

**INTEGRATIVE GENOME-WIDE ANALYSIS TO STUDY  
THE GERMLINE GENETICS  
OF MYELOPROLIFERATIVE NEOPLASMS**

by

Semanti Mukherjee

A Dissertation

Presented to the Faculty of the Louis V. Gerstner, Jr.

Graduate School of Biomedical Sciences,

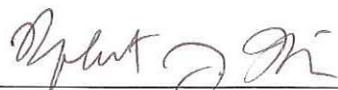
Memorial Sloan-Kettering Cancer Center

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

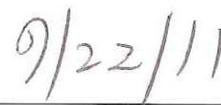
New York, NY

September, 2011



---

Robert J. Klein, PhD  
Dissertation Mentor



---

Date

Copyright by Semanti Mukherjee 2011

## DEDICATION

I would like to dedicate this thesis to my parents, Barin Kumar Mukherjee and Minakshi Mukherjee and my son Ayanava Ganguly. Without their constant support, perseverance and love, none of this work would have been possible.

## ABSTRACT

Myeloproliferative neoplasms (MPN) is a clonal disorder of hematopoietic lineage. MPN encompass three subtypes, namely polycythemia vera (PV), essential thrombocythemia (ET), and primary myelofibrosis (PMF) that are commonly associated with somatic mutation JAK2V617F. The family members of MPN patients are at high risk. There likely are additional genetic events that contribute to the pathogenesis of these phenotypically distinct disorders. To understand the etiology of the MPN phenotype and predisposition, we performed a genome-wide association (GWA) study followed by targeted sequencing using next generation sequencing technology. In a typical GWA study design, cases and controls are ideally matched for ethnicity, age, sex, socio-economic background and other environmental factors. Instead of using a matched control study design, we developed a method using principal component analysis to use controls from public databases. The optimum number of cases and controls were calculated analytically and type I error rate and power was determined by simulation. We applied this method for our MPN GWA study. A JAK2 SNP rs10974944 was significantly associated with MPN risk after correcting for residual population stratification and multiple testing. Further genetic analysis has shown that the risk allele - "G" allele (GG or CG) at rs10974944 preferentially acquires the V617F mutation. This illustrates a complex interplay between somatic and germline genetics in MPN. To dissect the functional variant(s) and to understand the haplotype-specific acquisition of somatic mutations, we carried out targeted sequencing of the 300kb haplotype block harboring *JAK2* using next generation sequencing technology (RainDance and SOLiD sequencing). We compared MPN cases that are homozygous for the risk allele (GG-MPN

cases) with the ones that are homozygous for wild type allele (CC- MPN cases). We found that there is no excess of single nucleotide variants in the *JAK2* locus in GG-MPN cases compared to CC-MPN cases using the ancestral sequence as reference. However, we further explored the existence of selection pressure at *JAK2* using HapMap phase III data and detected an excess of derived alleles at *JAK2* when compared to ancestral repeats. We further analyzed sequence specific differences between these two groups of patients and identified a candidate functional variant in the promoter region of *JAK2* gene that is predicted to bind to the transcription factor c-Fos in allele specific manner. We next analyzed the *JAK2* susceptibility haplotype in MPN (also referred as MPN risk haplotype) and reconstructed the phylogenetic tree using PHYLIP. We discovered that MPN risk haplotype forms a separate cluster from other haplotypes when using chimpanzee as out-group. The sequence similarity of MPN risk haplotype was more close to chimpanzee. Thus, we concluded that *JAK2* susceptibility haplotype in MPN is an ancestral haplotype compared to modern human population and is most compatible with the evolutionary model: ancestral susceptibility model of disease.

## BIOGRAPHICAL SKETCH

Semanti Mukherjee was born in Kolkata, India. She enjoyed her initial years in a small industrial town Korba, India, with her father Barin Kumar Mukherjee, her mother Minakshi Mukherjee and her sister Sudakshina. Her middle school, Kendriya Vidyalaya NTPC Korba had a strong influence in shaping her dreams. Her fascination about biology started in her early years and turned steadily into her passion. From her mother, a math teacher, she has imbibed the facet of learning as well as scientific reasoning; and from her father, an engineer, she has inherited the enthusiasm to work hard to succeed. She received her Bachelors and Masters in Biology from Nagpur University, India in 1997 and 1999, respectively. She was the summa-cum-laude in both the programs and received many awards of excellence. She immigrated to San Francisco, USA in 1999 with her husband Amitava Ganguly. She went to Ohlone College, California in 2000 to study computer programming. Living in different countries over the years, she was fortunate to get extensive and diverse exposure in different laboratories, namely: National Institute of Science, Singapore (2003), Indian Institute of Science Bangalore India (2004), University of California San Francisco (2005) and Stanford University California (2006-2007). While working at Stanford University, she decided to pursue the next step in her scientific career and joined the PhD program offered by Louis V. Gerstner, Jr. Graduate School of Biomedical Sciences at Memorial Sloan Kettering Cancer Center (MSKCC) in New York. During her term in MSKCC she enjoyed the wonderful experience and the research environment of the institute. And most importantly, she loved her life every day, spent together with her son Ayanava (Riki) Ganguly while finishing up her doctoral study.

## ACKNOWLEDGEMENTS

First and foremost, I offer my sincere gratitude to my mentor, Dr. Robert Klein, whose guidance and encouragement I will never forget. He has supported and motivated me throughout my thesis whilst allowing me the room to work in my own way. As a result, graduate school experience became rewarding for me.

I would like to thank the Klein lab members Xing (Dandan) Xu, Jason A. Willis and James E. Hayes for the stimulating discussions and their helpful comments on my dissertation. I thank Xiaoni Gao and Heriberto Moran for their technical assistance, and Concynella Graham-Wright for helping me with every administrative issue.

It is an honor for me to work in close collaboration with Dr. Ross Levine in the MPN project. I was delighted to interact with Levine lab members, especially Dr. Outi Kilpivaara and Alison Schram, who have contributed and extended their wet lab experience in our MPN functional genomics study.

Drs. Kenneth Offit and Christina Leslie deserve special thanks as my thesis committee members and advisors. They offered their valuable advice and suggestions whenever I needed them and helped me to envision my career beyond the graduate studies. I thank my clinical mentor Dr. Zsofia Stadler for giving me the opportunity to understand the role of medical genetics in clinics.

My deepest gratitude is due to Dr. Harold Varmus, Dr. Thomas Kelly, Dr. Kenneth Mariani and our benefactor Louis V. Gerstner, Jr for their vision to establish our graduate program. I would also like to thank our graduate school students, faculty members and administrative staff, especially Iwona Abramek for her help in formatting this dissertation.

I gratefully acknowledge the funding sources by the Geoffrey Beene Cancer Research Center at MSKCC, the Emerald Foundation, and NIH R03 CA141524 (to Dr. Klein). I am extremely grateful to all of the investigators and funding agencies responsible for the data deposited in dbGaP that made my PhD work possible.

I am indebted to my mentors at Stanford University Dr. Helen Blau and Dr. Jason Pomerantz for their encouragements to pursue my dream.

