Direct Assessment and Validation of Allele Specific Transcription Factor Binding in the Human Genome

Alicia Schiavi

Department of Human Genetics

McGill University

December 2012

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Masters of Science

Table of Contents

ABSTRACT

Characterization of human genetic variation has focused on expression quantitative trait loci (eQTL) mapping; however, direct assessment of *cis*-regulatory variation requires allele-specific approaches. Measuring allelic expression (AE) on a genome-wide scale appears more powerful as environmental and trans-acting influences are minimized. Results indicate that allele-specific differences in transcript expression within an individual can affect up to 30% of loci. The underlying variants can be identified by mapping differences in AE on Illumina BeadChips. Over 50% of population variance in AE is explained by mapped *cis*-rSNPs. Studies show that these *cis*-rSNPs have been implicated in differences in transcription factor (TF) binding, suggesting that TF action can be further investigated using population variation as a tool. In this thesis, these approaches have been extended to explore allele-specific TF binding using the model NF-κB by monitoring the consequences of gene knockdown in a genome-wide manner. NF-κB has been shown to be involved in the immune response and the NF-κB motif is enriched in lymphoblastoid cell lines (LCLs), mainly in promoters and strong enhancer elements. We intersected mapped candidate *cis*-rSNPs detected in LCLs in our above experiments as well as matched control SNPs from HapMap YRI and CEU populations with publicly available NF-κB Chromatin Immunoprecipitation (ChIP)-seq experiments from the ENCODE project. Preliminary analysis of regions surrounding candidate *cis*-rSNPs were enriched in NF-κB binding sites versus matched controls, with 39.0 % of top SNPs overlapping at least one NF-κB ChIP-seq peak. To elucidate the impact of candidate SNPs on AE imbalances, we performed TNF- α induction coupled to inhibition of NF-κB in LCLs followed by AE analysis on Illumina HumanOmni5-Quad

BeadChips. We used in house mapped *cis*-regulatory variants in the LCL population merged with data from the aforementioned experiment. Our data set, which consisted of loci associated to top 10 *cis*-rSNPs ranked by p-value (pv; top10= 10 most significant p-values) that showed diminished AE upon perturbation of NF-κB were overlapped with publicly available data. This data consisted of ENCODE ChIP-seq peaks and TRANSFAC binding sites for NF-κB and known cooperative TFs of NF-κB. Loci that had an AE change at greater than 3 SNPs upon perturbation of NF-κB and were associated to top heterozygous SNPs (rank 1, 2, 3) yielded 581 cases out of ~1700. Analysis of top 3 *cis*-r*SNPs* described showed a significant difference of over 5-fold between case and control SNPs, such that 64% of loci had a top 3 heterozygous SNP that was found in an LCL specific TF ChIP-seq peak or TRANSFAC binding site for NF-κB or a known cooperative TF of NF-κB. Bioinformatics analysis suggests that identified SNPs are essential for NF-κB binding. A case study was also done in order to perturb the TF SNAI1 because we had a strong hypothesis for the association of SNAI1 and *WNT4*, as well as, evidence for its role in fibroblasts (FBs). We were not able to reproduce the effect of SNAI1 on *WNT4 in vivo.* Upon comparison of the regulatory role of NF-κB and SNAI1 in LCLs and FBs, respectively; we observed that NF-κB had a regulatory effect on approximately 33% of loci in comparison to only approximately 2% of loci for SNAI1 in FBs. This study illustrates that key regulatory TFs, such as NF-κB in LCLs, can be globally studied at a single base resolution in living cells using a combination of perturbation and sensitive measurements with allelic resolution.

RÉSUMÉ

La caractérisation de la variation génétique a mis l'accent sur les loci d'expression de caractères quantitatifs (eLCQ); cependant, l'évaluation directe des variations régulatrices en cis nécessite des approches allèle-spécifique (*cis*-rSNPs). La mesure de l'expression allélique (EA) à l'échelle du génome est très efficace puisque les perturbations environnementales et les influences en trans sont réduites. Les résultats indiquent que les différences d'EA peuvent affecter jusqu'à 30% des loci chez un même individu. Les polymorphismes responsable de telles variations peuvent être identifiés par cartographie des différences d'EA en utilisant des puces de génotypage Illumina. Ainsi, plus de 50% de la variance en EA de la population est expliqué par la cartographie des *cis*-rSNPs. Des études ont montré que ces *cis*-rSNPs ont été impliqués dans des différences de liason de facteurs de transcription (FT). Celà suggère que l'étude du mode d'action de ces FT pourrait être approfondie par l'utilisation comme outil des polymorphismes présent dans la population. Dans cette thèse, nous avons appliqué cette approche au FT NF-κB et analysé les conséquences de l'inactivation de ce gène à l'échelle du genome. Des données récentes montrent l'implication de NF-κB dans la réponse immunitaire et son motif de liaison à l'ADN est retrouvé enrichi dans les cellules lymphoblastoïdes humaines (LCL), principalement au niveau des promoteurs et des activateurs transcriptionnels. Nous avons croisé les *cis*-rSNPs cartographiés dans des LCLs des populations HapMap YRI et CEU, ainsi que des SNPs de contrôles, avec des données d'immunoprécipitation de chromatine suivi de séquençage à haut débit (ChIP-seq) utilisant l'anticorps NF-κB du projet ENCODE et accessible au public. Des analyses préliminaires des régions contenant les *cis*-rSNPs candidats ont montré un enrichissement des sites de liaison pour le facteur NF-

κB par rapport aux sites contrôles. En effet, 39% des sites candidats sont situés dans un site de liaison pour NF-kB. Afin d'étudier le rôle potentiel des *cis*-rSNPs sur l'EA différentielle, nous avons réalisé des expériences d'induction de TNF-α couplé à l'inhibition de NF-κB dans les LCLs suivie par l'analyses de l'EA sur des puces de génotypage Illumina 5M. Nous avons ensuite comparé ces données avec la cartographie de cis-rSNPs dans des LCLs générée dans notre laboratoire. Nos données sont composées de *cis*-rSNPs associés à l'EA différentielle de loci suite à la perturbation de NF-κB et classés par valeur p (pv; top10= les 10 valeurs les plus significatives) et ont été croisées avec des données accessibles au public. Ces données sont composées des coordonnées de pics de ChIP-seq et des sites de liaison TRANSFAC pour le facteur NF-κB et de ses co-régulateurs transcriptionnels connus. Les loci montrant des différences d'EA suite à la perturbation de NF-κB et avec 1 ou plusieurs des *cis*-rSNPs (« top3 » pv) hétérozygotes dans les individus étudiés étaient aux nombre de 581. La recherche de ces *cis*-rSNPs classés « top 3 » dans les sites de liaison (pics de ChIP-seq) dans les LCLs ou dans un site de liaison TRANSFAC pour NF-κB ou pour un de ses co-régulateurs transcriptionnels montrent un enrichissement significatif (64% des loci) avec un ratio supérieur à 5 par rapport aux SNPs de contrôles. L'analyse bioinformatique suggère que les SNPs identifiés sont essentiels pour la liason de NF-κB. Une étude complémentaire à consisté à perturber le FT SNAI1 probablement associé au gène WNT4 et présentant un rôle important dans les fibroblastes (FB) selon des données précédemment obtenus au laboratoire. Cependant, nous n'avons pas été en mesure de reproduire l'effet de SNAI1 sur WNT4 in vivo. Nous avons ensuite comparé les rôles régulateurs de NF-κB et de SNAI1 dans les LCLs et les FBs respectivement (par inactivation de ces gènes). Nous avons

observé que NF-κB a un effet régulateurs sur environ 33% des loci dans les LCLs contre seulement 2 % pour SNAI1 dans des FBs. Cette étude supporte l'utilisation de perturbations de l'expression de FT, tels que NF-κB dans LCLs, associé à un contrôle des différences d'EA, pour étudier le rôle clé de FTs régulateurs dans un type cellulaire donné.

ABBREVIATIONS

A large-scale process to hunt for allele binding interacting transcription factors: ALPHABIT
Allele Binding Cooperativity: ABC
Allelic expression: AE
Allelic imbalance: AI
Base pair: bp
Binding region: BR
Bone mineral density: BMD
Caucasian population: CEU
Cis-regulatory element: CRE
Cis-regulatory single nucleotide polymorphism: *cis*-rSNP
Chromatin Immunoprecipitation: ChIP
Complementary DNA: cDNA
Cycle threshold: ct
Data Coordination Center: DCC
DNaseI hypersensitivity sites: DHS
Double stranded complementary DNA: dscDNA
Encyclopedia of DNA Elements: ENCODE
Expression quantitative trait loci: eQTL
Fibroblast: FB
Formaldehyde-assisted isolation of regulatory elements: FAIRE
DiGeorge Syndrome chromosome region 8: *DGCR8*
Genomic DNA: gDNA
Genome-wide Association Studies: GWAS
Hidden Markov Model: HMM
Human Genome Project: HGP
Ingenuity Pathway Analysis: IPA
Immunoglobulin G: IgG
Lymphoblastoid cell line: LCL
Messenger RNA: mRNA
Minor allele frequency: MAF
National Human Genome Research Institute: NHGRI
Next generation sequencing: NGS
Phenylmethylsulfonyl fluoride: PMSF
Phosphate buffered saline: PBS
Position weight matrix: PWM
P-value: pv
Radioimmunoprecipitation assay buffer: RIPA
Real-time Polymerase Chain Reaction: RT-PCR
Reticuloendotheliosis: Rel
RNA-interference: RNA*i*
RNA-induced silencing complex: RISC
RNA-sequencing: RNA-seq

Single nucleotide polymorphism: SNP
Trans-activation domain: TD
Transcription factor: TF
Transcription factor binding sites: TFBS
Transcription start site: TSS
Transcription termination site: TTS
Yoruban population: YRI

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to especially thank my supervisor, Dr. Tomi Pastinen for all the support and dedication he provided me throughout my graduate studies. Your raw love for genetics is magnetic and transfers to anyone that surrounds. I would like to thank you immensely for guidance and feedback over the last two years. I am grateful for the opportunity that you provided. Lastly, I appreciate your support and encouragement in all my future endeavors.

I would like to thank the members of my supervisory committee: Dr Anna Naumova and Dr. Guillaume Bourque for your help throughout my graduate studies, which greatly improved my work. My project could not have been possible without your constructive criti*cis*m, insight, and encouragement.

I would like to thank the various professors that I had the privilege to encounter: Dr. Yan Joly, Dr. Jacek Majewski and Dr. Aimee K. Ryan for advancing my knowledge in the field of human genetics. In addition, I would like to thank those that provided the weekly Journal Clubs, which allowed me to stay up to date on current discoveries, as well as, enhance my skills in public speaking.

I owe many thanks to the members of the Pastinen lab. To Dr. Veronique Adoue, who has been my mentor and mother in the laboratory. I think without your constant encouragement and support I would not have completed my degree. Your passion for science and knowledge were invaluable. To Dr. Stephan Busche, I especially appreciate your vast knowledgeable in the field and your ability to answer any questions. To Dr. Tony Kwan, I especially appreciate that I could rely on you for anything including advice for food options and comic relief. To Sherry Chen and Liliane Karemera, I really appreciate all the help on my projects. To Nicholas Light, your addition to the team was immediately felt. I appreciate your constant support and advice. I would like to thank Haig Djambazian and Bing Ge for their constant encouragement and help especially in terms of computer problems. I would like to especially thank Bing Ge for always taking the time to explain to me a concept with great detail.

I would like to thank Dr. Robert Sladek and the members of his team for their help and ideas in terms of my project, particularly to Dr. Albena Pramatarova. In addition, thank you to all of my colleagues at the Genome Quebec Centre who worked on my samples and whom I had the pleasure to befriend.

Last but not least, I would like to express my gratitude for my friends and family who supported me throughout my two years in graduate studies. This thesis could not have been possible without my friends understanding my very unreliable schedule and their constant encouragement. Lastly, I could not have finished my thesis without my parents and my brother Tim who beyond their support provided me food for fuel and many lifts to and from the laboratory.

CHAPTER 1: LITERATURE REVIEW

The Evolution of Functional Genomics

Human Genome Project (HGP)

A new era for the field of human genomics began with the completion of the HGP in 2003, 13 years after its initiation (*1, 2*). This enabled scientists to obtain the exact nucleotide sequence for any gene of interest and the location of the gene within the genome, as well as, within a particular chromosome. The international effort of the HGP enabled the shift from single-gene approaches to larger-scale, genome-wide "omics" strategies (*3*). As such, a surge of new analysis tools were developed, the majority of which used microarray technology in order to target millions of sites on the genome such as single nucleotide polymorphisms (SNPs) or exons. The aforementioned technologies allowed the potential to analyze transcript structure, gene expression and genetic markers on an unprecedented genome-wide scale. The impact of such analyses is evident for the field of human genetics including advances in the understanding of evolution, discovery of disease-susceptibility genes by linkage versus genome-wide association studies (GWAS), as well as, comprehension of both health and disease states (*3, 4*).

Encyclopedia of DNA Elements

The obstacle for the 21st century is the interpretation of the HGP (*5*). Following the completion of the HGP, the National Human Genome Research Institute (NHGRI) launched a project entitled the Encyclopedia of DNA Elements (ENCODE) aiming to catalogue and describe all of the functional elements encoded in the human genome sequence. The proportion of the human genome that encodes functional elements is unknown; however, it has been estimated using comparative genomics that 3%-8% of

base pairs (bp) are under negative selection (*6*).The term, "functional element," as described by ENCODE is used to denote a discrete region of the genome that encodes a defined product (e.g., protein), or a reproducible biochemical signature, such as a transcription factor (TF) or chromatin structure. In 2007, the pilot phase of ENCODE was completed, providing identification and analysis of functional elements for 1% of the human genome (*7*). Such signatures, either alone or in combination, are now known to mark genomic sequences with important functions, including transcriptional regulatory elements such as promoters and enhancers. The ENCODE Data Coordination Center (DCC) at the University of California, Santa Cruz is the central repository for ENCODE data, which is high-throughput, genome-wide data generated with technologies including Chromatin Immunoprecipitation (ChIP)-seq and RNA-sequencing (RNA-seq), (*8*). This repository is beneficial for scientists and researchers as data is publicly accessible for further analyses. The ENCODE project has been useful in mapping transcription factor binding sites (TFBS), histone marks, chromatin accessibility, DNA methylation, and RNA expression; however, additional analyses are still needed to deepen our understanding of functional elements (*8*). From the aforementioned ENCODE data, of particular interest to this project, is to identify *cis*-regulatory regions, including the study of TFBS in the human genome, in particular to understand their role with respect to linking genetic variation to changes in gene regulation (*5*).

The Transition to Functional Genomics

As described by Pevsner, functional genomics is the genome-wide study of the function of DNA, as well as, the nucleic acids and protein products encoded by DNA (*9*). This field is rapidly progressing towards the elucidation of elements that are crucial for the *cis*-

regulatory control of gene expression, which will be a main focus of our discussion. This transition includes diploid or allele-insensitive analyses to haploid or allele-specific examination (*10*). As such, the identification of DNA regulatory regions is a highly important yet challenging problem toward the functional annotation of genomes (*11*). Fortunately, the transition to functional genomics has been aided by the advent of new technologies.

Until recently, sequencing studies, including the HGP, relied on Sanger-based sequencing technologies (*12*). Since the completion of the HGP, considerable effort has been made in creating technologies capable of sequencing an entire human genome in a timely and cost-effective manner. As such, next-generation sequencing (NGS), has allowed for other genome studies aiming to elucidate a reference genome for organisms, as well as, various human populations, which are fundamental to continued research (e.g. 1000 Genomes (*13*) consortiums). NGS technologies can be used for global functional genomics assays in general or with respect to functional data with allelic resolution (*14-16*). In addition, functional data with allelic resolution can be extracted by analyzing variable sites (polymorphisms) using genome-wide genotyping arrays (*17, 18*). The advent of publicly available data, which can be combined together to increase the power of studies, has only further developed the field of functional genomics. As such, bioinformatics has become an intricate tool to studying functional genomics output.

The application of the aforementioned techniques will be described in greater detail below. As such, in this thesis, we present an application of functional genomics in terms of allele-specific analyses. This will be done by combining high-throughput genomic data with targeted approaches to study function in living cells in order to better

understand *cis*-regulatory variation and its impact on gene expression and disease. This will not only provide mechanisms for individual disease and trait associated SNPs but also a general paradigm of how to study the effect of single-base differences in an intact chromatin context.

Regulatory Variation in the Human Genome

Understanding complex disease and traits

Mendelian diseases are defined as those caused by single-gene mutations and which follow Mendel's laws of inheritance. Determining genes which underlie Mendelian diseases have also been valuable in assigning function to genes (*19*). Mutations associated with Mendelian disorders have been historically uncovered with linkage studies or with the recent development of SNP arrays, with homozygosity mapping (*19-22*). Advances in establishing projects for Mendelian disease are allowing a shift of the volume of resources onto the intricacies of complex diseases and traits *(Figure 1-1).*

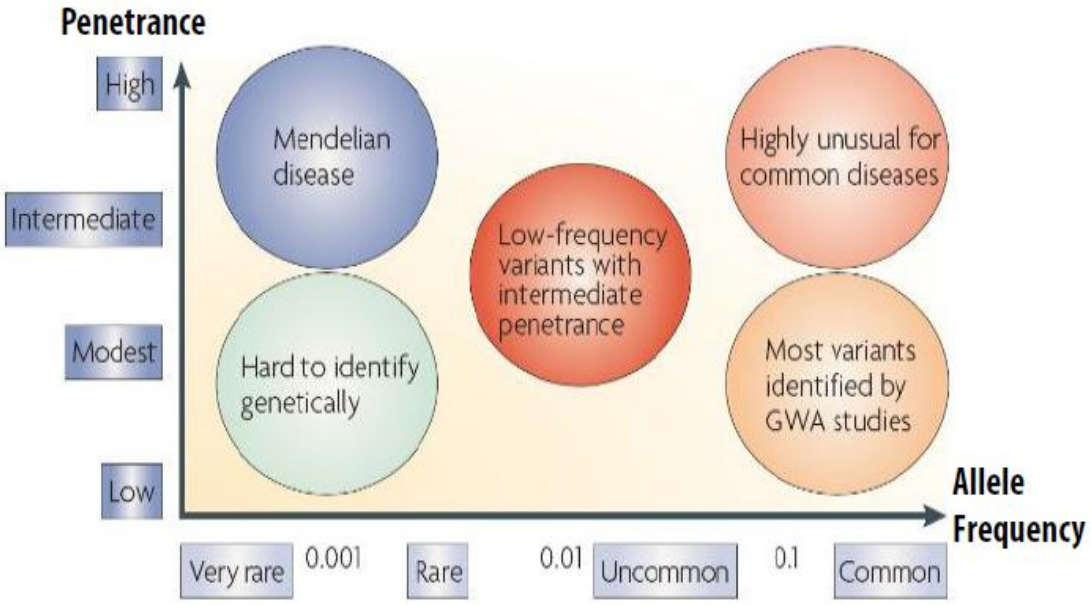


Figure 1-1. Identification of variants for complex diseases and traits. Various methods have been developed to study complex disorders one of which is GWAS, which identify the common genetic factors that underlie major complex diseases and traits (Adapted from Lobo et al., 2008).

The genetic basis for common complex diseases such as asthma, osteoporosis, or autoimmune diseases, indicates that genetic variants or SNPs, common at the population level often alter disease risk in a subtle manner (*23*). Gene expression is one of the complex traits known to be influenced by *cis (proximal)* and *trans (distal)*-acting genetic, epigenetic (e.g., methylation and histone modifications), and environmental influences (*10*). Variation in gene expression has also been shown as one mechanism underlying susceptibility to complex disease (*24*). Evolutionary constraint indicates that most of the functional DNA in the human genome is non-coding (*25*). In parallel, over 80 % of complex disease variants, which have been discovered by GWAS, are located in non-coding regions of the genome and seldom implicate common coding variants (*26*). GWAS have identified 100s of genetic variants associated with complex human diseases and traits also providing valuable insight into their genetic architecture (*26*). Furthermore, most of the validated disease SNPs, which are located outside the coding regions of the human genome (*27*) are presumed to impact gene regulation (*24*). As such, population-based studies of disease as well as gene expression traits are demonstrating the widespread evidence of the impact of non-coding variants on trait variance.

Deciphering the genetic code for regulatory DNA in our species has only recently begun (*7*) and the challenge now is to ameliorate our understanding of how and where non-coding variants act. Learning the relationship between genetic variation and variation in chromatin has the potential to bridge the gap between GWAS, which have linked disease to SNPs, as well as, better our understanding of how such polymorphisms, the majority of which are found in non coding regions, can underlie phenotypic variation (*28*).

Studying cis-regulatory evolution

The concept of evolutionary change was predicted over 40 years ago due to mutations, which altered the regulation of gene expression (*29, 30*). During the past 5-10 years, predictions have been validated by empirical evidence and identified that regulatory loci are the cause of divergent phenotypes. This evidence included expression divergence that correlates with phenotypic divergences, manipulations of gene expression that are able to recreate phenotypic differences as well as genetic mapping (*31, 32*). The view of transcriptional regulation is that *cis*-regulatory elements, such as promoters and enhancers, and proteins that bind to these elements control transcription of different genes (*33*). *Cis*-regulatory sequences can be clustered into *cis*-regulatory elements (CREs), which are a collection of TFBS and non-coding DNA that are sufficient to activate transcription in a defined spatial and/or temporal expression domain (*34*). There are long-standing questions with respect to the evolutionary process, which can be potentially addressed by identifying the genetic basis of divergent phenotypes (*31*, *32*). The current hypothesis is that *cis*-regulatory sequences are thought to be the most prevalent cause of phenotypic divergence. As such, identifying the sites that are responsible for divergent activity of *cis*-regulatory sequences can help to resolve these questions (*34*).

Regulatory variation is important not only in terms of evolution but there are well known biomedical traits caused by variation in non coding DNA that alters gene expression in *cis* (*10*) . There is 0.1 % heterozygosity in the human genome and there are a vast majority of polymorphic sites in non coding DNA (*35*) . Some of these differences contribute to risk of disease and other complex phenotypes (e.g., responses to drugs). It is known that a lot of functional DNA is located outside of the coding region notably by

looking at evolutionary conservation (*27*). One way to study variants underlying *cis*-regulatory variation is through mapping studies using expression quantitative trait loci (eQTLs).

Mapping regulatory variation by eQTLs

Initial studies of genome-wide mapping of gene expression began in the 1980s and 1990s by Damerval and de Viennes (*36, 37*). Studies in humans linking genetic variation to gene expression primarily focused on eQTLs (*10, 16*). EQTLs are used to describe a statistically significant genotype-gene expression level correlation. Transcript level correlation has been used to map *cis*-eQTLs in immortalized and primary cell panels (*38, 39*). Methods include specific expression platforms (*40*), and more comprehensively applying NGS of the transcriptome (RNA-seq) (*15, 16, 41*) is used to achieve higher resolution mapping of eQTLs. Therefore, genotyping data can be collected at a high density, which is needed for association-based mapping or at a lower density in order to perform family based linkage or eQTL mapping (*42, 43*).

EQTL data is available for a wide range of cell types and organisms. EQTL studies of human genes have implicated proximal regulatory variants (*cis-eQTLs)* as a prevalent cause of population variation in gene expression by co-localization of an associated signal with a gene of interest (*17, 40, 44-48*). Even though, recent studies have identified human eQTLs and elucidated their contribution to phenotypic variation, the list is still limited. Currently, only sporadic examples of causal SNPs in humans exist (*49*). A more efficient way to identify the causal variant underlying eQTLs is required since little is known about the regulatory mechanisms they act-by to alter gene regulation. A drawback of the eQTL mapping method is in the ability to specifically detect *cis*-

regulatory effects, which is reduced by background *trans*-acting variation. As such, even though characterization of human genetic variation, which affects gene expression, has focused on eQTL mapping, direct assessment of *cis*-regulatory variation necessitates allele-specific approaches (*10*).

Mapping regulatory variation by detecting allele-specific expression

An alternative approach, which directly assesses *cis*-acting components of expression variation, is through the mapping of differences in allelic expression (AE). AE is the measure of the relative expression between two allelic transcripts (*17*). In order to directly demonstrate that a variant acts in *cis* requires AE measurements followed by mapping of AE across samples to the genetic variants in the locus (*17*).The principle of AE is as follows: it is expected that in an autosomal locus the two allelic copies of the transcript will have equal expression. However, it is observed in some cases, biased AE of one of the two allelic transcripts in which one allele is differentially expressed *(Figure 1-2).* Allele-specific analyses rely on the power of using a within sample control, namely the other allele, which sensitively and specifically assesses the effects of genetic and/or epigenetic differences on *cis*-regulatory control in diploid genomes (*10*). The power of allele-specific analyses of gene expression (*50*) or transcriptional activity (*51*) has been elucidated by studying individual loci but currently it is possible to assess on a genome-wide scale due to recent advances in genomic technologies (*10*).

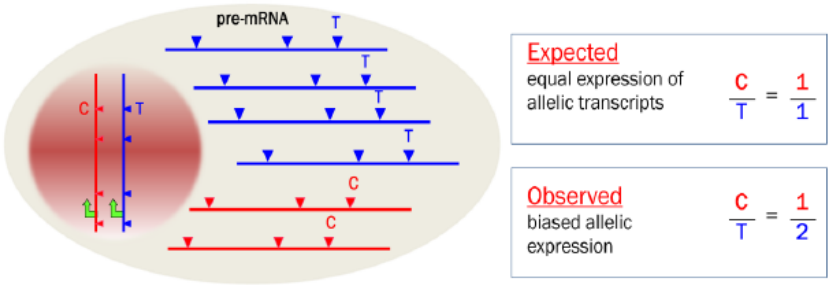


Figure 1-2. Allelic Expression (AE). The principle of AE is as follows. In an autosomal locus the two allelic copies of the transcript are expected to have equal expression of the two alleles (1:1 ratio) *(top)*. In many cases biased AE is actually observed and one allele is differentially expressed (1:2 ratio) *(bottom)*.

