ABSTRACT

IDAGHDOUR, YOUSSEF. Genetic and Environmental Components of Human Leukocyte Gene Expression Variation in Morocco. (Under the direction of Dr. Greg Gibson).

In humans, study of how environment and genome interact to shape phenotypic variation is particularly relevant to understanding the origins of complex diseases and the increase in their prevalence coinciding with major shifts from traditional to urbanized lifestyles. Gene expression is the first step in a complex and multi-step process towards the production of higher-level phenotypes. While genetic analysis of gene expression variation in humans using cell lines and clinical samples is subject to extensive research, little is known about the contributions of environment and geography to transcriptional variation.

To estimate these contributions, I examined gene expression in peripheral blood leukocyte samples from 46 desert nomadic, mountain agrarian and coastal urban Moroccan Amazigh individuals. Strikingly, as much as one third of the leukocyte transcriptome was found to be differentially expressed among lifestyles. Genome-wide polymorphism analysis indicates that genetic differentiation in the total sample is limited and is unlikely to explain the expression divergence. Methylation profiling of 1,505 CpG sites suggests limited contribution of methylation to the observed differences in gene expression.

To estimate the contributions of genetic and environmental factors jointly, I generated gene expression profiles and whole genome genotypic data from 208 and 203 individuals, respectively, from leukocyte samples of two groups of urban dwellers and two groups of rural villagers representing Arab and Amazigh ethnicities in southern Morocco. Again, the analysis revealed strong effects of environmental geography but also suggested that the interplay between non-genetic environmental factors and genes is a major modulator of the
transcriptome. Both studies confirm that genetic factors are neither the sole, nor even the major, source of variation affecting the leukocyte transcriptome. The amplitude and functional characteristics of the observed differences in both studies suggest a significant impact on immune response and disease susceptibility.

The analysis was extended by performing a gene expression genome-wide association where each of 516,972 genotypes was tested against each of 22,300 expressed transcripts while accounting for gender, location, genetic ethnicity, relatedness, and interaction effects. This analysis revealed 1,744 genome-wide significant associations involving 380 cis-eSNP and 16 trans-eSNP all robust to population structure and environmental modulation. No evidence for genotype-by-environment interactions modulating transcript abundance was detected for the genome-wide significant associations suggesting environmental geography and genotypes act in a largely additive manner. Further, I confirmed numerous previously reported regulatory signals and validated several previous trait and disease GWAS findings as being modulated by variation in gene expression levels. These results further our understanding of gene expression variation in humans and emphasize the biomedical importance of expression divergence within and among human populations.
Genetic and Environmental Components of Human Leukocyte Gene Expression Variation in Morocco

by
Youssef Idaghdour

A dissertation submitted to the Graduate Faculty of North Carolina State University
In partial fulfillment of the requirements for the Degree of Doctor in Philosophy

Genetics
Raleigh, North Carolina
2009

APPROVED BY:

Chair of Advisory Committee
Dr. Greg Gibson

Co-Chair of Advisory Committee
Dr. Owen McMillan

Committee Member
Dr. Eric Stone

Committee Member
Dr. Todd Klaenhammer
DEDICATION

In Loving Memory of my Mother, Rokaya Boubrik of Ighrem
BIOGRAPHY

I was born on February 11, 1975 in the heart of the Amazigh region of Souss in southern Morocco. I am the youngest of eight brothers and sisters, and I have been inspired throughout my life by the vision of two wonderful women, my sisters Aicha and Malika. I owe them my love of biology and my sense of wonder.

I joined a wildlife conservation institution after I received a BS in Biology. I developed an interest in the emerging field of conservation genetics and volunteered to help save the last wild population of bald ibis in the world. Afterwards, I took a year off and pursued a Masters degree in Molecular Genetics in England. I then moved back to Morocco and attempted to solve some of the challenges in wildlife conservation projects using genetics.

In spring of 2004 I was intrigued by a series of ideas in a paper published in the journal Molecular Ecology and contacted the author, Greg Gibson, and that is how the PhD chapter of my life started. In the following year, I was awarded a Fulbright Scholarship and relocated to the US with my family.
ACKNOWLEDGEMENTS

For the opportunity to be a Gibsonite and for taking a chance on me, I am forever grateful to Greg Gibson. I must thank him for giving me the freedom to switch to human genetics and for his demanding intellect that has been indispensable for the development of the ideas described in this thesis.

Thanks to my committee members. I want to thank Dr. Eric Stone for stimulating discussions and helping with aspects of data analysis. Thanks to Dr. Owen McMillan for providing biological insight and lab space after Greg’s move. Thanks also to Dr. Tod Klaenhammer and my former committee member Dr. Greg Dean for serving in my committee. Special thanks to Dr. Stephanie Curtis for her advice and guidance.

I cannot thank enough the many people in Morocco who directly contributed to my research. I am especially grateful to my family in particular, Malika, Aicha, Ahmed, Hassan, Abdelaziz, Nouzha, Fahd, Wafa, and Si Boujemaa, who provided crucial help from early stages of the project. This work would have not been possible without their contribution, and that of the study participants and many volunteers. Special thanks to Dr. Sami Jadallah for his encouragements and support to the project. His input was essential to obtaining collection permits in Morocco. Thanks to Zaid and Abdelatif for their involvement in sampling the Bedouins.

For useful discussions and assistance with data analysis, I would like to thank Rachel Myers, Dr. John Storey, Dr. Alison Motsinger-Reif, Dr. Russ Wolfinger, Wendy Czika, Dr.
Kelci Miclaus, Dr. Peter Vischer, Dr. Lee Hong and Dr. Dongliang Ge. For providing access to Illumina equipment, thanks to Dr. David Goldstein, Dr. Kevin Shianna, and Dr. Simon Gregory.

Thanks to the Gibson-McMillan group that has been a source of support and stimulating discussions. I have been fortunate to be part of a wonderful group. Erin Kennerly, Ian Dworkin, Julien Ayroles, Lisa Goering, Laura Reed, Stephanie Williams, Felix Araujo, Vinnie Izzi and Brian Counterman for the past four years, were an endless source of discussion…. and fun. There are numerous other friends and members of the Gibson-McMillan group that I cannot mention them all here. Thanks also to my colleagues and members at the Genetics Department, in particular Suzanne Quick for help at crucial moments.

For the opportunity to be a Fulbrighter, I thank the Moroccan-American Commission for Educational and Cultural Exchange. I hope my thesis work, at least in some small way, is in the spirit of the Fulbright legacy.

For sacrificing a great deal so that I have been able to pursue my dreams, thanks to my wife Nouzha. Without her love, none of this would have ever happened. Finally, I would like to express my love for my children, Samia and Reyad, for warming my heart.
# TABLE OF CONTENTS

List of Tables ......................................................................................................................... viii
List of Figures ........................................................................................................................ ix

## CHAPTER ONE: Introduction ................................................................. 1
   Overview .................................................................................................................... 2
   Phenotypic Variation ................................................................................................. 3
   The Human Legacy .................................................................................................... 6
   Gene Expression ........................................................................................................ 11
   Environmental Transcriptomics ................................................................................. 19
   Outline of the Thesis .................................................................................................. 27
   References .................................................................................................................. 28

## CHAPTER TWO: A Genome-wide Gene Expression Signature of Environmental Geography in Leukocytes of Moroccan Amazighs ................................................... 35
   Abstract ..................................................................................................................... 37
   Introduction ................................................................................................................ 37
   Results/Discussion ..................................................................................................... 37
   Conclusion .................................................................................................................. 42
   Materials and Methods ............................................................................................... 42
   References .................................................................................................................. 47

## CHAPTER THREE: Genetic and Environmental Components of the Human Leukocyte Transcriptome Variation ............................................................... 54
   Abstract ...................................................................................................................... 55
   Introduction ................................................................................................................ 56
   Materials and Methods ............................................................................................... 57
   Results ........................................................................................................................ 64
   Discussion .................................................................................................................. 77
   References .................................................................................................................. 86

## CHAPTER FOUR: Mapping the Genetic Architecture of the Leukocyte Transcriptome ........................................................ 102
   Abstract ..................................................................................................................... 103
   Introduction ................................................................................................................ 105
   Materials and Methods ............................................................................................... 107
   Results ........................................................................................................................ 110
   Discussion .................................................................................................................. 128
References ........................................................................................................................................ 136

CHAPTER FIVE: Conclusions and Future Possible Directions ........................................... 150
Conclusions .................................................................................................................................. 151
Possible Future Directions ....................................................................................................... 167
References ................................................................................................................................. 170
LIST OF TABLES

CHAPTER TWO
   Table 1: Number of differentially expressed genes ..................................................39

CHAPTER THREE
   Table 1: Differential expression................................................................................89
   Table 2: Panther enrichment analysis............................................................................90
LIST OF FIGURES

CHAPTER TWO

Figure 1: Geographic locations of sampled Amazighs groups in Morocco..............38
Figure 2: Volcano plots of statistical significance versus magnitude of differential expression between locations .................................................................39
Figure 3: Sex and location effects on methylation patterns ........................................41
Figure 4: Functional analysis of differentially expressed genes ...............................43
Figure S1: Two-way hierarchical clustering of expression .....................................47
Figure S2: Effect of surrogate variable inclusion on significance testing ...............48
Figure S3: Eigenstrat analysis of population structure .............................................49
Figure S4: Structure analysis of genotypic variation ..............................................50
Figure S5: Genomewide genetic and gene expression differentiation ......................51
Figure S6: Results of methylation analysis ...............................................................52
Figure S7: Correlations among Interleukin signaling components ...........................53

CHAPTER THREE

Figure 1: Geographic locations of sampled groups in the Souss region in southern Morocco ...................................................................................................91
Figure 2: Population structure analysis .....................................................................92
Figure 3: Two-way hierarchical clustering of expression data from Agadir and Boutroch individuals .........................................................................................93
Figure 4: Principal and variance component analyses of the leukocyte transcriptome .......................................................................................................................94
Figure 5: Principal component analyses of differentially expressed genes ..........95
Figure 6: Volcano plots of statistical significance versus magnitude of differential expression between regions .................................................................96
Figure 7: Overlap between differentially expressed genes .......................................97
Figure 8: Effects of Ethnicity, Gender, and their interaction in the urban sample ...98
Figure 9: Panther enrichment analysis .....................................................................99
Figure 10: KEGG Ribosome assembly map .............................................................100
Figure 11: KEGG Oxidative phosphorylation pathway ............................................101

CHAPTER FOUR

Figure 1: Gene expression data distribution .............................................................141
Figure 2: Genome-wide SNP (A) and gene-specific (B) Fst measures .................142
Figure 3: Genome-wide genetic and gene expression differentiation ..................143
Figure 4: Association model and effect comparisons ........................................144
Figure 5: Manhattan plots for genome-wide significant associations .................145
Figure 6: Properties of 1,744 genome-wide significant associations .................146
Figure 7: eSNP and target genes distribution ....................................................147
Figure 8: _cis_ versus _trans_ effects ....................................................................148
Figure 9: BCL11A _trans_ regulation of gamma globin genes ............................149
CHAPTER ONE

Introduction
Overview

In regard to human biology, much emphasis has recently been placed on our genetic make-up and little on the environment. This paradigm is often naively expressed by the media as the “nature versus nurture debate”. The two factors do not exclude each but our environment is changing, certainly faster and more than our genome does, suggesting that a “nature interacting with nurture discussion” would be more productive. The research presented here explores the mode and amplitude of effects of genetic ethnicity, the environment and their interactions on gene expression in humans by addressing the following questions. First, what is the magnitude of the impact the environment has on transcript abundance and can strongly affected aspects of immune function be detected? Second, is it possible to combine genotypic and gene expression data to better estimate the relative contributions of genetic ethnicity, environment and their interactions to gene expression variation? Third, what are the genetic factors that influence the environmental differences that are observed? Fourth, how robust is the genotype-transcript abundance relationship and what are the implications from both biomedical and evolutionary perspectives? The finding of this research might have important consequences for the study of the architecture of both normal and disease variation in humans.
In the following sections of this introduction, I first review the approaches that are used to try to partition sources of phenotypic variation in general. I then examine both the historical and genetic scope of human evolution and discuss the role the environment might be playing in shaping human biology taking a modern lifestyle transition as an example. I also discuss the current state of genotype-phenotype mapping in humans and some of the challenges that remain to be solved. This leads to a discussion of the role gene expression may play in shaping phenotypes and the advantages and the challenges studies of the transcriptome pose for unraveling the genetic basis of complex traits, emphasizing on the emerging trend of using gene expression network analysis to infer causation. Finally, I address some specific issues relating to transcriptional gene-environment interactions in human peripheral blood, which is the subject of the research presented in this thesis.

**Phenotypic Variation**

Evolutionary biologists celebrate Charles Darwin’s 200th birthday this year, as well as the anniversary of the founding of the field with the publication of Darwin and Wallace’s work 150 years ago. This also provides an occasion to evaluate the state of our contemporary understanding of evolutionary processes. Interestingly, decades after the re-discovery of Mendel’s principles and the establishment of the Modern Synthesis and as we are entering the post-genome era, the quest to understand each of the evolutionary
forces driving phenotypic variation and their proportionate contributions is still a fundamental component of modern evolutionary biology research.

Variation at all levels of biological organization is astonishing and is a result of the evolution of life that commenced nearly 3.8 billion years ago. Natural variation ranges from the simplest to the most complex forms. The field of quantitative genetics has emerged as an effort to characterize the components that contribute to biological variation as well as to understand the mechanisms through which it is shaped and evolves. The basis of quantitative genetic theory can be simplified as follows. Phenotypic variation is partitioned into three major components: variation due to genetic factors, variation due to the environment, and variation due to the interaction between these two factors (reviewed in Visscher et al. 2008). The integration of this central idea with evolutionary developmental biology concepts such as modularity, robustness and canalization sets the framework for the standard paradigm of modern evolutionary biology. Nonetheless, the pursuit of the ultimate goal of mapping genotype onto phenotypic or vice versa, and understanding the mechanisms involved is arguably in crisis, at least for complex traits. As a result, an appreciation of the complexity of the architecture of complex traits is rapidly gaining momentum.

For most traits, the genotype-phenotype relationship puzzle is highly complex. There are many components to it but one can only start to see the big picture by characterizing and then assembling all the little pieces that contribute to it. Most of the
complexity lies beneath the higher-level phenotypes, moving from the architecture of the genome itself, and transitioning through the complex modes of gene expression and post-transcriptional regulation to protein function. The particularly pervasive insight that is coming from the field of quantitative genetics is a new appreciation of the relevance of polygenic inheritance, epistasis, non-additive effects, as well as genotype-by-environment interactions that take effect and greatly contribute to phenotypic variation (Gibson 1996, reviewed in Anholt & Trudy 2004). More recently, our enhanced ability to detect the effects of these facets of biological complexity and to study the various whole-genome layers of biological processes collectively, rather than one or a few genes at a time, has fostered the adoption of systems biology approaches to the study of evolution.

Most of our current understanding of the evolutionary processes driving phenotypic variation has emerged from experiments conducted in model organisms. They provide a convenient framework to conceptualize the questions of interest, test hypotheses, set up genetic crosses, and perform functional tests. Although we have gained valuable insight into the processes of evolution in general, generalizations might not always be justified. Extrapolation to humans should be formulated especially carefully, given the uniqueness of many of the qualitative and quantitative characteristics that have shaped our biology for tens of thousands of years.
The Human Legacy

**Human Evolutionary History**

Whether *Homo sapiens* is a masterpiece of evolution or rather Pandora’s Box, it is ultimately an extraordinary species evolutionarily. While this might be a self-centric view of the organic world, we certainly have shaped life on earth and developed functions and levels of consciousness more than any other species since the origin of life. Anecdotally, as of 2008, humans are listed as a species of “least concern” for extinction by the International Union for Conservation of Nature (Mittermeier *et al.* 2008).

Although anthropologists have not settled on one model for the hominin evolutionary tree, the “Out of Africa” model of *Homo sapiens* origin is currently widely accepted based on fossil and genetic evidence. The model states that the species descended from *Homo erectus*, evolved around 200 kya in the African Savanna, migrated into the rest of Africa and Eurasia around 100-40 kya, and colonized the Americas approximately 20-10 kya (Templeton 2002). With the rise of civilization, modern humans have colonized all the continents, adapted to all climates and environments, and explored almost all niches on earth. Humans throughout their journey have also shown a great capacity for altering their habitats through urban planning, agriculture, deforestation, industrial activities, and other cultural innovations. A focus of this thesis is the impact of these human capacities on our own evolutionary potential as a species.
The Human Genetic Make-up

Several studies using different types of markers have reported small effective population sizes of non-African human populations (from a few hundred to a few thousand individuals) as a result of a series of population bottlenecks that are thought to have accompanied migrations of humans out of Africa (reviewed in Campbell & Tishkoff 2008). This scenario certainly resulted in a loss of genetic diversity (Ramachandran et al. 2005, Liu et al. 2006) as confirmed by the observation that non-Africans compared to Africans show lower levels of both mitochondrial and nuclear diversity (The International HapMap Consortium 2005, Jakobsson et al. 2008, Li et al. 2008). Nonetheless, the existing genotypic variation among human populations is sufficient to be used to statistically distinguish most ethnic groups from each other (Rosenberg et al. 2002, Wang et al. 2007, Wang et al. 2008, Friedlaender et al. 2008), and to generate phenotypic variation that has the potential to fuel evolutionary processes. Moreover, the emerging field of pharmacogenetics and the flood of disease association studies are ultimately an appreciation of the epidemiological relevance of variation in frequencies of variants influencing disease risk or drug response (Wilson et al. 2001, Risch et al. 2002).

Nevertheless, the over-appreciation of the role our genetic make-up has played as the primary factor shaping human phenotypic variation is currently being questioned. The bold and ambitious statements regarding the implications on human health that were
made in the early 1990s to push for increased funding for the human genome project certainly helped move the field forward, as they largely motivated genome sequencing and the International HapMap projects. The tremendous knowledge regarding Mendelian diseases that was gained from these projects is however arguably being overshadowed by the findings of studies attempting to discern the architecture of complex genetic disorders. For example, despite the flood of genetics data in the last decade, the identification of genetic variants that contribute to diseases is hampered by the complexity of the etiology of human disease. Although statistical significance in linkage and association studies has recently become less of an issue, most of the variants that have been detected account for only a small fraction of the phenotypic variation and often fail to replicate in different populations. This issue is arguably the most pressing challenge for genome-wide association studies (GWAS). The recent calls for performing meta-analyses on large sample size datasets strengthen the evidence that the genetic effect of single or even few markers to complex diseases is weaker than previously thought (NCI-NHGRI Working Group on Replication in Association Studies, et al. 2007) or is being weakened by other effects. The same observation is certainly true for other complex traits in humans and other species.

These findings should not diminish the importance of the genetic basis of phenotypic variation, and instead underscore the view that our grasp of the nature of our genetic make-up is too simplistic to account for most of the relevant genotypic variation.
It is worth noting for example that the markers used in association studies are of a minor allele frequency greater than 5% in the HapMap populations, namely Yorubans, Caucasians, Chinese, and Japanese (The International HapMap Consortium 2005) and perhaps should not be considered reliable enough to represent most of the genotypic variation that is shaping phenotypes. This idea is more appreciated as researchers begin to focus on the contributions of rare variants (Bodmer & Bonilla 2008) and attempt to integrate copy-number variants (McCarroll et al. 2008, Barnes et al. 2008, Korn et al. 2008), as well other structural and epigenetic variants in their analyses.

It might also be that the role other components contributing to phenotypic variation are playing is being underestimated and their effects are not as much appreciated as genetic variation. Environmental factors undoubtedly have shaped human biology through both direct effect and gene-environment interactions. It is therefore important to disentangle these effects to unravel the architecture of phenotypic traits of interest not just in laboratory settings but also in natural populations. In the next section, I discuss a major environmental shift in the history of humans that is certainly shaping the biology of our genome.

**The Human Lifestyle Transitions**

Humans have experienced multiple lifestyle transitions throughout history. Until the Neolithic period, early humans adopted a subsistence strategy of nomadic hunting and
gathering that underwent a gradual transition to agricultural and agrarian-based lifestyles. Likely the most important transition began around 9000-7000 BC in the Fertile Crescent. The transition was not a clear cut-off and it has continued up until very recent history. A yet more recent major transition in human lifestyles is urbanization. It is a process characterized by an increasing concentration of the population in towns and cities and it is associated with major changes in both human activities and surrounding environment. This process was accentuated by the advancement of knowledge and technology, most dramatically since the phenomenal population growth in the last century. There are different forms of urbanization that can be classified depending on the degree of concentration of the population, the type of land-use, and the style of the architecture. Quality of life is associated with either all or a combination of sub-optimal conditions that are known to have an effect on our physical and/or mental health.

In most countries, urbanization is fueled predominantly by rural immigration. Rural people move into cities to seek economic opportunities and improvement of their living standard. Nevertheless, newer rural immigrants often experience poverty and settle in slums characterized by sub-standard housing, unsanitary conditions, and high population density. The quality of life in urbanized settlements usually ranges from much deteriorated slums to high-end gated communities, but approaching one quarter of urban dwellers across the world can be classified in the low-range of life conditions. Only in the past decade has the emergence of a global middle class begun to reduce this
burden, though it too is associated with a global shift in the profile of chronic diseases. Thus, the phenomenon of urbanization coincides with the emergence of new diseases and with the increase in prevalence of complex diseases such as diabetes, asthma, depression, and hypertension that almost reached an epidemic status in the last few generations. A possible explanation of this observation can be attributed to an increasing number of incompatibilities between existing segregating genetic variation and perturbed environments. The genome of modern humans has certainly evolved to cope with different sources of perturbations under the effect of different evolutionary forces. However, the increased rate of accumulation of perturbations in the last few generations is probably beyond the pace at which the system can achieve stability.

Gene Expression

*Variation in Gene Expression*

The regulation of gene expression can be defined as the set of reactions that control the abundance of transcripts. It is the first step in a complex and multi-step process towards the production of phenotypes and arguably the most important component in the genetic basis of higher-level phenotypic variation (Pigliucci 1996, Wray et al. 2003, Caroll 2005). Gene expression is more immediately connected to genotype than the visible traits of an organisms and its variation has been observed since the early days of modern molecular biology. There is no doubt this variation is widespread and
present at all levels of biological organization including among cells within the same tissue (Raser & O’Shea 2005). This variation can be explained by the number, the complexity, and the nature of its intrinsic and extrinsic modulators that can be variable at the cell, tissue, organism, or population level. Modulation of gene expression can occur through multiple genetic and epigenetic mechanisms such as promoters, activators, enhancers, repressors, \textit{trans} effectors, chromatin and methylation state (reviewed in Wray \textit{et al.} 2003, and Jaenisch & Bird 2003). External stimuli and substances such as temperature, microorganisms, drugs, and chemicals also can modulate gene expression and therefore contribute to its variation.

\textit{Quantitative and Population Genetics of Gene Expression}

Genetical genomics refers to genetic mapping of genome-wide gene expression by taking advantage of the joint analysis of genetic and genome-wide gene expression data with the intention of gaining greater insight than can be provided by either type of data alone (Jansen & Nap 2001, Stamatoyannopoulos \textit{et al.} 2004, Gibson & Weir 2005, Li & Burmeister 2005). This relatively recent line of research was motivated by the basic idea that transcript abundance is a quantitative trait and consequently classical quantitative linkage mapping (QTL) methods can be used to map it to \textit{cis}- and \textit{trans}-acting sources of variation, also referred to as local and distant QTLs, respectively (Rockman & Kruglyak, 2006). The concept of genetical genomics and gene expression heritability are tightly linked. A heritable component to gene expression would necessarily mean the presence of determinants of gene expression
variation that can then be mapped. Expression QTLs (eQTLs) are therefore loci that account for a significant fraction of transcript variation within a cross, while eSNPs account for variation in transcript abundance among unrelated individuals.

E SNPs can now be reproducibly identified by genome-wide screens in multiple populations. For example, Stranger et al. (2007) used transcriptional profiles of lymphoblast cell lines from 210 unrelated individuals in the four HapMap populations (Europeans, Yoruban Africans, Han Chinese, and Japanese) and their genotypes at over 2 million SNPs to identify eSNPs. 831 genes showing a significant \textit{cis} association in at least one population, 310 genes in at least two populations, and 62 in all four populations were uncovered. Further association tests were performed on the full sample of all 210 unrelated individuals, a subset of three populations, and two populations while controlling for the effect of population differentiation by performing conditional permutations. 803, 735, and 651 genes were uncovered for the four-population, three-population and the two-population pools, respectively. This study uncovered at least 1,348 \textit{cis}- and 180 \textit{trans}-regulatory variants in total. 57 of the \textit{cis} eSNPs and 5 of the \textit{trans} eSNPs were found in all four populations. At least for 95\% of these, the direction of the allelic effect was the same across populations. This study was recently extended to 830 unrelated individuals from eight global populations of the extended HapMap samples using 1.6 million SNP genotypes from the HapMap3 Consortium and thousands of copy number variant (CNV) markers allowing the detection of \textit{cis}-acting genetic factors for
approximately 20% of expressed genes in all the populations combined, with 34% of detected eSNPs observed in at least two populations (Barbara Stranger, unpublished data). This follow-up study also allowed the identification of eCNVs, and it was observed that only 7-14% of CNV associations are also detectable with SNPs.