Lastly, my deepest gratitude goes to my family for their unconditional love and support throughout my life; my achievements are simply impossible without them. I would like to thank all people who have helped and inspired me in my journey.

Thank you.

Semanti Mukherjee  
*Gerstner Sloan Kettering Graduate School of Biomedical Science*  
New York  
September 2011

# TABLE OF CONTENTS

ABSTRACT.....	IV
LIST OF FIGURES .....	XIII
LIST OF TABLES.....	XV
LIST OF ABBREVIATIONS.....	XVI
<b>INTRODUCTION.....</b>	<b>1</b>
MYELOPROLIFERATIVE NEOPLASMS .....	1
JAK2 SOMATIC MUTATION .....	3
MPN PHENOTYPE PLEIOTROPY .....	5
GENOME WIDE ASSOCIATION STUDIES.....	9
LINKAGE STUDIES VERSUS GENOME WIDE ASSOCIATION STUDIES .....	11
GWAS STUDY DESIGN.....	13
GWAS CHALLENGES .....	20
EVOLUTIONARY MODEL FOR DISEASE THE SUSCEPTIBILITY LOCUS .....	24
<b>CHAPTER 1.....</b>	<b>29</b>
<b>USING ADDITION CONTROLS FROM PUBLIC DATABASES TO INCREASE POWER OF GWAS.....</b>	<b>29</b>
1.1 INTRODUCTION.....	29
1.2 SUBJECTS AND METHODS.....	32
<i>Ethics Statement.....</i>	<i>32</i>
<i>Analytical power calculation.....</i>	<i>32</i>
<i>Simulation study for empirical power and type I error rate calculations .....</i>	<i>32</i>
<i>Pancreatic cancer study samples and genotyping.....</i>	<i>34</i>
<i>Additional controls from dbGaP.....</i>	<i>35</i>

<i>Data processing and quality control</i> .....	35
<i>Principal components analysis</i> .....	37
<i>Additional quality control by control group comparisons</i> .....	37
<i>Association analysis and estimation of <math>\lambda</math></i> .....	38
<i>TaqMan genotyping assay</i> .....	38
1.3 RESULTS .....	40
<i>Analytical power</i> .....	40
<i>Power and type I error rate from simulation studies</i> .....	42
<i>Population stratification in New York based data</i> .....	44
<i>Additional quality control through comparison of control groups</i> .....	50
<i>Performance of known pancreatic cancer associated SNPs</i> .....	56
<i>Number of significant principal components</i> .....	60
1.4 DISCUSSION .....	64
<b>CHAPTER 2</b> .....	<b>69</b>
<b>GENOME WIDE ASSOCIATION STUDY OF MYELOPROLIFERATIVE NEOPLASMS</b> .....	<b>69</b>
2.1 INTRODUCTION.....	69
2.2 MATERIALS AND METHODS .....	71
<i>SNP Array Analysis of MPN Samples</i> .....	71
<i>Principal Component Analysis of MPN Patients/Controls</i> .....	71
<i>Statistical Analysis</i> .....	73
<i>Genotyping and Expression Analysis</i> .....	73
<i>JAK2 rs10974944/Mutation Clonal Analysis</i> .....	73
2.3 RESULTS .....	75

<i>Case-Control Analysis of Genome-Wide SNP Array Data Identifies JAK2 as a Major MPN Risk Allele</i> .....	75
<i>Germline Variation at the JAK2 Locus Influences MPN Predisposition</i> .....	78
<i>Germline Variation at JAK2 Specifically Predisposes to the Development of JAK2V617F-Positive MPN</i> .....	80
<i>JAK2V617F is Most Commonly Acquired in cis with JAK2 rs10974944</i> .....	82
2.4 DISCUSSION .....	85
<b>CHAPTER 3</b> .....	<b>88</b>
<b>MECHANISM FOR JAK2 SUSCEPTIBILITY HAPLOTYPE IN MPN</b> .....	<b>88</b>
3.1 INTRODUCTION.....	88
3.2 METHODS AND MATERIALS .....	92
<i>MPN case selection</i> .....	92
<i>JAK2 locus definition</i> .....	92
<i>Targeted amplification and next-generation sequencing</i> .....	93
<i>Single nucleotide variant analysis</i> .....	93
<i>Genotyping MPN cases and shared controls</i> .....	94
<i>Genotype data processing and association testing</i> .....	95
<i>Population stratification correction and association test</i> .....	95
<i>Imputation and association tests</i> .....	96
<i>Functional annotation</i> .....	97
<i>Allele-specific JAK2 expression in MPN cases</i> .....	97
<i>Targeted Sequencing of JAK2 locus</i> .....	99
<i>Analysis of the JAK2 risk locus in healthy individuals</i> .....	104

<i>Extended Genome Wide Association Study</i> .....	106
<i>Functional prediction of causal variant</i> .....	112
<i>Functional prediction of causal variant</i> .....	112
<i>Allele specific JAK2 expression in MPN cases</i> .....	115
3.4 DISCUSSION .....	117
<b>CHAPTER 4</b> .....	<b>120</b>
<b>AN EVOLUTIONARY MODEL FOR THE JAK2 SUSCEPTIBILITY LOCUS</b> .	<b>120</b>
4.1 INTRODUCTION.....	120
4.2 MATERIALS AND METHODS .....	122
<i>Study population and genotype data</i> .....	122
<i>Haplotype block definition and association test</i> .....	122
<i>Phylogenetic analysis</i> .....	123
<i>HapMap project data</i> .....	123
<i>Positive selection tests</i> .....	123
4.3 RESULTS .....	125
<i>Haplotype association test</i> .....	125
<i>Reconstruction of phylogenetic tree</i> .....	128
4.4 DISCUSSION .....	134
<b>IMPLICATIONS</b> .....	<b>137</b>
REFERENCES .....	140

## LIST OF FIGURES

FIGURE 1. ANALYTICAL POWER OF GWAS .....	41
FIGURE 2 POPULATION SUBSTRUCTURE OF MSKCC PANCREATIC CANCER CASES AND ADDITIONAL CONTROLS .....	47
FIGURE 3 QUANTILE -QUANTILE PLOT OF GWAS OF PANCREATIC CANCER CASES WITH ADDITIONAL CONTROLS .....	49
FIGURE 4 NORMALIZED SIGNAL INTENSITY PLOT FOR RS1975920 .....	53
FIGURE 5 GENOMIC INFLATION FACTOR LAMDA VERSUS NUMBER OF PRINCIPAL COMPONENTS ( PCs )USED FOR CORRECTION .....	62
FIGURE 6 PRINCIPAL COMPONENT ANALYSIS OF MPN CASES AND WTCCC CONTROLS ....	76
FIGURE 7 GENOME WIDE SNP ANALYSIS OF MPN CASES AND WTCCC CONTROLS .....	77
FIGURE 8 JAK2V617F IS ACQUIRED IN CIS WITH JAK2 SNP RS10974944 .....	84
FIGURE 9. TWO HYPOTHESIS TO EXPLAIN 46/1 MPN RISK HAPLOTYP E .....	91
FIGURE 10 SCHEMATIC DIAGRAM OF 300KB JAK2 RISK LOCUS .....	101
FIGURE 11 SINGLE NUCLEOTIDE VARIANT COUNTS FOR MPN CASES WITH AND WIHOUT 46/1 RISK HAPLOTYP E USING HUMAN ANCESTRAL SEQUENCE AS REFERENCE .....	103
FIGURE 12 MANHATTAN PLOT FOR EXTENDED MPN GWA STUDY .....	107
FIGURE 13 ASSOCIATION PLOT FOR IMPUTED SNPs AT 300 JAK2 LOCUS .....	110
FIGURE 14 PREDICTED FUNCTIONAL SNP RS1887428 .....	114
FIGURE 15 ALLELE-SPECIFIC EXPRESSION OF JAK2 IN HETEROZYGOUS MPN CASES .....	116
FIGURE 16 HAPLOTYP E PLOT FOR MPN CASES AND CONTROLS CONSTRUCTED USING HAPLOVIEW .....	126
FIGURE 17 PHYLOGENETIC TREE OF HAPLOTYPES IN BLOCK5.....	130