Methods with allelic discrimination are more powerful as environmental and *trans*-acting influences that alter gene expression or DNA-protein interactions are minimized. The challenge of environmental influences on connecting variation in genome sequence with variation in TF binding and gene expression (*52*) are minimized since both alleles are under the same internal control (*10*). This provides higher sensitivity for elucidating the direct influence of sequence or epigenetic variation in *cis*-regulatory elements (*10*). Results indicate that up to 30% of RefSeq transcripts exhibit AE differences. These variants can be identified by mapping differences in AE on Illumina HumanOmni1 BeadChips. Mapped *cis*-rSNPs explain >50% of population variance in AE. The primary discovery panel for this study consisted of 53 unrelated HapMap Caucasian (CEU) lymphoblastoid cell lines (LCLs) (*17*). We were able to further demonstrate that use of an ethnically distinct Yoruban African (YRI) population allows for fine-mapping of *cis*-rSNPs in certain cases (*17*). Using genome-wide maps of SNPs altering gene regulation (*17*) in conjunction with complex disease associated SNP catalogs (*27*) we are also able to build specific hypotheses of potential causal sites for various complex disease and traits. Therefore, another purpose of my research is to extend traditional approaches for validation of these hypotheses to living human cells. This will not only provide mechanisms for individual disease associated SNPs but also a general paradigm of how to study single-base differences in an intact chromatin context.

Global approaches to studying AE in the human genome are made possible due to the rapid development in sequencing technologies. This is done using NGS of the transcriptome (RNA-seq) or ChIP-seq *(to be discussed below).* RNA-seq has single nucleotide based resolution and it collects short sequence reads uniformly across

expressed transcripts and these reads correlate with the abundance of the RNA species. However, this technology only gives information about exonic sequences at a very high depth (*42*). This method yields information for only a small proportion of genes because of the unequal representation of the different RNA species and in certain cases the limited genetic variation in the mRNA. Therefore, the use of RNA containing unspliced primary transcripts in allele-specific expression analyses provides more information about a larger proportion of genes than if solely coding SNPs were used (*10*) *(Figure 1-3).*

Targeted approaches to measure AE allows the opportunity to investigate individual variant sites for allele-specific function. One advantage is the higher density of genomic data due to the fact that only sites, which are potentially informative (heterozygous) for allelic analyses are observed. Moreover, the sites are also targeted in genomic DNA (gDNA) control samples, which have an equal allelic content. The gDNA is used as a control for technical biases intrinsic in quantitative assessments of allele ratios (*10*). A low-cost and convenient method of assessing allele specificity of polymorphic sites is through the use of genome-wide genotyping arrays (*17*). One drawback is the coverage of allelic differences in terms of regulatory elements because standard SNP arrays encompass only a small subset of polymorphic regulatory elements (*10*) *(Figure 1-4).* Another consideration for study is variants underlying *cis*-regulatory variation, which affect AE, and are shown to be transmitted from parent to offspring (*48, 53*).

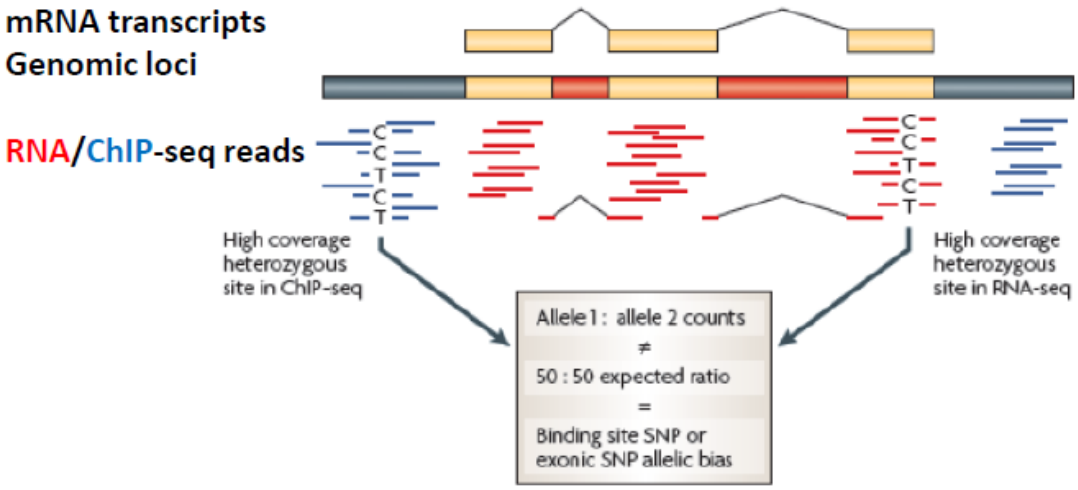


Figure 1-3. Global approaches to studying allele-specific function. Global approaches include NGS-based methods such as RNA-seq *(red)* and ChIP-seq *(blue)*. These methods can be used to investigate allelic effects for reads which overlap a site with a high coverage read depth. In turn, the ratio of reads from each allele is calculated and the allelic bias at each site is determined if there is a deviation from the expected 50:50 ratio (Adapted from Pastinen, T., 2010).

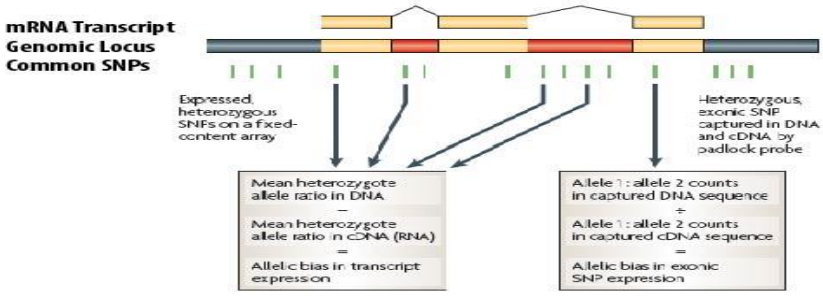


Figure 1-4. Targeted approaches to studying allele-specific function. Information based on SNPs *(green bars)* can be integrated in unspliced primary transcripts or target specific exons (yellow). In terms of studying primary transcripts, quantitative measurements are taken for multiple phased polymorphisms spanning the transcripts (mean allele ratio in cDNA) and normalized to genomic DNA from the same samples (mean allele ratio in DNA), which generates information on allelic biases in transcript expression (Adapted from Pastinen, T., 2010).

Heritability in the genomics era

Heritability is a population parameter, and consequently depends on population-specific factors, which includes allele frequency, the effects of gene variants, as well as, environmental factors. This measure allows for a comparison of the relative significance of genes and environment to the variation of traits within and across populations (*54*). In terms of complex traits, GWAS generally only explain a few percent of the estimated heritability, which suggests the necessity of further analysis into contributions to heritability (*55*).

Allele-specific gene expression is believed to play a role in phenotypic variation, but the genetic mechanisms responsible are not well understood. Heritable variants in gDNA include SNPs, insertions, and deletions, which may act to influence gene regulation (*39, 48*). As such, the previously mentioned heritable variations have been thought to affect the binding of sequence-specific TFs or to affect the conformation of chromatin. In earlier studies, McDaniell and colleagues analyzed the heritability of individual-specific and allele-specific binding of TFs CTCF and DNase I in two unrelated trio families (*28, 53*). It was shown that approximately 10% of active chromatin sites were individual-specific and a similar proportion were allele-specific. Both of these were commonly transmitted from parent to child (65%), which suggests that these are heritable features of the human genome. As such, it was suggested that in humans up to 11% of SNPs in sites involved in modulating chromatin activity demonstrate heritable allele-specificity and could directly affect TF binding and chromatin structure (*28, 53*). Consequently, there are many applications of heritability especially in terms of the genomic era (*54*). Therefore, we will utilize our expertise at allele-specific differences in

the genome to carry out direct assessment of allele differences in TF binding, which can be transmitted from parent to offspring.

The Mechanism of Action of TFs

TF Model

Comprehending how genomic information is translated into gene regulation has been investigated for decades now (*33*). An important question that needs to be further explored in the field of gene regulation is the extent by which variation in TF binding influences the effect and/or the mechanism by which it regulates transcription (*56*). In order to address the above, comprehensive TF binding maps are necessary.

A TF is a protein that binds to a sequence such as a promoter or enhancer element, to control different levels of transcription of genes (*33*). In humans, there are approximately 200-300 TFs that bind to core promoter elements and are components of the transcriptional machinery. Furthermore, there are approximately 1,400 TFs that contain sequence-specific DNA-binding properties. Due to the sequence specificity of some TFs they only regulate a subset of genes by binding to site-specific *cis* elements (*1, 2, 57*) *(Figure 1-5).*

TFs are central cellular components that control gene expression: their role is essential in determining how cells function and react to the surrounding environment. The transcriptional regulatory system also plays a major part in controlling a wide range of biological processes including cell progression, cellular differentiation and developmental time courses. Several diseases are caused in part or due to a breakdown of the regulatory system. Moreover, a large source of phenotypic diversity and the method by which organisms adapt to evolutionary changes are due to changes in the activity and

regulatory specificity of TFs (*57*). Multicellular organisms have complex TFs most of which work together with co-regulators in order to form networks of cooperating and interacting TFs (*58*).

The regulation of gene expression at the transcriptional level has become increasingly clear to be achieved by the complex interactions of TFs working cooperatively at their target genes. A necessary primary step is to better understand the evolution of transcriptional regulators, their relationship, as well as, their regulatory interactions with target genes. Although this idea has been elucidated the challenge remains to discover which TFs work cooperatively, the sites of cooperative action and to what extent does the influence of the cooperative action of TFs effect the regulation of target genes *(Figure 1-5)*. In previous studies, a variety of approaches such as *in vivo* and *in vitro* detection of protein-protein interactions, have been used to measure TF co-association. However, the above assays contain technical problems in terms of sensitivity and specificity (*59, 60*).

Recently a unique approach, the Allele Binding Cooperativity (ABC) test was used to identify TF co-association. The ABC test examines co-variation of motifs with variable binding regions (BR). (*56, 61*). BRs are clusters of binding peaks identified by ChIP-seq. The underlying concept is that variation in TF binding likely occurs because of sequence variation for associated TF binding sites and motifs. For instance, with respect to NF-κB, other associated DNA motifs, such as the STAT1 motif [previously associated with NF-κB (*62*)] and de novo searches for enriched DNA motifs in BRs were done. In addition, using the ABC approach, effects of genetic variation for each motif were analyzed. SNPs in the STAT1 motif were shown to elevate the

frequency of significant NF-κB binding differences (1.3 fold enrichment). Moreover, an improved STAT1 motif sequence increased the binding of NF-κB in 71% of cases (*56*). This suggests that there is functional interaction between STAT1 and NF-κB (*63*, *64*). The ABC approach has been used to direct a computational and experimental pipeline to identify targets from variation data. The aforementioned pipeline is a large-scale process to hunt for allele binding interacting transcription factors (ALPHABIT) and was applied to identify novel binding partners of NF-κB (p65) (*64*). The method successfully identified factors known to work with NF-κB (E2A, STAT1, IRF2), as well as, a unique association (EBF1). Furthermore, this approach highlighted functional information for TF coassociation indicating that variance in the motif of one factor correlates with the binding of the other factor (NF-κB). A cooperative mechanism was suggested for NF-κB and the aforementioned TFs due to the fact that binding of the putative coassociated factors were also shown to significantly predict binding of NF-κB. However, the global coassociation and sites of cooperative action for NF-κB are still relatively unknown and remain an area of further study (*64*).

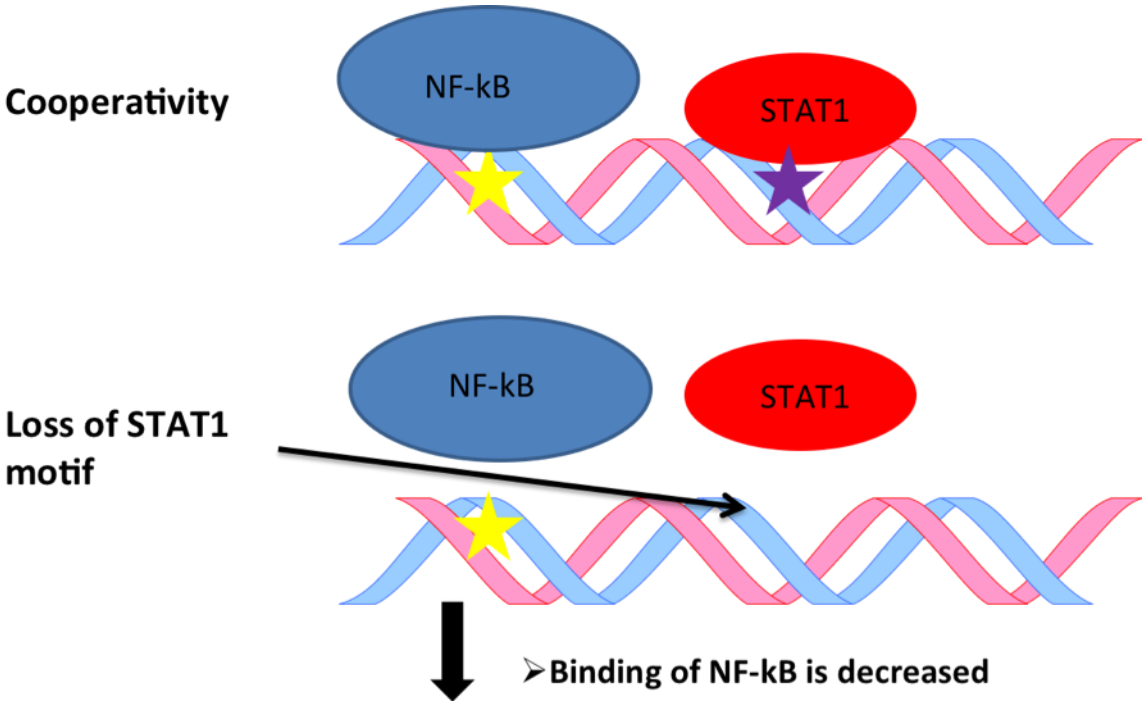


Figure 1-5. Model of cooperative TF associations. The model of cooperativity of TFs is outlined. For instance, if a STAT1 motif is present, both STAT1 and NF-κB are bound *(top)*. Due to the loss of the STAT1 motif, there is decreased binding of STAT1, as well as, NF-κB even though the NF-κB motif is still present *(bottom)*.

Variation in TF-DNA Binding

In higher eukaryotes, the majority of TFBS are organized into clusters called *cis*-regulatory modules (CRM), which are made up of DNA regions of up to a few hundred bp in length (100-900 bp) located in the neighborhood of the gene under regulation (*65*). The most prevalent cause of phenotypic divergence is currently thought to be mutations affecting the activity of *cis*-regulatory sequences, which can have a wide range of effects including altering TF binding (*31, 32*). One analysis by Kasowski and colleagues investigated whether BRs are population specific. Their work showed that ~ 0.1% to ~0.4% of events were population specific, which suggests that most alleles affecting TF binding are shared amongst populations (*56*). In contrast, the extent to which TF binding differences occur among individuals and the overall relationship between TF binding and genetic variation remains largely unexplored (*56, 66*).

In the past, several techniques were used to study variation in TF activity, the consequences on regulatory mechanisms, as well as, the effect of TF variation on gene expression. Assays have also analyzed TFs in terms of chromatin state for TF-DNA binding in order to identify active gene regulatory elements genome-wide. Widely used methods include DNaseI hypersensitivity sites (DHS), formaldehyde-assisted isolation of regulatory elements (FAIRE) and ChIP. DHSs are regions in the genome displaced by TFs; therefore, the regions are sensitive to DNaseI digestion. This method can identify different types of regulatory elements. Even though traditional DHS tools do not directly reveal which TFs are binding to a target region, it does uncover functional regulatory elements where TFs are likely to bind. In turn, FAIRE uses formaldehyde to biochemically separate DNA that is packaged into nucleosomes from DNA that is bound by non-nucleosomal proteins like TFs (*28*). Lastly, in order to assess TF occupancy ChIP

experiments can be used. Originally, the above three methods involved detection of specific signals using Southern blots or PCR; however, currently all the techniques have been adapted to use NGS technology (*28*).

The widely used application of NGS, in particular for ChIP-seq, as mentioned above, can provide highly specific information in terms of factor location but it is limited by factors for which high-grade antibodies are available and only one antibody can be assessed per experiment (*67*). ChIP only gives information about specific regions bound by the TF of interest and not the rest of the genome (*28*). There is also relatively low coverage of polymorphic sites in NGS studies, which remains an obstacle considering only a small subset of sites are informative for analysis (*10*) *(Figure 1-3)*. As described, differential allelic activity can affect the recruitment of TFs to DNA and thus alter disease phenotype. For instance, in a genome-wide study, SNPs were identified that correlate with population variation in DNA-protein interactions, which was assessed by ChIP-seq. This showed that *cis* variation has an effect on gene expression and regulatory DNA activity (*56*). In order to gain insight into the functional consequences of allelic differences in TF occupancy it is imperative to measure differences in AE using relevant cell types. Analysis of TFBS can also aid in understanding the mechanism of regulation, including the coordinated regulation of transcription factors acting cooperatively and as such, identify mutations that disrupt regulatory mechanisms. The model that will be further studied to better understand the aforementioned is the TF NF-κB.

Understanding NF-κB Function

Background and Rational for Study

For the past twenty five years NF-κB has been a well studied TF and thus provides a good model for further investigation into the intricacies of TF-DNA binding (*Figure 1-6)*.

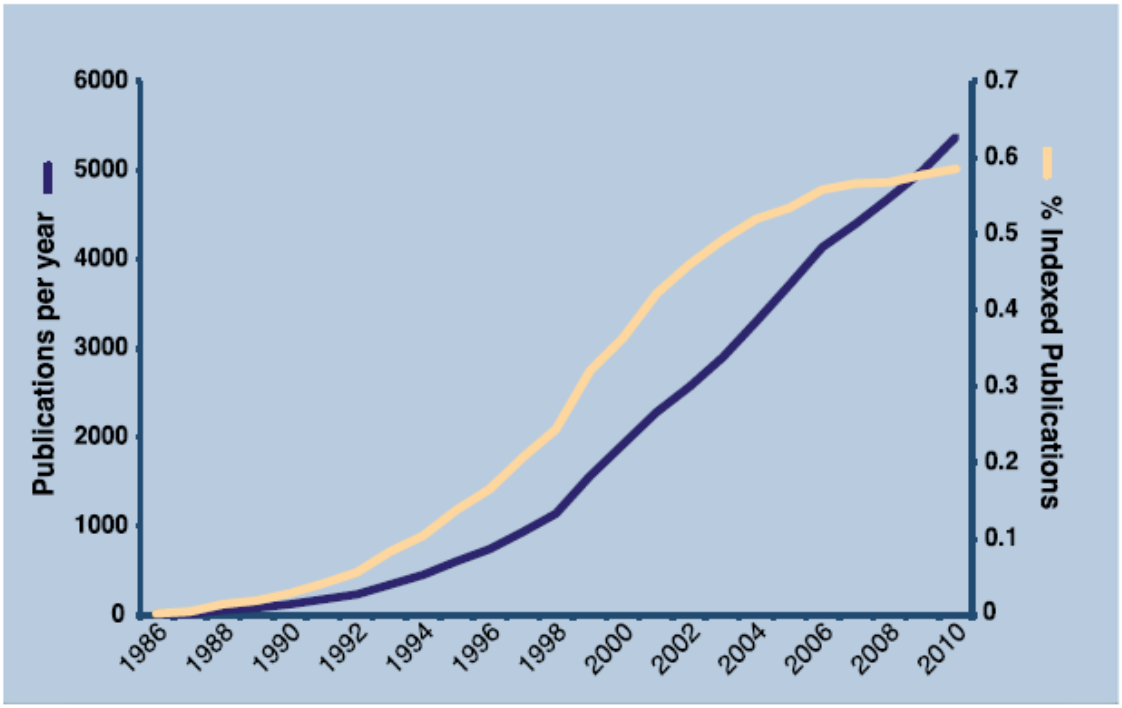


Figure 1-6. NF-κB literature. The graph demonstrates the total number of publications identified in PubMed using the keywords NF-κB, Rel, or IKK per year since 1986 (this is shown in the *left* axis). Also portrayed in the graph are the total publications identified with the aforementioned keywords as a percentage of all PubMed publications in the same calendar year (this is shown in the *right* axis) (Adapted from Hayden, et al., 2012).

NF-κB has served as a main model for evolutionarily conserved signal-activated TFs (*68*). This includes over 150 different stimuli mainly stress, cytokines, ultra violet light, viral and bacterial particles. Inducible activation of NF-κB plays a role in the control of transcription of over 150 target genes (*69*).The model of inducible regulation of gene expression enables organisms to more easily adapt to environmental, mechanical, microbiological, and chemical stresses (*70*). The NF-κB family of TFs have been studied due to their influence on gene expression for different biological processes such as maintenance of the immune system (*71*), epithelium, cell survival, apoptosis (*72*), differentiation, and proliferation (*70*). Dysregulation of NF-κB can lead to severe consequences including many diseases such as inflammatory diseases, autoimmune diseases (*73*), neurodegenerative diseases, diabetes, cardiovascular diseases and oncogenesis (*70-72, 74, 75*). However, the role of NF-κB is best understood in the context of chronic inflammatory and autoimmune diseases. Inflammation is the response of vascular tissues to any harmful injury including pathogens, irritants or damaged cells. Inflammation is part of the non-specific immune response, which acts as a protective mechanism for the body to remove a particular stimuli (*76*). Autoimmune diseases occur in response to an overactive immune response of the body against tissues and substances present normally in the body. Autoimmune disorders tend to be either systemic (such as systemic lupus erythematosus) or organ-specific (such as type 1 diabetes) and are characterized by prolonged inflammation and subsequent tissue destruction (*77*). As such, studying NF-κB can provide insight into the disease mechanism and pathogenesis for a wide range of diseases.

NF-κB is also a strong predictor of certain chromatin states. The NF-κB motif was most often found enriched (fold changes) in active promoters (40.7) and strong enhancers (49.0) *(Figure 1-7a).* Moreover, lymphoblastoid-specific enhancers enriched in the cluster "F," which is the immune response were preferentially bound by NF-κB in LCLs *(Figure 1-7b).* Therefore, NF-κB appears to control the expression of many genes and is potentially one of the key regulatory TFs in LCLs (*13*). As such, LCLs provide an optimal model to study TF-variation for NF-κB. Studying NF-κB was pursued by using two trios of LCLs from CEU and YRI populations. This was not only done for the above reasons but also because the cell lines have been well characterized by projects including the ENCODE project (*7*), the HapMap (*78*) and 1000 Genomes (*13*) consortiums. In addition, we have an in-house generated map of *cis*-regulatory variants associated to AE for LCLs, which provides a starting point for further analysis (*17*).

Kasowski and colleagues performed another relevant study, which analyzed binding of the NF-κB protein (p65) in stimulated LCLs. The work revealed binding sites for p65 in 10 LCLs. It was shown that 7.5% of binding sites differed between individuals and the binding differences were frequently due to variations in SNPs (*56*). However, further analysis is still needed in order to accomplish mapping of causal variants underlying binding differences.

Lastly, NF-κB can have a global effect on gene expression by shutting down and/or activating a wide range of pathways (*68*). As such, we decided to employ inhibition of NF-κB coupled to induction by TNF- α, followed by AE measurements in order to analyze NF-κB-DNA binding. Allele-specific assessment of expression as compared to total expression measurements is advantageous for the following reasons.

AE analysis is ideal as it uses the other allele as an internal control for other influences such as non-*cis*-acting factors (e.g., environment) (*10*). This diminishes the background noise created by indirect effects of inhibiting the action of NF-κB in the cell.

a) b)

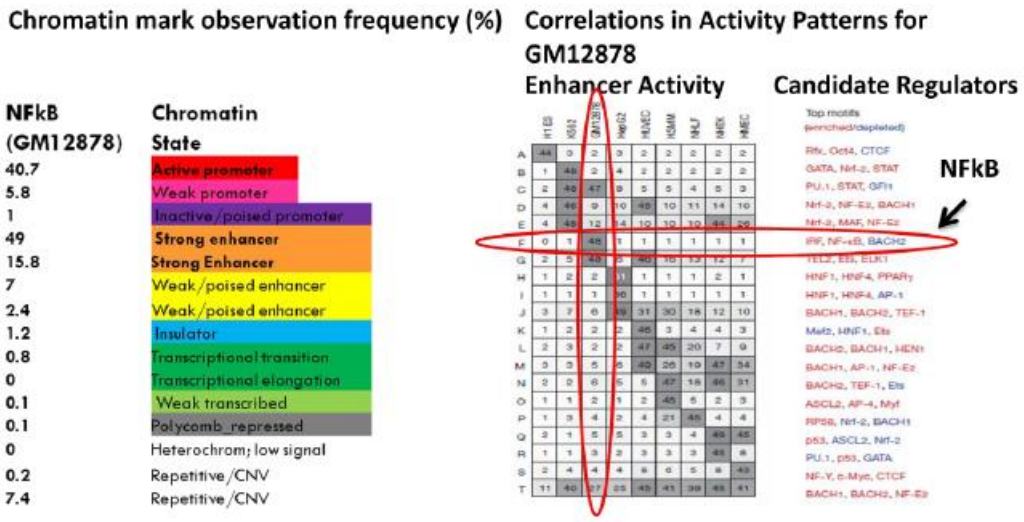


Figure 1-7. Chromatin state and NF-κB characterization. The table *(left)* shows chromatin states learned jointly across cell types *(by a HMM)*, which deciphered 15 chromatin states. This shows the functional enrichment of NF-κB was most often found in active promoter and strong enhancer elements. NF-κB was also shown to be a main factor binding enhancers in LCLs *(right)* (Adapted from Ernst et al., 2011).

The Mechanism of NF-κB Action

NF-κB plays a fundamental role in two different pathways, which include the canonical pathway and the non-canonical pathway. NF-κB functions as a homo- or heterodimers, which consist of the reticuloendotheliosis (Rel)-homology domain containing monomers from two sub-families: p50 and p52 (Type I subunits); RELA (p65), RELB and C-Rel (Type II subunits). Type II subunits have trans-activations domains (TDs) and can act as transcriptional activators alone, where as type I subunits can only activate transcription as a heterodimer with a type II subunit or as a homodimer in complex with co-factors such as IKBZ and BCL3 (*74*) . Members of the NF-κB family of TFs bind to a, "core motif," that is between 10 to 11 bases and are variations of the originally described consensus, GGRRNNYYCC(*79*).