Myriad studies have successfully identified regulatory variants in many other species (reviewed in Rockman & Kruglyak 2006, Williams et al. 2007) and new insight has emerged as to the estimates of transcript heritability (Visscher et al. 2008) and the proportionate contribution of cis- and trans-acting genetic variants (Dixon et al. 2007, Göring et al. 2007, Stranger et al. 2007). Interestingly, the broad-sense heritability estimates of gene expression levels in lymphoblastoid cell lines or lymphocytes from monozygotic twin pairs (McRae et al. 2007), a British asthma cohort (Dixon et al. 2007) and a Latin American heart disease cohort (Göring et al. 2007), were largely congruent with values varying between 0 and 0.8 and a median approximating 0.3. These studies and others (Stranger et al. 2007) reported that the vast majority of detected regulatory variants act locally in the genome, supporting the inference that there is an abundance of cis-regulatory variation in the human transcriptome. Also, trans-eSNPs were reported to act on multiple genes and also often affect the expression of key regulatory factors. However, little is known about the distribution of eSNPs in diverse human populations and almost nothing about their robustness to environmental perturbation.

Significant progress in the field of population genetics of gene expression has been hampered by the absence of a theoretical framework to interpret and develop
parameters that characterize population structure. A comprehensive neutral model for transcriptional evolution that is equivalent to Kimura’s neutral theory for molecular evolution is much needed to frame a null hypothesis to test hypotheses of interest against (Khaitovich et al. 2004). It is worth noting that an increasing number of studies are estimating the contribution of different evolutionary forces to shaping the evolution of the transcriptome (Rifkin et al. 2005, Denver et al. 2005, Landry et al. 2007) providing ground for new statistical genetics techniques.

Rosenberg et al. (2002) generated an accurate picture of human population structure at the genotypic level using the Structure algorithm. Population stratification refers to allele frequency differences between groups of individuals due to systematic ancestry differences and is accounted for by adjusting association statistics using different methods (Pritchard et al. 2000, Reich & Goldstein 2001, Satten et al. 2001, Zheng et al. 2005, Setakis et al. 2006). Because many regulatory polymorphisms in general are known to be segregating at different allelic frequencies in different populations, it is likely that a signature of population stratification is also present at the transcriptional level and therefore needs to be accounted for when performing transcript abundance-phenotype association, to avoid spurious signals. Also for this reason, it has been suggested that differences in frequency of transcript abundance classes between case and control groups could lead to the false inference that a particular transcript is causative in phenotype promotion (Gibson 2003).


**Gene Expression-Phenotype Mapping**

Although tremendous progress has been made in methods of quantifying gene expression levels that can simultaneously and robustly be measured for tens of thousands of transcripts at relatively affordable costs (Gibson 2002), discerning gene expression signatures that capture causal influences on the architecture of complex traits remains a major task for investigators. Studies on microorganisms have been quite successful in linking expression profiles to metabolic and other organismic-level phenotypes, primarily as a consequence of the tight link in time and space between transcript abundance and subsequent phenotypes. One might argue that unraveling transcript-phenotype associations in higher organisms might not be promising given the extent of biological complexity of the black box between genotype and phenotype. It might also be true that some transcriptional changes are silent and do not result in phenotypic change. Furthermore, distinct transcription profiles might converge and yield similar phenotypes as well as the possibility that mechanisms of post-transcriptional regulation may account for buffering transcriptional variation and therefore break the otherwise statistically significant associations between gene expression levels and higher-level phenotypes.

It is important to note that our current understanding of the nature of associations between gene expression changes and higher-level phenotypes is to a great extent dependent upon how gene expression variation is quantified and made sense of. Most of the standard methods that have been in use in the last two decades to analyze genome-
wide expression are typically laden with assumptions that are not trivial to demonstrate or refute, and therefore pose real challenges to avoidance of artifact results (Carlborg et al. 2005). These limitations are appreciated as researchers embrace systems biology approaches, sampling and integrating multiple types of biological data and incorporating innovative statistical approaches that have more power to discern patterns of causal relationship to phenotypes of interest (reviewed in Rockman 2008).

Network analysis-based methods are an emerging trend in the field, as opposed to classical clustering methods, and are used to discern biologically meaningful covariance patterns within the mass of gene expression variation with the intention of uncovering hidden signatures of interest (Rockman 2008). Different reverse-engineering computational algorithms have been developed to construct networks of co-expressed genes (reviewed in Bansal et al. 2007, Alten et al. 2008). Constructing gene expression networks is straightforward and requires the computation of all pairwise correlations from normalized expression data and a procedure to identify the edges of the network. Methods addressing the latter step are the subject of recent and ongoing research and differentiate among the algorithms available or being developed. Detailed description of gene expression network construction is beyond the scope of this introduction and the focus will rather be on the concepts behind using network analysis in understanding the genetic basis of complex traits in humans.
The ultimate goal of constructing networks of co-expressed genes is to infer and test hypotheses about the causative relationship between genes and biological phenotypes. The major challenge is to tease apart causal from reactive and/or independent effects. Causal analysis using Bayesian networks, systems of recursive equations, or influence diagram approaches, for example, have been used to this end (Opgen-Rhein & Strimmer 2007). In biomedical research, the soundest advances in inferring causality from networks of co-expressed genes have been made on the biological side of investigations by integrating genetical genomics concepts (Schadt et al. 2005, Alten et al. 2008) and perturbation experiments in the framework of network analysis. To date, the most comprehensive study combining the two aspects was performed by Chen et al. (2008). The authors characterized gene expression networks amenable to perturbation by previously identified susceptibility QTLs in a segregating mouse population. The authors reported the identification of expression modules of causal effect. These modules contain genes that have a significant effect on metabolic syndrome traits once they are knocked-out. A companion study by the same group integrated co-expression gene network analysis in mice and humans using adipose gene expression data, and identified a possibly causative association of a new core human network module enriched for immune response genes to obesity-related traits (Emilsson et al. 2008). More recently, similar approaches have been developed and are used to help determine the architecture of complex traits in Drosophila (Ayroles et al. 2009, Harbison et al. 2009). These studies promote the idea that deterministic changes in co-expression
networks have the potential to be translated into strong and true genotype-phenotype associations. This line of research will certainly be increasingly used to assist in the interpretation of disease association signals.

Environmental Transcriptomics

The Concept of Gene-Environment Interactions

Gene-by-environment interactions refer to the differential effect of a given genotype exposed to different environmental triggers. These interactions occur when there is variation among genotypes in the rank order or relative magnitude of effects in different environments (Falconer & Mackay 1996). This phenomenon relates to all organisms and virtually to all phenotypes. The range of phenotypes resulting from a given genotype is referred to as a norm of reaction or phenotypic plasticity and constitutes an important feature in agricultural, genetic, evolutionary, as well as biomedical research. Quantitative traits typically exhibit significant gene-by-environment interactions and have been reported in model organisms for a range of organismal, behavioral, and biochemical phenotypes (Mackay 2001). In humans, virtually all diseases exhibit gene-environment interactions to different extents. Asthma, cancer, heart disease, depression, and metabolic phenotypes (Hunter 2005, Holmes et al. 2008) are among diseases and conditions for which strong gene-environment interactions have been reported.
A remarkable case of the effect of a gene-by-environment interaction on a behavioral phenotype came out of the Dunedin Multidisciplinary Health and Development Study where a birth cohort was followed for 26 years. Caspi et al. (2003) used a regression of cognitive depression on the 5-HTTLPR genotype, stressful life events and their interaction, with sex as a covariate. They found a small main effect for stressful life events, no main effect for the 5-HTTLPR genotype, but a significant effect for the interaction between the two, such that individuals carrying at least one HTTLPR genotype experienced increased depression symptoms with increasing stressful life events but not for the individuals not carrying the HTTLPR genotype. This study provided evidence for a behavioral gene-environment interaction in which an individual’s response to an environmental perturbation is moderated by his or her genetic make-up. Similar scenarios but with an effect in different directions might be possible. For example the effect of a disease risk genetic variant might be moderated by an environmental stimulus or the effects of both the genotype and environmental might act on the same direction, aggravating the disease phenotype. Results such as these often fail to replicate in studies of other populations, and it is extremely difficult to gather cohorts of several thousand individuals. Since genes appear to have a stronger effect on transcript abundance than disease phenotypes, it will be revealing to measure genotype-by-environment interactions as the level of gene expression.
Transcriptional Gene-environment interactions

In humans, the role of environmental exposures in disease etiology is not questioned. There is increasing evidence for cases of specific environmental exposures causing disease conditions through deregulation of gene expression from normal state. Below, I outline a few examples of studies reporting a direct or an indirect link between the environment and a disease phenotype through gene expression modulation.

For Example, in follicular non-Hodgkin lymphoma the frequency of translocation of the gene B-cell leukemia/lymphoma 2 (BCL-2) from chromosome 18 to the immunoglobuline heavy chain locus on chromosome 14 was found to be significantly increased by exposure to a variety of pesticides. Overexpression of BCL-2 results from its placement under the control of the heavy chain enhancer that promotes lymphomagenesis (Roulland 2004).

Other environmental factors modulate gene expression by altering the epigenome state through processes such as methylation, histone modification, and chromatin remodeling resulting in higher-level phenotypic changes. Stages of methylation/demethylation can occur from gametogenesis, through fertilization, embryonic development, and aging (Reik & Walter 2001, Surani 2001, Weaver et al. 2004, reviewed in Richardson 2003). For example, in a remarkable study Weaver et al. (2004) showed that rats that experienced high quality maternal grooming exhibit
reduction in the methylation of the glucocorticoid receptor (GR) promoter region in the hippocampus, resulting in increased expression of GR and subsequently an improved response to stress later in life.

Alteration of the methylation state of certain genes can also occur through exposure to different types of diet. The Agouti viable yellow epiallele case in mice is a well documented example where a larger proportion of mice pups exhibit yellow, obese, and carcinogenic phenotypes as a result of the ubiquitous expression of the unmethylated AIP allele when pregnant mothers are fed on a standard diet as opposed to a diet supplemented with methyl donors (Wolff et al, 1998, Morgan et al. 1999). Aberrant methylation profiles and widespread chromatin modification resulting in specific alterations of gene expression profiles have also been reported in a variety of human diseases including monogenic diseases (i.e. Beckwith-Wiedemann syndrome, DeBaun et al. 2002) and cancers (reviewed in Feinberg & Tycko 2004).

Respiratory diseases such as asthma, bronchitis, and obstructive pulmonary disease are examples of conditions characterized by specific expression profiles of certain genes induced by direct exposure to air pollution through free radical/oxidative stress mechanisms. For example, in vitro exposure of human airway epithelial cells to aerosol liquids produced by industrial activities was reported to induce transcription of major pro-inflammatory transcription factors such as nuclear factor kappa B (NF-κB) and activator protein 1 (AP-1), which in turn promotes a cascade of events such as histone
acetylation, resulting in expression of inflammatory genes and increase in protein levels of interleukin 8 (IL-8), which is often used a marker of inflammation (Donaldson et al. 2003, Gilmour et al. 2003). Similar but also different cascades of events resulting in distinct expression profiles for specific batteries of genes after direct exposure to substances in the environment have been reported in chemical-specific signaling pathways (i.e. estrogen signaling, Vivacqua et al. 2003), allergic response pathways (i.e. T-cell allergic response, Lee et al. 2004) and other complex disease conditions such as breast cancer (Naderi et al. 2007) and Parkinson disease (Sun et al. 2005). Expression signatures in these studies have been tightly linked to the corresponding higher-level phenotypes and are often used as a diagnosis tool and to predict clinical outcome.

Psychosocial factors also can influence gene expression through their influence on hormones and neurotransmitters. For example, increased expression of the key transcription factors cJUN and cFOS was reported as a response to immobilization stress (Ueyama et al. 1999). Also, up-regulation of corticotrophiin-releasing hormone mRNA and down-regulation of mRNA levels of the 5-HT1 gene in rats is response to the same stress was reported (Givalois et al. 2000, Conrad & McEwen 2000). These molecules are used as biomarkers to measure the magnitude of response to stress.

Gibson (2008) reviewed and outlined the role the environment is playing in shaping gene expression profiles and emphasized on the importance of accounting for and integrating both genotypic and environmental variation to aid for the identification of
the genetic basis of traits in natural populations. The literature on the topic of transcriptional gene-by-environment interactions studied at the genomic level however is slim with the two most comprehensive studies reported in yeast. Landry et al. (2006) investigated whole-genome transcriptional plasticity of six yeast strains in four different conditions. They reported 127, 1,863, and 223 transcriptional traits out of 5258 genes assayed showing significant strain, condition, and strain-condition interaction effect, respectively. They also showed how different functional classes of genes are enriched for each effect. Intriguingly, Smith & Kruglyak (2008) reported effects an order of magnitude stronger. They compared expression profiles of two yeast strains in two different conditions. Out of over 4000 expression traits investigated, 2,996, 3,448, and 2,037 showed significant strain, condition, and strain-condition interaction effect, respectively. Further linkage analysis identified 1,555 eQTLs for 1,382 traits significant for the interaction term. Identified cis-eQTLs have been shown to be more stable across conditions while trans-eQTLs are predominantly condition-dependent. A study investigating olfactory behavior in Drosophila melanogaster reported that 50% of variation in this trait was attributable to gene-environment interactions while only 20 genes are significant for this effect at the transcriptional level (albeit in whole flies) despite using a relaxed statistical significance threshold (Sambandan et al. 2008). This is to say that very contrasting findings on the extent of transcriptional gene-by-environment interactions have been reported probably due to differential transcriptional responses to different environmental stimuli. But more investigations are needed to assess the
reproducibility of these experiments and improve study design and statistical methods used to estimate the extent of gene-by-environment effects. On the other hand, to my knowledge transcriptional gene-by-environment interactions at the genomic level have not been addressed in a natural human population, which is the focus of the research presented here.

The Immune System as a Model to Study Gene-Environment Interactions

The human immune system is composed of multiple inter-related and complementary sub-systems of defense that protect against pathogens and tumor cells and that consists of two major components: innate immunity and adaptive immunity. The former involves mechanisms that are not specific for each pathogen, and includes anatomical barriers, antimicrobial soluble proteins, phagocytic cells, and the inflammatory response. The latter is mediated by cell and humoral processes (B and T lymphocytes) that can recognize and target specific antigens. Lymphocytes display tremendous antigen-binding receptor diversity after their establishment. A range of stochastic processes can produce approximately $10^{14}$ millions different antigen-binding specificities during T-cell development (Janeway et al. 2001). Upon activation by exposure to the antigens presented by a pathogen, a lymphocyte that has a matching receptor will replicate preferentially through the process called clonal selection and amplification.
Given the extent of interplay between the environment and the immune system, the latter is ideal as a model to study gene-environment interactions in humans. The immune system develops and functions while incorporating information from the surrounding environment. The components of the immune system are responsive to a range of external factors include pathogens, stress, nutritional, social, and other environmental factors that together ensure that the development an individual’s immune system is context-dependent (McDade 2005).

A readily accessible component of the immune system is the total white blood cell population or leukocytes that can be isolated from peripheral blood. The latter is a complex body fluid composed of mixed cell populations of which leukocytes constitute a small fraction, yet they are a major component of the immune system. There are several types of cells among leukocytes, each of which can be further subdivided based on cellular and molecular characteristics. Variation in the relative proportions of different cell types is a response to numerous environmental and physiological stimuli and specific gene expression signatures have been traced to different cell types (Whitney et al. 2003, Palmer et al. 2006). Inferences from leukocytes gene expression profiles can still be made although there might be a confounding effects with differential cell proportions. The ideal study design would include cell counts/quantification as a covariate. This and other sources of individual expression heterogeneity pose a real challenge and should be taken into consideration when possible. Some of these sources of expression
heterogeneity can be controlled for (for example, age and sex). For unknown sources of expression heterogeneity, a statistical method, namely Surrogate Variable Analysis, was developed to correct for their effect on statistical inferences generated from gene expression data analysis (Leek & Storey 2007).

Outline of the Thesis

For my doctoral research presented here, the first chapter reports a pilot study where I collected peripheral blood samples from 46 Moroccan Amazigh (Berber) individuals representing three distinct lifestyles and geographic regions in Morocco (desert nomads, mountain villagers, and coastal urban dwellers), and conducted whole-genome gene expression, whole-genome genotyping, and a high throughput methylation assay. The aim was to quantify the effect environmental factors relating to lifestyle and geographic differences have on immune expression profiles of the three genetically relatively homogeneous populations. The second and third chapters report a fellow-up study where I sampled over 200 individuals in Morocco representing two ethnicities (Arab and Amazigh) and two distinct lifestyles (urban and rural) with the aims being to (i) attempt to replicate results of the pilot study, (ii) jointly use gene expression and genotypic data to partition the genotypic, environmental, and gene-by-environmental effects on transcriptional variation, (iii) explore the characteristics of genes significant for each effect, and finally (iv) unravel eSNP associations and test their robustness to environmental perturbations.
References


CHAPTER TWO

A Genome-wide Gene Expression Signature of Environmental Geography in

Leukocytes of Moroccan Amazighs
A Genome-wide Gene Expression Signature of Environmental Geography in Leukocytes of Moroccan Amazighs

Youssef Idaghdour, ¹ John D. Storey, ² Sami J. Jadallah, ³ and Greg Gibson ¹

¹ North Carolina State University, Raleigh, NC, USA
² University of Washington, Seattle, WA, USA
³ HRH Prince Sultan International Foundation for Conservation and Development of Wildlife, Agadir, Morocco

This chapter is a reprint of previously published open-access article distributed under the terms of the Creative Commons Attribution License. The article is included with the permission of the authors.

Reference:
A Genome-Wide Gene Expression Signature of Environmental Geography in Leukocytes of Moroccan Amazighs

Youssef Idaghdir1, John D. Storey2, Sami J. Jadallah3, Greg Gibson1,4*

1North Carolina State University, Raleigh, North Carolina, United States of America; 2Princeton University, Princeton, New Jersey, United States of America; 3HHR Prince Sultan International Foundation for Conservation and Development of Wildlife, Agadir, Morocco; 4University of Queensland, Brisbane, Queensland, Australia

Abstract

The different environments that humans experience are likely to impact physiology and disease susceptibility. In order to estimate the magnitude of the impact of environment on transcript abundance, we examined gene expression in peripheral blood leukocyte samples from 46 desert nomadic, mountain agrarian and coastal urban Moroccan Amazigh individuals. Despite great expression heterogeneity in humans, as much as one third of the leukocyte transcriptome was found to be associated with differences among regions. Genome-wide polymorphism analysis indicates that genetic differentiation in the total sample is limited and is unlikely to explain the expression divergence. Methylation profiling of 1,505 CpG sites suggests limited contribution of methylation to the observed differences in gene expression. Genetic network analysis further implies that specific aspects of immune function are strongly affected by regional factors and may influence susceptibility to respiratory and inflammatory disease. Our results show a strong genome-wide gene expression signature of regional population differences that presumably include lifestyle, geography, and biotic factors, implying that these can play at least as great a role as genetic divergence in modulating gene expression variation in humans.

Introduction

Understanding the contribution of genetic and environmental factors to variation in gene expression in humans is essential to interpreting the relationship between genotype and phenotype. Genetic differentiation has been the focus of several recent studies that have extensively mapped gene expression variation to specific genetic variants in lymphocyte samples [1] and Epstein-Barr virus-transformed lymphoblastoid cell lines [2-3]. The contribution of environmental factors to variation in gene expression in humans has not yet been explicitly investigated. Here we test the hypothesis that environmental factors can generate significant transcriptional variation by contrasting peripheral blood gene expression among three regionally distinct samples of Moroccan Amazighs, who are a genetically relatively homogeneous human population. The Amazighs, also known as Berber people, occupy northwest Africa and are thought to represent a genetically relatively homogenous human population [4-5]. They lead distinct ways of life and occupy diverse physical geographic habitats across Morocco thus providing an excellent opportunity to monitor the impact regional differences in living circumstances have on gene expression and therefore physiology.

Peripheral blood is a readily accessible tissue sample that integrates environmental factors such as immune exposure, diet, and psychological state. We collected peripheral blood samples and isolated total leukocytes for gene expression profiling. We set out to establish the extent of the effect environmental factors relating to lifestyle and geographic differences have on immune expression profiles. Whole-genome genotyping and methylation profiling were used to estimate the extent of population structure and methylation differentiation in our sample as a proxy for their effect on the observed expression differentiation.

Results/Discussion

Over a three week period, we obtained leukocyte samples from peripheral blood for gene expression profiling from 16 Bedouin living a traditional nomadic existence on the fringe of the Sahara desert near the town of Errachidia, 18 inhabitants of Aous, an urban slum-like settlement within the coastal city of Agadir, and 12 villages from Sidi-Nabhir, a remote rural mountain settlement south of Agadir (Figure 1). We isolated the total leukocyte population immediately after blood sampling [6], and extracted total RNA. Expression profiles were monitored with Illumina HumanRef-8 v2 BeadChip oligonucleotide arrays representing over 22,000 annotated genes [7], 10,177 of which were expressed in the samples.

Effect of Lifestyle and Geography on Gene Expression

We detected several distinct global profiles of expression, implying expression heterogeneity among individuals. This is seen in the analysis of all expressed genes, but is readily visualized in a heat map of two-way hierarchical clustering of the 1,000 most
Author Summary

The incidence of complex diseases such as diabetes, asthma, and depression is almost epidemic in many countries and coincides with transition in lifestyles. Clearly this is a result of interaction between modern cultural and environmental factors with the genetic legacy of human history. To estimate the extent of the effects of environmental factors, including lifestyle and geography, on gene expression, we examined gene expression differentiation in peripheral blood leukocyte samples from three Moroccan Amazigh groups leading distinct ways of life: desert nomadic, mountain agrarian, and coastal urban. Our data shows that as much as one third of the leukocyte transcriptome is associated with differences among the three regions. Network analysis implies that specific aspects of immune function are strongly affected by regional factors and may influence disease susceptibility. Genetic and methylation differentiation between the three regions is limited and is unlikely to explain the extent of the observed gene expression differentiation. Insight gained from this study highlights the impact transitions from traditional to modern lifestyles likely have on human disease susceptibility and further warrants the need to incorporate gene expression profiling alongside genetic association studies for the prediction of disease susceptibility.

significant differentially expressed genes (Figure S1), highlighting the dominant impact that locality has on gene expression profiles. Surrogate variable analysis (8) was employed to capture and control for additional sources of expression heterogeneity that may include circadian cycle and age. This has been shown to increase the accuracy of analyses and avoid confounding of signals due to hidden sources of heterogeneity (9) (Figure S2 and Table S1). An analysis of variance including terms for region, sex, batch, and six significant surrogate variables confirmed the significant effect regional factors have on gene expression. A total of 3,725 of 18,177 (20%) of the expressed transcripts were differentially expressed with respect to region at the 1% FDR cutoff in a full three-way comparison, rising to 6,215 transcripts (61%) at 5% FDR. Table 1 shows the number of significant genes for each pairwise comparison. In the same analysis 57 transcripts were differentially expressed with respect to sex, at 1% FDR (70% at 5% FDR).

Given the geographical distance between the sampled localities (urban to rural: 150 km; urban to nomadic: 560 km; rural to nomadic: 650 km), and the scope of the differences in lifestyle, we expected the nomadic sample to be the most differentiated. However, volcano plot of significance against the magnitude of gene expression divergence for each pairwise comparison of regions (Figure 2A) imply that the nomads diverge less from both the rural and urban individuals, than the urban diverge from rural individuals. We also examined the differentiation of all three regions. (Figure 2B), which shows that the majority of the differences between nomadic and rural individuals are also observed in contrasts of these two localities with the urban sample.

Comparison of Environmental and Genetic Effects on Gene Expression

It is usually difficult to contrast the genetic and environmental contributions to leukocyte expression variation simultaneously, since different ethnic groups occupy different environments, but comparison with published results from laboratory cultures of cell lines implies that the regional effect is relatively large. Transcript abundance has recently been measured in B lymphoblastoid cell lines derived from individuals of Asian, African and European ancestry [9-11]. The proportion of genes differentially expressed was for example jointly estimated within and among populations.

Figure 1. Geographic locations of sampled Amazigh groups in Morocco. A total of 52 peripheral blood samples were collected: 20 urban samples from the town of Anza (Latitude: 39.367°, Longitude: -9.635°), 15 rural samples from the rural village of Sbeit-Nabor (Latitude: 31.459°, Longitude: -8.050°), and 17 nomadic samples from the Sahara desert in eastern Morocco (Latitude: 31.809°, Longitude: -4.037°). Subsets of these were used in the gene expression profiling, genotyping and methylation profiling as described in Table S5.

doi:10.1371/journal.pgen.1000032.g001
Table 1. Number of differentially expressed genes.