FIGURE 18 DISTRIBUTION OF DERIVED ALLELE FREQUENCIES AT JAK2 LOCUS, TYRP1 AND ANCESTRAL REPEATS .....	132
FIGURE 19 DISTRIBUTION OF FST CALCULATED USING 11 HAPMAP III POPULATION COMPARING JAK2 LOCUS, TYRP1 AND ANCESTRAL REPEATS.....	133

## LIST OF TABLES

TABLE 1 EMPIRICAL POWER USING GWA SIMULATION STUDY .....	43
TABLE 2 CONTROLS FROM DBGAP USED IN THE PRESENT STUDY.....	46
TABLE 3 SNPs ASSOCIATED WITH PANCREATIC CANCER .....	52
TABLE 4 GENOMIC INFLATION FACTOR FOR ANALYSIS WITH VARIOUS DATASETS.....	55
TABLE 5 RANK AND P-VALUE OF FOUR PANCREATIC CANCER-ASSOCIATED SNPs .....	58
TABLE 6 EFFECT OF CHOICE OF CONTROLS ON ASSOCIATION STATISTICS.....	59
TABLE 7 RANK OF KNOWN PANCREATIC CANCER-ASSOCIATED SNPs.....	63
TABLE 8A GERMLINE GENOTYPE FOR JAK2 SNP rs10974944 AND MPN PREDISPOSITION .....	79
TABLE 9. GERMLINE GENOTYPE FOR JAK2 SNP rs10974744 IN JAK2 V617F -POSITIVE MPN CASES AND NEGATIVE MPN CASES COMPARED WITH WTCCC .....	81
TABLE 10 SINGLE NUCLEOTIDE VARIANT COUNT IN MPN CASES WITH OR WITHOUT 46/1 RISK HAPLOTYPE .....	102
TABLE 11 THE NUMBER OF SINGLE NUCLEOTIDE VARIANTS IN HAPMAP HEALTHY INDIVIDUALS FROM EUROPEAN ANCESTRY( CEU) OBTAINED FROM 1000 GENOMES PROJECT .....	105
TABLE 12 LIST OF SNPs ASSOCIATED WITH MPN RISK .....	108
TABLE 13 ASSOCIATION RESULTS FOR IMPUTED SNPs WITH THEIR FUNCTIONAL ANNOTATION.....	111
TABLE 14 HAPLOTYPE ASSOCIATION RESULTS OBTAINED FROM HAPLOVIEW .....	127
TABLE 15 LIST OF SNPs AND HAPLOTYPES PRESENT IN BLOCK 5 .....	129

## LIST OF ABBREVIATIONS

<b>CDCV</b>	Common disease common variant
<b>cDNA</b>	complementary DNA
<b>CEU</b>	Utah residents with Northern and Western European
<b>ChIP-seq</b>	chromatin immunoprecipitation -sequencing assay
<b>CLL</b>	chronic lymphocytic leukemia
<b>CML</b>	Chronic myelogenous leukemia
<b>DAF</b>	Derived Allele Frequency
<b>dbGaP</b>	The database of Genotypes and Phenotypes
<b>DNA</b>	Deoxyribonucleic acid
<b>ENCODE</b>	The Encyclopedia of DNA Elements
<b>ET</b>	Essential thrombocythemia
<b>FPTR</b>	Familial Pancreatic Tumor Registry
<b>gDNA</b>	genomic DNA
<b>GO</b>	gene ontology
<b>GRR</b>	Genotype relative risk
<b>GWAS (GWA stud</b>	Genome wide association study
<b>HSC</b>	Hematopoietic stem cells
<b>IBD</b>	identity by descent
<b>JAK2</b>	Janus kinase 2 (gene)
<b>KEGG</b>	Kyoto Encyclopedia of Genes and Genomes
<b>LD</b>	Linkage disequilibrium
<b>MPN</b>	Myeloproliferative neoplasms
<b>MSKCC</b>	Memorial Sloan Kettering Cancer Center
<b>NIH</b>	National Institute of Health
<b>OR</b>	Odds Ratio
<b>PC</b>	principal component
<b>PCA</b>	principal component analysis
<b>PMF</b>	Primary myelofibrosis
<b>PV</b>	Polycythemia vera
<b>PVSG</b>	Polycythemia Vera Study Group
<b>QC</b>	Quality control
<b>Q-Q plot</b>	Quantile-quantile plot
<b>SNP</b>	Single nucleotide polymorphism
<b>TSI</b>	Toscans in Italy
<b>UPD</b>	Uniparental disomy
<b>WHO</b>	World Health Organization
<b>WTCCC</b>	Wellcome Trust Case Control Consortium

## INTRODUCTION

### Myeloproliferative neoplasms

Myeloproliferative neoplasms (MPN) are a heterogeneous group of diseases characterized by aberrant proliferation of the myeloid lineages. They represent clonal hematopoietic stem cell disorders with an inherent tendency towards leukemic transformation. The classic BCR-ABL-negative MPNs include polycythemia vera (PV), essential thrombocythemia (ET) and primary myelofibrosis (PMF). They are uncommon tumors with yearly incident rates of 2.3 in 100,000 in the United States and primarily affect older adults, with a variable clinical presentation <sup>1</sup>.

In 1892, Louis Henri Vaquez first described PV in a patient and postulated that it was the result of hematopoietic cell proliferation <sup>2</sup>. Gustav Hueck, a German physician first described PMF and noted the presence of bone marrow fibrosis in patients with PMF <sup>3</sup>. In 1934 Emil Epstein and Alfred Goedel described ET and recognized that patients with thrombocytosis without marked erythrocytosis constituted a distinct clinical syndrome <sup>4</sup>. William Dameshek was the first to notice the clinical and bone marrow morphologic similarities between chronic myelogenous leukemia (CML), PV, ET, and PMF. He recognized their common trait of unregulated trilineage myeloproliferation and accordingly assigned the term myeloproliferative disorders (MPD) to describe them in a seminal 1951 commentary <sup>5</sup>. The first formal attempt in establishing diagnostic criteria for the classic BCR-ABL-negative MPNs was undertaken by the Polycythemia Vera Study Group (PVSG), in 1967 <sup>6</sup>. The PVSG subsequently published similar diagnostic criteria for ET <sup>7</sup>.