The canonical pathway consists mainly of dimers composed of p65:p50. This TF is sequestered in the cytoplasm by the inhibitor of kB (IκB) and in order to interact with DNA it needs to dimerize as either homo- or heterodimers in different combinations (*80*). The Iκb kinase (IKK) complex including catalytic subunits, IKK-α and IKK-β, and the regulatory NEMO protein, together degrade the NF-κB sequestering complex, IκB through phosphorylation on its serine components. This targets IκB proteins for degradative Lys48 linked polyubiquitination, resulting in their proteolysis, which in turn frees NF-κB dimers from inhibition in the cytoplasm and the dimers can enter the nucleus (*81*) in order to bind DNA (*Figure 1-8)*.

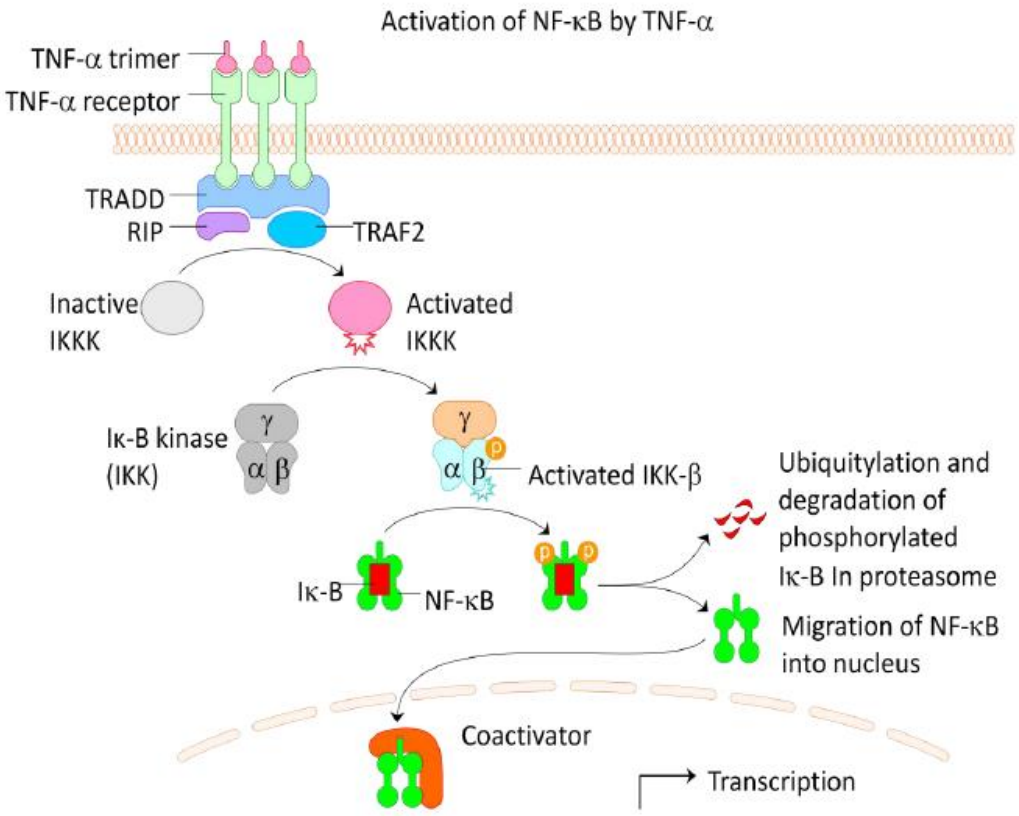


Figure 1-8. **Activation of NF-κB by TNF-α.** NF-κB is activated mainly in response to stimuli such as stress, cytokines, free radicals, ultra violet light viral and bacterial particles. Upon ubiquitylation and degradation of phosphorylated iκb in the proteasome, there is migration of NF-κB into the nucleus. NF-κB works cooperatively with other co-activators on target genes.

Objectives and Hypothesis

*Cis- r*SNPs alter *cis*-regulation of gene expression. Traditional tools for studying SNPs in regulatory regions typically isolated from their cell type dependent chromatin context are crucial for appropriate and coordinated regulation of gene expression. Consequently, such approaches may fall short in explaining differences in gene regulation observed in intact human cells (*82*). We hypothesize that combining high-throughput genomic data with targeted approaches to perturb TFs in living cells can be used to better understand DNA-protein interactions. We aim to optimize and develop a method of TF perturbation and apply it to NF-κB. We are specifically interested in studying the p65 subunit of the NF-κB family of TFs in order to better understand binding differences and to accomplish mapping of causal variants underlying binding differences.

In order to obtain a more in depth understanding of *cis*-regulatory variation and the impact of gene expression and consequences on disease studies a case study was also completed. We used the TF SNAI1 as we had a strong hypothesis for the involvement of SNAI1 and *WNT4*, as well as, evidence for its role in fibroblasts (FBs).

We hypothesize that using the aforementioned techniques will elucidate the regulatory role of NF-κB suggested in LCLs. In order to explore the above hypothesis the following objectives will be completed:

- Utilization of different approaches to carry out TF perturbation
- In vivo, DNA-protein assays of chromatin after perturbation of the genome
- Genome-wide AE assessment of genes involved
- Applying these approaches to better understand *cis*-rSNPs altering disease risk and mechanisms of disease associated SNPs
- Validation of allele-specific differences in TF binding in the human genome

CHAPTER 2: METHODS

Please Note: All analyses were done by myself with the following exceptions. Normalization of Illumina BeadChips using complementary DNA (cDNA) and gDNA was done by Bing Ge (Bioinformatician). Extraction of TRANSFAC binding sites was done by Dr. Tony Kwan (Research Associate). Graphs using the ENCODE data were done in conjunction with Dr. Tony Kwan. Double stranded cDNA (dscDNA) was done by Sherry Chen (Laboratory Technician).

Selection of Cell Lines and Mapping Datasets

The identification of *cis*-acting components of expression variation was done by mapping differences in AE. We used in-house generated data in two human cell types (lymphoblasts and fibroblasts) in which we have mapped potential *cis*-regulatory SNPs (*cis*-rSNP) for several thousand loci ((*17*) and unpublished). The *cis*-rSNPs are the top SNPs associated with allelic differences in gene expression, which we then ranked by p-value for each locus. AE assessment was carried out on trios from 3 different populations: 55 HapMap CEU LCLs and 63 HapMap YRI LCLs and 70 fibroblast (FB) cell lines. This was done by using Illumina BeadChips on gDNA and cDNA samples in order to analyze genotype and AE data in parallel. We used the genotypes and AE measurements to map top candidate *cis*-rSNPs, which control differential AE. Control SNPs were chosen with equivalent minor allele frequency (MAF), located at the same distance from the transcription start site (TSS) as the case SNPs but using another randomly chosen transcript with equivalent expression. Preliminary intersection of these variants with disease SNPs (http://www.genome.gov/26525384) indicates that there are

as many as 500 potential disease-relevant regulatory SNPs in our data. All the baseline data, i.e. genome-wide *cis*-rSNP maps and disease associated SNPs were available to me through our own database or public databases, respectively. Generation of the genome-wide maps of use in this thesis will be discussed in greater detail in the results section.

Normalization of Illumina BeadChip Readouts

Normalization of readouts from Illumina BeadChips were performed as in the article by Grundberg and colleagues (*38*). GDNA genotypes were extracted using BeadStudio. The parallel assessment of gDNA and cDNA heterozygote ratios was carried out essentially as described by Bing and colleagues, (*17*) but signal intensity normalization at heterozygous sites followed a somewhat different approach. For AE analysis we utilized the X_{raw} and Y_{raw} signal intensities, but since the variance in the two channels is not the same (i.e. it is a function of total intensity from both channels) this variation needed to be corrected through normalization in order to allow a comparison between gDNA and cDNA allele ratios. In this study, we only normalized β ratio (X_{raw}/ (X_{raw}+Y_{raw}) from heterozygous SNPs with total intensity (X_{raw} +Y_{raw}) higher than the threshold value of 1000. The scatter plot of β ratio against the $\log^{10}$ scaled total intensity fits well with the polynomial regression model (quadratic regression model). The quadratic model fits better than the linear regression model we employed earlier for normalization (*17*), which works well for the higher intensity component, but poorly in the lower intensity component in many samples. The normalization process can be summarized by the following key steps: 1) The β ratio is calculated in conjunction with total intensity in $\log_{10}$ scale for heterozygous SNPs. 2) All data points with greater than 1000 in total intensity are divided into 50 intensity bins. 3) A fitted curve from the median β ratio in each bin is

computed using a polynomial quadratic regression model, $y = \beta 1x + \beta 2x2 + a$, where y is the expected β ratio from the curve and x is the $\log^{10}$ scaled total intensity. 4) From the fitted curve, the expected β ratio based on total intensity is calculated. 5) The final normalized β ratio equals (β observed-β expected+0.5). After normalization, all median β ratio values in all intensity bins should be close, if not equal, to 0.5 (*38*).

Imputation

For the mapped list of *cis-r*SNPs, we used the 1000 Genomes Project as a reference set for imputation of genotypes from our panel of HapMap individuals. Untyped markers were inferred using algorithms implemented in MACH 1.0 (*83, 84*). The coefficient of determination, R^2 was used as an imputation quality control metric and estimates the squared correlation between imputed and true genotypes. All poorly imputed markers with $r^2<0.6$ were systematically removed.

Cell culture

The cells in which the regulatory effects were observed are cryopreserved and culturable in the Pastinen lab and provided me with the necessary biological samples for conducting much of the experiments outlined in this thesis. These cell lines were originally obtained from Coriell (http://www.coriell.org/). Due to the prevalent role of NF-κB in inflammatory disease, experiments were conducted in lymphoblastoid trios of the CEU and YRI populations. Lymphoblasts were grown in medium containing RPMI (SigmaAldrich, Suffolk, UK) supplemented with 2mmol/1 L-Glutamine, 100U/mL penicillin, 100U/mL streptomycin (National Veterinary Institute of Sweden, Uppsala,

Sweden), and 15% fetal bovine serum (SigmaAldrich, Suffolk, UK) at 37°C with 5% CO_2. FB and osteoblast cells were cultured in medium containing ∞-MEM (SigmaAldrich, Suffolk, UK) supplemented with 2mmol/1 L-Glutamine, 100U/mL penicillin, 100U/mL streptomycin (National Veterinary Institute of Sweden, Uppsala, Sweden), and 10% fetal bovine serum (SigmaAldrich, Suffolk, UK) at 37°C with 5% CO_2.

RNA and DNA preparation

DNA extraction

DNA was extracted from T_{75} tissue culture flasks upon confluence using the QIAGEN gDNA Extraction Kit for cultured cells. This was done according to the manufacturer's instructions (QIAGEN, Mississauga, Ontario). Concentrations were determined using NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, DE, USA). Extracted gDNA was stored at -20˚ Celsius.

RNA extraction

RNA was extracted from cell lysates using either the RNAeasy Mini Kit (QIAGEN, Mississauga, Canada) or the TRIzol reagent protocol (Invitrogen Corporation, Carlsbad, CA, USA) depending on the quantity of RNA. High RNA quality was confirmed for all samples using the Agilent 2100 BioAnalyzer (Agilent, technologies, Palo Alto, CA, USA), and the concentrations were determined using NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, DE, USA). RNA extraction, DNAse1 treatment, precipitation and cDNA synthesis were completed in order to perform real-time polymerase chain

reaction (RT-PCR). RNA samples were annealed to 500ng of random primers (Invitrogen Corporation, Carlsbad, CA, USA) at 70°C for 10 minutes.

CDNA synthesis and Real-time Polymerase Chain Reaction (RT-PCR)

First- and second-strand cDNA synthesis

First strand cDNA synthesis was performed using SuperScriptII reverse transcriptase (Invitrogen Corporation, Carlsbad, CA, USA) according to the manufacturer's instructions. The target genes as well as the *18S* housekeeping gene were analyzed in triplicates. The RT-PCR assays were performed on the Rotor-Gene™ 6000 real-time rotary analyzer (Corbett Life Sciences, Sydney, Australia) using the Platinum SYBR Green qPRC SuperMix-UDG (Invitrogen Corporation, Carlsbad, CA, USA) according to manufacturer's recommendations. Second strand cDNA synthesis was also performed using SuperScriptII reverse transcriptase (Invitrogen Corporation, Carlsbad, CA, USA) according to the manufacturer's instructions for the Illumina TotalPrep RNA Amplification Kit (Illumina Inc., San Diego, CA, US). This method required 50-500ng of RNA; however, approximately 3-5ug of RNA was used to ensure a sufficient quantity for subsequent steps. Second strand cDNA synthesis converts the single-stranded cDNA into a double-stranded DNA (dsDNA). This is used as a template for a reaction that employs DNA polymerase and RNase H to degrade the RNA and simultaneously synthesize the second strand of cDNA. The dscDNA was dissolved in 8-20ul DEPC treated water depending on the DNA microarray being used. The size distribution of the dscDNA samples (average 1.2–1.5kb) was confirmed using the Agilent BioAnalyzer DNA Kit (Agilent, Technologies, Palo Alto, CA, USA).

Primer design

Primers were designed using the Primer3 v. 0.4.0 software (http://frodo.wi.mit.edu/) and all primer sequences used can be found in Table 2-1.

Oligonucleotide name	Sequence	Amplicon length
RANK_F3	TTGCAGCTCAACAAGGACAC	1756 bp
RANK_R3	GATTTCTCTGTCCCATGATGTTC	
BCL2_F1	AAGCATACTCGAAGGCTCCA	211bp
BCL2_R1	GCGAGTGAGGAAAGGAGGTA	
IL1a_F1	CCGTGAGTTTCCCAGAAGAA	189 bp
IL1a_R1	ATCAGTACCTCACGGCTGCT	
IL1b_F2	TCTTTCAACACGCAGGACAG	133bp
IL1b_R2	TCCAGGGACAGGATATGGAG	
18S_F	TGTGGTGTTGAGGAAAGCAG	251 bp
18S_R	GGACCTGGCTGTATTTTCCA	
Cox2_RTF1	GCTGTCTAGCCAGAGTTTCACC	241bp
Cox2_RTR1	CCCTTGGGTGTCAAAGGTAA	
Il-8_RTF1	CTCTCTTGGCAGCCTTCCT	941 bp
Il-8_RTR1	AAATTTGGGGTGGAAAGGTT	
Il-6_RTF1	CCACACAGACAGCCACTCAC	1227 bp
Il-6_RTR1	TTTCAGCCATCTTTGGAAGG	
SNAI1_rs6684375_F1	TGCTCTATTGTGCTCCCTCA	290 bp
SNAI1_rs6684375_R1	GAAGCTCACACACCATGCAC	
SNAI1_rs6684375_F2	TGCTCTATTGTGCTCCCTCA	225 bp
SNAI1_rs6684375_R2	AGCCTCATCTCTCTGCATCC	
WNT4_RTF4	CGAGTCCATGACTTCCAGGT	162bp
WNT4_RTR4	CTCGTCTTCGCCGTCTTCT	
WNT4_RTF5	ACCTGGAAGTCATGGACTCG	235 bp
WNT4_RTR5	TCAGAGCATCCTGACCACTG	
SNAI1_RTF1	GCGAGCTGCAGGACTCTAAT	135bp
SNAI1_RTR1	GGACAGAGTCCCAGATGAGC	
rs909685_SYNGR1_F2	GCTGCGTTCACTGCTTTAGTC	112 bp
rs909685_SYNGR1_R2	AGGCATCAGAGGCAGAAATG	

Table 2-1. Primer sequences. The forward and reverse primer sequences are given for all primers used in PCR and RT-PCR reactions. The length of each amplicon is indicated.

PCR

PCR was performed in order to ascertain that the primers worked efficiently and the correct amplicon length was amplified. For each PCR reaction 4-10ng of gDNA and 10-15ng of cDNA was used. PCR product formation was assessed by electrophoresis in 2% agarose gel. The cycling protocol was as follows: initial denaturation was done at 95 °C for 15 minutes. The following steps were then repeated for 35 cycles: denaturation at 95°C for 30 seconds, annealing of the polymerase at 58°C for 30 seconds, and extension of the strand at 72 °C for 45 seconds. Following completion of these cycles, final extension was done at 72°C for 6 minutes.

RT-PCR

Validation of perturbation studies that will be discussed below were primarily analyzed by RT-PCR. Aliquots of RNA for each sample were annealed to 500ng of random primers. First-strand cDNA synthesis was performed using SuperScriptII reverse transcriptase (Invitrogen Corporation, Carlsbad, CA, USA) according to manufacturer's recommendations and as described above. The cycling conditions on the Rotor-Gene^{TM} 6000 real-time rotary analyzer were: 4 minutes at 95°C, 40 cycles x 20 seconds at 95°C, 30 seconds at 58°C and 30 seconds at 72°C followed by the dissociation protocol at 72°C (*38*) *(Table 2-3).* Results were analyzed using the comparative C_T method. The C_T mean and standard deviation of each technical replicate was calculated and the mean C_T values were then normalized to the 18S mean C_T value. Between two C_T values there is a twofold exponential difference in amplification.

Preliminary Intersection of in house and public ENCODE data

In order to assess the importance of studying NF-κB we used previously generated data, which consisted of samples from HapMap YRI and CEU populations. Mapping differential AE was determined by using Illumina HumanOmni1 BeadChips (Illumina Inc., SanDiego, CA, US), which generated AE and genotyping data. *Cis*-regulatory SNPs showing significant association were mapped within 250kb of genes exhibiting differential AE. The list of candidate *cis*-rSNPs, as well as a MAF-matched control data set, were intersected with publicly available whole-genome functional data (ChIP experiments) from the ENCODE project. This was done in order to observe if there was a significant difference in TF binding in regions with *cis*-rSNPs versus control SNPs. Discussed in the results section below are graphs based on these analyses. Data from the ENCODE NF-κB ChIP-seq peaks for LCLs were investigated further. From our original list of ranked candidate *cis*-rSNPs, we examined the top three SNPs (by p-value) for each locus in both the CEU/YRI populations that overlap regions of NF-κB binding. In addition, we observed the number of our genetically regulated loci that have a top SNP overlapping an NF-κB ChIP-seq peak. We further investigated loci from the ENCODE NF-κB ChIP-seq peaks for a subset of samples induced with TNF- α. Further analysis was done for the subset of loci, which have a top 3 SNP overlapping an NF-κB ChIP-seq peak for samples induced with TNF- α. Each gene was examined using a literature search for biological or biomedical relevance such as implications in the NF-κB pathway, in inflammation and/or autoimmune disease. Candidate *cis*-rSNPs versus control SNPs were also studied to determine in which chromatin states (e.g., enhancer, strong promoter, weak promoter). SNPs were most likely found (*85*).

Perturbation of NF-κB

The next step of this project was to perturb NF-κB in order to observe the effect genome-wide *(Figure 2-1)*. The LCLs used included two HapMap trios one from CEU (GM12891, GM12892, GM12878) and one from YRI (GM19239, GM19238, GM19240). To gain the ability to inhibit NF-κB in LCLs, the protocol had to be optimized based on methods and reagents previously used in the literature. Cells were plated in 6-well plates one day prior to the experiment. 500, 000 cells/ mL in 2mLs was used based on the growing conditions required for lymphoblasts which requires >300,000 cells/mL for proper growth. Cells were primarily transfected for one hour with a cell permeable small molecular compound, Helenalin (5uM) (EMD Chemicals, USA) in order to inhibit the activation of NF-κB (p65). Helenalin is a sesquiterpene lactone that acts as a specific NF-κB DNA binding inhibitor by irreversibly alkylating free sulfhydruls of the cysteine residues on the p65 subunit. Importantly, this compound exhibits no effect against cellular NF-κB activation, nuclear translocation or IκB dissociation/degradation (*86*). Following this inhibition, cells were stimulated with TNF- α (3ng/ul) at time points consisting of 4, 6, 8, 12, 24 and 48 hours in order to select for the ideal time point. Validation of the perturbation of NF-κB and induction by TNF- α was done by RT-PCR for genes targeted by NF-κB including IL-6, IL-8, IL-1a, and Bcl-2. The most optimal time point to stop the experiment was deemed 8 hours post- transfection *(Figure 3-3).*

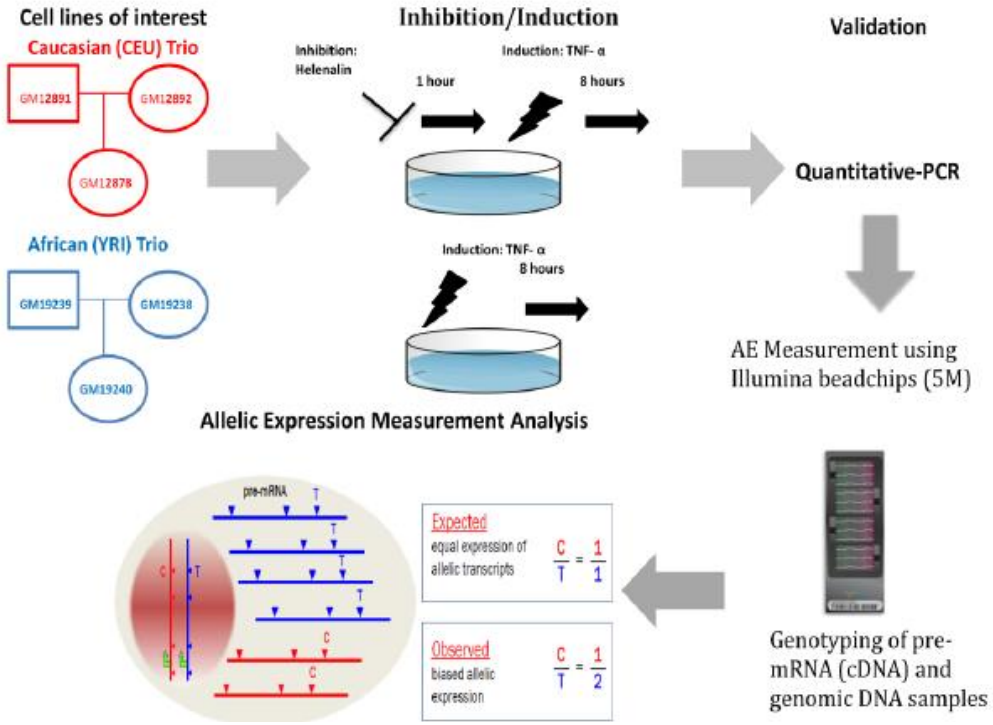


Figure 2-1. Experimental approach to perturb NF-κB and genome-wide AE assessment. Two trios from the HapMap CEU and YRI populations were used to perturb NF-κB using an inhibitor of NF-κB, Helenalin for 1 hour coupled to TNF- α induction for 8 hours. Validation of NF-κB perturbation was done using RT-PCR on known NF-κB gene targets. Samples were assessed on Illumina HumanOmni5-Quad BeadChips in order to analyze differential AE genome-wide.

Genotyping and Genome-wide AE Assessment (NF-κB samples)

The aforementioned inhibition of NF-κB was used to analyze differential AE. We were interested in AE differences in cells induced by TNF- α versus those induced by TNF- α and have inhibition of NF-κB. Genotyping and AE analysis were done using Illumina HumanOmni5-Quad BeadChip (Illumina Inc., SanDiego, CA, US) as per the manufacturer's instructors. Approximately 100ng/ul of gDNA and 50-300ng double-stranded cDNA was used for genotyping. Genotypes in gDNA were extracted using BeadStudio. The parallel assessment and normalization of gDNA and cDNA heterozygote ratios were carried out as described earlier (*38*) by genotyping of pre-mRNA (cDNA) and gDNA samples. Illumina HumanOmni5-Quad BeadChips use powerful tagSNPs from the International HapMap and 1000 Genomes Project that target common genetic variation down to approximately 1% MAF. Gene density of the SNPs on this BeadChip is approximately 100 SNPs/ RefSeq gene region (including 10Kb surrounding the gene) (http://www.illumina.com). Consequently, the redundancy of independent data in a given RefSeq gene enables us to show similar allelic deviation for different SNPs and this is important in order to build confidence in our results.

Bioinformatics Analysis of Genome-wide AE Assessment

Normalized gDNA and cDNA heterozygote ratios from the Illumina HumanOmni5-Quad BeadChip (Illumina Inc., SanDiego, CA, US) were generated as output for analysis. The overview of the bioinformatics approach is as follows and will be described in more detail in the results section. Our data set, which consisted of loci associated to top 10 *cis*-rSNPs that showed diminished AE upon perturbation of NF-κB were overlapped with

publicly available data on the UCSC Genome Browser. Functional data consisted of ENCODE ChIP-seq peaks and TRANSFAC binding sites for NF-κB and cooperative TFs of NF-κB (E2A, STAT1, IRF2, EBF1) *(Figure 2-2).* Further informatics analysis was done as per the below.

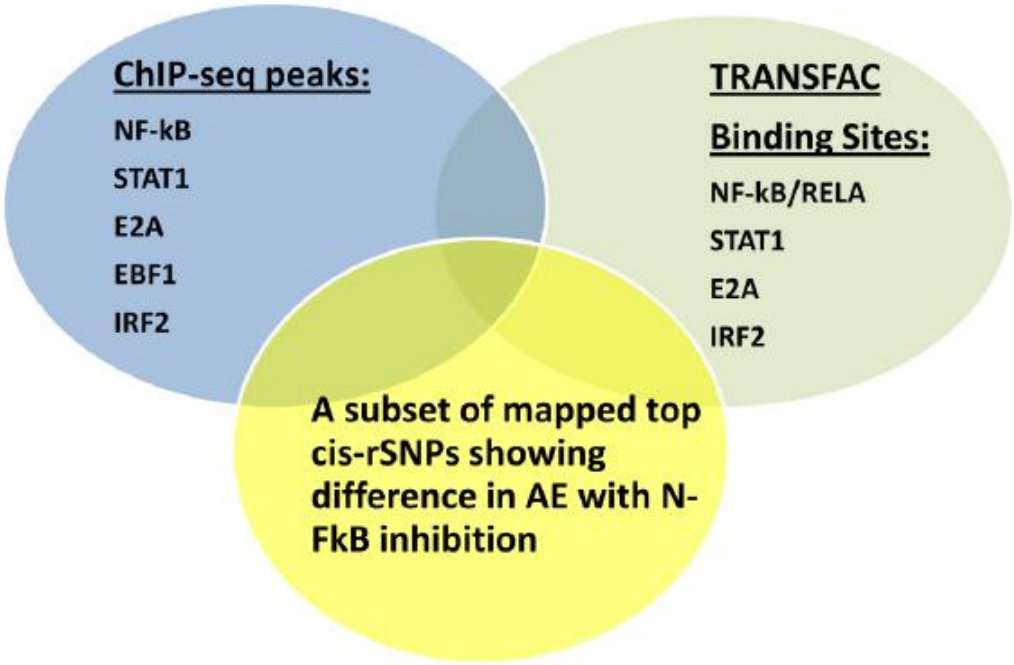


Figure 2-2. Overlapping data sets with the ENCODE ChIP-seq peaks and TRANSFAC binding sites for NF-κB and cooperative TFs. A subset of mapped *cis*-rSNPs, which showed diminished differential AE in inhibited versus induced NF-κB samples for associated loci were overlapped with publicly available data. This data includes ChIP-seq peaks and TRANSFAC binding sites for NF-κB, STAT1, E2A, EBF1, and IRF2.