<table>
<thead>
<tr>
<th>Significance Threshold</th>
<th>Urban vs. Rural</th>
<th>Urban vs. Nomadic</th>
<th>Nomadic vs. Rural</th>
<th>Male vs. Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonferroni P&lt;0.01</td>
<td>406 (40%)</td>
<td>103 (1.9%)</td>
<td>57 (0.5%)</td>
<td>22 (0.22%)</td>
</tr>
<tr>
<td>FDR&lt;0.01</td>
<td>2,770 (27.2%)</td>
<td>1,069 (10.5%)</td>
<td>704 (7.4%)</td>
<td>30 (2.9%)</td>
</tr>
<tr>
<td>P&lt;0.01</td>
<td>2,044 (20.9%)</td>
<td>1,897 (18.6%)</td>
<td>1,674 (16.4%)</td>
<td>365 (3.6%)</td>
</tr>
</tbody>
</table>

Numbers and percentages of differentially expressed genes at three different significance levels were obtained after fitting an analysis of variance model that included surrogate variables to account for unmodeled sources of differential expression, as well as fixed terms for location, sex, and batch.
doi:10.1371/journal.pgen.1000052.t001

[11], where it was concluded that approximately 17% of genes differentiate African and European populations. This percentage is based on the estimate that 63% of the genes are true nulls (that is, false positives if all genes are assumed to be significantly different) in the comparison. Using more stringent criteria, just 50 of 5,194 expressed genes were found to separate the two CEPH samples at an FDR of 20%, compared with 2,770 of 10,177 genes at an FDR of 1% in our similarly sized sample of leukocytes. In a reanalysis of the Asian and European comparison data set from ref. [2], it was shown that at least 94% of genes are differentially expressed with respect to a temporal batch effect and ancestry [13]. The reanalysis also estimated that 76% of genes are differentially expressed with respect to the batch effect among individuals of European ancestry. Therefore, of the 94% of genes differentially

![Volcano plots](image-url)

Figure 2. Volcano plots of statistical significance versus magnitude of differential expression between locations. (A) For each transcript, significance is shown as the negative logarithm of the P value on the y-axis, and the log base 2 of magnitude of mean expression difference is on the x-axis. Dashed lines indicate the threshold for significance (green: P<0.05; blue: 1% FDR, and red: Bonferroni adjusted P<0.05). The Venn diagram (B) shows the numbers of differentially expressed genes at 1% FDR for each comparison and the overlaps between them.
doi:10.1371/journal.pgen.1000052.g002
expressed with respect to year and ancestry, it can be estimated that 79% of these are due to the batch effect alone. This yields an estimate of 94%–95% is 15% of genes being differentially expressed with respect to ancestry, which is similar to the finding in ref. [11], and approximately half the level of differentiation we observe due to regional non-genetic factors.

The effect of regional factors in our study is by contrast similar to that of genetic influence on breast tumors [14] and of HIV infection on peripheral blood [15]. We have reanalyzed the Affymetrix dataset in ref. [15] contrasting PBMC expression in HIV-positive and negative mothers in a rural village in Botswana. HIV infection was significantly associated with expression of 10,000 most strongly transcribed transcripts at an FDR of 1% in this study of a similar size. Consequently, expression heterogeneity on a scale similar or greater to that observed among Moroccan lifestyles can arise within a single community due to an identifiable environmental agent. Furthermore, reanalysis of a breast cancer cDNA microarray dataset [16] shows that 33% of transcripts distinguish BRCA1 from BRCA2 mutant tumors [14]. Thus, when taken at face value, regional influences including lifestyle and geographic effects appear to make as large a contribution to leukocyte expression as common immune-related diseases.

It should be emphasized that though power to detect differential expression varies among studies, being heavily influenced by such technical factors as the array platform, sample size, and RNA handling methods, not to mention biological factors such as tissue homogeneity, genetic diversity, and individual environmental or cultural differences. Nevertheless, the studies contrasting lymphoblastoid lines from different populations should have been well powered to detect differences of the magnitude we attribute to lifestyle differences, and we also note that the estimates of the fraction of differentially expressed genes among lifestyles was not statistically affected by partial correlation of our dataset. Interestingly, an estimate of heritability of transcription in PBMC samples from Mexican American families [11] was congruent with an estimate of heritability of transcription in transformed lymphoblasts from a British cohort cohort [3], both recent studies finding that over one quarter of transcripts have heritabilities in excess of 0.3. Heritability provides an estimate of the genetic contribution to the expression trait, but is by definition a characteristic of a single population in a common environment. Taken together these and other studies thus establish that there is ample genetic variation within populations, and that while there is some divergence between human ethnic groups, the potential impact of environmental variables can be much greater than that of genetic divergence.

**Genetic Differentiation Is Limited in our Sample**

Although Amazighs are homogeneous relative to other human groups [4], having retained a distinct cultural identity in Morocco alongside Arabs and sub-Saharan Africans in the face of repeated invasions throughout recorded history, it is possible that there has been modest genetic differentiation among the three communities [5]. Whole genome genotype profiles [17] were obtained for eight randomly selected individuals from each region and principal component analysis was used to infer the extent of genetic variation from over 300,000 autosomal SNPs [18]. Out of the top 10 axes of variation, only one is significant ($P=0.0006$, Fray-Widom test). It distinguishes three residents of urban Arabic, possibly indicating recent Arab admixture (Figure S3 and Table S2). No significant axes of variation were detected when these three individuals were removed from the analysis. There was a suggestion of genetic divergence of the moznad in the second axis but it only explained a small fraction of the genotypic variation and was only marginally significant ($P=0.0167$). By contrast, application of principal component analysis to a European American sample [10] indicated that the top two axes were statistically significant ($P<0.01$ for each with the top two axes having $P<10^{-12}$, supporting the evidence that Moroccan Amazighs are a homogeneous group relative to other human groups.

The absence of meaningful population structure was also confirmed with *STRUCTURE* Version 2.2 [19]. We applied the program to 11,000 autosomal SNPs (500 randomly selected and approximately uniformly distributed from each of the 22 autosomal chromosomes) at $K=2–3$ (Figure 5A). No $K$ value separates those individuals (BI, 5D, and 5G in Figure S1) from the rest of samples; the same individuals were distinguished by *Eigensplit* analysis on the significant axis of variation. At $K=3$, two nomadic individuals cluster as a distinct unit but the rest of samples have high membership coefficient to one group. Increasing the number of SNPs had no effect on the results. As $K$ is increased, pairs of individuals show evidence for relatedness, but these are not members of the same population, further confirming the relative absence of population structure in the dataset. Comparison of pairwise differences among individuals confirms that, excluding the three outliers in Azilal, there is no notable difference in degree of relatedness of individuals within the different populations. Taken together these results indicate that the majority of the divergence in gene expression described here is unlikely to be explained by genetic divergence.

Despite the absence of genomewide differentiation between the study sites, it is possible that some of the expression divergence could be due to genetic differences at a small number of loci that regulate hundreds of downstream transcripts. These would likely need to be nearly fixed between the population to cause the almost complete correspondence between expression profile and locality. To explore this possibility, we estimated $F_{st}$ for each of over 300,000 sites for each of the three pairwise comparisons of populations, and plotted the values in a sliding window of 100 sites along each chromosome. $F_{st}$ values range from 0.015 to 0.05 but occasionally range up to just 0.11, averaging between 0.013 and 0.052 in each comparison, again confirming the low level of genetic divergence. No fixed differences were detected.

To explore the relationship between genetic and expression divergence at each locus, gene-specific $F_{st}$ measures were calculated by averaging $F_{st}$ values of all SNPs located within 1-Mb upstream and downstream of the expression probe. No correlation between $F_{st}$ and expression fold change (Figure S5A) or significance (Figure S5B) was detected when this analysis was performed for all 10,177 expressed genes and for all three pairwise comparisons. Although some cis-acting, presumably regulatory, variants have been shown to have large effects on transcript abundance [1–3], this result argues against the population expression differences being attributable to genetic divergence at thousands of locally acting polymorphisms in our study. While a few genes show both genetic and expression divergence, their number is no greater than expected by coincidence.

**Differential Methylation Analysis**

The observed differential expression could be due to differences in the ratios of the numerous cell types present in the total leukocyte population, to transient changes in activity of transcription factors and micro RNAs, or to longer-term epigenetic modification of chromatin structure. To test for one possible epigenetic mechanism, we measured the degree of methylation at 1,505 CpG sites [20] representing 895 genes of various classes, including tumor suppressor genes, oncogenes, genes involved in
DNA repair, cell cycle control, differentiation and apoptosis. 420 of these sites were within genes included in our list of transcripts differentially expressed among locations. 50 differences in CpG site methylation between males and females were detected at a P<0.001 with a mixed model analysis of variance (Figure 3A), the majority involving X-linked genes, as expected given the correspondence of methylation with X-inactivation [21] (Figure S6), thus providing a positive control for the methodology. However, only 18 CpG sites were found to differ between regions in this analysis at the same significance level, which is no more
than expected by chance. A small signature of a dozen or so genes differentiated half of the urban population (Figure 3B), but on the basis of our results epigenetic modification via methylation can only account for a small fraction of the expression divergence between regions.

Expression Differences and Impact on Immune Function

Global comparison of differential expression according to gene ontology classes and pathway analysis implicates divergence in core immune competence among the three groups. Ingenuity Pathway Analysis was used to explore connectivity among differentially expressed transcripts and identified significant differential expression networks anchored by key immunoregulatory factors. For example, all of the genes connected to the cell cycle and apoptosis regulators FOS and MYC [22] show significant effects of regional factors (Figure 4A and Table S3). The identities of the most differentially expressed genes also suggest that the expression divergence among the three Amazigh localities is likely to impact immune system function and disease susceptibility (Fig 4B and Table S4). The following specific examples illustrate this conclusion, which is further corroborated by network analysis of clusters of differentially expressed genes.

The most enriched Ingenuity Knowledge Database disease category in the differentially expressed gene list is the respiratory disease class (P<0.0001, Fisher’s exact test). Our data shows that IL-8, a cytokine involved in the immune response to infection as well as autoimmunity, is upregulated in the inhabitants of urban Azaa compared to the rural villagers (P=0.0005, ANOVA) and the nomads (P<0.0004). Additionally, four HLA-A, HLA-B, HLA-C, and HLA-DQ alleles associated with respiratory disease were differentially expressed by location (Figure S7). The Azaa residents are exposed to pollutants from a nearby concrete manufacturing plant and other industrial units, and to viral pathogens that are not experienced by nomads or villagers, and are known to have an enhanced incidence of respiratory illness. This result suggests that a genotypic-by-environment effect resulting in aberrant expression of IL-8 transcript abundance in leukocytes, should be considered in parallel with the use of genotypic variation as a possible diagnostic marker for respiratory illness.

The Human Leukocyte Antigen (HLA) complex is well known to mediate response to infection as well as autoimmunity, as it encodes the major histocompatibility molecules that present antigens to the immune system. Intriguingly, two non-classical members of the HLA complex, HLA-DQ and HLA-DQ, were differentially expressed among the urban population compared to the rural villagers (P<0.000009 and P<0.0002, respectively) and to the nomads (P<0.0009 and P<0.0006, respectively). These thought-to-be-invariant genes have been associated with susceptibility to auto-immune diseases [28], and are strongly co-regulated across our sample. This result warrants further exploration of association of these genes with immunologically disease.

One of the two most divergent transcripts encodes POLR2F (also known as OCT2), the octamer-binding transcription factor that regulates differentiation of leukocytes [27]. Transcription of this gene tends to be elevated in villagers relative to the other two groups. ELK1, another important lymphokinin transcription factor [29] that regulates FOS transcription [29], showed elevated expression only in individuals from the village. Potential and validated targets of ELK1 are enriched (P<0.006, Fisher’s exact test) in the list of differentially expressed genes as determined by counts of putative binding sites listed in the TRANSFAC database [30]. Consequently, modulation of expression of this single regulator probably has pleiotropic effects on immune function, both at the level of cellular differentiation and function.

Conclusion

The most plausible explanation for the dramatic differentiation of as much as a third of the transcriptome described here is that a combination of biotic, abiotic, and cultural differences is involved. A significant portion of these effects could be attributed to aspects related to lifestyle such as nutrition, history of immune exposure, and psychological stress. Consequently, insight gained from this study highlights the impact transitions from traditional to modern lifestyles likely have on human disease susceptibility, particularly through their impact on immune function. We speculate that diseases due to genetic factors in urban populations may bear little resemblance to the impact of the same genetic factors in traditional societies. Since the causal environmental factors and the mechanisms through which they remain to be identified, we advocate the incorporation of gene expression profiling alongside genetic association studies for the prediction of disease susceptibility.

Materials and Methods

Study Populations and Collection Protocol

Ethical approval for this study was granted by the Institutional Review Board for the Use of Human Subjects in Research at North Carolina State University, and the Moroccan Ministry of Health. Peripheral blood samples were collected under informed consent from 52 self-reported Moroccan Amazigh. All subjects were between the ages of 18 and 52 and were reported to be in good health at the time of blood sampling. The pastoral nomadic subjects of the study (average age = 35) are Bedouins from an Amazigh tribe inhabiting a remote area near Zaouiet Tafira (Latitude: 31.89°, Longitude: −4.68°), in the Sahara desert. The Bedouins live in traditional tents and their subsistence comes from domesticated animals. All but three subjects are unrelated (in Figure S1, individuals labeled 5A and 5E are siblings and 5A is their cousin). The rural subjects (average age = 32) are unrelated permanent residents of the Seh-Nabir village (Latitude: 31.45°, Longitude: −9.65°). The subjects have a traditional lifestyle based on traditional agriculture and herding with very little exposure to urbanized lifestyle, one indication of which is that access to the village was by a four hour donkey ride. The urban subjects of the study (average age = 31) are unrelated permanent residents of the urban town of Azaa (Latitude: 29.36°, Longitude: −9.63°). All the subjects have been living in Azaa for at least the last 10 years.

The same sample collection protocol was followed for the three collection sites in order to minimize the effect of this source on gene expression heterogeneity. Blood samples were collected within a period of three days for each locality in December 2006, with collection each day spread over a 6 hour period from mid-morning to mid-afternoon. Approximately 15 ml of peripheral whole blood was collected by a phlebotomist using vacutainer. The total leukocyte population was immediately within 2 minutes isolated from 9 ml, and its total RNA was stabilized with the LeukolOCR™ Total RNA Isolation System (Ambion, Austin, TX). The system incorporates depletion filter technology to isolate leukocytes and eliminate plasma, platelets, and red blood cells, and RNAStaer to stabilize the RNA in the cells captured in the filter. The remaining blood (approximately 6 ml) was stored in EDTA tubes for DNA extraction. The filters and blood samples were kept on ice, and then frozen at −80°C within three days of collection. While we cannot exclude the possibility that sampling differences...
Figure 4. Functional analysis of differentially expressed genes. (A) Differential expression of the FOS and MYC networks and enriched disease dataset. The Ingenuity Pathways Knowledge base (IPKB) was used to generate networks of interacting genes that are overrepresented in the set of transcripts differentially expressed (based on a 1% FDR cutoff) between the urban and rural samples. The top two networks are focused on the FOS and MYC transcription factors, and every one of the genes that the IPKB indicates is interacting either genetically or biochemically are differentially expressed in this comparison. Network connectivity is indicated as solid edges for direct interactions, and dashed edges for indirect interactions. Transcripts are displayed in green for down regulated and red for up regulated, while cellular compartments in which the gene products are localized are also indicated. Gold edges highlight shared interactions. The list of genes, their fold change and p-values are listed in Table S3. (B) Overrepresentation of disease classes affected by differentially expressed genes. Some of the Ingenuity Knowledge Database disease bio-function categories enriched (p<0.05) in differentially expressed transcripts (1% FDR) in the three lifestyle pairwise comparisons (gray, urban vs. rural; blue, nomadic vs. urban; green, nomadic vs. rural). Fisher's exact test was used to calculate the P-value associated with the probability that the number of genes in each biological function and/or disease assigned to that data set is greater or less than expected by chance given the numbers of genes expressed in leukocytes.

doi:10.1371/journal.pgen.1000052.g004
at each locality contribute to expression divergence, the fact that several individuals cluster outside their locality (Figure S1) argues against this.

RNA and DNA Preparation

Total RNA extraction, cDNA and cRNA synthesis were performed with the Illumina TotalPrep RNA Amplification Kit (Ambion, Austin, TX) following the manufacturer’s instructions. Size distribution of the extracted total RNA and the amplified cRNA was checked with Agilent’s RNA 6000 Nano LabChip kit and 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). The quality of RNA and cRNA was comparable among all samples used further in the experiment. DNA samples were extracted with QIAamp DNA Kit (Qagen, Valencia, CA). Standard gel electrophoresis and the ND-1000 (NanoDrop Technologies, Wilmington, DE) were used to check DNA quality and quantity, respectively. The list of samples used for each experiment described below is available in Table S5.

Gene Expression Profiling

Illumina’s HumanRef-8 v2 BeadChips (Illumina, San Diego, CA) were used to generate expression profiles of more than 22,000 transcripts with 500 ng of labeled cRNA for each sample and following manufacturer’s recommended protocols. All expression data are available at NCBI Gene Expression Omnibus (GEO) under series number GSE89347. The individual expression arrays are listed as GSM219088 through GSM220033. An Excel spreadsheet list of all differentially expressed gene is also available online at the PLoS Genetics website as Dataset S1, and significance criteria for all transcripts are provided as Dataset S2.

A randomized design was used to minimize chip effects. Four individuals were replicated in the two batches; these clusters adjusted for one another in hierarchical analysis and the expression intensities were averaged in the statistical analysis. Expression intensity measures were obtained from an average of 30 beads for each transcript. The BeadChips were imaged with an Illumina BeadArray Reader. The raw intensities were extracted with the Gene Expression Module in Illumina’s BeadStudio software. Expression intensities were log2 transformed and median-centered by subtracting the mean value of each array from each intensity value. 10,177 transcripts with expression at or above background levels averaged across all the arrays were retained for further analyses. These represent transcripts remaining after removal of 12,000 bead measurements that were considered to lay below background detection levels because they are less than the 10,177 most expressed transcripts, and the resulting relative fluorescence intensities were used in further analyses. List of transcripts considered expressed in leukocytes and lists of significance for differential expression for each comparison are available in Tables S6 and S7, respectively.

Statistical Modeling of Gene Expression

It has recently been shown that if there are additional sources of expression variation due to factors not included in the model, then this can lead to unreliable differential expression analyses due to large-scale dependence among genes and potential confounding with these unmodeled factors [8]. “Surrogate Variable Analysis” was developed to directly use the known variables (here, Location, Sex, and Batch) and the entire expression data set in order to estimate the signatures of these unmodeled factors, called “surrogate variables”. Table S1 lists the surrogate variables estimated and utilized in all of the analyses. After identifying surrogate variables, differential expression was estimated using an analysis of variance following standard methods [31] on the basis of the following model:

Expression = Baseline + Location + Sex + Location*Sex + Batch + Surrogate Variables + Noise.

For each of the Location and Sex variables, a P value (ANOVA) measuring significance of differential was obtained for each gene. False discovery rates were calculated according to the q-value software package [14]. It should be noted that even though the surrogate variables represent random effects, we were able to effectively treat them as fixed in the model fitting process, since all inference was performed conditional on the surrogate variables in a conditional likelihood framework [32]. The effect on P value distributions when including surrogate variables is shown in Figure S2.

Functional and Promoter Enrichment Analysis

The network, functional and biomarker analyses were generated through the use of Ingenuity Pathways Analysis. Genes whose expression was significantly differentially regulated between the three locations were included using a 1% false discovery rate cutoff from the surrogate variable analysis results. Fisher’s exact test was used to calculate a P value associated with the probability that the number of genes in each biological function and/or disease assigned to that data set is greater or less than expected by chance given the numbers of genes expressed in the leukocytes. The program PRIMAA [33] integrated in the software EpiInert [34] was used to find transcription factors whose binding sites are more frequent than expected by random in the promoters (spanning from 1000 bp upstream the TSS to 200 bp downstream the TSS) of the differentially expressed genes between locations. We used version 27 of the list of binding sites in the TRANSFAC database [30].

Genotyping and Population Structure

Twenty-four samples (eight randomly selected samples from each population) were assayed with Illumina’s Illumina Human-Hap550 SNIP Chip following standard procedures. The Human- Hap550 SNIP Chip contains over 318,000 SNPs derived from phase I of the International HapMap project [35]. The BeadChips were imaged using Illumina’s BeadArray Reader and genotype calls extracted with the Genotyping Module in Illumina’s BeadStudio software. Principal component analysis was used to infer the extent of genetic variation from over 300,000 autosomal SNPs using EIGENSTRAT as described in ref. [18]. Structure Version 2.2 [19] was used to infer population structure. We applied the program to 11,000 autosomal SNPs (500 randomly selected and approximately uniformly distributed from each of the 22 chromosomes) at K=2-3. We used a model with admixture and correlated allele frequency for 100,000 iterations after a burn-in length of 20,000. We used small K to analyze population structure given our prior knowledge about the Moroccan population. Structure runs at K=2-3 were repeated under the no admixture model with either correlated or uncorrelated allele frequencies with similar results.

Methylation Assay

A methylation profile was obtained for 96 samples (14 nomadic samples, 15 rural samples, and 39 urban samples, each sample
represented by two technical replicates) with Illumina's GoldenGate methylation Cancer Panel I array-based assay. DNA samples were first subjected to a bisulfite conversion reaction using an Illumina GoldenGate methylation ISH assay (Illumina, San Diego, CA) and then to Illumina's GoldenGate methylation assay [29]. The GoldenGate Methylation Cancer Panel I spans 1,505 CpG loci selected from 897 genes falling into various classes, including tumor suppressor genes, oncogenes, genes involved in DNA repair, cell cycle control, differentiation, and apoptosis. The raw methylation ratios were extracted using the Methylation Module in Illumina's BeadStudio after a background normalization that subtracts a background value derived by averaging the signals of built-in negative control bead types. A mixed model analysis of variance was applied on a GpG site-by-GpG site basis with the PROC MIXED procedure implemented in SAS ver. 9.1 (SAS Institute, Cary, NC).

\[ Y = \text{Location}_i + \text{Sex}_j + \text{Location} \times \text{Sex}_k + \epsilon \]

Where \( X \) and \( Z \) were considered fixed effects with the ith location (\( i = \) urban, rural or nomadic) and the jth sex (\( j = \) male or female). The effect of interaction between Location and Sex was included in this model, and the error \( \epsilon \) was assumed to be normally distributed with mean zero.

**Supporting Information**

**Figure S1** Two-way hierarchical clustering of expression. The heat map shows the clustering of expression profiles largely by location. Each row represents one of the top 1,000 transcripts for significance of the location effect and each column represents one individual. Intensity of red indicates relatively high expression relative to the sample mean, of blue relatively low expression. Individuals are identified by a code with the first letter representing gender (M: male, F: female), the second letter population (D: desert, N: nomadic, A: Anza), rural (R) or urban (U), the number corresponds to Illumina BeadArray, and last letter is a unique identifier within each array. Note that the highest level of clustering tends to be by population, while the two genders also tend to cluster within populations. The clustering was generated with Ward’s method in JMP Genomics ver. 3.0 implemented in JMP ver. 7.0.

**Figure S2** Effect of surrogate variable inclusion on significance testing. Quantile-quantile plots of the P-values resulting from differential expression analyses with and without SVA. The solid does represent P-values quantiles and the dashed line is the line of equality. Curves above the diagonal imply larger P-values without SVA, and hence a gain in power from the analysis that include surrogate variables. (A) Urban vs. Rural vs. Nomadic. (B) Urban vs. Rural. (C) Urban vs. Nomadic. (D) Nomadic vs. Rural.

**Figure S3** Eigengenesis of population structure. The plot shows the first two eigenvectors of genetic variance for the three Moroccan populations (red, urban; blue, rural; brown, nomadic). Only the first eigenvector is significant (P=0.0005, Tracy-Widom test); its eigenvalues within the urban sample as indicated in the figure. The second eigenvector suggests a separation of the nomads and villagers, but it is only marginally significant (P=0.00167) and explains only a very small proportion of the genetic variance. Each square represents one of the 24 individuals who were genotyped.

**Figure S4** Structure analysis of genotypic variation. Each individual is represented by a column that is partitioned into K colored segments representing the proportion of ancestry (Q-value) from each of the K clusters for each individual using 13,000 autosomal SNP markers. Two Structure runs at K = 2 and K = 3 are shown. At K = 3, 80% of individuals have high membership coefficient to one cluster.

**Figure S5** Genomic-wide genetic and gene expression differentiation. Average Fst measures plotted against log10 fold change (A) and negative log10 probability resulting from differential expression analyses with SVA (B), for each of the 10,177 expressed genes and for each individuals pairwise comparisons (Green, Nomadic vs. Rural; Blue, Nomadic vs. Urban; Red, Urban vs. Rural). Each open circle represents a gene. Gene-specific Fst values were calculated for each gene by averaging Fst values for all segregating SNPs within the gene and approximately 1MB upstream and downstream the gene. The percent variance explained is less than 0.001 with p<0.05 for all six regressions represented by these plots.

**Figure S6** Results of methylation analysis. Histograms show the distribution by chromosome of CpG sites for (A) the 1,505 CpG sites represented on Illumina’s GoldenGate Methylation Cancer Panel I array; (B) the 97 differentially methylated CpG sites for the sex effect at P<0.05 (ANOVA); (C) the 69 differentially methylated CpG sites between locations at P<0.05, and (D) the 24 differentially methylated CpG sites for the sex and location interaction effect at P<0.05. The X chromosome is shown in dark green. Panel B can be considered a positive control for the success of the analysis since methylation is known to be preferentially marked X-linked loci.

**Figure S7** Correlations among Interleukin signaling components. The three plots show the average relative transcript abundance (log base 2 scale) for the indicated genes in the Urban (U: Anza), Nomadic (N: Bedoin), and Rural (R: Sbe-Nabib) locations. P values in brackets indicate the significance of the 3-way location term from mixed model ANOVA. Across all 46 individuals, IL-8 is negatively correlated with IRAK2 (P = 0.0008), IRAK1 is negatively correlated with IRAK4 (P = 0.0008), and IRAK3 is positively correlated with IRAK4 (P = 0.0004).