PV is characterized by a proliferation of the erythroid lineage, resulting in increased erythroid cell mass, hemoglobin concentration, hematocrit value, and blood viscosity<sup>5</sup>. Patients with PV have an increased incidence of thromboses, hemorrhage, peptic ulcers, and stroke<sup>8</sup>. ET is characterized by dysregulated proliferation of megakaryocytes and platelets in the bone marrow and peripheral blood with an increased risk of thrombosis and bleeding<sup>9</sup>. In PMF, the dominant pathologic change is progressive bone marrow fibrosis and splenomegaly<sup>10,11</sup>.

Of the three classic MPNs, PMF is the most rare and has the worst outcome. The death of a PMF patient is frequently related to bone marrow failure with resultant systemic infection or fatal hemorrhage in many cases. PV, in contrast, has a more indolent course, but there is considerable associated mortality due to thromboses and/or hemorrhage which can be treated with moderate success by therapeutic phlebotomy<sup>12</sup>. The risk of transformation to an acute leukemia is highest in PMF (5%–30%) but can occur in ET and PV<sup>13</sup>.

## JAK2 somatic mutation

The genetic basis for MPN became known in 2005 when 4 separate groups identified a somatic gain-of-function mutation in the Janus kinase 2 (JAK2) gene on chromosome 9p in PV, ET, and PMF patients<sup>14-17</sup>. The JAK2 gene is a member of a large family of tyrosine kinases involved in cytokine receptor signaling. JAK2 is integral to intracellular signal transduction after the activation of receptors for erythropoietin, thrombopoietin, granulocyte-colony stimulating factor, and granulocyte-macrophage colony stimulating factor in the context of hematopoiesis. The signal transduction of these cytokines and their receptors is crucial for the coordinated proliferation and differentiation of the erythroid, megakaryocytic, and granulocytic lineages from pluripotent hematopoietic stem cells (HSCs). The JAK proteins have 2 adjacent kinase-like domains (JH1 and JH2), of which only the JH1 domain has enzymatic activity. The JH2 domain, or pseudokinase domain, is a negative regulator of kinase activity. The point mutation in JAK2 found by different group results in the substitution of valine for phenylalanine at position 617 in the JH2 regulatory domain of the JAK2 protein and is known as the JAK2V617F mutation. As a result of this substitution, JAK2 becomes constitutive activate and acts independent of ligand<sup>18</sup>. This mutation occurs at a primitive stem cell level, mostly HSCs and confers cytokine hypersensitivity and cytokine-independent signaling, leading to the downstream activation of multiple signaling cascades, such as the STAT proteins, phosphatidylinositol 3-kinase–AKT pathway, and mitogen-activated protein kinases and account for the proliferative component in the MPN<sup>19,20</sup>. Experiments performed in animal models confirmed that the mutated

JAK2V617F was constitutively active and the role of JAK2V617F in the pathogenesis of PV<sup>21</sup>.

JAK2V617F is the most prevalent mutation in *BCR-ABL1*-negative MPN: the mutational frequency is approximately 96% in PV, 55% in ET, and 65% in PMF<sup>14</sup>. This mutation is also found at lower percentages in a number of other myeloid malignancies, such as systemic mastocytosis, acute myeloid leukemia, and chronic myelomonocytic leukemia, but not in lymphomas or solid tumors. In 2008, the World Health Organization (WHO) included screening for JAK2V617F as diagnostic criteria for PV, ET and PMF<sup>22</sup>. The mutation cannot be used to distinguish one MPN from another, but it does complement histology in the diagnosis of PV, ET and PMF. The identification of the common JAK2V617F somatic mutation in ET, PV, and PMF has also led to targeted therapy using small-molecule JAK2 inhibitors<sup>23,24</sup>.

## MPN phenotype pleiotropy

The three subtypes of MPN are pathologically distinct disorders despite the shared genetic lesion JAK2V617F. The mechanism of how a single mutation can produce 3 different diseases is not clear. One explanation of MPN pleiotropy is a gene dosage effect of JAK2. JAK2V617F is an acquired hematopoietic stem cell mutation, yet, in many patients, hematopoiesis remains polyclonal since not all stem cell progenitors within an individual carry the mutation. In addition, many patients with the JAK2V617F mutation acquire two copies through an acquired uniparental disomy (UPD) at chromosome 9p24<sup>25</sup>. Thus, because of the variability in the number of cells that carry the mutation and the number of JAK2V617G alleles harbored within each cell, there is marked variability in JAK2V617F gene dosage. In murine and human studies, the JAK2V617F allele burden is lowest in ET compared with that of PV and PMF<sup>26-28</sup>. Sex is also an independent modifier of the MPN, with women having lower mutational burdens in JAK2V617F than men<sup>29</sup>.

Variation in JAK2V617F mutational burden alone, however, cannot explain the variability of disease phenotypes within the MPN. Host genetic background has been shown to play a significant role in the acquisition of the JAK2 mutation itself in mouse models<sup>30,31</sup>.

Additional somatic mutations have been identified in myeloproliferative neoplasm patients and may also contribute to the pathogenesis of JAK2V617F positive PV, ET, and PMF and MPN pleiotropy. Currently known MPN-associated mutations involve *JAK2* (exon 12)<sup>32-34</sup>, *MPL* (exon 10)<sup>35-37</sup>, *TET2*<sup>38,39</sup>, *ASXL1*<sup>40,41</sup>, *IDH1* and *IDH2*<sup>42,43</sup>, *CBL*<sup>44</sup>,

*IKZF1*<sup>45,46</sup>, *LNK*<sup>47,48</sup>, and *EZH2*<sup>49</sup>. Most of these mutations originate at the progenitor cell level but they do not necessarily represent the primary clonogenic event and are not mutually exclusive. *JAK2* exon 12, *MPL* and *LNK* mutations are relatively specific to *JAK2V617F*-negative MPN whereas the mutations observed in *TET2* (TET oncogene family member 2; 4q24) gene are seen in both *JAK2V617F* positive and negative MPN<sup>50</sup>. *TET2*, *EZH2* and *ASXL1* may contribute to epigenetic regulation of hematopoiesis<sup>39,49</sup>.

Thus, on the one hand, the sole *JAK2V617F* mutation is sufficient to induce an MPN, and the MPN phenotype depends on the cell targeted by the mutation or the genetic background of the patients or the intensity of *JAK2V617F* signaling. On the other hand, *JAK2V617F* can be an event secondary to a first hit that varies between the diseases.