Relative SNP distribution

Due to the fact that the NF-κB motif was most often found enriched in active promoter and strong enhancer elements we decided to investigate the relative distribution of SNPs, which may cause differential binding of NF-κB (*85*). We analyzed the relative distribution within the genome for the top 3 heterozygous *cis*-rSNPs for the CEU and YRI population combined. The SNPs were normalized against a reference consisting of the total SNPs found in the region of interest.

Gene network and pathway analysis

In order to visualize our data in the context of biological networks, functions or pathways the data was analyzed through the use of Ingenuity Pathway Analysis (IPA) system (Ingenuity Systems, Mountain View, CA, USA, www.ingenuity.com). The datasets containing differentially expressed genes with NF-κB inhibition coupled to induction with TNF- α compared to TNF- α induction only were uploaded to the application. The reference data set consisted of all 1753 genes from our mapped list of *cis*-regulatory variation for LCLs, which did not show differential AE. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base. These genes, called focus genes, were overlaid onto a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base. Furthermore, networks of the focus genes were subsequently generated based on an algorithm, which takes into account their connectivity. The functional analysis identified the biological functions that were most significant to the uploaded dataset. Genes from the dataset that met the required cutoff and were also associated with biological functions in the Ingenuity Pathway Knowledge Base were considered for the analysis. Finally, a score for

the network was given. The score was the negative exponent of the right tail of the Fisher's exact test, which calculates a p-value determining the probability that each biological function assigned to the dataset is only due to random chance.

GWAS

To assess the effect that mapped *cis*-rSNPs have for individual disease and trait associated SNPs we did the following analysis. Top 3 heterozygous SNPs associated to loci, which showed a change in AE for samples induced with TNF- α versus those with NF-κB inhibition were assessed for disease SNPs previously shown in the literature. This was done using a catalog of published GWASs (http://www.genome.gov/gwastudies/).

Methods Specific to the Case Study: SNAI1

Selection of Cell Lines

Cell lines included in the SNAI1 case study GM2317 and WG1657 were originally obtained from Coriell (http://www.coriell.org/). Primary cell cultures from human trabecular bone from the proximal femoral shaft, such as HOB642, were obtained from Uppsala University Hospital, Uppsala, Sweden from patients undergoing total hip or knee replacement. The cell lines were included as they were heterozygous for the SNP of interest, *rs6684375* in order to investigate the association between SNAI1 and *WNT4.*

Discovery of cis-rSNP

We possessed a great deal of support from previous studies in our laboratory, which will be discussed in the results section, to support the association between SNAI1 and *WNT4*. We identified a SNP that affects *WNT4 cis*-regulation in our FB panel and alters the risk for osteoporosis (*87*). Osteoporosis is a skeletal disorder characterized by compromised

bone strength and increased risk of fracture in which the regulation of bone remodeling is imbalanced. Clinical diagnosis of osteoporosis as well as the assessment of risk fracture is done using a heritable complex trait, bone mineral density (BMD). Numerous loci contributing to BMD and osteoporosis risk have been recently described by GWAS (*87*). This includes four loci which encode members of the Wnt and RANK-RANKL signaling pathways(*87*). However, the underlying biological effect of many of these variants remains unknown. Elucidating these effects and uncovering the remaining genetic variation is critical to understanding this complex disease.

SNAI1 is known for its involvement in mesenchymal cell development (*88*). As such, we decided to employ a case study in order to evaluate the association of a TF, SNAI1 and a SNP > 200Kb upstream of the gene *WNT4*. We did this by implementing our previously described approach to perturb TFs in order to better understand their relationship to DNA binding. This was done by pursuing *in vivo* validation of SNAI1 binding in living cells by carrying out ChIP in addition to *SNAI1* knockdown by RNA interference (RNA*i*) while monitoring its consequences in *WNT4* AE phenotype, as well as, genome-wide AE assessment.

RNAi targeting the TF SNAI1

RNA*i* is a process used *in vivo* for mRNA degradation that is induced by dsRNA in a sequence–specific fashion. RNA*i* is a powerful technique that specifically silences the expression of any gene for which the sequence is available. It is an invaluable tool in the field of reverse genetics. Traditional *RNAi* methods involve synthetic RNA duplexes made up of two unmodified 21mer oligonucleotides annealed to form short/small

interfering RNAs; however, stealth RNA*i* (Invitrogen Corporation, Carlsbad, CA, USA) was used and improves upon the aforementioned technology by using proprietary chemical modifications to ensure better RNA*i* results. This technology is advantageous in the following ways: providing effective knockdown, higher specificity, greater stability and less cellular toxicity. These small interfering RNAs make use of an endogenous RNAi pathway, which includes the enzymes Dicer and Argonaute, as well as, RNA-induced silencing complex (RISC) (*89*).

The reagents used included stealth RNA*i* targeting *SNAI1*, stealth RNA*i* targeting DiGeorge Syndrome chromosome region 8 (*DGCR8) (negative control)* (Invitrogen Corporation, Carlsbad, CA, USA), Opti-MEM (Invitrogen Corporation, Carlsbad, CA, USA) and the transfecting reagent, Lipofectamine RNA*i*MAX (Invitrogen Corporation, Carlsbad, CA, USA). Various stealth RNA*i* were tested in order to optimize for the sequence with the greatest knockdown of *SNAI1*. The protocol was done as per the manufacturer's instructions for forward transfection with the necessary changes implemented per cell line. Optimization included determining a valid protocol in terms of number of cells used, reagents and the time course of the experiment. Optimizations were done in 24-well plates and scaled up accordingly. Stealth RNA*i* (100uM) was transfected and the experiments were stopped at three different time points 48, 72, and 96 hours. This experiment was replicated with immortalized osteoblast and FB cell lines heterozygous for our SNP of interest (*rs6684375).* After transfections were completed samples were analyzed by quantitative RT-PCR using the protocol mentioned above to validate *SNAI1* knockdown.

Genotyping and Genome-wide AE Assessment

The aforementioned transfections of stealth RNA*i* targeting *SNAI1* versus the negative control were used to analyze differential AE. Genome-wide AE assessment was done using Illumina HumanOmni 2.5-Quad BeadChips (Illumina Inc., SanDiego, CA, US) by genotyping of cDNA and gDNA samples for GM2317, WG1657 and HOB642 (*Figure 2-3).* We hypothesized that there would be measurable differences in AE in cells transfected with stealth RNA*i* versus the negative control, which is used to control for the toxicity of using transfection reagents. Data from the high throughput experiments were normalized as previously explained (*38*) and subsequent analysis was performed, which will be discussed in the results section.

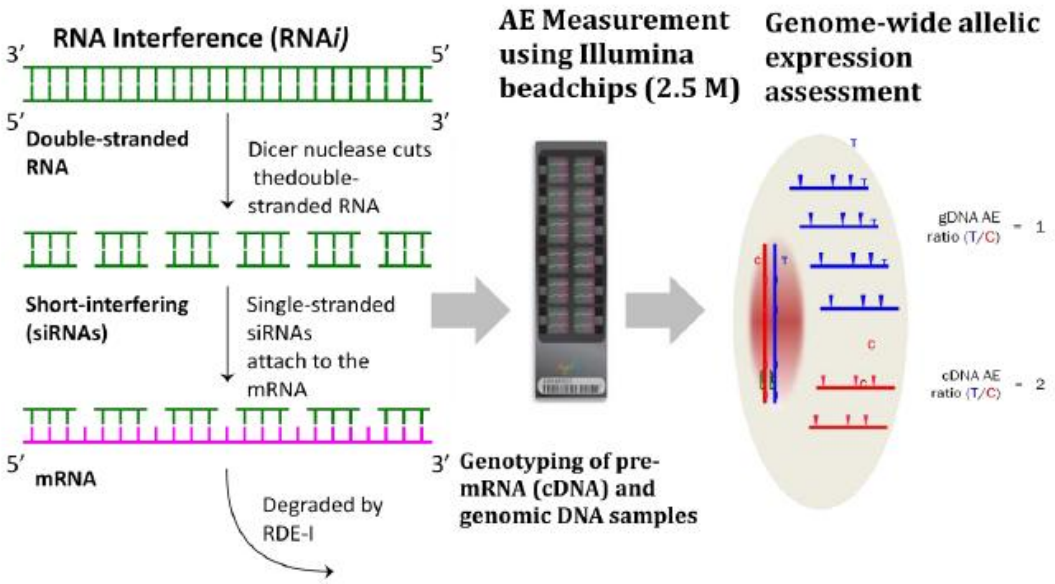


Figure 2-3. Schematic for genome-wide AE assessment (*SNAI1*). WG1657, GM02317, and HOB642 were transfected with RNA*i* targeting *SNAI1* and a negative control. CDNA and gDNA from the samples were analyzed on Illumina HumanOmni2.5-Quad BeadChips in order to assess AE genome-wide.

Primer design & ChIP-RT-PCR

The purpose of a ChIP experiment is to enrich for DNA fragments associated with a specific DNA-binding protein of interest (*90*). We used an in-house protocol in which optimizations were based on specific cell lines used. FB and osteoblast cells were grown in P15 dishes until 80 % confluent (10 dishes at a time) with 15 ml of media and subsequently cross-linked with 1% formaldehyde at room temperature for 10 minutes. After quenching with glycine for 5 minutes (125 mM glycine per mL of media), the cells were washed with ice-cold phosphate buffered saline (PBS). Cells were scraped with farnham lysis buffer (5mM PIPES pH8.0, 85mM KCl, 0.5% NP-40 and protease inhibitors), washed twice with PBS and subsequently collected after each wash by centrifugation at 2,000*g* for 5 minutes. Cell pellets were pooled into two cryotubes, flash frozen and stored at −80 °C. Frozen pellets were thawed and cells were lysed in farnham lysis buffer (5mM PIPES pH8.0, 85mM KCl, 0.5% NP-40 and protease inhibitors) for 10 minutes on ice. After centrifugation and wash with 1 mL of radioimmunoprecipitation assay buffer (RIPA) containing 50mM Tris HCl pH8, 150mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS and protease inhibitors, lysates were then diluted with 500 µl of RIPA to proceed to the sonication step. Cells were sonicated in non-stick tubes under conditions optimized to yield soluble chromatin fragments in a size range of 150 to 300 bp. Chromatin was sonicated for 5 minutes using a Branson 250 sonicator at 20% power amplitude (pulses: 10 s on and 30 s off). Lysates were cleared by centrifuging at 12,000g for 10 minutes at 4 °C to eliminate cellular debris. Chromatin was then flash frozen and stored at -80 °C or used immediately for the next step. Before each immunoprecipitation, the chromatin samples were pre-cleared with 50 µl of prewashed ProteinA-magnetic beads (Invitrogen; 100-02D) to avoid non-specific binding.

Immunoprecipitation was carried out for 12 hours by rotation at 4 °C in 500 µl of chromatin/RIPA buffer supplemented with protease inhibitor cocktails (Roche; 04 693 159 001) and phenylmethylsulfonyl fluoride (PMSF). We used 10 to 100 million cells and 2 to 20 µg of the following antibodies for each assay: SNAI1 (Abcam, Cambridge, MA, 85931), H3K4me1 (Abcam, Cambridge, MA; ab8895), and Normal Immunoglobulin G (IgG) (Cell Signaling Technology, Danvers, MA; #2729). After overnight incubation, samples were rotated with 100 µl of prewashed ProteinA-magnetic beads at 4 °C for 1 hour. The beads were then collected by brief centrifugation at 2,000*g* following by the use of a magnetic rack. Beads were washed five times with 1 mL of LiCl wash buffer (100mM Tris pH7.5, 500mM LiCl, 1% NP-40, 1% sodium deoxycholate) by resuspending the beads and keeping them on ice for 5 minutes. Bound chromatin was then eluted from the beads using 200 µl of elution buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, 1.0% SDS) by incubation at 65 °C for 1 hour with a vortex every 15 minutes. This was followed by centrifugation at 14,000*g* at room temperature for 3 minutes. The eluted chromatin and the 'input' samples were then incubated at 65 °C overnight with 0.2M of a 5M NaCl to remove the crosslinks. Samples were then treated with RNase at 37 °C for 30 minutes and digested with proteinase K at 55 °C for 1 hour. Immunoprecipitated DNA was then purified using QIAquick PCR Purification Kit (QIAGEN, Mississauga, Ontario) and eluted in 30 µl of elution buffer. Enrichments of relevant regions were validated using RT-PCR experiments for each antibody *(Figure 2-4).*

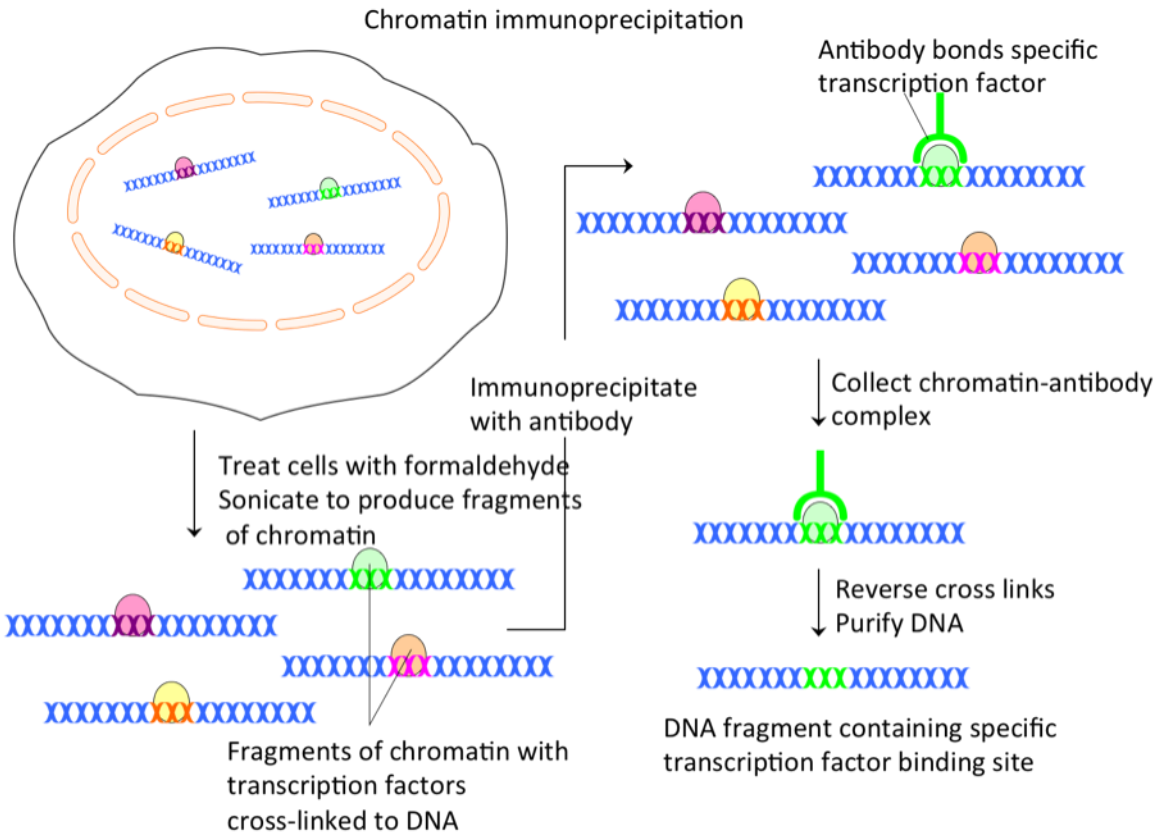


Figure 2-4. Schematic of ChIP. The purpose of ChIP experiments for DNA-binding proteins is to enrich for DNA fragments associated with a specific protein. Firstly, the DNA-binding protein is cross-linked to DNA *in vivo* by treating the cells with formaldehyde. The chromatin is then sheared by sonication into small fragments of approximately 150-300bp. Subsequently, the cross-links are reversed and the released DNA is assessed to determine the sequence that is bound by the protein.

We included cell lines WG1657 and HOB642 in the ChIP experiments as these cell lines were heterozygous for the SNP of interest, *rs6684375*. The antibodies used included SNAI1, IgG *(negative control)* and H3K4me1 *(positive control).* Primers for ChIP- RT-PCR were designed using primer3 (http://frodo.wi.mit.edu/primer3/) with a desired PCR amplicon length between 150 bp and 250bp *(see Table 2-1).* Two primers used which target amplification of the SNAI1 binding site *rs6684375,* as well as, *SYNGR,* which is a common region enriched using the antibody H3K4me1. Also, the selected region, *SYNGR* was used as a control locus since it did not show any SNAI1 binding. All primers were confirmed with BLAT (*S9*) (http://genome.ucsc.edu/) to avoid common SNPs, which could influence primer hybridization. The primers were tested to yield a unique product using insilico PCR (http://genome.ucsc.edu/). We determined the relative enrichment for each target locus compared to a reference locus using the Delta-Delta-Ct-Method (2-ΔΔCt). Results were analyzed using the comparative C_T method as described above and compared to the regions amplified by the IgG pull down (*38*).

CHAPTER 3: RESULTS

Quantitative AE Measurements and Mapping

Genome-wide quantitative AE measurements were employed on Human1M-Duo BeadChips (*17*). This assay is based on the quantitative assessment of allele ratios in expressed heterozygous SNPs in RNA (cDNA), which are normalized to corresponding gDNA heterozygote ratios. The primary discovery panel for this study consisted of 53 unrelated HapMap CEU LCLs. As before, we used both intronic and exonic SNPs, which passed the signal intensity threshold. Intronic SNPs present in primary transcripts supplied much of the information used. Analysis of differences in cDNA allele ratios were restricted to heterozygous sites averaged across full annotated primary transcripts after allele- ratio normalization (*38*). Phased genotypes were then used in association analysis with AE quantitative trait. Our aim was to increase specificity for detection of allelic differences impacting full transcripts, rather than other allelic differences observable in cases of differential splicing or 3' usage (*17*). Moreover, we included individuals from the YRI population to elucidate more *cis*-variants and to aid in fine-mapping of variants common to both CEU and YRI populations. The final panels we used consisted of 55 and 63 HapMap CEU and YRI LCLs, respectively. In addition a number of children not used in population mapping were applied in validation tests. Similar AE assessments were also carried out on 70 primary FB cell lines (Caucasian trios) in order to map *cis*-regulatory variants. A comprehensive list of >4000 transcripts mapped to top associated SNPs was generated for FBs and was used in the below analyses. However, the independent validations done in LCLs as described below were not completed for FBs.

Independent Validation of Associated Loci in LCLs

Our approach to isolate the maximum number of *cis*-regulated full-length transcripts for LCLs applied the power of AE mapping in primary transcripts. This was followed by replication in full length mature transcripts using total expression levels across all exons, generated independently by exon-arrays (*40, 91*) or RNA-seq (*15, 16*). Subsequently, we used independent AE tests that were not included in the LCL population AE mapping to validate our list of associations. A locus was defined as validated if it contained at least 1 heterozygous individual for a top SNP with differential AE of the associated transcript. The addition of these consecutive validations allowed us to establish a core set of high-confidence genetically regulated allelicly expressed transcripts. Overall, we mapped top associated SNPs to 1753 transcripts with significant AE in the YRI and/or CEU LCL populations.

Preliminary Validation of *cis*-rSNPs Affecting NF-κB Binding

Results from candidate *cis*-rSNPs showing significant association, which were mapped within 250kb of genes exhibiting differential AE for the LCL populations are described below. The list of the top LCL ranked candidate *cis*-rSNPs (by p-value) per loci, as well as, a MAF-matched control data set, were intersected with publicly available whole-genome functional data (ChIP experiments) from the ENCODE project. It was observed that there is a significant difference in TF binding for NF-κB in regions with top *cis*-rSNPs versus control SNPs (*Figure 3-1) (Table 3-1).* In order to increase the likelihood of finding the causal variants we investigated top 3 SNPs, which are highly associated to the top SNP due to linkage disequilibrium (LD), also termed, "allelic association". LD

arises due to selection or population history, which causes recombination to occur and in turn breaks down ancestral haplotypes. Haplotypes are the combinations of alleles, which are observed in a population. LD is the non-random association of alleles at two or more loci (*92*). As such, LD refers to the correlation among neighboring alleles, which reflects haplotypes descended from single, ancestral chromosomes (*93*).

A significant number of top 3 SNPs overlapped at least one ENCODE NF-κB ChIP-seq peak. We observed that 9.3% of top 3 SNPs overlapped at least one ENCODE NF-κB ChIP-seq peak (*Table 3-1*). In addition, 39.0% of loci contained a top 3 ranked SNP which overlapped at least one ENCODE NF-κB peak. A subset of loci were further investigated, which include ENCODE NF-κB ChIP-seq samples induced with TNF- α versus control samples. We observed that 91.5% of loci induced with TNF- α had a top SNP associated (*Table 3-2*). Results from these analyses suggest that there is a strong genetic component of allele-specific TF binding for NF-κB. Moreover, chromatin states based on the Bernstein Hidden Markov Model (HMM) classifications were downloaded for regions in which top 3 SNPs overlapped an NF-κB ChIP-seq peak as well as for control SNPs for samples induced with TNF- α. We observed similar results as seen in the Ernst *et al.*, 2011 article in which NF-κB binding sites for *cis*-rSNPs were significantly enriched compared to control SNPs in active promoter (Chi-squared test, pv= 0.004) and strong enhancer elements (Chi-squared test, pv= 0.0001) (*85*) *(Figure 3-2).* Finally, a literature search elucidated several loci, which have been associated to the NF-κB pathway and/or immune related diseases *(Table 3-3).*

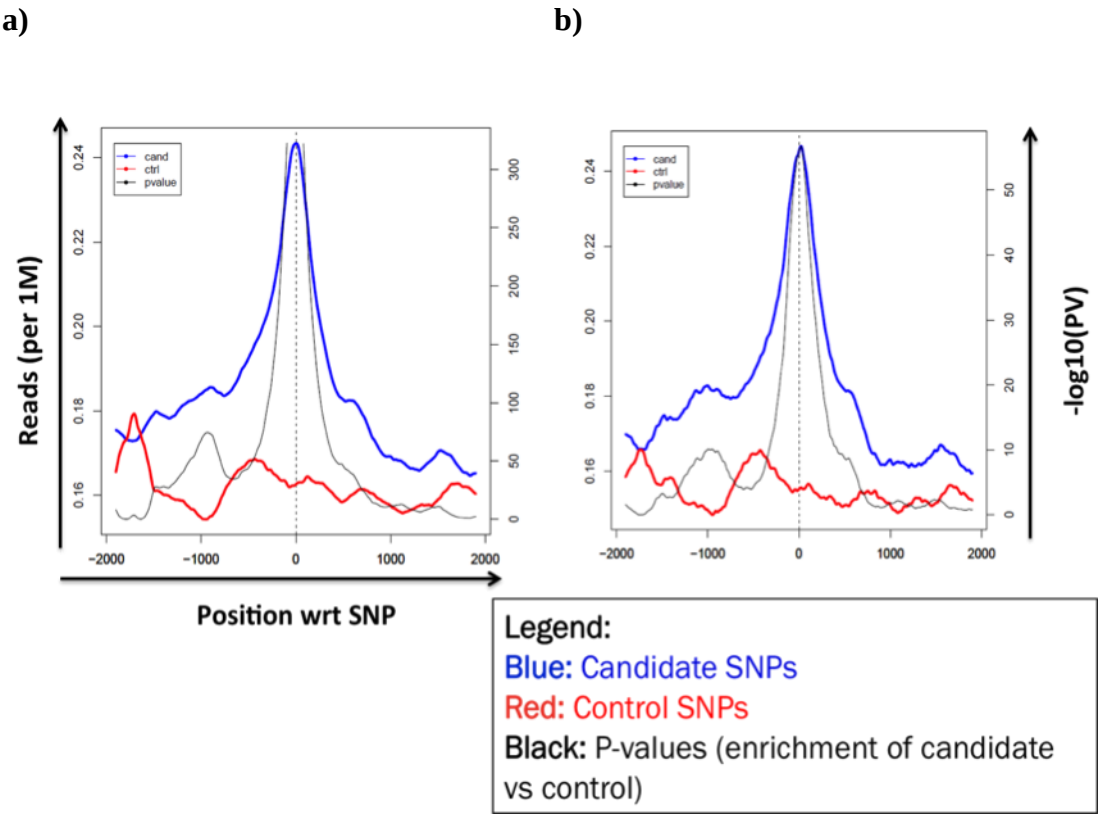


Figure 3-1. TF binding for NF-κB in regions with *cis*-rSNPs versus control SNPs. Distributions of reads density (1M) +/-2000bp around candidate (blue lines) and control sites (red lines) as well as significance of their differences (dark lines). **a)** LCL-specific NF-κB ChIP-seq data was analyzed. Regions surrounding candidate *cis*-rSNPs are significantly enriched in NF-κB binding sites versus matched controls **b)** GM12878 NF-κB ChIP-seq from samples induced with TNF-α. Similar results as seen in **a).**

Total unique top 3 ranked SNPs	11,505
Top 3 ranked SNPs overlapping an ENCODE NF-κB peak	1073/11,505 (9.3%)
Total unique regulatory loci/ genes	1,753
Loci with top 3 ranked SNPs overlapping at least one ENCODE NF-κB peak	684/1,753(39.0%)

Table 3-1. Summary of total data from ENCODE NF-κB ChIP-seq peaks. Data was extrapolated for top 3 SNPs (by p-value) from the original list of ranked candidate *cis*-rSNPs for each locus using CEU and YRI populations. From the data, 39.0% of top 3 SNPs overlap at least one ENCODE NF-κB ChIP-seq peak.