**Table S1** Surrogate variables

**Table S2** Eigenvector principal component statistics

**Table S3** List of genes in FOS and MYC networks shown in Figure 4A

**Table S4** Top 15 networks generated by Ingenuity from the list of genes differentially expressed (1% FDR) for each location pairwise comparison

**Table S5** List of individuals and analyses

---

**Environmental Geography and Gene Expression**

---

**Page 45**
Dataset S1 List of differentially expressed genes and their annotation. Found at: doi:10.1371/journal.pgen.1000520.s013 (1.49 MB XLS)

Dataset S2 List of significance of differential expression by various criteria. Found at: doi:10.1371/journal.pgen.1000520.s014 (0.83 MB XLS)

Acknowledgments
This study would not have been possible without the enthusiastic support of the study participants and assistance of many colleagues in Morroco, particularly Zaid Owair, Abdulatif Khakebose, Louisiane Largo, Hassan Boujidal, Noureddine Menou, Ahmed Belmassou, Abdelghazal Al-Khalidi, and the Idjabouza family. We also extend our thanks to members of Eunilinil Association and Taga Vet Association. We also gratefully acknowledge Ene Stone, Eric Kennedy, Marbta Curley, Josh Alco, Rachel Myers and Jeffrey Lork for discussions, and David Goldenz, Kevin Shatman and Simon Gregory for access to equipment at Duke University.

Author Contributions
Conceived and designed the experiments: YI GG. Performed the experiments: YI. Analyzed the data: YI JS GG. Wrote the paper: YI GG. Provided logistical support: SJ.

References
Figure S1. Two-way hierarchical clustering of expression. The heat map shows the clustering of expression profiles largely by location. Each row represents one of the top 1,000 transcripts for significance of the location effect and each column represents one individual. Intensity of red indicates relatively high expression relative to the sample mean, of blue relatively low expression. Individuals are identified by a code with the first letter representing gender (Male or Female), the second letter population (D, desert / nomadic; A, Anza / urban; V, Sebt-Nabor village / rural), the number corresponds to Illumina BeadArray, and last letter is a unique identifier within each array. Note that the highest level of clustering tends to be by population, while the two genders also tend to cluster within populations. The clustering was generated with Ward’s method in JMP Genomics ver. 3.0 implemented in JMP ver. 7.0.
Figure S2. Effect of surrogate variable inclusion on significance testing. Quantile-quantile plots of the $P$-values resulting from differential expression analyses with and without SVA. The solid dots represent $P$-value quantiles and the dashed line is the line of equality. Curves above the diagonal imply larger $P$-values without SVA, and hence a gain in power from the analysis that include surrogate variables. (A) Urban vs. Rural vs. Nomadic. (B) Urban vs. Rural. (C) Urban vs. Nomadic. (D) Nomadic vs. Rural.
Figure S3. Eigenstrat analysis of population structure. The plot shows the first two eigenvectors of genotypic variance for the three Moroccan populations (red, urban; blue, rural; brown, nomads). Only the first eigenvector is significant ($P < 0.0006$, Tracy-Widom test); it distinguishes three individuals within the urban sample as indicated in the figure. The second eigenvector suggests a separation of the nomads and villagers, but it is only marginally significant ($P < 0.0167$) and explains only a very small proportion of the genotypic variance. Each square represents one of the 24 individuals who were genotyped.
**Figure S4. Structure analysis of genotypic variation.** Each individual is represented by a column that is partitioned into $K$ colored segments representing the proportion of ancestry (Q value) from each of the $K$ clusters for each individual using 11,000 autosomal SNP markers. Two Structure run at $K = 2$ and $K = 3$ are shown. At $K = 3$, 80% of individuals have high membership coefficient to one cluster.
Figure S5. Genomewide genetic and gene expression differentiation. Average $F_{st}$ measures plotted against log2 fold change (A) and negative log10 probability resulting from differential expression analyses with SVA (B), for each of the 10,177 expressed genes and for each population pairwise comparison (Green, Nomadic vs. Rural; Blue, Nomadic vs. Urban; Red, Urban vs. Rural). Each open circle represents a gene. Gene-specific $F_{st}$ values were calculated for each gene by averaging $F_{st}$ values for all segregating SNPs within the gene and approximately 1Mb upstream and downstream the
gene. The percent variance explained is less than 0.001 with \( p > 0.05 \) for all six regressions represented by these plots.

Figure S6. Results of methylation analysis. Histograms show the distribution by chromosome of CpG sites of (A) the 1,505 CpG sites represented on Illumina’s GoldenGate Methylation Cancer Panel I array; (B) the 97 differentially methylated CpG sites for the sex effect at \( P < 0.05 \) (ANOVA); (C) the 69 differentially methylated CpG sites between locations at \( P < 0.05 \), and (D) the 24 differentially methylated CpG sites for the sex and location interaction effect at \( P < 0.05 \). The X chromosome is shown in dark green. Panel B can be considered a positive control for the success of the analysis since methylation is known to preferentially mark X-linked loci.
Figure S7. Correlations among Interleukin signaling components. The three plots show the average relative transcript abundance (log base 2 scale) for the indicated genes in the Urban (U: Anza), Nomadic (N: Bedouin), and Rural (R: Sebt-Nabor) locations. P values in brackets indicate the significance of the 3-way location term from mixed model ANOVA. Across all 46 individuals, IL-8 is negatively correlated with IRAK2 ($P = 0.0008$), IRAK1 is negatively correlated with IRAKBP1 ($P = 0.008$), and IRAK3 is positively correlated with IRAK4 ($P = 0.004$).
CHAPTER THREE

Genetic and Environmental Components of the Human Leukocyte Transcriptome Variation
Abstract

Transcript abundance can be thought of as any other quantitative trait. It is an endophenotype whose architecture is modulated through mechanisms that can incorporate genetic and environmental cues. But what are the relative magnitudes of these effects, and how important are interactions between genotype and environment in shaping immune gene expression profiles in humans? To address these questions, I generated and analyzed gene expression profiles and whole genome genotypic data from 208 and 203 individuals, respectively, from leukocytes of two groups of urban dwellers and two groups of rural villagers representing the two major ethnicities (Arab and Amazigh) in southern Morocco. Population structure analysis suggests the presence of a major genome-wide genotypic axis of variation across both ethnicities. Incorporation of estimates of genetic ethnicity in global quantitative analyses of gene expression profiles revealed strong effects of geography and lifestyle but also suggested that the interplay between non-genetic environment factors and genes is a major modulator of the transcriptome. The data further confirms that genetic factors are neither the sole, nor even the major, source of variation affecting leukocyte gene expression profiles in southern Morocco. The amplitude and functional characteristics of the observed differences suggest a significant impact on immune response and call for similar investigations in other ethnic groups.
Introduction

Genome-wide dissection of the effect of genetic variants on human phenotypes has become a subject of intensive research in the past two years. Despite the identification of loci contributing to normal and disease variation, generally only minor fraction of the genetic variance is explained, and progress has yet to be made in accounting for the majority of the observed phenotypic variance (Maher 2008, Bogardus 2009, Frazer et al. 2009, Ioannidis et al. 2009). Recent calls in this direction are pushing for a resurgence of attempts to quantify the contributions of the environment and gene-by-environment to phenotypic variation (Gibson 2008, Manolio 2008). Gene expression profiling provides an opportunity to address this question since transcript abundance is most immediately connected to genes and has the advantage of incorporating the effects of both genetic variants and environmental factors.

Elucidation of transcriptional gene-environment interactions in humans is particularly relevant to understanding modern human biology. Humans have evolved for tens of thousands of years living traditional lifestyles, but experienced major shifts to urbanized lifestyles in the last few generations. Human populations also live under diverse climatic conditions and in different geographic locations. In order to sample transcriptional variation of relevance to human biology it is therefore imperative to sample populations living under diverse and changing environmental conditions and not
be limited to cell lines or clinical samples as is common current practice (i.e. Stranger et al. 2007, Emilsson et al. 2008).

In this chapter, I report global quantitative analysis of the transcriptome of a sample that exhibits variation across the two major ethnic groups in the Moroccan population (Arabs and Amazighs) living in two urban and two rural locations. I jointly estimate the effect of genetic and non-genetic factors on global transcriptional variation and contrast the null hypotheses that there is no regional or ethnic differentiation in gene expression profiles, with the alternative hypotheses that (i) Gene expression differentiation between the two rural locations is greater than between the urban locations, (ii) Gene expression differentiation between rural and urban locations is greater than between location within each lifestyle, (iii) Environmental geography has a greater impact than ethnicity on gene expression differentiation, and (iv) Interaction effects between gender, ethnicity, and non-genetic factors also have a significant impact on gene expression variation.

Materials and Methods

Study population

The present-day Moroccan population consists predominantly of indigenous Amazighs, and Arabs who conquered the country between the 7th and 11th centuries. The Amazighs belong to a larger group usually referred to as Berbers who are a white Mediterranean group more closely related to Caucasians than to any other major human
ethnic group (Bosch et al. 2001). They have inhabited North Africa as far back as archaeological records exist, and speak languages that form a branch of the Afro-Asiatic family.

Sampling was designed so that four localities representing two major lifestyles were sampled, including both genders, while both Arab and Amazigh ancestries were represented in each locality. Sampling of the two ancestries relied originally on self-reported information. I had no clear expectation as to whether the real spectrum of genotypic variation in the combined sample displays discrete or continuous patterns. Furthermore, I could not rule out the possibility that a fraction of the subjects would be admixed, or have falsely reported their ancestry. Few subjects had dark skin, likely indicating a sub-Saharan component to their ancestry.

A large number of factors could contribute to environmental effects that are shared by individuals, so I took a holistic approach in this study and sampled typical urban and rural localities in a developing country. No attempt was made to make repeated measures from each participant, so individual-specific environmental factors will not be considered. The urban and rural lifestyle samples were separated by approximately 80 miles and are both located in the Souss region of southern Morocco as shown in Figure 1. The first group consists of urban residents who were sampled in two low income districts: Anza (Latitude = 30.447, Longitude -9.655, Altitude: 30 meters) and Dcheira (Latitude = 30.365, Longitude -9.532, Altitude: 12 meters). Anza and Dcheira are seven miles apart and both located within the urban Province of Agadir. All
of these subjects live a typical urban lifestyle characterized by relatively dense human population, frequent traffic, and the presence of industrial activities. The second group consists of villagers sampled from two rural sites (Ighrem and Boutroch) 26 miles apart. Although the subjects in both sites have a traditional lifestyle based on traditional agriculture and herding, the villagers in Boutroch are more isolated and have very limited exposure to urban activities, relative to the villagers in Ighrem. Obtaining samples from males from both villages was challenging and I noted that most of the males make occasional, or in some cases frequent, trips to neighboring cities and seemed to have a lifestyle less traditional compared to females particularly in Ighrem. Boutroch is known to be a predominantly Amazigh village and is in the low Atlas mountains (Latitude = 29.346, Longitude -9.368, Altitude: 1335 meters), whereas Ighrem is located in the foothills of the low Atlas mountains (Latitude = 29.459, Longitude -9.672, Altitude: 720 meters) and is historically Arab with a small fraction of Amazigh residents; self-report confirmed these ethnic differences.

**Collection protocol**

The study was approved by the ethical review committees of the Moroccan Ministry of Health, North Carolina State University, and the University of Queensland. A total of 284 peripheral blood samples were collected in the field under informed consent; 215 and 209 of which were gene expression and genotype profiled, respectively, but several samples were later discarded for quality control purposes as
described below. The subjects were between the ages of 18 and 50 and reported that they were in good health at the time of sampling.

Peripheral blood samples (~8 ml) were collected on six days during the course of the months of June and July 2008. The same collection protocol was followed for all samples in order to minimize heterogeneity due to technical reasons. All samples were collected within four hours between 8:00am and noon. The total leukocyte population was isolated from ~6 ml and within a few minutes its total RNA was stabilized using the Ambion Leukolock® Total RNA Isolation System (Applied Biosystems). The system incorporates depletion filter technology to isolate leukocytes and eliminate plasma, platelets, and red blood cells, and RNAlater® to stabilize the RNA in the cells captured in the filter. The remaining blood was stored in EDTA tubes for DNA extraction. The filters and blood samples were kept on ice, and then frozen at -45ºC within a few hours of collection at all study sites.

**RNA and DNA Preparation**

Total RNA extraction, and cDNA and cRNA synthesis were performed with the Illumina TotalPrep RNA Amplification kit (Ambion, Austin TX) following the manufacturer’s instructions. Total RNA samples were quality checked with the RNA 6000 Nano LabChip® kit and the 2100 Bioanalyzer (Agilent, Santa Carla CA). 215 samples with high RNA quality (Agilent’s RNA Integrity Number RIN > 8) were retained for expression profiling. 209 DNA samples were extracted with QIAamp DNA
Kit (Qiagen, Valencia, CA) and quantified using the ND-1000 (NanoDrop Technologies, Wilmington DE). All DNA samples have 260/280 and 260/230 ratio of optical density within the range 1.70-2.05.

**Gene Expression Profiling**

Illumina’s HumanHT-12 beadchips were used to generate expression profiles of more than 48,000 transcripts with 500ng of labeled cRNA for each of the 208 samples, again following manufacturer’s recommended protocols. The order in which the samples were processed was randomized to minimize chip effects. The beadchips were hybridized and scanned with an Illumina BeadArray reader by Dr Kevin Schianna’s laboratory at the Duke University Institute for Genomics and Science Policy (IGSP). The raw intensities were extracted with the Gene Expression Module in Illumina’s BeadStudio software. Expression intensities were log2 transformed and median-centered by subtracting the median value of each array from each intensity value. The top 22,300 transcripts with expression above background levels averaged across all the arrays were retained for further analyses as previously described (Idaghdour et al. 2008).

**Genotyping**

A total of 209 samples were assayed with Illumina’s Infinium Human 610-Quad beadchip following standard procedures, also at the Duke University IGSP. The Human 610-Quad SNP Chip contains over 610,000 markers based on HapMap release 23. The beadchips were imaged using Illumina’s BeadArray Reader and genotype calls extracted
with the Genotyping Module in Illumina’s BeadStudio software. Six samples with low intensity or low call rate using the Illumina cluster measure (<95%) were deleted and all SNPs that had a call frequency below 99% were deleted. SNPs with a cluster separation value below 0.3 were checked manually and those that could not be fixed manually were deleted. Next, to screen for departure from Hardy-Weinberg equilibrium, the quality of the raw and normalized data of autosomal SNPs with heterozygosity excess values between -1.0 to -0.1 and 0.1 to 1.0 was checked and any SNP cluster that was not clean was deleted. The process of quality control checks resulted in retention of over 570K SNPs in 203 individuals.

**Population Structure and Ancestry Inference**

Principal component analysis (PCA) and a Bayesian approach were implemented in *Eigenstrat* (Price *et al.* 2006) and *Structure* (Pritchard *et al.* 2000), respectively, to explore genetic structure among the samples. PCA was used to infer the extent of global genotypic variation from either the entire or subsets of over 570K SNPs dataset. *Structure* was used to infer population structure using a subset of approximately 16,000 autosomal SNPs (randomly selected and approximately uniformly distributed on the 22 autosomes) at \( K=2-5 \) using the admixture model with correlated allele frequencies with 20,000 iterations after a burn-in length of 20,000. Relatedness (IBS) between all pairs of individuals was estimated using PLINK (Purcell *et al.* 2007).
**Principal and variance component analyses and ANOVA**

Principal and variance component analyses were jointly performed on gene expression data using *JMP Genomics* v3.2 (SAS Institute, Cary NC). Expression principal components (ePC) were modeled as a function of various effects, assuming that each is a random term. A series of models were used to partition variance components into different combinations of the following factors and their pairwise combinations: Location, Gender, and gPC2. The magnitude and significance of differential expression of individual transcripts was evaluated using analysis of variance (ANOVA) through *JMP Genomics* using PROC MIXED as implemented in *SAS*.

**Clustering and functional enrichment annotation**

Clustering was generated with Ward's method in *JMP Genomics*. The gene ontology, pathway, and biomarker analyses were generated through the use of *Panther* (Thomas *et al.* 2003), *Kegg* (Kanehisa & Goto 2002, Kyoto Japan), and *Ingenuity Pathways Analysis* (IPA: Ingenuity, Seattle WA). Genes whose expression was significantly differentially regulated were included using stringent cutoffs as described in the Results section. Enrichment analysis was used to calculate the probability that the number of genes in each biological function, pathway and/or disease assigned to that data set is greater or less than expected by chance given the numbers of genes expressed in the samples. Corrections for multiple testing were achieved using Bonferroni or Benjamini-Hochberg methods depending on the analysis.
Results

Over a two month period, I obtained leukocyte and whole blood samples from a total of 284 individuals. These were used to generate gene expression profiles and whole genome genotypic data from 215 and 209 individuals, respectively. Seven expression-profiled and six genotyped samples were removed after quality control checks resulting in the use of 208 and 203 samples in subsequent analysis.

Population Structure in the Souss region of southern Morocco

I analyzed genotypic data for 203 individuals from four locations (Figure 1): 69 individuals from Anza, 41 individuals from Dcheira (both sites in Agadir), 54 individuals from Ighrem, and 39 individuals from Boutroch. The initial data set (598,828 SNPs) was curated following strict quality control checks as described in Materials and Methods, resulting in the use of 579,144 SNPs in the population structure analysis.

First, relatedness was calculated between all pairs of study participants using the identity by state (IBS) estimator in PLINK (Purcell et al. 2007). This analysis confirmed known family relationships in the sample, and revealed a small number of cryptic groups of related individuals unknown at the time of sample collection. Based on this analysis 65 individuals were found to belong to 25 clusters of related individuals, or 58 pairs with a pi-hat > 0.125 (Purcell et al. 2007) with first cousins having a pi-hat value of approximately 0.125. The 25 clusters correspond to 15 clusters of 2-3 full siblings (pi-hat
~ 0.5), 6 clusters of 2-3 related, but not full siblings, individuals (0.125 < pi-hat < 0.3), and 4 clusters of 4-5 clusters of a mix of both categories. In total, there are 163 unrelated individuals are present in the sample and those include 138 individuals with a pi-hat for all possible pairs < 0.13 plus a randomly chosen member of each of the 25 clusters of related individuals. Since family relationship biases estimates of population structure, only unrelated individuals were included in the final population structure analyses. In the next chapter I adopt a slightly more sophisticated measure of relatedness that does not influence the conclusions described here.

Initial Structure analysis (Pritchard et al. 2000) was run using approximately 16,000 markers on the entire sample and confirmed the effect of family relationship on the method, since the first dozen levels were populated by just a few relatives. Therefore, only the 163 unrelated individuals were used in subsequent population structure analysis. The strongest component (K=2) differentiated ten to twelve unrelated individuals that we assume, on the basis of dark skin color of four participants in this group, reflects a sub-Saharan genotypic contribution to the gene pool of the sample. At K=3, the results show that Boutroch samples are almost exclusively assigned to the second component (Amazigh) while most of the rest of the samples are a mixture at different degrees of the two or three components (Figure 2A).

I next applied principal component analysis as implemented in Eigenstrat (Price et al. 2006) to the full set of genotype data. The method is non-parametric and thus not burdened by the many assumptions concerning admixture of ancestral populations that
must be made when using the *Structure* algorithm. Again, 163 unrelated individuals were used in the analysis and these included one representative of each sibling cluster. The relatives from each sibling cluster that were not included were assigned the same principal component eigenvalue score as the included representative from the same cluster for downstream analyses of gene expression. Using 570K SNPs, *Eigenstrat* indicated the presence of seven significant principal components or eigenvectors \((p < 0.05)\) by the Tracy-Widom method of Patterson *et al.* (2006). Three individuals whose scores exerted the largest influence on gPC1 (genotypic principal component axis 1) have dark skin (the same individuals first identified by *Structure* at K=2), further supporting the inference that a sub-Saharan genotypic contribution present in relatively few individuals is the strongest population component in the data. A corollary is that the Arab and Amazigh groups are more closely related to one another than either are to sub-Saharan Africans. The sub-Saharan-associated gPC nevertheless became insignificant after the removal of seven individuals with the largest scores. Applying the *Eigenstrat* outlier removal function to the dataset resulted in the same conclusion. On the other hand, gPC2 is clearly capturing a major Arab-Amazigh continuous axis of variation (Figure 2B) as it almost perfectly separates Boutroch Amazighs (red squares) from Agadir Arabs (green pluses). Interestingly, Agadir Amazigh (red triangles) and both Ighrem Arabs and Amazigh (circles) are intermediate for gPC2.

gPC3 clearly differentiates the Arab villages from Ighrem (not shown). However, close inspection of SNP loadings for this PC axis and also for the remaining significant
principal components revealed a dominance of a small number of SNPs most of which map to the same region of the genome on Chromosome 1. Similar observations suggest that all gPC axes 3 through 7 reflect local chromosomal effects rather than whole genome patterns. Re-running Eigenstrat analysis on three subsets of 100,000 randomly selected marker revealed that only gPC axes 1 and 2 could be consistently recovered in the same rank order. Price et al. (2008) have recently shown that a few markers in strong linkage disequilibrium (LD) in a subset of samples can create significant principal components, and recommend ignoring such eigenvectors gPC3 when inferring effects of pattern structure on trait variation. Since the purpose of the gene expression analysis presented in this chapter is to jointly estimate genome-wide Arab-Amazigh ancestry and non-genetic regional effects on gene expression variation, only the gPC2 Eigenstrat scores were used as a quantitative estimate of Arab-Amazigh ethnicity in the analyses reported below.

Three-way comparison between regions

Illumina’s HumanHT12 beadchip was used to obtain expression data from 215 individuals. The chip contains over 48,802 transcript probes that include 12,646 non-annotated or poorly annotated transcripts and 36,156 probes for which comprehensive annotation information is available (RefSeq). Expression analysis was restricted to the RefSeq transcripts, and a further 13,856 probes were removed for which transcript abundance was deemed to be blow background. Seven individuals were removed after performing quality control checks as described in Materials and Methods.
I performed an initial analysis of variance including terms of Location, Gender and their interaction without accounting for genetic ethnicity differences between locations. The Location effect in this model encompasses both genetic ethnicity and non-genetic effects and I will be referring to the latter as the environmental geography effects later in this chapter. I first assigned four locations (the two villages of Ighrem and Boutroch, and the two urban sites of Anza and Dcheira, both within the city of Agadir). However, no significant differences were detected between Anza and Dcheira even when false discovery rate (FDR) was increased to 5%, so for all subsequent analyses these two sites were combined and are referred to as Agadir. Consequently, the Location term has three levels: Agadir, Ighrem, and Boutroch, whereas region refers to the urban (Agadir) versus rural (Ighrem plus Boutroch) comparison.

Re-running the analysis using three locations showed that a total of 8,911 of the 22,300 (40%) expressed transcripts were differentially expressed with respect to Location at the 1% FDR cutoff (Table 1), rising to 11,444 transcripts (51%) at 5% FDR. In the same analysis only 162 transcripts were differentially expressed with respect to Gender at 1% FDR (451 at 5% FDR). The large environmental geography influences on the leukocyte transcriptome are visible on two-way hierarchical clustering heat maps. Individuals particularly from Boutroch tend to cluster together even when all 22,300 expressed transcripts are visualized (Figure 3A). As expected, clustering of the majority of individuals from the same location became even more obvious when only
differentially expressed transcripts between locations were included in the analysis. For example the heat map in Figure 3B shows two major groups of genes that clearly differentiate Agadir and Boutroch individuals.

**Principal and variance component analyses (PCA/VCA)**

I next carried out principal component analysis on the total dataset of expressed transcripts. A feature of this analysis is the presence of strong correlation structure in the data such that expression PC axis 1 (ePC1) explains 21\% and ePC1 through gPC5 combined explain 50\% of the transcriptional variance. Remarkably, the distinctiveness of the expression profiles of each location within the entire sample is readily apparent in the plot of the two major principal components of the expression variance of the full dataset (Figure 4A). Boutroch villagers (blue) tend to have high values of PC2 and low values of PC1; Ighrem villagers (green) have high values of PC1 in general; and Agadir residents (red) typically have low PC2 scores.

Variance component analysis shows that almost half (47.6\%) of the variation captured by ePC1 through ePC5 can be decomposed into effects of the Arab-Amazigh axis of variation (gPC2), Location, Gender, and pairwise interactions among these factors (Figure 4B). This analysis is in large agreement with the gene-specific ANOVA, which revealed similar magnitudes of contribution of the various effects. For example, both methods shows that Gender alone has the weakest influence on gene expression, while the combination of Location and Arab-Amazigh genetic ethnicity affects 40\% of the transcripts by ANOVA (1\% FDR), similar to the effect of these factors in the PCA/VCA
analysis (Location + gPC2 + Location*gPC2 interaction = 35.6%). Overall, principal and variance component analyses using the entire dataset suggest that both genetic and non-genetic effects contribute significantly to the transcriptional variation in leukocytes.

