can map candidate mirSNPs to disease by filtering on genes that are dysregulated in a given disease, filtering on miRNAs that are dysregulated in a given disease, and filtering on disease-associated SNPs from the same genomic region as the candidate. As filtering on genes or miRNAs simply involves focusing on subsets of the UTRs or miRNAs, we detail the filtering on disease-associated SNPs.

Association studies can show association of marker SNPs with a disease, but not necessarily association of a causal SNP with the disease. Consequently, if we want to know whether a candidate mirSNP may be causal, we first have to map it to associated marker SNPs.

Mapping candidate SNPs to association studies consists in looking for GWAS top ranking SNPs that have been inherited together with our candidate SNPs; i.e. looking for candidate SNPs that have alleles that correlate with alleles of associated marker SNPs. This can be achieved by computing inheritance blocks.

Inheritance blocks are DNA regions with highly correlated alleles. Consequently, by knowing the alleles of one SNP of the block one can predict the alleles at another SNP of the block. This measure of inheritance is called linkage disequilibrium (LD). Given a candidate SNP, we can compute its inheritance block, according to HapMap data. The block is an area of strong linkage disequilibrium and shows SNPs that have high correlation between themselves and with the candidate SNP.

We can define a block as a set of successive SNPs:

$$Block = \{s_l, \ldots, s_r\},$$

where $s_l$ and $s_r$ are the left and right bound SNPs of the block.

A block spine is a set of LD values:

$$Spine = \{D'_{lj}\} \cup \{D'_{ir}\},$$

such that $l < j \leq r$ and $l < i < r$ and where $D'_{xy}$ is the linkage disequilibrium between the SNPs $s_x$ and $s_y$. In short, the spine consists of the borders of the block (the two borders of the triangle block).

A solid spine is a spine where a relative amount α of the spine's LD values is below a threshold $T$. For example, we can use α = 10% and $T$ = 0.8, to detect blocks with strong LD.

The block detection method (Figure 3) is called Solid Spine by Expansion and is an adaptation of the Solid Spine algorithm developed within the Haploview software (30). This expansion algorithm uses a candidate SNP as input. It starts the expansion from this SNP and then tries to expand the block successively in the downstream and upstream directions. An expansion occurs if the spine of the expanded block fits a rule depending on α and $T$. This algorithm needs an area of high LD to expand, which ensures that the algorithm returns few false positive blocks. The expansion can start on the left side as well as on the right side and the two directions can give different results. As we are interested in finding all SNPs that reside in blocks that have high LD with of the input SNP, we consider both resulting blocks.

Given a block of SNPs identified by the Solid Spine by Expansion algorithm above, we then extract GWAS top ranking SNPs from the block, to identify if the candidate SNP is correlated with any associated SNPs. We consider a SNP to be top-ranking when its rank is less than a given threshold.

We define three scores to assess the level of LD of the block defined by the candidate SNP and a top ranking SNP. The spine score is the mean of all LD values of the spine between the SNPs $s_x$ and $s_y$:

$$Sc_{spine} = \frac{1}{2(y - x) - 1}\left( \sum_{j=x+1}^{y} D'_{xj} + \sum_{i=x+1}^{y-1} D'_{iy} \right).$$

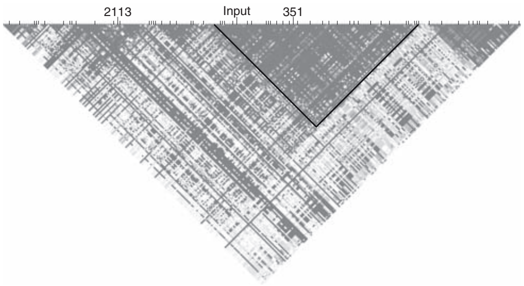


**Figure 3.** Example of a linkage disequilibrium block. Given an input SNP, we compute its linkage disequilibrium block (delimited by dark lines), and then look for top ranking SNPs in the block (here a SNP ranking as 351).

The triangle score is the mean of all LD values of the inner triangle between the SNPs $s_x$ and $s_y$:

$$Sc_{triangle} = \frac{2}{(y-x)(y-x+1)}\left(\sum_{i=x+1}^{y-2}\sum_{j=i+1}^{y-1} D'_{ij}\right).$$

A block score is the sum of the spine score and the triangle score:

$$Sc_{block} = Sc_{spine} + Sc_{triangle}.$$

## RESULTS

We first use data from allelic imbalance sequencing (31) to test our SNP scoring method and to compare our method with existing ones. Then we use two different GWAS data sets to evaluate the mapping method. Finally, we show that the method can find known altered miRNA targets associated with disease.

### Scoring method predicts effects of mirSNPs

Kim and Bartel (31) used allelic imbalance sequencing to measure for three miRNAs, *in vivo* miRNA-directed repression at polymorphic target sites in mice. They provide allelic ratios (target versus non-target allele) $AR = \frac{|target\ allele|}{|non\ target\ allele|}$ for 65 SNPs in 3′-UTRs that create or disrupt miRNA target sites in tissues expressing ($AR_E$) and not expressing ($AR_{NE}$) the considered miRNA. We used 47 of these SNPs (those that have both allelic ratios $AR_E$ and $AR_{NE}$) to test our method. For each of these 47 SNPs, we computed miRNA regulation scores for the target allele $S_T$ and non-target allele $S_{NT}$. We compared the difference of our scores between the two alleles $\Delta S = S_T - S_{NT}$ with the difference of logarithms of allelic ratios $\Delta AR = \log_2(AR_{NE}) - \log_2(AR_E)$ (Figure 4) and found a clear and significant correlation (Pearson's correlation *P*-value 0.0025, Spearman's rank correlation *P*-value 0.00019).

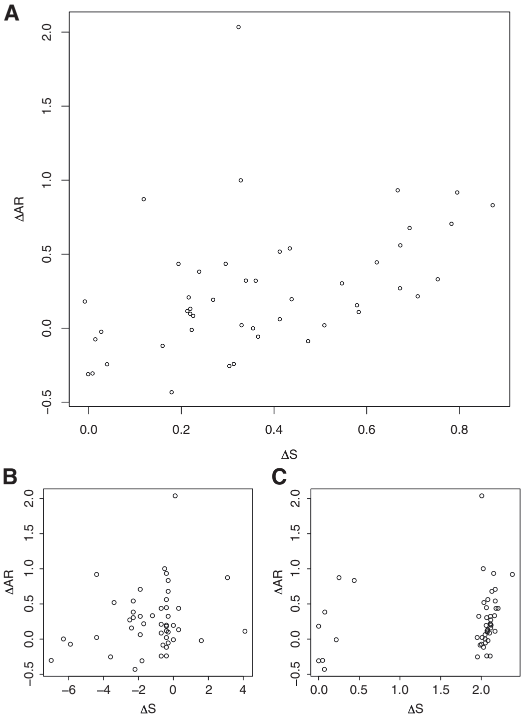


**Figure 4.** Predicted SNP effects correspond with observed effects. Correlation between the measured allelic ratio $\Delta AR$ and (**A**) the difference of our predicted allelic scores $\Delta S$ (with transformation), (**B**) MFE differences, and (**C**) TargetScan score differences (without transformation, but where the minimum TargetScan value represents the score for sequences without predicted target sites). See Table 1 for correlations and *P*-values.

In comparison, using MFE given by RNAhybrid 2.1 (32) to predict SNP effects gave insignificant correlations, whereas using TargetScan 5.0 context scores (13) (computed without taking conservation into account) gave significant but lower correlation (Table 1). Furthermore, our normalization method could improve the correlation based on TargetScan scores.

This result suggests that our scoring method for SNP effects fits data from allelic imbalance sequencing better than TargetScan context scores (13) or changes in MFE [for example, used in (17)]. Our method therefore appears to be the best choice for predicting effects of SNPs in microRNA target sites.

## ANALYSIS OF GWAS DATA

To generate a list of candidate SNPs involved in miRNA-based regulation, we computed differences of scores for all 3′-UTR haplotypes for all coding genes (UCSC RefSeq Genes hg18) and all miRNAs (from miRBase 13.0). Specifically, we analysed mRNAs that had more than 1 haplotype in their 3′-UTR (12 808 of the 26 963 coding transcripts) according to the CEU population from HapMap. Of the 12 808*698 = 89 39 984 mRNA/miRNA pairs, 396 851 had at least one haplotype score that differed from the other haplotype scores of the

**Table 1.** Correlations between the measured allelic ratio $\Delta AR$ and predicted SNP effects from several methods

| Method | Pearson's corr. | | Spearman's corr. | |
|---|---|---|---|---|
| | coeff. | *P*-value | coeff. | *P*-value |
| SVM (raw scores) | 0.383 | 0.0079 | 0.507 | 0.00033 |
| SVM (w/ transformation) | 0.431 | 0.0025 | 0.524 | 0.00019 |
| SVM (w/ transf, w/o 1 outlier) | 0.562 | $4.8*10^{-5}$ | 0.548 | 0.00010 |
| MFE (no helix constraint) | 0.223 | 0.1324 | 0.177 | 0.2345 |
| MFE (helix constraint 2–7) | 0.124 | 0.405 | 0.084 | 0.5736 |
| TargetScan (raw scores) | 0.168 | 0.2582 | 0.394 | 0.0062 |
| TargetScan (w/ transformation) | 0.299 | 0.0409 | 0.413 | 0.0039 |

same mRNA/miRNA pair. As explained in the methods, the haplotype score distribution has an exponential and a uniform part. Consequently, differences of scores also have a distribution with an exponential part, describing small differences in miRNA targeting. We used a threshold of 0.15 to filter out the exponential part. Of the 396 851 mRNA/miRNA pairs (which correspond to 401 983 $\Delta S$ values, as several mRNAs had several haplotype score differences), 55 707 pairs (60 751 $\Delta S$ values) had at least one $\Delta S > 0.15$. We selected the SNPs that generated a difference in score $\Delta S > 0.15$ as candidate SNPs (18 325 SNPs).

To further analyse the candidate mirSNPs, we mapped the mirSNPs to the breast cancer GWAS from CGEMS, as described in the methods. One would usually choose a high $T$ threshold as parameter for the mapping method to identify blocks with high LD. We chose $T = 0$, however, to have data with low LD to analyse the block score variation in relation to the SNP and GWAS scores, as the block scores quantify the link between the candidate mirSNPs and the GWAS SNPs. We computed block scores for each pair of candidate SNP and top ranking SNP detected by the mapping method.

Top-ranking SNPs are likely in strong LD with their causative SNP. Consequently, we would expect that if mirSNPs are a significant factor behind the top-ranking CGEMS SNPs, high $\Delta S$ scores would be enriched among the highest scoring blocks. Since a candidate SNP can have several corresponding $\Delta S$ due to several miRNAs and transcripts, we assigned to each SNP its maximum $\Delta S$ value: $\Delta S_M$. To test whether an increase in block score threshold between top-ranking SNPs and candidate SNPs causes any shift in the $\Delta S_M$ distribution, we computed the probability density of $\Delta S_M$ for different subsets of SNPs. These subsets were defined by a block score greater than a threshold, starting from all block scores and gradually reducing to only the best ones.

Figure 5 shows for SNPs mapped to the 2112 top-ranking CGEMS SNPs, the distributions of $\Delta S_M$ (from 0.15 to 1) for several subsets of SNPs based on different block score thresholds. The distributions show a shift of the main peak at $\Delta S = 0.33$ to $\Delta S = 0.53$ as the block score threshold increases. This shift is consistent with mirSNPs being significant causative factors behind the top-ranking CGEMS SNPs.

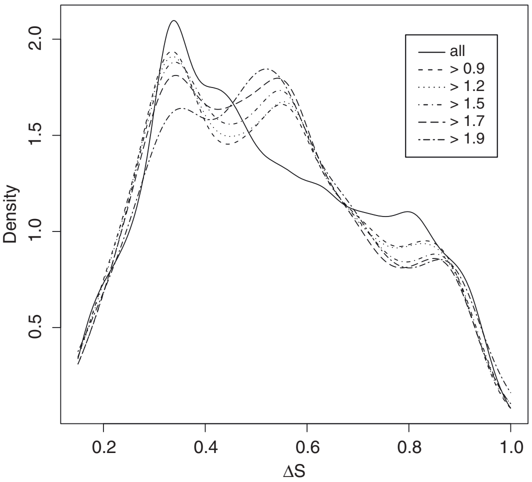


**Figure 5.** Distribution of mirSNP scores $\Delta S_M$ for SNPs mapped to high-ranking SNPs from the CGEMS breast cancer GWAS. $\Delta S_M$ is the maximum difference of scores for each SNP, where the scores are normalized scores from the SVM. Each curve shows the distribution for SNPs that have a block score greater than a given threshold. 'All' refers to $\Delta S_M$ of all SNPs. '>0.9' refers to $\Delta S_M$ of SNPs that have a block score >0.9 with one of the 2112 top-ranking CGEMS SNPs. The peak at 0.33 is decreasing as the block score threshold increases, whereas the peak at 0.53 is increasing with the block score threshold.

We would also expect that the shift will be less pronounced if we consider more candidate SNPs (by using a higher rank threshold on GWAS SNPs), as these SNPs will likely have a higher proportion of false positives. We therefore looked at different top-ranking thresholds to check that as the top-ranking threshold increases, the shift occurs later and later in terms of block score threshold. Figure 6A–D show 3D plots for top-ranking thresholds 528, 1056, 2112, and 4224. As in Figure 5, the plots show a shift of the main peak at $\Delta S = 0.33$ to $\Delta S = 0.53$ as the block score threshold increases.

The lower part of the plots shows all $\Delta S_M$ for all block scores—the background distribution of $\Delta S_M$ scores without taking LD into account. Increasing the block score threshold removes mirSNPs that are not linked to breast cancer-associated GWAS marker SNPs, thereby increasing the proportion of candidate mirSNPs that are associated with breast cancer. The shift in $\Delta S_M$ towards the right for high block score thresholds therefore shows that mirSNPs associated with breast cancer have a stronger effect on miRNA targeting than have the background of all mirSNPs.

As expected, increasing the threshold on top-ranking GWAS SNPs results in the shift occurring later and later on the *y*-axis. Using a higher top-ranking threshold gives a bigger proportion of false positive SNPs, whereas in contrast, a higher block score threshold gives a smaller proportion of false positives. Consequently, to compensate for the additional false positive SNPs that were added when increasing the rank threshold, a higher

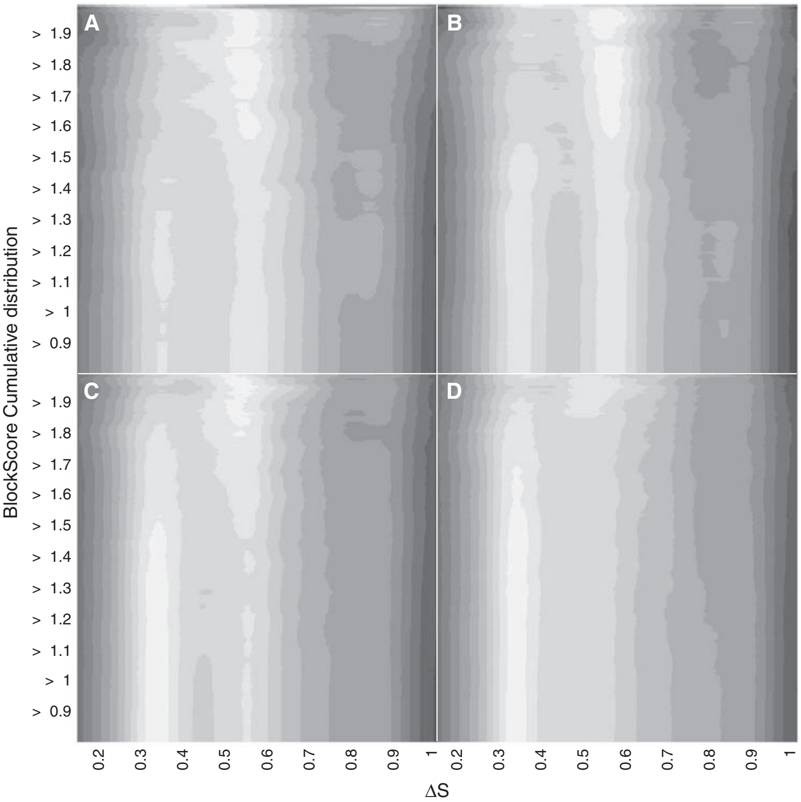


**Figure 6.** Distributions of $\Delta S_M$ for SNPs mapped to different numbers of high-ranking SNPs from the CGEMS breast cancer GWAS. The distributions vary with the number of candidate SNPs and block score thresholds. The graphs show $\Delta S_M$ on the *x*-axis (range [0.15, 1]), complementary cumulative distribution of block scores (from all block scores on the bottom, to gradually filtering to the best block scores on the top) on the *y*-axis, and density of $\Delta S_M$ for a given block score threshold (specifically, the distribution of $\Delta S_M$ for SNPs that have a block score > the value on the *y*-axis) on the *z*-axis (in grayscale). Dark grey, light grey and white are respectively low, intermediate, and high-density values. Panels (A), (B), (C) and (D) show 3D plots for top-ranking thresholds 528, 1056, 2112 and 4224, respectively. The plots show a shift of the main peak at $\Delta S_M = 0.33$ to $\Delta S_M = 0.53$, as the block score threshold increases.

block score threshold is needed to observe the shift in $\Delta S$. These results indicate a link between high $\Delta S$ and high-block score top-ranking SNPs. Furthermore, the analyses give a good overview of how our predicted scores $\Delta S$ fit some GWAS data and show that our approach can identify SNPs in regulatory elements that may be causal in disease.

Using TargetScan's context scores (13) computed for all 3′-UTR haplotypes (without considering conservation), gave similar results indicating that the analysis is robust to the choice of prediction method (Supplementary Figures S3 and S4).

We also repeated the analysis on a GWAS for Parkinson's disease. This analysis gave similar results, indicating that the method works with other data sets and diseases (Supplementary Figures S5 and S6). Finally, we analysed the significant trait-associated SNPs from the NHGRI GWAS Catalog (28) and found a similar shift in the $\Delta S$ distribution at very high-block scores between miRSNPs and associated SNPs from caucasian-based studies (Supplementary Figure S7; see Supplementary Table S1 for the list of the best-scoring miRSNPs strongly linked to caucasian-based trait-associated SNPs). This result is consistent with us using Hapmap CEU haplotypes and linkage disequilibrium data for the analysis and indicates that miRSNPs explain some of the trait-associations in the NHGRI GWAS Catalog.

### Disease-related examples

To further evaluate our methodology, we used it to analyse three miRNA/SNPs involved in breast cancer, asthma and Parkinson's disease.

Saetrom *et al.* (33) found that the SNP rs1434536 lies in the target site of the microRNA miR-125b within the gene *BMPR1b*, and is associated with breast cancer. In that study, we used the disease mapping method presented above to map the candidate SNP rs1434536 to the breast cancer GWAS from CGEMS. We computed the LD block of rs1434536, in which we found 5 SNPs that rank within the 500 best in the association study (ranks 67, 79, 291, 409 and 424) out of 528.000 SNPs; the candidate SNP lay in between the SNPs ranked 67 and 79 (Figure 7). The difference of scores for rs1434536 is 0.39. Saetrom *et al.* (33) verified that the SNP affects miR-125b's regulation of *BMPR1b* and verified the SNP's breast cancer association in an independent cohort.

Tan *et al.* (34) found that the SNP rs1063320 is associated with asthma, depending on the mother's disease status. rs1063320 lies in the 3′-UTR of *HLA-G*, and the authors showed that this SNP affects miR-148a, miR-148b and miR-152 targeting of the *HLA-G* gene. They suggested that this altered miRNA targeting increases the risk of asthma.

With our haplotype scoring method run genome-wide, we found 3 SNPs (rs1063320, rs1610696 and rs1707) in the 3′-UTR of *HLA-G* that can affect 28 miRNAs (data not shown). rs1063320 affects 10 miRNAs (data not shown), and its three largest differences of scores are given by the same three miRNAs reported by Tan *et al.* (34): 0.76, 0.78 and 0.81, respectively for miR-148b, miR-148a and miR-152. The other scores range from 0.33 to 0.55, indicating that the three miRNAs are clear candidates.

Wang *et al.* (35) found that the SNP rs12720208 is associated with Parkinson's disease. rs12720208 lies in the 3′-UTR of *FGF20*. They also showed that this SNP has an effect on miR-433 targeting of *FGF20*. They suggested that this altered targeting increases the risk of Parkinson's disease.

We identified two SNPs (rs1721100 and rs12720208) in the 3′-UTR of *FGF20* that can affect four miRNAs (data not shown). The largest difference of scores for this gene is 0.88 and is given by miR-433 at rs12720208—the same miRNA/SNP pair reported by Wang *et al.* (35). One other miRNA scores 0.44 with rs12720208, whereas SNP rs1721100 scores 0.24 and 0.43 with two miRNAs. Consequently, the pair rs12720208/miR-433 seems to be a clear candidate.

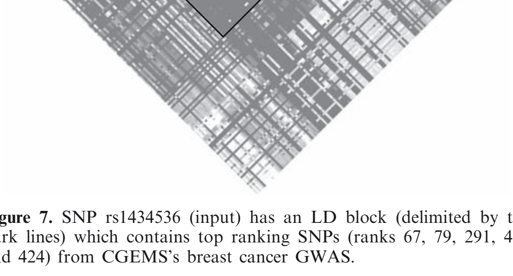


**Figure 7.** SNP rs1434536 (input) has an LD block (delimited by the dark lines) which contains top ranking SNPs (ranks 67, 79, 291, 409 and 424) from CGEMS's breast cancer GWAS.

## DISCUSSION

By evaluating our proposed method on allelic imbalance sequencing data, two different GWAS data sets, and validated mirSNPs, we have demonstrated that our method is useful for identifying potential causative SNPs in miRNA target sites. Specifically, our analyses of the allelic imbalance sequencing data show that our proposed method outperforms existing methods. Although the data set is limited as it contains only 47 SNPs, the data set should be of high quality as it was generated *in vivo* without artificially altering miRNA or target expression (31). Indeed, our results revealed clear differences between the methods. Especially, the method based on changes in predicted miRNA–mRNA hybridization MFE showed poor performance and could not predict the SNPs' effect on miRNA targeting. This result is consistent with overall miRNA–mRNA hybridization in itself being a poor predictor of miRNA targeting and support the model of target site context being essential for miRNA regulation (1).

The basic approach used by many existing tools for detecting SNPs in miRNA target sites looks for SNPs in seed regions of predicted target sites. Seed regions are known to be the most important regions for miRNA targeting efficacy (1). Focusing on seed regions reduces the amount of false positive SNPs predicted to alter miRNA-targeting, but will miss SNPs affecting non-canonical miRNA targeting such as 3′ supplementary sites. This basic method can however be used to filter the mRNA/miRNA pairs that are most likely affected by SNPs. Such filtered SNPs can then subsequently be analysed with our haplotype method.

SNPs outside the seed region can affect miRNA targeting, however, and some existing approaches based on computational RNA–RNA hybridization or thermodynamic calculations consider such SNPs. Our method can also detect SNPs in 3′ supplementary sites, but according to our analyses, such SNPs have a small predicted effect (Supplementary Figure S8). This result is consistent with the observation that conserved 3′ supplementary sites constitute 4.9% of all conserved pairing sites (36). As SNPs affecting seed site pairing have a bigger predicted effect than those affecting other miRNA features, our online database provide allelic sequences for SNPs in target seed sites.

A transcriptome-wide study of interactions between miRNAs and mRNAs estimated that sites with seed mismatches constitute <6.6% of all miRNA target sites (19). By excluding SNPs in mismatch sites, we only miss SNPs that change a mismatch target site (weak) into another mismatch site. Moreover, non-canonical sites appear to have a smaller regulatory effect than canonical target sites have (19). Thus, our method focuses on identifying the SNPs that are most likely to affect and to have the largest effect on miRNA targeting.

Our haplotype scoring method is based on HapMap haplotype data, and only 66% of the SNPs from HapMap have haplotype data. The 34% HapMap SNPs that do not have haplotype data have a very low minimum allele frequency (MAF), usually 0 in the considered hapmap population. Removing low MAF SNPs is an advantage in mapping SNPs to common diseases, resulting in less false positives (false causal SNPs), in a common variant common disease model.

Our haplotype approach also currently only focuses on analysing 3′-UTRs. Although miRNAs can target 5′-UTRs and coding regions, these sites have a limited effect compared to 3′-UTR sites (19,20).

The main advantage of our method compared to existing methods is that we analyse the regulatory effects of all linked genetic variations within regulatory regions, such as 3′-UTRs. Consequently, our method can be used to analyse how SNPs in multiple target sites together contribute to upregulate, downregulate, or compensate each other, through haplotype patterns.

## CONCLUSION

We have presented a tool that aims at identifying the causative variation within regions associated with diseases. Specifically, the tool identifies 3′-UTR SNPs that can affect miRNA targeting and predicts the SNPs' effect on miRNA regulation. Our main result is the SNP effect prediction method. The results suggest that the effect predictions are reliable, compare favourably to existing methods, and can be used to filter and identify causative SNPs.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## FUNDING

Norwegian Research Council; Interagon AS; the Norwegian Functional Genomics Program (FUGE) of the Norwegian Research Council. Funding for open access charge: Norwegian University of Science and Technology.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory Functions. *Cell*, **136**, 215–233.
2. Brennecke,J., Stark,A., Russell,R. and Cohen,S. (2005) Principles of MicroRNA-target recognition. *PLoS Biol.*, **3**, 404–418.
3. Didiano,D. and Hobert,O. (2006) Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat. Struct. Mol. Biol.*, **13**, 849–851.
4. Gaidatzis,D., van Nimwegen,E., Hausser,J. and Zavolan,M. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 248.
5. Saetrom,P., Heale,B.S.E., Snove,O. Jr, Aagaard,L., Alluin,J. and Rossi,J.J. (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res.*, **35**, 2333–2342.
6. Hirschhorn,J. and Daly,M. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.
7. Donnelly,P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, **456**, 728–731.
8. Mishra,P.J. and Bertino,J.R. (2009) MicroRNA polymorphisms: the future of pharmacogenomics, molecular epidemiology and individualized medicine. *Pharmacogenomics*, **10**, 399–416.
9. Borel,C. and Antonarakis,S.E. (2008) Functional genetic variation of human miRNAs and phenotypic consequences. *Mamm. Genome*, **19**, 503–509.
10. Landi,D., Gemignani,F., Naccarati,A., Pardini,B., Vodicka,P., Vodickova,L., Novotny,J., Foersti,A., Hemminki,K. and Canzian,F. (2008) Polymorphisms within micro-RNA-binding sites and risk of sporadic colorectal cancer. *Carcinogenesis*, **29**, 579–584.
11. Sethupathy,P. and Collins,F.S. (2008) MicroRNA target site polymorphisms and human disease. *Trends Genet.*, **24**, 489–497.
12. Bao,L., Zhou,M., Wu,L., Lu,L., Goldowitz,D., Williams,R.W. and Cui,Y. (2007) PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits. *Nucleic Acids Res.*, **35**, D51–D54.
13. Lewis,B., Burge,C. and Bartel,D. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
14. Georges,M., Clop,A., Marcq,F., Takeda,H., Pirottin,D., Hiard,S., Tordoir,X., Caiment,F., Meish,F., Bibe,B. *et al.* (2006) Polymorphic microRNA-target interactions: a novel source of phenotypic variation. *Cold Spring Harb. Symp. Quant. Biol.*, **71**, 343–350.
15. Xie,X., Lu,J., Kulbokas,E., Golub,T., Mootha,V., Lindblad-Toh,K., Lander,E. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
16. Barenboim,M., Zoltick,B.J., Guo,Y. and Weinberger,D.R. (2010) MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. *Hum. Mutat.*, **31**, 1223–1232.
17. Nicoloso,M.S., Sun,H., Spizzo,R., Kim,H., Wickramasinghe,P., Shimizu,M., Wojcik,S.E., Ferdin,J., Kunej,T., Xiao,L. *et al.* (2010) Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. *Cancer Res.*, **70**, 2789–2798.
18. Miranda,K.C., Huynh,T., Tay,Y., Ang,Y.-S., Tam,W.-L., Thomson,A.M., Lim,B. and Rigoutsos,I. (2006) A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.
19. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.-C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
20. Grimson,A., Farh,K.K.-H., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
21. Int HapMap Consortium. In Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
22. Int Human Genome Sequencing Conso. In Lander,E., Linton,L., Birren,B., Nusbaum,C., Zody,M., Baldwin,J., Devon,K., Dewar,K., Doyle,M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
23. Mouse Genome Sequencing Consor. In Waterston,R., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
24. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al.* (2009) The UCSC genome browser database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
25. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
26. Hunter,D.J., Kraft,P., Jacobs,K.B., Cox,D.G., Yeager,M., Hankinson,S.E., Wacholder,S., Wang,Z., Welch,R., Hutchinson,A. *et al.* (2007) A genome-wide association study identifies alleles in

FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genet.*, **39**, 870–874.

27. Maraganore,D., de Andrade,M., Lesnick,T., Strain,K., Farrer,M., Rocca,W., Pant,P., Frazer,K., Cox,D. and Ballinger,D. (2005) High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.*, **77**, 685–693.
28. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
29. Saito,T. and Saetrom,P. (2010) A two-step site and mRNA-level model for predicting microRNA targets. *BMC Bioinformatics*, **11**, 612.
30. Barrett,J., Fry,B., Maller,J. and Daly,M. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
31. Kim,J. and Bartel,D.P. (2009) Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression. *Nat. Biotechnol.*, **27**, 472–477.
32. Rehmsmeier,M., Steffen,P., Hochsmann,M. and Giegerich,R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA-Publ. RNA Soc.*, **10**, 1507–1517.
33. Saetrom,P., Biesinger,J., Li,S.M., Smith,D., Thomas,L.F., Majzoub,K., Rivas,G.E., Alluin,J., Rossi,J.J., Krontiris,T.G. *et al.* (2009) A risk variant in an miR-125b binding site in BMPR1B is associated with breast cancer pathogenesis. *Cancer Res.*, **69**, 7459–7465.
34. Tan,Z., Randall,G., Fan,J., Camoretti-Mercado,B., Brockman-Schneider,R., Pan,L., Solway,J., Gern,J.E., Lemanske,R.F. Jr and Nicolae,D. (2007) Allele-specific targeting of microRNAs to HLA-G and risk of asthma. *Am. J. Hum. Genet.*, **81**, 829–834.
35. Wang,G., van der Walt,J.M., Mayhew,G., Li,Y.-J., Zuechner,S., Scott,W.K., Martin,E.R. and Vance,J.M. (2008) Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein. *Am. J. Hum. Genet.*, **82**, 283–289.
36. Friedman,R.C., Farh,K.K.-H., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

# Supplementary Data: Inferring causative variants in microRNA target sites

Laurent F Thomas, Takaya Saito, and Pål Sætrom

Table 1: miRSNPs linked to trait-associated SNPs from the NHGRI GWAS catalog. The file "GWAScatMiRSNPs.xls" provides miRSNPs that have a maximum $\Delta S$ score greater than 0.45 (middle of the shift; see Supplementary Figure 7) and that have a block score greater than 1.988 (10% best percentile) with trait-associated SNPs from the NHGRI GWAS catalog (http://www.genome.gov/gwastudies; accessed Apr. 18, 2011). The trait-associated SNPs are strictly based on caucasian populations with Northern and Western European ancestry. For each miRSNP, we only provide the miRNA and the target site information for the miRNA that gives the maximum $\Delta S$ score. Other potentially affected miRNAs can be found online in our database (http://www.bigr.medisin.ntnu.no/mirsnpscore/). Within the file, columns A-H describe the following miRSNP information: chromosome, SNP ID, chromosome position in hg18, minimum allele frequency within the CEU Hapmap population, host gene, affected miRNA, $\Delta S$ score quantifying how the SNP can affect the miRNA, and the alleles and their respective target sites. Column I shows the block score between the miRSNP and the associated SNP. Columns J-M describe the following information for the associated SNP reported in the GWAS catalog: SNP ID, chromosome position in hg18, minimum allele frequency within the CEU Hapmap population, and the PubMed IDs that report the SNP as significantly associated with some trait.

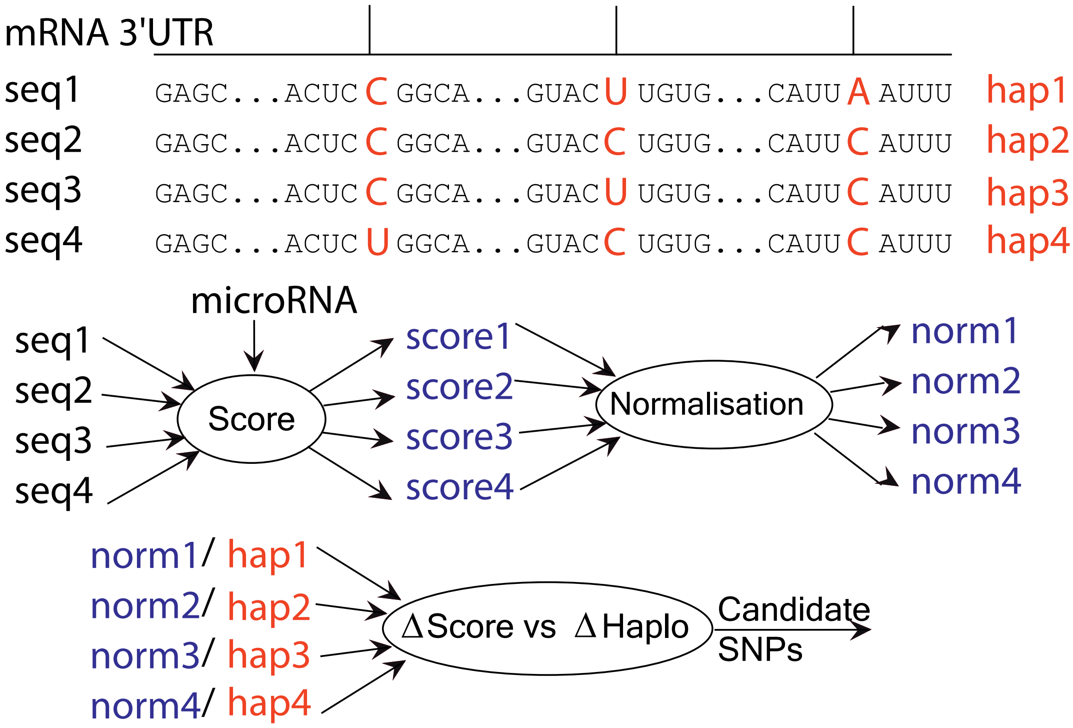

mRNA 3'UTR
seq1 GAGC...ACUC C GGCA...GUAC U UGUG...CAUU A AUUU hap1
seq2 GAGC...ACUC C GGCA...GUAC C UGUG...CAUU C AUUU hap2
seq3 GAGC...ACUC C GGCA...GUAC U UGUG...CAUU C AUUU hap3
seq4 GAGC...ACUC U GGCA...GUAC C UGUG...CAUU C AUUU hap4
microRNA
seq1
seq2
seq3
seq4
Score
score1
score2
score3
score4
Normalisation
norm1
norm2
norm3
norm4
norm1/ hap1
norm2/ hap2
norm3/ hap3
norm4/ hap4
ΔScore vs ΔHaplo
Candidate SNPs

Figure 1: Flowchart: from haplotypes to candidate SNPs in three steps. The example shows a 3′UTR that contains three SNPs that form four haplotypes (in red). Given a miRNA, we first use a miRNA target prediction tool to analyse and score each of the four complete 3′UTR haplotype sequences. Second, we normalise the scores to compare them to each other. Third, we compare the pairs score/haplotypes to infer candidate SNPs (see Methods and Supplementary Figure 2).

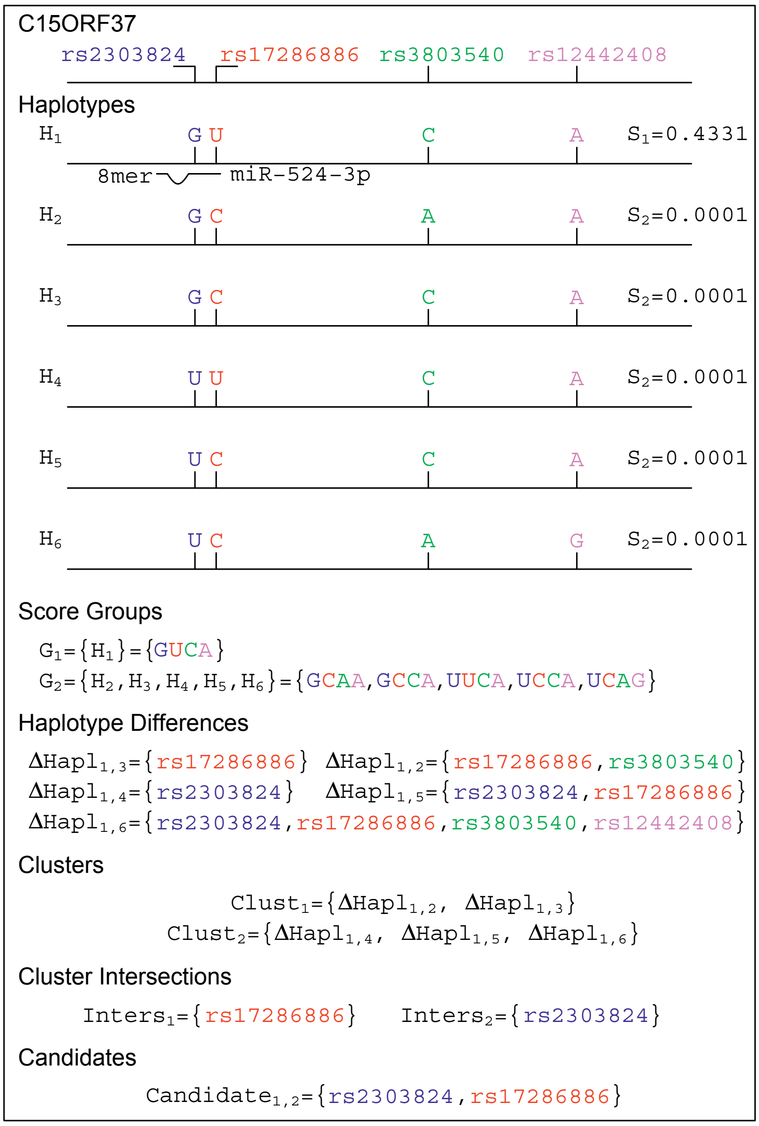

C15ORF37
rs2303824 rs17286886 rs3803540 rs12442408
Haplotypes
H1 G U C A S1=0.4331
8mer miR-524-3p
H2 G C A A S2=0.0001
H3 G C C A S2=0.0001
H4 U U C A S2=0.0001
H5 U C C A S2=0.0001
H6 U C A G S2=0.0001
Score Groups
G1={H1}={GUCA}
G2={H2,H3,H4,H5,H6}={GCAA,GCCA,UUCA,UCCA,UCAG}
Haplotype Differences
ΔHapl1,3={rs17286886} ΔHapl1,2={rs17286886,rs3803540}
ΔHapl1,4={rs2303824} ΔHapl1,5={rs2303824,rs17286886}
ΔHapl1,6={rs2303824,rs17286886,rs3803540,rs12442408}
Clusters
Clust1={ΔHapl1,2, ΔHapl1,3}
Clust2={ΔHapl1,4, ΔHapl1,5, ΔHapl1,6}
Cluster Intersections
Inters1={rs17286886} Inters2={rs2303824}
Candidates
Candidate1,2={rs2303824,rs17286886}

Figure 2: Scoring SNPs in miRNA target sites. The two SNPs rs2303824 and rs17286886 lie in one target site of miR-524-3p in the 3′UTR of the gene *C15ORF37*. There are two other SNPs in this 3′UTR: rs3803540 and rs12442408. Those four SNPs make six haplotypes in the CEU HapMap population. Haplotype $H_1$ has one target site and a score $S_1$. The five other haplotypes do not have any target sites for this miRNA and have identical scores $S_2$. We group haplotypes by score into two groups, where $G_1$ only consists of haplotype $H_1$. Then, we look at the differences between haplotypes from groups $G_1$ and $G_2$; *i.e.* differences between $H_1$ and $H_2$, $H_1$ and $H_3$, $H_1$ and $H_4$, $H_1$ and $H_5$, and $H_1$ and $H_6$. We cluster these haplotype differences into two clusters, because the intersection of all the haplotype differences is empty. The clusters are made so that the intersection within each cluster is not empty. Finally, we take the intersection within each cluster and merge the two resulting clusters, which gives the two SNPs that lie within the target site. These two SNPs are the candidates that explain the predicted miRNA-targeting differences between the haplotypes.