Total loci found	47
Loci with top 3 ranked SNPs overlapping at least one ENCODE NF-κB peak	43/47 (91.5%)

Table 3-2. Summary of data from the ENCODE NF-κB ChIP-seq peaks for samples induced with TNF- α. A subset of loci from the ENCODE NF-κB ChIP-seq peaks were further assessed, which consisted of 47 loci. Data was extrapolated for samples induced with TNF-α versus control samples for GM12878. 91.5% of the loci induced with TNF- α versus control samples have a top 3 SNP, which overlaps at least one ENCODE NF-κB ChIP-seq peak.

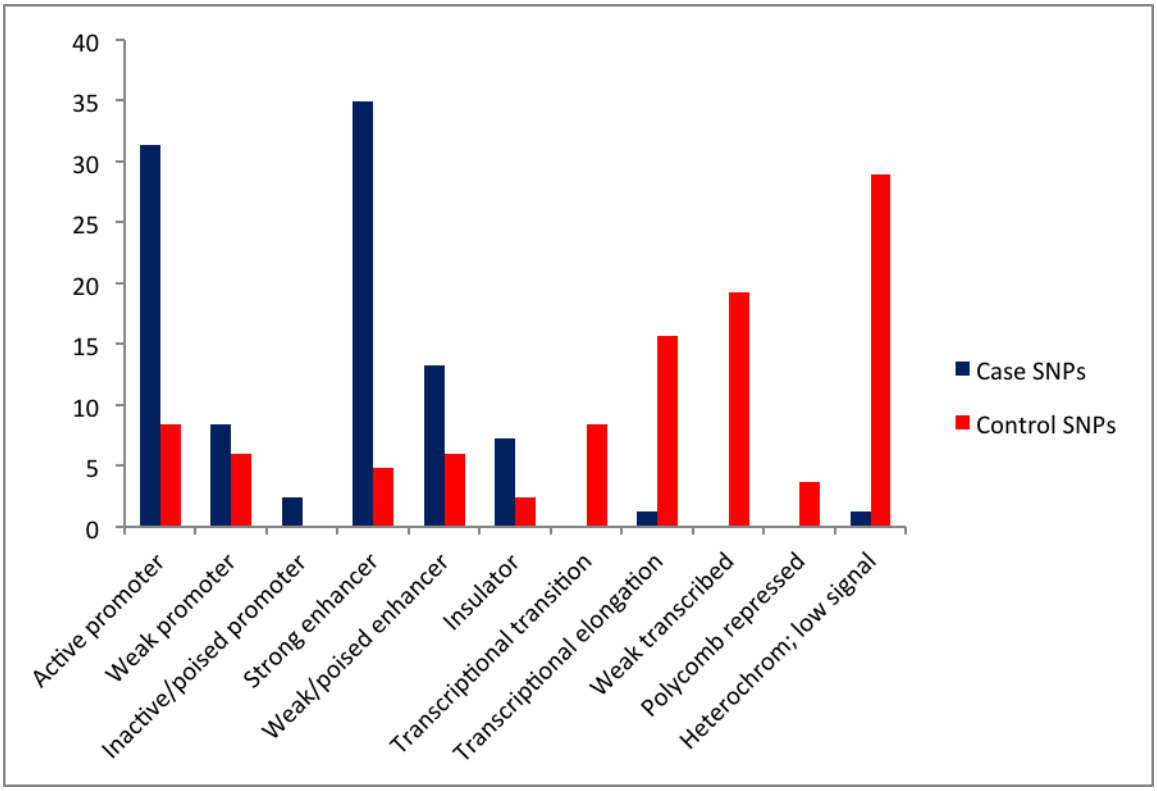


Figure 3-2. Chromatin classification for case *cis*-rSNPs and matched control SNPs overlapping NF-kB ChIP-seq peaks induced with TNF- α. Regions that contain a top 3 SNP overlapping an NF-κB ChIP- seq peak as well as control SNPs for a subset of samples induced with TNF- α were assessed. Bernstein HMM chromatin classifications were given for each region in which SNP was found. Results indicate a strong binding bias of *cis*-rSNPs towards regulatory regions such as active promoter and strong enhancer elements. In contrast, for control SNPs results indicate a depletion of SNPs in regulatory regions such as active promoter and strong enhancer elements. Furthermore, there is a binding bias of control SNPs in inactive heterochromatin regions.

Relevant Loci	Explanation
Oxidative stress-responsive 1 protein (*OXSR1*)	Target gene and activator of NF-κB (*94*)
Annexin A4 (ANXA4)	Interacts with the NF-κB subunit and modulates NF-κB transcriptional activity in a Ca^{2+}-dependent manner (*95*)
α/ β hydrolase domain-containing protein 5 (ABHD5)	Generates signaling lipids that regulate the balance between systemic inflammation and insulin action (*96*)
Hypoxia-inducible factor 1, alpha subunit (*HIF-α*) inhibitor (HIF1AN)	HIF-α is a subunit of the hypoxia inducible TF. Hypoxia is a regulator of angiogenesis and inflammation in rheumatoid arthritis. *HIF1AN* inhibits the alpha subunit of hypoxia factors (*97*).

Table 3-4. Subset of loci implicated in the NF-κB pathway or range of immune related diseases. A literature review using PubMed was done in order to determine loci which have been previously associated to the NF-κB pathway or immune related diseases. The subset of loci seen above consist of those from the preliminary analysis, which have a top 3 SNP overlapping an NF-κB ChIP-seq peak for samples induced with TNF- α (ENCODE experiment).

Validation of NF-κB Perturbation

Validation of NF-κB knockdown was widely done in the literature using RT-PCR of known gene targets. Results show significant changes in known targets of NF-κB for samples induced by TNF- α versus those induced by TNF- α coupled to inhibition of NF-κB *(Figure 3-3).*

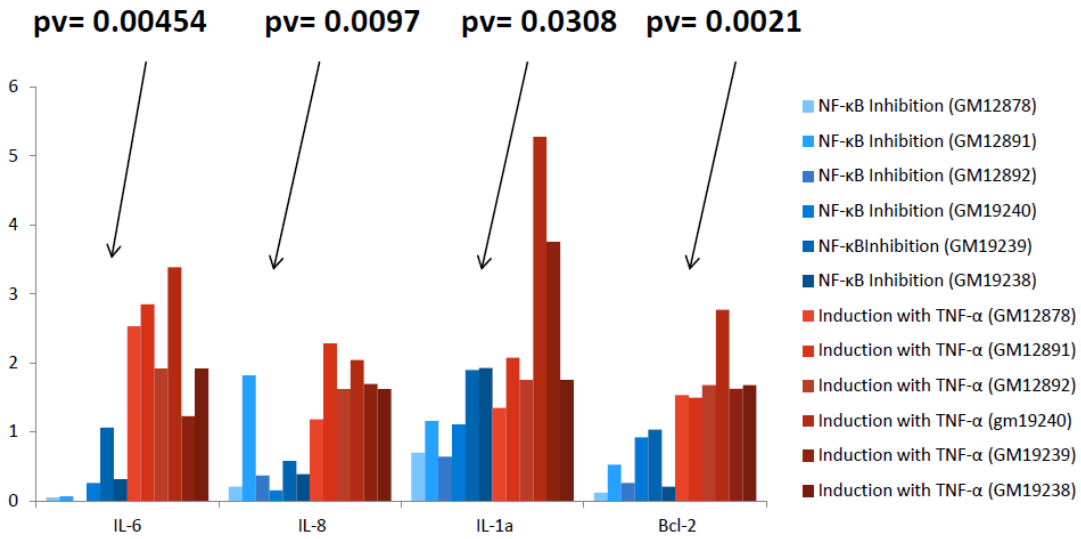


Figure 3-3. Validation of inhibition of NF-κB. Significant changes upon perturbation of NF-κB on known gene targets include Il-6 (p= 0.00454), Il-8 (p=0.0097), Il-1a (p=0.0308), and Bcl-2 (p=0.0021). P-values were calculated using a one-tail Fisher's exact test.

Genome-wide AE Analysis of Illumina HumanOmni5-Quad BeadChips

As described above, we had AE data from Illumina HumanOmni1 BeadChips, which was used for mapping *cis*-regulatory variants for 1753 loci in LCLs. This data was merged with mean AI values for each transcript from the Illumina HumanOmni5-Quad BeadChips, which was generated for each experimental condition. In order to obtain a manageable high confidence list of transcripts the following filters were implemented. The first criterion was to restrict loci to those showing an AE change of greater than the threshold of 0.05 from the Illumina HumanOmni1 BeadChip data (corresponding to 1.2-fold difference in expression between alleles) (*38*). Subsequently, loci were analyzed only if AE was diminished in samples in which NF-κB was inhibited in comparison to induction by TNF- α. In addition, mean AI values were compared and only loci with greater than 3 SNPs showing an AI change were retained. The aforementioned approach to perturb the TF, NF-κB and monitor consequences of the perturbation genome-wide can be generically extended to other TFs in the literature. Consequently, a specific case study was done targeting SNAI1 for knockdown in FB cells. This was done in order to test a plausible hypothesis between the association of SNAI1 and *WNT4,* as well as, observe the effect of SNAI1 perturbation on a genome-wide scale. We were also able to compare results of the perturbation of NF-κB and SNAI1 in LCLs and FBs, respectively. The results from the aforementioned analyses will be discussed below.

Bioinformatics Approach

Enrichment of top cis-rSNPs in ENCODE NF-κB ChIP-seq peaks

We intersected mapped top candidate *cis*-rSNPs (rank 1) detected in LCLs, as well as, matched control SNPs from HapMap YRI and CEU populations with publicly available NF-κB ChIP-seq experiments from the ENCODE project. Moreover, we also overlapped SNPs associated with the subset of loci showing differential AE upon perturbation of NF-κB. The mapped top *cis*-rSNPs described above were enriched in NF-κB binding sites versus the control SNPs. However, the subset of *cis*-rSNPs, which showed a difference in AE upon perturbation of NF-κB experiments showed a significant enrichment in NF-κB binding sites relative to all top mapped *cis-r*SNPs (Chi-squared test, pv = 1.0E-08). Similar results were observed overlapping the subset of *cis*-rSNPs and mapped top *cis*-rSNPs with NF-κB LCL-specific ChIP-seq experiments using only samples induced with TNF- α (Chi-squared test, pv= 9.2e-06) *(Figure 3-4*).

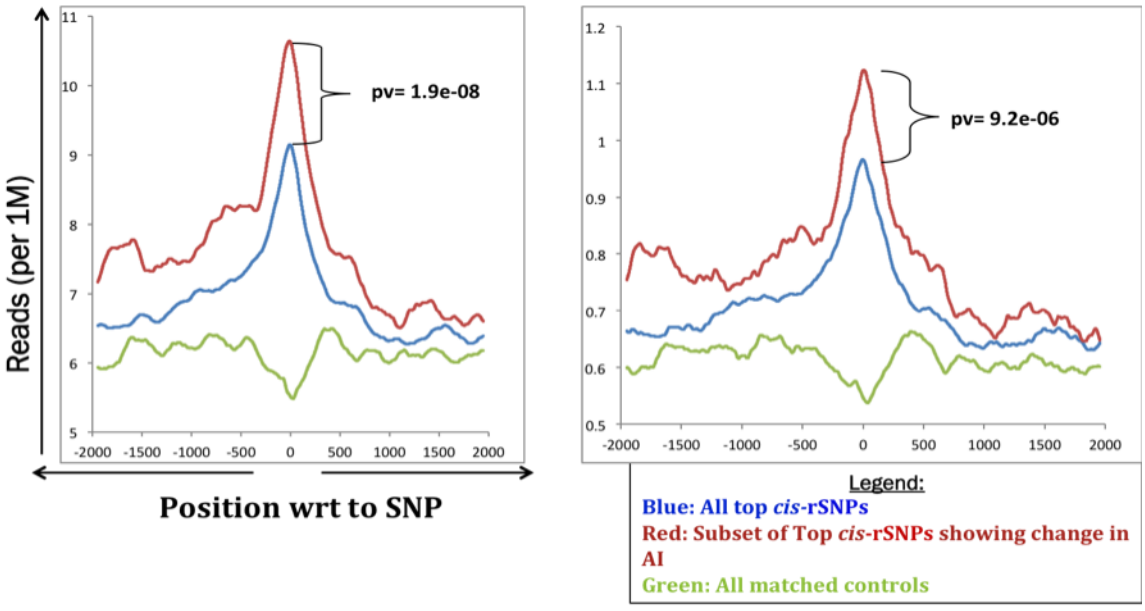


Figure 3-4. Enrichment of the top *cis-r*SNPs in ENCODE NF-κB ChIP-seq peaks. Significant enrichment of the top *cis*-rSNPs at NF-κB ChIP-seq peaks (Chi-squared test, pv =1.9e-08) *(left)*, as well as, NF-κB ChIP-seq peaks in which cells were treated with TNF-α (Chi-squared test, pv = 9.2e-06) *(right)*. P-values were calculated using the chi-squared test using the counts at the SNP position. The p-values show the significance of enrichment between the subset of top *cis*-rSNPs versus all top 10 mapped *cis*-rSNPs.

Analysis of cooperative action of TFs

Further analysis was done in order to understand differential AE observed upon perturbation of NF-κB. Bioinformatics analysis was performed to identify SNPs essential for NF-κB binding. The factors discussed (IRF2, STAT1, E2A and EBF1) that have been shown to work cooperatively with NF-κB were analyzed in conjunction with NF-κB. This was done using publicly available ChIP-seq data on the UCSC Genome Browser, as well as, binding motifs using the TRANSFAC database. The above was performed for top SNPs, as well as, top 3 SNPs associated to loci showing diminished AE upon inhibition of NF-κB. Therefore, using the TRANSFAC database, we searched for position weight matrix (PWM) motifs for NF-κB and known cooperative TFs that contained SNPs associated with the aforementioned loci. There were 631 unique loci, which showed differential AE relative to all mapping done in LCLs. Studying the top associated heterozygous SNPs associated to loci that had a change in AE at greater than 3 SNPs resulted in 581 cases. 210 cases contained a top SNP (rank 1), which overlapped with a ChIP-seq peak and/or disruption of a TRANSFAC binding site (using a window of SNP+/- 10bp) for NF-κB or one of the cooperative TFs. The same analysis was done for control SNPs and this yielded 29 loci associated to a control SNP, which overlapped with a ChIP-seq peak and/or disruption of a TRANSFAC binding site. There was significant enrichment of over 7-fold (Chi-squared test, p-value = 1.25E-26) between case and control heterozygous SNPs, which overlapped a functional element, such that 24.4% overlapped a functional element for NF-κB only. Functional analysis taking into consideration known cooperative TFs explained over 36% of loci harboring changes in differential AE upon perturbation.

Similar analysis was done for the most strongly associated candidate *cis*-rSNPs (rank 1, 2 and 3). This yielded 369 and 85 loci, containing a case or control SNP, respectively, which overlapped with a ChIP-seq peak and/or disruption of TRANSFAC binding site (using a window of SNP+/- 10bp) for NF-κB or one of the cooperative TFs. Bioinformatics analysis of top 3 SNPs, described above, revealed a significant difference of ~5-fold (Chi-squared test, p-value = 3.13E-30) between case and control SNPs, such that 47.5% overlapped a functional element for NF-κB. Moreover, considering the overlap between top 3 heterozygous *cis*-rSNPs and LCL-specific TF ChIP-seq peaks or TRANSFAC binding sites for NF-κB, as well as, known cooperative TFs, 64% of loci had a top 3 heterozygous *cis*-rSNP found in a functional element (*Figure 3-5*) *(Table 3-4).*

In addition, over 85% of cases that have a top 3 SNP which overlaps an NF-κB ChIP-seq peak or TRANSFAC binding site also overlaps or lies within 400bp of a ChIP-seq peak or TRANSFAC binding site for the other known cooperative TFs. There are several specific examples from our data, which demonstrate the aforementioned and are displayed below *(Figure 3-6)* (*64*).

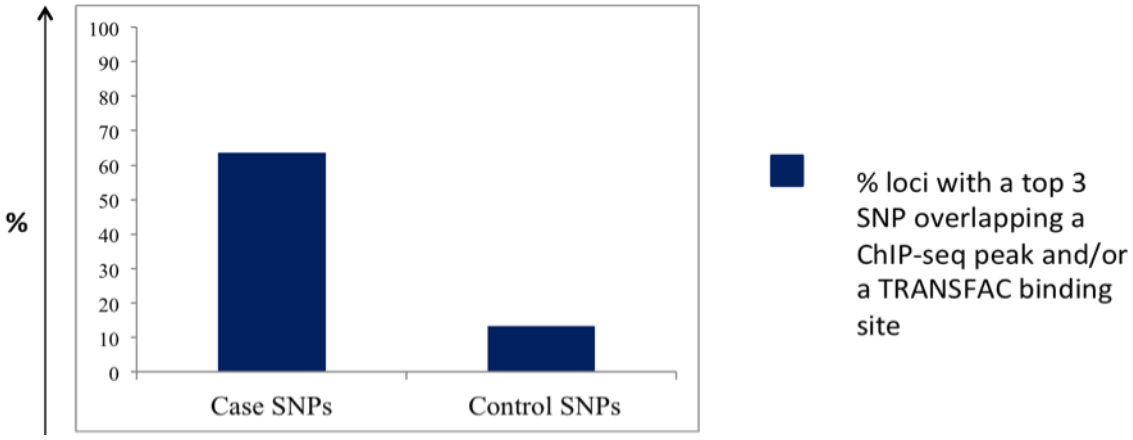


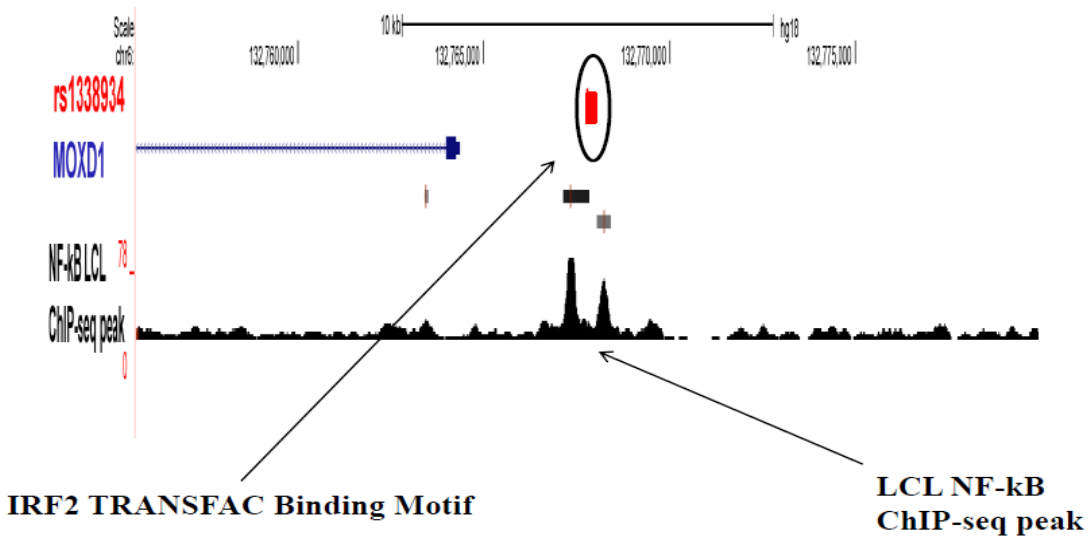
Figure 3-5. Comparison of top 3 heterozygous case and control SNPs. There were 581 loci, which showed differential AE relative to all mapping done in LCLs. Overlapping top 3 SNPs yielded 369 unique transcripts, which overlapped with a ChIP-seq peak and/or disruption of TRANSFAC binding site (using a window of SNP+/- 10bp) for NF-κB or one of the cooperative TFs. A similar analysis was done for control SNPs and this yielded 85 cases associated to a control SNP, which overlapped with a ChIP-seq peak and/or disruption of TRANSFAC binding site.

	Case SNPs	**Control SNPs (matched)**	**Fold difference (Case SNPs/Control SNPs)**	**Chi-squared test p-value**
Loci with a SNP that overlaps a functional element for NF-κB (top SNP)	24.4%	2%	~12	2.15e-20
Loci with a SNP that overlaps a functional element for a cooperative TF (top SNP)	11.6%	3%	~4	2.74e-09
Summary of loci explained by overlap with a functional element (top SNP)	36%	5%	~7	1.25e-26
Loci with a SNP that overlaps a functional element for NF-κB (top 3 SNP)	47.5%	8%	~6	2.09e-26
Loci with a SNP that overlaps a functional element for a cooperative TF (top 3 SNP)	16.5%	5%	~3	9.94e-07
Summary of loci explained by overlap with a functional element (top 3 SNP)	64%	13%	~5	3.13e-30
SNPs that overlap a TRANSFAC binding site for NF-κB (top SNP)	22 (corresponding to 21 unique loci)	6 (corresponding to 6 unique loci)	~4	0.007

SNPs that overlap a TRANSFAC binding site for a cooperative TF (top SNPs)	37 (corresponding to 28 unique loci)	6 (corresponding to 4 unique loci)	~6	6.14e-05
Summary of SNPs that overlap a TRANSFAC binding site (top SNPs)	59 (corresponding to 49 unique loci)	12 (corresponding to 10 unique loci)	~5	1.42e-06
SNPs that overlap a TRANSFAC binding site for NF-κB (top 3 SNP)	79 (corresponding to 70 unique loci)	24 (corresponding to 20 unique loci)	~3	6.62e-07
SNPs that overlap a TRANSFAC binding site for a cooperative TF (top 3 SNPs)	102 (corresponding to 70 unique loci)	30 (corresponding to 23 unique loci)	~3	4.54e-06
Summary of SNPs that overlap a TRANSFAC binding site (top 3 SNPs)	181 (corresponding to 140 unique loci)	54 (corresponding to 43 unique loci)	~3	4.04e-11

Table 3-4. Output of bioinformatics approach for loci of interest. Loci showing differential AE upon NF-κB perturbation associated to *cis*-rSNPs ranked by p-value were overlapped with functional data. Summary of overlap with ENCODE LCL ChIP-seq peaks and TRANSFAC binding sites for NF-κB and cooperative TFs is outlined.

a)



b)

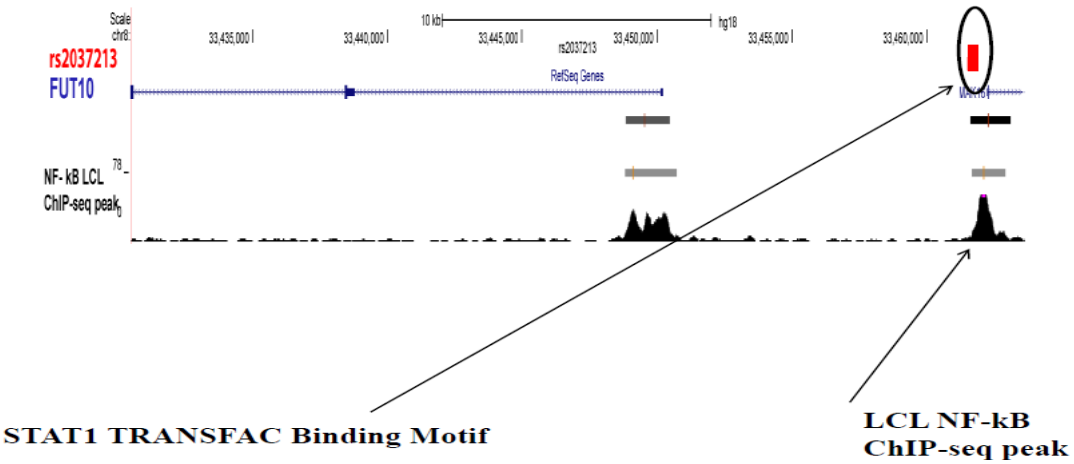


Figure 3-6. Top *cis-r*SNP overlaps functional data for NF-κB and cooperative TFs. a) A top *cis*-rSNP (*rs1338934)* associated to the transcript *MOXD1*overlaps both an LCL NF-κB peak and a TRANSFAC binding site for IRF2. The SNP is also upstream of the promoter region of the transcript *MOXD1*, which supports the notion of NF-κB as a regulatory TF. **b)** A top *cis*-rSNP (*rs2037213)* associated to the transcript *FUT10* overlaps a TRANSFAC binding site for the STAT1 motif and is located within a few Kb from an LCL specific NF-κB ChIP-seq peak. Also, the SNP is observed to be upstream of the promoter region of the transcript *FUT10,* which again supports the fact of NF-κB as a regulatory TF. The aforementioned examples fit the model of the cooperative action of TFs for NF-κB.