## MPN familial studies

Familial clusters of MPNs are characterized by clinical and genetic heterogeneity. First, within MPN families, distinct clinical entities are observed, the three main ones being PV, ET, and PMF. Second, disease evolution can be highly variable within families presenting with the same type of MPN. The primary familial congenital polycythemia and hereditary thrombocythemias, which are rare Mendelian disorders, are caused by mutations in the erythropoietin receptor gene and thrombopoetin gene, respectively<sup>51,52</sup>. These mutations have not been detected in the more common MPNs.<sup>53</sup>

The evidence for possible heritable component to MPN came from a small number of case reports and case series describing families with multiple affected individuals. JAK2 mutation analysis in these familial cases has led to several important observations. Even among familial cases in which all affected family members shared the V617F mutation, this mutation was identified as an acquired or somatic mutation and not an inherited mutation. Overall, the incidence of the V617F mutation is similar in familial and sporadic MPNs, and is found in 55% to 75% of familial cases of PV versus 95% of sporadic cases, 75%–90% of familial cases of PMF versus 50% of sporadic cases, and 50%–69% of familial cases of ET versus 50% of sporadic cases<sup>54</sup>.

In another study of 458 patients with apparently sporadic MPNs, 35 were found to be members of multiplex MPN families<sup>55</sup>. From this study, it was estimated that the prevalence of familial disease was 8.7%, 5.9%, and 8.2% for PV, ET, and PMF respectively.

In the only population-based study yet performed, investigators from Sweden found that the first-degree relatives of MPN patients had significantly increased risks of PV (RR = 5.7; 3.5-9.1) and ET (RR = 7.4; 3.7-14.8). The Swedish study's findings support the hypothesis that common, strong, shared susceptibility genes predispose to PV, ET, MF, and possibly CML. In many of these kindreds the inheritance pattern is consistent with autosomal dominant inheritance with incomplete penetrance<sup>56</sup>.

The evidence from familial MPN studies suggests that additional inherited alleles that predispose to MPN development or inherited modifiers that contribute to the clinical phenotype of MPN contribute to the pathogenesis of PV, ET, and PMF.

## Genome Wide Association Studies

In recent years, a genome wide association (GWA) study has been advocated as a method of choice to identify genetic variant(s) associated with various common diseases. The Human Genome Project stimulated the efforts to characterize the most abundant genetic variants in the human genome, single nucleotide polymorphisms (SNPs). SNPs are DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence is altered, and it must occur in at least 1% of the population. An estimated 3 million SNPs, which make up about 90% of all human genetic variation, occur every 100 to 300 bases along the 3-billion-base human genome. The nonrandom association between neighboring SNPs is called linkage disequilibrium; alleles of SNPs in high linkage disequilibrium are almost always inherited together and can serve as proxies for each other. Their correlation with each other in the population is measured by the  $r^2$  statistic, which is the proportion of variation of one SNP explained by the other, and ranges from 0 (no association) to 1 (perfect correlation). This approach relies on the foundation of data produced by the International Human HapMap Project<sup>57</sup>. Common genetic variation by and large is organized in “haplotype blocks,” local regions that have not been broken up by meiotic recombination and are separated by recombination “hot spots” that occur every 100–200 kb. These observations provided the empirical foundation for the construction of a haplotype map of the human genome for diverse populations. This haplotypic structure of the human genome makes it possible to survey the genome for common variability associated with the risk of disease simply by genotyping approximately 500,000 to 1 million judiciously chosen markers known as tagging SNPs<sup>58</sup>, in the genome of several thousand case subjects and control subjects.

The development of latest SNP chip technologies can now scan up to 1 million SNPs thus allowing GWAS results to hold the promise of revealing causal genes not previously suspected in disease etiology or genetic effects of non-genic DNA regions.

The impetus behind these studies can be traced back to two key papers from 1996<sup>59,60</sup>. These two papers argued that common variants may underlie many common diseases and would be more easily found using population-based association studies rather than family-based linkage analysis. This led to the common disease common variant hypothesis, first proposed in 2001<sup>53</sup>. It states that complex diseases are caused by the interaction of common alleles at a small group of susceptibility loci. These common alleles are not population specific, but are present at >1% minor allele frequency in multiple populations.

## Linkage studies versus Genome wide association studies

Early genetic mapping studies in humans utilized linkage mapping, a methodology that traces the transmission of phenotypes with genetic markers through pedigrees with positional cloning used to find gene mutations that lead to monogenic diseases. The linkage studies have been successful in identifying highly penetrant genetic variants of large effect [odds ratio >100] underlying hundreds of Mendelian diseases (for example, the *HTT* gene in Huntington's disease<sup>61,62</sup> and the *CFTR* gene in cystic fibrosis<sup>63,64</sup>). These searches have mostly led to identification of mutations that alter the amino acid sequence of a protein and enormously increase the risk of disease. Several common disease-predisposing variants that are associated with common disease variation were identified in early linkage/candidate gene studies, *e.g.*, Factor V<sup>Leiden</sup> in deep venous thrombosis<sup>64,65</sup> the *APOE* $\epsilon$ -4 allele in Alzheimer's disease<sup>65</sup>, and *PPAR* $\gamma$  in type 2 diabetes<sup>66</sup>. The major limitations of linkage studies are 1) relatively low power for complex disorders influenced by multiple genes, and 2) the large size of the chromosomal regions shared among family members (often comprising genomic regions of 5-10 Mb harboring hundreds of genes), in whom it can be difficult to narrow the linkage signal sufficiently to identify a causative gene.

In contrast to monogenic traits, complex traits have been more difficult to unravel using linkage approaches. GWA study is based on population-based samples (“common disease/common variant” (CDCV) hypothesis) and has power to identify common variants of modest effect, which could not be found using traditional linkage-based approaches. The association studies are able to refine genomic loci to roughly 10–100kb,

often just a few genes. Thus GWA studies build on the valuable lessons learned from family linkage studies, as well as the expanding knowledge of the relationships among SNP variants generated by the International HapMap Project<sup>57</sup>.

## GWAS study design

The typical GWA study has 4 parts: (1) selection of a large number of individuals with the disease or trait of interest and a suitable comparison group; (2) DNA isolation, genotyping, and quality control to ensure genotyping quality; (3) statistical tests for associations between the SNPs passing quality thresholds and the disease/trait; and (4) replication of identified associations in an independent population sample or examination of functional implications experimentally.

The most frequently used GWA study design has been the case-control design, in which allele frequencies in patients with the disease of interest are compared to those in a disease-free comparison group. Cohort studies involve collecting extensive baseline information in a large number of individuals who are then observed to assess the incidence of disease in subgroups defined by genetic variants.

Using the various genotyping platforms developed by commercial companies Affymetrix and Illumina, upto 1 Million SNPs can be genotyped at once. Genotyping platforms comprising 500,000 to 1,000,000 SNPs have been estimated to capture 67% to 89% of common SNP variation in populations of European and Asian ancestry and 46% to 66% of variation in individuals of African ancestry<sup>67</sup>. Genotyping errors, especially if occurring differentially between cases and controls, are an important cause of spurious associations and must be diligently sought and corrected<sup>68</sup>. A number of quality control features should be applied both on a per-sample and a per-SNP basis. Checks on sample identity to avoid sample mix-ups and a minimum rate of successfully genotyped SNPs per sample (usually 80%-90% of SNPs attempted) should be determined. The quality

control filters for probable genotyping errors, include the following: (1) the proportion of samples for which a SNP can be measured (the SNP call rate, typically >95%); (2) the minor allele frequency (often >1%, as rarer SNPs are difficult to measure reliably); (3) severe violations of Hardy-Weinberg equilibrium; and (4) concordance rates in duplicate samples (typically >99.5%) are regularly performed.