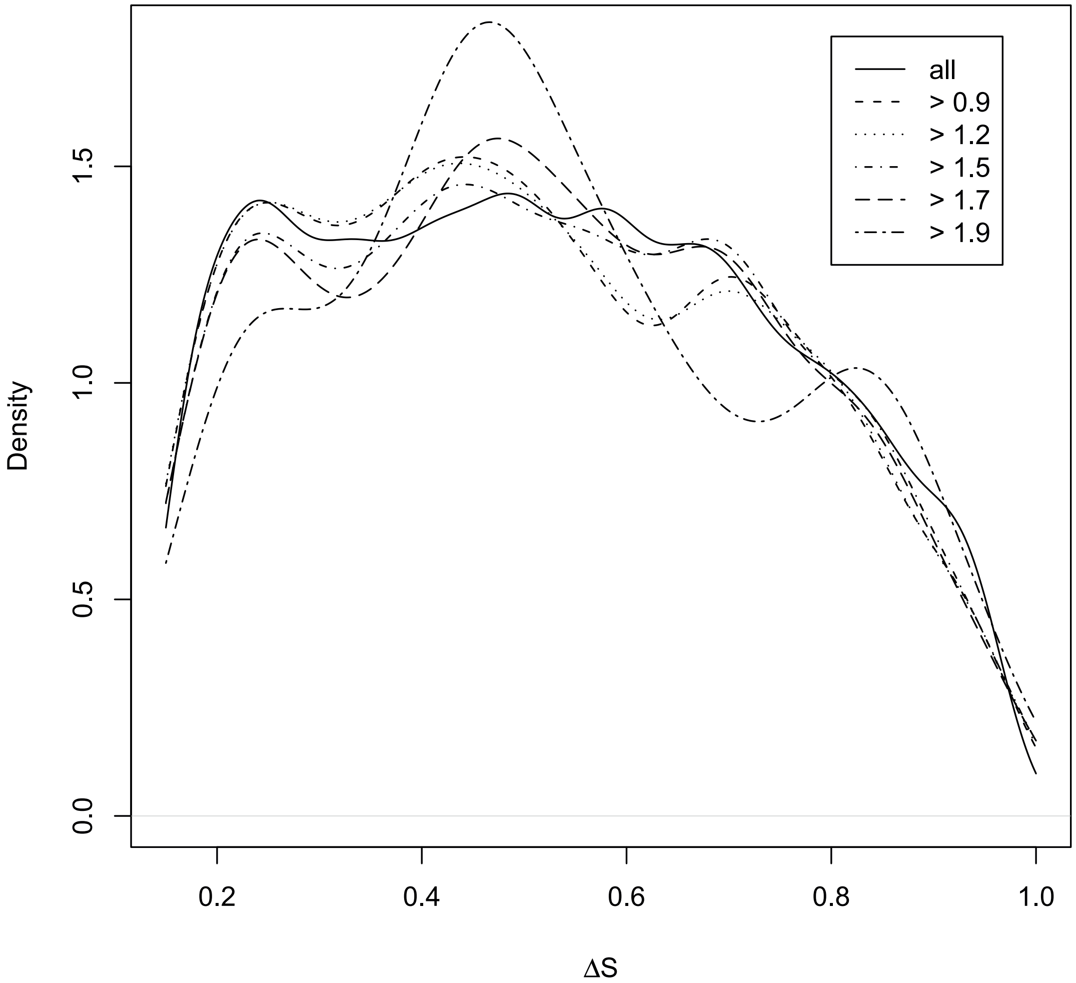


Figure 3: Distribution of TargetScan-based mirSNP scores $\Delta S_M$ for SNPs mapped to high-ranking SNPs from the CGEMS breast cancer GWAS. $\Delta S_M$ is the maximum difference of scores for each SNP, where the scores are normalised context scores from TargetScan. Each curve shows the distribution for SNPs that have a block score greater than a given threshold. 'All' refers to $\Delta S_M$ of all SNPs. '$> 0.9$' refers to $\Delta S_M$ of SNPs that have a block score $> 0.9$ with one of the 2112 top-ranking CGEMS SNPs. As with the SVM prediction tool, we see a shift from lower scores ($\Delta S = 0.25$) to higher scores ($\Delta S = 0.46$) as the block score threshold increases.

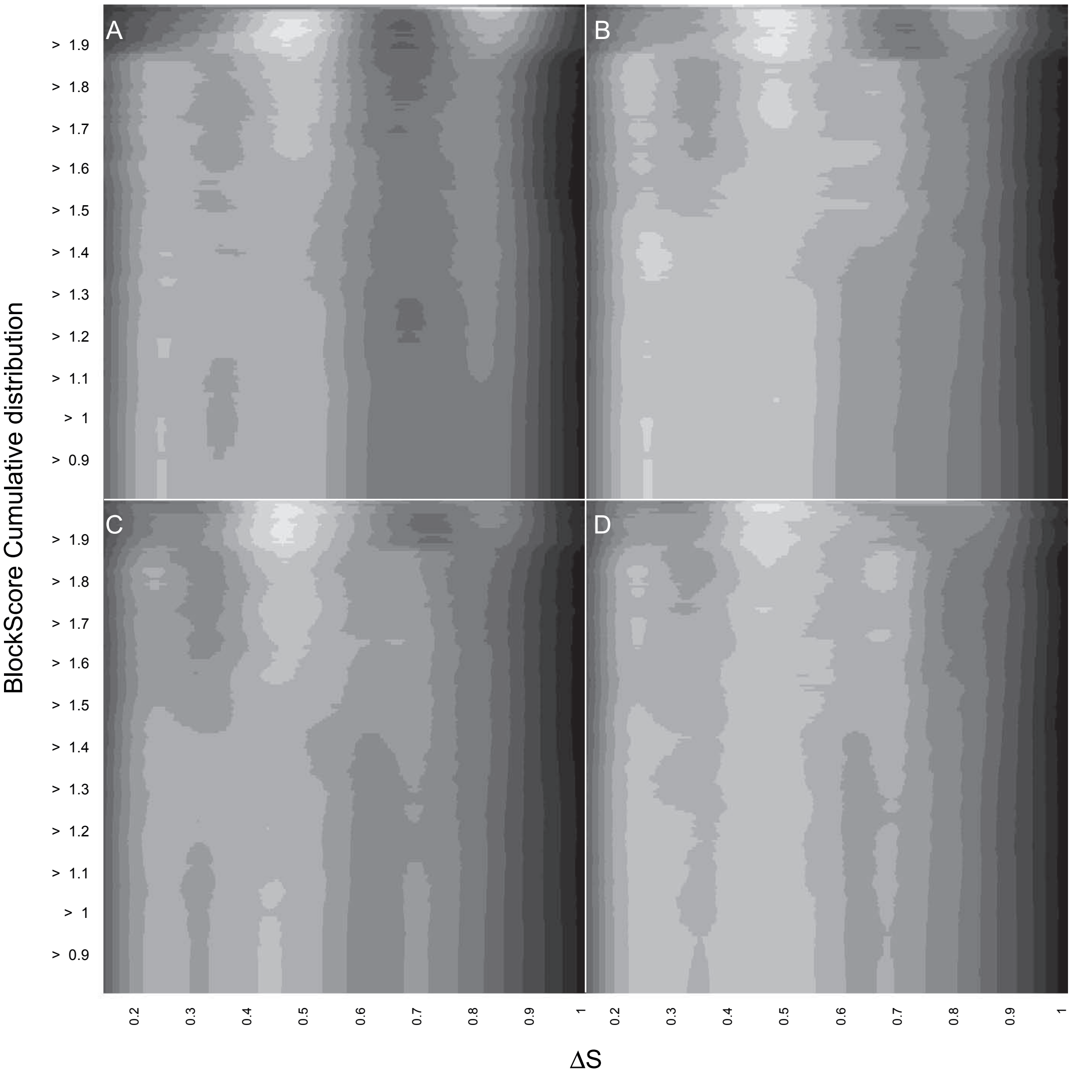


Figure 4: Distributions of TargetScan-based $\Delta S_M$ for SNPs mapped to different numbers of high-ranking SNPs from the CGEMS breast cancer GWAS. The distributions vary with the number of candidate SNPs and block score thresholds. The graphs show $\Delta S_M$ on the x-axis (range $[0.15, 1]$), complementary cumulative distribution of block scores (from all block scores on the bottom, to gradually filtering to the best block scores on the top) on the y-axis, and density of $\Delta S_M$ for a given block score threshold (specifically, the distribution of $\Delta S_M$ for SNPs that have a block score $>$ the value on the y-axis) on the z-axis (in grayscale). Dark gray, light gray, and white are respectively low, intermediate, and high density values. Panels (A), (B), (C), and (D) show 3-dimensional plots for top-ranking thresholds 528, 1056, 2112, and 4224, respectively. As with the SVM prediction tool, these plots based on TargetScan scores show a shift from lower scores ($\Delta S = 0.25$) to higher scores ($\Delta S = 0.46$) as the block score threshold increases.

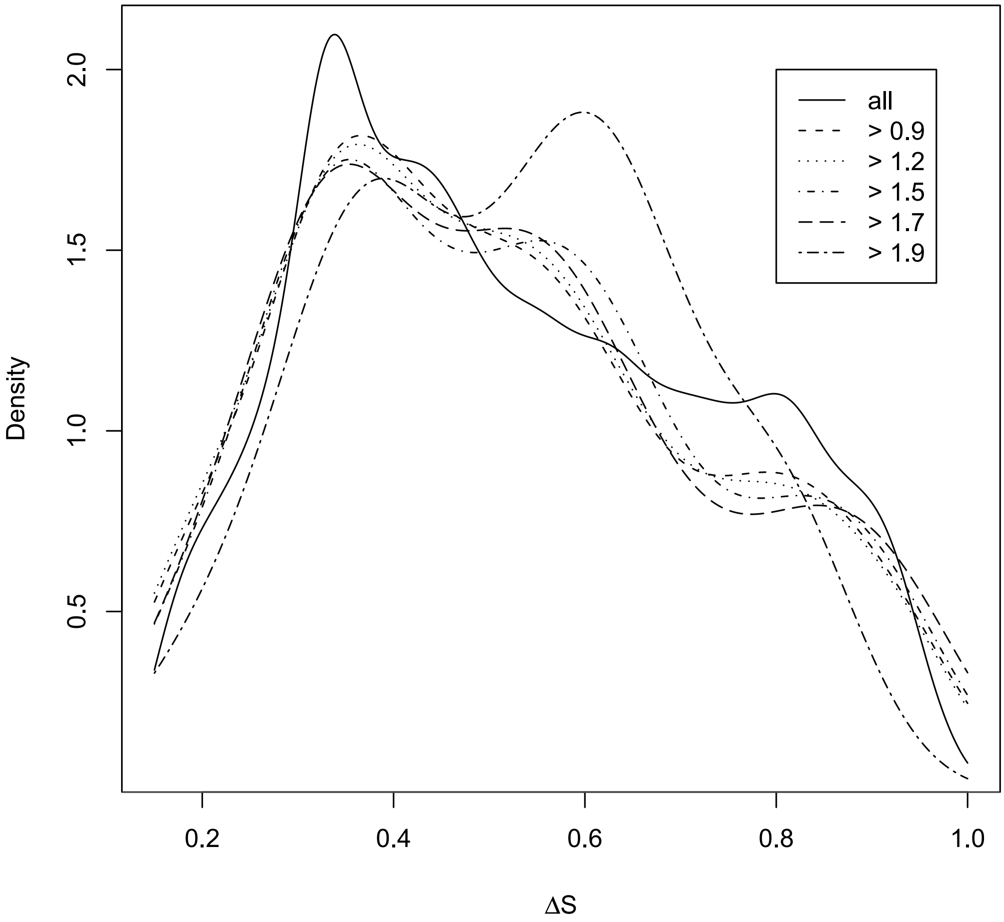


Figure 5: Distribution of SVM-based mirSNP scores $\Delta S_M$ for SNPs mapped to one of the 217 top-ranking SNPs from a Parkinson's disease GWAS (p-values $< 0.001$); see Supplementary Figure 3 for details. As with the CGEMS GWAS, we see a shift from lower scores ($\Delta S = 0.34$) to higher scores ($\Delta S = 0.60$), as the block score threshold increases.

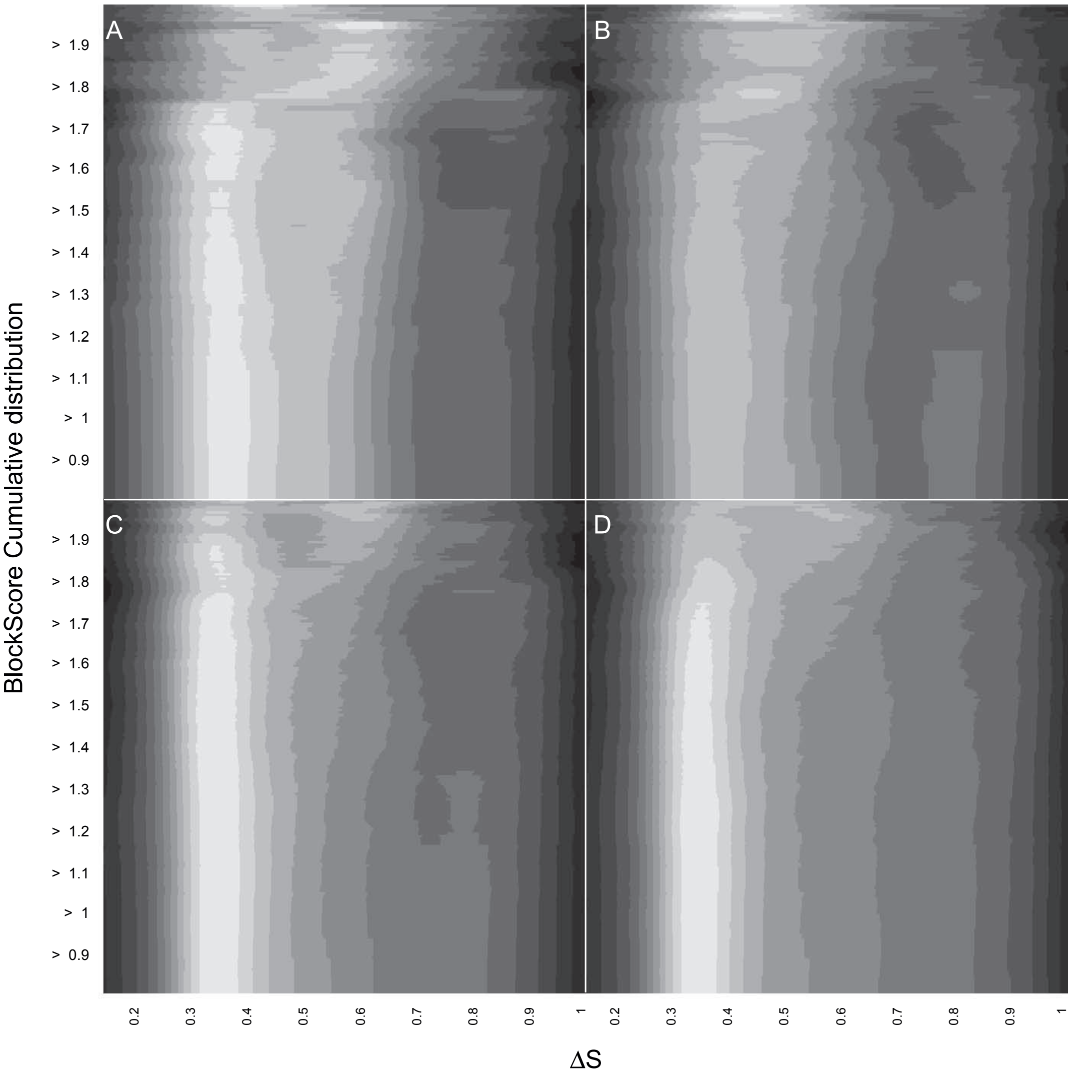


Figure 6: Distribution of SVM-based $\Delta S_M$ for SNPs mapped to different numbers of high-ranking SNPs from a Parkinson's disease GWAS; see Supplementary Figure 4 for details. Panels (A), (B), (C), and (D) show 3-dimensional plots for top-ranking thresholds 217, 413, 755, and 1517, respectively (corresponding to p-values $<$ 0.001, 0.002, 0.004, and 0.008 from a Parkinson's disease GWAS). As with the CGEMS GWAS, these plots show a shift from lower scores ($\Delta S = 0.34$) to higher scores ($\Delta S = 0.60$), as the block score threshold increases.

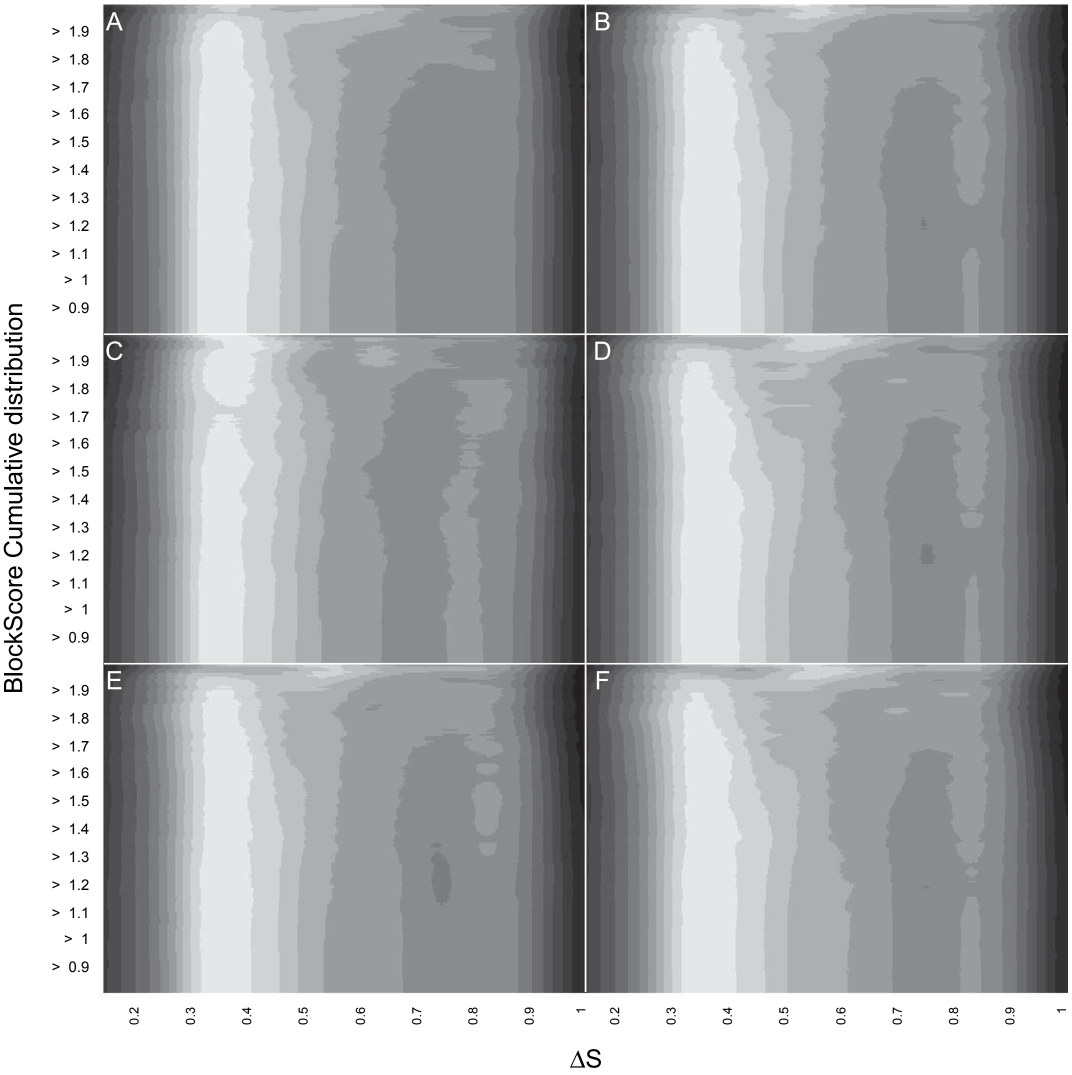


Figure 7: Distribution of SVM-based $\Delta S_M$ for SNPs mapped to 4304 trait-associated SNPs from the NHGRI GWAS Catalog; see Supplementary Figure 4 for details. Panels (A), (B), (C), (D), (E), and (F) show 3-dimensional plots for, respectively, all associated SNPs, associated SNPs based on caucasian populations, associated SNPs based on non-caucasian populations, associated SNPs strictly based on caucasian populations (study used only a caucasian population), associated SNPs based on caucasian populations with Northern and Western European ancestry (NWE), and associated SNPs strictly based on NWE (study used only a NWE population). As with the CGEMS GWAS, these plots show a shift from lower scores ($\Delta S = 0.35$) to higher scores ($\Delta S = 0.55$), with caucasian-based studies (panels A, B, D, E, and F). In the panel C, the distribution stays similar to background. This is because both block scores and miRSNPs are based on the CEU Hapmap population (Utah residents with Northern and Western European ancestry).

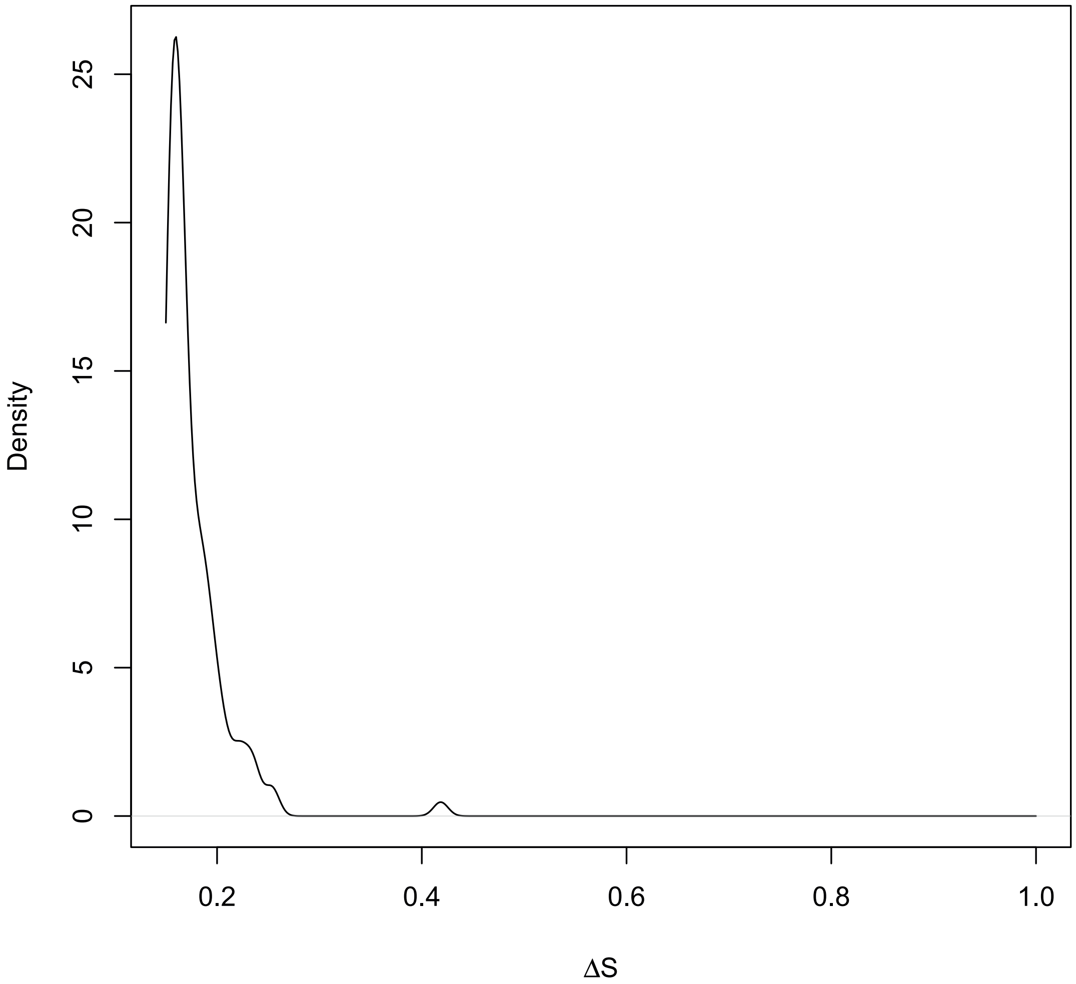


Figure 8: Distribution of score differences for mRNA/miRNA pairs that have no SNP in their seed sites, but one SNP in the $3'$supplementary region of the target site. The distribution is completely shifted to the left, showing that SNPs in $3'$-supplementary sites have a small predicted effect on miRNA-based regulation. The distribution is based on normalised scores from the SVM prediction tool.

# Paper II

Published OnlineFirst September 8, 2009; DOI:10.1158/0008-5472.CAN-09-1201

# A Risk Variant in an miR-125b Binding Site in *BMPR1B* Is Associated with Breast Cancer Pathogenesis

**Pål Sætrom,[1,2,3] Jacob Biesinger,[4] Sierra M. Li,[5] David Smith,[5] Laurent F. Thomas,[1,3] Karim Majzoub,[6] Guillermo E. Rivas,[7] Jessica Alluin,[6] John J. Rossi,[6] Theodore G. Krontiris,[7] Jeffrey Weitzel,[8] Mary B. Daly,[9] Al B. Benson,[11] John M. Kirkwood,[12] Peter J. O'Dwyer,[10] Rebecca Sutphen,[13] James A. Stewart,[14] David Johnson,[15] and Garrett P. Larson[7]**

Departments of [1]Cancer Research and Molecular Medicine and [2]Computer and Information Science, Norwegian University of Science and Technology; [3]Interagon AS, Laboratoriesenteret, Trondheim, Norway; [4]Department of Integrative Biology, Brigham Young University, Provo, Utah; [5]Department of Information Sciences, City of Hope National Medical Center; [6]Division of Molecular Biology, and Departments of [7]Molecular Medicine and [8]Clinical Cancer Genetics, Beckman Research Institute of the City of Hope, Duarte, California; [9]Department of Population Science, Fox Chase Cancer Center; [10]Division of Hematology/Oncology, Department of Medicine, University of Pennsylvania Cancer Center, Philadelphia, Pennsylvania; [11]Division of Hematology/Oncology, Department of Medicine, Robert J. Lurie Comprehensive Cancer Center, Northwestern University School of Medicine, Chicago, Illinois; [12]Division of Hematology/Oncology, Department of Medicine, University of Pittsburgh Cancer Center, Pittsburgh, Pennsylvania; [13]Interdisciplinary Oncology Program, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida, Tampa, Florida; [14]University of Wisconsin Comprehensive Cancer Center, University of Wisconsin School of Medicine, Madison, Wisconsin; and [15]Division of Hematology/Oncology, Department of Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee

## Abstract

**MicroRNAs regulate diverse cellular processes and play an integral role in cancer pathogenesis. Genomic variation within miRNA target sites may therefore be important sources for genetic differences in cancer risk. To investigate this possibility, we mapped HapMap single nucleotide polymorphisms (SNP) to putative miRNA recognition sites within genes dysregulated in estrogen receptor–stratified breast tumors and used local linkage disequilibirum patterns to identify high-ranking SNPs in the Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer genome-wide association study for further testing. Two SNPs, rs1970801 and rs11097457, scoring in the top 100 from the CGEMS study, were in strong linkage disequilibrium with rs1434536, an SNP that resides within a miR-125b target site in the 3′ untranslated region of the bone morphogenic receptor type 1B (*BMPR1B*) gene encoding a transmembrane serine/threonine kinase. We validated the CGEMS association findings for rs1970801 in an independent cohort of admixture-corrected cases identified from families with multiple case histories. Subsequent association testing of rs1434536 for these cases and CGEMS controls with imputed genotypes supported the association. Furthermore, luciferase reporter assays and overexpression of miR-125b–mimics combined with quantitative reverse transcription-PCR showed that *BMPR1B* transcript is a direct target of miR-125b and that miR-125b differentially regulates the C and T alleles of rs1434536. These results suggest that allele-specific regulation of *BMPR1B* by miR-125b explains the observed disease risk. Our approach is general and can help identify and explain the mechanisms behind disease association for alleles that affect miRNA regulation.** [Cancer Res 2009;69(18):7459–65]

## Introduction

MicroRNAs (miRNA) are a recently discovered class of short noncoding RNA genes that act posttranscriptionally as negative regulators of gene expression and play fundamental roles in cell growth, apoptosis, hematopoietic lineage differentiation, and differentiation (1, 2). Functional studies indicate that changes in miRNA expression patterns might underlie human pathologies, including malignancies (3, 4). In addition, variations in miRNA target sites mediated by single nucleotide polymorphisms (SNP) may be associated with human cancers (5, 6).

Gene expression profiling studies have identified specific signatures for breast cancer and are used to guide patient treatment with both the Oncotype Dx and Mammaprint tests in use clinically (7, 8). We previously described a meta-analysis of multiple independent breast cancer RNA expression studies whereby a unified set of dysregulated genes was identified in estrogen receptor (ER)+ and ER− tumors. The identification of germline variations in elements controlling RNA expression (i.e., transcription factor or miRNA recognition sites) may provide clues as to the mechanistic basis for the observed variations in gene expression patterns.

Genome-wide association studies (GWAS) have been used in many common diseases to identify SNPs associated with disease (9, 10). To date, four independent studies examining breast cancer patients have identified multiple SNPs associated with disease (10–13). Although some association signals seem universal in multiple studies (i.e., several SNPs within *FGFR2*), often these studies also yield vastly differing collections of SNPs associated with disease perhaps owing to differences in study design. Although many disease-associated SNPs are nongenic, and thus their contribution to disease pathogenesis is unclear, many are likely to reside in gene regulatory elements that may influence gene expression patterns observed in tumors.

We describe an integrative genomic approach leveraging gene expression patterns, miRNA targeting, breast cancer GWAS data, and biological testing to identify a disease-associated SNP in the

**Note:** Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

**Authors' contributions:** G.P. Larson conceived and supervised the study. P. Sætrom, J. Biesinger, and L.F. Thomas performed bioinformatics analyses and software development. G.P. Larson and G.E. Rivas performed genotyping. G.P. Larson, S.M. Li, and D. Smith carried out admixture and statistical analyses. G.P. Larson, K. Mazjoub, J. Alluin, and J.J. Rossi performed biological testing. T.G. Krontiris, J. Weitzel, M.B. Daly, A.B. Benson, J.M. Kirkwood, P. O'Dwyer, R. Sutphen, J.A. Stewart, and D. Johnson contributed patient materials. G.P. Larson and P. Sætrom wrote the manuscript and received important contributions from other authors. All authors read and approved the final manuscript.

**Requests for reprints:** Garrett P. Larson, Department of Molecular Medicine, Beckman Research Institute of the City of Hope, Duarte, CA 91010. Phone: 626-359-8111; Fax: 626-930-5330; E-mail: glarson@coh.org.

doi:10.1158/0008-5472.CAN-09-1201

Published OnlineFirst September 8, 2009; DOI:10.1158/0008-5472.CAN-09-1201

3′ untranslated region (UTR) of *BMPR1B* gene. To identify this SNP, we mapped a set of reference SNPs from the HapMap project to prospective miRNA target sites located in the 3′UTRs of a previously identified set of dysregulated ER+ and ER− genes (14). An analysis of local linkage disequilibrium (LD) patterns surrounding these SNPs identified one SNP (rs1434536) in strong LD with two SNPs showing a high degree of association in the Cancer Genetic Markers of Susceptibility (CGEMS) study. We replicated this association in an independent set of cases identified from families with multiple case histories and common CGEMS controls after controlling for population stratification with ancestry informative markers (AIM). We provide strong support that allelic variation at rs1434536 influences interactions with miR-125b leading to differences in *BMPR1B* expression levels. The approach described is generally applicable and provides clues to the role *cis*-acting allelic variation plays in tumor gene expression patterns via interactions with the miRNA machinery in disease pathogenesis.

## Materials and Methods

**Mapping SNPs to miRNA targets.** Our input data consisted of 275 candidate genes previously identified as constituting the top 1% of genes dysregulated in ER+ and ER− (130 and 145 genes, respectively) breast cancer tumors (14) and their annotated 3′UTR sequences from the University of California at Santa Cruz Table Browser (National Center for Biotechnology Information Build 36.1), mature human miRNA sequences from miRBase[16], and SNPs from HapMap.[17] Using custom python scripts, we (*a*) identified all unique 7mer seeds (nucleotides 2–8) within the mature miRNA sequences, (*b*) identified all seed sites—that is, locations with perfect reverse complementarity to a 7mer seed—within the candidate genes' 3′UTRs, (*c*) identified all HapMap SNPs that mapped to one of the seed site locations, and (*d*) removed all SNPs that had no reported minor allele in any HapMap population.

**Description of study populations.** Four hundred fifty-nine probands from a breast cancer–affected sibling pair cohort were recruited from a multi-institutional study [Eastern Cooperative Oncology Group (ECOG) Cancer in Sibling Study, E1Y97] under protocols approved by the respective Institutional Review Boards at each institution. The mean age of diagnosis for probands was 55 y (range, 16–87 y) and disease status was verified by pathology reports for 96.5% of cases (443 of 459; Supplementary Materials and Methods). We collected self-reported ethnicity data for both maternal and paternal grandparents from 78% (356) of our cases. CGEMS patients consisted of 1,142 controls and 1,145 cases of postmenopausal breast cancer and were gathered from the Nurses Health Study as described previously (10, 15). Self-reported ethnicity information was unavailable for these individuals.

**Genotyping and quality control.** DNA samples were prepared as previously described from peripheral leukocytes (16). SNP genotyping was performed using Sequenom MassARRAY genotyping technology and iPLEX chemistry according to manufacturer's instructions (17). AIMs were developed into two multiplex assays (Supplementary Table S1) as defined by the 64 $I_n4$ AIMs described by Kosoy and colleagues (18). Genotyping success ranged from 95.9% to 97.8% for the three association SNP in our cases. Patient samples were genotyped and samples demonstrating <80% completion rate (46 of 459) were subjected to a second round of genotyping. Quality control metrics for our cases included a minimum of 80% genotyping success, whereas SNPs with completion rates <90% were discarded. After two rounds of genotyping, four cases and nine AIMs were discarded from further analysis, having not met quality control metrics.

**Population structure analysis and association testing.** For admixture analysis, we used 45 AIMs and combined our cases (455), the CGEMS controls (1, 142) and seeded the analysis with a training set of 270 HapMap reference samples (CEU, YRI, and CHB+JTP) to perform STRUCTURE analysis with k = 3 populations (Supplementary Fig. S1). We observed general agreement between our patient's self-reported ethnicity and genetic ancestry as determined by our AIMs, although rarely a patient's self-identified ancestry was at odds with the calculated CEU ancestral component. In these instances, we relied on STRUCTURE results to determine genetic ancestry. For association testing, each SNP was analyzed using a logistic regression model where odds ratios (OR) are estimated for homozygous and heterozygous states of the indicated cases and CGEMS controls. For the causative SNP rs1434536, we directly genotyped our cases and imputed genotypes from CGEMS controls using HapMap CEU reference individuals (Supplementary Materials and Methods). IMPUTE and SNPTEST were used for genotype determination and association testing of rs1434536 as described (19).

**Cell lines, cloning, and dual luciferase reporter assays.** Cell lines were maintained in F12/DMEM, respectively, supplemented with 10% fetal bovine serum, and 1% Pen/Strep. Luciferase reporter targets were generated for the miR-125b target region of *BMPR1B* by cloning PCR products from HapMap NA18505 (rs1434536-C/T) into the 3′-UTR of the *Renilla* luciferase gene in the psiCheck2.2 dual reporter vector (Promega). Clones containing T or the C alleles at rs1434536 were verified by ABI fluorescent dideoxy sequencing and transiently transfected into MCF-7 and MD-MBA-231 cell lines. *Renilla* luciferase (hRluc) activity was measured 48 h after transfection. Cells were lysed with 120 μL Passive Lysis Buffer (Promega), and luciferase levels were analyzed from 10 μL lysates using the dual luciferase reporter assay (50 μL of each substrate reagent; Promega) on a Veritas Microplate Luminometer (Turner Biosystems). Changes in expression of *Renilla* luciferase (target) were normalized relative to Firefly luciferase.

**Transfection of miR-125b duplexes and qRT-PCR of BMPR1B.** siRNAs (IDT) were transfected into MCF-7 or MDA-MB-231 cells using RNAiMax (Invitrogen) using the manufacturers recommendations. Twenty-five pmol of each strand of the siRNA target were annealed by heating to 94°C for 2 min to form duplexes in buffer supplied by the manufacturer then allowed to cool to room temperature. Transfection efficiencies were monitored by transfecting in parallel a Cy3-labeled DS scrambled control siRNA duplex (IDT). Cells were harvested 24 h after transfection and RNAs were purified. cDNA was synthesized from 25 ng RNA using random hexamers and M-MLV reverse transcriptase, and was subsequently amplified with *BMPR1B* specific primers (Supplementary Materials and Methods). We calculated the SQ values and normalized *BMPR1B* transcript to *GAPDH*. RNA quantitation experiments were performed in triplicate from two independent transfection experiments.

## Results

**Multiple HapMap SNPs map to putative miRNA target sites in ER+ and ER− dysregulated genes.** Because allelic variations in miRNA binding sites have been shown to influence transcript levels (20), we examined if commonly occurring SNPs present in miRNA binding sites could be identified from the HapMap Consortium. Using a previously described set of genes dysregulated in ER+ and ER− breast tumors (14), we identified all HapMap SNPs residing within putative miRNA target sites in the genes' 3′UTRs (see Materials and Methods). We focused our search on the miRNA seed region, as the seed nucleates the miRNA to the complementary mRNA target region and is the main determinant for miRNA targeting (21). More specifically, we based our miRNA target site predictions on 7mer seed sites as we expected these would give an acceptable tradeoff between the number of false-negative and false-positive predictions (21). Our search identified 63 unique SNPs. Thirty-seven and 26 SNPs mapped to genes dysregulated in ER+ and ER− tumors, respectively (Supplementary Table S2). This collection of SNPs was considered for further analysis.

**A miR-125b target site SNP in *BMPR1B* is in strong LD with breast cancer–associated SNPs.** To prioritize the 63 SNPs for further biological testing, we mapped each to the publicly available

[16] http://microrna.sanger.ac.uk/; Release 9.1.
[17] http://www.hapmap.org/; Release #21.

Published OnlineFirst September 8, 2009; DOI:10.1158/0008-5472.CAN-09-1201

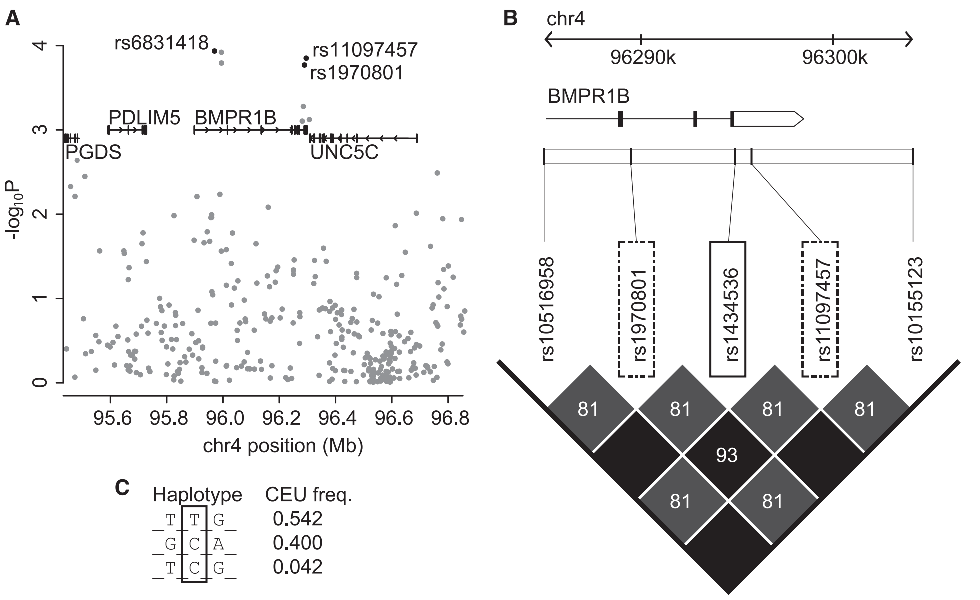


**Figure 1.** Regional CGEMS association data and LD structure in *BMPR1B* region. *A,* localized association data for CGEMS breast cancer data set (Chr 4, 95.3–96.8 Mb). Transcripts from the RefSeq database are shown in the top third part of the graph; *black,* selected SNPs. *B,* LD structure of *BMPR1B* (NM_001203) 3′UTR region. An ∼19-kb interval of the *BMPR1B* gene (*black boxes* and *white arrow-box,* coding sequence and 3′ UTR exons) and the surrounding region is depicted with select SNPs shown across the top. rs1434536 (*solid box*), located in 3′UTR of *BMPR1B*, is flanked by rs1970801 centromerically and rs11097457 telomerically (*dashed boxes*). *Shaded boxes,* pairwise LD values measured as $r^2$ with values listed; *black boxes,* perfect correlations ($r^2$ = 1). The direction of *BMPR1B* transcription, relative to the genome assembly, is from left to right. Panel adapted from Haploview (http://www.broad.mit.edu/mpg/haploview/). *C,* haplotype structure of three selected SNPs (boxed in *B*) with frequencies from HapMap CEU population where rs1970801 has been converted to + strand of University of California at Santa Cruz assembly.

CGEMS breast cancer GWAS[18] data set looking for SNPs that had signals of association. However, only seven mapped directly to this data set—none of which showed a statistically significant association. Twenty of the 63 target SNPs were either monomorphic (14 SNPs) in CEU samples or exhibited minor allele frequencies of <0.05 (6 SNPs) and were therefore not expected to be represented on the GWAS array, as rare SNPs (typically <5% minor allele frequency in CEU samples) are often excluded from these arrays. Moreover, the arrays typically contain only subsets of SNPs within haplotype blocks, but these SNPs can be used as proxies for the missing SNPs within blocks. To prioritize the remaining 43 SNPs, we therefore first used local LD structure from HapMap to identify proxy SNPs in the CGEMS data set and second observed such proxies' genome-wide association rank in the CGEMS set.

One SNP, rs1434536, showed high LD to rs1970801 and rs11097457 ($r^2$ = 0.81) in the HapMap CEU reference samples (Fig. 1). rs1970801 and rs11097457 ranked 79th and 67th in the CGEMS GWAS association data ($P$ = 0.00017 and $P$ = 0.00014, respectively, unadjusted score test). These SNPs exhibit extensive pairwise LD ($r^2$ = 0.93) in the CEU HapMap reference samples. We conclude that they likely represent the same association signal. The target site SNP rs1434536 lies 5.4 kb downstream of rs1970801 and 0.85 kb upstream of rs11097457 in the 3′UTR of the Bone Morphogenetic Protein Receptor 1B (*BMPR1B*) gene. The SNP's C→T change alters a 7mer seed site for miR-125b to a 6mer site—a change expected to reduce miR-125b's binding affinity to the site (Fig. 2). Moreover, miR-125b is differentially expressed in normal versus breast cancer in general, and in ER+ versus ER− tumors in particular (22–24). The combined observations that miR-125b and *BMPR1B* are differentially expressed in breast cancer, that allelic variation of rs1434536 likely disrupts miR-125b's regulation of *BMPR1B*, and that the SNP is in LD with two breast cancer–associated SNPs, suggest that rs1434536 has a pathogenic role in breast cancer.