Heritable NF-κB -mediated AE

From the mapped list of *cis*-rSNPs there are several examples, which showed a decrease in AE upon inhibition of NF-κB in both the child and the parent compared to the samples induced with TNF-α. Overall, a transmission of AE from parent to child was observed in 54% of informative cases. In many instances, bioinformatics analysis of the SNPs suggested essentiality for NF-κB binding. The analysis consisted of assessing publicly available LCL-specific ChIP-seq peaks and TRANSFAC binding sites for NF-κB and cooperative TFs. An example of a transmission of differential AE from parent to child is illustrated below *(Figure 3-7, 3-8).*

Relative distribution of SNPs

The relative distribution within the genome for top 3 heterozygous *cis*-rSNPs in CEU and YRI populations combined was assessed. It was observed that the subset of top SNPs of interest was enriched in regulatory regions compared to all SNPs in that region. Regulatory regions include around the TSS, particularly 2Kb upstream *(Figure 3-9).*

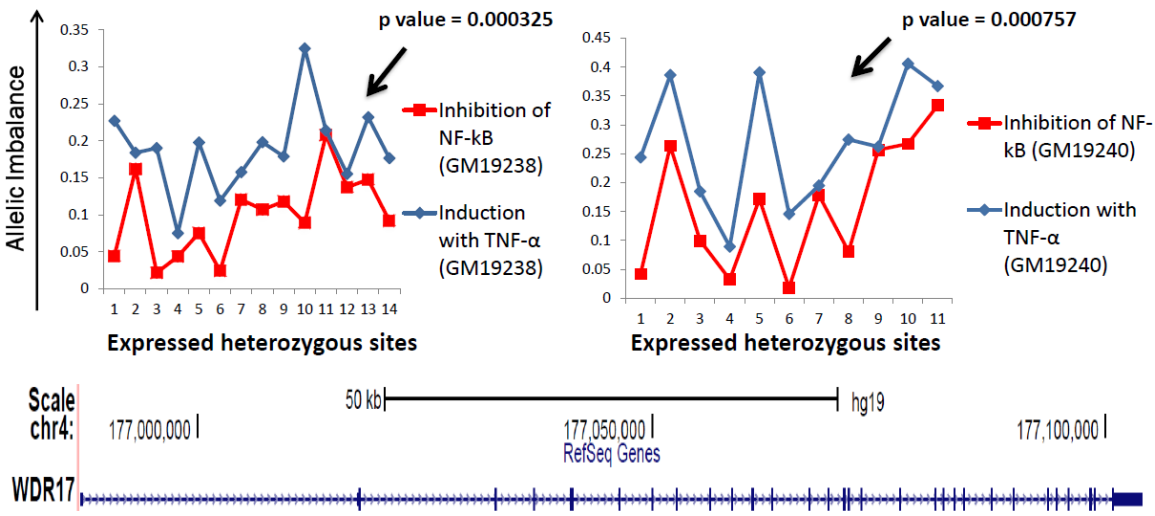


Figure 3-7. Example of NF-κB mediated AI in the African trio. Locus *WDR17* on chromosome 4 shows heritable NF-κB mediated AI for the African samples (GM19238 and 19240) upon perturbation of NF-κB. AI is diminished in samples in which NF-κB is inhibited versus stimulation with TNF-α. *Left:* AI is diminished for the locus *WDR17* in the parent (GM19238) in samples upon inhibition of NF-κB coupled to induction by TNF- α. *Right:* AI is diminished for locus *WDR17* for the child (GM19240) in samples upon inhibition of NF-κB coupled to induction by TNF- α

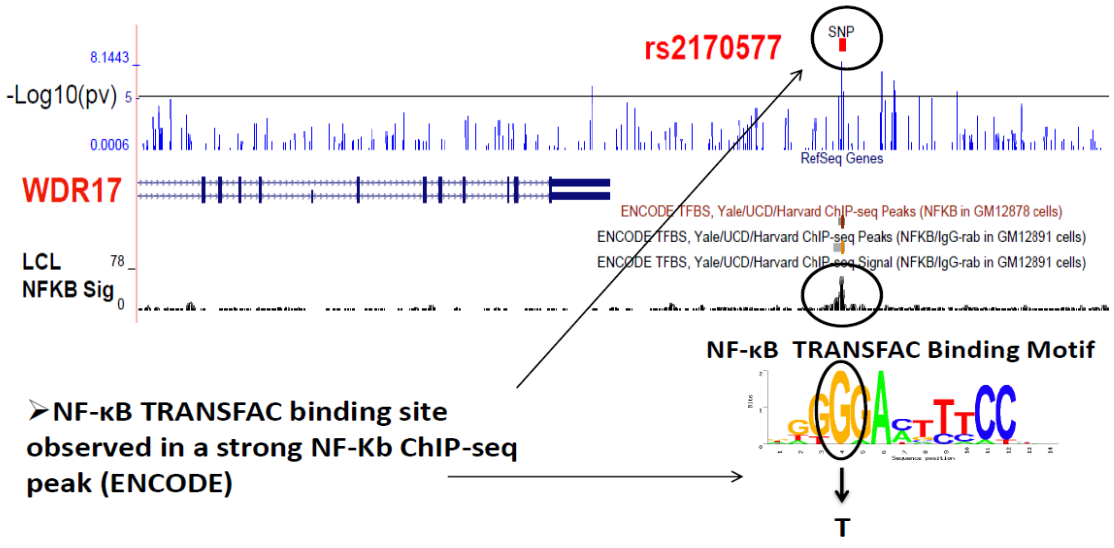


Figure 3-8. Example of SNP overlapping NF-κB ChIP-seq peak and TRANSFAC binding site. One of the top associated SNPs (*rs2170577)* with locus *WDR17* is also in a NF-κB TRANSFAC binding site. Disruption from G to T results in a loss of the NF-κB binding motif. In addition, this SNP is observed in a strong NF-κB LCL specific ChIP-seq peak (ENCODE).

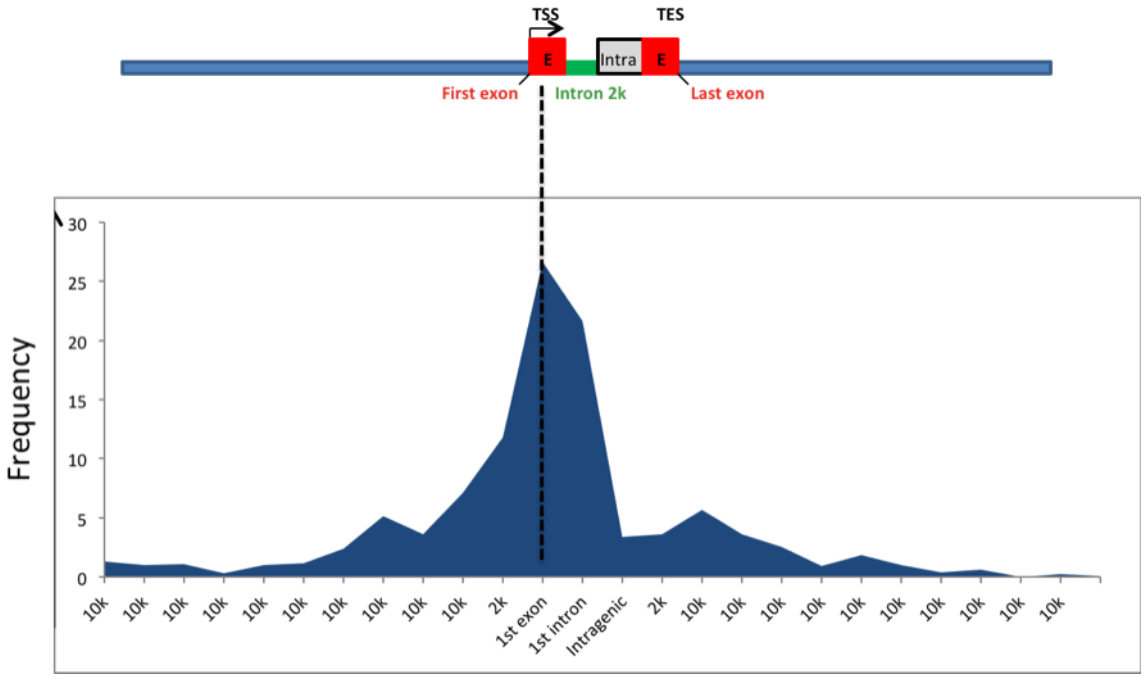


Figure 3-9. Relative distribution of SNPs. Analysis of heterozygous top 3 SNPs associated to transcripts showing differential allelic expression upon inhibition of NF-kB versus induction with TNF-α. There is significant enrichment of SNPs in regulatory regions including 10Kb upstream of the promoter, 2Kb upstream of the promoter, the first exon, and the 1st intron.

Network Analysis

Functional analysis of loci showing differential AE associated to top 3 heterozygous *cis*-rSNPs was done using IPA. The output was 3 networks consistent with the literature for NF-κB function within the cell (*70-75*) (*Table 3-5)*. The first network was characterized by functions relating to cell death, cellular compromise and inflammatory response ($p= 10^{-41}$). The second network indentified in the set of loci involved functions relating to dermatological disease and conditions, infectious disease and lipid metabolism ($p= 10^{-35}$). The second network also contained a top hub, which was NF-κB (*Figure 3-10).*

Network	Molecules in Network	Score	Focus Molecules	Top Functions
1	Adaptor protein 1, **ASAH1**, atypical protein kinase C, BCR (complex), **CCL22, COL16A1, CYBA, DGKI, DYRK4**, ERK1/2, **EXOC2**, Fc gamma receptor, Fcer1, Fgf, **FLCN, HDGFRP3, IL6R, MRPL19, MYO5B, NEDD4, NPLOC4, NTN1**, PLC gamma, **PMEPA1, RALB, SH2B3, SIRPB1, SMURF2**, Sos, **SPRED2, SPRY1, STK17B, SYK, TSPO**, VAV	41	25	**Cell Death, Cellular Compromise, Inflammatory Response**
2	aldehyde dehydrogenase [NAD(P)], ALDH, **ALDH1L1, ALDH3A1, ALDH3B1, ALDH7A1, BBS2, BLK, CBR3**, CCDC155, **CR2, EIF2AK1, FUCA1, GNL2**, Ifn, IFN alpha/beta, Igg3, IL-1R/TLR, **IL1R1**, Immunoglobulin, IRAK, **LY86, MOG**, NFkB (complex), PI3K p85, **RAB31, RGS20, RNF141, SYK/ZAP, TBC1D4, TLR1, TLR6**, Tlr, **TREML4, ZNF71**	35	3	**Dermatological Diseases and Conditions, Infectious Diseases, Lipid Metabolism**
3	**ABL2**, Akt, Alpha catenin, **ARHGAP24, AXIN1, CCNB1**, CD3, Cdc2, **CDC7 (includes EG:12545), CNIH4**, Cyclin A, Cyclin B, Cyclin E, **DDX11/DDX12P, DHRS4**, E2f, F Actin, Hdac, **ICAM3, IQSEC1, LMO2, MAD1L1, MYBL2, MYO5A, PYGL**, Ras homolog, Rb, Rock, **SMPDL3B, STEAP2, TFDP2, TJP2, TNFRSF10A, TOP2B, ZNF665 (includes others)**	34	22	**Cancer, Gatrointestinal Disease, Cellular Assembly and Organization**

Table 3-5. Functional analysis using IPA. Loci showing differential AE were analyzed using IPA. This yielded three top networks consistent with NF-κB function. The score is based on a p-value calculation. This calculates the likelihood that the network eligible molecules that are part of the network are found there by random chance. The value is the negative exponent of the right-tailed Fisher's exact test result.

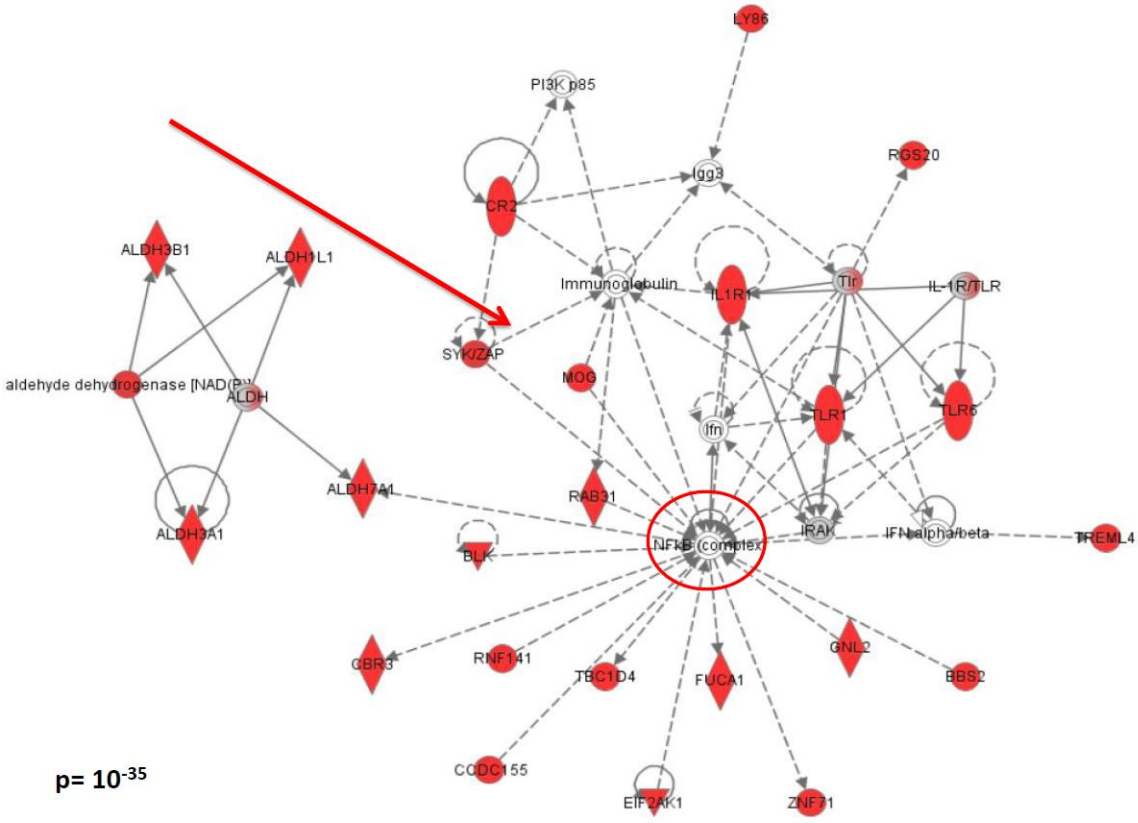


Figure 3-10. Network 2: Dermatological Disease and Conditions, Infectious Disease and Lipid Metabolism. NF-κB was a top hub in the second network associated with the set of differentially expressed loci (p-value=10^{-35}).

GWAS analysis

We investigated whether top 3 SNPs in our list of mapped *cis*-regulatory variants, which were associated to the subset of loci that showed differential AE, were also associated to complex diseases based on previous GWAS results available in the literature. Top 3 *cis*-rSNPs were associated to complex diseases including immune related and/or autoimmune diseases such as systemic lupus erythematosus, ulcerative colitis, and rheumatoid arthritis *(Table 3-6).*

Gene ID/ Accession	SNP	Association (p-value)	Complex Disease (GWAS)	Association (p-value)	Popu-lation
BLK (NM_001715)	rs13277113	7.09E-28	Systemic lupus erythematosus	1.0E-10	CY
BLK (NM_001715)	rs2618476	1.27E-30	Systemic lupus erythematosus	2.0E-8	CY
BLK (NM_001715)	rs2736340	4.47E-30	Systemic lupus erythematosus Rheumatoid arthritis	3.0E-7	CY
FAM167A (NM_053279)	rs2618476	9.26E-21	Systemic lupus erythematosus	2.0E-8	CY
FAM167A (NM_053279)	rs2736340	7.12E-21	Systemic lupus erythematosus Rheumatoid arthritis	3.0E-7	CY
FAM167A (NM_053279)	rs13277113	9.26E-21	Systemic lupus erythematosus	1.0E-10	CY
GSDMB (NM_001042471)	rs8067378	5.59E-18	Ulcerative colitis	1.0E-7	CY
HPCAL4 (NM_016257)	rs873917	8.40E-08	Amyotrophic lateral sclerosis	8.0E-6	Y
MYO1B (NM_012223)	rs13030978	3.81E-05	Liver enzyme levels (gamma-glutamyl transferase)	1.0E-11	C
C8orf12 (uc003wtu.2)	rs2618476	9.49E-20	Systemic lupus	3.0E-7	CY

			erythematosus		
C8orf12 (uc003wtu.2)	rs13277113	9.49E-20	Systemic lupus erythematosus	1.0E-10	CY
C8orf12 (uc003wtu.2)	rs2736340	1.11E-19	Systemic lupus erythematosus Rheumatoid arthritis	3.0E-7	CY
854 (OTTHUMT00000256997)	rs8067378	7.36E-18	Ulcerative colitis	1.0E-7	CY

Table 3-3. GWAS Analysis. Candidate *cis*-rSNP most strongly associated (rank 1, 2 and 3) to loci harbouring changes in differential AE were analyzed to investigate association to immune related and/or autoimmune diseases previously seen in the literature by GWAS.

Bernstein HMM Classifications

We investigated chromatin states based on Bernstein HMM classifications (*85*) for case and control SNPs associated with loci showing differential AE upon perturbation of NF-κB. Case SNPs were twice as likely to be in active promoter or strong enhancer regions as compared to controls SNPs. This significant enrichment was observed when the investigation was restricted to top 10 SNPs (Chi-squared test, p-value = 5.3E -40), top 3 SNPs (Chi-squared test, p-value = 2.7E-24) and also for top SNPs (Chi-squared test, p-value = $6.8E^{-17}$) per loci. In addition, a similar proportion of case and control SNPs were seen in all of the other chromatin states (*Figure 1-7).*

Investigation of relevant loci

Using a literature search, transcripts were characterized by biological or biomedical relevance to the NF-κB pathway and implications in inflammation and/or autoimmune disease. Top 3 *Cis*-rSNPs associated to loci showing differential AE in our experiment composed the subset under investigation *(Table 3-7).*

Relevant Loci	Explanation
Complement component 3 (C3)	Target gene of NF-κB (*69*), part of the complement system, which contributes to innate immunity (*98*)
Complement component receptor 2 (CR2)	Part of the complement system, which contributes to innate immunity and activates the NF-κB pathway (*99*)
***Interleukin 6 receptor (IL-6R*)**	IL-6 plays a central role in vascular inflammation by the activation of NF-κB and downstream events including production of IL-6 by targeting the receptor component (*100*)
Gasdermin B (GSDMB)	*GSDMB* transcript has been implicated in asthma studies, as well as, autoimmune diseases (e.g., rheumatoid arthritis, Crohns disease and ulcerative colitis) with SNPs causing opposite effects on the immunopathogenesis of the aforementioned (*101*)
B lymphocyte kinase (BLK)	*Cis*-rSNPs have been associated to BLK and to systemic lupus erythematosus (*17*)

Table 3-7. Implication of loci showing differential AE in the NF-κB pathway or range of immune related diseases. A literature review using the search engine, PubMed was done in order to determine loci previously associated to the NF-κB pathway or immune related diseases. The subset of loci investigated includes transcripts showing differential AE associated to top 3 ranked SNPs for the LCL populations.

Case Study: SNAI1/WNT4 Model

Intersection of datasets for discovery of cis-rSNP

We had a strong hypothesis for the association of *SNAI1* and *WNT4*. We identified a SNP that affects *WNT4 cis*-regulation in our FB panel and alters the risk for osteoporosis (*87*). The signaling pathway of *WNT4* has been implicated in bone development (*102, 103*). Combination of multiple GWAS datasets, eQTLs in osteoblasts, AE in primary FBs and DHS-seq from mesenchymal stem cell (MSC) lineage allowed determination of a single SNP > 200kb upstream of the *WNT4* gene. Primarily, overlapping BMD detected by Decode (*104*) and our human osteoblast eQTL (21) data indicated a WNT4 5' regulatory variant specific for MSC lineage, and therefore was directly relevant to bone disease (BJ and NHDF cells). A common variant overlapping the lineage specific DHS-site was fine-mapped (sequencing functional sites + follow up genotyping in cell panels) in independent AE data explaining increased BMD association in GEFOS1 and *WNT4* expression in independent data sets. Bioinformatics analysis of the *cis*-rSNP indicated that it altered a *SNAI1* binding site. Consequently, this region was thought to be *cis*-regulated since it is found in a DHS region, which is a site of active chromatin *(Figure 3-11)*. We have shown *in vitro* allele-specific EMSA signals in nuclear extracts from MSC lineage (MG-63 cells) and we are now pursuing *in vivo* validation of SNAI1 binding in living cells. As previously mentioned, this was completed by carrying out ChIP, as well as *SNAI1* knockdown by RNA*i* with monitoring of its consequences in *WNT4* AE phenotype, as well as, genome-wide AE assessment.

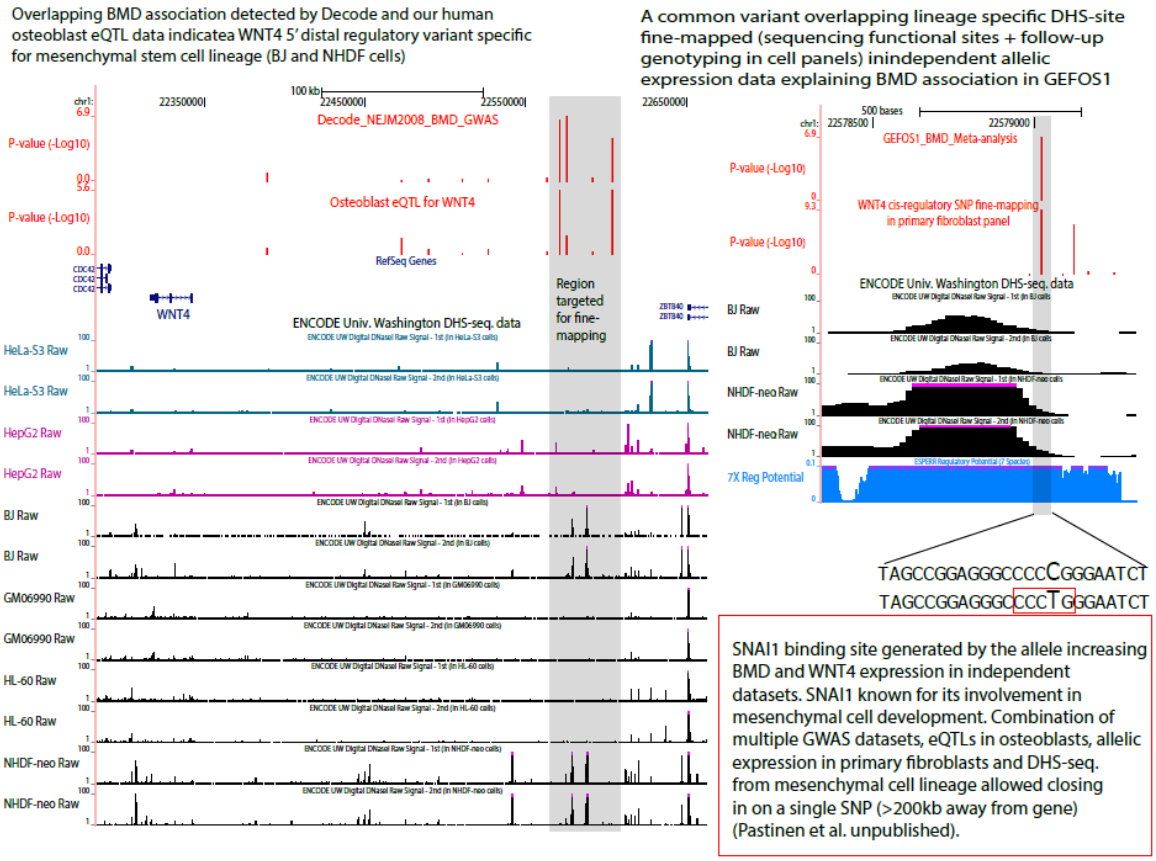


Figure 3-11. Intersection of datasets for discovery of *cis*-rSNP. The combination of multiple datasets enabled the identification of a *cis*-rSNP which increases the expression of *WNT4* and BMD in independent datasets. This *cis*-rSNP is >200kb upstream of the *WNT4* gene. Bioinformatics analysis of this SNP using TRANSFAC showed that it altered a SNAI1 binding site.

Validation of SNAI1 knockdown by RNAi

Consistent inhibition of *SNAI1* using *RNAi* in transfection studies has been observed as compared to the negative control for various cell lines heterozygous for *rs6684375*, including WG1657, GM02317, HOB642 using RT-PCR *(Figure 3-12)*. Validation of the effect of *SNAI1* on the expression of *WNT4* was also done using RT-PCR using the Rotor-Gene^{TM} 6000 real-time rotary analyzer (Corbett Life Sciences, Sydney, Australia). Ambivalent results were observed in terms of verifying the knock down of *WNT4* expression. This included inconsistencies in technical replicates as per the example using the cell line WG1657 *(Figure 3-13).* In addition, the majority of the discrepancies were seen in replicates done in RT- PCR; consequently, many of the analysis could not be expressed quantitatively.

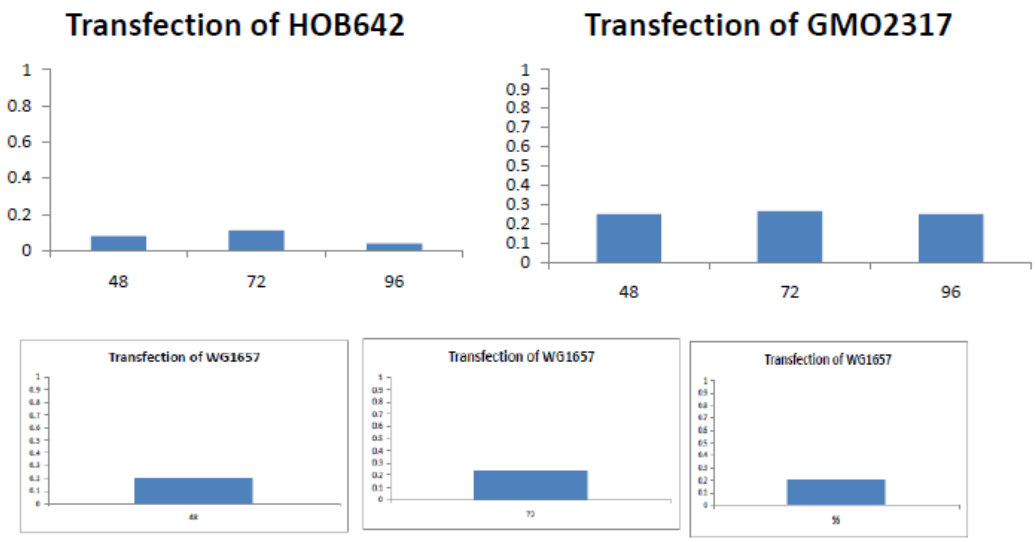


Figure 3-12. Validation of *SNAI1* knockdown by RT-PCR. Validation of *SNAI1* knockdown versus the negative control was done by using known primers for SNAI1, as well as, housekeeping genes (*18s ribosomal RNA* and *exportin-5)* using the comparative C_T method. This was done at various time points *(x-axis)* and the fold change is indicated *(y-axis).* Approximately 85 % inhibition of *SNAI1* was observed for WG1657, HOB642 and GMO2317.