#### *Statistical tests for association*

Associations with the two alleles of each SNP are tested in a relatively straightforward manner by comparing the frequency of each allele in cases and controls. The most powerful tool for the analysis of GWA data has been a single-point, one degree of freedom test of association, such as the Cochran–Armitage test. Such tests allow comparison of the genotype distributions of cases and controls at each SNP in turn, and can be conducted with or without adjustment for relevant covariates. The different genetic models (dominant, recessive, or additive) may be included in the analysis, although additive models, in which each copy of the allele is assumed to increase risk by the same amount, tend to be the most common. Odds ratios (OR) of disease associated with the risk allele or genotype(s) can then be calculated and are typically modest, often in the range of 1.2 to 1.5. Many studies also calculate population attributable risk, classically defined as the proportion of disease in the population associated with a given risk factor. Many software are available for analysis with PLINK being the most popular GWAS analysis package <sup>69</sup>.

When testing 1 million SNPs for association, 50,000 SNPs will appear to be “associated” with disease at the conventional  $P < .05$  level of significance. Almost all are

false positives and due to chance alone- this is known as the multiple testing problem. The most common manner of dealing with this problem is to reduce the false-positive rate by applying the Bonferroni correction, in which the conventional  $P$  value is divided by the number of tests performed <sup>70</sup>. A 1 million SNP survey would thus use a threshold of  $P < .05/10^6$ , or  $5 \times 10^{-8}$ , to identify associations unlikely to have occurred by chance. Other approaches have been proposed, including estimation of the false discovery rate or proportion of significant associations that are actually false positive associations, false-positive report probability <sup>71,72</sup>, calculation of probability that the null hypothesis is true given a statistically significant finding, <sup>73</sup> and/or estimation of Bayes factors that incorporate the prior probability of association based on characteristics of the disease or the specific SNP <sup>74</sup>.

#### *Replication stage*

An important strategy has been replication of GWA results in independent samples to separate the many false-positive associations from the few true-positive associations with disease in GWA studies. The consensus criteria for replication is to test the SNP reported in the initial study in the same or very similar phenotype and population, and demonstration of a similar magnitude of effect and significance (in the same genetic model and same direction) for the same SNP and the same allele as the initial report <sup>75</sup>.

## Fine mapping of disease locus to identify causal variant(s)

The causal variant is usually not identified by GWA studies and may be more strongly associated (and explain more of the risk) than the marker detected in the initial GWA. To generate a comprehensive list of potential causal variants that could explain an association signal, resequencing across the entire region of association (at least out to the point at which LD has substantially decayed) and confirmatory genotyping efforts is generally required <sup>76</sup>. Next generation sequencing technologies like SOLiD sequencing or Illumina can be used to sequence the region identified to be associated with disease, both in depth and breadth, to fully interrogate nearby variants /genes for possible susceptibility alleles.

Imputation, a statistical method can be used to predict /generate statistically all SNPs in HapMap <sup>67,77</sup> or 1000 Genomes Project and can be tested for association <sup>78</sup>. Numerous methods like Impute <sup>78,79</sup>, Mach <sup>80,81</sup>, Beagle<sup>82</sup> are regularly used to impute millions of SNP for association test. Efforts are currently being directed toward implementation of novel analytic approaches and testing rare variants for association with complex traits using imputed variants from the publicly available 1000 Genomes Project <sup>77</sup> resequencing data and from direct resequencing of clinical samples.

Another method of identifying causal variants in an extended strong LD block is to perform fine mapping in populations of different ancestries. The pairwise correlation coefficients will not be equally high in all populations .By genotyping all of the equivalently associated variants in multiple populations, it is possible that a subset of variants may emerge that show a more consistent pattern of association across

populations, making these as more likely candidates for being causal. Individuals of African ancestry may be particularly helpful because of the lower levels and distinct patterns of LD<sup>83,84</sup>.

Functional annotation of the genome can shed light on mechanisms of the trait's biology. One common approach is to determine whether trait-associated variants cluster into groups of specific biological functions more than would be expected by chance, *e.g.*, within gene ontology (GO) terms. Large-scale databases integrate various types of data from the literature to build pathways, and commercial and public tools exist to facilitate access *e.g.*, Ingenuity, Kyoto Encyclopedia of Genes and Genomes (KEGG), ENCODE<sup>85</sup>. Similarly, other data types such as methylation / acetylation, protein–protein interactions, and miRNA regulatory networks, can be integrated with GWAS results<sup>85</sup>. Through integration with annotations and functional genomic data as well as *in vitro* and *in vivo* experimentation, mapping studies continue to characterize functional variants associated with complex traits.

## Success in GWAS

In 2005, the first successful GWA study was of age-related macular degeneration, with 100,000 SNPs tested for association in 96 cases and 50 healthy controls<sup>86</sup>, followed by GWA studies for Crohn's Disease<sup>87</sup>, myocardial infarction<sup>88</sup>, inflammatory bowel disease<sup>89</sup>, and type 2 diabetes<sup>90</sup>. A landmark study by the Wellcome Trust Case Control Consortium (2007) (WTCCC)<sup>74</sup> reported successfully the GWAS results for seven common diseases, including bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type I diabetes, and type II diabetes. Several common variants influencing continuous traits, such as lipids<sup>91</sup>, height<sup>92,93</sup> and fat mass<sup>94,95</sup>, have also been found. GWA studies have also proven to be successful in identifying more than 200 mostly common low-penetrance susceptibility loci for a range of different cancer types. These included for example breast<sup>96-103</sup>, prostate<sup>104-110</sup>, lung<sup>111-113</sup>, colorectal<sup>114-120</sup>, urinary bladder<sup>121,122</sup>, pancreatic cancer<sup>123,124</sup>, hematological malignancies<sup>125-127</sup>, gliomas<sup>128</sup> and ovarian cancers<sup>129</sup>. Breast cancer and prostate cancer GWAS have been at forefront of cancer GWAS with many loci associated with these diseases. As of March, 2011, 1319 human GWAS at  $p\text{-value} < 5 \times 10^{-8}$  have been published on 221 traits. The National Institute of Health (NIH) keeps a weekly updated a catalog of published GWAS results (<http://www.genome.gov/gwastudies>)<sup>130</sup>.