An examination of the nature of the interactions by plotting the two major principal components for the 1500 most differentially expressed genes for main effects and their interactions indicates the presence of complex interactions (Figure 5). In this filtered dataset, the samples largely separate by lifestyle along ePC1 (urban red, positive values; rural green and blue, negative values; Arabs circles; Amazighs squares) and gender along ePC2 (females light colors, negative values; males dark colors, positive values). Closer analysis highlights further structure since all 15 men, as well as the 10 Amazigh women, from Ighrem cluster with urban samples. Almost all of the strongly negative values for ePC1 are observed for the Arab women from Ighrem and the predominantly Amazigh men and women from Boutroch. Ighrem males more often commute to the neighboring cities of Tiznit and Agadir and certainly experience more urban-like environmental and cultural influences compared to Ighrem females. Clustering of Ighrem Amazigh women with Agadir females is unexpected and there is no clear indication of the factors contributing to this pattern although it is likely due to an ethnicity-by-location/lifestyle or a gender-by-ethnicity-by-location effect. The presence of such complex correlations in the data is also reflected in the VCA since interaction effects account nearly for as much as main effects account for (Figure 4B). Similar principal component analyses on the entire subset of the differentially expressed genes
and also on random subsets of genes were performed and overall, all reveal the presence of a strong and complex signature of interaction effects seemingly implicating genetic and environmental geography interactions with no evidence for a significant role for gender effect.

Taken together these observations, along with the observed differences between rural and urban lifestyles, suggest that environmental geography directly influences leukocyte gene expression in the Souss region of southern Morocco. Statistically speaking, there is only weak evidence for significant contributions of the effects of gender, genetic ethnicity, or interactions between these two factors in the total sample. However, Ighrem villagers show differences in the profiles of men and women, the latter according to whether they are Arab or Amazigh, implicating cultural and behavioral factors rather than intrinsic genetic or sexual ones.

**Pairwise comparisons of regions**

Breaking down the analysis of variance into pairwise comparisons shows that the Location effect remains the strongest followed by the Females-by-Location and the Males-by-Location, with only a relatively weak effect of Gender alone (Table 1). The amplitude and significance of differential expression for each effect is visually shown in the volcano plots in Figure 6. On inspection of the Location effect stratified by Gender, gene expression is much more strongly influenced in females (17.7% on average at 1% FDR) than males (4.9% on average at 1% FDR), suggesting that most of the differences between locations can be attributed to the differences between females. This observation
is readily visible in the volcano plots in Figure 6: columns A, B and C show contrasts among the Locations, while the rows indicate the overall, female, and male-specific effects. As pointed out above, the reduced differential gene expression between males in Agadir and Ighrem might be explained by the similarity of their lifestyle.

Pairwise comparisons within each major effect revealed that the strongest regional effect is observed between the rural village of Boutroch and either the village of Ighrem or the city of Agadir, contrary to a priori expectation that the two rural villages would show similar expression profiles. The Venn diagram in Figure 7 shows that the majority of transcripts differentially expressed between Boutroch and Agadir (75%) are also differentially expressed between Boutroch and Ighrem (69%) further illustrating the differentiation of Boutroch and the resemblance of expression profiles of Ighrem and Agadir relative to Boutroch.

*Estimating the effect of genetic ethnicity within a single region*

Partitioning the sources of variance for various effects in the rural villages is complicated given the confounding of Ethnicity with Location in both villages. However, since no significant expression differences were detected between the urban sites of Anza and Dcheira and given that both Arabs and Amazigh individuals are equally represented in both sites, partitioning the sources of variance for Ethnicity and Gender in the urban sample only does not suffer from such caveat. A mixture of 96 Arab and Amazigh urban individuals were selected for this analysis after excluding nine urban
individuals that separate from the rest of urban samples along gPC1 as expression in these may be affected by their sub-Saharan component of genetic variation.

First, PCA of the entire transcriptome in this dataset showed over 53% of the variance is captured by ePC1 through ePC5. Variance component analysis after weighting the average proportion for each effect across ePC1 through ePC5 indicated that a total of just 10.6% of the variance is explained by gPC2 and its interaction with Gender (Figure 8A). To test for the effect of ethnicity as a categorical variable (Arab/Amazigh), I used self-reported ethnicity but after re-assigning ethnicity to 13 individuals based on gPC2 scores. Although some individuals are admixed, the majority of urban samples can be clustered within either ethnic group. Re-running the variance component analysis of expression principal components revealed only 1.4% of the variance is explained by categorical ethnicity alone, an effect as weak as that of gender (Figure 8B). Second, an analysis of variance using categorical ethnicity failed to detect any differences even at 5% FDR (Figure 8C). This analysis suggests that the effect of Arab-Amazigh ethnicity is weaker compared to the effect of non-genetic regional influences, and hence that the differences between Boutroch and Ighrem are more likely due to geographic than genetic factors. As argued earlier, ethnic differences contribute to differentiation between the villages, but it is unlikely that they are the major factor.

**Functional annotation of transcriptional differences**

*Ingenuity IPA* and *Panther* were used to investigate differential expression according to gene ontology classes and pathway analysis. *IPA* analysis allows the
application of very stringent criteria where highly constrained filters can be applied to the data. I increased the stringency of these analyses so as to only include those transcripts and relationships that are curated in Ingenuity’s knowledge database and are very relevant to human immune system cells (B and T lymphocytes, Dendritic cells, Macrophages, Monocyte-derived macrophage, Monocytes, Neutrophils, and NK cells). The objective of the analysis is to depict the most highly relevant molecular processes, functions and diseases related to differentially expressed genes.

First, a comparison of global differential expression between Agadir and Ighrem (763 transcripts at 1% FDR) implicates divergence in key immune functions. Interestingly, antigen presentation is the most significantly enriched cellular function (15 genes, \( p < 0.0007 \), Benjamini-Hochberg) in the list of differentially expressed transcripts between Agadir and Ighrem, likely reflecting differential exposure to different types and populations of pathogens. This observation is consistent with the observation that two of the three disease categories that show significant enrichment are infection mechanisms and inflammatory response, suggesting differences in core immune responses to specific surrounding microbial environments. IPA also revealed that a single disease, atherosclerosis, exhibits significant enrichment for gene expression in this analysis. In fact, five atherosclerosis marker genes (\( FOSB, CDKN1A, EGR1, KLF4 \) and \( PTGS2 \)), upregulation of which in circulating monocytes and plaque macrophages purified from atherosclerosis patients has been reported by Patino et al. (2006). Strikingly, four of
these genes are consistently upregulated in Ighrem and downregulated in Boutroch with Agadir being intermediate as shown by analysis of variance.

The analysis of variance between locations demonstrated the striking differentiation of the village of Boutroch relative to Ighrem and Boutroch. Also, as shown in the Venn diagram in Figure 7, most of the differences between Boutroch and Ighrem are also shared between Boutroch and Agadir. To explore the functional characteristics of the genes differentiating Boutroch and Agadir, I ran Panther enrichment analysis on the 1399 most significant differentially expressed genes between Agadir and Boutroch ($p < 2.4 \times 10^{-6}$, significant after Bonferroni correction for multiple comparisons). This analysis revealed very strong enrichment for biological processes involving protein biosynthesis, signal transduction, and oxidative phosphorylation. The subcategories for biological processes and molecular functions are shown in Figure 9, and Table 2.

Inspection of subgroups of genes involved in protein biosynthesis shows the presence of a very strong enrichment for a group of genes encoding a functionally and structurally closely related group of ribosomal proteins. The entire list of expressed transcripts (22,300) contains 186 ribosomal protein genes, 156 of which (83%) are differentially expressed at 1% FDR. Intriguingly, 76 genes of those are in the list of the top 1399 differentially expressed transcripts (at Bonferroni significance) while only 16 genes are expected by chance. The striking enrichment for a closely related group of ribosomal proteins was confirmed by an exploration of the translation machinery using
*Kegg* analysis (Figure 10). Furthermore, a family of SNORD small nucleolar RNA genes is also strongly differentiated between the three locations probably indicating a link to ribosome biogenesis that takes place in the nucleolus (Olsen *et al.* 2000).

Ribosomes are the central entity of the protein synthesis machinery and their production is a major metabolic activity coordinated in the nucleolus. In humans, ribosome production is a highly coordinated process that aims to end up with equimolar production of about 80 ribosomal proteins that play a major role in this process that expends a large proportion of the energy produced in the cell (Warner 1999). In fact, a strong signature of high production of energy along with upregulation of ribosomal protein (RP) genes is evident from the enrichment analysis as I observed a strong upregulation (*p* < 2 x 10^{-6}, Bonferroni) of 21 genes involved in oxidative phosphorylation, the major biological process involved in energy production in the mitochondrion (Figure 11). Further, upregulation of the mitochondrial ribosomal protein transcripts that are encoded by nuclear genes indicate that the mitochondrial translation machinery also is active and it is known for instance that several genes involved in oxidative phosphorylation are translated in the mitochondrion. These expression patterns are consistent with each other and most likely indicate differential regulation of cell growth and proliferation among the locations.
Discussion

Population structure of the Souss region

This genetic analysis of a southern Moroccan population has revealed specific variation patterns in the Souss, some of which were unexpected. Self-reported ethnicity of Arab and Amazigh individuals proved useful in guiding the interpretation of Eigenstrat and Structure analyses results allowing me to address the source of inconsistencies among previous population genetic studies in the region (Bosch et al. 2001, Brakez et al. 2001, Chbel et al. 2003, Coudray et al. 2006). These are likely due to discrepancies between different types of markers, to the difficulty in inferring degrees of admixture from self-reported information, and to confusion between cultural and genetic identities. In fact, population structure analysis presented herein supports all three scenarios.

It is not always trivial to distinguish between present-day cultures and genetic ancestry, particularly when the physical characteristics of the founding population are similar, as in Morocco, leading to inadequate classification of individuals. It is well known for instance that a process of Arabization occurred in many North African populations, certainly resulting in some Amazigh groups identifying themselves as Arabs. In fact, to my surprise the village of Ighrem is genetically intermediate between Agadir Arabs and Boutroch Amazigh, contrary to my a priori expectation that Ighrem Arabs would cluster with Agadir Arabs. I had also assumed that Agadir would consist of
Amazigh individuals who would cluster with Boutroch, as well as of individuals who would be a clear mixture of the two ethnicities.

The data does suggest that admixture is occurring in a fraction of the sample, notably in Agadir and probably in the village of Ighrem. On the other hand, almost all Boutroch individuals seem to truly represent a non-admixed sample of southern Moroccan Amazighs. This observation supports the idea that many present-day Moroccan tribes and villages can incontestably be identified as of Amazigh ancestry given their known history of isolation in the Atlas Mountains, their ancient and Amazigh-specific cultural customs, and also their total ignorance of the Arabic language. This is also true for the first-generation of urban dwellers that moved from known isolated Amazigh villages.

Generally, Arab and Amazigh ethnicities seem to have genetic affinities that may reflect shared past history even though some groups such as Agadir Arab and Boutroch Amazigh show clear genome-wide differentiation. Several studies have emphasized the genetic affinities of Amazigh with Eurasians in contrast to sub-Saharan Africans. For example, a mitochondrial DNA control region survey of 31 bone remains from an archaeological site in Morocco suggested that the early inhabitants of Morocco arrived at least 30,000 years ago, with subsequent migrations all being from groups with Eurasian affinities, the major group being Arabs (Kéfi et al. 2005). This said, comprehensive genome-wide surveys of the Moroccan population genetic structure that include control
individuals from both ethnicities and also other ethnic groups remain to be carried out systematically.

**Sources of gene expression variation**

Comparative analysis of leukocytes gene expression profiles within and between the four sampled Moroccan communities was conducted, revealing striking patterns of variation. Four specific alternative hypotheses regarding these patterns were tested:

**Hypothesis 1:** Gene expression differentiation between the two rural locations is greater than between the urban locations.

Gene expression differentiation is indeed greater between the two rural villages than between the urban sites as the initial analysis of variance failed to detect significant differentiation between Anza and Dcheira. This was not surprising as both sites are within the city of Agadir and only seven miles apart. Individuals from both sites have a similar lifestyle and are exposed to similar environmental influences. The presence of both ethnicities in both sites could also be masking any transcriptional differences caused by environmental effects.

**Hypothesis 2:** Gene expression differentiation between rural and urban lifestyle is greater than between locations within each lifestyle.

This hypothesis was rejected only for the comparison contrasting the differentiations between lifestyles on the one hand, and between urban sites on the other hand. Contrary to my expectations that the strongest regional effect would be observed between the city of Agadir and both villages, striking differences were observed between
the two villages of Ighrem and Boutroch. This conclusion is anchored by the observation that Ighrem is less differentiated than Agadir relative to Boutroch. The fact that Ighrem and Boutroch are not equally differentiated from Agadir can be attributed to several reasons. One possible explanation is the geographic differences between the two villages. Boutroch is located within the low Atlas mountains and has an altitude 615 meters higher than Ighrem that is located 26 miles far away in the foothills of the mountains. Another possibility is the closer location of Ighrem and its accessibility to the city relative to Boutroch. Further, it is plausible that the observed relatively greater genetic differentiation of Boutroch individuals is adding transcriptional differentiation to pronounced non-genetic effects in the more elevated and isolated region. Ighrem individuals are certainly genetically more similar to Agadir individuals on gPC2 used in the analysis, but this affinity is unlikely to be accounting for the much closer expression profiles.

**Hypothesis 3:** Ethnicity has a greater impact than environmental geography on gene expression differentiation.

Testing the hypothesis was somewhat complicated by confounding of Ethnicity with Location in rural villages. Nevertheless, strong evidence suggests that genetic ethnicity is neither the sole nor the major source of the striking gene expression differentiation observed between the three groups of southern Moroccans. The main observation that supports this conclusion is the lack of significant differentiation between Arab and Amazigh individuals in the city despite a clear separation along gPC2, even
when a relaxed FDR significance threshold (5%) was considered. Further observations relating to the issue of relative contributions of genetic ethnicity and environmental geography on gene expression differentiation are addressed below.

**Hypothesis 4:** Interaction effects between gender, genetic ethnicity, and environmental geography have a significant impact on gene expression variation.

Strong evidence suggests that this hypothesis can only be rejected only for the gender effect and its interactions with both genetic ethnicity and environmental geography factors. Principal and variance component analyses combined suggest that individuals cluster most under the effects of location/lifestyle or location/lifestyle-by-ethnicity. For example, both Ighrem Arab males and Ighrem Amazigh females clearly separate from their respective genetically closest groups (Ighrem Arab females and Boutroch Amazigh, respectively), with the rest of samples largely clustering by lifestyle regardless of genetic ethnicity. Further, since we know that differences between males and females due to the Gender effect are minor (i.e. Chapter 2 results, urban females/urban males and Boutroch males/Boutroch females comparisons in this chapter), any major differences between the two genders are most likely due either to location/lifestyle or its interaction with genetic ethnicity. There is no indication as to what factors are causing a dozen Ighrem Amazigh females to cluster with urban females along ePC1, but it is important to bear in mind that PCA is by definition based solely on gradually capturing the maximum variation along orthogonal axes that may or may not have a biological interpretation. In any case, the data clearly shows that the leukocyte
transcriptome of southern rural Moroccans is strongly affected by complex environmental geography and genetic interactions.

Overall, the results reported here are consistent with the findings reported in Chapter 2. For example, the amplitude of differential expression observed between regions reported here and in Chapter 2 for the three-way comparisons are comparable (40% and 37%, respectively). The regional effect reported in Chapter 2 was estimated while minimizing the effects of genetic differences as only Amazighs were sampled but the possibility that the Arab-Amazigh genetic differences are contributing to the observed increase of the magnitude of the location effect between Ighrem and Boutroch could not be ruled out.

All the groups sampled in both studies are southern Moroccans, except the nomadic group in the first study, and live in the Souss region. In both studies over one third of the expressed transcripts are significant for regional effects, with several hundred to thousands of transcripts significant at experiment wide levels. The magnitudes of regional effects are clearly consistent despite sampling different sites, significantly increasing the sample size, and sampling in a different season of the year. These results further confirm that a person's environment considerably impacts their immune system expression profile and hence, it can be inferred, susceptibility to a range of diseases.

**Nature of the expression differences**

The massive transcription of both cytoplasmic and mitochondrial ribosomal protein genes along with a signature of high energy production is intriguing. Global
upregulation of ribosomal proteins may be a read-out of up-regulated cell growth and proliferation, given that synthesis and assembly of ribosome components are prerequisite to the increased protein synthesis required to sustain such a process. This observation further warrants the exploration of the quantitative and qualitative characteristics of various immune cell types in future investigations. The process of active cell growth and proliferation is likely a response to a trigger or series of triggers encountered in both Agadir and Ighrem, and to a less extent in Boutroch. A possible scenario is that microbial load is lower in the mountain or that Boutroch individuals encounter less abiotic environmental stressors. Further, alteration of the expression of oxidative phosphorylation genes might be pinpointing clinical manifestations mediated by oxidative stress such as that which contributes to renal/kidney disease (Di Donato 2009) but further work is needed to confirm this.

Alteration of the expression of ribosomal proteins has consistently been associated with tumorigenesis and several RP genes have been reported to be over-expressed in cancer cells (reviewed in Ruggero & Pandolfi, 2003). It has also been shown in yeast that disrupting the expression of only one RP (*RPL16*) can result in a defect in cellular growth through a global alteration of the entire ribosomal biogenesis (Rotenberg 1988). Similar scenarios have been reported in Drosophila, mice and several human cell lines (reviewed in Ruggero & Pandolfi 2003). Alteration of the expression of a single or few RP genes can therefore have an effect on protein synthesis but it is not clear to what
extent it contributes to the co-expression patterns of the other RP genes I have observed in the data.

Gene expression regulation is a complex process and most genes are regulated by multiple transcription factors. Nonetheless, it is possible that the observed global co-expression pattern of RPs is a result of the effect of master regulator of RPs. Intriguingly, upregulation of \textit{cMYC} and \textit{nMYC} is correlated with over-expression of RPs in both Agadir ($p < 0.002$ and $p < 0.01$, respectively) and Ighrem ($p < 0.006$ and $p < 0.00002$, respectively) but not Boutroch. This observation is consistent with previous finding that RPs are \textit{MYC} gene targets (Coller \textit{et al.} 2000), and the demonstration in Chapter 2 that the \textit{MYC-FOS} network is a highly differentially expressed module between regions. Furthermore, it has been demonstrated that the expression of a large set of RPs in neuroblastoma cell lines is enhanced by \textit{nMYC} (Boon \textit{et al.} (2001) as well in several different cell types (Ruggero and Pandolfi 2003). Further investigations are needed however to explain the trigger causing the upregulation of \textit{cMYP} and \textit{nMYC} in specific regions.

In summary, the examination of regional differences in the peripheral blood leukocyte transcriptome from three southern Moroccan groups confirms striking regional differentiation in both Arabs and Amazigh. Environmental geography influences through direct and interaction effects with genetic factors are the major determinant of those transcriptional profiles. The amplitude and functional features of these differences suggest a significant impact on disease susceptibility and call for similar investigations in
other ethnic groups. Additional insight will be gained by including a systematic
collection of individual phenotypic and endophenotypic information relevant to disease
susceptibility and immune response, further dissection of the environmental and cultural
characteristics of each location, and the inclusion of cell immunotyping of peripheral
blood.
References


Table 1. Differential expression. Numbers and percentages of differentially expressed genes at 1% FDR obtained after fitting an analysis of variance that included Location (I for Ighrem, B for Boutroch, A for Agadir), Gender (F for females, M for males), and Gender-by-Location as fixed terms. Anza and Dcheira are combined in this analysis. Both three-way and two-way comparisons are shown.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Comparison</th>
<th>FDR 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Three-way comparison</td>
<td>8,911 (40%)</td>
</tr>
<tr>
<td>Gender</td>
<td>Gender = (F)-(M)</td>
<td>162 (0.7%)</td>
</tr>
<tr>
<td>Location</td>
<td>Pairwise comparison</td>
<td></td>
</tr>
<tr>
<td>Location = (A)-(B)</td>
<td></td>
<td>7,143 (32%)</td>
</tr>
<tr>
<td>Location = (A)-(I)</td>
<td></td>
<td>760 (3.4%)</td>
</tr>
<tr>
<td>Location = (B)-(I)</td>
<td></td>
<td>7,833 (35.1%)</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>5,245 (23.5%)</td>
</tr>
<tr>
<td>Gender*Location</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location*Gender = (A F)-(A M)</td>
<td></td>
<td>23 (0.1%)</td>
</tr>
<tr>
<td>Location*Gender = (I F)-(I M)</td>
<td></td>
<td>677 (3%)</td>
</tr>
<tr>
<td>Location*Gender = (B F)-(B M)</td>
<td></td>
<td>13 (0.05%)</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>237 (1%)</td>
</tr>
<tr>
<td>Location*Males</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location* Males = (A M)-(B M)</td>
<td></td>
<td>576 (2.6%)</td>
</tr>
<tr>
<td>Location*Males = (A M)-(I M)</td>
<td></td>
<td>8 (0.03%)</td>
</tr>
<tr>
<td>Location* Males = (B M)-(I M)</td>
<td></td>
<td>2,706 (12.1%)</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>1,096 (4.9%)</td>
</tr>
<tr>
<td>Location*Females</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location* Females = (A F)-(B F)</td>
<td></td>
<td>5,292 (23.7%)</td>
</tr>
<tr>
<td>Location* Females = (A F)-(I F)</td>
<td></td>
<td>1,525 (6.8%)</td>
</tr>
<tr>
<td>Location* Females = (B F)-(I F)</td>
<td></td>
<td>5,055 (22.66%)</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>3,957 (17.7%)</td>
</tr>
</tbody>
</table>
**Table 2. Panther enrichment analysis.** Panther enrichment analysis for the top 1399 most differentially expressed transcripts between Agadir and Boutroch. Only biological processes (A) and Molecular functions (B) enriched at Bonferroni significance are shown. Number of genes among the 1399 top differentially expressed, the number of genes expected by chance, and enrichment at Bonferroni significance are shown.