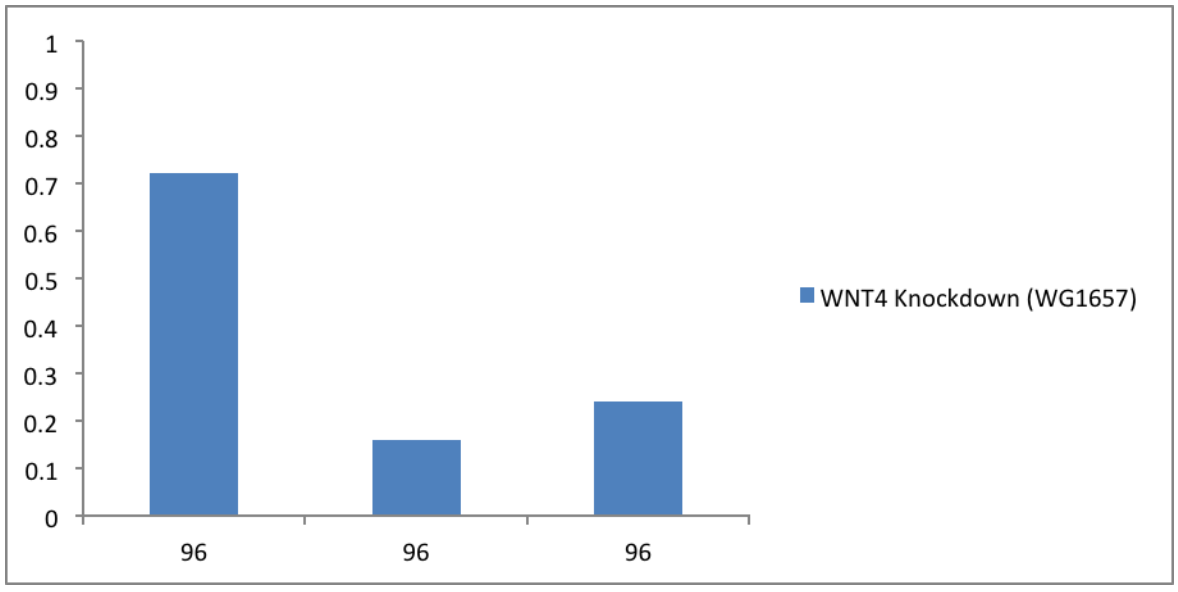


Figure 3-13. Independent analyses of *WNT4* knockdown. Knockdown of *SNAI1* and its effect on the expression of *WNT4* was analyzed using RT-PCR for primers that enriched for *WNT4*. Independent analysis of *WNT4* in three experiments with the cell line WG1657 were done at 96 hours post-transfection. *WNT4* knockdown was not consistent across technical replicates.

Results from assessment of SNAI1 binding at rs6684375 by ChIP

ChIP-RT-PCR experiments show the enrichment of *SNAI1*-specific binding site (*rs6684375)* in MSC lineages. Enrichment of the *SNAI1*-specific binding region was demonstrated by using primers, which surround and amplify the SNP. The enrichment from the SNAI1 antibody was compared to that of IgG antibody. Relative enrichment using two different primers, as well as, two cell lines is shown in Figure 3-14.

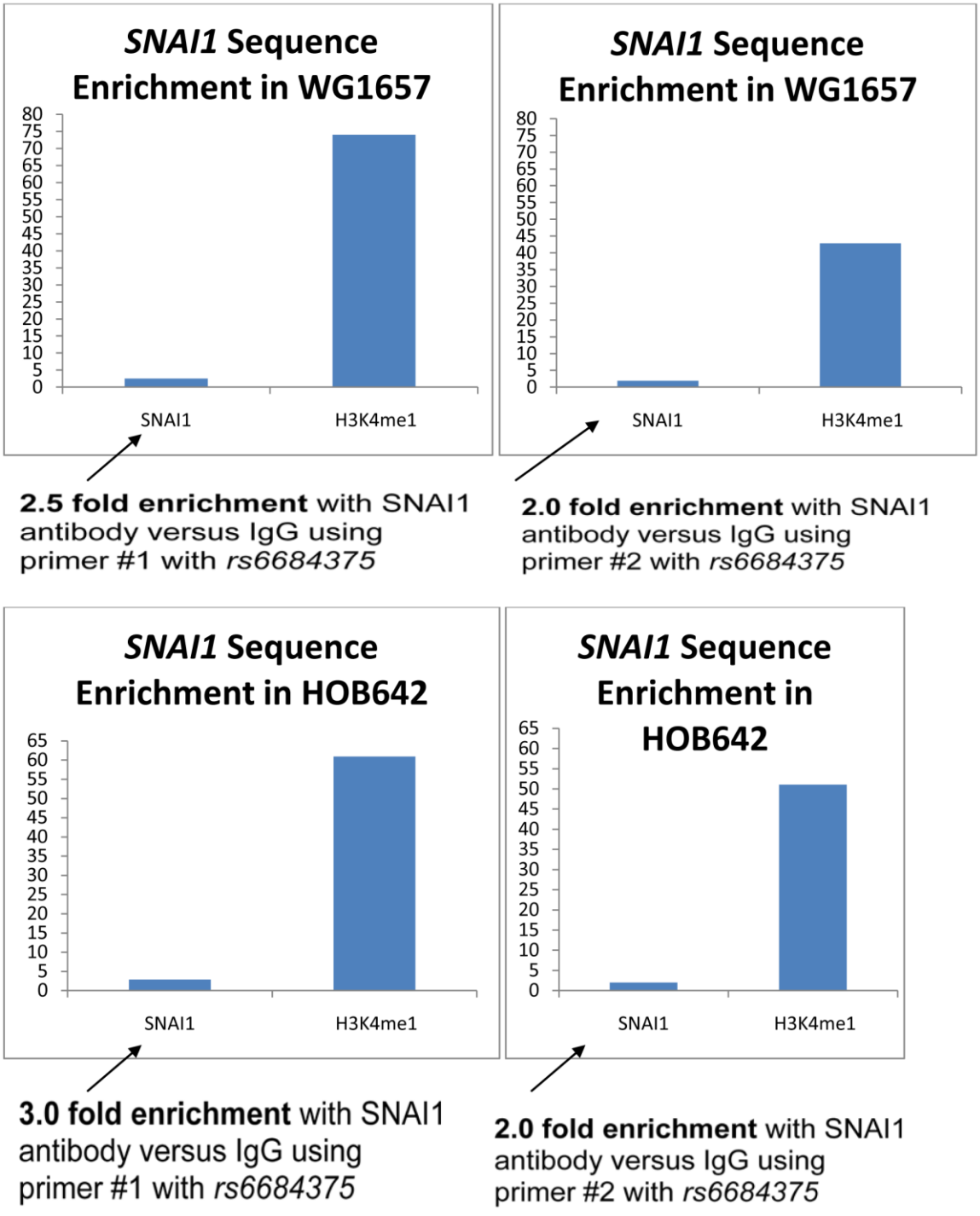


Figure 3-14. SNAI1 binding site (*rs6684375)* enrichment assessed by ChIP-RT-PCR. ChIP-RT-PCR was done using the antibodies SNAI1, IgG, and H3k4me1. Primers in which *rs6684375* was amplified were used to validate the enrichment of SNAI1 binding. 2-3 fold enrichment was seen for SNAI1versus immunoprecipitation with IgG.

Bioinformatics analysis of genome-wide AE

Preliminary analysis of data generated from the high-throughput experiment in which SNAI1 was perturbed, was analyzed as follows. Analysis consisted of determining the multiplicative probability of observing 3 SNPs in a row using different significant cutoff p-values (i.e. the likelihood of seeing 3 consecutive SNPs with high change in heterozygous magnitude ratio). After which we compared treated samples and control samples in order to observe a change in AE. The change in AE was only significant for values over the 90^{th} percentile. We first investigated SNPs across the *WNT4* transcript, which did not show a consistent change in AE between treated and control samples. We proceeded to analyze other transcripts in our data as mentioned above and we observed in FBs, as per the following example, SNAI1-mediated specific AE differences in the transcript *GRIN3B* on chromosome 19 *(Figure 3-15).* In addition, in an independent analysis all overexpressed alleles were phased to the same chromosome. Phasing was done for all SNPs using 1M data. The independent analysis showed a similar result, which confirmed the differential AE seen in the transcript *GRIN3B*.

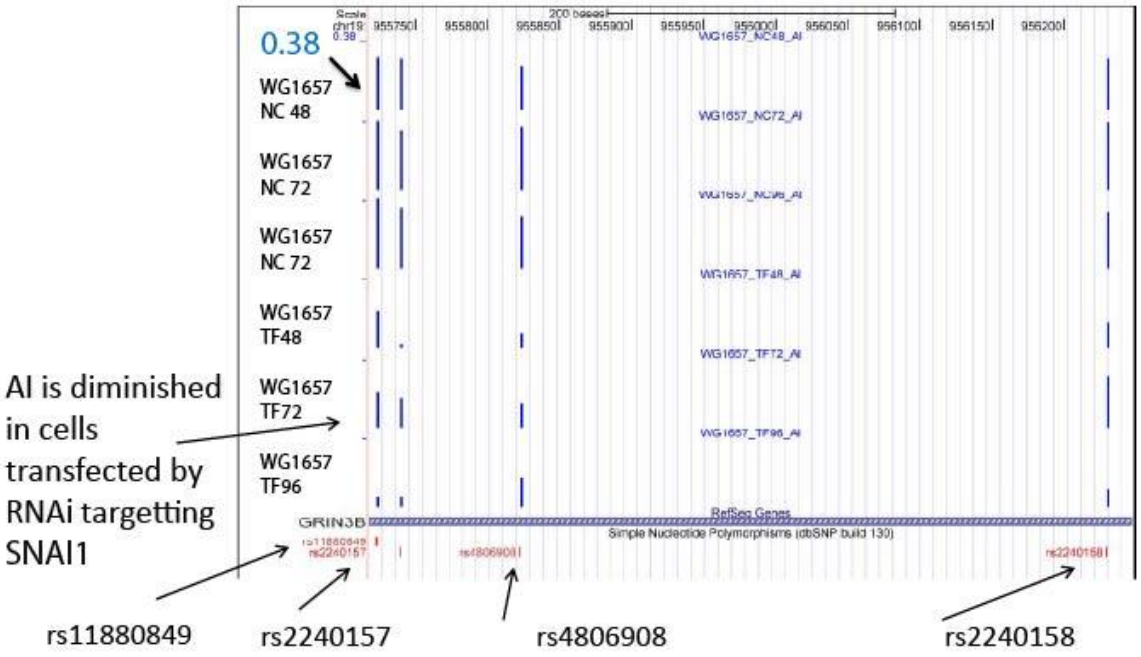


Figure 3-15. *SNAI1*-mediated AE change for *GRIN3B* on chromosome 19. There was diminished AE in cells transfected with RNA*I* targeting *SNAI1* as compared to the control samples for *GRIN3B.* This was observed at 3 time points (48, 72, 96 hours). In an independent analysis all overexpressed alleles were phased to the same chromosome. Phasing was done for all SNPs using data from the Illumina HumanOmni 1M-Quad BeadChips.

An alternative approach to assess AE (SNAI1)

Due to the fact that the TF NF-κB is well understood throughout the literature than, we believed that the method of analysis used would provide insight into studying SNAI1. In addition, we observed encouraging results from the study with NF-κB, in which specific loci changed upon perturbation, which furthered our reasoning for analyzing SNAI1 data with a similar methodology. The in-house map of *cis*-regulatory variation for FB cells, which is similar to the one created for LCLs, was integrated with FB mean AE data for experimental samples in which SNAI1 was targeted by RNA*i* and the negative control. As can be seen in Figure 3-16, we replicated a similar approach both to perturb SNAI1 using RNA*i* and genome-wide AE assessment on Illumina HumanOmni2.5-Quad BeadChips as for NF-κB. We were searching for loci, which showed differential AE upon targeting SNAI1 versus the control samples. Consequently, we observed diminished AE for only 63 loci in samples transfected with RNA*i* targeting SNAI1 versus control samples at 48, 72 and 96 hours post treatment.

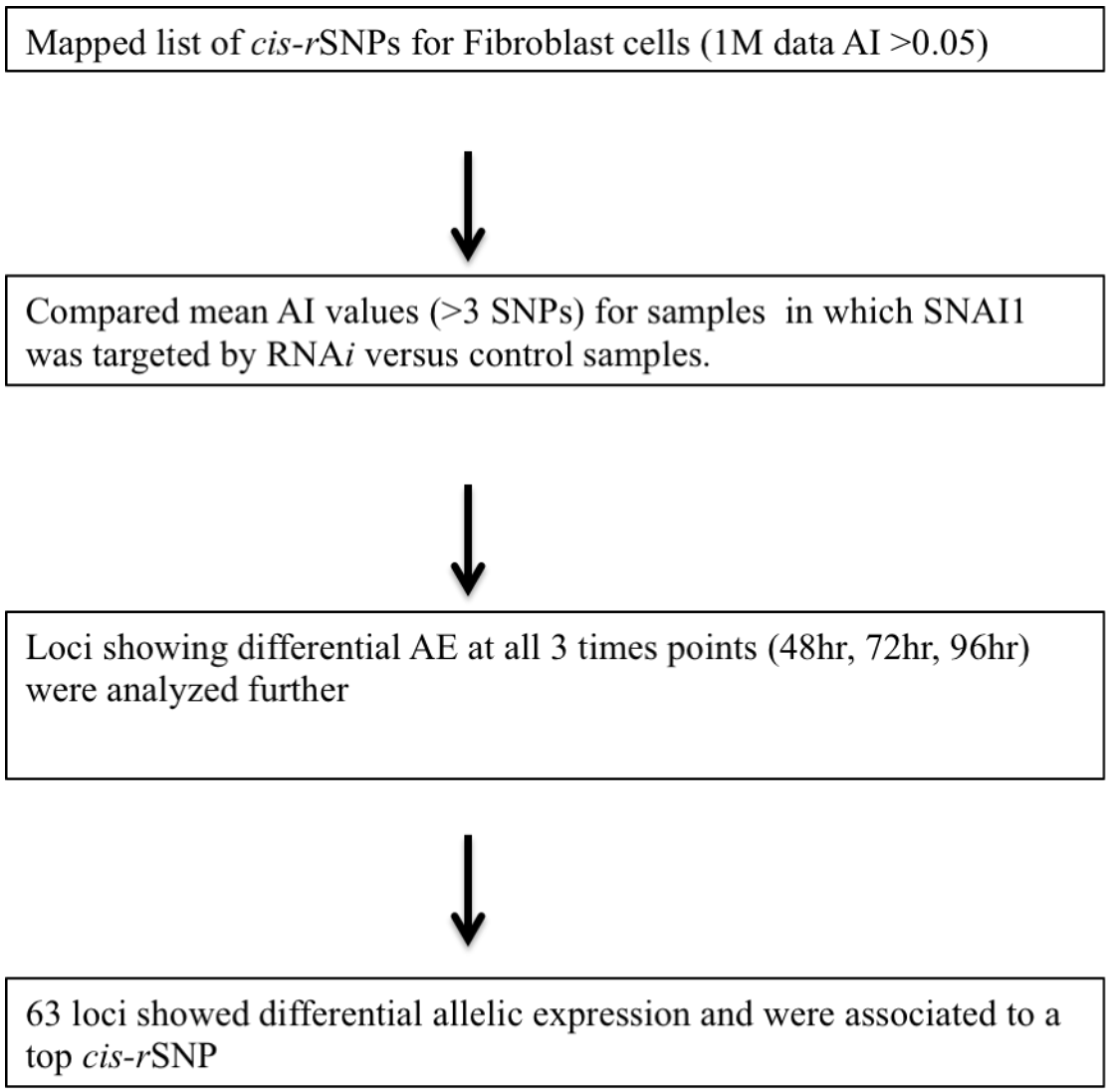


Figure 3-16. Filtered analysis of genome-wide AE data. An in-house map of *cis*-regulatory variation for FB cells was merged with our mean AI data from samples in which *SNAI1* was targeted by RNA*i.* The first filter implemented was to use loci from the list showing an AE change of greater than the threshold of 0.05 from the Illumina HumanOmni1-Quad BeadChip data. Subsequently, loci were analyzed only if AE was diminished in samples in which SNAI1 was knockdown versus the control. Mean AI values were compared and only loci with an AE change at greater than >3 SNPs were retained. Moreover, loci showing differential AE at the 3 times points (48, 72 and 96 hours) were further investigated.

CHAPTER 4: DISCUSSION

Study Conclusions

Over the preceding 25 years, major progress has been made in elucidating the ubiquitous yet intricate role the NF-κB family of inducible TFs play in gene regulation (*70*). However, numerous questions have remained unanswered, several of which we have attempted to shed light on. This thesis primarily investigated the effect of TF binding variation of NF-κB on differential AE. Preliminary analysis of LCL in house *cis*-rSNP data revealed enrichment in NF-κB binding sites versus control SNPs highlighting the potential value of an in depth study on the role of NF-κB on AE in LCLs. Genome-wide AE assessment using the Illumina HumanOmni5-Quad BeadChips highlighted transcripts from our mapped list of *cis*-regulatory variants that showed differential AE upon perturbation experiments of NF-κB in LCLs. *Cis*-rSNPs (rank 1,2,3) associated with this subset of loci were subsequently overlapped with in-house and publicly available functional genomic data. The functional data included LCL-specific ChIP-seq peaks and TRANSFAC binding sites for the following TFs: NF-κB, E2A, STAT1, IRF2, and EBF1. Significant enrichment was observed for top *cis*-rSNPs (rank 1) at NF-κB ChIP-seq peaks for the subset of loci versus all mapped loci.

Further investigation revealed loci exhibiting differential AE were associated to a heterozygous top 1 or top 3 ranked *cis*-rSNP, which in 36% of cases and 64% of cases, respectively, overlapped a functional element for NF-κB or one of the known cooperative TFs. Results consistent with conclusions previously reported in the literature for NF-κB were also observed for loci, which showed differential AE upon inhibition of NF-κB coupled to induction by TNF- α. *Cis*-rSNPs were significantly enriched in active promoter and strong enhancer elements as compared to control SNPs. The distribution of

top 3 *cis*-rSNPs across the genome showed an increased likelihood of SNPs found in functional elements relative to all SNPs in the region. In addition, network analysis using IPA uncovered 3 top networks, which were consistent with NF-κB function within the cell and one of which contained NF-κB as a top hub in the network. Finally, many of the transcripts exhibiting differential AE upon perturbation were themselves part of the NF-κB pathway and/or were related to inflammation and/or autoimmune disease.

Challenges for the cooperative binding mechanism of NF-κB

One challenge that remains is to discover the remaining TF coassociations at specific target sites since the vast majority has yet to be uncovered especially in terms of NF-κB. Previous studies have shown that TF coassociations are evident within about 1-2kb of the factor under analysis. However, distal coassociations present complex problems in part due to the likelihood that more nonfunctional but related motifs will be found in these regions, or potentially that there will be a weaker effect observed for distal compared to proximal regions (63). Another consideration is that the majority of previous analyses have focused on linear coassociations; yet, biological processes are complex and likely involve a thermodynamic TF coassociation model (*64*, *105*). Investigation of thermodynamic models is based on the equilibrium binding of the TF to DNA as well as to one another and has shown promising results in eukaryotic systems. Thermodynamic models make use of synthetic expression libraries, which consists of a random combination of three to four TFBS. For instance, using a yeast genome, *Saccharomyces cerevisiae*, a synthetic promoter library was constructed and results showed a number of Mig1-regulated genes that lack significant binding sites for Mig1 in the promoter region (*105*). The above reasoning could in part account for the observation that some transcripts

associated to a *cis*-rSNP (rank 1, 2 ,3) do not overlap with a functional element for NF-κB or a cooperative TF.

Caveats for observed AE differences

Limitations in explaining the observed AE differences in our experiments can be, to a certain extent, attributed to the publicly available ChIP-seq data used in the bioinformatics approach. ChIP-seq is one of the earliest applications of NGS, which although advantageous in many respects such as the ability to sequence thousands or millions of short DNA fragments per run, there are still systematic difficulties with the technology. Even though errors due to sequencing have decreased considerably, this is still an important consideration particularly, near the end of reads (*90*). Another challenge in ChIP studies, which was alluded to above, is the low coverage of polymorphic sites, thus decreasing the number of informative sites for study. In order to circumvent this problem, higher coverage is needed to provide the power to detect allelic biases (*10*). Detecting enriched regions is also challenging when there is an insufficient number of reads as it compromises sensitivity and specificity (*90*). Studies published to date have not yet achieved the coverage to include the comprehensive range of allelic biases in the genome (*10*).

Future Studies

In order to validate the differences in AE observed, high-throughput ChIP-seq experiments primarily using NF-κB (Santa Cruz Biotechnology; sc-372) could be done with samples from the experiment in which LCLs (GM19239, GM19240, GM19238, GM12891, GM12892, GM12878) were treated with TNF- α induction coupled to inhibition of NF-κB. In parallel, factors known to be coassociated to NF-κB that we

evaluated E2A, STAT1, IRF2, and EBF1, ChIP-seq could be done on the same samples. ChIP-seq would be done at a higher coverage and used in our bioinformatics approach to increase our understanding of the cooperative mechanism of NF-κB binding. Future studies may extend to include more cell lines in order to increase the power of results observed for NF-κB. Analysis of NF-κB and other cooperating factors may potentially better explain the role of such SNPs in control of regulatory variation. Combining other analyses will be essential for understanding differences in NF-κB binding, which will in turn allow for comprehensive annotations of genomic variants and developments in systems biology (*64*). This can be extended as an approach to study other transcription factor binding using a suitable model system.

Development and Refinement of SNAI1 Approach

Experiments were performed in order to validate the hypothesis for the association of the TF SNAI1 and *WNT4*, in which a *cis*-rSNP forming a SNAI1 binding site was shown to alter the expression of *WNT4. WNT4* is relatively lowly expressed in most cell lines as can be observed via BioGPS (www. http://biogps.org), including in FB and osteoblast cell lines. Due to this, RNA*i* studies of expression are difficult to undertake with *WNT4.* Discordant results were observed for *WNT4* expression, constituting a key reason for the inability to fully address our hypothesis in the case study of the association between SNAI1 and *WNT4.* Our initial genome-wide AE approach using the Illumina HumanOmni 2.5M-Quad BeadChips, in order to search for loci with differential AE in samples in which *SNAI1* was inhibited versus control samples did not elucidate a large number of loci. Even though applying the alternative approach, based on encouraging results from analysis of NF-κB, resulted in an increased number of loci exhibiting

changes in AE upon inhibition of *SNAI1* in FB cell lines, the experiment still seemed to alter AE for relatively few loci.

Validation of the Regulatory Role of NF-κB in LCLs

The above point was made evident due to a parallel comparison of differential AE with NF-κB in LCLs, as well as, SNAI1 in FB cell lines. This comparison is shown in Figure 3-16. There is over a 16-fold difference in the relative number of loci that were affected between LCLs and FBs. From our mapped list of *cis*-regulatory variation, NF-κB had a regulatory effect on approximately 33% of loci in comparison to approximately 2% of loci for SNAI1 in FBs using a similar approach to filter the loci. We observed a significant difference (Chi-squared test, pv= 1.83E-195) between perturbations of NF-κB versus SNAI1.

The statistical significant difference in the perturbation studies between the two transcription factors provides evidence for the importance of using a specific transcription factor in an appropriate cell line in order to assess the TF's control of regulatory variation. Even though SNAI1 is known for its involvement in mesenchymal stem cell development and we have in-house evidence demonstrating the role of the TF SNAI1 in fibroblasts; fibroblast cells may not provide an ideal system for studying SNAI1 (*88*). SNAI1 is known to play a part in epithelial-mesenchymal transition (EMT) and as such epithelial cells could be used to better study the regulatory role of SNAI1 (*88, 106*). Conversely, the regulatory role of transcription factors SOX2 and C-MYC could be better elucidated in fibroblast cells (*107*).

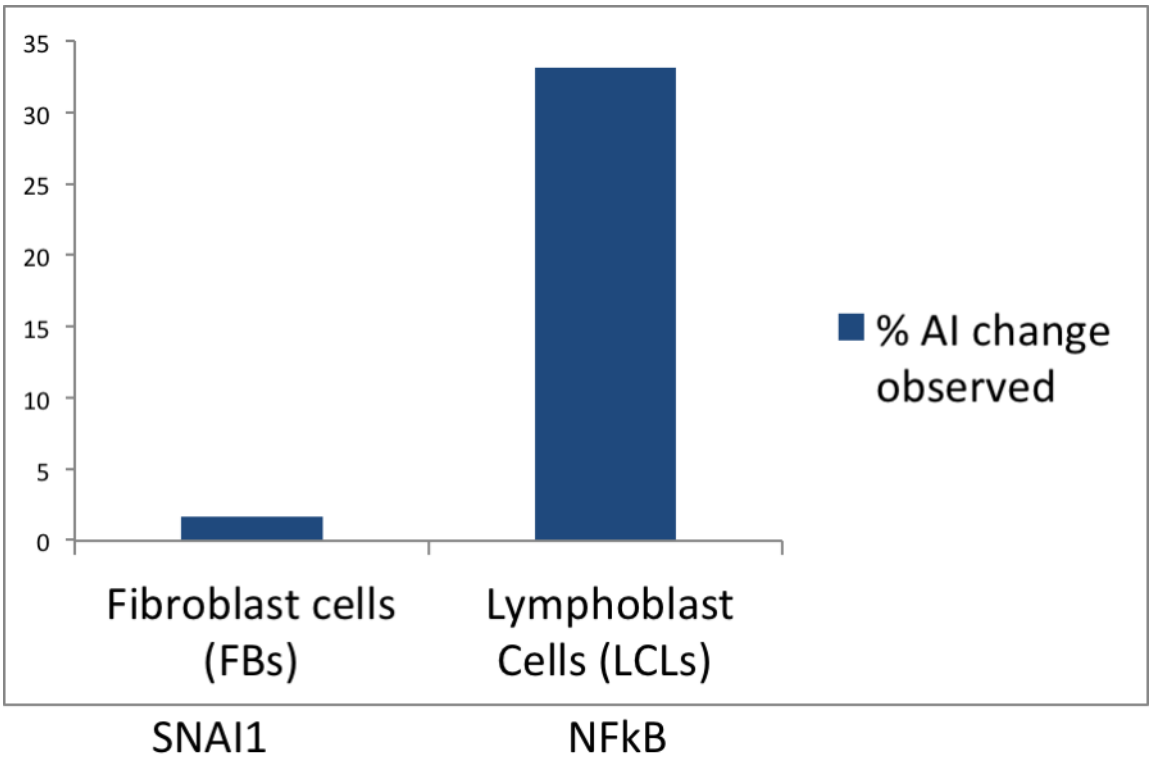


Figure 3-16. Comparison of NF-κB and SNAI1 effect on AE genome-wide. In terms of the mapped list of *cis*-regulatory variants for LCLs and FBs, >1700 and >4000 loci were associated to *cis*-rSNPs, respectively. From the list of loci for LCLs, 581 were heterozygous at a top SNP and had an AE change at greater than 3 SNPs compared to 63 loci for FBs.