**Independent cohort confirms the association of *BMPR1B* SNP with breast cancer.** Although the CGEMS results did not reach genome-wide significance for either rs1970801 or rs11097457, we elected to replicate the CGEMS results by screening rs1970801 in an independent cohort of genetically enriched breast cancer cases. In parallel, we screened two additional SNPs for association with disease: rs1219648 and rs6831418, which ranked 2 and 52, respectively, in the unadjusted CGEMS genomewide rankings (Supplementary Table S3). SNP rs6831418 resides within an intron of *BMPR1B*, ∼320 kb upstream of rs1970801 ($r^2$ = 0.118 with rs1970801), and a regional association plot of the CGEMS data (Fig. 1*A*) also indicated potential disease association. SNP rs1219648, present in intron 2 of *FGFR2*, was previously shown to be the most strongly associated SNP with breast cancer in multiple GWAS studies including the CGEMS, Wellcome Trust (rs2981582,

[18] http://caintegrator.nci.nih.gov/cgems/

Published OnlineFirst September 8, 2009; DOI:10.1158/0008-5472.CAN-09-1201

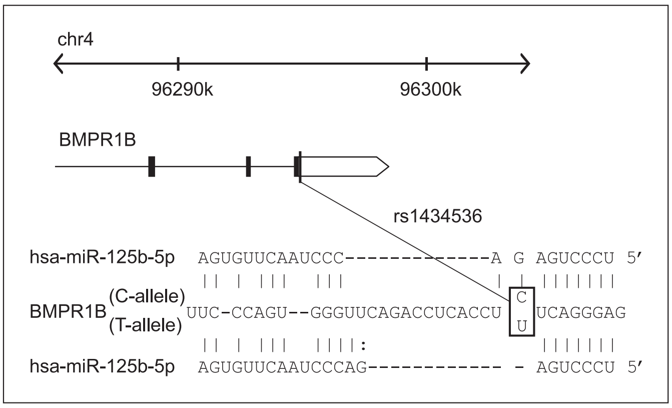

chr4
96290k
96300k
BMPR1B
rs1434536
hsa-miR-125b-5p AGUGUUCAAUCCC--------------A G AGUCCCU 5′
BMPR1B (C-allele) (T-allele) UUC-CCAGU--GGGUUCAGACCUCACCU C U UCAGGGAG
hsa-miR-125b-5p AGUGUUCAAUCCCAG------------ - AGUCCCU 5′

**Figure 2.** Predicted effect of allelic variation at rs1434536 on miR-125b recognition. *Top, BMPR1B* gene as described in Fig. 1 (*white box,* 3′UTR). *Bottom,* partial sequence of *BMPR1B* 3′UTR and SNP rs1434536 (boxed). *Bottom,* seed pairing of miR125-b (nucleotides 2–8 at 5′ end) with C (*top sequence*) and T(U) (*bottom sequence*) alleles of rs1434536.

$r^2$ = 1.0 with rs1219648), and Memorial Sloan-Kettering Cancer Center Ashkenazi Jewish (rs1078806, $r^2$ = 0.96 with rs1219648) studies (10, 11, 13). We used rs1219648/*FGFR2* as a positive control for association in our cases, as the three previous studies indicated this SNP is a universal marker for disease. Our breast cancer cases consisted of probands ascertained by virtue of a living, affected full sibling with disease, whereas we used admixture-corrected, shared disease-free controls from the CGEMS study. The use of cases ascertained by virtue of family history served to enrich for alleles with a strong genetic etiology. In addition, the use of shared controls has recently been described for multiple common disease scenarios (9, 25, 26).

Before comparing allele frequencies between our cases and CGEMS controls for the three SNPs, we sought to eliminate two potential biases: population differences between cases and controls and technical artifacts (e.g., errors in genotype scoring). To reduce the likelihood that any observed associations could be mediated by differences in the genetic ancestry of our cases and the controls, we elected to use AIMs and only analyze cases and controls with a high percentage (>90%) of Caucasian ancestry as defined by HapMap CEU reference samples. Recently, AIMs useful for determining intercontinental admixture have been described to facilitate structured association testing in case-control studies (27). We selected 59 AIMs (Supplementary Table S2) based on the 64 most informative $I_n4$ markers as described by Kosoy and colleagues (18) for population structure analysis. These markers have a high discriminatory power to distinguish CEU, YRI, CHB+JPT, and AMI (American Indian) continental populations. After STRUCTURE analysis (Supplementary Materials and Methods), we observed 94.1% (428 of 455) of our cases exhibited >90% CEU ancestry, whereas CGEMS controls showed 93.3% (1,064 of 1,142) CEU ancestry (Supplementary Fig. S1).

**Table 1.** Association testing in ECOG breast cancer cases and common CGEMS controls

| SNP | GT | ECOG cases<br>(*n* = 428; admixture adjusted) | | | CGEMS controls<br>(*n* = 1064; admixture adjusted) | | | ECOG cases +<br>CGEMS controls | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Count | Prop | HWE | Count | Prop | HWE | OR | 95% C.I | *P** |
| rs1219648 | A/A | 129 | 0.31 | 0.694 | 405 | 0.38 | 0.698 | 1.00 | | 5.2E-3 |
| Het | A/G | 204 | 0.48 | | 497 | 0.47 | | 1.29 | 1.00–1.67 | |
| Minor | G/G | 88 | 0.21 | | 161 | 0.15 | | 1.72 | 1.24–2.38 | |
| rs1970801 | G/G | 58 | 0.14 | 0.834 | 203 | 0.19 | 0.292 | 1.00 | | |
| Het | T/G | 192 | 0.46 | | 543 | 0.51 | | 1.24 | 0.88–1.73 | 4.8E-4 |
| Major | T/T | 167 | 0.40 | | 317 | 0.30 | | 1.84 | 1.30–2.61 | |
| rs6831418 | C/C | 121 | 0.29 | 0.427 | 349 | 0.33 | 0.037 | 1.00 | | 1.8E-1 |
| Het | C/T | 213 | 0.52 | | 549 | 0.52 | | 1.12 | 0.86–1.45 | |
| Minor | T/T | 79 | 0.19 | | 165 | 0.16 | | 1.38 | 0.98–1.94 | |
| rs1434536 † | C/C | 67 | 0.16 | 1.000 | 203 | 0.19 | 0.320 | 1.00 | | 1.6E-4 ‡ |
| Het | T/C | 200 | 0.48 | | 543 | 0.51 | | 1.29 | 0.95–1.74 | |
| Major | T/T | 148 | 0.36 | | 318 | 0.30 | | 1.94 | 1.40–2.71 | |

NOTE: Association testing with ECOG breast cancer cases (*n* = 428) and common CGEMS controls (*n* = 1064), which have been corrected for genetic admixture with ancestry informative markers (see Materials and Methods).
*Unadjusted *P* value from the score test with degrees of freedom (df) of 2 in logistic regression.
†Genotypes in CGEMS controls imputed by IMPUTE.
‡Score test with df of 2 (performed by SNPTEST) and considering imputation uncertainty.

Published OnlineFirst September 8, 2009; DOI:10.1158/0008-5472.CAN-09-1201

We next used logistic regression and calculated ORs testing independently for both heterozygotes and homozygotes carrier states omitting both ECOG cases and CGEMS controls not exhibiting >90% CEU class membership (Table 1). rs1219648/*FGFR2* showed association in the ECOG cases exhibiting a heterozygote and homozygote OR of 1.29 and 1.72 ($P$ = 0.0052) for the minor allele (G). These ORs are similar to those observed for the original CGEMS findings of 1.25 and 1.81 for heterozygotes and homozygotes. rs1970801-T also exhibited association in the ECOG cases with an OR of 1.24 and 1.84 for heterozygotes and homozygotes ($P$ = 0.00048). These ORs are comparable with those previously observed in the CGEMS study (1.21 for T/G and 1.65 for T/T). In contrast, rs6831418 did not exhibit significant association ($P$ = 0.177) in our cases. One explanation for the lack of confirmatory association with rs6831418 may stem from departure from Hardy Weinberg equilibrium (HWE) in both CGEMS cases and controls, whereas all three SNPs were in HWE for our ECOG cases (Supplementary Table S3). As final verification of association, we genotyped rs1434536 in our cases and, using the CEU HapMap LD structure, imputed genotypes for rs1434536 in admixture-corrected CGEMS controls. We observed association in our cases with an OR of 1.29 for T/C heterozygotes and an OR of 1.94 for major allele homozygote T/T (Table 1). Based on the replication of association in our breast cancer cases for rs1219648/*FGFR2*, rs1970801-T and rs1434536-T we concluded that rs1434536 was indeed associated with disease risk.

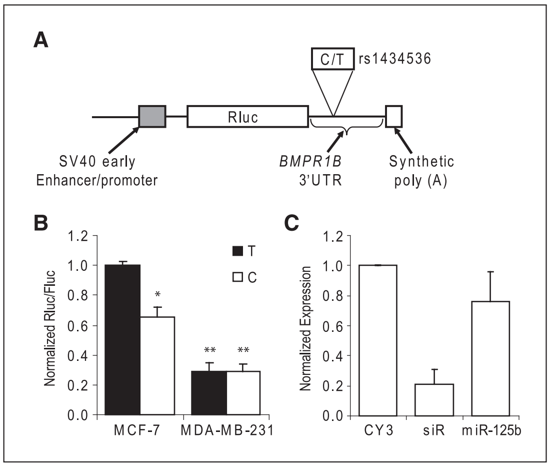


**Figure 3.** Allelic variation of rs1434536 influences luciferase reporter activity and miR-125b targeting. *A*, structure of *luc* allelic reporter constructs depicting psiCheck-2.2 (Promega) dual luciferase reporter constructs. *B*, luciferase reporter assays to measure C→T allele differences at rs1434536. Cells were transiently transfected with C- or T-bearing reporters into MCF-7 or MDA-MB-213 cells, which is predicted to influence the recognition of the miR-125b seed sequence in the *BMPR1B* 3′UTR. After 48 h, *Renilla* luciferase (hRluc) activity was measured and normalized to Firefly luciferase. Results are shown as percentage relative to luciferase activity. Data are from four independent transfection experiments with assays performed in triplicate ($n$ = 4). *, $P$ < 0.05; **, $P$ > 0.05. *C*, miR-125b weakly down-regulates *BMPR1B*. MCF-7 cells were transfected with siRNA duplexes and RNAs were harvested 24 h after transfection. cDNA was synthesized and used for real-time qRT-PCR analysis of *BMPR1B* expression normalized to a *GAPDH* standard. *CY3,* scrambled negative control siRNA; *siR,* siRNA duplex targeting position 867 in *BRPR1B*; *miR-125b,* duplex mimicking hsa-miR-125b and targeting the C allele at rs1434536. Expression levels are relative to the CY3 control ($n$ = 3).

**miR-125b differentially regulates the allelic variants of rs1434536.** Next, we tested a biological model where miR-125b differentially regulates the C/T allelic variants of rs1434536 in *BMPR1B*. In this model, rs1434536-T results in increased *BMPR1B* transcript levels, which gives an increased breast cancer risk as shown by the association testing. Computational models of miRNA target interactions predicted that miR-125b would down-regulate the C allele more strongly than the T allele, as the T allele forms a weaker 6mer seed site for miR-125b binding (Fig. 2; ref. 21). The PITA thermodynamic model of miRNA binding supports this allelic difference. The algorithm models miRNA targeting as a competition between the free energy gained by miRNA binding and the energetic cost of displacing existing RNA secondary structure at the target site (28). PITA summarizes this model in the ΔΔG score, where smaller values indicate stronger miRNA binding. Inputting the 200 nucleotides centered on rs1434536-C/T alleles to PITA gave ΔΔG values of −0.53 and 3.09, which suggested reduced binding of miR-125b to *BMPR1B* for the T allele.

To test our model, we cloned partial *BMPR1B* 3′UTR fragments from a rs1434536 heterozygote into the luciferase 3′UTR reporter vector psiCHECK-2 to compare the luciferase activities between the two alleles at rs1434536. Vectors containing either C or T alleles were transiently transfected into ER+ and ER− cell lines and *Renilla* luciferase activity was measured. When transfected into MCF-7 (ER+) cells the C-allele gave a 38% reduced luciferase activity relative to the T allele consistent with our model (Fig. 3*B*). However, when we tested luciferase activity in MD-MBA-231 (ER−) cells, we observed no difference between the C and T alleles. Additionally, the overall luciferase activities observed were lower in MDA-MB-231 cells relative to MCF-7 cells, which may reflect the higher levels of miR-125b in this cell line (22).

As an additional test of our model, we transiently transfected synthetic miR-125b oligos into MCF-7 and MDA-MB-231 cells, and quantitated endogenous *BMPR1B* transcript levels by qRT-PCR. Prior genotyping MCF-7 and MDA-MB-231 cells revealed homozygous T and C genotypes at rs1434536, respectively. The oligonucleotides, which mimicked the annotated hsa-miR-125b:hsa-miR-125b-1* duplex, only weakly down-regulated *BMPR1B* in MCF-7 (Fig. 3*C*), which is consistent with our model. In contrast, transfection with a miRNA mimic (siR), not targeting the miR-125b site, resulted in an 80% reduction in *BMPR1B* transcript levels. When we tested these duplexes in MDA-MB-231 cells, *BMPR1B* levels were <1/50 of the levels in MCF-7 and below the assay's detection limit (data not shown). The low levels of *BMPR1B* levels in ER− MDA-MB-231 cells were consistent with our prior meta-analysis from ER+ and ER− tumors and with increased levels of endogenous miR-125b in 231 cells (22).

## Discussion

Both rs1434536-T and rs1970801-T were shown to be associated with increased risk in an independent cohort of admixture-corrected cases and controls. We have shown that miR-125b negatively regulates *BMPR1B* and that C/T allelic variation (rs1434536) within the target site disrupts the regulation of miR-125b. The presence of rs1434536-T leads to loss of miR-125b regulation, increased *BMPR1B* expression, and ultimately elevated disease risk. Consistent with this, increased *BMPRIB* expression is associated with high tumor grade, proliferation, cytogenetic instability, and a poor prognosis in ER+ breast carcinomas (29). Moreover, breast cancers in general and ER+ tumors in particular

Published OnlineFirst September 8, 2009; DOI:10.1158/0008-5472.CAN-09-1201

seem to have reduced levels of miR-125b (22–24), which in light of these results, at least partly explain why ER+ tumors have increased *BMPR1B* expression (14).

BMPR1B binds bone morphogenetic proteins (BMP) and are multifunctional signaling molecules that belong to the transforming growth factor β superfamily and were first identified based on their ability to form bone at extraskeletal sites (30). Once activated, BMP/receptor complexes phosphorylate cytosolic SMAD proteins that translocate to nucleus and regulate target genes (31). Our findings indicate that ER+ patients harboring elevated *BMPR1B* transcript levels may have poorer outcomes when carrying the risk-associated rs1434536-T allele. Whereas not only identifying a new target for further study, these results show the importance of combining tumor expression profiles and genotype data in determining a patients' clinical prognosis.

More generally, our methodology has identified a set of allelic variants present in miRNA recognition sites within a set of dysregulated ER responsive genes. Independent of our efforts, Adams and colleagues (20) identified rs9341070 in a miR-206 site in the *ESR* 3′UTR. Allelic variation at this SNP was shown to influence *ESR* expression over 3-fold in HeLa cells. This SNP resides in a domain upstream of the miRNA seed targeting sequence (nucleotides 2–8), yet we identified this same SNP by virtue of its presence in a miR-122 seed region (Supplementary Table S1). However, due to the low frequency of rs9341070 in CEU samples (<0.01) this SNP is not represented in any GWAS array. This illustrates a common deficiency of GWAS data sets: the absence of low frequency/rare SNPs that may also play a role in disease risk (32). One would anticipate that appropriately powered future association studies of these potential miRNA interacting rare variants may support their role in risk.

We found that T/T homozygotes at rs1970801 had slightly higher ORs in our ECOG cases (1.84) compared with the CGEMS cases (1.65), and this could be explained by differences in case ascertainment. CGEMS cases were recruited from a prospective cohort study where only 22% (274 of 1,145) reported first-degree family history as opposed to our cases whereby all cases exhibited first-degree family history, namely an affected sibling. Second, all CGEMS cases were of postmenopausal disease, whereas only half of the ECOG cases indicated an age of diagnosis of <50. These differences indicate that the genetic contribution to risk may have been underestimated for rs1970801-T in the CGEMS study reinforcing the importance of family history in confirmatory replication studies as this may be valuable for later risk-assessment predictions. We observed a higher OR for TT homozygotes at rs1434536 when we tested for association with imputed genotypes in the CGEMS controls compared with rs1970801 TT homozygotes (1.94 versus 1.84; Table 1). These results also highlight both the merits of the tagging SNPs used in the GWAS studies and the utility of imputation for deriving missing genotypes.

Our replication of prior disease associations for two SNPs (Table 1) relied on using a set of AIMs to correct for differences in genetic admixture between our cases and CGEMS controls. Approximately 6% to 7% of CGEMS controls and CGEMS cases (data not shown) showed <90% CEU ancestry as defined by HapMap reference samples. This indicates that population substructure introduced by intercontinental admixture may have contributed to potential false positives or missed associations in the original CGEMS data. To rectify this, it has been proposed that AIM panels should be used before GWA tests (26). More subtle levels of admixture within both European and Chinese populations have recently been described, which will necessitate the continued use of extended AIM panels to discern finer levels of population substructure as a prelude to association testing and biological testing (33–35).

The usefulness of GWAS data for identifying breast cancer susceptibility alleles is premised on the common disease–common variant hypothesis whereby SNPs (>5% frequency) may act as surrogates to identify causal variants. Replication studies of the very top tier signals in breast cancer have firmly established some associations; however, modest signals in first round GWAS screens may not be selected for rescreening (36). Thus, we feel it is likely that future meta-analyses of multiple GWAS data sets will provide additional candidates for examination (37).

These findings have implicated a germline variant in breast cancer susceptibility and provided a strong model for biological causality via miRNAs. Our approach relies on integrating association data, expression profiles, and testable biological models to evaluate potential disease alleles in pathogenesis (38). As GWAS have identified only common SNPs as genetic risk factors, it is likely that many rare alleles present within motifs for miRNAs and additional *trans*-acting regulators (i.e., transcription factors) remain to be identified. In addition, approaches such as whole genome sequencing and the identification of common recurrent somatic mutations in breast tumors may provide a large collection of potential disease alleles for exploration (39, 40).

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Acknowledgments

Received 3/31/09; revised 6/29/09; accepted 7/22/09; published OnlineFirst 9/8/09.

**Grant support:** Susan G. Komen for the Cure Basic, Clinical and Translational Research Grant BCTR0504486 (G.P. Larson), and the Norwegian Functional Genomics Program (FUGE) of the Norwegian Research Council (P. Sætrom and L.F. Thomas). J. Biesinger was a participant in the Southern California Bioinformatics Institute Summer Academy (NSF-NIH EEC-0609454). J.J. Rossi is funded by (HL074704).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank the ECOG study subjects in E1Y97 for their participation.

## References

1. Ambros V. The functions of animal microRNAs. Nature 2004;431:350–5.
2. Kloosterman WP, Plasterk RH. The diverse functions of microRNAs in animal development and disease. Dev Cell 2006;11:441–50.
3. Esquela-Kerscher A, Slack FJ. Oncomirs — microRNAs with a role in cancer. Nat Rev Cancer 2006;6:259–69.
4. Kumar MS, Lu J, Mercer K, et al. Impaired microRNA processing enhances cellular transformation and tumorigenesis. Nat Genet 2007;39:673–7.
5. Yu Z, Li Z, Jolicoeur N, et al. Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers. Nucleic Acids Res 2007;35:4535–41.
6. Landi D, Gemignani F, Naccarati A, et al. Polymorphisms within micro-RNA-binding sites and risk of sporadic colorectal cancer. Carcinogenesis 2008;29:579–84.
7. Loi S, Piccart M, Sotiriou C. The use of gene-expression profiling to better understand the clinical

Published OnlineFirst September 8, 2009; DOI:10.1158/0008-5472.CAN-09-1201

heterogeneity of estrogen receptor positive breast cancers and tamoxifen response. Crit Rev Oncol Hematol 2007;61:187–194.

**8.** Dobbe E, Gurney K, Kiekow S, et al. Gene-expression assays: new tools to individualize treatment of early-stage breast cancer. Am J Health Syst Pharm 2008;65:23–8.

**9.** Consortium WTC C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78.

**10.** Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39:870–4.

**11.** Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447:1087–93.

**12.** Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007;39:865–9.

**13.** Gold B, Kirchhoff T, Stefanov S, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci U S A 2008;105:4340–5.

**14.** Smith DD, Saetrom P, Snøve O, Jr., et al. Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. BMC Bioinformatics 2008;9:63.

**15.** Tworoger SS, Eliassen AH, Sluss P, et al. A prospective study of plasma prolactin concentrations and risk of premenopausal and postmenopausal breast cancer. J Clin Oncol 2007;25:1482–8.

**16.** Larson GP, Zhang G, Ding S, et al. An allelic variant at the ATM locus is implicated in breast cancer susceptibility. Genet Test 1997;1:165–70.

**17.** Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 2007;39:596–604.

**18.** Kosoy R, Nassir R, Tian C, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 2009;30:69–78.

**19.** Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39:906–13.

**20.** Adams BD, Furneaux H, White BA. The micro-ribonucleic acid (miRNA) miR-206 targets the human estrogen receptor-α (ERα) and represses ERα messenger RNA and protein expression in breast cancer cell lines. Mol Endocrinol 2007;21:1132–47.

**21.** Grimson A, Farh KK, Johnston WK, et al. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007;27:91–105.

**22.** Iorio MV, Ferracin M, Liu CG, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res 2005;65:7065–70.

**23.** Blenkiron C, Goldstein LD, Thorne NP, et al. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome Biol 2007;8:R214.

**24.** Mattie MD, Benz CC, Bowers J, et al. Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies. Mol Cancer 2006;5:24.

**25.** Plenge RM, Cotsapas C, Davies L, et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet 2007;39:1477–82.

**26.** Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet 2008;17:R143–50.

**27.** Halder I, Shriver M, Thomas M, et al. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat 2008; 29:648–58.

**28.** Kertesz M, Iovino N, Unnerstall U, et al. The role of site accessibility in microRNA target recognition. Nat Genet 2007;39:1278–84.

**29.** Helms MW, Packeisen J, August C, et al. First evidence supporting a potential role for the BMP/SMAD pathway in the progression of oestrogen receptor-positive breast cancer. J Pathol 2005;206:366–76.

**30.** Wozney JM. Overview of bone morphogenetic proteins. Spine 2002;27:S2–8.

**31.** Kawabata M, Imamura T, Miyazono K. Signal transduction by bone morphogenetic proteins. Cytokine Growth Factor Rev 1998;9:49–61.

**32.** McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356–69.

**33.** Novembre J, et al. Genes mirror geography within Europe. Nature 2008;2008:e3862.

**34.** Tian C, et al. Analysis of East Asia genetic substructure using genome-wide SNP arrays. PLoS ONE 2008;3:e3862.

**35.** Seldin MF, Price AL. Application of ancestry informative markers to association studies in European Americans. PLoS Genet 4:e5.

**36.** Garcia-Closas M, Hall P, Nevanlinna H, et al. Heterogeneity of breast cancer associations with five susceptibility Loci by clinical and pathological characteristics. PLoS Genet 2008;4:e1000054.

**37.** Zintzaras E, Lau J. Trends in meta-analysis of genetic association studies. J Hum Genet 2008;53:1–9.

**38.** McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 2008;17:R156–65.

**39.** Wood LD, Parsons DW, Jones S, et al. The genomic landscapes of human breast and colorectal cancers. Science 2007;318:1108–13.

**40.** Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–9.

## SUPPLEMENTARY METHODS

### Description of Study Populations and Genotyping

Our ascertainment criteria for cases included a living affected sister with disease willing to participate in the study. Among families that provided information, 96% reported Caucasian ancestry, 2% Ashkenazi, ~1% African American, 1.7% Native American, and <1% other. Our sib pairs consisted of 8 sets of self-reported monozygotic twins, (1.7%) and 6 pairs (1.3%) having non-shared paternity based upon allele sharing at the X-linked androgen receptor (*AR*) microsatellite (het=0.89) [1]. Non-shared paternity for a given sib pair was defined as an index case and her sibling sharing 0 alleles at *AR*. Index cases from these pairs were nonetheless retained for association testing. SNP assays were designed with assay Design 3.1 software into 3 separate assays (Supplementary Table S2). After a 6ml multiplex PCR amplification the resulting products were treated with Shrimp Alkaline Phosphatase (SAP) and single-base primer extension chemistry was conducted with mass modified dideoxyribonucleotides (iPlex Gold Chemistry). Extension products were processed with SpectroCLEAN resin, and spotted onto SpectroCHIPs and analyzed via MALDI-TOF mass spectrometry [2]. Automated genotype calls were made with SpectroTYPER v3.4 software. To reduce the likelihood of scoring errors due to genotyping platform disparities we genotyped both our AIM and association SNPs in a reference panel of HapMap samples and observed a concordance rate of 99.36% (2161/2175). Genotype data for all association SNPs were tested for deviations from Hardy-Weinberg proportions in cases and we observed no deviation for the 3 SNPs tested ($p$<0.001). Five AIMs showed deviations from Hardy-Weinberg proportions but were nonetheless retained for analysis with STRUCTURE 2.2 for population admixture.

### Population Structure Analysis and Association Testing

We genotyped 59 AIMs utilizing Sequenom iPLEX mass spectrometry technology. We downloaded equivalent genotypes for HapMap self-identified reference samples and for controls (n=1,142) and cases (1,145) from the CGEMS database. After genotyping our ECOG BrCa cases we retained 455 individuals for STRUCTURE analysis. Five AIMs (rs1040045, rs6451722, rs3907047, rs4746136 and rs798443) exhibited prior association to disease in the CGEMS dataset ($p<0.006$ to $p<0.037$) and were omitted from STRUCTURE analysis [3]. STRUCTURE analyses were run without any prior population assignment using 50,000 iterations with 10,000 burn-in cycles under the admixture model with initial □ =1.0 without specifying population membership. We utilized the infer □ option and estimated a separate a for each population under the F model (□ is defined as the Dirichlet parameter for degree of admixture). When we included 105 AMI (AmerInd) individuals as described by Kosoy, et. al [4] and increased the defined population cluster parameter to k=4 populations we observed no appreciable difference in the clustering of our ECOG cases or CGEMS controls to CEU HapMap reference samples (data not shown). More importantly we were also able to identify the most likely genetic ancestry of our cases and the CGEMS controls for which we lacked self-reported ethnicity. Association testing for imputed rs1434536 genotypes were performed by the method of Marchini [5]. Briefly, rs1434536 genotypes from CGEMS controls were imputed with IMPUTE from HapMap CEU SNPs from chromosome 4 region 96,289 – 96,296 kb, which surround rs1434536. IMPUTE uses a hidden Markov model and known HapMap haplotypes to impute missing data. Association testing with SNPTEST includes imputation uncertainty in the subsequent association test by modeling the observed data likelihood using the full data likelihood integrated over missing data.

### Cloning, Luciferase Assays, and qRT-PCR of *BMPR1B*

MB-MDA-231 and MCF-7 cell lines were obtained through American Type Culture Collection (Manassas, VA). All tissue culture reagents were purchased from Invitrogen (Carlsbad, CA) and Sigma (St. Louis, MO). PCR primers for *BMPR1B* were (forward primer: 5'-CCGCTCGAG GTCCCAGGACATTAAACTCTG-3', Reverse primer: 5'-TTTTCCTTTTGCGGCCGCGCATCA

TATCTTGAACAAGTT-3') containing *Xho I* and *Not I* restriction sites respectively for directional cloning into the MCS site Psi-CHECK-2.2. Twenty-five nanograms of genomic DNA were PCR amplified (95°C, 5min, 95°C 30sec, 55°C, 30sec, 72°C, 40sec, for 35 cycles, 72°C 3 min final extension) with 1□l of Taq (5U/□l Roche), 1X Taq Buffer, 1□M primers, and 200□M dNTPs. PCR products (~0.28kb) were restricted with the aforementioned enzymes, purified via gel electrophoresis and cloned into PsiCheck-2.2. Genotypes for MCF-7 and MDA-MB-231 cell lines at rs1434536 were determined by sequencing PCR products derived from 25ng genomic DNA isolated from cells grown in culture and the aforementioned primers.

MDA-MB-231 and MCF-7 cells seeded one day before, were transfected with plasmids bearing the T or C alleles in triplicates in 24-well plates at 80% confluency with a Lipofectamine 2000 (Invitrogen) complexed with a mixture of 25 ng psiCheck reporter plasmid and 75 ng stuffer DNA (pBlueScript) per well. miR125-b target site cleavage results in degradation of reporter mRNA, with a concomitant decrease in translated product, which can be detected by a luminescence-based assay system. Firefly luciferase expressed from psiCheck2.2 served as an internal normalization control.

### Transfection of miR-125b Duplexes and qRT-PCR of *BMPR1B*

Sequences for siRNA duplexes: siR 5'- GGACUAUAGCUAAGCAGAUUCAGat-3' and 3'-UUCCUGAUAUCGAUUCGUCUAAGUCUA-5' (RNA nucleotides are shown in uppercase and DNA nucleotides shown in lower case) and has-miR-125b:hsa-miR-125b-1* (targets C allele of rs1434536) duplex: 5'-UCCCUGAGACCCUAACUUCUGA-3' and 5'-ACGGGUUAGGCUCUUGGGAGCU-3'. These duplexes target positions chr4:96,270,043 and chr4:96,294,738 respectively in *BMPR1B*. Cells were purified with RNA STAT-60 (IsoTex Diagnostics, Inc.) according to manufacturers directions. *BMPR1B* specific primers 5'-CAACAAAATTCTTCCCAGGAACT-3' and 5'-TGGTTCACAGAGTGCAACAATA-3' were used to amplify cDNAs. Samples were treated with DNase I (Ambion, Turbo DNA-free) and control reactions omitting M-MLV were also included to rule out genomic DNA contamination. SYBR green technology was utilized for transcript quantitation. *GAPDH* intron spanning primers ('5-CATTGACCTCAACTACATG-3' and 5'-TCTCCATGGTGGTGAAGAC-3') were utilized as normalization controls. PCR conditions were: 95$^{o}$C for 5 min, followed by 40 cycles of 95$^{o}$C, 15 sec, 55$^{o}$C for 30sec, 72$^{o}$C for 30sec.

## References

1. Haiman, C.A., et al., *The androgen receptor CAG repeat polymorphism and risk of breast cancer in the Nurses' Health Study.* Cancer Res, 2002. **62**(4): p. 1045-9.
2. Ragoussis, J., et al., *Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research.* PLoS Genet, 2006. **2**(7): p. e100.
3. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.
4. Kosoy, R., et al., *Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America.* Hum Mutat, 2009. **30**(1): p. 69-78.
5. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.

Supplementary Table S1 Sequenom MassARRAY assay designs of association and AIM SNPs

| | | | | | UEP_ | EXT1_ | EXT2_ |
|---|---|---|---|---|---|---|---|
| WELL* | SNP_ID† | 2nd-PCRP‡ | 1st-PCRP | UP_SEQ§ | DIR | CALL | CALL |
| Assay_1 | rs9522149 | AGAAAGGAGAGGAAACACCG | TCAGCAACTTCTAGTCCTCG | GGTCCTTGCAGCTCC | F | C | T |
| Assay_1 | rs11652805 | CCCTCAAAGTTTGGTGCATC | CTCTTCCTGGTCCTGAGATG | CTTTCTCTCTCCCAGC | F | C | T |
| Assay_1 | rs9530435 | ATCAGGCACATGGTAAGCAC | CTCCATCTGGTACATATGTG | gAAGCACTCAGCGAAG | R | T | C |
| Assay_1 | rs2416791 | TATAGCATCTACCATCAGCC | ATACTGCCCCATAAGGAGTG | aACCATCAGCCCAATTC | F | A | G |
| Assay_1 | rs9855638[2] | GGTTAGTTTTGGTGAAGTCC | GACCTTGGCTTTTACCATAG | TTGTTGCTCATGCATTT | R | G | C |
| Assay_1 | rs10108270 | AACAGCATCTGAGACGCTTC | AGTGACCCTGGACACAATTC | TCAGGTGAGGACTTAGC | R | C | A |
| Assay_1 | rs4666200 | CCCTATCCTTGGTGATTTGG | CAGTCACAATTGGCAAGCAC | tACTTCAGAGCTATTGGC | R | G | A |
| Assay_1 | rs9319336 | ATGCAAGGTAATGCACCCTC | TCTACCTGCAGGTAAGTGTC | ACCCTCTCCCTGCTTCTAT | F | C | T |
| Assay_1 | rs3907047 | CAGAATCGGACATGATACCC | GAAAGTCCAGGAAGTTCAGG | ctTCAGCTCTCTGATCTCC | F | C | T |
| Assay_1 | rs4908343 | CCAAACCCCACAAGCTTAAC | AGGGAGAGAAGGTCAGTTAC | AACCCCTGGGCTATGACAA | F | A | G |
| Assay_1 | rs1513181 | CAGATTTCCCATAGCCTCTC | AGGTGAGACAGTTGGACAAG | GTTGAGCTTGAAAAATTCCC | F | C | T |
| Assay_1 | rs3737576 | GGTCCTGGTTCTTGTCAAAG | AGGAGGAAGAGCATAGTGAG | AGATTGTGAAAGACTGAAAT | F | A | G |
| Assay_1 | rs1040045 | GAGAGAAAGGGACACAGAAC | CCTCACCCCATCTACTCTTG | tagtATGGGGATTGGGGTAA | F | C | T |
| Assay_1 | rs7803075 | AATCCACATGAACTGCGCTC | ATCTATCACGTGGACCTTTG | cctCGCTCCTGGATCTTTTAC | F | A | G |
| Assay_1 | rs731257 | GTTGAGGTATCAGTGGAATC | TCCATCCTAATTGGAACTGC | TGGAATCACAAATTGTATCTC | R | G | A |
| Assay_1 | rs7997709 | TAACTGTGTTTCCCTCAGTG | ATGTGGATGGATTGCTCAAC | agaTTTCCCTCAGTGGTTAGC | R | T | C |
| Assay_1 | rs4918842 | ATTGCTCAGAAATGCTGTGG | TTAATCGGATGCTGAGCCTG | AAATGCTGTGGATATTGACTTA | F | C | T |
| Assay_1 | rs10496971 | AATGTCACCTTTAGGCAGAG | GGGAACATTGAGTCCTCAAG | ggggaTTTAGGCAGAGGCATTT | F | G | T |
| Assay_1 | rs12629908 | CCAATCTCTTACATCTCCTG | TCCATCATCCAGGAGCTTAG | CTTACATCTCCTGAAAAGAAATT | R | G | A |
| Assay_1 | rs10007810 | TCTTCTCTTGTAGACAGGGC | CGTGGCACATGCCATGTTTT | acttgAGACAGGGCCCTCTATCT | F | A | G |
| Assay_1 | rs772262 | CACTTTTGACTTAAGACGGG | TTCAATACCTCTGGCCTCTC | cGACGGGTTTTTATCAGGACATA | F | A | G |
| Assay_1 | rs4746136 | GGTATGTCCTAGATGACAAG | AGCACATACCTGCAAGCACG | TGGACAGATAAACTTATTTGTGTA | R | G | A |
| Assay_1 | rs2125345 | AGTGTATGGTTTCTTTGTGG | TAACGTGAGTCAGACTGTAG | ggtagGGTTTCTTTGTGGGATTCT | R | G | A |
| Assay_1 | rs6451722 | CTCTGTAAGCAGCTATTGCC | TCTGCTCCTAAGGAAGATGC | tCAGCTATTGCCATTTTTTTCTCAT | F | A | G |
| Assay_1 | rs7554936 | AACCAGGGACTGCATACAAC | CATCCTAGTGAATGCCATCC | ccctTAAAGTCATAGGTGAACCTTC | R | T | C |
| Assay_1 | rs7657799 | ACAAGGCCCAATTGCTGAAG | AGCCAACTTGATTCTCTTTC | cccTGATCTACCTTGCAGGTATAATG | F | G | T |
| Assay_1 | rs260690 | CTCATAGTTGCTATGAACAG | TCTGTGGCCAACGTAAAAGG | ggcGTTGCTATGAACAGTTTAACAGT | R | C | A |
| Assay_1 | rs4891825 | GTGTAACAATCTCAATCCCC | CTAGGGTTGGTAAAGGATGG | atcgCAATCCCCCTTAATGTTTTCATC | F | A | G |
| Assay_1 | rs6104567 | ACAAGGCCCAGTATGATTG | GCTTGGCTTTAATATGGAGG | CAGTATGATTGATACATATCTAATTAA | F | G | T |
| Assay_1 | rs1471939 | TACCACCCATCTTAAACAGC | TGTTAACTCCAGAACAAGTG | cctCATCTTAAACAGCTATAGATATAGT | R | T | C |
| Assay_2 | rs1407434 | CCCATATCATCTCCACTCAG | TGAACCTAAAAAGCAAAGGG | GCCCTCAGTCCCTCT | R | T | C |
| Assay_2 | rs2504853 | CATCCTGAAGGTGATGGAAG | GAAATTCACAGGCTCCAGAC | ATGGAAGCCTTGCAT | F | C | T |
| Assay_2 | rs870347 | ACCTTTTTCAGCCTGACTCC | ATCATGCGACATCCAGGTAG | TGCTAAGTCCCTCACT | F | G | T |
| Assay_2 | rs4821004 | CTTGCAAGTGTGAAGAGCAG | CAAGGGCCGATGATATTTGC | GGGGAGGGAGCAAGCC | F | C | T |
| Assay_2 | rs9845457 | TTGCACTAGATCCGGGAAGC | CTTACTCCATCCCAGTACAG | ggCCGGGAAGCCGCTGC | R | G | A |
| Assay_2 | rs2946788 | TATCTACTCTGGCCAAACTC | CATTCCAATGAGCTTAAGCC | CCAAACTCAATAGCCACA | R | G | T |
| Assay_2 | rs8113143 | TGTGGGTTCTTGCTGTGTTG | AAGTGAGAGGATGAGAGGAG | GTTGGATAACACATCCCC | R | C | A |
| Assay_2 | rs2030763 | CTTCCTTTTCTTACCAACTGC | ATCCATGCGGATGGCTTAAC | ATGAATAAGCTGAGCTTCT | R | G | A |
| Assay_2 | rs9809104 | AAAACCACAGGACAGGACAG | TGACGTGGAGTGATTTGGAG | CAGGACAGTTATTCAGGAA | F | C | T |
| Assay_2 | rs798443 | GGTATTGCTAACATCTCCAG | CTCAGTGCAGATGGGAAATG | cAATTTCCACTAACAACGCA | F | A | G |
| Assay_2 | rs2397060 | AAAACATGTTTAGGGTTTG | CCTTCATTACAACCCAGGTA | ATGTTTAGGGTTTGAAGAAT | F | C | T |
| Assay_2 | rs4984913 | GGAAGTGGTCCTCTTCTTAC | ACCCGGAACTTTCGTGGTGT | aagaaCAGGAAGTGGGCACA | F | A | G |
| Assay_2 | rs2627037 | AGCGCCGAACTTCAATTATC | GTGCCTTCCTTTTCGGAATC | TTGTCTGAATCTCCAGTTTAC | R | G | A |
| Assay_2 | rs3943253 | TGTGGCTTAGGAGTGACATC | ATCCAGTGTAGAAAGAGCCG | TGACATCGTAATACCACTTGG | R | G | A |
| Assay_2 | rs13400937 | CTTACCACCCGTGAAATAAC | CCAAAGTTTGTTCCAAATCTG | aaACATTTCAGGAAGTTGAATT | F | G | T |
| Assay_2 | rs734873 | ACTGTCCTGTGTCAAGAACC | GATGTCTTGATGATTCCTCC | gggaCCTAGGGCAAGAGAGTAA | R | T | C |
| Assay_2 | rs3745099 | CAGTTACTTTTCTCCCCTGC | AAGTAGAAGGTGAGTGAGGG | ggccGCTATTTTCTCGGCACCTT | R | G | A |
| Assay_2 | rs1040404 | AACTCAAGTGTCTCCTGAGC | CAGCTGAGCATTTTGTAGTG | tgGTGATACTATTTTCTACCACA | F | C | T |
| Assay_2 | rs10236187 | AGAAGGAACGGCAGACAAAG | CCTAGGTGGGAGTAAAAGTG | ggagtCAGACAAAGCCTCACATTA | F | C | A |
| Assay_2 | rs1325502 | TCTGGATAAACATTCTGGCG | CATCACCCAGAATGGCAAAC | gagtgCTGGCGTTGCTGCATGTTT | R | G | A |
| Assay_2 | rs10513300 | TACCTCTGCAATGCCCTATC | AAGAGCACATACTCCATACC | CCCTATCTTATTATCATATGAGTTC | F | C | T |
| Assay_2 | rs12130799 | GTGTTACTCAATGGAGCTCT | GGTTCTGGGATATTGTTGGG | gTTCTCTATTGTATCTCCAATGTCT | F | A | G |
| Assay_2 | rs6422347 | TGAAGGCCGACTTCACGGA | ATGTTGACCTCCCTCTCCC | ggtcCGGAGCTGGTGACATTTTAAC | F | C | T |
| Assay_2 | rs1408801 | GTGATAGTTTTACAGTTTCC | ACATGCATGTGTATTGCAGG | ggcaTTTTACAGTTTCCTAAACCATG | R | G | A |
| Assay_2 | rs4463276 | TCGGCTTGTTTCCTTTTTTG | ACAACAAGGAAAATGAGCCC | tttgGTGGGTACACAGTAAGTGTATA | R | G | A |
| Assay_2 | rs4717865 | GTTCTAGATTCAGACCCTGC | CATCGGAGAGGCAAATTGAC | ggggaCCTGCTGCTGCTACCCAGCCTC | R | G | A |
| Assay_2 | rs1760921 | ATACGCAAAACCACTGCCAC | TACTGGCCATATTCTCTCTC | ccacaACTGCCACATCCGTCCCATACCT | F | A | G |
| Assay_2 | rs6556352 | CAATGCATATGTACTGCTTCC | AGCATTCTATAAACCGACAG | ccTCCATAAAAATGAAATATCATTTAAC | R | T | C |
| Assay_2 | rs7421394 | AGTTTAAGAGGTTTGACAGG | TTTTCCACGTGAACATACCC | cctctGTTTGACAGGATAATTTCTGAGA | R | G | A |
| Assay_3 | rs1219648 | GACAAGCCATGGCCATCCTT | TCTTCCATGGTACCGGTTTC | GGCCATCCTTGAAGAG | R | G | A |
| Assay_3 | rs1970801 | CAGTAGGCCATAAATGTGGG | CAAATTGCTTTATGGGGAAG | GACACCCATTTCTTACCT | R | C | A |
| Assay_3 | rs6831418 | GGACTTCCTTACTAGAGCAC | CCTCACAGAATTAAGAGTGC | TGTTTCCTTTCCTCTCC | R | T | C |