<table>
<thead>
<tr>
<th>Biological Process</th>
<th># Expressed Genes</th>
<th># Genes in Diff List</th>
<th>Expected</th>
<th>Enrichment</th>
<th>Significance Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein biosynthesis</td>
<td>300</td>
<td>98</td>
<td>26.71</td>
<td>+</td>
<td>3.55E-25</td>
</tr>
<tr>
<td>Signal transduction</td>
<td>2125</td>
<td>117</td>
<td>189.2</td>
<td>-</td>
<td>2.35E-08</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>51</td>
<td>21</td>
<td>4.54</td>
<td>+</td>
<td>0.00000214</td>
</tr>
<tr>
<td>Developmental processes</td>
<td>1380</td>
<td>75</td>
<td>122.87</td>
<td>-</td>
<td>0.0000226</td>
</tr>
<tr>
<td>Protein metabolism and modification</td>
<td>2070</td>
<td>244</td>
<td>184.3</td>
<td>+</td>
<td>0.000071</td>
</tr>
<tr>
<td>Cell communication</td>
<td>723</td>
<td>32</td>
<td>64.37</td>
<td>-</td>
<td>0.000547</td>
</tr>
<tr>
<td>Cell surface receptor mediated signal transduction</td>
<td>951</td>
<td>48</td>
<td>84.67</td>
<td>-</td>
<td>0.000792</td>
</tr>
<tr>
<td>Electron transport</td>
<td>168</td>
<td>33</td>
<td>14.96</td>
<td>+</td>
<td>0.00103</td>
</tr>
<tr>
<td>Cell adhesion-mediated signaling</td>
<td>220</td>
<td>4</td>
<td>19.59</td>
<td>-</td>
<td>0.0042</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Molecular Function</th>
<th># Expressed Genes</th>
<th># Genes in Diff List</th>
<th>Expected</th>
<th>Enrichment</th>
<th>Significance Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ribosomal protein</td>
<td>186</td>
<td>76</td>
<td>16.56</td>
<td>+</td>
<td>6.58E-25</td>
</tr>
<tr>
<td>Signaling molecule</td>
<td>507</td>
<td>17</td>
<td>45.14</td>
<td>-</td>
<td>0.0000299</td>
</tr>
<tr>
<td>Receptor</td>
<td>845</td>
<td>39</td>
<td>75.24</td>
<td>-</td>
<td>0.0000498</td>
</tr>
<tr>
<td>Nucleic acid binding</td>
<td>1867</td>
<td>224</td>
<td>166.23</td>
<td>+</td>
<td>0.0000585</td>
</tr>
<tr>
<td>Oxidoreductase</td>
<td>416</td>
<td>62</td>
<td>37.04</td>
<td>+</td>
<td>0.00243</td>
</tr>
<tr>
<td>Zinc finger transcription factor</td>
<td>631</td>
<td>29</td>
<td>56.18</td>
<td>-</td>
<td>0.00563</td>
</tr>
<tr>
<td>Transcription factor</td>
<td>1354</td>
<td>85</td>
<td>120.55</td>
<td>-</td>
<td>0.00612</td>
</tr>
<tr>
<td>Membrane-bound signaling molecule</td>
<td>92</td>
<td>0</td>
<td>8.19</td>
<td>-</td>
<td>0.0433</td>
</tr>
</tbody>
</table>
Figure 1. Geographic locations of sampled groups in the Souss region in southern Morocco. Peripheral blood samples were collected from two lifestyles: urban (two sites within the city of Agadir: Anza (Latitude = 30.447, Longitude -9.655, Altitude: 30 meters) and Dcheira (Latitude = 30.365, Longitude -9.532, Altitude: 12 meters), 7 miles apart) and rural (two rural villages: Ighrem and Boutroch, 26 miles apart). The urban and rural lifestyles are separated by approximately 80 miles. Boutroch village is known to be a predominantly Amazigh village and is in the low Atlas mountains (Latitude = 29.346, Longitude -9.368, Altitude: 1335 meters), whereas Ighrem is located in the foothills of the low Atlas mountains (Latitude = 29.459, Longitude -9.672, Altitude: 720 meters) and is historically Arab with a small fraction of Amazigh residents. The white spots indicated the cities of Agadir and Tiznit.
Figure 2. Population structure analysis. (A). *Structure* analysis of genotypic variation using an admixture model with correlated allelic frequencies and using 16,000 autosomal SNP markers. Each of the 163 unrelated individuals is represented by a column that is partitioned into up to 3 colored segments representing the proportion of ancestry (Q value) from each of the 3 clusters for each individual. The four sampled groups are labeled as 1 (Anza), 2 (Dcheira), 3 (Boutroch), and 4 (Ighrem). (B). *Eigenstrat* analysis of population structure. The plot shows the first two eigenvectors of genotypic variance for 159 unrelated individuals (red, self-reported Amazighs; green, self-reported Arabs; black, unknown or dark skinned individuals; open squares, Boutroch; + and triangles, Agadir; circles, Ighrem). Five outlier individuals on PC axis 1 are not shown.
Figure 3. Two-way hierarchical clustering of expression data from Agadir and Boutroch individuals. In (A), each row represents one of the 22,300 expressed transcripts and each column represents one individual. (B) shows clustering using the top 1399 most significantly differentiated transcripts. Intensity of red indicates relatively high expression relative to the sample mean, of blue relatively low expression. The red and the black bars indicate Agadir and Boutroch individuals, respectively, with the exception of individuals circled in black (Boutroch) or red (Agadir). Note the two major cliques differentiating Agadir and Boutroch in B. The clustering was generated with Ward's method in *JMP Genomics ver. 3.2*. 
Figure 4. Principal and variance component analyses of the leukocyte transcriptome. Principal (A) and variance (B) component analyses of 22,300 genes expressed in the sample. The two major principal components explaining 31.6% of the expression variance are plotted for 194 individuals. In (A) Location is indicated by color (red; urban, green; Ighrem, and blue squares; Boutroch). Shape refers to lifestyle (circle; urban and square; rural). Males and females are colored in dark or light, respectively. Note clustering of Boutroch individuals readily visible. Weighted average contributions of various effects to transcriptional variation for the first five expression principal components (ePC1-5) combined explaining about 50% of the variance of the entire expression dataset. The effects are Gender, Location, Eigenstart’s gPC2, and their pairwise interactions. The unexplained variance is referred to as Residual.
Figure 5. Principal component analyses of genes differentially expressed. Principal component analyses of top 1,500 genes differentially expressed for all effects in the three-way ANOVA test. The two major principal components explaining 40% of the expression variance are plotted for 194 individuals. Lifestyle is indicated in color (urban red, rural green and blue). Squares and circles indicate Arabs and Amazigh, respectively. Groups of individuals discussed in main text are further labeled based on location and self-reported ethnicity. Circled groups indicate where the majority of those individuals are on the plot: cluster 1: Ighrem Arab females, cluster 2: Boutroch females (Amazigh), cluster 3: Ighrem Amazigh females, cluster 4: Ighrem males (Arabs), cluster 5: Boutroch males (Amazigh).
Figure 6. Volcano plots of statistical significance versus magnitude of differential expression between regions. For each transcript, significance is shown as the negative logarithm of the p-value on the y-axis, and the log base 2 of magnitude of mean expression difference is on the x-axis. Dashed lines indicate the threshold for significance at 1% FDR. Three pairwise comparisons between locations (Agadir, A; Boutroch, B; and Ighrem, I), and gender within locations (Females, F, and Males M) are shown. A-B, A-I, and B-I comparisons are shown in columns A, B, and C, respectively. Note that the greater differentiation between Regions is observed in females.
Figure 7. **Overlap between differentially expressed genes.** The Venn diagram shows the numbers of differentially expressed genes at 1% FDR for each pairwise regions comparison (A, Agadir; B, Boutroch, I, Ighrem) and the overlaps between them. The majority of transcripts differentially expressed between Boutroch and Agadir (75%) are also differentially expressed between Boutroch and Ighrem (69%).
Figure 8. Effects of Ethnicity, Gender, and their interaction in the urban sample. Variance component analysis using Eigenstart’s gPC2 (A) and adjusted self-reported ethnicity –adSREth- (B). In (C), Volcano plots show the effect of Gender and adjusted self-reported Ethnicity from an analysis of variance.
Figure 9. Panther enrichment analysis. Enrichment analysis for the top 1399 most differentially expressed transcripts between Agadir and Boutroch. Only biological processes (A) and Molecular functions (B) enriched at Bonferroni significance are shown. The red bars indicate the number of genes among the 1399 relevant to the corresponding process or function while the blue bars indicate the number of genes expected by chance. Numbers of genes and p-values for significance are shown in Table 2.
Figure 10. *KEGG* Ribosome assembly map. The figure shows the cytoplasmic ribosomal protein genes that are upregulated in the city of Agadir (red stars). The green rectangles correspond to *Homo sapiens* genes.
Figure 11. *KEGG* Oxidative phosphorylation pathway. The figure shows the genes that are upregulated in the city of Agadir (red stars). The green rectangles correspond to *Homo sapiens* genes.
CHAPTER FOUR

Mapping the Genetic Architecture of the Leukocyte Transcriptome
Abstract

Genetic and environmental alterations of gene expression regulation are likely to impact physiology and disease susceptibility. Mapping the genetic architecture of transcript abundance has opened up the opportunity to discover thousands of regulatory genetic variants, however, little is known about the modulation of these effects by the environment. Here I report a gene expression genome-wide association study on 194 leukocyte samples collected from three Moroccan communities living in three distinct geographic locations. Associations were tested for each of 516,972 genotypes against each of 22,300 expressed transcripts while accounting for gender, location, genetic ethnicity, relatedness, and interaction effects. The analysis revealed 1,744 genome-wide significant associations involving 396 unique autosomal transcripts and 1,427 eSNPs. Annotation information was available for eSNPs and genes involved in 1,636 associations. The majority of these are either cis local (within 500Kb of the expression probe; 1,242 or 76%) or cis distant (over 500Kb of the expression probe; 327 or 20%), all involving an eSNP and a target on the same chromosome, and only 4% (67) in trans with an eSNP and its target located in different chromosomes. All these associations are robust to population structure and environmental perturbation, and no evidence for genotype-by-environment interactions modulating transcript abundance was detected. Further, I confirmed numerous previously reported regulatory signals and validated several previous trait and disease GWAS findings as being modulated by variation in
gene expression levels. These results further our understanding of gene expression variation in humans, and show the utility of genome scans of eSNPs in helping to understand the role gene expression variation pay in promoting disease.
Introduction

Major advances in large-scale genotyping technologies have provided the opportunity to characterize hundreds to thousands of genetic markers, thereby improving our ability to map variants that underlie variation in heritable phenotypes and to understand their mechanisms of action. Gene expression levels are no exception, as studies mapping expression quantitative trait loci in several species have shown that transcript abundance is heritable, facilitating mapping gene expression regulatory variants. As a result, the large number of genetical genomics studies published in the last few years (reviewed in Rockman & Kruglyak, 2006, Williams et al. 2007, and Gilad et al. 2008) is indicative of the increased interest in characterizing the genotypic modulators of gene expression.

In humans, this line of research has focused on genome-wide association mapping of variants that influence transcript abundance in transformed lymphoblast cell lines (Monks et al. 2001, Stranger et al. 2007, Dixon et al. 2007, Smirnov et al. 2009), buffy coats (Göring et al. 2007), peripheral blood mononucleated cells (Heinzen et al. 2009), leukocytes (Heap et al. 2009), cortical brain tissue (Myers et al. 2007, Heinzen et al. 2008), adipose tissue and whole blood (Emilsson et al. 2008), and liver tissue (Schadt et al. 2008). These studies indicated the presence of significant local (also referred to as cis), and to a much less extent distant (also referred to as trans) control of gene expression. However, distant effects are certainly less tractable because of the
complexity of transcriptional interactions and the caveat of multiple testing. It has also been shown that a fraction of these regulatory effects could be detected in various cell types and/or tissues but evidence for tissue-specific effects has also been provided for mice (Cowels et al. 2002, Petretto et al. 2006) and humans (Schadt et al. 2008, Heinzen et al. 2008, Kwan et al. 2008). Nevertheless, estimates of repeatability reported in these studies for these effects across cells and tissues vary considerably.

Variation in gene expression levels can also be attributed to the interplay between environmental effects and gene expression regulatory variants, as documented in C. elegans and yeast (Li et al. 2006, Smith & Kruglyak 2008). A linkage study in human lymphoblast cell lines also found a predominance of trans-acting factors to be responsible for widespread transcriptional response to irradiation (Smirnov et al. 2009), arguing for an interaction between genotype and environment. Yet no systematic genome-wide investigation of such interactions in humans has yet been performed. Missing also is knowledge about the robustness of genetic regulatory effects to lifestyle and geographic variation in humans. The research presented so far in this thesis demonstrated that the environment can profoundly shape gene expression profiles. Modulation of the effects of genetic regulatory variants is a possible mechanism of action of such alterations in gene expression, with broad implications for understanding the heritability of transcription and hence disease.

Here I follow up of the analysis presented in Chapter 3 and address the effect that each of 516,792 common genetic variants have on gene expression level variation in
leukocytes sampled from 194 Moroccan individuals living distinct lifestyles in different geographic locations, and estimate the robustness of these effects to environmental perturbation. This study includes: (i) a contrast of gene-specific expression and gene-specific genetic differentiation, (ii) a genome scan for \textit{cis} and \textit{trans} expression single nucleotide polymorphisms (eSNPs), (iii) an investigation of environmental modulation of significant genotypic effects, and (iv) a contrast of eSNPs and variants reported in several published trait and disease GWAS.

\textbf{Materials and Methods}

\textit{Gene expression and genotypic data}

Expression and genotypic data were obtained for 194 individuals, using procedures and quality control checks as described in Chapter 3. Briefly, genotypic data was generated using Illumina’s Infinium Human 610-Quad beadchip following standard procedures. The Human 610-Quad beadchip contains approximately 600K SNPs based on HapMap release 23. The beadchips were imaged using Illumina’s BeadArray Reader and genotype calls extracted with the Genotyping Module in Illumina’s BeadStudio software. Illumina’s HumanHT-12 beadchips were used to generate expression profiles of more than 48,000 transcripts with 500ng of labeled cRNA for each of the 194 samples and following manufacturer’s recommended protocols.
Statistical Analyses

Fst estimates between locations were calculated by Kelci Miclaus in Russ Wolfinger’s group at the SAS institute for each of the 516,972 SNPs using PROC ALLELE in SAS 9.2 (SAS, Cary NC). This implementation uses the method of moments approach in an ANOVA framework using expected mean squares to estimate Fst. The method assumes “random” (in contrast to “fixed”) populations and accounts for common evolutionary history. Gene-specific Fst estimates were calculated by averaging Fst measures of all SNPs within each gene and in flanking 5’ and 3’ UTR regions. Various single SNP measures including tests for Hardy-Weinberg equilibrium and allele and genotype frequencies were calculated using PROC ALLELE as implemented in JMP Genomic v3.2 (SAS). Tests for association of gene expression levels with each genotype were performed by Wendy Czika also at the SAS Institute, using both an ANOVA and regression (testing for a linear trend) as implemented in PROC MIXED and PROC LOGISTIC, respectively, using SAS 9.2 and JMP Genomics. First, the entire allelic data set was coded as 0, 1, or 2 where each number represents the number of copies of the major allele. Each SNPs was tested for association with each of the 22,300 expressed transcripts. First the following basic model was used, where \( \mu \) is the mean measure of transcript abundance and the error \( \varepsilon \) is assumed to be normally distributed with a mean of zero:

\[
\text{Expression} = \mu + \text{SNP} + \varepsilon
\]
The results from this model provided a subset of SNPs and transcripts involved in the 10,000 most significant associations to focus on. Gender (Male or Female), Location (Agadir, Ighrem, or Boutroch), and the interaction of location with both SNP and gender were accounted for in Model 2:

\[
\text{Expression} = \mu + \text{Location} + \text{Gender} + \text{SNP} + \text{SNP} \times \text{Location} + \text{Gender} \times \text{Location} + \epsilon
\]

Effects of relatedness and genetic ethnicity were accounted for using model 3:

\[
\text{Expression} = \mu + \text{Location} + \text{Gender} + \text{Relatedness} + \text{gPC1} + \text{gPC2} + \text{gPC3} + \text{gCluster} + \text{SNP} + \text{SNP} \times \text{Location} + \text{Gender} \times \text{Location} + \epsilon
\]

where \(\text{gPC1-3}\) correspond to genotypic principal component eigenvalues of axis 1, 2 and 3 (see Chapter 3 for detailed description); \(\text{gCluster}\) represents clustered ethnicity where the 194 sample are clustered into four groups corresponding largely to Agadir Arabs, Ighrem Arabs, Boutroch Amazighs and finally admixed individuals from Agadir and Ighrem; Relatedness was calculated for all individual pairs using Goddard’s method (Hayes et al. 2009):

\[
\hat{A}_{ij} = (x_{it} - 2p) \cdot \frac{x_{jt} - 2p}{2pq}
\]

where \(x_{it} = 0, 1\) or \(2\) according to whether individual \(i\) has genotype \(aa, Aa\) or \(AA\) at locus \(l\), \(p (q)\) is allele frequency of A (a), and \(2p\) is the mean of \(x_{i}^l\).
Results

Gene expression and genotypic data

High-quality gene expression data was obtained for 194 individuals using Illumina’s Infinium HumanHT12 beadchip: 105 from the city of Agadir (56 females and 49 males), 51 from the village of Ighrem (36 F and 15 M), and 38 samples from the village of Boutroch (27 F and 11 M). The chip contains 48,802 transcript probes but the analysis was restricted to 36,156 RefSeq transcripts for which comprehensive annotation information is available.

A total of 22,300 RefSeq transcripts were considered expressed beyond background levels (56.14%), comparable to a previously reported figure for RefSeq transcripts in human lymphocyte samples (Göring et al. 2006). Over 6,000 Unigene transcripts were expressed beyond background levels but are not included in the analysis detailed in this chapter. Frequency distributions for the 22,300 transcripts in all 194 individuals before and after normalization are shown in Figure 1. The similarity in the distribution of median-centered log base 2 transformed expression data for all 194 arrays shows the high quality of the data and justifies the use of median-centering standardization without normalizing variances across arrays. Some RefSeq genes in the dataset are represented by more than one probe and others by more than one transcript. In total, 16,738 unique genes are represented in the set of transcripts expressed beyond
background levels. Results are reported below either in the context of genes or probes, based on the specific analysis in question.

Initial genotype quality control was performed so that only samples with genotyping call rate > 95% and SNPs that had a call frequency above 99% were retained. Monomorphic SNPs (n = 8,430) were excluded. 52,349 SNPs had a minor allele frequency (MAF) greater than zero but ≤ 5%. In total 516,972 SNPs that had a MAF > 5%, the majority of which (88%) had a MAF > 10%. The results of the association tests results were examined so that eSNPs with a MAF < 5% or significantly departing from Hardy-Weinberg equilibrium (p < 0.001) were filtered. Further, six eSNPs overlapping with expression probe location were also removed to minimize the possibility that significant associations are not due to hybridization artifacts.

**Genetic versus expression gene-specific differentiation between regions**

In order to evaluate whether population differences in gene expression could be explained by local divergence in allele frequencies, I first looked for covariance of genotypic and transcriptional divergence. *Eigenstrat* analysis detailed in Chapter 3 revealed the presence of a genome-wide genetic ancestry axis of variation between Arabs and Amazighs. However, the same analysis also revealed the presence of five other significant axes of variation (p < 0.05, TW statistic) that seemed to be caused by local rather than genome-wide effects. A close inspection of SNP loadings for each of those five *Eigenstrat* axes corroborated this conclusion since each of axes 3 through 7 were dominated by a small number of SNPs that mapped to the same region of the genome.
To investigate the extent to which local genetic effects might impact gene expression, I contrasted transcript expression differentiation and gene-specific genetic differentiation, tested for the presence of a correlation between the two measures for all transcripts that are differentially expressed between each pair of the three regions (Agadir, Boutroch and Ighrem), and sought genes that have relatively high $Fst$ and expression differentiation. No correlation across all or the majority of genes was expected given that both Agadir and Ighrem samples are a mixture of Arab, Amazigh or admixed individuals and therefore were expected to have generally low $Fst$ estimates between locations. Instead, a simple expectation is that if the number of genes with both high $Fst$ and high expression divergence is more than expected by chance, it is likely then that expression differentiation of those genes is due to genetic factors, specifically allele frequency differences across the sample.

To achieve this, $Fst$ estimates were calculated for each 516,972 SNPs (MAF > 5%). Gene-specific $Fst$ estimates were then generated by averaging $Fst$ measures of all SNPs located within and flanking (5’ and 3’ UTR regions) the corresponding gene. These regions are known to be involved in both regulation and stability of transcript abundance (Shyu et al. 2008). First, to explore local genetic divergence at each locus, SNP-specific $Fst$ estimates for all 516,972 sites for each of the three pairwise comparisons of locations were plotted. These revealed $Fst$ values that are typically below 0.08, 0.10 and 0.12 for Agadir-Ighrem, Boutroch-Ighrem, and Agadir-Boutroch comparisons, respectively. However, $Fst$ estimates occasionally ranged between 0.12
and 0.30 depending on the comparison (Figure 2A). No fixed differences were detected between the three populations. Overall, Agadir-Ighrem differentiation is lower than Boutroch compared to either Agadir or Ighrem, in agreement with the Eigenstrat analysis results detailed in Chapter 3.

The use of $Fst$ estimates to contrast genetic and transcriptional differentiation is ideal given that both $Fst$ and differential expression analyses are conducted within an analysis of variance framework where the magnitude and significance of differentiation of each is defined by taking into account the variance within and between classes (Gibson & Weir, 2005). Gene-specific $Fst$ estimates were then calculated by averaging $Fst$ measures of SNPs located within each gene and in flanking 5’ and 3’ UTR regions (Figure 2B). On average $Fst$ measures from 23 SNPs were available for each of 22,172 genes. Averaged gene-specific $Fst$ values are smaller than marker-specific measures, but relative differentiation patterns between the three regions are similar for both measures.

To explore the relationship between genetic and expression divergence at each locus, gene-specific $Fst$ measures were plotted against expression fold change and significance. This analysis was performed for differentially expressed genes (FDR < 1%) for each location pairwise comparison and revealed no correlation between $Fst$ and expression fold change (Figure 3A) or significance (Figure 3B) as the percent variance explained is less than 0.001 with $P > 0.047$ for all six regressions shown in Figure 3. Further, there is no evidence for outlier genes that show both high $Fst$ and high expression differentiation. In fact, almost all of the top 10% most differentially
expressed genes are among the least genetically differentiated. The results of this 
analysis suggest that observed expression differences between locations are for the most 
part not attributable to gene-specific allelic frequency differences between locations. 
However, no inferences about genetic control of gene expression variation across the 
total sample can be made.

**Genome-wide association model comparisons**

Genome-wide levels of significance were used to identify the most significant 
associations. These correspond to a threshold of NLP (negative log_{10} p-value) = 8 (p = 1 
\times 10^{-8}) for *cis* associations (~ 0.05 / (22,300 probes * 200 SNPs)), and to NLP = 11.36 (p 
= 4.34 \times 10^{-12}) for *trans* associations (~ 0.05 / (22,300 probes * 516,792 SNPs)). A 
slightly relaxed Bonferroni threshold of NLP = 10.5 (~ 0.05 / (22300 probes * 75,000 
tagging SNPs)) was considered to account for LD based on a rough estimation of 75,000 
tagging SNPs. However, only two independent *trans* associations, or target transcripts, 
were uncovered by dropping the *trans* associations threshold from NLP = 11.36 to NLP = 
10.5. The significance thresholds considered might be overly conservative given the 
presence of correlation structures among expression traits and the dependence among 
SNPs in a given linkage disequilibrium block, but I use them in order to detect and focus 
on the strongest effects as is standard in the GWAS literature.

Three association models and two statistical methods were contrasted. First, a 
basic regression model (Model 1) was used and all but monomorphic SNPs were tested to 
check the effect of MAF on the results. A close examination of the top 10,000
(corresponding to NLP > 7.5 or \( p < 2.8 \times 10^{-8} \)) revealed that associations involving SNPs with a MAF less than 2% often showed inflated significance due to one or two minor allele homozygote individuals also having markedly different transcript abundance levels for the target transcript. Since most transcripts have outliers, and there are 20 times as many SNPs as transcripts, the likelihood of false positive associations of this type is high, and consequently I decided only to retain SNPs with a MAF > 5%. Further, only transcripts and SNPs involved in the associations significant at NLP > 8 \( (p < 1 \times 10^{-8}) \), corresponding to genome-wide significance for \( cis \) effects, were retained for further analysis. This first pass of the analysis provided a subset of 3,430 associations for careful examination.

Next, Model 2 was run for the SNPs and transcripts involved in the 3,430 pairs of associations. This model was used to account for the effects of SNP, location, gender, and the interactions between location and both gender and SNP. The inclusion of the interaction effect between genotype and location is of special interest to test the robustness of genotype signals detected in Model 1, and to uncover any location-specific genotypic effects. A comparison of \( p \)-values from Model 1 and Model 2 for the 3,430 pairs showed a significant drop in statistical significance for the genotype effect only for the associations involving a target transcript on the sex chromosomes (Figure 4A). In fact, approximately 70% of the significant associations in Model 1 involving a target gene located on the X chromosome dropped below the threshold of NPL = 8 suggesting
that these associations are largely affected by gender-specific differences and haplotype frequencies of the genes located on this chromosome.

Similarly, I contrasted \( p \)-values for the genotype effect generated by Model 2 to those generated by Model 3, again excluding all associations involving target genes located on sex chromosomes (Figure 4B). This comparison revealed that the vast majority of associations for the genotype effect remained significant after accounting for relatedness and genetic ethnicity. Also, I compared the results of ANOVA (which tests for differences among the three genotype classes) and regression (which is an allele trend test that requires that heterozygotes have intermediate expression levels between homozygote classes). A comparison of the results from the genotype and trend tests shows a large agreement of both methods (Figure 4C). However, the ANOVA is prone to false positives for several hundred associations, in cases where one or a few individuals with \( 0.05 < \text{MAF} < 0.1 \) happen to also have outlier transcript abundance values. Since there is otherwise a very high correlation between genotype and trend statistics, the trend test based on regression was considered more robust.

Importantly, none of the significant associations (NLP > 8) for the SNP effect are significant for the SNP-by-location effect using Model 3, and only three cases are significant from Model 2, but all involve SNPs with a MAF < 0.1. In fact, a close examination of these three associations shows their significance is due to an outlier effect which is inflating the significance for one location relative to all populations combined. Model 3 is robust to these cases as it specifically adjusts for population stratification. A
comparison of $p$-values for both the SNP and SNP-by-location effects from Models 2 and 3 and for all significant associations (NLP > 8) is shown in Figure 4C. This observation is striking and shows the consistency of genome-wide significant eSNP effects to environmental modulation.

Failure to detect SNP-by-location effects might be due to lack of power given the relatively reduced sample size of the rural locations as both Ighrem and Boutroch combined consist of 89 samples. Power analysis using JMP (SAS) shows that this study has 70% power to detect 1.3-fold difference in the mean of expression level differences between a homozygote genotype and a heterozygote genotype (assuming a sample size of 194, an average MAF of 0.25, an error standard deviation of 2, and $\alpha = 0.0001$). However, the power drops to 22% for a sample size of 100 samples (assuming 50 homozygotes and 50 heterozygotes). This analysis indicated that a sample size of 200 is sufficient to detect a 1.36-fold change of expression levels between two genotype classes at 80% power but only fold change as high as 2 for a sample size of 100. Considering an observed average effect size of approximately 0.3, this analysis indicates that this study has sufficient power to detect association signals consistent across the entire sample but has reduced power to detect signals of average and lower size effect.

**General patterns of SNP-expression level associations**

As described above, Model 3 was used to analyze transcript abundance as a function of genotype, gender, location, ethnicity, relatedness, the interaction of genotype and location, and the interaction of gender and location. Using this model, 1,744 SNP-
expression associations were uncovered at genome-wide significance. These correspond to the associations involving autosomal target transcripts. Associations involving target transcripts located on sex chromosomes were excluded given the potential artifactual association signals due to haplotype frequencies of the genes located on the X chromosome and also because gender has a minor effect on the transcriptome variation as shown in Chapters 2 and 3. The 1,744 genome-wide significant autosomal associations are summarized in the Manhattan plot in Figure 5. These associations correspond to 1,427 unique eSNPs and 396 unique autosomal probes (382 unique genes) and combine both cis and trans associations.

The 1,744 genome-wide significant associations provided a reduced set of associations, eSNPs and target transcripts to be further examined. The magnitude of the effects for the 1,744 associations ranged between 16% and 66% of explained expression trait variance with 25% explaining more than 30% of the variance (Figure 6A). The 1,427 unique eSNPs have an average heterozygocity of 39% and an average MAF of 31% (Figure 6B). MAF distribution within the three locations (Agadir, Ighrem and Boutroch) is similar with a mean ranging between 27 and 28%. An examination of the location and categories of the 1,427 eSNPs showed that 683 eSNPs (47.8%) are located in 3’ and 5’ UTR regions, 650 (45.6%) are intronic, and 94 (6.6%) are exonic, 62 of which are non synonymous (Figure 6C) warranting an investigation of their effect on protein function and probably an involvement if a feed-back regulatory mechanism. Counts, distribution and percentages of the 1,427 unique eSNP and 382 unique genes by
chromosome relative to the number of SNPs and genes initially tested for association are shown in Figure 7. Chromosomes 5, 6, 12, 17 and 19 clearly stand out as being enriched with eSNPs. The 1,744 eSNPs correspond to 0.33% of all tested SNPs while 2.28% of genes expressed in leukocytes are targets for at least one eSNP at genome-wide significance.