Examples of experimentally confirmed functional variants underlying validated GWAS hits are accumulating, revealing a variety of functional mechanisms underlying trait variation. For example the IRF5 locus includes variants that disrupt intron splicing, decrease mRNA transcript stability, and delete part of the interferon regulating factor

(IRF) protein <sup>131</sup>, all of which together explain the independent associations with systemic lupus erythematosus <sup>132,133</sup>, inflammatory bowel disease <sup>134</sup>, and rheumatoid arthritis <sup>135</sup>. Allele-specific chromatin remodeling affecting the expression of several genes in the ORMDL3 locus region<sup>136</sup> explains its association with asthma <sup>137</sup>, Crohn's disease <sup>138</sup> and type 1 diabetes <sup>139</sup>. At a locus associated with elevated LDL-cholesterol levels in the blood and myocardial infarction, a common nonprotein-coding variant was found to create a transcription factor binding site that alters the expression of the *SORT1* gene in the liver <sup>140</sup>.

A striking example of functional data of cancer GWAS SNPs is the noncoding variants concentrated within a 1.2 Mb gene desert on chromosome 8q24, where numerous studies have reported associations between multiple types of cancer—including prostate, colorectal, breast, and urinary bladder. Various groups have studied the 8q24 locus and identified two functional SNPs and several transcriptional enhancers. Two of them in a prostate cancer risk region were occupied by androgen receptor and a SNP lies within a FoxA1 binding site <sup>141</sup>. In a separate study a 8q24 SNP in colorectal cancer was found situated within a transcriptional enhancer and its activity is affected by the risk SNP <sup>142</sup>. The risk SNP has been shown to physically interact with the MYC gene with allele-dependent binding of transcription factor 7-like 2(TCFL2) <sup>142</sup>. Researchers demonstrated that the 8q24 cancer-associated variant lies within an *in vivo* prostate enhancer whose expression mimics that of the nearby *MYC* proto-oncogene in mouse model <sup>143</sup>. Thus GWA results have provided unprecedented views into the contribution of common variants to complex traits, illuminated genome function, and have opened new possibilities for the development of therapeutic interventions.

## GWAS challenges

Although GWA studies have proven successful in identifying regions of the genome harboring variants that contribute to complex phenotypes and diseases, several challenges have been encountered:

### *Power and cost of the GWAS*

The statistical power of a GWAS is a function of its total sample size, effect size, causal allele frequency, marker allele frequency and the strength of correlation between marker alleles and causal variants. GWA studies need a large number of cases and controls to be genotyped to attain power to identify genetic variants with small effect sizes. To address this problem, many groups have joined efforts to create large consortia with DNA samples from thousands or tens of thousands of individuals to conduct studies that are well powered to detect even a modest genetic effect. Even with large consortia, however, the cost of genotyping such a large number of samples can be prohibitive. When the genotyping is performed across various institutions and later combined, technical errors and batch effects may be introduced. In 2007, the Wellcome Trust Case Control Consortium (WTCCC)<sup>74</sup> used the "shared controls" approach to study seven common diseases. Rather than using controls individually matched to the cases for each disease, the WTCCC genotyped a common set of controls representative of the self-identified white European population of Great Britain and compared allele frequencies from this group with each set of case individuals. This approach has increased the power of GWAS in a cost effective manner.

### *Population Stratification*

Confounding due to population stratification (also called population structure) has been cited as a major threat to the validity of genetic association studies. The presence of population stratification (PS)—allele frequency differences between cases and controls due to systematic ancestry differences—can lead to greater than nominal type I error rate. Differences in the origin of populations of cases and controls can arise if the two groups are recruited independently or have different inclusion criteria. Population structure can be assessed in GWA studies by examining the distribution of test statistics generated from the thousands of association tests performed (eg, the  $\chi^2$  test) and assessing their deviation from the null distribution (that are expected under the null hypothesis of no SNP associated with the trait) in a quantile-quantile or “Q-Q” plot. The extent and impact of PS on case-control association studies in practice, particularly in GWAS, can now be thoroughly investigated by a strategy that leverages the fact that in the context of GWAS, the vast majority of the SNPs are not associated with the trait under study and therefore can be used to infer ancestry and evaluate/adjust for PS. One popular type of method e.g. EIGENSTRAT<sup>144</sup> constructs principal components (PCs) on the genotype data and infer a continuous axis of genetic disparity. Afterwards, GWAS tests are corrected by adjusting simultaneously for top-ranked PCs.

### *Insensitivity to rare variants*

Evaluation of the contribution of rare variants to common disease susceptibility raises issues related to detection and functional assessment. The rare variants are poorly

captured by the standard GWA SNP chip arrays and the sheer number of such variants has the limited power to test them for association.

#### *Environmental exposure and other non-genetic factors*

There is a need for improved methods to estimate the joint effects of multiple genes (G) and/or environmental exposures (E) on disease predisposition. Such analyses raise both computational, statistical and study design issues, related to the scale and complexity of the data and the large number of hypotheses that could be addressed. Most GWASs have not investigated  $G \times E$ , primarily due to lack of data on environmental exposures.

#### *Source of heterogeneity*

The interpretation of a failure to replicate GWA results is difficult. If it is clear that the replication studies were well powered and well performed, and that there is genuine divergence between the effect-size estimates, then the possible explanation can be attributed to some source of heterogeneity. The list of potential causes of heterogeneity is long: it includes variable patterns of LD between the genotyped SNP and untyped causal alleles (although this is unlikely if the samples are of similar ancestry); differences in the distribution, frequency or effect size of the causal alleles at a given locus (due to, for example, differences in case ascertainment); and the impact of non-additive interactions with other genetic variants or environmental exposures.

### *Missing heritability*

For most traits or complex disease studies in GWA study, the effects of all associated loci account for a small proportion of the estimated heritability. With the exception of age-related macular degeneration and type 1 diabetes, for which collectively the proportion of heritability explained to date is approximately 50% and 80%<sup>86,139</sup>, respectively, most complex disease variants identified to date together account for much less of the trait variance. However, these loci in combination typically explain only a fraction of the inherited contribution to risk, raising the question of how best to find the variation responsible for the remainder.

### *Poorly understood genotype –phenotype mechanism*

For most associated loci, there is substantial ignorance regarding the mechanisms by which genetic variation could influence phenotype: the identity of the gene(s) affected by the susceptibility variant(s) at each locus is often uncertain, and the mechanisms by which the causal variants (also often unknown) influence phenotype is usually unclear. This lack of knowledge is a substantial impediment to the understanding needed to make progress towards new therapies or preventive measures. This obstacle highlights the need to pinpoint the causal variants and the genes affected by those variants, as well as for informative functional and computational studies to move from gene identification to possible mechanisms that could guide translational progress.

## Evolutionary model for disease the susceptibility locus

To explain the evolutionary framework of disease susceptibility locus, discovered by linkage studies or GWAS, there are two models: 1) Mutation-selection-balance model and 2) Ancestral susceptibility model.