Relevant Contributions to the Field

The ability to not only effectively perturb NF-κB in a relevant cell type but also use a bioinformatics approach to narrow down on likely causal variants makes this an attractive method for further studies with other TFs. Integrating the parallel collection and assessment of allelic data at several levels of genomics is crucial to bettering our understanding of TF-DNA interactions. Our results demonstrate extensive contributions of genetic variation on TF binding for NF-κB, as well as, SNPs underlying the allele-specific sites, which could likely affect TF binding and chromatin structure. Furthermore, the majority of loci displaying changes in AE upon perturbation of NF-κB are shown to be associated to top *cis*-rSNPs overlapping functional elements. The genetic variants overlapping LCL-specific ChIP-seq peaks and/or TRANSFAC binding sites for NF-κB or a known cooperative TF can have an effect on the observed AE. As such, we have provided evidence that assessment of TF binding for gene regulation can be translated from traditional tools to the direct assessment of allele-specific TF binding.

REFERENCES

1. Initial sequencing and analysis of the human genome. *Nature* **409**, 860 (2001).
2. J. C. Venter *et al.*, The sequence of the human genome. *Science* **291**, 1304 (February 16, 2001, 2001).
3. F. S. Collins, M. Morgan, A. Patrinos, The human genome project: lessons from large-scale biology. *Science* **300**, 286 (April 11, 2003).
4. M. L. Metzker, Sequencing technologies the next generation. *Nat Rev Genet* **11**, 31 (2010).
5. E. P. C., A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**, e1001046 (2011).
6. S. C. J. Parker, L. Hansen, H. O. Abaan, T. D. Tullius, E. H. Margulies, Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324**, 389 (April 17, 2009).
7. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799 (2007).
8. B. J. Raney *et al.*, ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Research*, (October 30, 2010).
9. J. Pevsner, in *Wiley-Blackwell*. (N.J., 2009).
10. T. Pastinen, Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**, 533 (2010).
11. V. Ferretti *et al.*, PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Research* **35**, D122 (January 1, 2007).
12. J. Shendure, H. Ji, Next-generation DNA sequencing. *Nat Biotech* **26**, 1135 (2008).
13. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010).
14. G. A. Heap *et al.*, Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human Molecular Genetics* **19**, 122 (January 1, 2010).
15. S. B. Montgomery *et al.*, Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773 (2010).
16. J. K. Pickrell *et al.*, Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768 (2010).
17. B. Ge *et al.*, Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat Genet* **41**, 1216 (2009).
18. A. Gimelbrant, J. N. Hutchinson, B. R. Thompson, A. Chess, Widespread monoallelic expression on human autosomes. *Science* **318**, 1136 (November 16, 2007).
19. D. Botstein, N. Risch, Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*.
20. G. Kuhlenbäumer, J. Hullmann, S. Appenzeller, Novel genomic techniques open new avenues in the analysis of monogenic disorders. *Human Mutation* **32**, 144 (2011).

21. S. B. Ng, D. A. Nickerson, M. J. Bamshad, J. Shendure, Massively parallel sequencing and rare disease. *Human Molecular Genetics* **19**, R119 (October 15, 2010, 2010).
22. I. Lobo, Multifactorial inheritance and genetic disease. *Nature Education* **1**, (2008).
23. D. Altshuler, M. J. Daly, E. S. Lander, Genetic mapping in human disease. *Science* **322**, 881 (November 7, 2008, 2008).
24. W. Cookson, L. Liang, G. Abecasis, M. Moffatt, M. Lathrop, Mapping complex disease traits with global gene expression. *Nat Rev Genet* **10**, 184 (2009).
25. E. V. Davydov *et al.*, Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).
26. T. A. Manolio *et al.*, Finding the missing heritability of complex diseases. *Nature* **461**, 747 (2009).
27. L. A. Hindorff *et al.*, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362 (June 9, 2009).
28. E. Birney, J. D. Lieb, T. S. Furey, G. E. Crawford, V. R. Iyer, Allele-specific and heritable chromatin signatures in humans. *Human Molecular Genetics*, (September 16, 2010).
29. R. J. Britten, E. H. Davidson, Gene regulation for higher cells: a theory. *Science* **165**, 349 (July 25, 1969).
30. M. King, A. Wilson, Evolution at two levels in humans and chimpanzees. *Science* **188**, 107 (April 11, 1975).
31. D. L. Stern, V. Orgogozo, The loci of evolution: how predictable is genetic evolution?*Evolution* **62**, 2155 (2008).
32. S. B. Carroll, Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25 (2008).
33. P. J. Farnham, Insights from genomic profiling of transcription factors. *Nat Rev Genet* **10**, 605 (2009).
34. P. J. Wittkopp, G. Kalay, Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**, 59 (2012).
35. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928 (2001).
36. L. Consoli, A. Lefèvre, M. Zivy, D. de Vienne, C. Damerval, QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize. *Plant Molecular Biology* **48**, 575 (2002).
37. C. Damerval, A. Maurice, J. M. Josse, D. de-Vienne, Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. *Genetics* **137**, 289 (May 1, 1994).
38. E. Grundberg *et al.*, Global analysis of the impact of environmental perturbation on *cis*-regulation of gene expression. *PLoS Genet* **7**, e1001279 (2011).
39. M. Morley *et al.*, Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743 (2004).
40. T. Kwan *et al.*, Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40**, 225 (2008).

41. M. Griffith *et al.*, Alternative expression analysis by RNA sequencing. *Nat Meth* **7**, 843 (2010).
42. J. Majewski, T. Pastinen, The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in genetics : TIG* **27**, 72 (2011).
43. T. Zeller *et al.*, Genetics and Beyond – The transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5**, e10693 (2010).
44. V. Emilsson *et al.*, Genetics of gene expression and its effect on disease. *Nature* **452**, 423 (2008).
45. B. E. Stranger *et al.*, Population genomics of human gene expression. *Nat Genet* **39**, 1217 (2007).
46. A. L. Dixon *et al.*, A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202 (2007).
47. H. H. H. Goring *et al.*, Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* **39**, 1208 (2007).
48. V. G. Cheung *et al.*, Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365 (2005).
49. J. F. Degner *et al.*, DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390 (2012).
50. H. Yan, W. Yuan, V. E. Velculescu, B. Vogelstein, K. W. Kinzler, Allelic variation in human gene expression. *Science* **297**, 1143 (August 16, 2002, 2002).
51. J. C. Knight, B. J. Keating, K. A. Rockett, D. P. Kwiatkowski, In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat Genet* **33**, 469 (2003).
52. T. E. Reddy *et al.*, Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research* **22**, 860 (May 1, 2012).
53. R. McDaniell *et al.*, Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235 (April 9, 2010).
54. P. M. Visscher, W. G. Hill, N. R. Wray, Heritability in the genomics era-concepts and misconceptions. *Nature Rev. Genet.* **9**, 255 (2008).
55. G. Gibson, G. Wagner, Canalization in evolutionary genetics: a stabilizing theory? *Bioessays* **22**, 372 (2000).
56. M. Kasowski *et al.*, Variation in transcription factor binding among humans. *Science* **328**, 232 (April 9, 2010).
57. J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, N. M. Luscombe, A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252 (2009).
58. K. Chen, N. Rajewsky, The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8**, 93 (2007).
59. J. J. Moresco, P. C. Carvalho, J. R. Yates Iii, Identifying components of protein complexes in C. elegans using co-immunoprecipitation and mass spectrometry. *Journal of Proteomics* **73**, 2198 (2010).
60. T. Ravasi *et al.*, An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744 (2010).
61. W. Zheng, H. Zhao, E. Mancera, L. M. Steinmetz, M. Snyder, Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464**, 1187 (2010).

62. J. Rozowsky *et al.*, PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotech* **27**, 66 (2009).
63. L. J. Jensen *et al.*, STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research* **37**, D412 (January 1, 2009).
64. K. J. Karczewski *et al.*, Cooperative transcription factor associations discovered using regulatory variation. *Proceedings of the National Academy of Sciences* **108**, 13353 (August 9, 2011).
65. M. Levine, E. H. Davidson, Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 4936 (April 5, 2005).
66. A. R. Borneman *et al.*, Divergence of Transcription Factor Binding Sites Across Related Yeast Species. *Science* **317**, 815 (August 10, 2007).
67. C. R. Lickwar, F. Mueller, S. E. Hanlon, J. G. McNally, J. D. Lieb, Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **484**, 251 (2012).
68. B. Yan *et al.*, Systems biology-defined NF-kappaB regulons, interacting signal pathways and networks are implicated in the malignant phenotype of head and neck cancer cell lines differing in p53 status. *Genome Biology* **9**, R53 (2008).
69. H. L. Pahl, Activators and target genes of Rel/NF-kB transcription factors. *Oncogene* **18**, 6853 (22 November 1999).
70. M. S. Hayden, S. Ghosh, NF-κB, the first quarter-century: remarkable progress and outstanding questions. *Genes & Development* **26**, 203 (February 1, 2012).
71. S. L. Beinke, Steven C., Functions of NF-kappaB1 and NF-kappaB2 in immune cell biology. *Biochemical Journal* **382**, 393 (2004).
72. J. Dutta, Y. Fan, N. Gupta, G. Fan, C. Gelinas, Current insights into the regulation of programmed cell death by NF-kappaB. *Oncogene* **25**, 6800 (2000).
73. S. Bacher, M. L. Schmitz, The NF-kB Pathway as a Potential Target for Autoimmune Disease Therapy. *Current Pharmaceutical Design* **10**, 2827 (2004).
74. T. D. Gilmore, Introduction to NF-kappaB: players, pathways, perspectives. *Oncogene* **25**, 6680 (2000).
75. M. S. Hayden, A. P. West, S. Ghosh, NF-kappaB and the immune response. *Oncogene* **25**, 6758 (2000).
76. L. Ferrero-Miliani, O. H. Nielsen, P. S. Andersen, S. E. Girardin, Chronic inflammation: importance of NOD2 and NALP3 in interleukin-1β generation. *Clinical & Experimental Immunology* **147**, 227 (2007).
77. E. Rai, E. K. Wakeland, Genetic predisposition to autoimmunity - What have we learned? *Seminars in Immunology* **23**, 67 (2011).
78. The International HapMap Project. *Nature* **426**, 789 (2003).
79. I. A. Udalova, R. Mott, D. Field, D. Kwiatkowski, Quantitative prediction of NF-κB DNA– protein interactions. *Proceedings of the National Academy of Sciences* **99**, 8167 (June 11, 2002).
80. J. Hiscott, H. Kwon, P. Génin, Hostile takeovers: viral appropriation of the NF-kB pathway. *The Journal of Clinical Investigation* **107**, 143 (2001).
81. J. Ruland, Return to homeostasis: downregulation of NF-kappaB responses. *Nat Immunol* **12**, 709 (2011).

82. E. T. Cirulli, D. B. Goldstein, In vitro assays fail to predict in vivo effects of regulatory polymorphisms. *Human Molecular Genetics* **16**, 1931 (August 15, 2007).
83. Y. Li, C. Willer, S. Sanna, G. Abecasis, Genotype imputation. *Annual Review of Genomics and Human Genetics* **10**, 387 (2009).
84. B. Li, S. M. Leal, Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311 (2008).
85. J. Ernst *et al.*, Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43 (2011).
86. G. Lyß, A. Knorre, T. J. Schmidt, H. L. Pahl, I. Merfort, The anti-inflammatory sesquiterpene lactone helenalin inhibits the transcription factor NF-κB by directly targeting p65. *Journal of Biological Chemistry* **273**, 33508 (December 11, 1998).
87. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet* **41**, 1199 (2009).
88. D. Olmeda *et al.*, SNAI1 Is required for tumor growth and lymph node metastasis of human breast carcinoma MDA-MB-231 cells. *Cancer Research* **67**, 11721 (December 15, 2007).
89. R. W. Carthew, E. J. Sontheimer, Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**, 642 (2009).
90. P. J. Park, ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669 (2009).
91. W. Zhang *et al.*, Evaluation of genetic variation contributing to differences in gene expression between populations. *The American Journal of Human Genetics* **82**, 631 (2008).
92. S. B. Gabriel *et al.*, The structure of haplotype blocks in the human genome. *Science* **296**, 2225 (June 21, 2002).
93. D. E. Reich *et al.*, Linkage disequilibrium in the human genome. *Nature* **411**, 199 (2001).
94. T. G. Canty *et al.*, Oxidative stress induces NF-κB Nuclear translocation without degradation of IκBα. *Circulation* **100**, II (November 9, 1999).
95. Y.-J. Jeon *et al.*, Annexin A4 interacts with the NF-κB p50 subunit and modulates NF-κB transcriptional activity in a Ca2+-dependent manner. *Cellular and Molecular Life Sciences* **67**, 2271 (2010).
96. C. C. Lord *et al.*, CGI-58/ABHD5-derived signaling lipids rgulate systemic inflammation and insulin action. *Diabetes* **61**, 355 (February 1, 2012).
97. S. Konisti, S. Kiriakidis, E. M. Paleolog, Hypoxia-a key regulator of angiogenesis and inflammation in rheumatoid arthritis. *Nat Rev Rheumatol* **8**, 153 (2012).
98. D. Ricklin, G. Hajishengallis, K. Yang, J. D. Lambris, Complement: a key system for immune surveillance and homeostasis. *Nat Immunol* **11**, 785 (2010).
99. N. Sugano, W. Chen, M. L. Roberts, N. R. Cooper, Epstein-barr virus binding to CD21 activates the initial viral promoter via NF-κB induction. *The Journal of Experimental Medicine* **186**, 731 (August 29, 1997).
100. A. R. Brasier, The nuclear factor-κB–interleukin-6 signalling pathway mediating vascular inflammation. *Cardiovascular Research* **86**, 211 (May 1, 2010).

101. X. Li *et al.*, Genome-wide association studies of asthma indicate opposite immunopathogenesis direction from autoimmune diseases. *Journal of Allergy and Clinical Immunology*.
102. Y. Chen, B. A. Alman, Wnt pathway, an essential role in bone regeneration. *Journal of Cellular Biochemistry* **106**, 353 (2009).
103. J. Chang *et al.*, Noncanonical Wnt-4 Signaling enhances bone regeneration of mesenchymal stem cells in craniofacial defects through activation of p38 MAPK. *Journal of Biological Chemistry* **282**, 30938 (October 19, 2007).
104. U. Styrkarsdottir *et al.*, Multiple genetic loci for bone mineral density and fractures. *New England Journal of Medicine* **358**, 2355 (2008).
105. J. Gertz, E. D. Siggia, B. A. Cohen, Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**, 215 (2009).
106. D. Medici, E. D. Hay, B. R. Olsen, Snail and Slug promote epithelial-mesenchymal transition through β-Catenin–T-cell factor-4-dependent expression of Transforming Growth Factor-β3. *Molecular Biology of the Cell* **19**, 4875 (November 1, 2008).
107. N. T. Dong Wook Han, Andreas Hermann, Kathrin Hemmer, Susanne Höing, Marcos J. Araúzo-Bravo, Holm Zaehres, Guangming Wu, Stefan Frank, Sören Moritz, Boris Greber, Ji Hun Yang, Hoon Taek Lee, Jens C. Schwamborn, Alexander Storch, Hans R. Schöler, Direct reprogramming of fibroblasts into neural stem cells by defined factors. *Cell Stell Cell* **10**, (6 April 2012).

APPENDIX A
Oral Presentations

1. Oral Presentation at Canadian Human and Statistical Genetics Meeting, White Oaks Conference Centre, Niagara Falls, Ontario, Monday April 30th, 2012

Published Abstracts

1. 4th Annual Canadian Human Genetics Conference, Banff, Alberta, April 2011

Validation of *cis*-regulatory SNPs altering disease risk for osteoporosis

Alicia Schiavi, Veronique Adoue, Stephan Busche, Bing Ge, Tony Kwan, Tomi Pastinen

Department of Human Genetics, McGill University, Montréal, Canada

Osteoporosis is a skeletal disorder characterized by compromised bone strength and increased risk of fracture in which the regulation of bone remodeling is imbalanced. Numerous loci contributing to bone mineral density and osteoporosis risk have been recently described by GWAS (Rivanedeira et al. NG 2009); however, the underlying biological effect of many of these variants remains unknown. Elucidating these effects and uncovering the remaining genetic variation is critical to understanding this complex disease.

Using a powerful and highly sensitive method developed in our lab to map *cis*-regulatory variants in human cells (lymphoblasts, monocytes, and fibroblasts), we identified a SNP that affects *WNT4 cis*-regulation in our fibroblast panel and alters the risk for osteoporosis (Rivadeneira et al. 2009). The signaling pathway of *WNT4* has been implicated in bone development. The *cis*-rSNP maps >200kb upstream of the *WNT4* gene and overlaps a site of active chromatin observed in fibroblasts. Consistent with a cell type restricted chromatin signal, we observed the regulatory association only in fibroblasts, indicating that the SNP alters gene regulation in mesenchymal stem cell (MSC) lineage and therefore is directly relevant to bone disease. Bioinformatics analysis of the *cis*-rSNP indicates that it alters a *SNAI1* binding site. We have shown *in vitro* allele-specific EMSA signals in nuclear extracts from MSC lineage (MG-63 cells) and are now pursuing *in vivo* validation of SNAI1 binding in living cells by carrying out ChIP with allele-specific readouts, as well as *SNAI1* knockdown by RNAi with monitoring of its consequences in *WNT4* allelic expression phenotype. Consistent inhibition (85%) of *SNAI1* using RNAi in transfection studies has been observed. We are pursuing on-going allelic expression imbalance assessment of *WNT4* in cells heterozygous for this *cis*-rSNP and treated with efficient *SNAI1* RNAi.

These approaches will be generically extended to other *cis*-rSNPs altering osteoporosis disease risk and we will monitor the consequences of these gene knockdowns in a genome-wide manner.

2. Human Genetics Research Day, McGill University June 2011

Validation of *cis*-regulatory SNPs altering disease risk for osteoporosis

Alicia Schiavi, Veronique Adoue, Stephan Busche, Bing Ge, Tony Kwan, Tomi Pastinen
Department of Human Genetics, McGill University, Montréal, Canada
M.Sc.1, Tomi Pastinen (supervisor)

Osteoporosis is a skeletal disorder characterized by compromised bone strength and increased risk of fracture in which the regulation of bone remodeling is imbalanced. Numerous loci contributing to bone mineral density and osteoporosis risk have been recently described by GWAS (Rivanedeira et al. NG 2009); however, the underlying biological effect of many of these variants remains unknown. Elucidating these effects and uncovering the remaining genetic variation is critical to understanding this complex disease.

Using a powerful and highly sensitive method developed in our lab to map *cis*-regulatory variants in human cells (lymphoblasts, monocytes, and fibroblasts), we identified a SNP that affects *WNT4 cis*-regulation in our fibroblast panel and alters the risk for osteoporosis (Rivadeneira et al. 2009). The signaling pathway of *WNT4* has been implicated in bone development. The *cis*-rSNP maps >200kb upstream of the *WNT4* gene and overlaps a site of active chromatin observed in fibroblasts. Consistent with a cell type restricted chromatin signal, we observed the regulatory association only in fibroblasts, indicating that the SNP alters gene regulation in mesenchymal stem cell (MSC) lineage and therefore is directly relevant to bone disease. Bioinformatics analysis of the *cis*-rSNP indicates that it alters a *SNAI1* binding site. We have shown *in vitro* allele-specific EMSA signals in nuclear extracts from MSC lineage (MG-63 cells) and are now pursuing *in vivo* validation of SNAI1 binding in living cells by carrying out ChIP with allele-specific readouts, as well as *SNAI1* knockdown by RNAi with monitoring of its consequences in *WNT4* allelic expression phenotype. Consistent inhibition (85%) of *SNAI1* using RNAi in transfection studies has been observed. We are pursuing on-going allelic expression imbalance assessment of *WNT4* in cells heterozygous for this *cis*-rSNP and treated with efficient *SNAI1* RNAi.

These approaches will be generically extended to other *cis*-rSNPs altering osteoporosis disease risk and we will monitor the consequences of these gene knockdowns in a genome-wide manner.

3 key words/phrases
1) *cis*-regulatory variants affecting allelic expression
2) Genome-wide association studies
3) risk of osteoporosis

3. ICHG/ASHG 2011, Montreal, Quebec, October 2011

A. Schiavi[1,2], V. Adoue[1,2,] S. Busche [1,2,] B. Ge[2], T. Kwan[1,2,] T. Pastinen[1,2]
1)Human Genetics, McGill University, Montreal, Quebec, Canada; 2) McGill University and Genome Quebec Innovation Centre, Montreal, Quebec,Canada

Direct assessment and validation of allele-specific transcription factor binding in the human genome

Characterization of human genetic variation, which effects gene expression, has focused on expression quantitative trait loci (eQTL) mapping; however, direct assessment of *cis*-regulatory variation necessitates allele-specific approaches. Measuring allelic expression (AE) on a genome-wide scale is more powerful as environmental and trans-acting influences are minimized. Results indicate that allele-specific differences among transcripts within an individual can affect up to 30% of loci. These variants can be identified by mapping differences in AE on Illumina HumanOmni-1M/2.5M BeadChips. Over 50% of population variance in AE is explained by mapped *cis*-rSNPs. Studies show that these *cis*-rSNPs have been implicated in differences in transcription factor binding, suggesting a strong genetic component that needs to be further investigated. Combination of multiple GWAS datasets, eQTLs in osteoblasts, AE in primary fibroblasts and DHS-seq from mesenchymal stem cell lineage (MSC) allowed determination of a single SNP >200kb upstream of the *WNT4* gene. Consistent with a cell type restricted chromatin signal, we observed the regulatory association only in fibroblasts, indicating that the SNP alters gene regulation in MSC lineage and therefore is directly relevant to bone disease. Numerous loci contributing to bone mineral density and osteoporosis risk have been described by GWAS; however, the underlying biological effect of many of these variants remains unknown. Bioinformatic analysis of the *cis*-rSNP indicates that it alters a *SNAI1* binding site. We have shown in vitro allele-specific EMSA signals in nuclear extracts from MSC lineage and are now pursuing in vivo validation of *SNAI1* binding in living cells by carrying out ChIP with allele-specific readouts, as well as, *SNAI1* knockdown by RNAi with monitoring of its consequences in *WNT4* allelic expression phenotype. Consistent inhibition (85%) of *SNAI1* using RNAi in transfection studies has been observed. We are pursuing allelic expression imbalance assessment of *WNT4* in cells heterozygous for this *cis*-rSNP and treated with efficient *SNAI1* RNAi. These approaches will be generically extended to other allele-specific transcription factor binding and the consequences of these gene knockdowns will be monitored in a genome-wide
manner. In progress is work on the NF-κB transcription factor that has been shown to be involved in the immune response and where the NF-κB motif is enriched in lymphoblastoid cell lines.

4. Canadian Human and Statistical Genetics Meeting, Niagara Falls, Ontario, April 2012

Direct assessment and validation of allele-specific transcription factor binding in the human genome

Alicia Schiavi, Veronique Adoue, Stephan Busche, Bing Ge, Tony Kwan, Tomi Pastinen
Department of Human Genetics, McGill University, Montréal, Canada

Characterization of human genetic variation has focused on expression quantitative trait loci (eQTL) mapping; however, direct assessment of *cis*-regulatory variation requires allele-specific approaches. Measuring allelic expression (AE) on a genome-wide scale appears more powerful as environmental and trans-acting influences are minimized (Pastinen, *Nat. Rev. Gen.*, 2010). Results indicate that allele-specific differences in transcript expression within an individual can affect up to 30% of loci. The underlying variants can be identified by mapping differences in AE on Illumina BeadChips. Over 50% of population variance in AE is explained by mapped *cis*-rSNPs. Studies show that these *cis*-rSNPs have been implicated in differences in transcription factor binding, suggesting a strong genetic component that needs to be further investigated.

These approaches were extended to analyze allele-specific transcription factor binding by monitoring the consequences of gene knockdowns in a genome-wide manner. NF-κB has been shown to be involved in the immune response and the NF-κB motif is enriched in lymphoblastoid cell lines, mainly in promoters and strong enhancer chromatin states (Bernstein et al. *NG* 2011). We intersected mapped candidate *cis*-rSNPs detected in lymphoblastoid cells in our above experiments as well as matched control SNPs from HapMap YRI and CEU populations with publicly available NF-κB Chromatin Immunoprecipitation (ChIP)-seq experiments from the ENCODE project. We observed that regions surrounding candidate *cis-r*SNPs are enriched in NF-κB binding sites versus matched controls, with 37.4 % of top SNPs overlapping at least one NF-κB ChIP-seq peak. To elucidate the impact of candidate SNPs on AE imbalances, we performed TNF-a induction coupled to inhibition of NF-κB in lymphoblastoid cell lines followed by AE analysis on Illumina HumanOmni5-Quad BeadChips. We detected enrichment in NF-κB binding sites in samples induced with TNF-α versus control. On-going validation includes two SNPs, *rs11204415* and *rs2170577* associated with loci *ALDH3A1* and *WDR17*, respectively, which show a decrease in AE upon inhibition of NF-κB. Bioinformatic analysis suggests the identified SNPs to be essential for NF-κB binding.

APPENDIX B
Publications

Schiavi, A., N. Light, et al. (2011). "Human genetics in full resolution." Genome Biology **12**(11): 309.