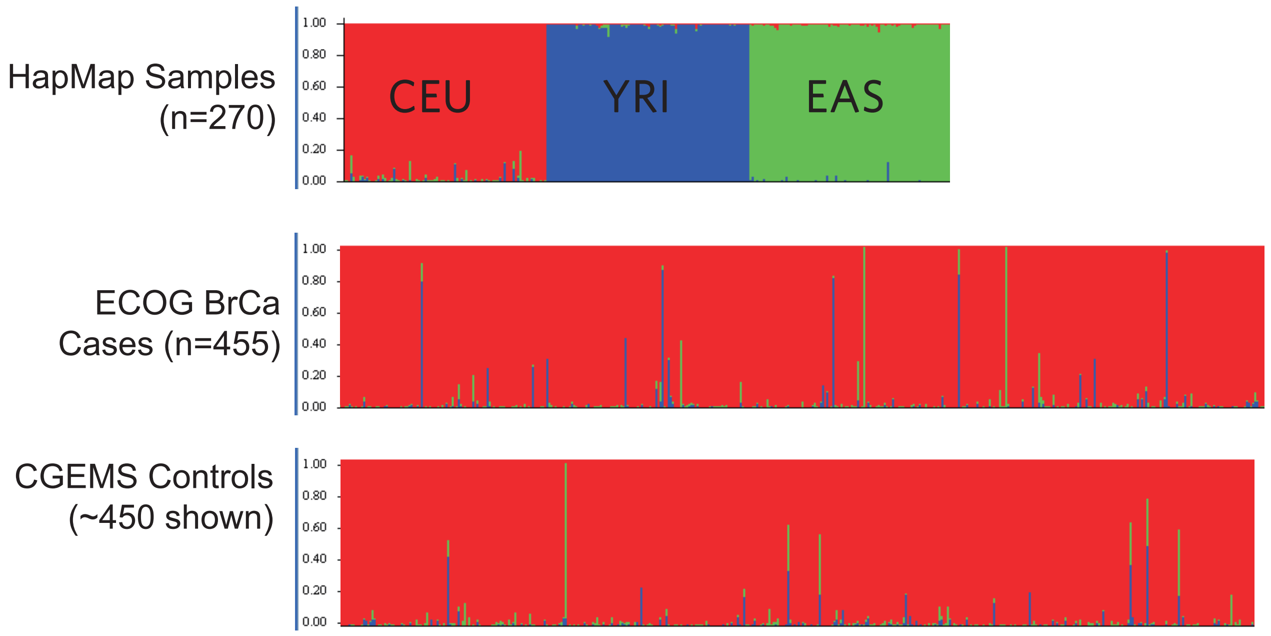
Supplementary Table S1 Sequenom MassARRAY assay designs of association and AIM SNPs

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Assay_3 | rs1434536 | CTTGAACATCGTCCTGCTTC | TGGGAGCTTCTCTGTCTTTG | GGTTCAGACCTCACCT | R | G | A |
| Assay_3 | rs11097457 | TGCTCTTGTGTTGTAAGAGG | AGATACAAGCCATCTGCACG | gggACATGTCAACAAAGATAGG | F | A | G |

Notes

| | |
|---|---|
| * | Assays 1 and 2 dervied from In4 markers of Kosoy, et al, Assay 3 SNP for association testing |
| † | rs9855638 replaced In$_4$ SNP rs6548616 with $r^2$ >0.9 in 3 HapMap populations |
| ‡ | ACGTTGGATG mass tags were appended to each PCR primer |
| § | lower case nucleotides in extension primer are non-complementary to amplified region |

Supplemental Figure 1



| Patient Sample Set | Number Patients | >80% YRI membership | >80% EAS membership | >90% CEU membership |
|---|---|---|---|---|
| CGEMS Controls | 1142 | 0 | 9 | 1064 (0.932) |
| ECOG BrCa Cases | 455 | 4 | 2 | 428 (0.941) |

Supplementary Table 2
SNPs Mapping to miRNA Sites

| SNP_ID | miRNA recogniton site | chrom | SNP_Pos* | Motif_Chrom:position | genelist† |
|---|---|---|---|---|---|
| rs10157828 | GCAGCCA | chr1 | 17267200 | chr1:17267200-17267206 | ER- overexpressed |
| rs4366378 | GGCCAGT | chr1 | 208914363 | chr1:208914359-208914365 | ER+ overexpressed |
| rs6565 | AACCATA | chr1 | 28085744 | chr1:28085739-28085745 | ER- overexpressed |
| rs6619 | TTGGGAG | chr1 | 37803283 | chr1:37803283-37803289 | ER+ overexpressed |
| rs1058240 | CCGTTGA | chr10 | 8156604 | chr10:8156602-8156608 | ER+ overexpressed |
| rs11191382 | AATGGGT | chr10 | 104488596 | chr10:104488593-104488599 | ER+ overexpressed |
| rs1334891 | CCAGGTT | chr10 | 99070861 | chr10:99070857-99070863 | ER+ overexpressed |
| rs7074516 | GTGTGAG | chr10 | 98344914 | chr10:98344912-98344918 | ER- overexpressed |
| rs7922412 | CAAGGGA | chr10 | 124805384 | chr10:124805379-124805385 | ER+ overexpressed |
| rs10279 | GTATTAT | chr11 | 8925885 | chr11:8925881-8925887 | ER+ overexpressed |
| rs1056562 | AAGGGAT | chr11 | 117630835 | chr11:117630830-117630836 | ER- overexpressed |
| rs10790248 | ACACTAC | chr11 | 117630882 | chr11:117630877-117630883 | ER- overexpressed |
| rs12288903 | AAGTCCA | chr11 | 45860170 | chr11:45860164-45860170 | ER+ overexpressed |
| rs3741325 | GTGCCAT | chr11 | 117911199 | chr11:117911194-117911200 | ER+ overexpressed |
| rs3741360 | TCCAGAG | chr11 | 66056924 | chr11:66056919-66056925 | ER+ overexpressed |
| rs8432 | AGCTGCT | chr11 | 66056091 | chr11:66056091-66056097 | ER+ overexpressed |
| rs8995 | CCACCCC | chr11 | 63351648 | chr11:63351645-63351651 | ER- overexpressed |
| rs2857672 | CACCAGC | chr12 | 50994544 | chr12:50994544-50994550 | ER- overexpressed |
| rs859147 | CTCTATG | chr12 | 25152535 | chr12:25152535-25152541 | ER+ overexpressed |
| rs1327179 | ATACTGT | chr13 | 20626320 | chr13:20626318-20626324 | ER- overexpressed |
| rs403904 | AAGGCAT | chr13 | 35144233 | chr13:35144228-35144234 | ER+ overexpressed |
| rs1565970 | AGTCTTA | chr14 | 51967826 | chr14:51967824-51967830 | ER+ overexpressed |
| rs10468050 | AGGCACT | chr15 | 69860993 | chr15:69860990-69860996 | ER+ overexpressed |
| rs16956198 | CTGTTGA | chr15 | 69858995 | chr15:69858992-69858998 | ER+ overexpressed |
| rs17811309 | AAAGGGA | chr15 | 69860243 | chr15:69860239-69860245 | ER+ overexpressed |
| rs2072692 | GGGATGC | chr15 | 87816037 | chr15:87816035-87816041 | ER- overexpressed |
| rs30122 | CATTAAC | chr16 | 14266734 | chr16:14266731-14266737 | ER+ overexpressed |
| rs30126 | GAGACGG | chr16 | 14263266 | chr16:14263262-14263268 | ER+ overexpressed |
| rs1051443 | TTAGCTC | chr17 | 6294757 | chr17:6294751-6294757 | ER- overexpressed |
| rs7687 | TTCCCCC | chr17 | 41459142 | chr17:41459141-41459147 | ER+ overexpressed |
| rs1046294 | ACAACCT | chr19 | 40352386 | chr19:40352380-40352386 | ER- overexpressed |
| rs12427 | GCTGGAG | chr19 | 48962659 | chr19:48962655-48962661 | ER- overexpressed |
| rs12891 | GAGCCAG | chr19 | 8233196 | chr19:8233196-8233202 | ER+ overexpressed |
| rs7257398 | AAGCACA | chr19 | 59433680 | chr19:59433674-59433680 | ER- overexpressed |
| rs2287086 | GTGCAAA | chr2 | 60539999 | chr2:60539993-60539999 | ER- overexpressed |
| rs6729137 | AATGCAT | chr2 | 5757912 | chr2:5757907-5757913 | ER- overexpressed |
| rs6737419 | GTGCAAA | chr2 | 3570576 | chr2:3570570-3570576 | ER- overexpressed |
| rs873033 | TCTAGAG | chr2 | 85390883 | chr2:85390879-85390885 | ER- overexpressed |
| rs1048055 | GTGCCAT | chr20 | 1558062 | chr20:1558057-1558063 | ER- overexpressed |
| rs2281807 | GAGCCAG | chr20 | 1558201 | chr20:1558195-1558201 | ER- overexpressed |
| rs6043374 | GTAAACC | chr20 | 1557952 | chr20:1557946-1557952 | ER- overexpressed |
| rs6091230 | ACTGCAG | chr20 | 48926602 | chr20:48926602-48926608 | ER+ overexpressed |
| rs2834602 | TTACTAG | chr21 | 35012300 | chr21:35012294-35012300 | ER+ overexpressed |
| rs12172608 | AGGTGCA | chr22 | 45137237 | chr22:45137236-45137242 | ER+ overexpressed |
| rs6007891 | CAGTTTT | chr22 | 45135497 | chr22:45135493-45135499 | ER+ overexpressed |

Supplementary Table 2

SNPs Mapping to miRNA Sites

| | | | | | |
|---|---|---|---|---|---|
| rs495702 | CACTTCA | chr3 | 173598334 | chr3:173598328-173598334 | ER- overexpressed |
| rs1046322 | GAGTGAC | chr4 | 6355349 | chr4:6355347-6355353 | ER+ overexpressed |
| rs1434536 | CTCAGGG | chr4 | 96294988 | chr4:96294988-96294994 | ER+ overexpressed |
| rs1141538 | ATGCTGC | chr5 | 137303098 | chr5:137303092-137303098 | ER+ overexpressed |
| rs1438688 | CCCCGCC | chr5 | 148619255 | chr5:148619255-148619261 | ER+ overexpressed |
| rs1062577 | ATTCTTT | chr6 | 152465598 | chr6:152465594-152465600 | ER+ overexpressed |
| rs1225737 | TGCCTTA | chr6 | 11090638 | chr6:11090633-11090639 | ER+ overexpressed |
| rs508477 | GGTGTGT | chr6 | 13472323 | chr6:13472320-13472326 | ER- overexpressed |
| rs7756717 | CTGAGCC | chr6 | 11090925 | chr6:11090921-11090927 | ER+ overexpressed |
| rs8523 | ATTTCTC | chr6 | 11089039 | chr6:11089037-11089043 | ER+ overexpressed |
| rs9341070 | ACACTCC | chr6 | 152461890 | chr6:152461884-152461890 | ER+ overexpressed |
| rs9341074 | AATGGGT | chr6 | 152464148 | chr6:152464142-152464148 | ER+ overexpressed |
| rs10263074 | TTTATCT | chr7 | 87375999 | chr7:87375994-87376000 | ER- overexpressed |
| rs6616 | ATTTCTC | chr7 | 16790503 | chr7:16790502-16790508 | ER+ overexpressed |
| rs4986994 | AACTGAC | chr8 | 18124767 | chr8:18124761-18124767 | ER+ overexpressed |
| rs12710570 | GTGCAAT | chrX | 115506264 | chrX:115506260-115506266 | ER- overexpressed |
| rs6567569 | TGCAGAA | chrX | 3534309 | chrX:3534306-3534312 | ER- overexpressed |
| rs741500 | TGTTACT | chrX | 10377020 | chrX:10377016-10377022 | ER- overexpressed |

* UCSC Build 36 coordinates

† Defined in Smith, et al BMC Bioinformatics 9:63,2008

Supplementary Table S3-Original CGEMS Association Findings

| | | CGEMS cases (n=1145) | | | CGEMS controls (n=1142) | | | CGEMS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SNP | GT | Count | Prop | HWE | Count | Prop | HWE | OR | 95% C.I. | P-value[*] | Rank[†] |
| rs1219648 | A/A | 351 | 0.31 | 0.152 | 432 | 0.38 | 0.851 | 1.00 | | 8.0E-06 | 2 |
| Het | A/G | 542 | 0.47 | | 534 | 0.47 | | 1.25 | 1.04-1.50 | | |
| Minor | G/G | 249 | 0.22 | | 169 | 0.15 | | 1.81 | 1.42-2.04 | | |
| rs1970801 | G/G | 165 | 0.15 | 0.950 | 215 | 0.30 | 0.337 | 1.00 | | 1.7E-04 | 79 |
| Het | T/G | 534 | 0.47 | | 577 | 0.51 | | 1.21 | 0.95-1.52 | | |
| Major | T/T | 436 | 0.38 | | 344 | 0.19 | | 1.65 | 1.29-2.11 | | |
| rs6831418 | C/C | 250 | 0.22 | 0.005 | 374 | 0.33 | 0.028 | 1.0 | | 1.2E-04 | 52 |
| Het | C/T | 515 | 0.45 | | 588 | 0.52 | | 0.88 | 0.73-1.06 | | |
| Minor | T/T | 373 | 0.33 | | 175 | 0.15 | | 1.43 | 1.13-1.82 | | |

[*] Unadjusted P-value from the score test with df=2 in Logistic regression from original CGEMS GWAS

[†] Rank of the SNPs from original CGEMS GWAS, for rs1970801, test was for minor allele G

# Paper III

# Single Nucleotide Polymorphisms Can Create Alternative Polyadenylation Signals and Affect Gene Expression through Loss of MicroRNA-Regulation

Laurent F. Thomas[1,3], Pål Sætrom[1,2,3,*]
**1 Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, N-7489 Trondheim, Norway**
**2 Department of Computer and Information Science, Norwegian University of Science and Technology, N-7489 Trondheim, Norway**
**3 Interagon AS, Laboratoriesenteret, NO-7006 Trondheim, Norway**
**∗ E-mail: pal.satrom@ntnu.no**

## Abstract

Alternative polyadenylation (APA) can for example occur when a protein-coding gene has several polyadenylation (polyA) signals in its last exon, resulting in messenger RNAs (mRNAs) with different 3' untranslated region (UTR) lengths. Different 3'UTR lengths can give different microRNA (miRNA) regulation such that shortened transcripts have increased expression. The APA process is part of human cells' natural regulatory processes, but APA also seems to play an important role in many human diseases. Although altered APA in disease can have many causes, we reasoned that mutations in DNA elements that are important for the polyA process, such as the polyA signal and the downstream GU-rich region, can be one important mechanism. To test this hypothesis, we identified single nucleotide polymorphisms (SNPs) that can create or disrupt APA signals (APA-SNPs). By using a data-integrative approach, we show that APA-SNPs can affect 3'UTR length, miRNA regulation, and mRNA expression—both between homozygote individuals and within heterozygote individuals. Furthermore, we show that a significant fraction of the alleles that cause APA are strongly and positively linked with alleles found by genome-wide studies to be associated with disease. Our results confirm that APA-SNPs can give altered gene regulation and that APA alleles that give shortened transcripts and increased gene expression can be important hereditary causes for disease.

## Author Summary

Variants in DNA that affect gene expression—so-called regulatory variants—are thought to play important roles in common complex diseases, such as cancer. In contrast to variants in protein-coding regions, regulatory variants do not affect protein sequence and function. Instead, regulatory variants affect the amount of protein produced. The 3' untranslated region (UTR) is one gene region that is critically important for gene regulation; cancers for example, often express genes with shortened 3'UTRs that, compared with full-length 3'UTRs, have higher and more stable expression levels. We have investigated one kind of regulatory variant that can affect the 3'UTR length and thereby cause disease. We identified several such variants in different genes and found that these variants affected the genes' expression. Some of these variants were also strongly linked with known markers for disease, suggesting that these regulatory variants are important hereditary causes for disease.

## Introduction

In protein-coding genes, the polyadenylation process consists of cleaving the end of 3' untranslated regions (UTR) of precursor messenger RNAs (pre-mRNA) and adding a polyadenylation (polyA) tail. Alternative polyadenylation (APA) can occur when several polyadenylation (polyA) signals lie in the last exon of a protein-coding gene. Many APA signals are evolutionary conserved [1], and Expressed Sequence Tag

(EST) data suggest that 54% of human genes have alternative polyadenylation signals [1]. The polyA signals themselves are hexamer DNA sequences that usually lie 10 to 30 nucleotides upstream from the cleavage site [2], but a GU-rich region 20 to 40 nucleotides downstream of the cleavage site is also important for the polyA-process [2].

One functional consequence of APA is transcripts with different 3'UTR lengths and different microRNA (miRNA) regulation [3, 4]. Shortened transcripts tend to have increased expression compared with longer transcripts, and the same expression increase can be achieved by deleting miRNA target sites in non-shortened transcripts [5].

Data on APA can be used as an efficient biomarker for distinguishing between cancer subtypes and for prognosis [6], and seems to play an important role in gene deregulation and in many human diseases [7]. One such mechanism for deregulation is mutations in the polyA signal or GU-rich downstream region [7]. A single nucleotide polymorphism (SNP) in the GU-rich region downstream of an alternative polyA signal in the *FGG* gene has for example been shown to affect the usage of this polyA site, and has been associated with increased risk for deep-venous thrombosis [8]. Similarly, a mutation in the 3'UTR of the *CCND1* gene has been shown to create an alternative polyA signal and is associated with increased oncogenic risk in mantle cell lymphoma [9].

Hypothesizing that mutations in DNA elements such as the polyA signal can be an important cause of altered APA, we investigated to what extent SNPs can create or disrupt APA signals (APA-SNPs). Specifically, we tested whether APA-SNPs can give shorter 3'UTRs, increased gene expression through loss of miRNA regulation (Fig. 1), and be associated with disease. Our hypothesis focuses on shorter 3'UTRs rather than longer ones, because the loss of functional miRNA sites in the 3'UTR is more likely than the gain of new sites downstream of the gene.

First, by analysing EST data, we found that SNPs can create polyA motifs and affect 3'UTR length. Second, differential allelic expression from RNA-seq data, as well as mRNA and miRNA microarray expression data revealed an association between alternative polyA site strength (signal and GU-content), loss of miRNA target sites, and transcript expression. Third, based on these analyses we also identified significant APA-SNPs. Fourth, we mapped the identified SNPs to disease-associated SNPs and found that APA alleles were significantly correlated with disease-risk alleles. Together, these results suggest that APA-SNPs can be a significant causative mechanism in disease (Fig. S1).

# Results

## SNPs can create and delete polyadenylation signal motifs

The distribution of SNPs within 3'UTRs is fairly uniform [10] (Fig. S2A). The main exceptions are microRNA target sites and the start and end of the 3'UTR, which have decreased SNP diversity that is consistent with these regions containing functional elements under selective pressure [10]. Indeed, when specifically investigating the region around the transcription end site, we found that the position containing the polyA signal has a markedly decreased SNP density (Fig. S2B,C), indicating that SNPs arising there could have a high functional impact.

To analyse SNPs in alternative polyadenylation signals, we first identified a set of SNPs that potentially create new APA signals in 3'UTRs. Specifically, we searched for any Hapmap SNP [11] that could create or disrupt one of the 13 known polyA signal hexamers [1] in any coding gene's 3'UTRs (see Methods). We found 1954 SNPs, including 755 SNPs that are mono-allelic in the CEU population from Hapmap [11] (see Datasets). We kept only APA-SNPs that change from no signal to one signal in the locus, by discarding loci with several signals in the 40 nucleotides around the SNP, discarding SNPs that change one signal into another, and discarding mono-allelic SNPs. After filtering, 412 SNPs that can create or delete potential polyadenylation signals remained. We will from now refer to them as our candidate SNPs.

## EST data indicate that SNPs can give functional alternative polyA sites

To investigate whether SNPs can create functional alternative polyA sites, we analysed the EST-based polyA sites from the PolyA_Db database [12, 13]. In the PolyA_Db database, there are several polyA sites which do not have any noticeable polyA signal (according to the reference genome) in the 40, 80, and 100 nucleotides upstream from the reported cleavage site position (Table S1). In those regions, we used different SNP data to look for SNPs that could create a polyA signal with the non-reference allele. When considering regions of 100 nucleotides and SNPs from NCBI dbSNP Build 130 [14], we could identify polyA signals with the alternative allele for more than 2% of the missing signals. Some of the remaining sites can probably be explained by SNPs further upstream, and some other by exon splicing, by alterations in ESTs that are not registered in dbSNP, or as false positive sites.

Since EST-based annotated polyA sites can be affected by SNPs, we wanted to test whether alleles in polyA sites could be associated with EST ending positions. Specifically, we first took the intersection between the polyA signals from our 412 candidate SNPs, and the polyA sites from PolyA_Db database [12, 13]. We identified 18 intersecting polyA sites, that have a polyA signal with either the reference or the non-reference allele, corresponding to 18 genes, and 18 SNPs. Five SNPs were discarded because they lie within the 20 last nucleotides of the reference 3'UTR. The following 13 genes remained: *ABCC4, AKAP13, FANCD2, KY, MIER1, OSTM1, PNN, RASGRP3, RHOJ, SELS, SHMT1, SLBP*, and *SLC11A2*. Second, for each of these genes, we identified and imputed (see Methods) alleles at the SNPs in the EST sequences when possible, and tested if the proportion of alleles with polyA signal (APA alleles) was different for EST sequences ending within the interval $[-30, +50]$ nucleotides around the polyA site, compared to EST sequences ending further downstream (see Methods). The two genes *MIER1* (SNP rs17497828) and *PNN* (SNP rs532) were significant (Fig. 2, Table 1). After correcting for multiple testing (Benjamini & Hochberg correction), the genes remained significant when including alleles imputed based on haplotype (Table 1).

For *MIER1*, of the 16 EST sequences ending near the annotated APA site, 12 had the APA allele (including 2 with a clear polyA tail), whereas 3 had the non-APA allele (none of them had a clear polyA tail). Similarly, for *PNN*, of the 34 EST sequences ending near the annotated APA site, all had the APA allele (including 10 with a clear polyA tail). Together, these results suggest that SNPs can create functional APA sites and thereby affect 3'UTR length.

## RNA-seq data indicate association between SNPs in polyA sites and both transcript length and increased transcript expression

EST data can be used to identify alleles and transcript ending positions (Fig. 1), but EST data seldom have sufficient resolution to quantify transcript expression levels. In contrast, RNA-seq data can both be used to genotype SNPs [15] and to analyse transcript length and expression patterns. The main challenge with RNA-seq data compared with ESTs, however, is the shorter sequence reads, which makes it challenging to distinguish between homozygotes, heterozygotes with strong expression differences between its alleles (allelic imbalance), sequencing errors, and alignment errors.

To explore whether RNA-seq data could reveal whether APA-SNPs affect transcript expression, we therefore developed and validated an RNA-seq-based genotyping approach (see Supporting Text S2). We then used this approach to show that APA-SNPs can affect transcript expression and that this effect is associated with loss of miRNA regulation. Specifically, we first show that homozygous APA-SNPs have significantly shorter 3'UTRs than have heterozygous or homozygous wildtype SNPs. Second, we show an association between allelic imbalance of heterozygous APA-SNPs and the two following important features of polyA sites: signal strength and GU level downstream of the cleavage site. Third, we show that the loss of miRNA target sites can be the missing link in this association. Fourth, we use allelic imbalance to detect potential functional APA-SNPs. Fifth, we show that APA-SNPs at strong sites (strong APA signal and high GU level) that have a strong predicted effect on miRNA regulation, have higher allelic imbalance and higher transcript expression than have other APA-SNPs.

### Transcripts are shorter for genes with homozygous APA-SNPs

RNA-seq data give the opportunity to both genotype exonic SNPs and determine transcript structure. We therefore decided to use the Burge RNA-seq data to determine whether APA-SNPs had a significant effect on transcript length. Moreover, as the Burge RNA-seq data set consists of samples from both highly proliferating cell lines and highly differentiated human tissues and as transcripts in proliferating cells tend to have shorter 3'UTRs, we also wanted to determine the effect that the cell's proliferation status had on transcript length. Specifically, we first estimated for each gene and RNA-seq sample, the transcript's 3' end position and its distance from the 3' end of the longest annotated transcript (see Methods). Second, we divided the RNA-seq samples into two groups such that the five cancer cell lines and one immortalized cell line were defined as proliferating, whereas the 16 other tissue samples were defined as non-proliferating.

Third, from the 412 candidate APA-SNPs, we discarded those that share the same gene and those that lie upstream of the longest 3'UTR (to avoid combinations of alternative splicing and alternative polyadenylation), resulting in 362 SNPs. In total, 262 unique SNPs had 3' end estimates and genotypes available (6852 data points). To analyse the impact of APA-SNPs on 3' end positions, we only considered SNPs that lie far enough (at least 1500 bp upstream) from the annotated 3'end. This final requirement gave 93 unique SNPs (2340 data points).

Fourth, we ran correlation analyses between the genotype (WT homozygous: 0, heterozygous: 1, and APA homozygous: 2) and the negative logarithm of the distance between the estimated and the annotated transcript end (see Methods). As expected, we found a significant negative correlation (Pearson's correlation coefficient $r = -0.15$, p-value $p = 3.9 * 10^{-13}$, sample size $n = 2340$), which shows that APA homozygotes are shorter than the WT ones. Then, we tested the correlation between the negative log distance and the proliferation status of the cell types (proliferating: 1; non-proliferating: 0). Again, as expected, we found a significant negative correlation ($r = -0.19$, $p < 2.2 * 10^{-16}$, $n = 2340$). When subgrouping the samples based on proliferation status (Fig. S3), we could not detect a significant genotype correlation in the proliferating cells—possibly because transcripts are already short in these cells due to other factors. For non-proliferating cells, however, we found that APA homozygotes were significantly shorter than the two other genotypes ($r = -0.17$, $p = 1.01 * 10^{-13}$, $n = 1883$). This result confirms our previous EST results that APA SNPs can affect transcript length.

### Heterozygous SNPs affecting strong polyA sites have an increased imbalance towards APA alleles

Since RNA-seq data can genotype our candidate SNPs and at the same time determine transcript expression levels, we decided to analyse ratios of allele expression (allelic imbalance). According to our hypothesis (Fig. 1), APA alleles can shorten transcripts, resulting in loss of miRNA targeting and increased transcript expressions. To test this hypothesis, we investigated allelic imbalance of our APA-SNPs in 19 of the samples from the Burge RNA-seq data; we excluded the three samples (MAQC, MAQC UHR, and MD435) that were a mixture of several individuals. We expected increased transcript expression for the APA allele compared to the non-APA allele; that is, a positive log ratio of the APA allele expression over the non-APA allele expression. Moreover, we expected this allelic imbalance to depend on two important polyA site features: polyA signal strength and downstream GU level.

**Signal strength** Some polyadenylation signals occur more frequently upstream of known polyA sites than other signals do [1]. By assuming that this frequency of occurrence correlate with signal strength, such that frequent signals have a higher probability of causing polyadenylation than have rare signals (Table S4), we expected that frequent (strong) signals would have a higher allelic ratio (AR) than rare (weak) signals. We compared the distribution of allelic ratios of APA allele over non-APA allele for each signal, ordered by strength rank such that strong (frequent) signals had a low rank. As expected, we found that signal rank is negatively correlated with log allelic ratio ($r = -0.144$, $p = 0.013$, $n = 300$) (Fig. 3A). Strong signals tend to have high and positive log AR; that is, a higher expression of the APA allele than of the non-APA allele. This fits our hypothesis that transcripts with an APA allele can escape miRNA targeting, resulting in increased gene expression.

**GU-content** In addition to having strong polyA signals, functional polyA sites tend to have a GU-rich region downstream of the cleavage site [2]. We therefore expected that SNPs creating alternative polyadenylation signals with a GU-rich region downstream of the signal had a higher allelic imbalance than the ones with no particular GU-rich region.

We computed the GU level for each of our candidate SNPs. As the background value outside the GU-rich region is about 0.51 (Fig. S4), we used a threshold of 0.55 to define SNPs that have a downstream GU-rich region. Then, in each of the two GU groups, we investigated the allelic ratio distribution for each signal. We still found a negative correlation between the signal rank and the log allelic ratio for the SNPs with a GU-rich region ($r = -0.195$, $p = 0.032$, $n = 122$) (Fig. 3B). In contrast, for the SNPs without a GU-rich region, log AR did not correlate with signal rank ($r = -0.104$, $p = 0.17$, $n = 178$) (Fig. 3C). This indicates that increased allelic imbalance at APA-SNPs requires both a strong signal and a GU-rich downstream region.

To further evaluate the connection between signal strength, GU level, and allelic imbalance, we grouped the SNPs according to their GU level and their signal strength (Fig. 4; strong: rank $\leq 6$; weak: rank $> 6$). Compared with the other three groups, APA-SNPs with a strong signal and a GU-rich region had a significantly higher mean and median log AR (Student's t-test, $p = 0.025$; Wilcoxon rank sum test, $p = 0.036$). Together, these results suggested that alternative polyadenylation can give increased expression of APA alleles.

### The loss of miRNA target sites can explain an important part of allelic imbalance

Increased expression of APA allele transcripts is consistent with loss of miRNA regulation, but other factors such as RNA-binding proteins could potentially also explain these results. We therefore wanted to test whether loss of miRNA regulation could be a significant factor in the increased allelic imbalance. Specifically, we matched the miRNA expression data of the MCF7, BT474, and T47D breast cancer cell lines from Landgraf *et al.* [16] with the allelic ratios from the corresponding cell lines from the Burge dataset (24 unique SNPs, 34 allelic imbalance values in the 3 cell lines). Given the miRNA profile of the considered cell line, we then for each SNP computed a score which predicted the potential effect that a cleavage site at the SNP locus would have on miRNA regulation (see Methods). Finally, we ran several linear regression analyses with the log allelic ratios as response variable and the signal rank, the GU level, and the miRNA score difference as dependent variables.

Basic linear models with signal rank, GU level, and miRNA score alone showed that these variables could explain 3.84%, 3.55%, and 6.73% of the response variance, respectively. A model with signal rank and GU level decreased the partial explained variance for each of the two variables compared to the two individual models. In contrast, adding the miRNA variable to the Signal rank model, or the GU level model increased all partial $r^2$ values, indicating that the dependent variable is a conjunction of these variables. Similarly, adding the miRNA variable to the Signal rank + GU level model could increase all the partial $r^2$ as well. In that full model, the miRNA variable could explain 12.39% of the response variance (p-value $p = 0.04$). This indicates that loss of miRNA target sites can partly explain the increased allelic imbalance for APA-SNPs in strong APA signals with high downstream GU content.

### Allelic expression can detect potentially functional APA-SNPs

Having established that APA-SNPs can give allelic imbalance by affecting miRNA regulation, we set out to identify functional candidate APA-SNPs. We identified SNPs from the 19 non-mixed samples from the Burge dataset that were classified as heterozygous when mapping reads to both the reference and non-reference allele-based genomes and that had at least 10 allele counts in total. This resulted in 75 individual/SNP pairs (36 unique SNPs), which we tested individually for significant positive imbalance; that is, an APA allele count significantly greater than the non-APA allele count. We used a $\chi^2$ goodness-of-fit test (1 degree of freedom) to test if the allele counts fit the hypothesis of an equal proportion. Three heterozygotes were significant and after correcting for multiple testing by using the Benjamini & Hochberg correction, two heterozygotes remained significant. The two individual/SNP pairs had both a positive log-ratio, a GU-rich region and a strong APA signal. After correcting with the more stringent Bonferonni method, the same two pairs remained. Those two individual/SNP pairs were actually the same SNP (rs2269123 in gene *MRPS34*) from two breast cancer cell lines (BT474 and MCF-7; p-values $5.53 * 10^{-3}$ and $1.09 * 10^{-11}$, respectively), suggesting that this SNP gives a functional APA signal that strongly affects host gene expression.

### MicroRNAs link higher proportion of APA alleles to higher gene expression

Since heterozygous SNPs in strong APA signals can have an increased imbalance towards APA alleles, we investigated whether positive allelic imbalance can be associated with increased gene expression; that is whether a higher proportion of APA alleles than non-APA alleles was associated with an increased total allele count. We focused on the 12 samples from the Burge dataset that we could match to miRNA expression data in similar cell types from Landgraf *et al.* [16]; these were the 3 breast cancer cell lines (MCF7, BT474 and T47D), and the liver, heart, testis, and 6 cerebellum samples. In those 12 samples, we identified 174 allelic ratios (97 unique SNPs) that were classified as heterozygous when mapping to both the reference and non-reference allele based genomes. Given the miRNA profile, we then assigned a miRNA score which predicted the potential effect that a cleavage site at the SNP locus would have on miRNA regulation (see Methods).

Based on the 174 allelic ratios, we compared SNP expression (sum of APA and non-APA allele counts) for groups with higher APA allele proportion (positive log AR) with groups with higher non-APA allele proportion (negative log AR; Fig. 5). We found that SNPs with strong APA signal, high GU level, and high miRNA score had a significant log SNP expression difference between positive log ratios and negative log ratios. This indicates that APA alleles of SNPs with strong APA sites and high miRNA scores can upregulate gene expression (Fig. 6). This links positive allelic imbalance to higher gene expression.

## MicroRNAs link genotype to increased gene expression

To confirm the results from the RNA-seq-based allelic imbalance analyses, we turned to gene expression data from the well characterised Hapmap population. We looked at human gene expression profiling of EBV-transformed lymphoblastoid cell lines from 270 unrelated Hapmap individuals [17], and genotypes of the same individuals, from the Hapmap database [11]. Specifically, we first investigated whether genotypes of SNPs in strong polyA sites that affect miRNA targeting in general are associated with increased gene expression. Second, we investigated whether individual APA-SNP genotypes correlate significantly with gene expression.

### Genotype of SNPs in strong polyA sites and the loss of miRNA target sites can explain increased gene expression

From the Hapmap expression profiles and our 412 potential APA-SNPs, we identified 333 SNPs that could be mapped to 315 unique probe IDs. Discarding SNPs sharing the same probe IDs, resulted in 299 unique SNPs and probe IDs. We then used human miRNA expression profiles from EBV-transformed lymphoblastoid cell lines [18], to compute a miRNA score that quantifies the potential effect of a cleavage site at each SNP locus on miRNA regulation (see Methods).

Simple regression analyses with mRNA expression as response variable and with each of genotype, signal rank, local GU level downstream of the signal, and miRNA score as dependent variables, found that the GU level explained the most of the mRNA variance ($r^2 = 3.3\%$). We therefore computed the GU level in the whole 3'UTR and ran a regression of the mRNA expression on this variable. Surprisingly, we found that this variable was positively correlated with higher gene expression for our 299 genes ($\rho = 0.285$, $p = 5.3 * 10^{-7}$) and could explain 7% of the response variance. One possible explanation is that non-canonical polyA sites are thought to rely mostly on downstream GU-rich elements [19]. If this explanation is true we could expect that genes with increased GU level in 3'UTR can have a higher number of APA sites, which could result in generally higher mRNA expression. Indeed, based on polyA_Db, we found that 3'UTR GU level is positively correlated with the number of polyA sites in each gene ($\rho = 0.193$, $p < 2.2*10^{-16}$, $n = 13181$). Moreover, the number of polyA sites is also positively correlated with mRNA expression from microarray data ($\rho = 0.200$, $p < 2.2 * 10^{-16}$, $n = 11756$). Expectedly, longer 3'UTRs are more likely to have more polyA sites (correlation coefficient $\rho = 0.333$, $p < 2.2 * 10^{-16}$, $n = 17298$). However, we also found that the GU level is correlated with 3'UTR length ($\rho = 0.192$, $p < 2.2 * 10^{-16}$, $n = 17934$). All these results suggest that the 3'UTR GU level is a confounding variable giving increased APA and thereby mRNA expression. We therefore analysed mRNA expression data after correcting for the general 3'UTR GU level; *i.e.* we regressed the mRNA expression on the 3'UTR GU content and used the residuals as the new response variable.

When comparing residual gene expression medians for the 3 genotypes (Fig. 7), we found that increased expression correlates with the number of APA alleles in the genotype and that SNPs with strong APA signal (S) had a significant gene expression median difference between the 3 genotypes (Fig. 7 A). This was particularly evident for SNPs with high miRNA score (Fig. 7 B), which are those that are supposed to have the strongest effect on miRNA regulation. Furthermore, a multiple regression on transcript length from the Burge RNA-seq data showed that APA homozygotes, cell proliferation, strong signals, and local and global GU levels, all contribute significantly to reduced transcript lengths (Table S5). Together, these results indicate that APA alleles of SNPs with strong APA sites and high miRNA scores can upregulate gene expression and link APA homozygotes to increased gene expression.

### Gene expression can detect potential functional APA-SNPs

Since genotype of SNPs in strong polyA sites and the loss of miRNA target sites can be associated with increased gene expression, we decided to use correlation to detect potential functional APA-SNPs. Of the 333 candidate SNPs that mapped to gene probes, we discarded SNPs that were in genes whose maximum expression value among the 270 individuals was lower than the total expression median, to remove from the analysis genes that are very low or unexpressed in all the individuals. 243 SNPs remained and we tested these separately in a correlation analysis of genotype and mRNA expression.

We found 47 SNPs (on 47 genes) that were significantly different from 0 (see Table S6). All had a positive coefficient, indicating a positive correlation between genotype and gene expression. This fits both previous results where APA was associated with increased expression levels [4] and our RNA-seq results. After correcting for multiple testing with the Benjamini & Hochberg correction, 19 SNPs remained significant; 13 SNPs remained if correcting with stringent Bonferroni correction.

## Potential functional APA alleles are positively correlated with risk alleles from disease-associated SNPs

Since SNPs can alter polyadenylation and affect miRNA target sites and gene expression, we wondered whether they can also play an important role in human diseases. We therefore tested if any of our APA-

SNPs were linked to trait-associated SNPs from the NHGRI GWAS catalogue [20, 21], which consists of SNP-trait associations from published genome-wide association studies (GWAS) (accessed Apr. 18, 2011). Specifically, we mapped our 412 APA-SNPs to the 4304 GWAS SNPs, by using the mapping method described in Thomas *et al.* [22]. The mapping was based on linkage disequilibrium (LD) data from the Hapmap database (CEU population release 27). We identified 135 APA-SNP/GWAS-SNP pairs (consisting of 84 unique APA-SNPs and 123 unique GWAS SNPs) that had available haplotype data in Hapmap and one known and unique risk allele in the GWAS catalogue. For each APA-SNP/GWAS-SNP pair, we computed the correlation between the APA allele and risk allele as the LD value $r = \frac{p_{AR}-p_A*p_R}{\sqrt{p_A*(1-p_A)*p_R*(1-p_R)}}$ [23], where $p_A$, $p_R$, and $p_{AR}$ are respectively the APA allele frequency of the APA-SNP, the risk allele frequency of the GWAS SNP, and the "APA allele risk allele" haplotype frequency in the CEU Hapmap population. For each of the 84 unique APA-SNPs, we computed $\hat{r}$ as the mean of $r$ when an APA-SNP was linked to several GWAS SNPs, and similarly $\hat{r^2}$ as the mean of $r^2$.

We hypothesised that if APA-SNPs play a role in diseases, then APA alleles would be positively ($\hat{r} > 0$) and strongly (high $\hat{r^2}$) correlated with risk alleles, particularly for the significant APA-SNPs that we identified in the previous sections, as they are more likely to be functional, and particularly those that are linked to GWAS-SNPs from CEU-population-related studies, since the $r$ values are based on CEU haplotypes.

Among the 84 APA-SNPs, 60 were paired to GWAS-SNPs that are trait-associated in CEU-related populations. Nine of those SNPs were identified in the previous sections as significant APA-SNPs, and those nine SNPs had a significantly high number of positive $\hat{r}$ (more positive correlations between APA and risk alleles than expected) and a significantly high number of $\hat{r^2}$ greater than 0.2 (higher number of correlations between APA and risk alleles than expected) (Table 2). In contrast, for $\hat{r}$ computed from CEU haplotypes but for GWAS-SNPs that are trait-associated in non-CEU-related populations, binomial test p-values were not significant, suggesting that GWAS and haplotype data should be matched according to population, to detect potential disease-related APA-SNPs.

Those results show that a significantly high proportion of our candidate SNPs is in LD with trait-associated SNPs and their APA alleles are positively correlated with risk alleles of trait SNPs. This suggests that those APA-SNPs can potentially be the cause of their corresponding disease-association signals measured and registered in the GWAS catalogue.

## Discussion

Our analyses confirmed the hypothesis (presented in Fig. 1) that SNPs can create functional alternative polyadenylation signals and thereby affect miRNA-based gene regulation and give increased gene expression. Both EST and RNA-seq analyses supported our hypothesis, despite some limitations. Additionally, the microarray analysis could also confirm those results and strengthen our hypothesis. Given those results, we estimate the proportion of functional APA-SNPs to be $(2 + 1 + 13)/(13 + 36 + 243) = 0.055$ (5.5%).

The EST analysis supports our hypothesis but has some limitations. Specifically, we analysed EST data for 13 genes and found that 2 of them had an APA-SNP that could create polyA motifs and affect 3'UTR length. However, the EST analysis does not take into account the presence of a polyA tail in the EST sequence. Moreover, the ESTs came from a mix of tissues, which could also affect the results. Segregating ESTs based on tissue origin or filtering on sequences with clear tails in the "short" group, reduces sample size and affects statistical power. However, when combining sequences from our two significant genes, all of the 12 EST sequences ending at the alternative cleavage site and that have a polyA tail, had the APA allele. This number is significant (binomial test p-value of 0.024, where the expected proportion of the APA allele is the combination of weighted allele frequencies of APA alleles for the 2 SNPs), and tells that the shortened transcripts arose from functional APA signals from the APA alleles.

Similarly, RNA-seq data from the Burge Lab, matched to miRNA expression data showed association between alternative polyA site strength (signal and GU-content), loss of miRNA target sites, allelic imbalance, and transcript expression. The Burge dataset was generated by using cDNA fragmentation, which gives a good coverage of 3'UTRs [24]. An increased allelic imbalance towards the APA allele could come from the loss of miRNA target sites, but also from the fragmentation method. This is because cDNA fragmentation gives a good coverage at the end of the transcript, and, in case of alternative polyadenylation, the transcript is shorter for the APA allele, which results in a high coverage at the SNP locus. In contrast, a longer transcript with the non-APA allele could have a higher coverage downstream, but a lower coverage at the SNP locus. Bias from cDNA fragmentation would therefore give an increased allelic ratio towards the APA allele simply because of transcript length differences. Consequently, we cannot exclude that some of the overall RNA-seq trends can be attributed to cDNA fragmentation bias.