**Local versus distant eSNPs**

Local effects are likely to reflect the activity of regulatory sequences such as enhancers, promoters, intronic and exonic regulatory sites, and splice sites (Veyrieras et al. 2008). Local eSNPs correspond to SNPs that cause local regulatory effects and are located within the gene or in 5’ and 3’ flanking regions. Distant or *trans* effects on the other hand involve genetic variants that associate with expression level variation of a transcript through an intermediary protein or RNA that in turn regulates the expression of that transcript (Wray et al. 2003). There is no consensus as to how distant an eSNP should be to be considered *trans* effects, but investigators typically use cutoffs between 100Kb and 5Mb from the midpoint, start or end of genes or expression probe. A more strict definition limits *trans* effects to associations involving SNPs located on different chromosomes than the target gene whose transcript abundance they are associated with. The process of defining the size of a *cis*-window is for the most part subjective and is prone to confounding *trans*-signals as previously shown (Ronald et al. 2005, reviewed in Rockman & Kruglyak 2006; Williams et al. 2007). In this chapter, significant association is classified into one of the following three categories:
(i) **trans**: The eSNP and the gene transcript are located on different chromosomes.

(ii) **cis Distant**: The eSNP and the target gene are located on the same chromosome but the eSNP is located at least 500Kb farther away from the expression probe and is closer to a different gene than the target gene.

(iii) **cis Local**: The eSNP and the target gene are located on the same chromosome and the eSNP is located within 500kb from the expression probe of the target transcript.

To explicitly address the strongest regulatory effects I focused on the 1,636 (out of 1,744) genome-wide significant associations involving autosomal target transcripts and for which reliable positional information was available. An examination of these significant associations shows that only 67 (4%) associations are in **trans** involving eSNPs and transcripts located on different chromosomes. Breaking down the rest of associations shows that 1349 (86%) are **cis** local and 220 (10%) are **cis** distant. The **cis** versus **trans** pattern is visible in the plot contrasting chromosomal coordinates of eSNPs and expression probes (Figure 8) as the majority of associations lay on the diagonal indicating a proximate position of the eSNPs and the transcript whose expression level associates to the eSNP. In summary, the genome scan for genetic regulators of transcript abundance uncovered at least 396 independent effects. This figure corresponds to a conservative scenario where linked eSNPs associated to expression levels of the same transcript tag the same regulatory signal and all at genome-wide significance. The farthest and the closest eSNPs from the expression probe of the transcript it associates to on the same chromosome are 112Mb and 3 bp, respectively. An examination of the cases
where a gene is target to multiple eSNPs in *cis* shows clearly that the eSNPs lie either within or close to the gene. This pattern is typical for *cis*-acting eSNPs located within core and proximal promoters, with distant *cis* eSNPs located within or close to enhancers and repressors usually not proximal to the gene (Wray *et al.* 2003). Veyrieras *et al.* (2008) produced a high-resolution eQTL map consistent with this observation.

**eSNPs and LD patterns**

Gene ontology and functional categories of a large proportion of transcripts expressed in leukocytes reflect their key involvement in immune system response warranting a close examination of genetic control of expression of specific classes of those transcripts. The first class of genes I examined is the major histocompatibility complex (MHC) given their clustering on chromosome 6, along with the observation that this chromosome is the most enriched in terms of the number of detected eSNP-expression associations. It is readily visible in the Manhattan plot in Figure 5C (as indicated by the red arrow) that multiple eSNP hotspots are located within or in the vicinity of the chromosomal section containing the MHC genes. In fact, 115 of 195 eSNP-expression associations involving eSNPs located on chromosome 6 involve HLA class I (9 associations, 2 genes) and HLA class II (106 associations, 7 genes). All of these 115 associations are local with no overlap between HLA class I and class II eSNPs suggesting that mechanisms of genetic control of each HLA genes are local and largely if not exclusively contained within proximate physical distance of either class. It is worth
noting that several eSNPs associate to more than one HLA class II gene suggesting that some HLA class II genes have common genetic regulatory mechanisms.

Out of 396 transcripts involved in the 1,744 genome-wide significant associations, 107 are associated to a single eSNPs while the rest associate with 2 or more eSNPs and 49 transcripts have between 10 and 28 associations. In almost all cases, the multiple SNPs are in the same linkage region: only 4 cases of a transcript being regulated by loci on two chromosomes were observed (LOC656304, MKNK1, UTS2, and NRXN2). A close examination of the association clusters reveals that eSNPs associated to the same transcript show various LD patterns. For example when multiple eSNPs associate to the same HLA class II gene they tend to be in high LD: for example, the gene HLA-DRB4 is associated with 45 eSNPs, all showing high LD (|D'| ~ 1) with at least one adjacent eSNP. The entire set of HLA-DRB4’s eSNP pairs have an average |D'| of 0.73. Performing a forward stepwise regression shows that after fitting the strongest eSNP for HLA-DRB4, only a slight improvement of effect on the model fit is observed as more eSNPs are added to the model ($r^2 = 0.64$ up from 0.57 using only rs28366298) strongly suggesting all 45 eSNPs are capturing the same causal effect(s) and are indicative of the presence of blocks of eSNP haplotypes. The forward stepwise regression analysis was extended to five other genes associated to multiple eSNPs and again little or no improvement to the model fit was observed.

Extending the same analysis to other hotspots in other chromosomes showed similar patterns. For example, a physically linked cluster of 20 trans eSNPs located on
chromosome 17 is associated with the expression level of the gene MAPK8IP1 located on chromosome 11. The 20 eSNPs are in high LD (average $|D'| = 0.85$) and are located within the vicinity of 10 genes or adjacent to them. This observation warrants further investigation of those genes in the regulation of MAPK8IP1, a gene that has been associated with Type 1 Diabetes (Waeber et al. 2000).

**trans associations: the example of gamma-globin genes**

Two of the 16 distant target genes are the gamma-globin genes ($HBG1$ or hemoglobin A and $HBG2$ or hemoglobin G) that are located on chromosome 11 and normally expressed in the fetal spleen, bone marrow and liver. Gene products of both genes along with the products of two alpha globin genes located on chromosome 16 constitute fetal hemoglobin (HbF). Normally the production of gamma-globin genes declines after birth in concert with a rise in beta globin synthesis. The production of gamma globin genes at adulthood is associated with blood disorders such as beta-thalassemia and related conditions. Expression levels of both $HBG1$ and $HBG2$ are highly correlated ($r^2 = 0.99$, Figure 9A) and both genes are associated in trans to the same distant eSNP (rs766432) located on the second intron of the gene $BCL11A$ on chromosome 2 that explains the same percent variation for each gene (21.6 and 21.8%, Figure 9B). Examination of the expression probes of both genes show that are different in one bp position very close to the end of the probes, indicating that they are almost certain to cross-hybridize to both transcripts, and hence to report their combined transcript abundance. Balanced expression of both genes is expected given that
equimolar quantities of both corresponding globin proteins need to be produced and to
precisely match expression of the two alpha globin genes to make HbF.

In fact a literature search revealed that a GWAS identified \textit{BCL11A} as being
associated with F cells (the fraction of erythrocytes containing measurable amounts of
HbF) and reported rs766432 as one of the most significant variants associated to the trait
\cite{Menzel2007}. More recently, Sankaran \textit{et al.} \cite{Sankaran2008} investigated the role
\textit{BCL11A} plays in regulation of the expression of gamma-globin genes using genetic,
biochemical and developmental approaches. The authors genotyped a SNP (rs4671393)
and found it to be associated with transcript levels of two isoforms of \textit{BCL11A} in human
lymphoblastoid cell lines from the HapMap CEU and YRI populations with the high-HbF
genotype being associated with low expression of those variants, suggesting a silencing
role of \textit{BCL11A} in regulation of the gamma-globin genes.

The SNP rs4671393 was not represented in my dataset but it is likely to be in
perfect LD with the eSNP rs766432 as they are 981bp apart and the former is located
between rs766432 and another SNP (rs 6706648, located 2070bp away) that are in near-
perfect LD ($\mid D \mid = 0.98$). Three \textit{BCL11A} probes are detected in my samples, two at low
levels and one at a higher level across all samples (> 3 fold change) independent of
genotype (Figure 9C). There is strong evidence that the eSNP rs766432 is altering
expression levels of gamma-globin genes, however the observation that \textit{BCL11A}
expression levels are not altered by genotype along with the high expression of only one
isoform suggests that the mechanism through which \textit{BCL11A} alters gamma-globin genes
is independent of \textit{BCL11A} expression level variation and is more likely mediated by a genotype-dependent functional alteration of \textit{BCL11A} protein isoform(s). Sankaran \textit{et al.} (2008) reported that they were surprised by the observations that the embryonic erythroleukemia cell line K562 expressed no or very little of the two isoforms expressed in lymphoblastoid cell lines but rather expressed shorter variants of the \textit{BCL11A} proteins and follow-up work indicated that fetal liver and primitive erythroblasts expressed predominantly shorter \textit{BCL11A} variants. Based on these findings the authors concluded that \textit{BCL11A} is also developmentally regulated, such that full-length isoforms are expressed almost exclusively in adult-stage erythroblasts under strong genetic regulatory control, while shorter isoforms are present in primitive or fetal cells.

The results of my study indicate however that there is no evidence for genetic regulatory control of \textit{BCL11A} expression levels in adult-stage leukocyte samples and argue instead for a probable role of a 3’ UTR sequence variant in a post-transcriptional mode of regulation of the expressed isoform(s) that is yet to be elucidated. Post-transcriptional mechanisms of surveillance of gene products involving 3’ UTRs are known to exist and have been reported to influence the stability and/or translation of specific mRNAs (Shyu \textit{et al.} 2008). The most abundantly expressed isoform translates to an 835 amino acid protein that is among the largest of the \textit{BCL11A} proteins. This observation indicates that the scenario of the expression of variants translated to shorter proteins, similar to the case of embryonic erythroleukemia cell line K562 in the Sankaran \textit{et al.} study (2008), is unlikely in my samples. These considerations indicate how
valuable information can be obtained from genome scans for genetic regulatory variants and that can guide the discovery process of mechanisms of gene expression regulation.

**Trait and biomedical GWAS versus gene expression GWAS**

An alternative approach to data mining of the results of eSNP surveys is to examine eSNPs and target transcripts that have also previously been identified by GWAS as being associated to diseases/traits (Schadt et al. 2008)). GWAS have reported hundreds of SNP associations to a wide range of phenotypes from complex diseases to normal variation for phenotypic and biochemical traits. In order to examine possible association of GWAS SNPs to local or distant regulation of gene expression I extracted a list of 1628 independent SNP-trait associations ($p$-values $> 10^{-5}$) from the “Catalog of Published Genome-Wide Association Studies” maintained by the National Human Genome Research Institute (NHGRI). The 1628 associations were reported in the literature between the dates of March 10, 2005 and April 15, 2009 and each of them correspond to only one SNP within a gene or region of high linkage disequilibrium unless there was evidence of independent association. Contrasting this list to the list of unique local eSNPs ($NLP > 8$, $p < 10^{-8}$) in my study revealed that 10 of these have previously been associated to 10 diseases/traits ranging from LDL cholesterol to rheumatoid arthritis. All the 10 uncovered eSNPs have a local effect explaining between 16.78% and 55.3% of expression level variance of their respective gene. Some of these eSNPs were reported as being among the strongest detected associations in their respective studies. This observation strongly suggests that variation of the traits in questions might be
initiated at the transcriptional level by an alteration of transcript abundance that is readily detectable using the genome-wide survey of gene expression variation presented here and that required a sample size of only 194 individuals relative to thousands usually used in well-powered GWAS.

Five of the 10 GWAS involve a disease condition (rheumatoid arthritis, celiac disease, T1D, ulcerative colitis and systemic lupus erythematosus) while the other five involve an endophenotype (PAFAH1B2 protein level, triglycerides, Soluble ICAM-1, LDL cholesterol and hip bone mineral density). Four of the five endophenotypes were measured in serum and/or plasma that are derived from blood samples that include leukocytes, the transcriptome of which is the subject of this study. The cases of PAFAH1B2 protein level and soluble ICAM-1 are certainly interesting. Reporting their association to rs7112513 and rs281437, respectively, was novel and has not been replicated to date (Melzer et al. 2008 and Paré et al. 2008). For example, soluble ICAM-1 is found in plasma and consists of the extracellular domains of ICAM-1. The association of this protein to rs281437 located on the 3’ UTR region of the ICAM-1 gene at a genome-wide significance level of 3.2 x 10^{-10} was reported in a cohort of 6,578 apparently healthy women (Paré et al. 2008). The association of the same SNP to transcript abundance of ICAM-1 (p = 1.23 x 10^{-10}) is readily detectable in my survey using only 194 samples, further demonstrating the power of joint expression-whole genome genotyping surveys that require smaller sample sizes. In light of this result it highly plausible that transcript abundance and serum levels of soluble ICAM-1 are
correlated. Furthermore, it is probable that transcript abundance of ICAM-1 is exclusively under genetic control, as I detected no location effect on the association \( p = 0.908 \). I reached the same conclusions for PAFAH1B2 transcript abundance and the SNP rs7112513, thus validating and extending prior results from recently published studies.

**Discussion**

Contrasting gene-specific genetic and expression differentiation measures suggests that observed expression differences between locations are for the most part not attributable to gene-specific allele frequency differences between locations. This observation is in line with the results of the analysis detailed in Chapter 3 and further confirms that genetic factors are not the sole, nor the major source of variation affecting leukocytes gene expression profiles in southern Morocco. Testing local genetic effects on gene expression was extended by conducting finer analyses testing the effect of each common variant on expression variation while accounting for other effects that potentially modulate gene expression variation.

The results of genome-wide association of leukocyte gene expression levels in 194 individuals were presented in this chapter. To my knowledge, this study is the first to map regulatory variants of gene expression using a full model that accounts for gender, environmental, genetic ethnicity, relatedness, and interaction effects. In total, 1,744 genome-wide significant \textit{cis} and \textit{trans} associations were revealed. The vast majority of
theses associations are robust to both population structure and environmental effects. Many of the uncovered eSNPs and target genes have previously been reported in other studies but my study also revealed new and robust regulatory signals in leukocyte samples. It is certain that hundreds of true regulatory effects are missed at the considered stringent genome-wide level of statistical significance. Nevertheless, a minimum of approximately 400 independent genetic regulatory effects are detected. This number of associations is largely in line with the most powered published expression-GWAS studies.

Three of several observations stand out in my study. First, genome-wide significant $trans$ and $cis$ genetic regulatory effects are readily detectable in a sample of about 200 individuals. This conclusion is supported by power calculations showing that this study has 80% power to detect a 1.2-fold change ($p = 0.0001$) in expression levels means between two genotype classes. However, the power drops to only 22% for a sample size of 100 indicating reduced power to detect Location-specific effects. This observation might also partly explain the low cross-study replication figures reported (Peirce et al. 2006, Gilad et al. 2008) warranting the need for comparisons of similarly-powered datasets to make valid conclusions. The detectability of genetic regulatory effects in my sample suggests that most of the regulatory signals of average and large effects, and that are consistent across locations, are detected at this sample size. In fact, relaxing the statistical significance threshold increases the number of eSNPs but has less of an effect on the number of new target transcripts detected. Focusing on genome-wide
significant eSNPs-target transcript pairs is justified because it certainly is more interesting to identify new target transcripts than eSNPs of smaller size effect if eSNPs in the same LD block but with a larger effect have already been identified. Although almost all of the genome-wide significant associations involve one eSNP (or LD block of eSNPs) to a single target, there are 4 cases where a target gene is associated to both linked and unlinked eSNPs, indicating the presence of shared regulatory mechanisms. This observation warrants an investigation of the extent and nature of epistatic and pleiotropic effects in our dataset.

Second, my data suggests the robustness of genome-wide significant associations (NLP > 8 and NLP > 10.5 for cis and trans effects, respectively) to environment modulation. This observation is undoubtedly promising for potential use of eSNPs in epidemiological and clinical genetic applications. It is worth emphasizing that the effect of the environment on the associations was only tested for genome-wide significant associations for the genotype effect. The possibility that significant transcriptional genotype-by-environment interactions that are not significant for the genotype effect exist cannot be excluded, though statistical genetic theory suggest that this is generally unlikely. It is also possible that strong genotype-by-environment interactions are tractable only for modest effects of significance below genome-wide thresholds but this possibility is not addressed in this thesis.

Third, follow up examination of the results of genome-wide associations of transcript abundance is promising and can complement GWAS of biomedical and normal
traits and also guide the interpretation of their findings (Schadt et al. 2008). Two observations bare emphasis with regard to this issue. First, eSNPs that are both genome-wide significant and explain a large proportion of transcript variance are readily detectable in a sample size of approximately 200 individuals, in contrast to the much larger sample sizes required to detect robust signals and yet of modest to moderate effects in GWAS of other phenotypes (NCI-NHGRI Working Group on Replication in Association Studies, 2007). For example, I identified 10 eSNPs that have also been identified by GWAS associated to disease conditions or traits other than expression levels, some of which were reported as being among the strongest detected associations. My data most likely has the power to uncover more than just 10 eSNPs/GWAS variants given that only genome-wide eSNPs were considered for comparison and no eSNPs in LD with GWAS variants were included. Intriguingly, the 10 eSNP/GWAS variants explain fractions of variance on average an order of magnitude larger for expression levels than for disease conditions or serum levels, further speaking to the power of expression-genotype association studies (Chen et al. 2008, Emilsson et al. 2008, Nica & Dermiyzikis 2008, Yang et al. 2009). One possible explanation is the fact that transcript abundance is closely linked to genetic regulatory variants and therefore a large fraction of genetic effects would be most tractable at this level using relatively modest sample sizes. On the other hand the detection of genetic influences becomes more challenging the farther one departs from genotype. Second, eSNP surveys can provide valuable insights into the mechanisms through which variants uncovered by GWAS act. This is
particularly promising for traits influenced by alterations of transcript abundance and alternative splicing of genetic origin.

In relation to the issue of GWAS and eSNPs, it is intriguing that several GWAS variants associated to metabolic disorders are detected as eSNPs in the context of the leukocyte transcriptome. This might indicate either a tissue-independent effect of those eSNPs or most likely reflects the interplay between immune function and metabolic regulation as previously suggested (Hotamisligil 2006). This observation is certainly interesting, as my samples were collected from apparently healthy individuals and this observation might be pinpointing a genetic basis of disease susceptibility that is tractable using transcriptional markers in peripheral blood.

Genetical genomics surveys have a great potential to contribute to the process of uncovering the genetic basis of phenotypic variation. Integration of gene expression data itself in the process has several advantages. One is that the number of transcript abundance traits is small compared to the hundreds of thousands of SNPs usually needed in GWAS, and therefore severe multiple testing problems are not inherited particularly as researchers turn to integrating network analysis and other methods to reduce the dimensionality of expression data. In fact, a recent trend is the combination of co-expression patterns with genotypic data in human phenotype mapping studies (Chen et al. 2008, Emilsson et al. 2008, Farber et al. 2009). The question however is how much is missed by focusing on co-expression patterns and how often two genes involved in the same process are not co-expressed. The case of BCL11A transcripts and gamma-globin
genes would certainly be missed using methods based on co-expression patterns, as neither of the \textit{HBG} transcripts is correlated with the candidate regulatory transcript. These methods are complementary to each other and should be used in parallel to depict gene expression regulatory mechanisms. Another advantage is that the gene expression level of a given gene largely sums the effects of various genetic and epigenetic regulatory factors, whether they are common polymorphisms, low frequency alleles, structural variants or chromatin modifications. It is therefore likely to be revealing to contrast transcriptional profiles to traits or disease states and incorporate eSNP data to bridge the two (Nica & Dermitzakis 2008, Chen \textit{et al.} 2008, Emilsson et al. 2008, Farber \textit{et al.} 2009).

One of the challenges for genetical genomics studies is to infer likely biological consequences of eSNP-expression level associations (Schadt \textit{et al.} 2005, Zhu \textit{et al.} 2007). This problem is an extension of a larger debate over what transcriptional changes are neutral and which translate to functionally relevant phenotypes. Divergent views are emerging on this issue (Hoakstra & Coyne 2007). Some believe that most of the functionally relevant mutations occur in coding sequences and affect gene products functions instead of transcript abundance. Others consider that gene expression regulation is the most important component in the genetic basis of higher-level phenotypic variation. Both regulatory and coding variants are certainly relevant but questions remain as to which one is more important in the context of the condition being considered. Documenting eSNP variants and target genes can undoubtedly help tease
apart these scenarios for certain phenotypes. It will be interesting to see how often
eSNPs and their target genes coincide with genetic variants and genes discovered by
GWAS studies, as more and more of these will be discovered.

The example of fetal hemoglobin and gamma-globin genes is a clear case of a
heritable phenotype regulated by a regulatory mechanism in *trans* that is insensitive to
environmental perturbation. Serum levels of fetal hemoglobin were not measured in this
study but published work has shown that five common variants explain about the same
percent variance of serum levels of fetal hemoglobin (Lettre *et al.* 2008 and Uda *et al.*
2008) as the percent variance of transcript abundance of the two gamma-globin genes
encoding the two HbF subunits proteins explained by one eSNP identified in my survey.

It is important to stress the fact that the identified eSNPs are not necessarily of
causal effect and might just be tagging a nearby causal genetic variant that is truly
involved in gene expression regulation of the gene in question. Follow-up work is
therefore necessary to identify causative genetic regulatory variants (Schadt *et al.* 2005,
Zhu *et al.* 2007). *trans*-acting eSNPs and by extension the genes they are located in are
certainly the most interesting cases for validation and follow-up work. Two of the *trans-
acting events* I detected (gamma-globin genes and RPS26) have in fact been under
scrutiny in two recent studies (Sarakan *et al.* 2008, Schadt *et al.* 2008) and revealed
valuable information.
In summary, the results presented here clearly show that there is a heritable component of gene expression levels, but strengthen the conclusion that expression differences between locations in southern Morocco are largely attributable to environmental factors. I documented a list of genome-wide significant genetic regulatory signals in leukocyte samples and demonstrated their robustness to population structure and environmental perturbation. Also, I confirmed numerous cases of SNPs reported in other gene expression GWAS association studies, several of which are also associated with disease conditions or traits. This work shows how the human transcriptome can be shaped under the effects of both the environment and genetic regulatory variants, and also the promise of integrating gene expression in disease mapping studies.
References


Figure 1. Gene expression data distribution. Data distribution of log2 transformed data before (A) and after median-centering standardization (B). The upper graph shows overlayed kernel density measures of 22,300 transcripts for all 194 arrays. The lower graph shows data distribution for each array as a column, array mean (green dash), and three standard deviations of the data (red rectangle).
Figure 2. Genome-wide SNP (A) and gene-specific (B) $Fst$ measures. $Fst$ were calculated for each of 516,792 segregating SNPs (MAF > 5%) for each population pairwise comparison (B, Boutroch; I, Ighrem and A, Agadir). Each point represents a SNP in (A) and a gene in (B). Gene-specific $Fst$ values were calculated by averaging $Fst$ values for all SNPs located within the gene and in flanking 3’ and 5’ regions. Horizontal reference lines in (A) correspond to $Fst$ of 0.1 (black) and 0.15 (blue). Colored blocks correspond to the 23 chromosome.
Figure 3. **Genome-wide genetic and gene expression differentiation.** Average $Fst$ measures plotted against absolute values of Log2 fold change (A) and negative Log10 probability resulting from differential expression analyses (B), for each of the differentially expressed genes and for each population pairwise comparison (Red, Boutroch vs. Ighrem; Green, Agadir vs. Boutroch; Blue, Agadir vs. Ighrem). Each point represents a gene. Gene-specific $Fst$ values were calculated for each gene by averaging $Fst$ values for all 516,792 segregating SNPs (MAF > 5%) located within the gene and in flanking 3’ and 5’ regions. The percent variance explained is less than 0.001 with $p > 0.047$ for all six regressions represented by these plots.
Figure 4: **Association model and effect comparisons.** Effect of fitting Model 2 on significance of the Genotype effect (A). Associations involving target genes located on the X chromosome (red circles) are shown only in (A.1). (B.1) shows the effect of fitting Model 3 on significance of the Genotype effect, and the plot in (B.2) contrast ANOVA and regression test NLP from Model 2. The plots (C.1) and (C.2) show the effect of fitting the term Genotype*Location on the significance of the Genotype effect for Models 2 and 3, respectively. NLP > 32 generated by Model 2 are not shown and are all assigned the value 32, as shown by the plateau in the plots.
Figure 5. Manhattan plots for genome-wide significant associations. Each circle represents one eSNP-target transcript association. The horizontal red and black lines are Bonferroni thresholds for trans and cis associations, respectively. The 1,744 genome-wide associations are shown in (A): in the upper plot, trans associations are colored in red, cis distant in blue, and cis local in black. In the lower plots, circles are colored by eSNP chromosomes from 1 to 22. The 195 associations involving target transcripts on chromosome 6 are shown in (B). The position of the MHC region on chromosome 6 is shown by the red arrow.
**Figure 6. Properties of 1,744 genome-wide significant associations.** Distribution of percent variance explained by 1,744 genome-wide significant associations (A) and of minor allele frequencies (MAF) of 1,427 unique eSNPs involved (B). The magnitude of the eSNP effects ranged between 16% and 66% of explained expression trait variance with a mean of 26% and a 75% percentile of 30%. Average MAF equals 31%. Genotype classes (red, major allele homozygote; blue, minor allele homozygote, and grey, heterozygote) across 194 samples for the 1,427 eSNPs are shown in the cell plot (C) where each column represents one eSNP (sorted by minor allele frequency) and each column one individuals. eSNP categories and locations are shown in (D).
**Figure 7. eSNP and target genes distribution.** Distribution of 1,427 unique eSNPs (A.1) and 382 unique target autosomal genes (B.1) involved in 1,744 genome-wide significant associations. The counts are contrasted to the distribution of 516,792 SNPs and 15,067 genes tested for association (A.2 and B.2, respectively). Note the enrichment of certain chromosomes for eSNPs and target genes.
**Figure 8.** *cis versus trans effects.* Contrasting chromosomal coordinates of eSNPs and expression probes of target autosomal transcripts involved in genome-wide local (circles, colored by chromosome) and distant (red squares) associations. The majority of associations are local laying on the diagonal indicating a proximate position of the eSNPs and the transcript whose expression level associates to the eSNP.
Figure 9. BCL11A \textit{trans} regulation of gamma globin genes. eSNP rs766432 located on 3’ UTR region of BCL11A and its targets in \textit{trans} the gamma globin genes HBG1 and HBG2. Co-expression ($r^2 = 0.99$) of HBG1 and HBG2 across all samples is shown in (A). Regression of eSNP rs766432 genotype classes against HBG1 and HBG2 expression levels (B). Expression patterns of HBG1 and HBG2, and four probes of three BCL11A isoforms (C). Only one isoform (colored in purple) encoding a 835 aa protein is expressed at high levels but its expression is not dependent on genotype.
CHAPTER FIVE

Conclusions and Possible Future Directions
Conclusions

In humans, study of how the environment and genome interact to shape phenotypic variation is particularly relevant to understanding the origins of complex diseases and the increase in their prevalence coinciding with major shifts from traditional to urbanized lifestyles. Efforts to tackle this challenge have been extended to various layers of biological organization and to the genome-wide level as major advances in high-throughput technologies and genomic methods have emerged in the last decade. Motivated by a desire to understand the impact of lifestyle transitions on physiological variation in my own country, I have used a combination of genomic techniques and quantitative and population genetic approaches to survey patterns of genetic and gene expression variation in southern Morocco and attempted to extend knowledge about the environmental and genetic basis of gene expression variation in general.