The independent microarray data strongly support the EST and RNA-seq results, however. Specifically, the mRNA and miRNA expression data from microarray showed association between alternative polyA site strength (signal and GU-content), loss of miRNA target sites, and transcript expression. This microarray analysis had the advantage of directly using genotype data from Hapmap, instead of genotyping SNPs through mapped RNA-seq reads. Furthermore, the microarray analysis focused on transcript expression differences between individuals and therefore required data from a unique cell type, whereas the RNA-seq analysis focused on allelic expression differences within a sample and could therefore involve different cell types. As expected, the microarray analysis showed similar results as the RNA-seq analysis, suggesting that the increased allelic ratios from RNA-seq data did not come from a potential bias due to the cDNA fragmentation method, but from the loss of functional miRNA target sites.

One clear disadvantage of using the RNA-seq data for genotyping and allelic-imbalance-based detection, was false positive homozygotes. We could detect potential functional candidate SNPs by testing for allelic imbalance, which takes into account the number of reads and their quality, while testing for unusual allele proportion patterns. The difficulty was to find extreme allelic imbalance, as we could miss extreme imbalance by classifying a locus as homozygote because of too few reads ($< 15\%$) corresponding to the alternative allele. This was a conscious trade-off, however, since we wanted to maximise true positive heterozygotes and avoid false positives (*i.e.* predicted heterozygotes that were in fact homozygous).

RNA-seq data enabled us to genotype SNPs in expressed genes and compute allelic imbalance. Genotype classification could be checked with known genotypes from the Heap dataset and with mono-allelic SNPs. However the Heap dataset could not be used in the allelic imbalance analysis, because the library was generated by using RNA fragmentation, which gives a good coverage for the coding regions [24], but not for the UTRs. Since we were interested in SNPs in 3'UTRs, and particularly at the end of potential alternative transcripts, RNA fragmentation would affect allelic imbalance.

The whole analysis is limited to SNPs that can make the reference 3'UTR shorter, lose miRNA sites and upregulate genes, because the loss of functional miRNA sites within the 3'UTR is more likely than the gain of new ones downstream of the annotated 3'UTR. However, it could be interesting to consider the hypothesis where SNPs in the signals at the end of the reference transcript could make 3'UTR longer having more miRNA target sites further downstream, and down-regulate the gene.

Alternative polyadenylation alleles play a role in 3'UTR shortening, gene deregulation, and increased disease risk (Fig. 1). Our analyses confirm that APA is an important factor for miRNA-mediated gene regulation [4]. EST data suggest that SNPs can create polyA motifs and affect 3'UTR length, and allelic imbalance from RNA-seq data coupled to miRNA expression data suggest an association between alternative polyA site strength (signal and GU-content), loss of miRNA target sites, allelic imbalance and transcript expression. Similarly, mRNA expression data from microarray and genotype of the same individuals, coupled with miRNA expression data could confirm association between alternative polyA site strength (signal and GU-content), loss of miRNA target sites, genotype and transcript expression.

Each of our analyses could also be used to detect potential functional APA-SNPs. Those detected APA-SNPs could be linked to GWAS-SNP markers. A significant part of those APA-SNPS had their

APA allele positively correlated with disease-risk alleles and we propose that these are potential disease-causative variants.

# Methods

## Datasets

We used SNP data from the CEU population (CEPH - Utah residents with ancestry from northern and western Europe) from the human haplotype map project (HapMap database [11]), release 22 for haplotype data, and release 27 for the genotype, allele, frequency, and linkage disequilibrium data. We used the human genome assembly version 18 (hg18) [25], RefSeq gene annotations (hg18 version), and EST sequences from the UCSC Genome browser [26]. We used human APA sites from PolyA_Db [12,13]. We used disease-associated SNPs from the NHGRI GWAS catalogue [20,21]. RNA-seq data came from Heap *et al.* [15] and from the Burge Lab [27]. Human miRNA profiles came from Landgraf *et al.* [16] (their Table S5) and from Wang *et al.* [18]. MicroRNA data came from the MirBase database release 16 [28].

## Candidate SNPs in alternative polyadenylation signals

Thirteen polyA signal motifs are known in human genes: AAUAAA, AUUAAA, UAUAAA, AGUAAA, AAGAAA, AAUAUA, AAUACA, CAUAAA, GAUAAA, AAUGAA, UUUAAA, ACUAAA, and AAUAGA [1] (ordered by strength ranks). We detected SNPs in potential APA signals, by a motif search that looks if any CEU Hapmap SNP in the 3'UTR of any coding gene would create/disrupt one of those 13 motifs. For a given SNP, the motif search looks for a given motif in an mRNA sub-sequence consisting of the SNP and its flanking sequences (6 nucleotides up/downstream), for each allele.

## PolyA_Db

We downloaded the 28.857 APA sites (human) from PolyA_Db [12,13] from the UCSC track (hg18) [26]. We downloaded knownToLocusLink.txt and knownToRefSeq.txt from UCSC (hg18) [26] to convert entrez gene ID to RefSeq gene ID. We took the intersection between our APA signals and polyA sites from PolyA_Db, by taking all the sites from PolyA_Db that lie up to 40 bp downstream of our signals.

## EST

For each of the 13 candidate genes, we downloaded the EST sequences (Expressed sequence tag) from UCSC (hg18, tables 'all_mrna' and 'all_est') [26] that lie within their 3'UTR region. We also downloaded their alignment to their reference mRNA sequence from UCSC [26], and the list of EST that support the considered polyA site from PolyA_Db2 [13]. We used sequence alignment to identify the allele and haplotype of each sequence, when possible. Otherwise, the APA-SNP allele was imputed, by using haplotypes from the CEU Hapmap population [11] (see Dataset). We tested the proportion of APA alleles that support the candidate APA site, versus longer transcripts, by using a 2x2 contingency table. If the 4 expected values were greater than 5: we used the 2x2 $\chi^2-$test, and Fisher's exact test otherwise.

## Allele imputation in EST data

Given a 3'UTR region of a gene of interest, we took all the phased SNPs from Hapmap [11] in that region, as well as their haplotypes in the CEU population [11]. For each of those SNPs, we identified the allele in the EST sequence when possible, to identify the EST haplotype. We discarded EST haplotypes that had zero identified allele. For each remaining EST haplotype, we selected haplotypes from Hapmap that fit the identified alleles in the EST haplotype. The APA-SNP could be imputed if there was only one unique allele at that SNP in all the selected haplotypes from Hapmap.

## RNA-seq data

We downloaded RNA-seq data from human primary *CD4*$^{+}$ *T* cells from 4 individuals [15] (Short read archive accession number: SRA008367), reads in FASTQ format, length of 45 bp. We downloaded Burge lab RNA-seq [27] (Short read archive: SRA002355, and Gene expression omnibus: GSE12946): Human tissue samples (brain, liver, heart, skeletal muscle, colon, adipose, testes, lymph node, breast, MAQC, 6 Cerebellum), immortalised and cancer cell lines (BT474, HME, MCF-7, MD435, T47D, MAQC UHR), reads in FASTQ format, length of 36 bp. MAQC is a mixture of brain cell types from several donors, MAQC UHR is a mixture of several cancer cell lines, and MD435 is thought to be contaminated by the M14 melanoma cell line. Therefore those 3 cell lines were discarded from the allelic imbalance analysis.

## RNA-seq mapping

We mapped RNA-seq reads using the RMAP software [29], with option '-Q' for position weight matrix matching, based on quality score. Alignment was stored in BED files. We used the default options: 2 mismatches allowed in the 32 first nucleotides, 10 mismatches allowed in the whole read. Ambiguous reads were discarded. Paired-End reads were mapped as Single-End reads.

We mapped those reads to 3'UTR ± 50bp: the reference sequence is all 3'UTR DNA sequences (from the human genome assembly HG18 [25]) from all coding genes (excluding Y chromosome because of overlap with X), including introns, extended of 50 nucleotides up- and downstream. Overlapping sequences were merged (19012 regions). We mapped reads to a second version of the reference sequence, where reference alleles of APA-SNPs were replaced by non-reference alleles.

## RNA-seq genotyping

We counted base calls based on base quality probability score and sequence alignment score: We discarded reads mapped with an alignment score $s > 4$, and reads that had a quality score $< 99\%$ accuracy at the SNP. Quality score probability of accuracy at a SNP was computed as follows: $p = 1 - 10^{-ord(Q-33)/10}$, where $Q$ is the ASCII character of one base call in a read in FASTQ file format [30]. We computed the mapping score as $m = 1 - (s/5)$, where $s$ is the alignment score given by RMAP. We counted alleles as $\sum p * m$ for each allele (for all the FASTQ files of each individual). We discarded alleles that do not fit Hapmap bi-allelic SNPs. If there was only one allele left, we classified the SNP as homozygous. If there were two alleles left, with both proportions greater than 0.15, we classified the SNP as heterozygous. If there were two alleles but one had its proportion lower than 0.15, we classified the SNP as homozygous with the allele having the biggest proportion.

## RNA-seq transcript end estimation

We mapped reads from the Burge dataset using the alignment software Bowtie [31] version 0.12.7 with default options. Bowtie generated alignments in the SAM format [32]. The transcript assembly software Cufflinks version 1.3.0 [33] was then used with the SAM files to generate a list of expressed exons for each run (default options). Those exons were then mapped back to annotated RefSeq genes. Exons that mapped to several different genes were discarded; the corresponding genes they overlapped were also discarded. For a given gene and a given run, the 3' end of the exon that mapped the most downstream on the gene was used as an estimate of the gene's 3' end. Finally, the distance between the estimate and the annotated transcript end was computed for each gene and each run. This distance $D$ is negative when the transcript is shorter than the annotation and had a logarithmic distribution for negative $D$s.

Few transcripts were longer than the annotated transcription end site, resulting in positive $D$ values. To handle these few positive $D$ values, we put a threshold at 30, so that all $D \geq 30$ were truncated to 29. We then converted the $D$s to the logarithm scale by using the following formula: $-\log(-D + 30)$.

## Allelic imbalance

Log Allelic Ratio for each heterozygous SNP is defined as $\log AR = \log \frac{|APAallele|}{|nonAPAallele|}$, where counts of alleles are computed in a similar way as in the genotyping section (by taking base quality and alignment score into account). $\log AR$ is positive when the transcripts with APA alleles are up-regulated compared to non-APA allele.

However, to avoid that a mapping bias towards reference alleles affects allelic ratios, we used a corrected allelic imbalance in our analyses, by combining allelic ratios computed from reads mapped to the reference genome with reference alleles, and allelic ratios computed from reads mapped to the same genome but with non-reference alleles at candidate SNPs. We defined it as the mean of the two log-ratios:

$$\log_2 AR = \log_2 \sqrt{\frac{A_R}{B_R} \frac{A_{NR}}{B_{NR}}}$$

where $AR$ is the allelic ratio, $A_R$ and $A_{NR}$ are the counts of APA alleles mapped to respectively the genome with reference alleles, and the one with non-reference alleles. Similarly $B_R$ and $B_{NR}$ are the counts of non-APA alleles.

## GU-rich regions

We took all the known coding genes from the UCSC RefSeq gene database (hg18) [26]. To define the precise region of GU-analysis, for each gene, we computed the GU proportion in a 5-nucleotide long window sliding from the polyA signal downstream in a 70-nucleotide long region. Those curves represent the variation of GU proportion in the region for each gene. We then took the mean of all the curves, which showed that the increased GU region was from the $25^{th}$ window to the $45^{th}$ window (Fig. S4). We therefore defined the GU level as the mean of the GU-proportions in the 5-nucleotide windows, from the $25^{th}$ to the $45^{th}$ downstream of the polyA signal.

## Scoring APA for miRNA regulation

### MicroRNA expression in Burge samples

Human miRNA profiles from Landgraf *et al.* [16] (their Table S5) were matched to Burge samples. We grouped and summed miRNA expression for mature miRNAs that have the same seed sequence and identified 117 seeds having a non-null expression.

### MicroRNA expression in Hapmap cell lines

We took human miRNA profile from Wang *et al.* [18] (Gene expression omnibus: GSE14794), consisting of miRNA expression for EBV-transformed lymphoblastoid cell lines for 90 samples. For each of the 735 miRNA probes, we took the mean expression value among the 90 samples, resulting in one expression value per probe. We then computed the mean expression value among miRNA probes, and discarded all probes being smaller than the mean: 275 probes remained. We mapped probe IDs to miRNA seeds using the Illumina annotation file HumanMI_V1_R2_XS0000122-MAP. A total of 215 miRNA seeds remained. For each seed, we summed the exponential of expression values of the corresponding probes, since they were at a logarithm scale. We used these scores to compute the proportion of expression for each seed. We discarded seeds that do not have reference mature miRNAs in the MirBase database release 16 [28]. 163 seeds corresponding to 285 mature miRNAs remained.

### MicroRNA scores

For each of the candidate SNPs and their corresponding RefSeq genes, we defined a short 3'UTR as the exonic region from the mRNA stop codon to the SNP, and a long 3'UTR, as the reference 3'UTR. We computed miRNA target predictions on those short and long sequences using the prediction tool from Saito *et al.* [34] for all mature miRNA sequence corresponding to the seed sequences identified in the considered cell line. The tool scores the mRNA/miRNA pairs, according to how a miRNA targets an mRNA: a high score means that the miRNA is more likely to down-regulate the mRNA. To compare scores for long and short UTRs, we normalised scores using the normalising method described in Thomas *et al.* [22]. Then for a given pair of miRNA seed and a UTR sequence, we took the score mean when one miRNA seed motif corresponded to several mature miRNAs, to have one score per seed. Then for a given UTR sequence, we computed a global score taking all expressed miRNAs into account: we summed scores for all the seeds, weighted by their proportion of expression in the considered cell line. When a gene corresponded to several RefSeq transcripts we took the score mean, resulting in having one long UTR score and one short UTR score for each candidate SNP. We could then compute the score difference for each SNP: this quantifies the potential effect of a cleavage site at the SNP locus on miRNA regulation.

## Messenger RNA expression and genotype

We downloaded human gene expression profiling of EBV-transformed lymphoblastoid cell lines from 270 unrelated Hapmap individuals [17] (Gene expression omnibus: GSE6536, data normalised across populations), and genotypes for the same individuals, from the Hapmap database release 27.

We mapped probe IDs to RefSeq genes using the BioConductor package for R [35, 36] (R version 2.10.1, AnnotationDbi package version 1.8.2 [37] and the annotation file illuminaHumanv1.db version 1.4.0). One candidate SNP could have one or several RefSeq gene IDs, which could be mapped to one or several probe IDs. Among those probe IDs, we selected the one with maximum variance across all the individuals in the dataset, and assigned it to the given SNP in the 3'UTR.

Genotype was encoded as 0, 1, and 2 for non-APA homozygotes, heterozygotes, and APA homozygotes, respectively.

## Bootstrapping median differences

We computed bootstraps of median differences: Given two groups with different sizes, we resampled with replacement in each group with their actual original size. We took the median in each resampling and computed the difference. We repeated this procedure 1000 times to create a median difference distribution, which was then used to compute the 95% confidence interval (95% CI).

## Mapping APA-SNPs to GWAS

We mapped APA-SNPs to GWAS SNPs, using the mapping method described in Thomas *et al.* [22]. The mapping was based on linkage disequilibrium (LD) data from the Hapmap database (CEU population release 27). The mapping parameter was the threshold $T = 0$ (see Thomas *et al.* [22]), to identify all neighbouring APA-SNP/GWAS-SNP pairs.

# Supporting information

**Text S1** Translation of the Abstract into French by LFT

**Text S2** RNA-seq data successfully genotype known SNPs.

**Figure S1** Diagram showing the workflow of our analyses and summarizing the number of SNPs investigated in each analysis.

**Figure S2** Distribution of Hapmap SNPs within 3'UTRs of all RefSeq genes. Panel (A) shows the SNP distribution as a function of relative position within the 3'UTR (coding end site at position 0 and transcript end site at position 1). The SNP distribution, which is based on a kernel density estimate, is relatively uniform across the 3'UTR. Panels (B) and (C) show the SNP distribution from, respectively, 500 bp and 200 bp upstream of the transcription end position to the first 50 bp outside the gene. The SNP density is uniform within the 3'UTR except at the polyA signal position around 30 bp upstream of the transcript end.

**Figure S3** Distribution of distance $D$ between estimated and annotated transcript ends within the Burge RNA-seq data, grouped into six sub-groups by the samples' cell proliferation state (non-proliferating vs. proliferating) and the APA SNPs' genotype (WT Hom.: homozygous wildtype; Het.: heterozygous; APA Hom.: homozygous APA). The distance $D$ is shown on a negative logarithmic scale to reflect that the estimated transcript ends are shorter than the annotated ends. As expected, transcripts in proliferating cells are shorter than in non-proliferating cells. Moreover, transcripts that have homozygous APA SNPs are shorter than other genotypes; particularly for non-proliferating cells.

**Figure S4** GU content around transcription end site, based on all RefSeq genes. Mean of curves defined as GU proportion in a 5-nucleotide window sliding from the polyA signal to 70 nucleotides downstream. The GU-rich region is located between the 25th window and the 45th window.

**Table S1** A portion of the EST-based polyA sites from PolyA_Db that do not have any signal in $N$ nucleotides upstream of the cleavage site when looking at the reference genome, can be explained by a SNP in the region creating a signal from the SNP's non-reference allele.

**Table S2** Checking genotyping of 755 mono-allelic SNPs in 2 datasets (Heap and Burge). Columns correctHOM, incorrectHOM, and incorrectHET show the number and proportion of correctly classified homozygotes and of incorrectly classified homozygotes and heterozygotes among the total number of genotypes, respectively; 'correct|classified' shows the proportion of correctly classified homozygotes among classified genotypes. Row Burge CEU corresponds to individuals in the Burge dataset that are Caucasian.

**Table S3** Genotyping results for the 412 candidate APA-SNPs in the Heap and Burge datasets.

**Table S4** PolyA signal frequencies. The first three columns show polyA signal ranks, signal hexamers, and their frequencies in human genes from Tian *et al.* [1]; columns four and five show the hexamers' absolute and relative frequencies in human RefSeq 3'UTRs; column six shows the signal frequencies divided by the signals' relative frequencies in human 3'UTRs; and columns seven and eight show the counts and frequencies of our 412 candidate APA-SNPs. PolyA signal frequency ("PAS frequency") corresponds well with how frequently the signal causes polyadenylation ("PAS frequency/Motif frequency").

**Table S5** Multiple regression on distance between the estimated and the annotated transcript end ($D$; see Methods) and APA SNP genotype, cell proliferation status, APA signal strength, and local and global GU level. We only considered SNPs that lie at least 1500 kb from the annotated 3' end. (A) All the dependent variables contribute significantly and negatively to the response variable ($D$), which means that homozygous APA SNPs, proliferating cells, strong signals, local and global GU levels all contribute to shortened 3'UTRs. (B) We get similar results when controlling for the global GU level. Specifically, the response variable in this analysis was the residuals from regressing global GU level on $D$.

**Table S6** Significant APA-SNPs from the microarray, EST and RNA-seq analyses.

## Acknowledgments

## References

1. Tian B, Hu J, Zhang H, Lutz C (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res 33: 201-212.

2. Colgan D, Manley J (1997) Mechanism and regulation of mRNA polyadenylation. Genes Dev 11: 2755-2766.

3. Legendre M, Ritchie W, Lopez F, Gautheret D (2006) Differential repression of alternative transcripts: A screen for miRNA targets. PLoS Comput Biol 2: 333-342.

4. Mayr C, Bartel DP (2009) Widespread Shortening of 3 ' UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. Cell 138: 673-684.

5. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB (2008) Proliferating cells express mRNAs with shortened 3 ' untranslated regions and fewer microRNA target sites. Science 320: 1643-1647.

6. Singh P, Alley TL, Wright SM, Kamdar S, Schott W, et al. (2009) Global Changes in Processing of mRNA 3 ' Untranslated Regions Characterize Clinically Distinct Cancer Subtypes. Cancer Res 69: 9422-9430.

7. Danckwardt S, Hentze MW, Kulozik AE (2008) 3 ' end mRNA processing: molecular mechanisms and implications for health and disease. Embo J 27: 482-498.

8. Uitte De Willige S, Rietveld IM, De Visser MCH, Vos HL, Bertina RM (2007) Polymorphism 10034c>t is located in a region regulating polyadenylation of fgg transcripts and influences the fibrinogen $\gamma'/\gamma a$ mrna ratio. J Thromb Haemost 5: 1243–1249.

9. Wiestner A, Tehrani M, Chiorazzi M, Wright G, Gibellini F, et al. (2007) Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. Blood 109: 4599-4606.

10. Mu XJ, Lu ZJ, Kong Y, Lam HYK, Gerstein MB (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res 39: 7058-7076.

11. Int HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-U3.

12. Zhang H, Hu J, Reccel M, Tian B (2005) PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res 33: D116-D120.

13. Lee JY, Yeh I, Park JY, Tian B (2007) PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. Nucleic Acids Res 35: D165-D168.

14. Sherry S, Ward M, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.

15. Heap GA, Yang JHM, Downes K, Healy BC, Hunt KA, et al. (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Hum Mol Genet 19: 122-134.

16. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, et al. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 129: 1401-1414.

17. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848-853.

18. Wang L, Oberg AL, Asmann YW, Sicotte H, McDonnell SK, et al. (2009) Genome-Wide Transcriptional Profiling Reveals MicroRNA-Correlated Genes and Biological Processes in Human Lymphoblastoid Cell Lines. PLoS One 4.

19. Nunes NM, Li W, Tian B, Furger A (2010) A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. Embo J 29: 1523-1536.

20. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.

21. Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed Apr. 18, 2011.

22. Thomas LF, Saito T, Sætrom P (2011) Inferring causative variants in microRNA target sites. Nucleic Acids Res 39.

23. Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38: 226-231.

24. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

25. Int Human Genome Sequencing Conso, Lander E, Linton L, Birren B, Nusbaum C, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

26. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2010) The ucsc genome browser database: update 2011. Nucl Acids Res .

27. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470-476.

28. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36: D154-D158.

29. Smith AD, Xuan Z, Zhang MQ (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9.

30. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38: 1767-1771.

31. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10.

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

33. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-U174.

34. Saito T, Sætrom P (2010) A two-step site and mRNA-level model for predicting microRNA targets. BMC Bioinformatics 11: 612.

35. R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

36. Gentleman RC, Carey VJ, Bates DM, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 5: R80.

37. Pages H, Carlson M, Falcon S, Li N AnnotationDbi: Annotation Database Interface. R package version 1.8.2.

## Figures

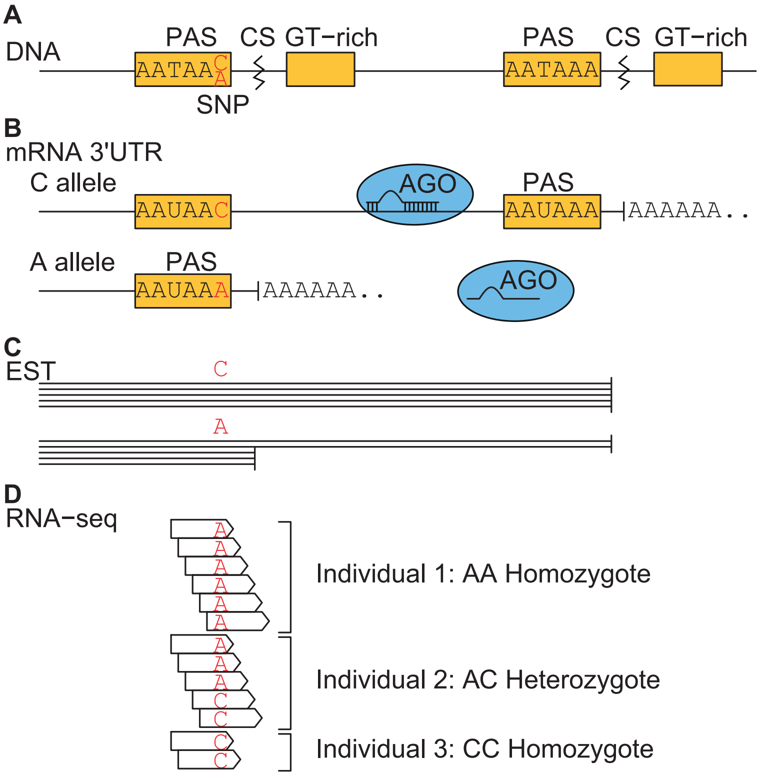


**Figure 1. A model of the effect of APA-SNPs in the 3'UTR of a gene.** Top panel: For the C allele, the second cleavage site (CS) is used, because the first polyA signal is not functional. For the A allele, the first polyA signal is functional, therefore the pre-mRNA can be cleaved at the first CS, resulting in a loss of functional miRNA target sites downstream, and increased gene expression. Middle panel: EST sequences enable identifying APA-SNP alleles and 3'UTR length. Bottom panel: RNA-seq reads enable genotyping APA-SNPs and quantifying expression patterns.

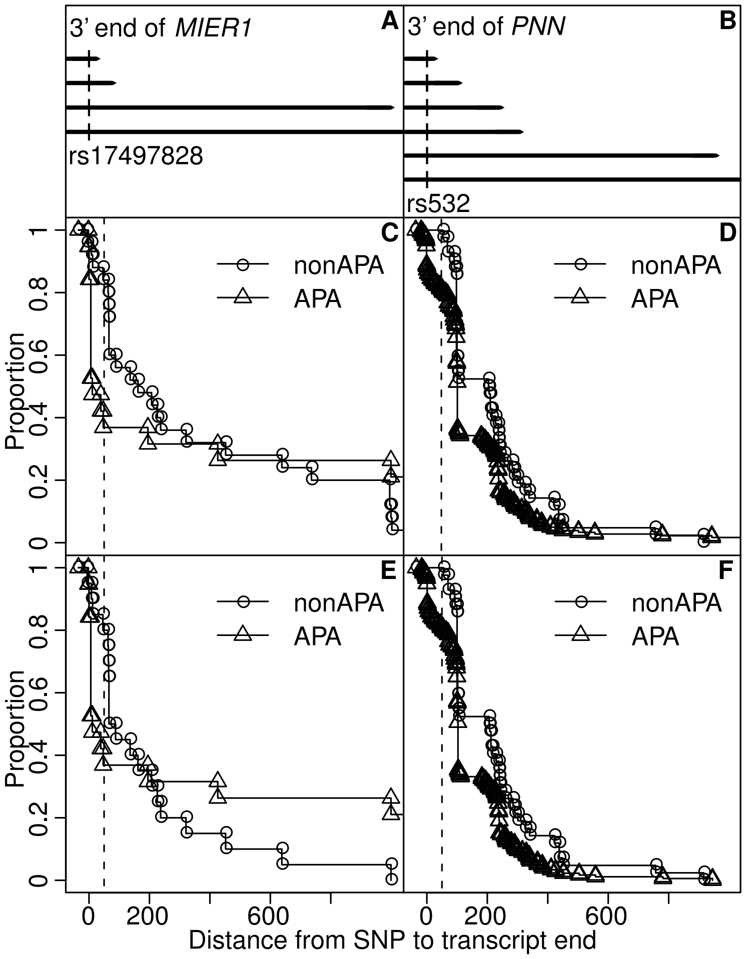


**Figure 2. SNPs can affect 3'UTR length.** Panels (A) and (B) show 3' ends of the *MIER1* and *PNN* genes as annotated in PolyA_Db (3' ends of the horizontal lines), and their candidate APA SNP. The four other graphs show the inverse cumulative distribution of EST sequence ending position for APA alleles (triangles) and non-APA alleles (circles). The dashed vertical line shows the threshold separating short and long transcripts. The transcript proportion is decreasing before the threshold for APA alleles, compared to non-APA alleles. This decrease indicates that APA alleles are more likely to produce shorter transcripts. Panels (A), (C) and (E) show the *MIER1* gene. Panels (B), (D) and (F) show the *PNN* gene. Several unknown alleles could be imputed through haplotypes (included in Panels (C) and (D)).

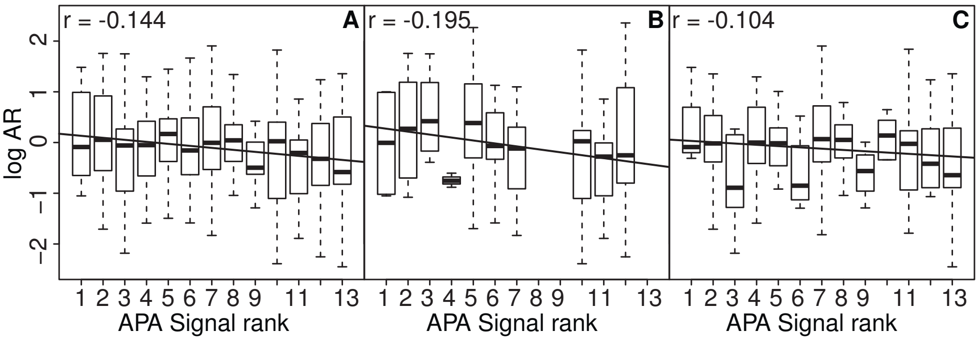


**Figure 3. Increased allelic imbalance correlates with signal strength and depends on downstream GU-content.** Log allelic ratio distribution of APA allele over non-APA allele for each polyA signal ordered by strength. Panel (A): log allelic ratio is negatively correlated with signal rank for all APA-SNPs. Compared with all APA-SNPs, APA-SNPs with a GU-rich region (Panel (B)) have a stronger negative correlation between log allelic ratio and signal rank. For APA-SNPs without a GU-rich region (Panel (C)), there is no significant correlation between signal rank and log allelic ratio. The graphs include data from the 19 non-mixed cell lines and tissues. The line in each panel shows the linear regression line; the corresponding Pearson correlation coefficient $r$ is in the panel's upper left corner.

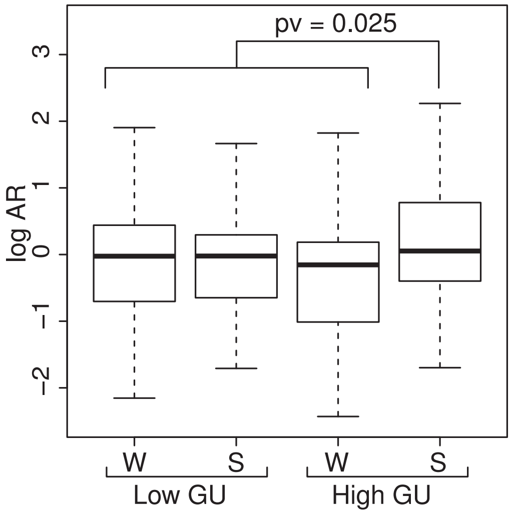


**Figure 4. Allelic imbalance distributions according to signal strength and downstream GU levels.** Allelic imbalance is increased towards APA alleles for APA-SNPs in strong (S) signals with high downstream GU levels. The graph shows a box-plot of the log AR distribution of APA-SNPs grouped by signal strength (weak (W) and strong (S)) and downstream GU levels.

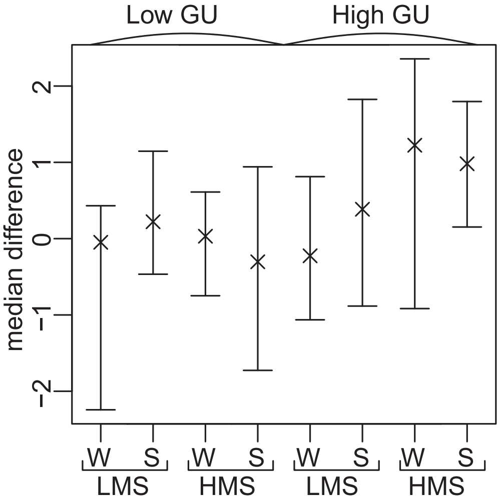


**Figure 5. SNP expression difference between SNPs with positive and negative log allelic ratios.** Logarithm of SNP expression median difference between SNPs with positive log allelic ratios and those with negative log allelic ratios, in several groups (low and high GU level, low (LMS) and high (HMS) miRNA score, and weak (W) and strong (S) signal). Crosses show median differences. Bootstrapping median differences gives 95% CI. Only one CI does not contain zero: the one with high GU, HMS and S, indicating that positive allelic imbalance for SNPs in strong polyA sites and affecting miRNA target sites, is associated with increased SNP expression, and therefore increased gene expression.

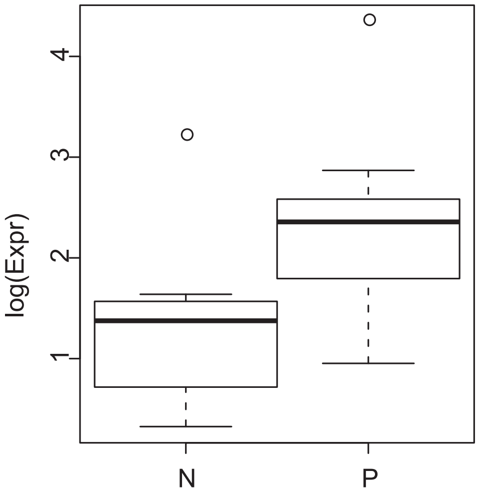


**Figure 6. SNP expression distributions according to allelic imbalance direction.** SNPs in strong APA signal, with high GU level and high miRNA score, have a significantly higher logarithm of SNP expression for SNPs with imbalance towards APA allele (positive (P) log allelic ratio), compared to SNPs with imbalance towards non-APA allele (negative (N) log allelic ratio)

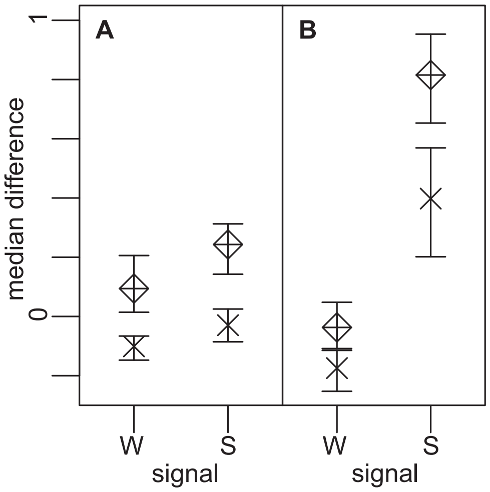


**Figure 7. APA homozygotes have an increased gene expression for strong polyA signals and high miRNA score.** Gene expression medians in several groups are shown: Median differences between the APA homozygotes and non-APA homozygotes (Rhombus), and between heterozygotes and non-APA homozygotes (Cross). 95% CI for median differences are shown. Expression of APA homozygotes is generally higher, followed by heterozygotes, and then finally non-APA homozygotes. (A): genes where alternative polyadenylation does not affect miRNA targeting (low miRNA score). Strong signals (S) have a slightly higher median difference compared to weak signals (W). (B): genes where alternative polyadenylation affects miRNA targeting (high miRNA score). Strong signals have a significantly higher median difference

# Tables

**Table 1. Significant genes in the EST analysis**

| | no correction | | Benjamini&Hochberg correction | |
|---|---|---|---|---|
| Gene | imputation | no imputation | imputation | no imputation |
| *MIER1* | 0.004* | 0.016* | 0.032* | 0.103 |
| *PNN* | 0.005* | 0.004* | 0.032* | 0.058 |

* shows significant p-values.

P-values for 2x2 $\chi^2-$test comparing the proportion of alleles with APA signal for short versus long EST sequences. The *MIER1* and *PNN* genes were significant (including and not including imputed alleles). After correcting for multiple testing, the proportions including imputed alleles remained significantly different between short and long ESTs.

**Table 2. Potential functional APA alleles are positively correlated with risk alleles from GWAS SNPs.**

| Predicate | Success count | Trial count | Success probability under $H_0$ | p-value (>) |
|---|---|---|---|---|
| $[\hat{r} > 0]$ | 8 | 9 | $32/60 = 0.533$ | 0.03 |
| $[\hat{r}^2 > 0.2]$ | 5 | 9 | $15/60 = 0.25$ | 0.049 |

Two predicates were tested in a binomial setting: $[\hat{r} > 0]$ for positive trend correlation between APA and risk alleles, and $[\hat{r}^2 > 0.2]$ for the strength of the correlation. For the 60 APA-SNPs paired to GWAS-SNPs, the proportions of $\hat{r} > 0$ and $\hat{r}^2 > 0.2$ were respectively 0.53 and 0.25. Among the 9 SNPs identified in the previous sections as functional candidate, respectively 8 and 5 succeeded the Bernoulli trial. Both null hypotheses were rejected.

Supplementary Text S1

# Supporting Abstract:

## Single Nucleotide Polymorphisms Can Create Alternative Polyadenylation Signals and Affect Gene Expression through Loss of MicroRNA-Regulation

Laurent F. Thomas and Pål Sætrom

### Translation of the Abstract into French by LFT

La polyadénylation alternative (APA) est un mécanisme qui peut se produire par exemple lorsqu'un gène codant pour une protéine présente plusieurs signaux de polyadénylation (polyA) dans son dernier exon, résultant ainsi en ARN messagers (ARNm) de différentes longueurs au niveau de leur région 3 ' non traduite (UTR). Différentes longueurs de 3 ' UTR peuvent perturber la régulation des gènes par microARNs (miARNs) de telle sorte que l'expression des transcrits écourtés augmente. L'APA fait partie des mécanismes naturels de régulation des cellules humaines, mais semble également jouer un rôle important dans de nombreuses maladies humaines. Bien qu'une polyadenylation altérée dans le cadre de pathologies puisse avoir plusieurs causes, nous avons présupposé que des mutations d'ADN au niveau d'éléments particulièrement importants dans le processus de polyA, tels que le signal de polyA ainsi que la région en aval riche en GU, pouvaient être un important mécanisme d'altération. Pour tester cette hypothèse, nous avons identifié des polymorphismes nucléotidiques simples (SNP) qui peuvent créer ou perturber des signaux de polyA alternative (APA-SNP). En utilisant une approche d'intégration de données, nous montrons que les APA-SNPs peuvent affecter la longueur du 3 ' UTR, la régulation par miARN et l'expression d'ARNm — et ce, en comparant aussi bien l'expression des gènes d'individus homozygotes que l'expression allélique d'individus hétérozygotes. Par ailleurs, nous montrons qu'une proportion significative d'allèles causant l'APA est fortement et positivement liée aux allèles identifiées comme étant à risque par des études pangénomiques d'association à diverses maladies. Nos résultats confirment que l'APA-SNP peut modifier la régulation des gènes et que les allèles d'APA donnant des transcrits raccourcis ainsi qu'une augmentation de l'expression des gènes peuvent être une importante cause de maladies héréditaires.

# Supporting Result:
## Single Nucleotide Polymorphisms Can Create Alternative Polyadenylation Signals and Affect Gene Expression through Loss of MicroRNA-Regulation

Laurent F. Thomas and Pål Sætrom

### RNA-seq data successfully genotype known SNPs

We used our genotyping approach (see Methods) to analyse Heap and colleagues' RNA-seq data [1], which are based on human primary *CD4*$^+$ *T* cells from 4 individuals. After mapping the reads to the reference genome, we could genotype our 755 candidate SNPs that are mono-allelic in the Hapmap CEU population, since the 4 individuals are known to be Caucasian. Of the $755 * 4 = 3020$ possible genotypes, 1650 were correctly classified as homozygous with the expected Hapmap allele, 1360 could not be classified because of the lack of reads (unexpressed genes), only 3 were misclassified as heterozygous, and 7 were misclassified as homozygous with the unexpected allele (minor allele frequency (MAF) allele) (Table S2). We also took the intersection between the known heterozygous SNPs reported in Heap *et al.* [1], and our candidate SNPs (26 genotypes, 19 SNPs), and could classify all of them as heterozygous (Table S2).

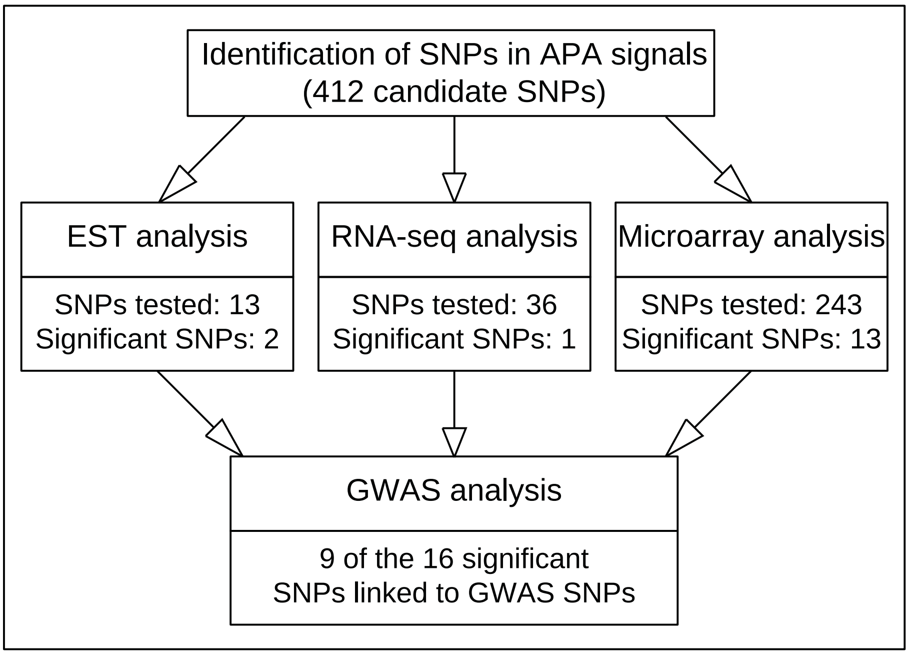
We also analysed the Burge Lab's RNA-seq data [2], which are based on 22 unrelated individuals; specifically, 7 cancer cell lines and 15 tissue samples. Again we genotyped SNPs that are mono-allelic in the CEU population and got similar results as for the Heap data (Table S2). Discarding samples that are not Caucasian increased the fraction of correctly classified genotypes (Table S2), which is consistent with us using the CEU Hapmap population to assess correctness. Specifically, by using the Hapmap CEU population to evaluate our genotyping approach, we got an upper-bound estimate of our method's accuracy, as the CEU population only approximates our samples' true genetic variations. Table S3 shows the number of classified genotypes in the 2 datasets for our candidate SNPs, which exclude mono-allelic SNPs. Based on the CEU-based validations, we expected most of these genotypes to be correct.

## References

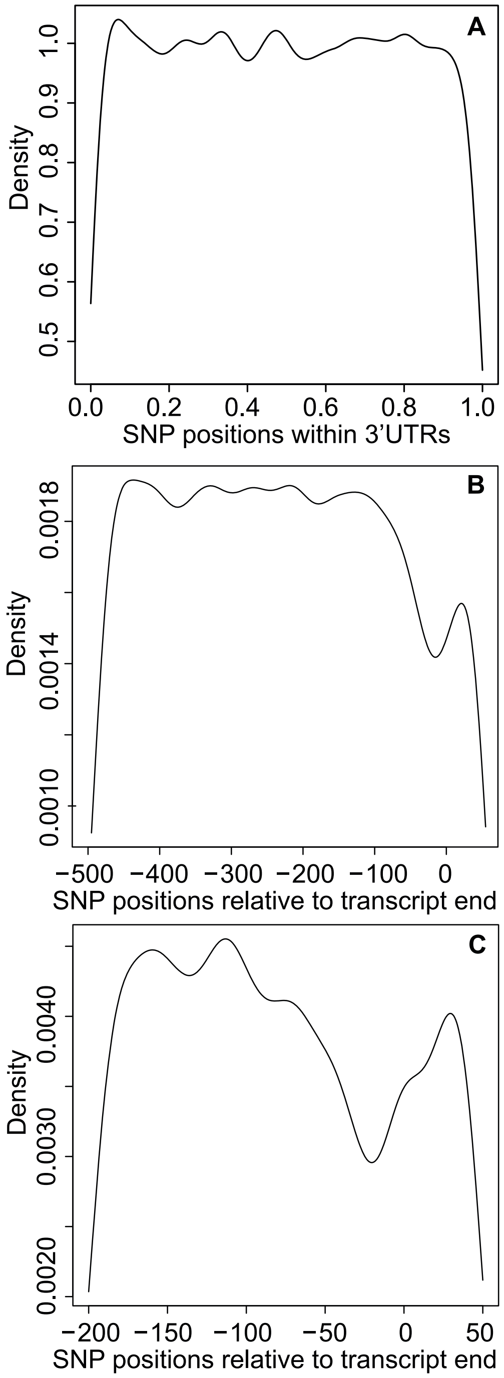
[1] Heap GA, Yang JHM, Downes K, Healy BC, Hunt KA, et al. (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. HUMAN MOLECULAR GENETICS 19: 122-134.

[2] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. NATURE 456: 470-476.

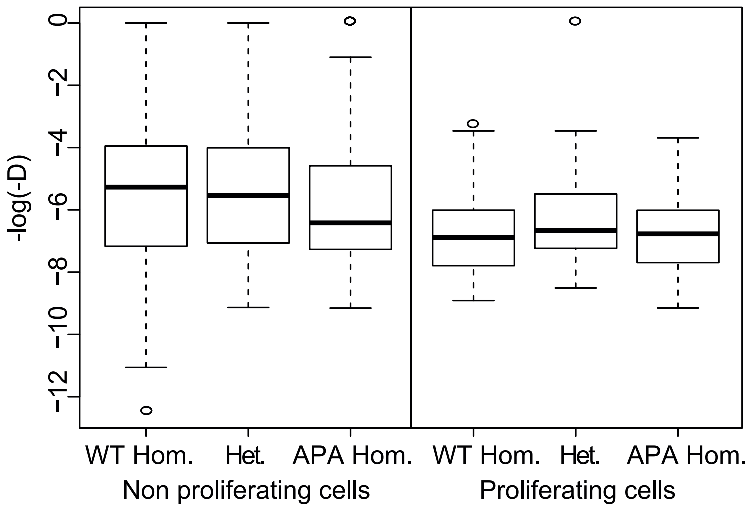
Supplementary Figure S1



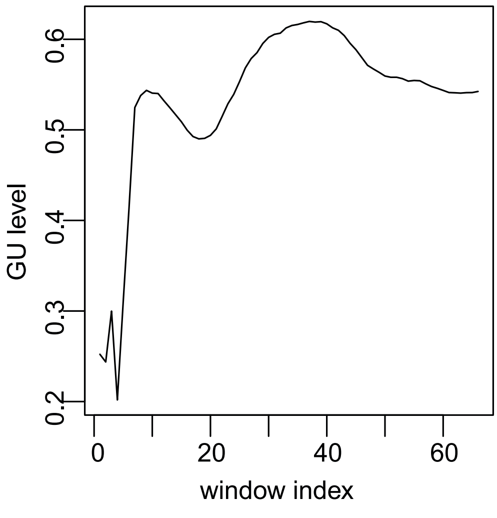
Supplementary Figure S2



Supplementary Figure S3



Supplementary Figure S4



Supplementary Table S1

| Region Size | # PolyA sites | # PolyA sites with SNP-created Signal | | |
|---|---|---|---|---|
| $N$ | without Signal | CEU Hapmap | dbSNP126 | dbSNP130 |
| 40 | 1728 | 6/1728 | 21/1728 | 24/1728 |
| 80 | 1343 | 9/1343 | 20/1343 | 26/1343 |
| 100 | 1210 | 10/1210 | 22/1210 | 26/1210 |

Supplementary Table S2

| Dataset | n | total genotypes | correctHOM | incorrectHOM | incorrectHET | correct\|classified |
|---|---|---|---|---|---|---|
| Heap | 4 | $4 * 755 = 3020$ | 1650(54.6%) | 7(0.23%) | 3(0.1%) | 99.4% |
| Burge | 22 | $22 * 755 = 16610$ | 5748(34.6%) | 42(0.25%) | 51(0.31%) | 98.41% |
| Burge CEU | 18 | $18 * 755 = 13590$ | 4753(35%) | 20(0.15%) | 33(0.24%) | 98.9% |

Supplementary Table S3

| Dataset | n | total genotypes | classified |
|---|---|---|---|
| Heap | 4 | $4 * 412 = 1648$ | 865(52.5%) |
| Burge | 22 | $22 * 412 = 9064$ | 3156(34.8%) |

Supplementary Table S4

| | | PAS | Motifs in 3'UTRs | | $\frac{\text{PAS frequency}}{\text{Motif frequency}}$ | APA-SNPs | |
|---|---|---|---|---|---|---|---|
| rank | signal | frequency | count | frequency | | count | frequency |
| 1 | AAUAAA | 53.18% | 24436 | 15.90% | 3.35 | 10 | 2.43% |
| 2 | AUUAAA | 16.78% | 13614 | 8.86% | 1.89 | 27 | 6.55% |
| 3 | UAUAAA | 4.37% | 11434 | 7.44% | 0.59 | 33 | 8.01% |
| 4 | AGUAAA | 3.72% | 7459 | 4.85% | 0.77 | 23 | 5.58% |
| 5 | AAGAAA | 2.99% | 17767 | 11.56% | 0.26 | 55 | 13.35% |
| 6 | AAUAUA | 2.13% | 9818 | 6.39% | 0.33 | 23 | 5.58% |
| 7 | AAUACA | 2.03% | 7667 | 4.99% | 0.41 | 42 | 10.19% |
| 8 | CAUAAA | 1.92% | 6507 | 4.23% | 0.45 | 27 | 6.55% |
| 9 | GAUAAA | 1.75% | 5914 | 3.85% | 0.45 | 23 | 5.58% |
| 10 | AAUGAA | 1.56% | 11005 | 7.16% | 0.22 | 44 | 10.68% |
| 11 | UUUAAA | 1.20% | 25949 | 16.88% | 0.07 | 55 | 13.35% |
| 12 | ACUAAA | 0.93% | 6570 | 4.27% | 0.22 | 24 | 5.83% |
| 13 | AAUAGA | 0.60% | 5565 | 3.62% | 0.17 | 26 | 6.31% |
| | total | 93.16% | 153705 | 100% | | 412 | 100% |

Supplementary Table S5

**A**

| Variables | $\beta_i$ estimates | p-values |
|---|---|---|
| Genotype (WT:0, HET:1, APA:2) | -0.30010 | $5.9 * 10^{-12}$ |
| Proliferating (True: 1, False: 0) | -0.89453 | $< 2 * 10^{-16}$ |
| Signal (Strong: 1, Weak: 0) | -0.18289 | $1.7 * 10^{-2}$ |
| Local GU level | -1.06154 | $9.3 * 10^{-4}$ |
| Global GU level | -10.30803 | $1.5 * 10^{-4}$ |

Multiple $R^2$: 0.0726

**B**

| Variables | $\beta_i$ estimates | p-values |
|---|---|---|
| Genotype (WT:0, HET:1, APA:2) | -0.29752 | $8.4 * 10^{-12}$ |
| Proliferating (True: 1, False: 0) | -0.89865 | $< 2 * 10^{-16}$ |
| Signal (Strong: 1, Weak: 0) | -0.18015 | $1.8 * 10^{-2}$ |
| Local GU level | -0.93212 | $1.9 * 10^{-3}$ |

Multiple $R^2$: 0.0616

Supplementary Table S6. Excel sheet for microarray results

| SNP | Gene | 3'UTRlengt | SNPprop | Signal | GU | DS Lympho | pearson r | pvalue | BH pv | bonf pv |
|---|---|---|---|---|---|---|---|---|---|---|
| rs3763406 | FAM62B | 3209 | 66,50 % | 2 | 0,5 | 0,102 | 0,418 | 3,86E-12 | 9,37E-10 | 9,37E-10 |
| rs986475 | NCR3 | 172 | 77,91 % | 1 | 0,7 | 0,000 | 0,410 | 1,05E-11 | 1,27E-09 | 2,55E-09 |
| rs3743955 | ITPRIPL2 | 5561 | 90,92 % | 9 | 0,59 | 0,041 | 0,378 | 3,94E-10 | 3,20E-08 | 9,59E-08 |
| rs10793442 | ZNF239 | 358 | 92,46 % | 9 | 0,76 | 0,000 | 0,342 | 1,78E-08 | 1,08E-06 | 4,33E-06 |
| rs1060379 | ZNF117 | 3667 | 82,00 % | 2 | 0,52 | 0,053 | 0,337 | 2,86E-08 | 1,39E-06 | 6,94E-06 |
| rs15062 | BCKDHB | 2466 | 92,09 % | 2 | 0,48 | 0,033 | 0,283 | 3,23E-06 | 1,31E-04 | 7,86E-04 |
| rs6972005 | CALU | 2310 | 80,48 % | 6 | 0,59 | 0,017 | 0,267 | 8,52E-06 | 2,96E-04 | 2,07E-03 |
| rs9162 | CCDC74A | 268 | 74,25 % | 8 | 0,48 | 0,000 | 0,241 | 3,76E-05 | 1,07E-03 | 9,14E-03 |
| rs6777019 | CGGBP1 | 3523 | 51,49 % | 7 | 0,67 | 0,132 | 0,247 | 4,23E-05 | 1,07E-03 | 1,03E-02 |
| rs4612984 | EXOC5 | 6133 | 71,20 % | 2 | 0,47 | 0,175 | 0,249 | 4,39E-05 | 1,07E-03 | 1,07E-02 |
| rs1052873 | PBK | 684 | 4,09 % | 8 | 0,51 | 0,075 | 0,238 | 6,11E-05 | 1,35E-03 | 1,48E-02 |
| rs3209335 | PPM1A | 6606 | 90,37 % | 6 | 0,69 | 0,010 | 0,222 | 1,31E-04 | 2,64E-03 | 3,17E-02 |
| rs1942 | RTF1 | 2875 | 53,50 % | 9 | 0,49 | 0,104 | 0,228 | 0,0002 | 2,82E-03 | 3,67E-02 |
| rs1043881 | BCAT1 | 6663 | 98,42 % | 13 | 0,44 | -0,007 | 0,222 | 0,0002 | 0,0039 | |
| rs29069 | VAPA | 5810 | 30,15 % | 8 | 0,49 | 0,097 | 0,210 | 0,0005 | 0,0073 | |
| rs11920 | C10ORF18 | 1948 | 44,30 % | 4 | 0,51 | 0,176 | 0,204 | 0,0007 | 0,0107 | |
| rs9242 | SRGAP2 | 3018 | 87,14 % | 13 | 0,45 | 0,026 | 0,186 | 0,0011 | 0,0155 | |
| rs1188401 | MATN1 | 2303 | 68,22 % | 7 | 0,21 | 0,025 | 0,178 | 0,0027 | 0,0362 | |
| rs7305647 | SUDS3 | 3602 | 92,39 % | 9 | 0,7 | 0,011 | 0,174 | 0,0030 | 0,0390 | |
| rs1156 | CHD6 | 2063 | 58,31 % | 5 | 0,53 | 0,156 | 0,157 | 0,0051 | | |
| rs702530 | PDE4D | 5625 | 52,14 % | 3 | 0,37 | 0,024 | 0,163 | 0,0052 | | |
| rs3745008 | SLC14A2 | 576 | 2,95 % | 7 | 0,31 | 0,079 | 0,161 | 0,0056 | | |
| rs1053489 | WDR48 | 1647 | 63,39 % | 13 | 0,63 | 0,077 | 0,157 | 0,0067 | | |
| rs1653589 | CAMKK2 | 3007 | 32,03 % | 10 | 0,56 | 0,069 | 0,154 | 0,0077 | | |
| rs12608564 | ZNF551 | 1513 | 55,58 % | 8 | 0,55 | 0,076 | 0,154 | 0,0077 | | |
| rs1057403 | BTK | 431 | 44,32 % | 5 | 0,4 | 0,003 | 0,146 | 0,0111 | | |
| rs10921309 | TROVE2 | 1582 | 18,52 % | 7 | 0,37 | 0,107 | 0,142 | 0,0119 | | |
| rs703258 | VCL | 1987 | 79,72 % | 11 | 0,52 | 0,084 | 0,136 | 0,0127 | | |
| rs11948089 | WDR36 | 3619 | 50,07 % | 7 | 0,54 | 0,138 | 0,138 | 0,0127 | | |
| rs1061686 | NUDT19 | 1839 | 91,41 % | 2 | 0,35 | 0,000 | 0,140 | 0,0132 | | |
| rs10686 | SEC23IP | 7966 | 98,93 % | 11 | 0,72 | 0,004 | 0,130 | 0,0162 | | |
| rs2833955 | C21ORF62 | 3004 | 85,09 % | 9 | 0,64 | 0,061 | 0,129 | 0,0171 | | |
| rs1061646 | ZNF276 | 2678 | 49,40 % | 5 | 0,55 | 0,095 | 0,131 | 0,0198 | | |
| rs506619 | DTNA | 2651 | 11,69 % | 13 | 0,53 | 0,154 | 0,128 | 0,0220 | | |
| rs4145905 | SORBS1 | 3286 | 40,66 % | 10 | 0,67 | 0,200 | 0,127 | 0,0233 | | |
| rs10143429 | C14ORF129 | 1542 | 92,93 % | 10 | 0,57 | 0,001 | 0,121 | 0,0270 | | |
| rs4558 | TJP2 | 830 | 74,70 % | 8 | 0,63 | 0,052 | 0,118 | 0,0281 | | |
| rs27194 | NLRC5 | 996 | 7,43 % | 9 | 0,5 | 0,152 | 0,117 | 0,0295 | | |
| rs3731661 | WDR35 | 3290 | 74,83 % | 10 | 0,7 | 0,026 | 0,116 | 0,0324 | | |
| rs12479 | HSPA13 | 2501 | 26,71 % | 13 | 0,56 | 0,117 | 0,117 | 0,0327 | | |
| rs3750992 | TRIM68 | 1604 | 11,35 % | 11 | 0,49 | 0,155 | 0,116 | 0,0343 | | |
| rs158688 | SYK | 2950 | 32,95 % | 5 | 0,25 | 0,111 | 0,116 | 0,0349 | | |
| rs10476052 | ICHTHYIN | 1762 | 32,12 % | 5 | 0,53 | 0,161 | 0,115 | 0,0352 | | |
| rs8970 | LTBP1 | 963 | 11,21 % | 7 | 0,45 | 0,119 | 0,115 | 0,0355 | | |
| rs3748983 | FLJ11151 | 5087 | 9,49 % | 10 | 0,53 | 0,275 | 0,109 | 0,0367 | | |
| rs15563 | UBE2Z | 1926 | 36,19 % | 4 | 0,58 | 0,075 | 0,112 | 0,0397 | | |
| rs11708200 | NPHP3 | 1300 | 20,38 % | 8 | 0,49 | 0,219 | 0,107 | 0,0410 | | |

SNP: SNP rsid
Gene: gene name
3'UTRlength: 3'UTR length
SNPprop: proportion of the 3'UTR upstream of the SNP
Signal: APA signal rank
GU: GU level downstream of the SNP
DS Lymphobla miRNA score of the SNP based on lymphoblastoid miRNA expression and target prediction
pearson r: pearson correlation coefficient between the SNP and gene expression
pvalue: p-value without correction for multiple testing
BH pv: p-value after benjamini correction
bonf pv: p-value after bonferonni correction

Supplementary Table S6. Excel sheet for EST results

| | | | | | | p-value without Correction | | p-value after BH correction | |
|---|---|---|---|---|---|---|---|---|---|
| SNP | Gene | 3'UTRlen | SNPprop | Signal | GU | no hapl imputation | imputation | no hapl imputation | imputation |
| rs17497828 | MIER1 | 2146 | 7,74 % | 2 | 0,48 | 0,016 | 0,004 | 0,103 | 0,032 |
| rs532 | PNN | 1355 | 9,30 % | 7 | 0,66 | 0,004 | 0,005 | 0,058 | 0,032 |

| | |
|---|---|
| SNP: | SNP rsid |
| Gene: | gene name |
| 3'UTRlength: | 3'UTR length |
| SNPprop: | proportion of the 3'UTR upstream of the SNP |
| Signal: | APA signal rank |
| GU: | GU level downstream of the SNP |
| p-value without Correction: | pvalue of 2x2 chi^2 test, without correction |
| p-value after BH correction | pvalue after benjamini correction |
| no hapl imputation | pvalue without including alleles imputated through haplotypes |
| imputation | pvalue including alleles imputated through haplotypes |

Supplementary Table S6. Excel sheet for RNA-seq results

| snp | gene | cell line | data | 3'UTRle | SNPprop | Sig | GU | Ds | APAal | APAcount | al2 | count2 | N | logAR | APAal prop | chisq pv | BH pv | Bonf pv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs2269123 | MRPS34 | BT474 | burge | 326 | 19,02 % | 5 | 0,56 | 0,004 | T | 53,57 | C | 19,65 | 73,23 | 1,45 | 73,16 % | 7,38E-05 | 2,77E-03 | 5,53E-03 |
| rs2269123 | MRPS34 | MCF-7 | burge | 326 | 19,02 % | 5 | 0,56 | 0,004 | T | 113,89 | C | 26,36 | 140,25 | 2,11 | 81,21 % | 1,45E-13 | 1,09E-11 | 1,09E-11 |

| | |
|---|---|
| SNP: | SNP rsid |
| Gene: | gene name |
| cell line | cell line name |
| data | dataset name |
| 3'UTRlengt | 3'UTR length |
| SNPprop: | proportion of the 3'UTR upstream of the SNP |
| Signal: | APA signal rank |
| GU: | GU level downstream of the SNP |
| DS | miRNA score of the SNP based on matched cell line miRNA expression and target prediction |
| APAal | APA allele |
| APAcount | APA allele counts (read quality based) |
| al2 | nonAPA allele |
| count2 | nonAPA allele counts (read quality based) |
| N | total count |
| logAR | allelic log ratio |
| APAal prop | propotion of APA allele |
| chisq pv | 1df chi^2 test pvalue |
| BH pv: | p-value after benjamini correction |
| bonf pv: | p-value after bonferonni correction |

Supplementary Table S6. Excel sheet for GWAS results

| | APA SNP | | | alleles | | | PolyA motif | GWAS SNP | | alleles | | | | | | | LD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| chr | rsid | MAF | gene | APA | WT | Analysis | motif | rsid | MAF | risk a | other | PUBMED | Trait | Sampl | Replic | p-value | r | r^2 | mean(r ) | mean(r^2) |
| 10 | rs11920 | 0,233 | C10ORF18 | A | C | microarra | AGU[A]AA | rs2380205 | 0,492 | C | T | 20453838 | Breast can | British | Europ | 5,00E-07 | 0,167 | 0,028 | 0,167 | 0,028 |
| 16 | rs1061646 | 0,4 | ZNF276 | A | G | microarra | AAGAA[A] | rs258322 | 0,175 | A | G | 19578364 | Melanoma | Europ | Europ | 3,00E-27 | 0,564 | 0,318 | 0,564 | 0,318 |
| 16 | rs1061646 | 0,4 | ZNF276 | A | G | microarra | AAGAA[A] | rs258322 | 0,175 | A | G | 18483556 | hair color | Europ | Indivi | 2,00E-23 | 0,564 | 0,318 | | |
| 16 | rs2269123 | 0,103 | MRPS34 | T | C | RNA-seq | AAG[A]AA | rs1065656 | 0,292 | G | C | 21216879 | Insulin-like | Europ | NR | 1,00E-11 | 0,543 | 0,295 | 0,543 | 0,295 |
| 17 | rs15563 | 0,483 | UBE2Z | A | G | microarra | [A]GUAAA | rs46522 | 0,483 | T | C | 21378990 | Coronary h | Europ | Indivi | 2,00E-08 | -1 | 1 | -0,449 | 0,202 |
| 17 | rs15563 | 0,483 | UBE2Z | A | G | microarra | [A]GUAAA | rs9674544 | 0,442 | G | A | 20195514 | Primary toc | Europ | NR | 2,00E-08 | -0,449 | 0,202 | | |
| 17 | rs15563 | 0,483 | UBE2Z | A | G | microarra | [A]GUAAA | rs9674544 | 0,442 | G | A | 20195514 | Primary toc | Europ | NR | 8,00E-07 | -0,449 | 0,202 | | |
| 18 | rs3745008 | 0,458 | SLC14A2 | T | C | microarra | AA[U]ACA | rs1050286 | 0,092 | C | T | 19260141 | Biochemica | Croati | NR | 7,00E-06 | 0,114 | 0,013 | 0,114 | 0,013 |
| 21 | rs12479 | 0,292 | HSPA13 | T | C | microarra | A[A]UAGA | rs1006899 | 0,158 | A | G | 19079262 | Bone mine | Individ | Indivi | 6,00E-06 | 0,077 | 0,006 | 0,077 | 0,006 |
| 5 | rs1047605 | 0,067 | ICHTHYIN | A | C | microarra | AAG[A]AA | rs1174056 | 0,043 | G | A | 20889312 | Bipolar dis | Europ | NR | 1,00E-06 | 0,931 | 0,867 | 0,3 | 0,58 |
| 5 | rs1047605 | 0,067 | ICHTHYIN | A | C | microarra | AAG[A]AA | rs2277027 | 0,308 | A | C | 20010835 | Pulmonary | Europ | Europ | 1,00E-10 | -0,4 | 0,16 | | |
| 5 | rs1194808 | 0,133 | WDR36 | A | T | microarra | [A]AUACA | rs2416257 | 0,133 | C | T | 19198610 | Asthma an | Europ | Europ | 1,00E-06 | 1 | 1 | 1 | 1 |
| 6 | rs986475 | 0,036 | NCR3 | A | G | microarra | AA[U]AAA | rs2844479 | 0,317 | A | C | 19079260 | Obesity | Individ | Indivi | 2,00E-08 | 0,306 | 0,094 | 0,033 | 0,002 |
| 6 | rs986475 | 0,036 | NCR3 | A | G | microarra | AA[U]AAA | rs3117582 | 0,092 | G | T | 19836008 | Lung aden | Europ | Europ | 5,00E-12 | 0,066 | 0,004 | | |

APA SNP: SNP id, position, minimum allele frequency, gene where the SNP is located, the APA and WT alleles,associated with APA
and the analysis where the SNP was significantly

PolyA motif: signal hexamer showing the APA allele, its start and end

GWAS SNP: SNP id, position, minimum allele frequency, risk and non risk alleles, pubmed id, trait, samples, p-value

LD: r and r^2 values. Positive r shows a positive correlation between APA and risk alleles. High r^2 shows LD

APA alleles, WT alleles, Risk alleles, and non risk alleles are shown on the positive strand of DNA

mean(r ) and mean(r^2) are mean values computed for each APA SNP by taking the average r's when an APA SNP is paired to several GWAS SNPs

## Dissertations at the Faculty of Medicine, NTNU

**1977**

1. Knut Joachim Berg: EFFECT OF ACETYLSALICYLIC ACID ON RENAL FUNCTION
2. Karl Erik Viken and Arne Ødegaard: STUDIES ON HUMAN MONOCYTES CULTURED *IN VITRO*

**1978**

3. Karel Bjørn Cyvin: CONGENITAL DISLOCATION OF THE HIP JOINT.
4. Alf O. Brubakk: METHODS FOR STUDYING FLOW DYNAMICS IN THE LEFT VENTRICLE AND THE AORTA IN MAN.

**1979**

5. Geirmund Unsgaard: CYTOSTATIC AND IMMUNOREGULATORY ABILITIES OF HUMAN BLOOD MONOCYTES CULTURED IN VITRO

**1980**

6. Størker Jørstad: URAEMIC TOXINS
7. Arne Olav Jenssen: SOME RHEOLOGICAL, CHEMICAL AND STRUCTURAL PROPERTIES OF MUCOID SPUTUM FROM PATIENTS WITH CHRONIC OBSTRUCTIVE BRONCHITIS

**1981**

8. Jens Hammerstrøm: CYTOSTATIC AND CYTOLYTIC ACTIVITY OF HUMAN MONOCYTES AND EFFUSION MACROPHAGES AGAINST TUMOR CELLS *IN VITRO*

**1983**

9. Tore Syversen: EFFECTS OF METHYLMERCURY ON RAT BRAIN PROTEIN.
10. Torbjørn Iversen: SQUAMOUS CELL CARCINOMA OF THE VULVA.

**1984**

11. Tor-Erik Widerøe: ASPECTS OF CONTINUOUS AMBULATORY PERITONEAL DIALYSIS.
12. Anton Hole: ALTERATIONS OF MONOCYTE AND LYMPHOCYTE FUNCTIONS IN REALTION TO SURGERY UNDER EPIDURAL OR GENERAL ANAESTHESIA.
13. Terje Terjesen: FRACTURE HEALING AND STRESS-PROTECTION AFTER METAL PLATE FIXATION AND EXTERNAL FIXATION.
14. Carsten Saunte: CLUSTER HEADACHE SYNDROME.
15. Inggard Lereim: TRAFFIC ACCIDENTS AND THEIR CONSEQUENCES.
16. Bjørn Magne Eggen: STUDIES IN CYTOTOXICITY IN HUMAN ADHERENT MONONUCLEAR BLOOD CELLS.
17. Trond Haug: FACTORS REGULATING BEHAVIORAL EFFECTS OG DRUGS.

**1985**

18. Sven Erik Gisvold: RESUSCITATION AFTER COMPLETE GLOBAL BRAIN ISCHEMIA.
19. Terje Espevik: THE CYTOSKELETON OF HUMAN MONOCYTES.
20. Lars Bevanger: STUDIES OF THE Ibc (c) PROTEIN ANTIGENS OF GROUP B STREPTOCOCCI.
21. Ole-Jan Iversen: RETROVIRUS-LIKE PARTICLES IN THE PATHOGENESIS OF PSORIASIS.
22. Lasse Eriksen: EVALUATION AND TREATMENT OF ALCOHOL DEPENDENT BEHAVIOUR.
23. Per I. Lundmo: ANDROGEN METABOLISM IN THE PROSTATE.

1986

24. Dagfinn Berntzen: ANALYSIS AND MANAGEMENT OF EXPERIMENTAL AND CLINICAL PAIN.
25. Odd Arnold Kildahl-Andersen: PRODUCTION AND CHARACTERIZATION OF MONOCYTE-DERIVED CYTOTOXIN AND ITS ROLE IN MONOCYTE-MEDIATED CYTOTOXICITY.
26. Ola Dale: VOLATILE ANAESTHETICS.

**1987**

27. Per Martin Kleveland: STUDIES ON GASTRIN.
28. Audun N. Øksendal: THE CALCIUM PARADOX AND THE HEART.
29. Vilhjalmur R. Finsen: HIP FRACTURES

**1988**

30. Rigmor Austgulen: TUMOR NECROSIS FACTOR: A MONOCYTE-DERIVED REGULATOR OF CELLULAR GROWTH.
31. Tom-Harald Edna: HEAD INJURIES ADMITTED TO HOSPITAL.
32. Joseph D. Borsi: NEW ASPECTS OF THE CLINICAL PHARMACOKINETICS OF METHOTREXATE.
33. Olav F. M. Sellevold: GLUCOCORTICOIDS IN MYOCARDIAL PROTECTION.
34. Terje Skjærpe: NONINVASIVE QUANTITATION OF GLOBAL PARAMETERS ON LEFT VENTRICULAR FUNCTION: THE SYSTOLIC PULMONARY ARTERY PRESSURE AND CARDIAC OUTPUT.
35. Eyvind Rødahl: STUDIES OF IMMUNE COMPLEXES AND RETROVIRUS-LIKE ANTIGENS IN PATIENTS WITH ANKYLOSING SPONDYLITIS.
36. Ketil Thorstensen: STUDIES ON THE MECHANISMS OF CELLULAR UPTAKE OF IRON FROM TRANSFERRIN.
37. Anna Midelfart: STUDIES OF THE MECHANISMS OF ION AND FLUID TRANSPORT IN THE BOVINE CORNEA.
38. Eirik Helseth: GROWTH AND PLASMINOGEN ACTIVATOR ACTIVITY OF HUMAN GLIOMAS AND BRAIN METASTASES - WITH SPECIAL REFERENCE TO TRANSFORMING GROWTH FACTOR BETA AND THE EPIDERMAL GROWTH FACTOR RECEPTOR.
39. Petter C. Borchgrevink: MAGNESIUM AND THE ISCHEMIC HEART.
40. Kjell-Arne Rein: THE EFFECT OF EXTRACORPOREAL CIRCULATION ON SUBCUTANEOUS TRANSCAPILLARY FLUID BALANCE.
41. Arne Kristian Sandvik: RAT GASTRIC HISTAMINE.
42. Carl Bredo Dahl: ANIMAL MODELS IN PSYCHIATRY.

**1989**

43. Torbjørn A. Fredriksen: CERVICOGENIC HEADACHE.
44. Rolf A. Walstad: CEFTAZIDIME.
45. Rolf Salvesen: THE PUPIL IN CLUSTER HEADACHE.
46. Nils Petter Jørgensen: DRUG EXPOSURE IN EARLY PREGNANCY.
47. Johan C. Ræder: PREMEDICATION AND GENERAL ANAESTHESIA IN OUTPATIENT GYNECOLOGICAL SURGERY.
48. M. R. Shalaby: IMMUNOREGULATORY PROPERTIES OF TNF-$\alpha$ AND THE RELATED CYTOKINES.
49. Anders Waage: THE COMPLEX PATTERN OF CYTOKINES IN SEPTIC SHOCK.
50. Bjarne Christian Eriksen: ELECTROSTIMULATION OF THE PELVIC FLOOR IN FEMALE URINARY INCONTINENCE.
51. Tore B. Halvorsen: PROGNOSTIC FACTORS IN COLORECTAL CANCER.

**1990**

52. Asbjørn Nordby: CELLULAR TOXICITY OF ROENTGEN CONTRAST MEDIA.
53. Kåre E. Tvedt: X-RAY MICROANALYSIS OF BIOLOGICAL MATERIAL.
54. Tore C. Stiles: COGNITIVE VULNERABILITY FACTORS IN THE DEVELOPMENT AND MAINTENANCE OF DEPRESSION.
55. Eva Hofsli: TUMOR NECROSIS FACTOR AND MULTIDRUG RESISTANCE.
56. Helge S. Haarstad: TROPHIC EFFECTS OF CHOLECYSTOKININ AND SECRETIN ON THE RAT PANCREAS.
57. Lars Engebretsen: TREATMENT OF ACUTE ANTERIOR CRUCIATE LIGAMENT INJURIES.
58. Tarjei Rygnestad: DELIBERATE SELF-POISONING IN TRONDHEIM.
59. Arne Z. Henriksen: STUDIES ON CONSERVED ANTIGENIC DOMAINS ON MAJOR OUTER MEMBRANE PROTEINS FROM ENTEROBACTERIA.
60. Steinar Westin: UNEMPLOYMENT AND HEALTH: Medical and social consequences of a factory closure in a ten-year controlled follow-up study.
61. Ylva Sahlin: INJURY REGISTRATION, a tool for accident preventive work.
62. Helge Bjørnstad Pettersen: BIOSYNTHESIS OF COMPLEMENT BY HUMAN ALVEOLAR MACROPHAGES WITH SPECIAL REFERENCE TO SARCOIDOSIS.
63. Berit Schei: TRAPPED IN PAINFUL LOVE.
64. Lars J. Vatten: PROSPECTIVE STUDIES OF THE RISK OF BREAST CANCER IN A COHORT OF NORWEGIAN WOMAN.

**1991**

65. Kåre Bergh: APPLICATIONS OF ANTI-C5a SPECIFIC MONOCLONAL ANTIBODIES FOR THE ASSESSMENT OF COMPLEMENT ACTIVATION.
66. Svein Svenningsen: THE CLINICAL SIGNIFICANCE OF INCREASED FEMORAL ANTEVERSION.
67. Olbjørn Klepp: NONSEMINOMATOUS GERM CELL TESTIS CANCER: THERAPEUTIC OUTCOME AND PROGNOSTIC FACTORS.
68. Trond Sand: THE EFFECTS OF CLICK POLARITY ON BRAINSTEM AUDITORY EVOKED POTENTIALS AMPLITUDE, DISPERSION, AND LATENCY VARIABLES.
69. Kjetil B. Åsbakk: STUDIES OF A PROTEIN FROM PSORIATIC SCALE, PSO P27, WITH RESPECT TO ITS POTENTIAL ROLE IN IMMUNE REACTIONS IN PSORIASIS.
70. Arnulf Hestnes: STUDIES ON DOWN´S SYNDROME.
71. Randi Nygaard: LONG-TERM SURVIVAL IN CHILDHOOD LEUKEMIA.
72. Bjørn Hagen: THIO-TEPA.
73. Svein Anda: EVALUATION OF THE HIP JOINT BY COMPUTED TOMOGRAMPHY AND ULTRASONOGRAPHY.

**1992**

74. Martin Svartberg: AN INVESTIGATION OF PROCESS AND OUTCOME OF SHORT-TERM PSYCHODYNAMIC PSYCHOTHERAPY.
75. Stig Arild Slørdahl: AORTIC REGURGITATION.
76. Harold C Sexton: STUDIES RELATING TO THE TREATMENT OF SYMPTOMATIC NON-PSYCHOTIC PATIENTS.
77. Maurice B. Vincent: VASOACTIVE PEPTIDES IN THE OCULAR/FOREHEAD AREA.
78. Terje Johannessen: CONTROLLED TRIALS IN SINGLE SUBJECTS.
79. Turid Nilsen: PYROPHOSPHATE IN HEPATOCYTE IRON METABOLISM.
80. Olav Haraldseth: NMR SPECTROSCOPY OF CEREBRAL ISCHEMIA AND REPERFUSION IN RAT.
81. Eiliv Brenna: REGULATION OF FUNCTION AND GROWTH OF THE OXYNTIC MUCOSA.

**1993**

82. Gunnar Bovim: CERVICOGENIC HEADACHE.
83. Jarl Arne Kahn: ASSISTED PROCREATION.
84. Bjørn Naume: IMMUNOREGULATORY EFFECTS OF CYTOKINES ON NK CELLS.
85. Rune Wiseth: AORTIC VALVE REPLACEMENT.
86. Jie Ming Shen: BLOOD FLOW VELOCITY AND RESPIRATORY STUDIES.
87. Piotr Kruszewski: SUNCT SYNDROME WITH SPECIAL REFERENCE TO THE AUTONOMIC NERVOUS SYSTEM.
88. Mette Haase Moen: ENDOMETRIOSIS.
89. Anne Vik: VASCULAR GAS EMBOLISM DURING AIR INFUSION AND AFTER DECOMPRESSION IN PIGS.
90. Lars Jacob Stovner: THE CHIARI TYPE I MALFORMATION.
91. Kjell Å. Salvesen: ROUTINE ULTRASONOGRAPHY IN UTERO AND DEVELOPMENT IN CHILDHOOD.

**1994**

92. Nina-Beate Liabakk: DEVELOPMENT OF IMMUNOASSAYS FOR TNF AND ITS SOLUBLE RECEPTORS.
93. Sverre Helge Torp: *erb*B ONCOGENES IN HUMAN GLIOMAS AND MENINGIOMAS.
94. Olav M. Linaker: MENTAL RETARDATION AND PSYCHIATRY. Past and present.
95. Per Oscar Feet: INCREASED ANTIDEPRESSANT AND ANTIPANIC EFFECT IN COMBINED TREATMENT WITH DIXYRAZINE AND TRICYCLIC ANTIDEPRESSANTS.
96. Stein Olav Samstad: CROSS SECTIONAL FLOW VELOCITY PROFILES FROM TWO-DIMENSIONAL DOPPLER ULTRASOUND: Studies on early mitral blood flow.
97. Bjørn Backe: STUDIES IN ANTENATAL CARE.
98. Gerd Inger Ringdal: QUALITY OF LIFE IN CANCER PATIENTS.
99. Torvid Kiserud: THE DUCTUS VENOSUS IN THE HUMAN FETUS.
100. Hans E. Fjøsne: HORMONAL REGULATION OF PROSTATIC METABOLISM.
101. Eylert Brodtkorb: CLINICAL ASPECTS OF EPILEPSY IN THE MENTALLY RETARDED.
102. Roar Juul: PEPTIDERGIC MECHANISMS IN HUMAN SUBARACHNOID HEMORRHAGE.
103. Unni Syversen: CHROMOGRANIN A. Phsysiological and Clinical Role.

**1995**

104. Odd Gunnar Brakstad: THERMOSTABLE NUCLEASE AND THE *nuc* GENE IN THE DIAGNOSIS OF *Staphylococcus aureus* INFECTIONS.
105. Terje Engan: NUCLEAR MAGNETIC RESONANCE (NMR) SPECTROSCOPY OF PLASMA IN MALIGNANT DISEASE.
106. Kirsten Rasmussen: VIOLENCE IN THE MENTALLY DISORDERED.
107. Finn Egil Skjeldestad: INDUCED ABORTION: Timetrends and Determinants.
108. Roar Stenseth: THORACIC EPIDURAL ANALGESIA IN AORTOCORONARY BYPASS SURGERY.
109. Arild Faxvaag: STUDIES OF IMMUNE CELL FUNCTION *in mice infected with* MURINE RETROVIRUS.

**1996**

110. Svend Aakhus: NONINVASIVE COMPUTERIZED ASSESSMENT OF LEFT VENTRICULAR FUNCTION AND SYSTEMIC ARTERIAL PROPERTIES. Methodology and some clinical applications.
111. Klaus-Dieter Bolz: INTRAVASCULAR ULTRASONOGRAPHY.
112. Petter Aadahl: CARDIOVASCULAR EFFECTS OF THORACIC AORTIC CROSS-CLAMPING.
113. Sigurd Steinshamn: CYTOKINE MEDIATORS DURING GRANULOCYTOPENIC INFECTIONS.
114. Hans Stifoss-Hanssen: SEEKING MEANING OR HAPPINESS?
115. Anne Kvikstad: LIFE CHANGE EVENTS AND MARITAL STATUS IN RELATION TO RISK AND PROGNOSIS OF CANCER.
116. Torbjørn Grøntvedt: TREATMENT OF ACUTE AND CHRONIC ANTERIOR CRUCIATE LIGAMENT INJURIES. A clinical and biomechanical study.
117. Sigrid Hørven Wigers: CLINICAL STUDIES OF FIBROMYALGIA WITH FOCUS ON ETIOLOGY, TREATMENT AND OUTCOME.
118. Jan Schjøtt: MYOCARDIAL PROTECTION: Functional and Metabolic Characteristics of Two Endogenous Protective Principles.
119. Marit Martinussen: STUDIES OF INTESTINAL BLOOD FLOW AND ITS RELATION TO TRANSITIONAL CIRCULATORY ADAPATION IN NEWBORN INFANTS.
120. Tomm B. Müller: MAGNETIC RESONANCE IMAGING IN FOCAL CEREBRAL ISCHEMIA.
121. Rune Haaverstad: OEDEMA FORMATION OF THE LOWER EXTREMITIES.
122. Magne Børset: THE ROLE OF CYTOKINES IN MULTIPLE MYELOMA, WITH SPECIAL REFERENCE TO HEPATOCYTE GROWTH FACTOR.
123. Geir Smedslund: A THEORETICAL AND EMPIRICAL INVESTIGATION OF SMOKING, STRESS AND DISEASE: RESULTS FROM A POPULATION SURVEY.