When I started this research it was just beginning to be demonstrated that in humans, as well as in all the other organisms, gene expression is highly heritable (Dixon et al. 2007, Göring et al. 2007, Stranger et al. 2007) but little was known about the modulation of gene expression variation by environmental factors (Gibson & Weir 2005, Gibson 2006). To tackle this issue I used three approaches as detailed in Chapters 2, 3, and 4. Here I will highlight several original aspects of this work, synthesize the findings and discuss some of the challenges encountered and possible future directions.
Human functional genomics studies and sampling in field settings

Obtaining samples from remote populations is a difficult challenge for human geneticists. Setting up efficient recruitment and sampling strategies in Morocco, particularly for the rural and nomadic populations, was a major effort that is only touched on in this thesis. It was equally challenging to ensure that the samples are suitable for functional genomics experiments and could be shipped to Raleigh appropriately. Peripheral blood, specifically the leukocyte fraction, was chosen as the tissue for my study since it is a readily accessible component of the immune system that is also ideal as a model to study gene-by-environment interactions given that the immune system develops and functions while incorporating information from the surrounding environment.

The use of the Leukolock® filters (Ambion, Austin TX) in a non-clinical context was a novel contribution of my research. Previous studies that used peripheral blood either stored whole blood samples in RNALater® solution, or used protocols (notably, PaxGene tubes) that require a centrifugation step to fractionate components of blood prior to storage. The Leukolock system offers several advantages over these standard methods as it incorporates depletion filter technology to isolate leukocytes and eliminate plasma, platelets, and the vast majority of red blood cells within minutes after blood collection. The method therefore allows for rapid isolation of leukocytes without the need for centrifugation, while also removing the highly variable globin-rich cellular fraction that greatly increases noise in immunoprofiling experiments. Most importantly, my protocol
ensured that neither the *in vivo* state of the leukocyte transcriptome nor the integrity of RNA molecules were altered by sample manipulation and storage conditions. This was apparent as both RNA and gene expression microarray data of high quality were obtained. My study was the first application of the Leukolock system in a study involving sample collection from remote locations, and therefore demonstrated the usefulness of the filters for studies in epidemiological genetics. Adoption of the Leukolock protocol should definitely help extend functional genetics research to understudied populations in remote areas, particularly in non-developed countries and on isolated islands.

*Effect of environmental geography on leukocytes of Moroccan Amazigh*

Since little was known about the impact that environmental factors have on gene expression in humans, I first set out to estimate the magnitude of the impact environmental geography has on the transcriptome of three Moroccan Amazigh communities each of 12-18 individuals living in three regions with distinct environmental geography characteristics. The term “environmental geography” is usually used in social sciences, particularly the discipline of geography, to refer to both human and physical geography. I introduced the use of this phrase in the context of human genetics as a convenient term to capture all environmental factors affecting human biology including biotic, abiotic and cultural influences.

Several studies in humans have demonstrated that non-genetic factors like HIV infection status (Montano *et al*. 2006), exposure to arsenic (Fry *et al*. 2007), and obesity
(Emilsson et al. 2008) significantly impact gene expression profiles. However, my study was the first explicit effort to quantify the effect of environmental geography on genome-wide gene expression variation in humans outside clinical settings. Only individuals who appeared to be healthy at the time of collection and who stated that they were not suffering from a chronic condition were included in the study. Further, genetic influences on gene expression variation were minimized by sampling only individuals of presumed Amazigh ancestry.

The most striking result of this pilot study (Idaghdour et al. 2008) was the observation that up to one third of the leukocyte transcriptome is differentiated among regions (FDR < 1%). Genome-wide polymorphism analysis confirmed that genetic differentiation among regions was limited, further supporting the presence of a large contribution of regional non-genetic population differences to gene expression variation in southern Morocco and likely in other human populations. Further, network and annotation analyses implied that specific aspects of immune function are strongly affected by regional factors and may influence susceptibility to respiratory and inflammatory disease.

The results of this study firmly rejected the null hypothesis that genetically similar populations show similar expression profiles. Further, I rejected the hypothesis that genetic factors make a greater contribution to gene expression variation than environmental factors. This was achieved by contrasting the magnitude of expression differentiation in my study to results from a study that compared lymphoblastoid cell
lines derived from individuals of Asian and European ancestry (Storey et al. 2007), as well as to those from a re-analysis of an African and European comparison also using lymphoblastoid cell lines (Akey et al. 2007). I also noted similar magnitudes of differential expression for arsenic exposure on newborn cord blood in a Thai urban and rural population (Fry et al. 2007), HIV infection status of mothers in a rural village in Botswana (Montano et al. 2007), and for BRCA1 and BRCA2 mutant tumors when compared to each other (Hedenfalk et al. 2001).

Three concerns rose with the pilot study. First, power to detect differential expression varies among studies, being heavily influenced by such technical factors as the array platform, sample size, and sample handling methods, not to mention biological factors such as tissue homogeneity, genetic diversity, and individual environmental or cultural differences. Although every attempt was made to make valid comparisons, this problem emphasized for me the need to estimate the genetic and the environmental contributions to gene expression variation jointly in the same sample. Second, the sample size was modest, even though I had sufficient power to see a substantial effect of environmental geography, as supported by the observation that the estimates of the fraction of differentially expressed genes among regions were not strongly affected by partial reduction of the dataset. This said, it is plausible that expression heterogeneity would significantly increase if sample size is increased and increased variability might swamp the effect observed with the relatively small dataset or at least reduce its size. Third, the issue of replication arose and questions regarding the possibility that such
effects might not be observed if different individuals or locations were sampled in a different season of the year remained unanswered. Follow-up work was therefore essential, partly to validate the results obtained from the pilot study by addressing the specific points mentioned above, and partly to test further hypotheses concerning the environmental and genetic basis of gene expression variation.

Subsequently and following the same methodology as for the pilot study (Idaghdour et al. 2008), I sampled over 200 urban and rural individuals in summer of 2008 (as opposed to winter of 2006) and from similar locations as two of those sampled in the pilot study. An exception is that the urban site of Anza that was sampled in both campaigns was represented by different individuals in the follow-up study and complemented by a second community on the opposite side of Agadir. Each lifestyle was represented by approximately 100 samples from two villages and two urban sites with both men and women in the sample. Further, both Arab and Amazigh ethnicities were represented in both lifestyles. Genome-wide gene expression profiling and genotyping were performed and quantitative genetic analysis of gene expression variation was performed to estimate the genetic and environmental contributions jointly.

*Genome-wide estimates of ethnicity*

One caveat with sampling understudied human population is that often genetic and cultural ethnicities are confused, leading to inadequate classification of individuals when self-reported information solely is used. The choice of the study sites was based on my *a priori* information of southern Morocco while family and ethnicity information
relied on self-reported information collected at the time of collection. However, genome-wide genetic analysis of the entire sample is a much more comprehensive way to approximate the true picture of population structure in the sample. In fact, this analysis revealed specific patterns of variation that confirmed some of what was known but that also unveiled some unexpected cryptic patterns. Two examples bear mentioning here:

First, relatedness measures calculated using the identity by state (IBS) estimator in PLINK (Purcell et al. 2007) and the Goddard method (Hayes et al. 2009) confirmed known family relationships in the sample, but also revealed a small number of cryptic groups of related individuals unknown at the time of sample collection. Second, Eigenstrat analysis (Price et al. 2006) revealed that the village of Ighrem is genetically intermediate between Agadir Arabs and Boutroch Amazigh, contrary to my a priori expectation that Ighrem Arabs would cluster with Agadir Arabs. The data also does suggest that admixture is occurring in a fraction of the sample, notably in Agadir and probably in the village of Ighrem but it is not the only possibility.

In fact, the presence of a small fraction of Amazighs individuals in the village of Ighrem suggests that a low level of gene flow has been occurring in the village probably for hundreds of years and might explain the shift of Ighrem towards Boutroch relative to Agadir along the main Arab-Amazigh genotypic axis of variation. Another plausible scenario is the divergence of both Arab and Amazighs groups in southern Morocco from a common ancestral population with traditionally Amazigh villages, such as Boutroch, being more differentiated given their isolation, probably for thousands of years. Villages
like Ighrem also might have been isolated for a period of time and started to differentiate along a different genotypic axis of variation. Generally, Arab and Amazigh ethnicities seem to have genetic affinities that may reflect shared past history as also suggested by previous studies (Bosch et al. 2000, Kéfi et al. 2005). These observations reflect population history and the complex nature of population structure in most human populations. Systematic genome-wide analysis of population structure of southern Moroccans with the inclusion of control populations is required to tease apart the different scenarios but it is beyond the scope of the work presented in this thesis and remains to be carried out. Gratifyingly, self-reported ethnicity of Arab and Amazigh individuals agreed in general with genetic clusters revealed by Eigenstart analysis and proved useful in guiding the interpretation of the results.

Eigenvectors of genotypic variance were conveniently used for the purpose of the analysis detailed in Chapter 3. Contrary to self-reported ethnicity, principal component estimates of genotypic variance act as a genome-wide continuous quantitative measure that reflects the proportion of each ancestry and therefore are more suitable to represent the continuous gradient of genetic variation ranging from non-admixed Arabs to non-admixed Amazigh and including admixed individuals at various degrees. Although seven significant axes of genotypic variation were detected by Eigenstrat analysis, only one axis reflects genome-wide differentiation of Arabs from Amazighs independent of local genotypic effects. The other axes were each strongly influenced by a small number of
SNPs located in the same chromosomal region, as described by other authors (Price et al. 2008, Need et al. 2008, Biswas et al. 2009).

**Ethnicity and environmental contributions to gene expression variation**

Various contrasts were performed in Chapter 3 using a combination of analysis of variance, principal component analysis and variance component analysis. First and foremost I have replicated the results of the first study as again I observed a striking effect of environmental geography. Compared to the figures found in the first study the magnitude of the effect of gender was comparable, and astonishingly that of location was even higher. In fact, a three-way contrast of Ighrem, Boutroch and Agadir including males and females revealed that 40% of all expressed genes are differentially expressed at the 1% FDR threshold. The location effect varied depending on the contrast but was greatest when Boutroch females were compared to either Ighrem or Agadir females. One observation that was not replicated though is a significant enrichment for respiratory disease biomarkers in the set of differentially expressed genes, possibly reflecting seasonal variation of expression profiles between summer and winter. Nevertheless, one important observation echoed by the functional enrichment analysis is the apparent specificity of differential expression patterns and the likely impact they have on disease susceptibility.

Estimating the contribution of ethnicity relative to location turned out to be a difficult task since both effects are confounded in the villages and also because differences in geography and lifestyle were observed between the villages despite the fact
that both are considered rural. Since no significant differentiation of Arabs and Amazigh in the city of Agadir was detected, whether eigenvectors of genotypic variance or self-reported ethnicity were used, it can be inferred that environmental geography impacts the leukocyte transcriptome more strongly than the contributions of gender and ethnicity. Furthermore, the magnitude of the effects of gender and ethnicity were found to be comparable when sources of variance of the urban dataset were partitioned among ethnicity estimates, gender and their interaction using quantitative analysis of the variance components.

An important observation in both rural locations was the complexity of interaction effects that statistically involve gender effects. These could either be intrinsic to sexual differences between men and women, or reflect cultural practices adopted by the two genders. The evidence suggests that these interactions are more likely indicative of strong lifestyle interactions. For example, Ighrem males have expression profiles similar to Agadir male residents as only eight genes differentiated (1% FDR) the two groups. This difference between Arab men and women in Ighrem is almost certainly a lifestyle effect rather than a biologically determined gender difference, first because most of the Ighrem males commute to the city and second because in both studies the overall gender effect is minor within locations. On the other hand, a dozen Ighrem Amazigh females are not more similar to females from the Amazigh village of Boutroch indicating an interaction effect of ethnicity and environmental geography.
**Genome-wide mapping of genetic regulators of the leukocyte transcriptome**

Dissection of the genetic and environmental basis of gene expression variation was extended in Chapter 4 by conducting a genome scan for genome-wide significant eSNP effects and testing their robustness to population structure and differences in environmental geography. The SNP-expression level association tests were computed in a collaboration with Wendy Czika and Russ Wolfinger using newly developed JMP software at the SAS Institute in Cary, NC. An initial model was run for each of 516,972 genotypes against each of 22,300 expressed transcripts. Further models were run on the subset of SNPs and transcripts that were implicated in associations with NLP > 7.5 to account for gender, location, genetic ethnicity, relatedness, and various pairwise interaction effects. The analysis revealed 1,744 genome-wide significant associations involving 396 unique autosomal genes and 1,427 eSNPs. 96% of the associations involve eSNPs and target genes located on the same chromosome with the majority of eSNPs presumably acting in *cis* as they lie close to the actual target genes. Only 4% of the associations are in *trans* involving target genes located in different chromosomes than the eSNP they associate to. These results are largely consistent with finding reported by the handful published human genetical genomic studies (Williams *et al.* 2007, Gilad *et al.* 2008). My study however was original both in scope (sampling multiple ethnicities and localities) and with regard to some of its analytical aspects, and thus provided some novel findings.
As described in Chapter 4 we ran an initial statistical model and identified a subset of significant associations with $p < 10^{-8}$. However, it was important to adjust for gender and location effect but also for interaction effects, population structure and relatedness. Nowadays the use of principal components of genotypic variance to adjust for population stratification is a standard practice in GWAS. I extended this approach by further adjusting for genetic effects using clustered ethnicity based on Eigenstrat analysis (Agadir Arabs, Boutroch, Ighrem Arabs and Admixed individuals), and in collaboration with Hong Lee and Peter Visscher at the Queensland Institute for Medical Research, by using a relatedness kinship matrix (Hayes et al. 2009). Extension of the statistical model importantly controls for hidden genetic structure in the dataset, providing better estimates of the non-genetic influences on gene expression structure. Similar approaches have been previously used to adjust for relatedness is GWAS traits other than gene expression (maize; Yu et al. 2006, dogs Kennerly et al. 2008, humans; Lowe et al. 2009). Since relatedness was minimal in our data, the impact of the adjustment was not great, but we nevertheless strongly advocate adoption of this approach in gene expression GWAS of human populations.

One of the most pressing questions I set to answer concerns the robustness of eSNP effects to population structure and most importantly to environmental modulation. Only few studies have previously addressed gene-by-environment interactions at the transcriptional level, all in model organisms (Landry et al. 2006, Smith & Kruglyak 2008, Sambandan et al. 2008). Previous studies in humans have only addressed
interactions in the context of the tissue specificity of genetic regulatory effects. For example, Emilsson et al. (2008) and Heinzen et al. (2008) have reported that only 50% of genetic control signals of transcript abundance in blood and adipose tissue, and in brain tissue and PBMC, respectively, are detected in both tissues. Accounting for population structure and relatedness in our genome scan did not alter the genome-wide significant genotypic effects. More intriguingly, we have demonstrated for the first time the robustness of the genome-wide significant eSNP effects to environmental modulation and showed striking consistency of the magnitude of these effects across three locations. This observation is by far one of the most remarkable in my study.

I did not expect this conclusion given the observation that environmental geography influences in the three locations hugely impact the leukocyte transcriptome. I assumed that environmental triggers would modulate gene expression profiles by altering many of the genome-wide significant genetic regulatory effects. This observation implies that the regulatory mechanisms that are under strong genetic effects are independent from the mechanisms underlying differential expression in response to environmental changes. It also highlights the robustness of the regulatory mechanisms involved and that have evolved to cope with environmental perturbations.

However, this observation is in striking contrast with a recent study in Nature by Smirnov et al. (2009) that reported a case where genetic individual differences mediate response to an environmental stimulus. The authors performed a genome scan of genetic regulatory variants of radiation-responsive 3,280 genes in CEPH cell lines and identified
more than 1,200 genome-wide regulatory effect almost all *trans*-acting (target transcript 5Mb off the target gene). The authors hypothesized that the large number of *trans*-acting regulators (that are under genetic control) provides cells with multiple mechanisms to mount responses to different types of stressors as also previously reported in *C. elegans* and yeast (Li *et al.* 2006, Smith & Kruglyak 2008). The results of this study are at odds with all published human eSNP or eQTL studies in that very few *cis*-eQTL were reported and it is not clear if this pattern is biologically true of if it is due to a reason related to the experiment or the analysis.

For computational reasons I have not been able to generate genome-wide estimates for the genotype-by-location effect, so the question if these effects are present when the main genotype effect is below genome-wide significance is unsolved and remains plausible. It should be noted, though, that significant interaction effects are generally only expected to be observed in the presence of a significant main genotype effect. Further, future investigations of genotype-by-environment interactions should involve much larger sample sizes, as my study was only powered to detect interactions of unusually large effect size.

Changes in gene expression levels that have a genetic basis have been found to associate to a wide spectrum of phenotypes in humans (Kleinjan & van Heyningen 2005, Wray 2007). The results of our genome scan of those changes have important implications particularly with regard to studies addressing the basis of disease susceptibility and disease mapping in general. The robustness of the genome-wide
significant regulatory effects implies in fact that the variants implicated can predict across populations both transcript abundance levels and traits or disease phenotypes in the majority of cases. More ambitiously, transcript abundance of those target genes implicated in the robust association can potentially be used as predictors of other phenotypes such as protein levels or even organismal traits and disease conditions. This observation becomes even more promising given the relatively small sample size required to detect transcriptional associations. However, the interpretative power of expression studies for mapping complex traits will certainly depend on the underlying architecture on the network model of the trait and if it is causal, independent or reactive (Zhu et al. 2007, Schadt et al. 2005). Nevertheless, Schliekelman (2008) explicitly addressed the issue of statistical power for expression-trait associations under a variety of assumptions about dominance, number of segregating loci, and other parameters, and showed for example that if the disease risk depends directly on transcript level the power to detect association between transcript and disease is quite good with sample sizes on the order of 100 being sufficient for 80% power. However the power is generally low if the transcript is not causal (i.e. the transcript is reactive to a disease locus).

Another issue that might arise with the use of expression traits for mapping of phenotypes in human populations is population stratification at the transcriptional level. This is likely to happen when regulatory polymorphisms are segregating at different allelic frequencies in different populations. In fact, it has been suggested that differences in frequency of transcript abundance classes between case and control groups could lead
to the false inference that a particular transcript is causative in phenotype promotion (Gibson 2003). Therefore such patterns need to be accounted for when performing transcript abundance-phenotype association, to avoid spurious signals. In this study, I demonstrated that there is no correlation between allele frequency divergence ($F_{st}$) and expression differentiation, so local genetic divergence is unlikely to have had a measurable impact on the results.

In summary, this thesis is to my knowledge the first attempt to partition genetic and environmental geography contributions to gene expression variation jointly and to test the robustness of genetic regulatory effects in humans. Overall, my results have provided ample evidence that environmental geography is a major modulator of the leukocyte transcriptome and that both gender and genetic factors are relatively minor sources of variation. The leukocyte transcriptome dataset analyzed in Chapter 3 and 4 consisted of over 16,000 expressed genes. Transcript abundance of about half of those genes (~8000) is significantly altered by environmental geography, while only about 400 genes are under the influence of genome-wide significant genetic regulatory effects. The mechanisms that mediate the environmental response and variation within locations are likely independent. Both genome-wide significant cis and trans eSNPs are robust to environmental modulation but it remains possible that some genetic eSNP-by-environment interaction are more pronounced for regulatory effects below genome-wide significance. Perhaps more importantly, lifestyle, geography and ethnicity interact in a complex but statistically predominately additive manner to alter a large fraction of the
transcriptome. Further, I have shown that eSNP mapping can help dissect the genetic basis of gene expression regulation and more specifically provide valuable information to connect phenotypes to regulatory mechanisms.

**Possible future directions**

The research presented in this thesis lays the foundation for follow-up work in several directions. For example, the observation that the expression profiles of Ighrem villagers are significantly different from those of Boutroch villagers calls for a survey of expression profile in more villages and cities. Another direction would be the extension of this research to other countries and ethnic groups. Since Arab and Amazigh ethnicities seem to have genetic affinities, further insight will be gained by studying pairs of populations that are more genetically differentiated. It is ultimately interesting to test if environmental geography influences will always prevail over those effects with a genetic basis. Another interesting question would concern the variation in magnitude and specificity of both environmental and genetic regulatory effect among ethnic groups and also lifestyles. Unfortunately, extended phenotypic and endophenotypic data was not collected in my study. Several non-invasive or minimally invasive tests developed for field setting are currently available such as those for measuring levels of various molecules in saliva, urine or blood and it would definitely be useful to integrate such a component in future studies.
It is noteworthy that the data that I have generated have not been fully exploited yet. At least three aspects deserve careful examination. First, the robustness of the eSNP effects has been demonstrated only for genome-wide significant associations but the possibility of subtle eSNP-by-environment signals cannot be ruled out. This issue can be properly examined only by increasing sample size as shown in Chapter 4. Nevertheless, an examination of the presence of eSNP-by-environment signals in the associations with significance levels below genome-wide merits follow-up with the aim to identify at least a fraction of those effects, if any.

Second, integrating the recently developed MMC method (Stone & Ayroles 2009) for the construction of modules co-expressed genes with eSNP information is worth pursuing. For example, it would be interesting to identify modules of co-expressed genes that are consistent across the three locations and examine their genetic basis. I predict that integration of some of those networks that are robust to environmental perturbation will have a strong genetic basis that can be elucidated by integrating eSNP data. On the contrary genes in modules constructed from differentially expressed gene would likely not be enriched with genes under strong genetic regulatory control. Further, follow-up analysis to attempt to predict causal regulators and construct gene regulatory network using similar approaches used in yeast by Zhu et al. (2008) merits consideration. The authors generated genotype and gene expression data and integrated them with available transcription factor binding sites and protein-protein interaction data in a Bayesian approach to identify and validate several regulators of eQTL hotspots.
Third, the population genetics of eSNPs should be explored with special emphasis to detect signals of selection using the analytical framework introduced by Kudaravalli et al. (2008). The authors used a haplotype-based approach (the iHS method) and detected at least 20 cis eQTL in the Yuruba CEPH dataset as being targets of positive selection. The Moroccan population is understudied and has never been integrated in surveys of positive selection even in the most extended worldwide surveys (e.g., Pickrell et al. 2009).

Further, a preliminary examination of genetic differentiation of eSNPs in my dataset compared to the four HapMap populations showed that many in fact have high $F_{st}$ estimates probably indicative of selection. It is therefore worth following up on this population genetic analysis aspect with particular emphasis on eSNPs that have been identified by disease GWAS, those coinciding with candidate genes for selection.

Furthermore, particular attention should be dedicated to trans eSNPs partly because Kudaravalli et al. (2008) did not examine them but most importantly because they are more likely to affect gene expression of more than one gene and thus are more likely to have deleterious effects compared to cis eSNPs.

Finally, while basic research such as I have reported here is fascinating, future studies may increasingly emphasize the biomedical importance of expression divergence within and among human populations.