**1997**

124. Torstein Vik: GROWTH, MORBIDITY, AND PSYCHOMOTOR DEVELOPMENT IN INFANTS WHO WERE GROWTH RETARDED *IN UTERO*.
125. Siri Forsmo: ASPECTS AND CONSEQUENCES OF OPPORTUNISTIC SCREENING FOR CERVICAL CANCER. Results based on data from three Norwegian counties.
126. Jon S. Skranes: CEREBRAL MRI AND NEURODEVELOPMENTAL OUTCOME IN VERY LOW BIRTH WEIGHT (VLBW) CHILDREN. A follow-up study of a geographically based year cohort of VLBW children at ages one and six years.
127. Knut Bjørnstad: COMPUTERIZED ECHOCARDIOGRAPHY FOR EVALUTION OF CORONARY ARTERY DISEASE.
128. Grethe Elisabeth Borchgrevink: DIAGNOSIS AND TREATMENT OF WHIPLASH/NECK SPRAIN INJURIES CAUSED BY CAR ACCIDENTS.
129. Tor Elsås: NEUROPEPTIDES AND NITRIC OXIDE SYNTHASE IN OCULAR AUTONOMIC AND SENSORY NERVES.
130. Rolf W. Gråwe: EPIDEMIOLOGICAL AND NEUROPSYCHOLOGICAL PERSPECTIVES ON SCHIZOPHRENIA.
131. Tonje Strømholm: CEREBRAL HAEMODYNAMICS DURING THORACIC AORTIC CROSSCLAMPING. An experimental study in pigs

**1998**

132. Martinus Bråten: STUDIES ON SOME PROBLEMS REALTED TO INTRAMEDULLARY NAILING OF FEMORAL FRACTURES.

133. Ståle Nordgård: PROLIFERATIVE ACTIVITY AND DNA CONTENT AS PROGNOSTIC INDICATORS IN ADENOID CYSTIC CARCINOMA OF THE HEAD AND NECK.
134. Egil Lien: SOLUBLE RECEPTORS FOR **TNF** AND **LPS**: RELEASE PATTERN AND POSSIBLE SIGNIFICANCE IN DISEASE.
135. Marit Bjørgaas: HYPOGLYCAEMIA IN CHILDREN WITH DIABETES MELLITUS
136. Frank Skorpen: GENETIC AND FUNCTIONAL ANALYSES OF DNA REPAIR IN HUMAN CELLS.
137. Juan A. Pareja: SUNCT SYNDROME. ON THE CLINICAL PICTURE. ITS DISTINCTION FROM OTHER, SIMILAR HEADACHES.
138. Anders Angelsen: NEUROENDOCRINE CELLS IN HUMAN PROSTATIC CARCINOMAS AND THE PROSTATIC COMPLEX OF RAT, GUINEA PIG, CAT AND DOG.
139. Fabio Antonaci: CHRONIC PAROXYSMAL HEMICRANIA AND HEMICRANIA CONTINUA: TWO DIFFERENT ENTITIES?
140. Sven M. Carlsen: ENDOCRINE AND METABOLIC EFFECTS OF METFORMIN WITH SPECIAL EMPHASIS ON CARDIOVASCULAR RISK FACTORES.

**1999**

141. Terje A. Murberg: DEPRESSIVE SYMPTOMS AND COPING AMONG PATIENTS WITH CONGESTIVE HEART FAILURE.
142. Harm-Gerd Karl Blaas: THE EMBRYONIC EXAMINATION. Ultrasound studies on the development of the human embryo.
143. Noèmi Becser Andersen:THE CEPHALIC SENSORY NERVES IN UNILATERAL HEADACHES. Anatomical background and neurophysiological evaluation.
144. Eli-Janne Fiskerstrand: LASER TREATMENT OF PORT WINE STAINS. A study of the efficacy and limitations of the pulsed dye laser. Clinical and morfological analyses aimed at improving the therapeutic outcome.
145. Bård Kulseng: A STUDY OF ALGINATE CAPSULE PROPERTIES AND CYTOKINES IN RELATION TO INSULIN DEPENDENT DIABETES MELLITUS.
146. Terje Haug: STRUCTURE AND REGULATION OF THE HUMAN UNG GENE ENCODING URACIL-DNA GLYCOSYLASE.
147. Heidi Brurok: MANGANESE AND THE HEART. A Magic Metal with Diagnostic and Therapeutic Possibilites.
148. Agnes Kathrine Lie: DIAGNOSIS AND PREVALENCE OF HUMAN PAPILLOMAVIRUS INFECTION IN CERVICAL INTRAEPITELIAL NEOPLASIA. Relationship to Cell Cycle Regulatory Proteins and HLA DQBI Genes.
149. Ronald Mårvik: PHARMACOLOGICAL, PHYSIOLOGICAL AND PATHOPHYSIOLOGICAL STUDIES ON ISOLATED STOMACS.
150. Ketil Jarl Holen: THE ROLE OF ULTRASONOGRAPHY IN THE DIAGNOSIS AND TREATMENT OF HIP DYSPLASIA IN NEWBORNS.
151. Irene Hetlevik: THE ROLE OF CLINICAL GUIDELINES IN CARDIOVASCULAR RISK INTERVENTION IN GENERAL PRACTICE.
152. Katarina Tunòn: ULTRASOUND AND PREDICTION OF GESTATIONAL AGE.
153. Johannes Soma: INTERACTION BETWEEN THE LEFT VENTRICLE AND THE SYSTEMIC ARTERIES.
154. Arild Aamodt: DEVELOPMENT AND PRE-CLINICAL EVALUATION OF A CUSTOM-MADE FEMORAL STEM.
155. Agnar Tegnander: DIAGNOSIS AND FOLLOW-UP OF CHILDREN WITH SUSPECTED OR KNOWN HIP DYSPLASIA.
156. Bent Indredavik: STROKE UNIT TREATMENT: SHORT AND LONG-TERM EFFECTS
157. Jolanta Vanagaite Vingen: PHOTOPHOBIA AND PHONOPHOBIA IN PRIMARY HEADACHES

**2000**

158. Ola Dalsegg Sæther: PATHOPHYSIOLOGY DURING PROXIMAL AORTIC CROSS-CLAMPING CLINICAL AND EXPERIMENTAL STUDIES
159. xxxxxxxxx (blind number)
160. Christina Vogt Isaksen: PRENATAL ULTRASOUND AND POSTMORTEM FINDINGS – A TEN YEAR CORRELATIVE STUDY OF FETUSES AND INFANTS WITH DEVELOPMENTAL ANOMALIES.
161. Holger Seidel: HIGH-DOSE METHOTREXATE THERAPY IN CHILDREN WITH ACUTE LYMPHOCYTIC LEUKEMIA: DOSE, CONCENTRATION, AND EFFECT CONSIDERATIONS.

162. Stein Hallan: IMPLEMENTATION OF MODERN MEDICAL DECISION ANALYSIS INTO CLINICAL DIAGNOSIS AND TREATMENT.
163. Malcolm Sue-Chu: INVASIVE AND NON-INVASIVE STUDIES IN CROSS-COUNTRY SKIERS WITH ASTHMA-LIKE SYMPTOMS.
164. Ole-Lars Brekke: EFFECTS OF ANTIOXIDANTS AND FATTY ACIDS ON TUMOR NECROSIS FACTOR-INDUCED CYTOTOXICITY.
165. Jan Lundbom: AORTOCORONARY BYPASS SURGERY: CLINICAL ASPECTS, COST CONSIDERATIONS AND WORKING ABILITY.
166. John-Anker Zwart: LUMBAR NERVE ROOT COMPRESSION, BIOCHEMICAL AND NEUROPHYSIOLOGICAL ASPECTS.
167. Geir Falck: HYPEROSMOLALITY AND THE HEART.
168. Eirik Skogvoll: CARDIAC ARREST Incidence, Intervention and Outcome.
169. Dalius Bansevicius: SHOULDER-NECK REGION IN CERTAIN HEADACHES AND CHRONIC PAIN SYNDROMES.
170. Bettina Kinge: REFRACTIVE ERRORS AND BIOMETRIC CHANGES AMONG UNIVERSITY STUDENTS IN NORWAY.
171. Gunnar Qvigstad: CONSEQUENCES OF HYPERGASTRINEMIA IN MAN
172. Hanne Ellekjær: EPIDEMIOLOGICAL STUDIES OF STROKE IN A NORWEGIAN POPULATION. INCIDENCE, RISK FACTORS AND PROGNOSIS
173. Hilde Grimstad: VIOLENCE AGAINST WOMEN AND PREGNANCY OUTCOME.
174. Astrid Hjelde: SURFACE TENSION AND COMPLEMENT ACTIVATION: Factors influencing bubble formation and bubble effects after decompression.
175. Kjell A. Kvistad: MR IN BREAST CANCER – A CLINICAL STUDY.
176. Ivar Rossvoll: ELECTIVE ORTHOPAEDIC SURGERY IN A DEFINED POPULATION. Studies on demand, waiting time for treatment and incapacity for work.
177. Carina Seidel: PROGNOSTIC VALUE AND BIOLOGICAL EFFECTS OF HEPATOCYTE GROWTH FACTOR AND SYNDECAN-1 IN MULTIPLE MYELOMA.

**2001**

178. Alexander Wahba: THE INFLUENCE OF CARDIOPULMONARY BYPASS ON PLATELET FUNCTION AND BLOOD COAGULATION – DETERMINANTS AND CLINICAL CONSEQUENSES
179. Marcus Schmitt-Egenolf: THE RELEVANCE OF THE MAJOR hISTOCOMPATIBILITY COMPLEX FOR THE GENETICS OF PSORIASIS
180. Odrun Arna Gederaas: BIOLOGICAL MECHANISMS INVOLVED IN 5-AMINOLEVULINIC ACID BASED PHOTODYNAMIC THERAPY
181. Pål Richard Romundstad: CANCER INCIDENCE AMONG NORWEGIAN ALUMINIUM WORKERS
182. Henrik Hjorth-Hansen: NOVEL CYTOKINES IN GROWTH CONTROL AND BONE DISEASE OF MULTIPLE MYELOMA
183. Gunnar Morken: SEASONAL VARIATION OF HUMAN MOOD AND BEHAVIOUR
184. Bjørn Olav Haugen: MEASUREMENT OF CARDIAC OUTPUT AND STUDIES OF VELOCITY PROFILES IN AORTIC AND MITRAL FLOW USING TWO- AND THREE-DIMENSIONAL COLOUR FLOW IMAGING
185. Geir Bråthen: THE CLASSIFICATION AND CLINICAL DIAGNOSIS OF ALCOHOL-RELATED SEIZURES
186. Knut Ivar Aasarød: RENAL INVOLVEMENT IN INFLAMMATORY RHEUMATIC DISEASE. A Study of Renal Disease in Wegener’s Granulomatosis and in Primary Sjögren’s Syndrome
187. Trude Helen Flo: RESEPTORS INVOLVED IN CELL ACTIVATION BY DEFINED URONIC ACID POLYMERS AND BACTERIAL COMPONENTS
188. Bodil Kavli: HUMAN URACIL-DNA GLYCOSYLASES FROM THE UNG GENE: STRUCTRUAL BASIS FOR SUBSTRATE SPECIFICITY AND REPAIR
189. Liv Thommesen: MOLECULAR MECHANISMS INVOLVED IN TNF- AND GASTRIN-MEDIATED GENE REGULATION
190. Turid Lingaas Holmen: SMOKING AND HEALTH IN ADOLESCENCE; THE NORD-TRØNDELAG HEALTH STUDY, 1995-97
191. Øyvind Hjertner: MULTIPLE MYELOMA: INTERACTIONS BETWEEN MALIGNANT PLASMA CELLS AND THE BONE MICROENVIRONMENT

192. Asbjørn Støylen: STRAIN RATE IMAGING OF THE LEFT VENTRICLE BY ULTRASOUND. FEASIBILITY, CLINICAL VALIDATION AND PHYSIOLOGICAL ASPECTS
193. Kristian Midthjell: DIABETES IN ADULTS IN NORD-TRØNDELAG. PUBLIC HEALTH ASPECTS OF DIABETES MELLITUS IN A LARGE, NON-SELECTED NORWEGIAN POPULATION.
194. Guanglin Cui: FUNCTIONAL ASPECTS OF THE ECL CELL IN RODENTS
195. Ulrik Wisløff: CARDIAC EFFECTS OF AEROBIC ENDURANCE TRAINING: HYPERTROPHY, CONTRACTILITY AND CALCUIM HANDLING IN NORMAL AND FAILING HEART
196. Øyvind Halaas: MECHANISMS OF IMMUNOMODULATION AND CELL-MEDIATED CYTOTOXICITY INDUCED BY BACTERIAL PRODUCTS
197. Tore Amundsen: PERFUSION MR IMAGING IN THE DIAGNOSIS OF PULMONARY EMBOLISM
198. Nanna Kurtze: THE SIGNIFICANCE OF ANXIETY AND DEPRESSION IN FATIQUE AND PATTERNS OF PAIN AMONG INDIVIDUALS DIAGNOSED WITH FIBROMYALGIA: RELATIONS WITH QUALITY OF LIFE, FUNCTIONAL DISABILITY, LIFESTYLE, EMPLOYMENT STATUS, CO-MORBIDITY AND GENDER
199. Tom Ivar Lund Nilsen: PROSPECTIVE STUDIES OF CANCER RISK IN NORD-TRØNDELAG: THE HUNT STUDY. Associations with anthropometric, socioeconomic, and lifestyle risk factors
200. Asta Kristine Håberg: A NEW APPROACH TO THE STUDY OF MIDDLE CEREBRAL ARTERY OCCLUSION IN THE RAT USING MAGNETIC RESONANCE TECHNIQUES

**2002**

201. Knut Jørgen Arntzen: PREGNANCY AND CYTOKINES
202. Henrik Døllner: INFLAMMATORY MEDIATORS IN PERINATAL INFECTIONS
203. Asta Bye: LOW FAT, LOW LACTOSE DIET USED AS PROPHYLACTIC TREATMENT OF ACUTE INTESTINAL REACTIONS DURING PELVIC RADIOTHERAPY. A PROSPECTIVE RANDOMISED STUDY.
204. Sylvester Moyo: STUDIES ON STREPTOCOCCUS AGALACTIAE (GROUP B STREPTOCOCCUS) SURFACE-ANCHORED MARKERS WITH EMPHASIS ON STRAINS AND HUMAN SERA FROM ZIMBABWE.
205. Knut Hagen: HEAD-HUNT: THE EPIDEMIOLOGY OF HEADACHE IN NORD-TRØNDELAG
206. Li Lixin: ON THE REGULATION AND ROLE OF UNCOUPLING PROTEIN-2 IN INSULIN PRODUCING ß-CELLS
207. Anne Hildur Henriksen: SYMPTOMS OF ALLERGY AND ASTHMA VERSUS MARKERS OF LOWER AIRWAY INFLAMMATION AMONG ADOLESCENTS
208. Egil Andreas Fors: NON-MALIGNANT PAIN IN RELATION TO PSYCHOLOGICAL AND ENVIRONTENTAL FACTORS. EXPERIENTAL AND CLINICAL STUDES OF PAIN WITH FOCUS ON FIBROMYALGIA
209. Pål Klepstad: MORPHINE FOR CANCER PAIN
210. Ingunn Bakke: MECHANISMS AND CONSEQUENCES OF PEROXISOME PROLIFERATOR-INDUCED HYPERFUNCTION OF THE RAT GASTRIN PRODUCING CELL
211. Ingrid Susann Gribbestad: MAGNETIC RESONANCE IMAGING AND SPECTROSCOPY OF BREAST CANCER
212. Rønnaug Astri Ødegård: PREECLAMPSIA – MATERNAL RISK FACTORS AND FETAL GROWTH
213. Johan Haux: STUDIES ON CYTOTOXICITY INDUCED BY HUMAN NATURAL KILLER CELLS AND DIGITOXIN
214. Turid Suzanne Berg-Nielsen: PARENTING PRACTICES AND MENTALLY DISORDERED ADOLESCENTS
215. Astrid Rydning: BLOOD FLOW AS A PROTECTIVE FACTOR FOR THE STOMACH MUCOSA. AN EXPERIMENTAL STUDY ON THE ROLE OF MAST CELLS AND SENSORY AFFERENT NEURONS

**2003**

216. Jan Pål Loennechen: HEART FAILURE AFTER MYOCARDIAL INFARCTION. Regional Differences, Myocyte Function, Gene Expression, and Response to Cariporide, Losartan, and Exercise Training.

217. Elisabeth Qvigstad: EFFECTS OF FATTY ACIDS AND OVER-STIMULATION ON INSULIN SECRETION IN MAN
218. Arne Åsberg: EPIDEMIOLOGICAL STUDIES IN HEREDITARY HEMOCHROMATOSIS: PREVALENCE, MORBIDITY AND BENEFIT OF SCREENING.
219. Johan Fredrik Skomsvoll: REPRODUCTIVE OUTCOME IN WOMEN WITH RHEUMATIC DISEASE. A population registry based study of the effects of inflammatory rheumatic disease and connective tissue disease on reproductive outcome in Norwegian women in 1967-1995.
220. Siv Mørkved: URINARY INCONTINENCE DURING PREGNANCY AND AFTER DELIVERY: EFFECT OF PELVIC FLOOR MUSCLE TRAINING IN PREVENTION AND TREATMENT
221. Marit S. Jordhøy: THE IMPACT OF COMPREHENSIVE PALLIATIVE CARE
222. Tom Christian Martinsen: HYPERGASTRINEMIA AND HYPOACIDITY IN RODENTS – CAUSES AND CONSEQUENCES
223. Solveig Tingulstad: CENTRALIZATION OF PRIMARY SURGERY FOR OVARAIN CANCER. FEASIBILITY AND IMPACT ON SURVIVAL
224. Haytham Eloqayli: METABOLIC CHANGES IN THE BRAIN CAUSED BY EPILEPTIC SEIZURES
225. Torunn Bruland: STUDIES OF EARLY RETROVIRUS-HOST INTERACTIONS – VIRAL DETERMINANTS FOR PATHOGENESIS AND THE INFLUENCE OF SEX ON THE SUSCEPTIBILITY TO FRIEND MURINE LEUKAEMIA VIRUS INFECTION
226. Torstein Hole: DOPPLER ECHOCARDIOGRAPHIC EVALUATION OF LEFT VENTRICULAR FUNCTION IN PATIENTS WITH ACUTE MYOCARDIAL INFARCTION
227. Vibeke Nossum: THE EFFECT OF VASCULAR BUBBLES ON ENDOTHELIAL FUNCTION
228. Sigurd Fasting: ROUTINE BASED RECORDING OF ADVERSE EVENTS DURING ANAESTHESIA – APPLICATION IN QUALITY IMPROVEMENT AND SAFETY
229. Solfrid Romundstad: EPIDEMIOLOGICAL STUDIES OF MICROALBUMINURIA. THE NORD-TRØNDELAG HEALTH STUDY 1995-97 (HUNT 2)
230. Geir Torheim: PROCESSING OF DYNAMIC DATA SETS IN MAGNETIC RESONANCE IMAGING
231. Catrine Ahlén: SKIN INFECTIONS IN OCCUPATIONAL SATURATION DIVERS IN THE NORTH SEA AND THE IMPACT OF THE ENVIRONMENT
232. Arnulf Langhammer: RESPIRATORY SYMPTOMS, LUNG FUNCTION AND BONE MINERAL DENSITY IN A COMPREHENSIVE POPULATION SURVEY. THE NORD-TRØNDELAG HEALTH STUDY 1995-97. THE BRONCHIAL OBSTRUCTION IN NORD-TRØNDELAG STUDY
233. Einar Kjelsås: EATING DISORDERS AND PHYSICAL ACTIVITY IN NON-CLINICAL SAMPLES
234. Arne Wibe: RECTAL CANCER TREATMENT IN NORWAY – STANDARDISATION OF SURGERY AND QUALITY ASSURANCE

**2004**

235. Eivind Witsø: BONE GRAFT AS AN ANTIBIOTIC CARRIER
236. Anne Mari Sund: DEVELOPMENT OF DEPRESSIVE SYMPTOMS IN EARLY ADOLESCENCE
237. Hallvard Lærum: EVALUATION OF ELECTRONIC MEDICAL RECORDS – A CLINICAL TASK PERSPECTIVE
238. Gustav Mikkelsen: ACCESSIBILITY OF INFORMATION IN ELECTRONIC PATIENT RECORDS; AN EVALUATION OF THE ROLE OF DATA QUALITY
239. Steinar Krokstad: SOCIOECONOMIC INEQUALITIES IN HEALTH AND DISABILITY. SOCIAL EPIDEMIOLOGY IN THE NORD-TRØNDELAG HEALTH STUDY (HUNT), NORWAY
240. Arne Kristian Myhre: NORMAL VARIATION IN ANOGENITAL ANATOMY AND MICROBIOLOGY IN NON-ABUSED PRESCHOOL CHILDREN
241. Ingunn Dybedal: NEGATIVE REGULATORS OF HEMATOPOIETEC STEM AND PROGENITOR CELLS
242. Beate Sitter: TISSUE CHARACTERIZATION BY HIGH RESOLUTION MAGIC ANGLE SPINNING MR SPECTROSCOPY
243. Per Arne Aas: MACROMOLECULAR MAINTENANCE IN HUMAN CELLS – REPAIR OF URACIL IN DNA AND METHYLATIONS IN DNA AND RNA

244. Anna Bofin: FINE NEEDLE ASPIRATION CYTOLOGY IN THE PRIMARY INVESTIGATION OF BREAST TUMOURS AND IN THE DETERMINATION OF TREATMENT STRATEGIES
245. Jim Aage Nøttestad: DEINSTITUTIONALIZATION AND MENTAL HEALTH CHANGES AMONG PEOPLE WITH MENTAL RETARDATION
246. Reidar Fossmark: GASTRIC CANCER IN JAPANESE COTTON RATS
247. Wibeke Nordhøy: MANGANESE AND THE HEART, INTRACELLULAR MR RELAXATION AND WATER EXCHANGE ACROSS THE CARDIAC CELL MEMBRANE

**2005**

248. Sturla Molden: QUANTITATIVE ANALYSES OF SINGLE UNITS RECORDED FROM THE HIPPOCAMPUS AND ENTORHINAL CORTEX OF BEHAVING RATS
249. Wenche Brenne Drøyvold: EPIDEMIOLOGICAL STUDIES ON WEIGHT CHANGE AND HEALTH IN A LARGE POPULATION. THE NORD-TRØNDELAG HEALTH STUDY (HUNT)
250. Ragnhild Støen: ENDOTHELIUM-DEPENDENT VASODILATION IN THE FEMORAL ARTERY OF DEVELOPING PIGLETS
251. Aslak Steinsbekk: HOMEOPATHY IN THE PREVENTION OF UPPER RESPIRATORY TRACT INFECTIONS IN CHILDREN
252. Hill-Aina Steffenach: MEMORY IN HIPPOCAMPAL AND CORTICO-HIPPOCAMPAL CIRCUITS
253. Eystein Stordal: ASPECTS OF THE EPIDEMIOLOGY OF DEPRESSIONS BASED ON SELF-RATING IN A LARGE GENERAL HEALTH STUDY (THE HUNT-2 STUDY)
254. Viggo Pettersen: FROM MUSCLES TO SINGING: THE ACTIVITY OF ACCESSORY BREATHING MUSCLES AND THORAX MOVEMENT IN CLASSICAL SINGING
255. Marianne Fyhn: SPATIAL MAPS IN THE HIPPOCAMPUS AND ENTORHINAL CORTEX
256. Robert Valderhaug: OBSESSIVE-COMPULSIVE DISORDER AMONG CHILDREN AND ADOLESCENTS: CHARACTERISTICS AND PSYCHOLOGICAL MANAGEMENT OF PATIENTS IN OUTPATIENT PSYCHIATRIC CLINICS
257. Erik Skaaheim Haug: INFRARENAL ABDOMINAL AORTIC ANEURYSMS – COMORBIDITY AND RESULTS FOLLOWING OPEN SURGERY
258. Daniel Kondziella: GLIAL-NEURONAL INTERACTIONS IN EXPERIMENTAL BRAIN DISORDERS
259. Vegard Heimly Brun: ROUTES TO SPATIAL MEMORY IN HIPPOCAMPAL PLACE CELLS
260. Kenneth McMillan: PHYSIOLOGICAL ASSESSMENT AND TRAINING OF ENDURANCE AND STRENGTH IN PROFESSIONAL YOUTH SOCCER PLAYERS
261. Marit Sæbø Indredavik: MENTAL HEALTH AND CEREBRAL MAGNETIC RESONANCE IMAGING IN ADOLESCENTS WITH LOW BIRTH WEIGHT
262. Ole Johan Kemi: ON THE CELLULAR BASIS OF AEROBIC FITNESS, INTENSITY-DEPENDENCE AND TIME-COURSE OF CARDIOMYOCYTE AND ENDOTHELIAL ADAPTATIONS TO EXERCISE TRAINING
263. Eszter Vanky: POLYCYSTIC OVARY SYNDROME – METFORMIN TREATMENT IN PREGNANCY
264. Hild Fjærtoft: EXTENDED STROKE UNIT SERVICE AND EARLY SUPPORTED DISCHARGE. SHORT AND LONG-TERM EFFECTS
265. Grete Dyb: POSTTRAUMATIC STRESS REACTIONS IN CHILDREN AND ADOLESCENTS
266. Vidar Fykse: SOMATOSTATIN AND THE STOMACH
267. Kirsti Berg: OXIDATIVE STRESS AND THE ISCHEMIC HEART: A STUDY IN PATIENTS UNDERGOING CORONARY REVASCULARIZATION
268. Björn Inge Gustafsson: THE SEROTONIN PRODUCING ENTEROCHROMAFFIN CELL, AND EFFECTS OF HYPERSEROTONINEMIA ON HEART AND BONE

**2006**

269. Torstein Baade Rø: EFFECTS OF BONE MORPHOGENETIC PROTEINS, HEPATOCYTE GROWTH FACTOR AND INTERLEUKIN-21 IN MULTIPLE MYELOMA
270. May-Britt Tessem: METABOLIC EFFECTS OF ULTRAVIOLET RADIATION ON THE ANTERIOR PART OF THE EYE
271. Anne-Sofie Helvik: COPING AND EVERYDAY LIFE IN A POPULATION OF ADULTS WITH HEARING IMPAIRMENT

272. Therese Standal: MULTIPLE MYELOMA: THE INTERPLAY BETWEEN MALIGNANT PLASMA CELLS AND THE BONE MARROW MICROENVIRONMENT
273. Ingvild Saltvedt: TREATMENT OF ACUTELY SICK, FRAIL ELDERLY PATIENTS IN A GERIATRIC EVALUATION AND MANAGEMENT UNIT – RESULTS FROM A PROSPECTIVE RANDOMISED TRIAL
274. Birger Henning Endreseth: STRATEGIES IN RECTAL CANCER TREATMENT – FOCUS ON EARLY RECTAL CANCER AND THE INFLUENCE OF AGE ON PROGNOSIS
275. Anne Mari Aukan Rokstad: ALGINATE CAPSULES AS BIOREACTORS FOR CELL THERAPY
276. Mansour Akbari: HUMAN BASE EXCISION REPAIR FOR PRESERVATION OF GENOMIC STABILITY
277. Stein Sundstrøm: IMPROVING TREATMENT IN PATIENTS WITH LUNG CANCER – RESULTS FROM TWO MULITCENTRE RANDOMISED STUDIES
278. Hilde Pleym: BLEEDING AFTER CORONARY ARTERY BYPASS SURGERY - STUDIES ON HEMOSTATIC MECHANISMS, PROPHYLACTIC DRUG TREATMENT AND EFFECTS OF AUTOTRANSFUSION
279. Line Merethe Oldervoll: PHYSICAL ACTIVITY AND EXERCISE INTERVENTIONS IN CANCER PATIENTS
280. Boye Welde: THE SIGNIFICANCE OF ENDURANCE TRAINING, RESISTANCE TRAINING AND MOTIVATIONAL STYLES IN ATHLETIC PERFORMANCE AMONG ELITE JUNIOR CROSS-COUNTRY SKIERS
281. Per Olav Vandvik: IRRITABLE BOWEL SYNDROME IN NORWAY, STUDIES OF PREVALENCE, DIAGNOSIS AND CHARACTERISTICS IN GENERAL PRACTICE AND IN THE POPULATION
282. Idar Kirkeby-Garstad: CLINICAL PHYSIOLOGY OF EARLY MOBILIZATION AFTER CARDIAC SURGERY
283. Linn Getz: SUSTAINABLE AND RESPONSIBLE PREVENTIVE MEDICINE. CONCEPTUALISING ETHICAL DILEMMAS ARISING FROM CLINICAL IMPLEMENTATION OF ADVANCING MEDICAL TECHNOLOGY
284. Eva Tegnander: DETECTION OF CONGENITAL HEART DEFECTS IN A NON-SELECTED POPULATION OF 42,381 FETUSES
285. Kristin Gabestad Nørsett: GENE EXPRESSION STUDIES IN GASTROINTESTINAL PATHOPHYSIOLOGY AND NEOPLASIA
286. Per Magnus Haram: GENETIC VS. AQUIRED FITNESS: METABOLIC, VASCULAR AND CARDIOMYOCYTE ADAPTATIONS
287. Agneta Johansson: GENERAL RISK FACTORS FOR GAMBLING PROBLEMS AND THE PREVALENCE OF PATHOLOGICAL GAMBLING IN NORWAY
288. Svein Artur Jensen: THE PREVALENCE OF SYMPTOMATIC ARTERIAL DISEASE OF THE LOWER LIMB
289. Charlotte Björk Ingul: QUANITIFICATION OF REGIONAL MYOCARDIAL FUNCTION BY STRAIN RATE AND STRAIN FOR EVALUATION OF CORONARY ARTERY DISEASE. AUTOMATED VERSUS MANUAL ANALYSIS DURING ACUTE MYOCARDIAL INFARCTION AND DOBUTAMINE STRESS ECHOCARDIOGRAPHY
290. Jakob Nakling: RESULTS AND CONSEQUENCES OF ROUTINE ULTRASOUND SCREENING IN PREGNANCY – A GEOGRAPHIC BASED POPULATION STUDY
291. Anne Engum: DEPRESSION AND ANXIETY – THEIR RELATIONS TO THYROID DYSFUNCTION AND DIABETES IN A LARGE EPIDEMIOLOGICAL STUDY
292. Ottar Bjerkeset: ANXIETY AND DEPRESSION IN THE GENERAL POPULATION: RISK FACTORS, INTERVENTION AND OUTCOME – THE NORD-TRØNDELAG HEALTH STUDY (HUNT)
293. Jon Olav Drogset: RESULTS AFTER SURGICAL TREATMENT OF ANTERIOR CRUCIATE LIGAMENT INJURIES – A CLINICAL STUDY
294. Lars Fosse: MECHANICAL BEHAVIOUR OF COMPACTED MORSELLISED BONE – AN EXPERIMENTAL IN VITRO STUDY
295. Gunilla Klensmeden Fosse: MENTAL HEALTH OF PSYCHIATRIC OUTPATIENTS BULLIED IN CHILDHOOD
296. Paul Jarle Mork: MUSCLE ACTIVITY IN WORK AND LEISURE AND ITS ASSOCIATION TO MUSCULOSKELETAL PAIN

297.Björn Stenström: LESSONS FROM RODENTS: I: MECHANISMS OF OBESITY SURGERY – ROLE OF STOMACH. II: CARCINOGENIC EFFECTS OF *HELICOBACTER PYLORI* AND SNUS IN THE STOMACH

**2007**

298.Haakon R. Skogseth: INVASIVE PROPERTIES OF CANCER – A TREATMENT TARGET ? IN VITRO STUDIES IN HUMAN PROSTATE CANCER CELL LINES

299.Janniche Hammer: GLUTAMATE METABOLISM AND CYCLING IN MESIAL TEMPORAL LOBE EPILEPSY

300.May Britt Drugli: YOUNG CHILDREN TREATED BECAUSE OF ODD/CD: CONDUCT PROBLEMS AND SOCIAL COMPETENCIES IN DAY-CARE AND SCHOOL SETTINGS

301.Arne Skjold: MAGNETIC RESONANCE KINETICS OF MANGANESE DIPYRIDOXYL DIPHOSPHATE (MnDPDP) IN HUMAN MYOCARDIUM. STUDIES IN HEALTHY VOLUNTEERS AND IN PATIENTS WITH RECENT MYOCARDIAL INFARCTION

302.Siri Malm: LEFT VENTRICULAR SYSTOLIC FUNCTION AND MYOCARDIAL PERFUSION ASSESSED BY CONTRAST ECHOCARDIOGRAPHY

303.Valentina Maria do Rosario Cabral Iversen: MENTAL HEALTH AND PSYCHOLOGICAL ADAPTATION OF CLINICAL AND NON-CLINICAL MIGRANT GROUPS

304.Lasse Løvstakken: SIGNAL PROCESSING IN DIAGNOSTIC ULTRASOUND: ALGORITHMS FOR REAL-TIME ESTIMATION AND VISUALIZATION OF BLOOD FLOW VELOCITY

305.Elisabeth Olstad: GLUTAMATE AND GABA: MAJOR PLAYERS IN NEURONAL METABOLISM

306.Lilian Leistad: THE ROLE OF CYTOKINES AND PHOSPHOLIPASE $A_2$s IN ARTICULAR CARTILAGE CHONDROCYTES IN RHEUMATOID ARTHRITIS AND OSTEOARTHRITIS

307.Arne Vaaler: EFFECTS OF PSYCHIATRIC INTENSIVE CARE UNIT IN AN ACUTE PSYCIATHRIC WARD

308.Mathias Toft: GENETIC STUDIES OF LRRK2 AND PINK1 IN PARKINSON'S DISEASE

309.Ingrid Løvold Mostad: IMPACT OF DIETARY FAT QUANTITY AND QUALITY IN TYPE 2 DIABETES WITH EMPHASIS ON MARINE N-3 FATTY ACIDS

310.Torill Eidhammer Sjøbakk: MR DETERMINED BRAIN METABOLIC PATTERN IN PATIENTS WITH BRAIN METASTASES AND ADOLESCENTS WITH LOW BIRTH WEIGHT

311.Vidar Beisvåg: PHYSIOLOGICAL GENOMICS OF HEART FAILURE: FROM TECHNOLOGY TO PHYSIOLOGY

312.Olav Magnus Søndenå Fredheim: HEALTH RELATED QUALITY OF LIFE ASSESSMENT AND ASPECTS OF THE CLINICAL PHARMACOLOGY OF METHADONE IN PATIENTS WITH CHRONIC NON-MALIGNANT PAIN

313.Anne Brantberg: FETAL AND PERINATAL IMPLICATIONS OF ANOMALIES IN THE GASTROINTESTINAL TRACT AND THE ABDOMINAL WALL

314.Erik Solligård: GUT LUMINAL MICRODIALYSIS

315.Elin Tollefsen: RESPIRATORY SYMPTOMS IN A COMPREHENSIVE POPULATION BASED STUDY AMONG ADOLESCENTS 13-19 YEARS. YOUNG-HUNT 1995-97 AND 2000-01; THE NORD-TRØNDELAG HEALTH STUDIES (HUNT)

316.Anne-Tove Brenne: GROWTH REGULATION OF MYELOMA CELLS

317.Heidi Knobel: FATIGUE IN CANCER TREATMENT – ASSESSMENT, COURSE AND ETIOLOGY

318. Torbjørn Dahl: CAROTID ARTERY STENOSIS. DIAGNOSTIC AND THERAPEUTIC ASPECTS

319.Inge-Andre Rasmussen jr.: FUNCTIONAL AND DIFFUSION TENSOR MAGNETIC RESONANCE IMAGING IN NEUROSURGICAL PATIENTS

320.Grete Helen Bratberg: PUBERTAL TIMING – ANTECEDENT TO RISK OR RESILIENCE ? EPIDEMIOLOGICAL STUDIES ON GROWTH, MATURATION AND HEALTH RISK BEHAVIOURS; THE YOUNG HUNT STUDY, NORD-TRØNDELAG, NORWAY

321.Sveinung Sørhaug: THE PULMONARY NEUROENDOCRINE SYSTEM. PHYSIOLOGICAL, PATHOLOGICAL AND TUMOURIGENIC ASPECTS

322.Olav Sande Eftedal: ULTRASONIC DETECTION OF DECOMPRESSION INDUCED VASCULAR MICROBUBBLES

323.Rune Bang Leistad: PAIN, AUTONOMIC ACTIVATION AND MUSCULAR ACTIVITY RELATED TO EXPERIMENTALLY-INDUCED COGNITIVE STRESS IN HEADACHE PATIENTS

324. Svein Brekke: TECHNIQUES FOR ENHANCEMENT OF TEMPORAL RESOLUTION IN THREE-DIMENSIONAL ECHOCARDIOGRAPHY
325. Kristian Bernhard Nilsen: AUTONOMIC ACTIVATION AND MUSCLE ACTIVITY IN RELATION TO MUSCULOSKELETAL PAIN
326. Anne Irene Hagen: HEREDITARY BREAST CANCER IN NORWAY. DETECTION AND PROGNOSIS OF BREAST CANCER IN FAMILIES WITH *BRCA1*GENE MUTATION
327. Ingebjørg S. Juel : INTESTINAL INJURY AND RECOVERY AFTER ISCHEMIA. AN EXPERIMENTAL STUDY ON RESTITUTION OF THE SURFACE EPITHELIUM, INTESTINAL PERMEABILITY, AND RELEASE OF BIOMARKERS FROM THE MUCOSA
328. Runa Heimstad: POST-TERM PREGNANCY
329. Jan Egil Afset: ROLE OF ENTEROPATHOGENIC *ESCHERICHIA COLI* IN CHILDHOOD DIARRHOEA IN NORWAY
330. Bent Håvard Hellum: *IN VITRO* INTERACTIONS BETWEEN MEDICINAL DRUGS AND HERBS ON CYTOCHROME P-450 METABOLISM AND P-GLYCOPROTEIN TRANSPORT
331. Morten André Høydal: CARDIAC DYSFUNCTION AND MAXIMAL OXYGEN UPTAKE MYOCARDIAL ADAPTATION TO ENDURANCE TRAINING

**2008**

332. Andreas Møllerløkken: REDUCTION OF VASCULAR BUBBLES: METHODS TO PREVENT THE ADVERSE EFFECTS OF DECOMPRESSION
333. Anne Hege Aamodt: COMORBIDITY OF HEADACHE AND MIGRAINE IN THE NORD-TRØNDELAG HEALTH STUDY 1995-97
334. Brage Høyem Amundsen: MYOCARDIAL FUNCTION QUANTIFIED BY SPECKLE TRACKING AND TISSUE DOPPLER ECHOCARDIOGRAPHY – VALIDATION AND APPLICATION IN EXERCISE TESTING AND TRAINING
335. Inger Anne Næss: INCIDENCE, MORTALITY AND RISK FACTORS OF FIRST VENOUS THROMBOSIS IN A GENERAL POPULATION. RESULTS FROM THE SECOND NORD-TRØNDELAG HEALTH STUDY (HUNT2)
336. Vegard Bugten: EFFECTS OF POSTOPERATIVE MEASURES AFTER FUNCTIONAL ENDOSCOPIC SINUS SURGERY
337. Morten Bruvold: MANGANESE AND WATER IN CARDIAC MAGNETIC RESONANCE IMAGING
338. Miroslav Fris: THE EFFECT OF SINGLE AND REPEATED ULTRAVIOLET RADIATION ON THE ANTERIOR SEGMENT OF THE RABBIT EYE
339. Svein Arne Aase: METHODS FOR IMPROVING QUALITY AND EFFICIENCY IN QUANTITATIVE ECHOCARDIOGRAPHY – ASPECTS OF USING HIGH FRAME RATE
340. Roger Almvik: ASSESSING THE RISK OF VIOLENCE: DEVELOPMENT AND VALIDATION OF THE BRØSET VIOLENCE CHECKLIST
341. Ottar Sundheim: STRUCTURE-FUNCTION ANALYSIS OF HUMAN ENZYMES INITIATING NUCLEOBASE REPAIR IN DNA AND RNA
342. Anne Mari Undheim: SHORT AND LONG-TERM OUTCOME OF EMOTIONAL AND BEHAVIOURAL PROBLEMS IN YOUNG ADOLESCENTS WITH AND WITHOUT READING DIFFICULTIES
343. Helge Garåsen: THE TRONDHEIM MODEL. IMPROVING THE PROFESSIONAL COMMUNICATION BETWEEN THE VARIOUS LEVELS OF HEALTH CARE SERVICES AND IMPLEMENTATION OF INTERMEDIATE CARE AT A COMMUNITY HOSPITAL COULD PROVIDE BETTER CARE FOR OLDER PATIENTS. SHORT AND LONG TERM EFFECTS
344. Olav A. Foss: "THE ROTATION RATIOS METHOD". A METHOD TO DESCRIBE ALTERED SPATIAL ORIENTATION IN SEQUENTIAL RADIOGRAPHS FROM ONE PELVIS
345. Bjørn Olav Åsvold: THYROID FUNCTION AND CARDIOVASCULAR HEALTH
346. Torun Margareta Melø: NEURONAL GLIAL INTERACTIONS IN EPILEPSY
347. Irina Poliakova Eide: FETAL GROWTH RESTRICTION AND PRE-ECLAMPSIA: SOME CHARACTERISTICS OF FETO-MATERNAL INTERACTIONS IN DECIDUA BASALIS
348. Torunn Askim: RECOVERY AFTER STROKE. ASSESSMENT AND TREATMENT; WITH FOCUS ON MOTOR FUNCTION
349. Ann Elisabeth Åsberg: NEUTROPHIL ACTIVATION IN A ROLLER PUMP MODEL OF CARDIOPULMONARY BYPASS. INFLUENCE ON BIOMATERIAL, PLATELETS AND COMPLEMENT

350. Lars Hagen: REGULATION OF DNA BASE EXCISION REPAIR BY PROTEIN INTERACTIONS AND POST TRANSLATIONAL MODIFICATIONS
351. Sigrun Beate Kjøtrød: POLYCYSTIC OVARY SYNDROME – METFORMIN TREATMENT IN ASSISTED REPRODUCTION
352. Steven Keita Nishiyama: PERSPECTIVES ON LIMB-VASCULAR HETEROGENEITY: IMPLICATIONS FOR HUMAN AGING, SEX, AND EXERCISE
353. Sven Peter Näsholm: ULTRASOUND BEAMS FOR ENHANCED IMAGE QUALITY
354. Jon Ståle Ritland: PRIMARY OPEN-ANGLE GLAUCOMA & EXFOLIATIVE GLAUCOMA. SURVIVAL, COMORBIDITY AND GENETICS
355. Sigrid Botne Sando: ALZHEIMER'S DISEASE IN CENTRAL NORWAY. GENETIC AND EDUCATIONAL ASPECTS
356. Parvinder Kaur: CELLULAR AND MOLECULAR MECHANISMS BEHIND METHYLMERCURY-INDUCED NEUROTOXICITY
357. Ismail Cüneyt Güzey: DOPAMINE AND SEROTONIN RECEPTOR AND TRANSPORTER GENE POLYMORPHISMS AND EXTRAPYRAMIDAL SYMPTOMS. STUDIES IN PARKINSON'S DISEASE AND IN PATIENTS TREATED WITH ANTIPSYCHOTIC OR ANTIDEPRESSANT DRUGS
358. Brit Dybdahl: EXTRA-CELLULAR INDUCIBLE HEAT-SHOCK PROTEIN 70 (Hsp70) – A ROLE IN THE INFLAMMATORY RESPONSE ?
359. Kristoffer Haugarvoll: IDENTIFYING GENETIC CAUSES OF PARKINSON'S DISEASE IN NORWAY
360. Nadra Nilsen: TOLL-LIKE RECEPTOR 2 –EXPRESSION, REGULATION AND SIGNALING
361. Johan Håkon Bjørngaard: PATIENT SATISFACTION WITH OUTPATIENT MENTAL HEALTH SERVICES – THE INFLUENCE OF ORGANIZATIONAL FACTORS.
362. Kjetil Høydal : EFFECTS OF HIGH INTENSITY AEROBIC TRAINING IN HEALTHY SUBJECTS AND CORONARY ARTERY DISEASE PATIENTS; THE IMPORTANCE OF INTENSITY,, DURATION AND FREQUENCY OF TRAINING.
363. Trine Karlsen: TRAINING IS MEDICINE: ENDURANCE AND STRENGTH TRAINING IN CORONARY ARTERY DISEASE AND HEALTH.
364. Marte Thuen: MANGANASE-ENHANCED AND DIFFUSION TENSOR MR IMAGING OF THE NORMAL, INJURED AND REGENERATING RAT VISUAL PATHWAY
365. Cathrine Broberg Vågbø: DIRECT REPAIR OF ALKYLATION DAMAGE IN DNA AND RNA BY 2-OXOGLUTARATE- AND IRON-DEPENDENT DIOXYGENASES
366. Arnt Erik Tjønna: AEROBIC EXERCISE AND CARDIOVASCULAR RISK FACTORS IN OVERWEIGHT AND OBESE ADOLESCENTS AND ADULTS
367. Marianne W. Furnes: FEEDING BEHAVIOR AND BODY WEIGHT DEVELOPMENT: LESSONS FROM RATS
368. Lene N. Johannessen: FUNGAL PRODUCTS AND INFLAMMATORY RESPONSES IN HUMAN MONOCYTES AND EPITHELIAL CELLS
369. Anja Bye: GENE EXPRESSION PROFILING OF *INHERITED* AND *ACQUIRED* MAXIMAL OXYGEN UPTAKE – RELATIONS TO THE METABOLIC SYNDROME.
370. Oluf Dimitri Røe: MALIGNANT MESOTHELIOMA: VIRUS, BIOMARKERS AND GENES. A TRANSLATIONAL APPROACH
371. Ane Cecilie Dale: DIABETES MELLITUS AND FATAL ISCHEMIC HEART DISEASE. ANALYSES FROM THE HUNT1 AND 2 STUDIES
372. Jacob Christian Hølen: PAIN ASSESSMENT IN PALLIATIVE CARE: VALIDATION OF METHODS FOR SELF-REPORT AND BEHAVIOURAL ASSESSMENT
373. Erming Tian: THE GENETIC IMPACTS IN THE ONCOGENESIS OF MULTIPLE MYELOMA
374. Ole Bosnes: KLINISK UTPRØVING AV NORSKE VERSJONER AV NOEN SENTRALE TESTER PÅ KOGNITIV FUNKSJON
375. Ola M. Rygh: 3D ULTRASOUND BASED NEURONAVIGATION IN NEUROSURGERY. A CLINICAL EVALUATION
376. Astrid Kamilla Stunes: ADIPOKINES, PEROXISOME PROFILERATOR ACTIVATED RECEPTOR (PPAR) AGONISTS AND SEROTONIN. COMMON REGULATORS OF BONE AND FAT METABOLISM
377. Silje Engdal: HERBAL REMEDIES USED BY NORWEGIAN CANCER PATIENTS AND THEIR ROLE IN HERB-DRUG INTERACTIONS
378. Kristin Offerdal: IMPROVED ULTRASOUND IMAGING OF THE FETUS AND ITS CONSEQUENCES FOR SEVERE AND LESS SEVERE ANOMALIES

379.Øivind Rognmo: HIGH-INTENSITY AEROBIC EXERCISE AND CARDIOVASCULAR HEALTH
380. Jo-Åsmund Lund: RADIOTHERAPY IN ANAL CARCINOMA AND PROSTATE CANCER

**2009**

381.Tore Grüner Bjåstad: HIGH FRAME RATE ULTRASOUND IMAGING USING PARALLEL BEAMFORMING
382.Erik Søndenaa: INTELLECTUAL DISABILITIES IN THE CRIMINAL JUSTICE SYSTEM
383.Berit Rostad: SOCIAL INEQUALITIES IN WOMEN'S HEALTH, HUNT 1984-86 AND 1995-97, THE NORD-TRØNDELAG HEALTH STUDY (HUNT)
384.Jonas Crosby: ULTRASOUND-BASED QUANTIFICATION OF MYOCARDIAL DEFORMATION AND ROTATION
385.Erling Tronvik: MIGRAINE, BLOOD PRESSURE AND THE RENIN-ANGIOTENSIN SYSTEM
386.Tom Christensen: BRINGING THE GP TO THE FOREFRONT OF EPR DEVELOPMENT
387.Håkon Bergseng: ASPECTS OF GROUP B STREPTOCOCCUS (GBS) DISEASE IN THE NEWBORN. EPIDEMIOLOGY, CHARACTERISATION OF INVASIVE STRAINS AND EVALUATION OF INTRAPARTUM SCREENING
388.Ronny Myhre: GENETIC STUDIES OF CANDIDATE TENE3S IN PARKINSON'S DISEASE
389.Torbjørn Moe Eggebø: ULTRASOUND AND LABOUR
390.Eivind Wang: TRAINING IS MEDICINE FOR PATIENTS WITH PERIPHERAL ARTERIAL DISEASE
391.Thea Kristin Våtsveen: GENETIC ABERRATIONS IN MYELOMA CELLS
392.Thomas Jozefiak: QUALITY OF LIFE AND MENTAL HEALTH IN CHILDREN AND ADOLESCENTS: CHILD AND PARENT PERSPECTIVES
393.Jens Erik Slagsvold: N-3 POLYUNSATURATED FATTY ACIDS IN HEALTH AND DISEASE – CLINICAL AND MOLECULAR ASPECTS
394.Kristine Misund: A STUDY OF THE TRANSCRIPTIONAL REPRESSOR ICER. REGULATORY NETWORKS IN GASTRIN-INDUCED GENE EXPRESSION
395.Franco M. Impellizzeri: HIGH-INTENSITY TRAINING IN FOOTBALL PLAYERS. EFFECTS ON PHYSICAL AND TECHNICAL PERFORMANCE
396.Kari Hanne Gjeilo: HEALTH-RELATED QUALITY OF LIFE AND CHRONIC PAIN IN PATIENTS UNDERGOING CARDIAC SURGERY
397.Øyvind Hauso: NEUROENDOCRINE ASPECTS OF PHYSIOLOGY AND DISEASE
398.Ingvild Bjellmo Johnsen: INTRACELLULAR SIGNALING MECHANISMS IN THE INNATE IMMUNE RESPONSE TO VIRAL INFECTIONS
399.Linda Tømmerdal Roten: GENETIC PREDISPOSITION FOR DEVELOPMENT OF PREEMCLAMPSIA – CANDIDATE GENE STUDIES IN THE HUNT (NORD-TRØNDELAG HEALTH STUDY) POPULATION
400.Trude Teoline Nausthaug Rakvåg: PHARMACOGENETICS OF MORPHINE IN CANCER PAIN
401.Hanne Lehn: MEMORY FUNCTIONS OF THE HUMAN MEDIAL TEMPORAL LOBE STUDIED WITH fMRI
402.Randi Utne Holt: ADHESION AND MIGRATION OF MYELOMA CELLS – IN VITRO STUDIES –
403.Trygve Solstad: NEURAL REPRESENTATIONS OF EUCLIDEAN SPACE
404.Unn-Merete Fagerli: MULTIPLE MYELOMA CELLS AND CYTOKINES FROM THE BONE MARROW ENVIRONMENT; ASPECTS OF GROWTH REGULATION AND MIGRATION
405.Sigrid Bjørnelv: EATING– AND WEIGHT PROBLEMS IN ADOLESCENTS, THE YOUNG HUNT-STUDY
406.Mari Hoff: CORTICAL HAND BONE LOSS IN RHEUMATOID ARTHRITIS. EVALUATING DIGITAL X-RAY RADIOGRAMMETRY AS OUTCOME MEASURE OF DISEASE ACTIVITY, RESPONSE VARIABLE TO TREATMENT AND PREDICTOR OF BONE DAMAGE
407.Siri Bjørgen: AEROBIC HIGH INTENSITY INTERVAL TRAINING IS AN EFFECTIVE TREATMENT FOR PATIENTS WITH CHRONIC OBSTRUCTIVE PULMONARY DISEASE
408.Susanne Lindqvist: VISION AND BRAIN IN ADOLESCENTS WITH LOW BIRTH WEIGHT
409.Torbjørn Hergum: 3D ULTRASOUND FOR QUANTITATIVE ECHOCARDIOGRAPHY

410. Jørgen Urnes: PATIENT EDUCATION IN GASTRO-OESOPHAGEAL REFLUX DISEASE. VALIDATION OF A DIGESTIVE SYMPTOMS AND IMPACT QUESTIONNAIRE AND A RANDOMISED CONTROLLED TRIAL OF PATIENT EDUCATION
411. Elvar Eyjolfsson: 13C NMRS OF ANIMAL MODELS OF SCHIZOPHRENIA
412. Marius Steiro Fimland: CHRONIC AND ACUTE NEURAL ADAPTATIONS TO STRENGTH TRAINING
413. Øyvind Støren: RUNNING AND CYCLING ECONOMY IN ATHLETES; DETERMINING FACTORS, TRAINING INTERVENTIONS AND TESTING
414. Håkon Hov: HEPATOCYTE GROWTH FACTOR AND ITS RECEPTOR C-MET. AUTOCRINE GROWTH AND SIGNALING IN MULTIPLE MYELOMA CELLS
415. Maria Radtke: ROLE OF AUTOIMMUNITY AND OVERSTIMULATION FOR BETA-CELL DEFICIENCY. EPIDEMIOLOGICAL AND THERAPEUTIC PERSPECTIVES
416. Liv Bente Romundstad: ASSISTED FERTILIZATION IN NORWAY: SAFETY OF THE REPRODUCTIVE TECHNOLOGY
417. Erik Magnus Berntsen: PREOPERATIV PLANNING AND FUNCTIONAL NEURONAVIGATION – WITH FUNCTIONAL MRI AND DIFFUSION TENSOR TRACTOGRAPHY IN PATIENTS WITH BRAIN LESIONS
418. Tonje Strømmen Steigedal: MOLECULAR MECHANISMS OF THE PROLIFERATIVE RESPONSE TO THE HORMONE GASTRIN
419. Vidar Rao: EXTRACORPOREAL PHOTOCHEMOTHERAPY IN PATIENTS WITH CUTANEOUS T CELL LYMPHOMA OR GRAFT-vs-HOST DISEASE
420. Torkild Visnes: DNA EXCISION REPAIR OF URACIL AND 5-FLUOROURACIL IN HUMAN CANCER CELL LINES

**2010**

421. John Munkhaugen: BLOOD PRESSURE, BODY WEIGHT, AND KIDNEY FUNCTION IN THE NEAR-NORMAL RANGE: NORMALITY, RISK FACTOR OR MORBIDITY ?
422. Ingrid Castberg: PHARMACOKINETICS, DRUG INTERACTIONS AND ADHERENCE TO TREATMENT WITH ANTIPSYCHOTICS: STUDIES IN A NATURALISTIC SETTING
423. Jian Xu: BLOOD-OXYGEN-LEVEL-DEPENDENT-FUNCTIONAL MAGNETIC RESONANCE IMAGING AND DIFFUSION TENSOR IMAGING IN TRAUMATIC BRAIN INJURY RESEARCH
424. Sigmund Simonsen: ACCEPTABLE RISK AND THE REQUIREMENT OF PROPORTIONALITY IN EUROPEAN BIOMEDICAL RESEARCH LAW. WHAT DOES THE REQUIREMENT THAT BIOMEDICAL RESEARCH SHALL NOT INVOLVE RISKS AND BURDENS DISPROPORTIONATE TO ITS POTENTIAL BENEFITS MEAN?
425. Astrid Woodhouse: MOTOR CONTROL IN WHIPLASH AND CHRONIC NON-TRAUMATIC NECK PAIN
426. Line Rørstad Jensen: EVALUATION OF TREATMENT EFFECTS IN CANCER BY MR IMAGING AND SPECTROSCOPY
427. Trine Moholdt: AEROBIC EXERCISE IN CORONARY HEART DISEASE
428. Øystein Olsen: ANALYSIS OF MANGANESE ENHANCED MRI OF THE NORMAL AND INJURED RAT CENTRAL NERVOUS SYSTEM
429. Bjørn H. Grønberg: PEMETREXED IN THE TREATMENT OF ADVANCED LUNG CANCER
430. Vigdis Schnell Husby: REHABILITATION OF PATIENTS UNDERGOING TOTAL HIP ARTHROPLASTY WITH FOCUS ON MUSCLE STRENGTH, WALKING AND AEROBIC ENDURANCE PERFORMANCE
431. Torbjørn Øien: CHALLENGES IN PRIMARY PREVENTION OF ALLERGY. THE PREVENTION OF ALLERGY AMONG CHILDREN IN TRONDHEIM (PACT) STUDY.
432. Kari Anne Indredavik Evensen: BORN TOO SOON OR TOO SMALL: MOTOR PROBLEMS IN ADOLESCENCE
433. Lars Adde: PREDICTION OF CEREBRAL PALSY IN YOUNG INFANTS. COMPUTER BASED ASSESSMENT OF GENERAL MOVEMENTS
434. Magnus Fasting: PRE- AND POSTNATAL RISK FACTORS FOR CHILDHOOD ADIPOSITY
435. Vivi Talstad Monsen: MECHANISMS OF ALKYLATION DAMAGE REPAIR BY HUMAN AlkB HOMOLOGUES
436. Toril Skandsen: MODERATE AND SEVERE TRAUMATIC BRAIN INJURY. MAGNETIC RESONANCE IMAGING FINDINGS, COGNITION AND RISK FACTORS FOR DISABILITY

437. Ingeborg Smidesang: ALLERGY RELATED DISORDERS AMONG 2-YEAR OLDS AND ADOLESCENTS IN MID-NORWAY – PREVALENCE, SEVERITY AND IMPACT. THE PACT STUDY 2005, THE YOUNG HUNT STUDY 1995-97
438. Vidar Halsteinli: MEASURING EFFICIENCY IN MENTAL HEALTH SERVICE DELIVERY: A STUDY OF OUTPATIENT UNITS IN NORWAY
439. Karen Lehrmann Ægidius: THE PREVALENCE OF HEADACHE AND MIGRAINE IN RELATION TO SEX HORMONE STATUS IN WOMEN. THE HUNT 2 STUDY
440. Madelene Ericsson: EXERCISE TRAINING IN GENETIC MODELS OF HEART FAILURE
441. Marianne Klokk: THE ASSOCIATION BETWEEN SELF-REPORTED ECZEMA AND COMMON MENTAL DISORDERS IN THE GENERAL POPULATION. THE HORDALAND HEALTH STUDY (HUSK)
442. Tomas Ottemo Stølen: IMPAIRED CALCIUM HANDLING IN ANIMAL AND HUMAN CARDIOMYOCYTES REDUCE CONTRACTILITY AND INCREASE ARRHYTHMIA POTENTIAL – EFFECTS OF AEROBIC EXERCISE TRAINING
443. Bjarne Hansen: ENHANCING TREATMENT OUTCOME IN COGNITIVE BEHAVIOURAL THERAPY FOR OBSESSIVE COMPULSIVE DISORDER: THE IMPORTANCE OF COGNITIVE FACTORS
444. Mona Løvlien: WHEN EVERY MINUTE COUNTS. FROM SYMPTOMS TO ADMISSION FOR ACUTE MYOCARDIAL INFARCTION WITH SPECIAL EMPHASIS ON GENDER DIFFERECES
445. Karin Margaretha Gilljam: DNA REPAIR PROTEIN COMPLEXES, FUNCTIONALITY AND SIGNIFICANCE FOR REPAIR EFFICIENCY AND CELL SURVIVAL
446. Anne Byriel Walls: NEURONAL GLIAL INTERACTIONS IN CEREBRAL ENERGY – AND AMINO ACID HOMEOSTASIS – IMPLICATIONS OF GLUTAMATE AND GABA
447. Cathrine Fallang Knetter: MECHANISMS OF TOLL-LIKE RECEPTOR 9 ACTIVATION
448. Marit Følsvik Svindseth: A STUDY OF HUMILIATION, NARCISSISM AND TREATMENT OUTCOME IN PATIENTS ADMITTED TO PSYCHIATRIC EMERGENCY UNITS
449. Karin Elvenes Bakkelund: GASTRIC NEUROENDOCRINE CELLS – ROLE IN GASTRIC NEOPLASIA IN MAN AND RODENTS
450. Kirsten Brun Kjelstrup: DORSOVENTRAL DIFFERENCES IN THE SPATIAL REPRESENTATION AREAS OF THE RAT BRAIN
451. Roar Johansen: MR EVALUATION OF BREAST CANCER PATIENTS WITH POOR PROGNOSIS
452. Rigmor Myran: POST TRAUMATIC NECK PAIN. EPIDEMIOLOGICAL, NEURORADIOLOGICAL AND CLINICAL ASPECTS
453. Krisztina Kunszt Johansen: GENEALOGICAL, CLINICAL AND BIOCHEMICAL STUDIES IN *LRRK2* – ASSOCIATED PARKINSON'S DISEASE
454. Pål Gjerden: THE USE OF ANTICHOLINERGIC ANTIPARKINSON AGENTS IN NORWAY. EPIDEMIOLOGY, TOXICOLOGY AND CLINICAL IMPLICATIONS
455. Else Marie Huuse: ASSESSMENT OF TUMOR MICROENVIRONMENT AND TREATMENT EFFECTS IN HUMAN BREAST CANCER XENOGRAFTS USING MR IMAGING AND SPECTROSCOPY
456. Khalid S. Ibrahim: INTRAOPERATIVE ULTRASOUND ASSESSMENT IN CORONARY ARTERY BYPASS SURGERY – WITH SPECIAL REFERENCE TO CORONARY ANASTOMOSES AND THE ASCENDING AORTA
457. Bjørn Øglænd: ANTHROPOMETRY, BLOOD PRESSURE AND REPRODUCTIVE DEVELOPMENT IN ADOLESCENCE OF OFFSPRING OF MOTHERS WHO HAD PREECLAMPSIA IN PREGNANCY
458. John Olav Roaldset: RISK ASSESSMENT OF VIOLENT, SUICIDAL AND SELF-INJURIOUS BEHAVIOUR IN ACUTE PSYCHIATRY – A BIO-PSYCHO-SOCIAL APPROACH
459. Håvard Dalen: ECHOCARDIOGRAPHIC INDICES OF CARDIAC FUNCTION – NORMAL VALUES AND ASSOCIATIONS WITH CARDIAC RISK FACTORS IN A POPULATION FREE FROM CARDIOVASCULAR DISEASE, HYPERTENSION AND DIABETES: THE HUNT 3 STUDY
460. Beate André: CHANGE CAN BE CHALLENGING. INTRODUCTION TO CHANGES AND IMPLEMENTATION OF COMPUTERIZED TECHNOLOGY IN HEALTH CARE
461. Latha Nrugham: ASSOCIATES AND PREDICTORS OF ATTEMPTED SUICIDE AMONG DEPRESSED ADOLESCENTS – A 6-YEAR PROSPECTIVE STUDY

462.Håvard Bersås Nordgaard: TRANSIT-TIME FLOWMETRY AND WALL SHEAR STRESS ANALYSIS OF CORONARY ARTERY BYPASS GRAFTS – A CLINICAL AND EXPERIMENTAL STUDY

Cotutelle with University of Ghent: Abigail Emily Swillens: A MULTIPHYSICS MODEL FOR IMPROVING THE ULTRASONIC ASSESSMENT OF LARGE ARTERIES

**2011**

463. Marte Helene Bjørk: DO BRAIN RHYTHMS CHANGE BEFORE THE MIGRAINE ATTACK? A LONGITUDINAL CONTROLLED EEG STUDY

464. Carl-Jørgen Arum: A STUDY OF UROTHELIAL CARCINOMA: GENE EXPRESSION PROFILING, TUMORIGENESIS AND THERAPIES IN ORTHOTOPIC ANIMAL MODELS

465. Ingunn Harstad: TUBERCULOSIS INFECTION AND DISEASE AMONG ASYLUM SEEKERS IN NORWAY. SCREENING AND FOLLOW-UP IN PUBLIC HEALTH CARE

466. Leif Åge Strand: EPIDEMIOLOGICAL STUDIES AMONG ROYAL NORWEGIAN NAVY SERVICEMEN. COHORT ESTABLISHMENT, CANCER INCIDENCE AND CAUSE-SPECIFIC MORTALITY

467. Katrine Høyer Holgersen: SURVIVORS IN THEIR THIRD DECADE AFTER THE NORTH SEA OIL RIG DISASTER OF 1980. LONG-TERM PERSPECTIVES ON MENTAL HEALTH

468. MarianneWallenius: PREGNANCY RELATED ASPECTS OF CHRONIC INFLAMMATORY ARTHRITIDES: DISEASE ONSET POSTPARTUM, PREGNANCY OUTCOMES AND FERTILITY. DATA FROM A NORWEGIAN PATIENT REGISTRY LINKED TO THE MEDICAL BIRTH REGISTRY OF NORWAY

469. Ole Vegard Solberg: 3D ULTRASOUND AND NAVIGATION – APPLICATIONS IN LAPAROSCOPIC SURGERY

470. Inga Ekeberg Schjerve: EXERCISE-INDUCED IMPROVEMENT OF MAXIMAL OXYGEN UPTAKE AND ENDOTHELIAL FUNCTION IN OBESE AND OVERWEIGHT INDIVIDUALS ARE DEPENDENT ON EXERCISE-INTENSITY

471. Eva Veslemøy Tyldum: CARDIOVASCULAR FUNCTION IN PREECLAMPSIA – WITH REFERENCE TO ENDOTHELIAL FUNCTION, LEFT VENTRICULAR FUNCTION AND PRE-PREGNANCY PHYSICAL ACTIVITY

472. Benjamin Garzón Jiménez de Cisneros: CLINICAL APPLICATIONS OF MULTIMODAL MAGNETIC RESONANCE IMAGING

473. Halvard Knut Nilsen: ASSESSING CODEINE TREATMENT TO PATIENTS WITH CHRONIC NON-MALIGNANT PAIN: NEUROPSYCHOLOGICAL FUNCTIONING, DRIVING ABILITY AND WEANING

474. Eiliv Brenner: GLUTAMATE RELATED METABOLISM IN ANIMAL MODELS OF SCHIZOPHRENIA

475. Egil Jonsbu: CHEST PAIN AND PALPITATIONS IN A CARDIAC SETTING; PSYCHOLOGICAL FACTORS, OUTCOME AND TREATMENT

476. Mona Høysæter Fenstad: GENETIC SUSCEPTIBILITY TO PREECLAMPSIA : STUDIES ON THE NORD-TRØNDELAG HEALTH STUDY (HUNT) COHORT, AN AUSTRALIAN/NEW ZEALAND FAMILY COHORT AND DECIDUA BASALIS TISSUE

477. Svein Erik Gaustad: CARDIOVASCULAR CHANGES IN DIVING: FROM HUMAN RESPONSE TO CELL FUNCTION

478. Karin Torvik: PAIN AND QUALITY OF LIFE IN PATIENTS LIVING IN NURSING HOMES

479. Arne Solberg: OUTCOME ASSESSMENTS IN NON-METASTATIC PROSTATE CANCER

480. Henrik Sahlin Pettersen: CYTOTOXICITY AND REPAIR OF URACIL AND 5-FLUOROURACIL IN DNA

481. Pui-Lam Wong: PHYSICAL AND PHYSIOLOGICAL CAPACITY OF SOCCER PLAYERS: EFFECTS OF STRENGTH AND CONDITIONING

482. Ole Solheim: ULTRASOUND GUIDED SURGERY IN PATIENTS WITH INTRACRANIAL TUMOURS

483. Sten Roar Snare: QUANTITATIVE CARDIAC ANALYSIS ALGORITHMS FOR POCKET-SIZED ULTRASOUND DEVICES

484. Marit Skyrud Bratlie: LARGE-SCALE ANALYSIS OF ORTHOLOGS AND PARALOGS IN VIRUSES AND PROKARYOTES

485.Anne Elisabeth F. Isern: BREAST RECONSTRUCTION AFTER MASTECTOMY – RISK OF RECURRENCE AFTER DELAYED LARGE FLAP RECONSTRUCTION – AESTHETIC OUTCOME, PATIENT SATISFACTION, QUALITY OF LIFE AND SURGICAL RESULTS;

HISTOPATHOLOGICAL FINDINGS AND FOLLOW-UP AFTER PROPHYLACTIC MASTECTOMY IN HEREDITARY BREAST CANCER

486. Guro L. Andersen: CEREBRAL PALSY IN NORWAY – SUBTYPES, SEVERITY AND RISK FACTORS

487. Frode Kolstad: CERVICAL DISC DISEASE – BIOMECHANICAL ASPECTS

488. Bente Nordtug: CARING BURDEN OF COHABITANTS LIVING WITH PARTNERS SUFFERING FROM CHRONIC OBSTRUCTIVE PULMONARY DISEASE OR DEMENTIA

489. Mariann Gjervik Heldahl: EVALUATION OF NEOADJUVANT CHEMOTHERAPY IN LOCALLY ADVANCED BREAST CANCER BASED ON MR METHODOLOGY

490. Lise Tevik Løvseth: THE SUBJECTIVE BURDEN OF CONFIDENTIALITY

491. Marie Hjelmseth Aune: INFLAMMATORY RESPONSES AGAINST GRAM NEGATIVE BACTERIA INDUCED BY TLR4 AND NLRP12

492. Tina Strømdal Wik: EXPERIMENTAL EVALUATION OF NEW CONCEPTS IN HIP ARTHROPLASTY

493. Solveig Sigurdardottir: CLINICAL ASPECTS OF CEREBRAL PALSY IN ICELAND. A POPULATION-BASED STUDY OF PRESCHOOL CHILDREN

494. Arne Reimers: CLINICAL PHARMACOKINETICS OF LAMOTRIGINE

495. Monica Wegling: KULTURMENNESKETS BYRDE OG SYKDOMMENS VELSIGNELSE. KAN MEDISINSK UTREDNING OG INTERVENSJON HA EN SELVSTENDIG FUNKSJON UAVHENGIG AV DET KURATIVE?

496. Silje Alvestad: ASTROCYTE-NEURON INTERACTIONS IN EXPERIMENTAL MESIAL TEMPORAL LOBE EPILEPSY – A STUDY OF UNDERLYING MECHANISMS AND POSSIBLE BIOMARKERS OF EPILEPTOGENESIS

497. Javaid Nauman: RESTING HEART RATE: A MATTER OF LIFE OR DEATH – PROSPECTIVE STUDIES OF RESTING HEART RATE AND CARDIOVASCULAR RISK (THE HUNT STUDY, NORWAY)

498. Thuy Nguyen: THE ROLE OF C-SRC TYROSINE KINASE IN ANTIVIRAL IMMUNE RESPONSES

499. Trine Naalsund Andreassen: PHARMACOKINETIC, PHARMACODYNAMIC AND PHARMACOGENETIC ASPECTS OF OXYCODONE TREATMENT IN CANCER PAIN

500. Eivor Alette Laugsand: SYMPTOMS IN PATIENTS RECEIVING OPIOIDS FOR CANCER PAIN – CLINICAL AND PHARMACOGENETIC ASPECTS

501. Dorthe Stensvold: PHYSICAL ACTIVITY, CARDIOVASCULAR HEALTH AND LONGEVITY IN PATIENTS WITH METABOLIC SYNDROME

502. Stian Thoresen Aspenes: PEAK OXYGEN UPTAKE AMONG HEALTHY ADULTS – CROSS-SECTIONAL DESCRIPTIONS AND PROSPECTIVE ANALYSES OF PEAK OXYGEN UPTAKE, PHYSICAL ACTIVITY AND CARDIOVASCULAR RISK FACTORS IN HEALTHY ADULTS (20-90 YEARS)

503. Reidar Alexander Vigen: PATHOBIOLOGY OF GASTRIC CARCINOIDS AND ADENOCARCINOMAS IN RODENT MODELS AND PATIENTS. STUDIES OF GASTROCYSTOPLASTY, GENDER-RELATED FACTORS, AND AUTOPHAGY

504. Halvard Høilund-Kaupang: MODELS AND METHODS FOR INVESTIGATION OF REVERBERATIONS IN NONLINEAR ULTRASOUND IMAGING

505. Audhild Løhre: WELLBEING AMONG SCHOOL CHILDREN IN GRADES 1-10: PROMOTING AND ADVERSE FACTORS

506. Torgrim Tandstad: VOX POPULI. POPULATION-BASED OUTCOME STUDIES IN TESTICULAR CANCER

507. Anna Brenne Grønskag: THE EPIDEMIOLOGY OF HIP FRACTURES AMONG ELDERLY WOMEN IN NORD-TRØNDELAG. HUNT 1995-97, THE NORD-TRØNDELAG HEALTH STUDY

508. Kari Ravndal Risnes: BIRTH SIZE AND ADULT MORTALITY: A SYSTEMATIC REVIEW AND A LONG-TERM FOLLOW-UP OF NEARLY 40 000 INDIVIDUALS BORN AT ST. OLAV UNIVERSITY HOSPITAL IN TRONDHEIM 1920-1960

509. Hans Jakob Bøe: LONG-TERM POSTTRAUMATIC STRESS AFTER DISASTER – A CONTROLLED STUDY OF SURVIVORS' HEALTH 27 YEARS AFTER THE CAPSIZED NORTH SEA OIL RIG

510. Cathrin Barbara Canto, Cotutelle with University of Amsterdam: LAYER SPECIFIC INTEGRATIVE PROPERTIES OF ENTORHINAL PRINCIPAL NEURONS

511. Ioanna Sandvig: THE ROLE OF OLFACTORY ENSHEATHING CELLS, MRI, AND BIOMATERIALS IN TRANSPLANT-MEDIATED CNS REPAIR

512. Karin Fahl Wader: HEPATOCYTE GROWTH FACTOR, C-MET AND SYNDECAN-1 IN MULTIPLE MYELOMA
513. Gerd Tranø: FAMILIAL COLORECTAL CANCER
514. Bjarte Bergstrøm: INNATE ANTIVIRAL IMMUNITY – MECHANISMS OF THE RIG-I-MEDIATED RESPONSE
515. Marie Søfteland Sandvei: INCIDENCE, MORTALITY, AND RISK FACTORS FOR ANEURYSMAL SUBARACHNOID HEMORRHAGE. PROSPECTIVE ANALYZES OF THE HUNT AND TROMSØ STUDIES
516. Mary-Elizabeth Bradley Eilertsen: CHILDREN AND ADOLESCENTS SURVIVING CANCER: PSYCHOSOCIAL HEALTH, QUALITY OF LIFE AND SOCIAL SUPPORT
517. Takaya Saito: COMPUTATIONAL ANALYSIS OF REGULATORY MECHANISM AND INTERACTIONS OF MICRORNAS

Godkjent for disputas, publisert post mortem: Eivind Jullumstrø: COLORECTAL CANCER AT LEVANGER HOSPITAL 1980-2004

518. Christian Gutvik: A PHYSIOLOGICAL APPROACH TO A NEW DECOMPRESSION ALGORITHM USING NONLINEAR MODEL PREDICTIVE CONTROL
519. Ola Storrø: MODIFICATION OF ADJUVANT RISK FACTOR BEHAVIOURS FOR ALLERGIC DISEASE AND ASSOCIATION BETWEEN EARLY GUT MICROBIOTA AND ATOPIC SENSITIZATION AND ECZEMA. EARLY LIFE EVENTS DEFINING THE FUTURE HEALTH OF OUR CHILDREN
520. Guro Fanneløb Giskeødegård: IDENTIFICATION AND CHARACTERIZATION OF PROGNOSTIC FACTORS IN BREAST CANCER USING MR METABOLOMICS
521. Gro Christine Christensen Løhaugen: BORN PRETERM WITH VERY LOW BIRTH WEIGHT – NEVER ENDING COGNITIVE CONSEQUENCES?
522. Sigrid Nakrem: MEASURING QUALITY OF CARE IN NURSING HOMES – WHAT MATTERS?
523. Brita Pukstad: CHARACTERIZATION OF INNATE INFLAMMATORY RESPONSES IN ACUTE AND CHRONIC WOUNDS

**2012**

524. Hans H. Wasmuth: ILEAL POUCHES
525. Inger Økland: BIASES IN SECOND-TRIMESTER ULTRASOUND DATING RELATED TO PREDICTION MODELS AND FETAL MEASUREMENTS
526. Bjørn Mørkedal: BLOOD PRESSURE, OBESITY, SERUM IRON AND LIPIDS AS RISK FACTORS OF ISCHAEMIC HEART DISEASE
527. Siver Andreas Moestue: MOLECULAR AND FUNCTIONAL CHARACTERIZATION OF BREAST CANCER THROUGH A COMBINATION OF MR IMAGING, TRANSCRIPTOMICS AND METABOLOMICS
528. Guro Aune: CLINICAL, PATHOLOGICAL, AND MOLECULAR CLASSIFICATION OF OVARIAN CARCINOMA
529. Ingrid Alsos Lian: MECHANISMS INVOLVED IN THE PATHOGENESIS OF PRE-ECLAMPSIA AND FETAL GROWTH RESTRICTION. TRANSCRIPTIONAL ANALYSES OF PLACENTAL AND DECIDUAL TISSUE
530. Karin Solvang-Garten: X-RAY REPAIR CROSS-COMPLEMENTING PROTEIN 1 – THE ROLE AS A SCAFFOLD PROTEIN IN BASE EXCISION REPAIR AND SINGLE STRAND BREAK REPAIR
531. Toril Holien: BONE MORPHOGENETIC PROTEINS AND MYC IN MULTIPLE MYELOMA
532. Rooyen Mavenyengwa: *STREPTOCOCCUS AGALACTIAE* IN PREGNANT WOMEN IN ZIMBABWE: EPIDEMIOLOGY AND SEROTYPE MARKER CHARACTERISTICS
533. Tormod Rimehaug: EMOTIONAL DISTRESS AND PARENTING AMONG COMMUNITY AND CLINIC PARENTS
534. Maria Dung Cao: MR METABOLIC CHARACTERIZATION OF LOCALLY ADVANCED BREAST CANCER – TREATMENT EFFECTS AND PROGNOSIS
535. Mirta Mittelstedt Leal de Sousa: PROTEOMICS ANALYSIS OF PROTEINS INVOLVED IN DNA BASE REPAIR AND CANCER THERAPY
536. Halfdan Petursson: THE VALIDITY AND RELEVANCE OF INTERNATIONAL CARDIOVASCULAR DISEASE PREVENTION GUIDELINES FOR GENERAL PRACTICE
537. Marit By Rise: LIFTING THE VEIL FROM USER PARTICIPATION IN CLINICAL WORK – WHAT IS IT AND DOES IT WORK?

538. Lene Thoresen: NUTRITION CARE IN CANCER PATIENTS. NUTRITION ASSESSMENT: DIAGNOSTIC CRITERIA AND THE ASSOCIATION TO SURVIVAL AND HEALTH-RELATED QUALITY OF LIFE IN PATIENTS WITH ADVANCED COLORECTAL CARCINOMA
539. Berit Doseth: PROCESSING OF GENOMIC URACIL IN MAN AND MOUSE
540. Gro Falkenér Bertheussen: PHYSICAL ACTIVITY AND HEALTH IN A GENERAL POPULATION AND IN CANCER SURVIVORS – METHODOLOGICAL, OBSERVATIONAL AND CLINICAL ASPECTS
541. Anne Kari Knudsen: CANCER PAIN CLASSIFICATION
542. Sjur Urdson Gjerald: A FAST ULTRASOUND SIMULATOR
543. Harald Edvard Mølmen Hansen: CARDIOVASCULAR EFFECTS OF HIGH INTENSITY AEROBIC INTERVAL TRAINING IN HYPERTENSITIVE PATIENTS, HEALTHY AGED AND YOUNG PERSONS
544. Sasha Gulati: SURGICAL RESECTION OF HIGH-GRADE GLIOMAS
545. John Chr. Fløvig: FREQUENCY AND EFFECT OF SUBSTANCES AND PSYCHOACTIVE MEDICATIONS THE WEEK BEFORE ADMISSION TO AN ACUTE PSYCHIATRIC DEPARTMENT
546. Kristin Moksnes Husby: OPTIMIZING OPIOID TREATMENT FOR CANCER PAIN – CLINICAL AND PHARMACOLOGICAL ASPECTS
547. Audun Hanssen-Bauer: X-RAY REPAIR CROSS-COMPLEMENTING PROTEIN 1 ASSOCIATED MULTIPROTEIN COMPLEXES IN BASE EXCISION REPAIR
548. Marit Saunes: ECZEMA IN CHILDREN AND ADOLESCENTS – EPIDEMIOLOGY, COURSE AND IMPACT. THE PREVENTION OF ALLERGY AMONG CHILDREN IN TRONDHEIM (PACT) STUDY, YOUNG-HUNT 1995-97
549. Guri Kaurstad: CARDIOMYOCYTE FUNCTION AND CALCIUM HANDLING IN ANIMAL MODELS OF INBORN AND ACQUIRED MAXIMAL OXYGEN UPTAKE
550. Kristian Svendsen: METHODOLOGICAL CHALLENGES IN PHARMACOEPIDEMIOLOGICAL STUDIES OF OPIOID CONSUMPTION
551. Signe Nilssen Stafne: EXERCISE DURING PREGNANCY
552. Marius Widerøe: MAGNETIC RESONANCE IMAGING OF HYPOXIC-ISCHEMIC BRAIN INJURY DEVELOPMENT IN THE NEWBORN RAT – MANGANESE AND DIFFUSION CONTRASTS
553. Andreas Radtke: MOLECULAR METHODS FOR TYPING *STREPTOCOCCUS AGALACTIAE* WITH SPECIAL EMPHASIS ON THE DEVELOPMENT AND VALIDATION OF A MULTI-LOCUS VARIABLE NUMBER OF TANDEM REPEATS ASSAY (MLVA)
554. Thor Wilhelm Bjelland: PHARMACOLOGICAL ASPECTS OF THERAPEUTIC HYPOTHERMIA
555. Caroline Hild Hakvåg Pettersen: THE EFFECT OF OMEGA-3 POLYUNSATURATED FATTY ACIDS ON HUMAN CANCER CELLS – MOLECULAR MECHANISMS INVOLVED
556. Inga Thorsen Vengen: INFLAMMATION AND ATHEROSCLEROSIS – RISK ASSOCIATIONS IN THE HUNT SURVEYS
557. Elisabeth Balstad Magnussen: PREECLAMPSIA, PRETERM BIRTH AND MATERNAL CARDIOVASCULAR RISK FACTORS
558. Monica Unsgaard-Tøndel: MOTOR CONTROL EXERCISES FOR PATIENTS WITH LOW BACK PAIN
559. Lars Erik Sande Laugsand: INSOMNIA AND RISK FOR CARDIOVASCULAR DISEASE
560. Kjersti Grønning: PATIENT EDUCATION AND CHRONIC INFLAMMATORY POLYARTHRITIS – COPING AND EFFECT
561. Hanne Gro Wenzel: PRE AND POST-INJURY HEALTH IN PERSONS WITH WHIPLASH: THE HUNT STUDY. EXPLORATION OF THE FUNCTIONAL SOMATIC MODEL FOR CHRONIC WHIPLASH
562. Øystein Grimstad: TOLL-LIKE RECEPTOR-MEDIATED INFLAMMATORY RESPONSES IN KERATINOCYTES
563. Håkon Olav Leira: DEVELOPMENT OF AN IMAGE GUIDANCE RESEARCH SYSTEM FOR BRONCHOSCOPY
564. Michael A. Lang: DIVING IN EXTREME ENVIRONMENTS: THE SCIENTIFIC DIVING EXPERIENCE

565. Helena Bertilsson: PROSTATE CANCER-TRANSLATIONAL RESEARCH. OPTIMIZING TISSUE SAMPLING SUITABLE FOR HISTOPATHOLOGIC, TRANSCRIPTOMIC AND METABOLIC PROFILING

566. Kirsten M. Selnæs: MR IMAGING AND SPECTROSCOPY IN PROSTATE AND COLON CANCER DIAGNOSTICS

567. Gunvor Steine Fosnes: CONSTIPATION AND DIARRHOEA. EFFECTIVENESS AND ADVERSE EFFECTS OF DRUGS

568. Areej Elkamil: SPASTIC CEREBRAL PALSY: RISK FACTORS, BOTULINUM TOXIN USE AND PREVENTION OF HIP DISLOCATION

569. Ruth Derdikman Eiron: SYMPTOMS OF ANXIETY AND DEPRESSION AND PSYCHOSOCIAL FUNCTION IN MALES AND FEMALES FROM ADOLESCENCE TO ADULTHOOD: LONGITUDINAL FINDINGS FROM THE NORD-TRØNDELAG HEALTH STUDY

570. Constantin Sergiu Jianu: PROTON PUMP INHIBITORS AND GASTRIC NEOPLASIA IN MAN

571. Øystein Finset Sørdal: THE ROLE OF GASTRIN AND THE ECL CELL IN GASTRIC CARCINOGENESIS

572. Lisbeth Østgaard Rygg: GROUP EDUCATION FOR PATIENTS WITH TYPE 2 DIABETES – NEEDS, EXPERIENCES AND EFFECTS

573. Viola Lobert: IDENTIFICATION OF NOVEL REGULATORS OF EPITHELIAL POLARITY AND CELL MIGRATION

574. Maria Tunset Grinde: CHARACTERIZATION OF BREAST CANCER USING MR METABOLOMICS AND GENE EXPRESSION ANALYSIS

575.Grete Kjelvik: HUMAN ODOR IDENTIFICATION STUDIES IN HEALTHY INDIVIDUALS, MILD COGNITIVE IMPAIRMENT AND ALZHEIMER’S DISEASE

576.Tor Eivind Bernstein: RECTAL CANCER SURGERY. PROGNOSTIC FACTORS RELATED TO TREATMENT

577. Kari Sand: INFORMED CONSENT DOCUMENTS FOR CANCER RESEARCH: TEXTUAL AND CONTEXTUAL FACTORS OF RELEVANCE FOR UNDERSTANDING

578. Laurent Francois Thomas: EFFECTS OF SINGLE-NUCLEOTIDE POLYMORPHISMS ON microRNA-BASED GENE REGULATION AND THEIR ASSOCIATION WITH DISEASE