Mendelian traits are controlled by genes of large effect and show simple patterns of inheritance within families. They are usually caused by rare strongly deleterious and new mutations. The new mutations, usually referred to as ‘derived’ alleles, can be inferred by comparing the allele observed at any given human polymorphic site with its orthologous nucleotide position in a close outgroup species (e.g. the chimpanzee). The mutation-selection-balance model can explain such disease causing variants in which disease alleles are continuously generated by new mutation and eliminated by purifying selection. This framework has been also used to model the genetic risk to common diseases based on the observation that most common diseases have a late age of onset. Thus, mutation-selection-balance model for common disease suggests that disease variants are derived (new mutation) and slightly deleterious<sup>53,145,146</sup>. At such loci, the total frequency of susceptibility mutations may be quite high, and there is likely to be extensive allelic heterogeneity at many of these loci due to weak purifying selection acting on these loci. The reason that weak purifying selection increases polymorphism is that it greatly reduces the probability that susceptibility alleles will be at or near fixation. The situation is different when the susceptibility alleles are very deleterious—as seen at Mendelian disease loci—in which case, selection dominates the effects of mutation pressure and drift and keeps susceptibility alleles at low frequency. For association or

linkage-disequilibrium mapping, it is important to know about the frequencies and ages of individual mutations within the susceptible class. Thus the genetic variation at disease-susceptibility loci may possibly be determined by taking into account the evolutionary processes such as mutation, genetic drift, and the possibility of selection.

The second evolutionary framework to explain disease-susceptibility locus was proposed by Di Rienzo A and Hudson RR in 2005<sup>147</sup>. They observed that unlike rare Mendelian diseases, which are due to new mutations (i.e. derived alleles), several alleles that increase the risk to common diseases are ancestral alleles, whereas the derived alleles are protective. Examples include variants involved in biological processes such as energy metabolism and sodium homeostasis. The  $\epsilon 4$  allele of the gene encoding Apolipoprotein E (*APOE*), which increases the risk to coronary artery disease<sup>148,149</sup> and Alzheimer's disease<sup>65,150</sup> carries the ancestral allele at two common amino acid polymorphisms. These observations can be explained in which ancestral alleles reflect ancient adaptations to the lifestyle of ancient human populations, whereas the derived alleles were deleterious<sup>151</sup>. However, with the shift in environment and lifestyle, the ancestral alleles now increase the risk of common diseases in modern populations.

## Introduction to my thesis project

The goal of my thesis project was to understand germline genetics of MPN using GWA study followed by fine mapping of the MPN susceptibility locus. As discussed earlier in this section, MPN shows remarkable molecular heterogeneity and the etiology of this disease remains unclear. The story of MPN pathogenesis started with the discovery of the JAK2V617F mutation<sup>14-17</sup>; followed by identification of many other mutations of MPN some involving JAK-STAT signaling activation, others chromatin remodeling and others still leukemic transformation. A role for inherited genetic factors in the etiology of MPNs has also been suggested from smaller case studies showing evidence of familial clustering of PV, ET, MF, and chronic myeloid leukemia (CML) as well as in largest population-based case-control study. The central hypothesis of this project is that there are common, strong, shared germline susceptibility loci that predispose to all three MPN - PV, ET and PMF. To test our hypothesis, we performed GWA study of patients diagnosed with MPN. Given that our MPN dataset lacks genotype data for healthy controls, we used a “shared control approach” in our GWA study. The shared controls are a group of healthy individuals that can be used as controls in GWA studies of different diseases. The shared controls approach was first used by the Wellcome Trust Case-Control Consortium (WTCCC) study<sup>74</sup>. We developed a systematic method to match genetically diverse cases with controls from public database instead of using matched control study design. Thus, my thesis work capitalizes on the concept of shared controls in GWA studies and established methodologies for analyzing cases with shared controls in GWA study to identify the germline variant in JAK2 gene associated with MPN. We studied the JAK2 susceptibility locus using phylogenetic analysis tools.

Fine mapping of the MPN disease locus was done using imputation and next generation technologies, and we identified candidate functional variant(s) that may play an important role in the etiology of MPN.

I have organized my findings into the following chapters:

CHAPTER 1: *Using additional controls from public database to increase power of GWAS*

We developed a pipeline to match genetically diverse cases with shared controls on the basis of their genetic variation. We used analytical methods to calculate the optimum number of cases and controls. To determine the type I error rate and power of the method, a whole genome simulation study was used. As proof of principle, we used a pancreatic cancer dataset to test the power of this method.

CHAPTER 2: *Identification of genetic variant(s) associated with MPN predisposition*

The GWA study was performed comparing MPN patients with controls from public database. We identified the genetic variant rs10974944, SNP located in JAK2 gene to be associated with MPN predisposition. The MPN associated haplotype is known as the 46/1 haplotype.

### CHAPTER 3: *Mechanism for JAK2 susceptibility haplotype in MPN*

We explored the two suggested hypotheses – the hyper-mutability hypothesis or the activation hypothesis to explain the mechanism of the well-established finding that the 46/1 JAK2 haplotype predisposes to JAK2V617F positive MPN. We used targeted sequencing and fine mapping to understand the role of 46/1 susceptibility haplotype in predisposition to MPN.

### CHAPTER 4: *Evolutionary framework of JAK2 susceptibility locus*

We investigated the JAK2 haplotype that is associated with MPN using Haploview and reconstructed the phylogenetic tree using chimpanzee as outgroup to understand the relationship of various haplotypes present in the JAK2 locus. Even though there is no evidence of recent positive selection at the JAK2 locus, we observed an excess of derived alleles at the JAK2 locus. We concluded that the JAK2 susceptibility locus exhibits the ancestral-susceptibility model.

## CHAPTER 1

### Using addition controls from public databases to increase power of GWAS

#### 1.1 Introduction

A typical GWA study involves a case-control design in which the investigator analyzes DNA samples from both affected case individuals and matched, healthy control individuals. One hurdle in conducting such studies, in which hundreds of thousands of SNPs are independently tested for association with disease, is the large sample size required to obtain adequate power to detect a modest effect after correcting for multiple testing. To address this problem, many groups have joined efforts to create large consortia with DNA samples from thousands or tens of thousands of individuals to conduct studies that are well powered to detect even a modest genetic effect. Even with large consortia, however, the cost of genotyping such a large number of samples can be prohibitive.

One potential solution to the sample size requirement of GWAS that has been proposed is the use of a common set of control individuals in numerous studies. In 2007, the Wellcome Trust Case Control Consortium (WTCCC) used this "shared controls" approach to study seven common diseases<sup>74</sup>. This approach has been used by others with case individuals who come from both the UK and elsewhere, including the United States<sup>74,125,128,152,153</sup>. Recently Zhuang *et al.* reported a simulation study in which they showed the theoretical potential for expanding the control group with publicly available

disease or reference samples to increase the power of GWAS<sup>154</sup>; we refer to the use of such controls from the database as "additional or shared controls."

Despite the apparent practical success of this approach and simulation studies suggesting its effectiveness, both the power and pitfalls of using additional controls from databases in the genetically heterogeneous United States population remains unclear. Genome-wide genotype information, along with limited phenotypic data, is available for numerous healthy individuals from the U.S. in the dbGaP database at NIH. Therefore, in theory it should be possible to combine these data with genome-wide SNP profiles from a smaller number of cases that an individual investigator is studying to identify disease susceptibility loci. However, population stratification due to differences in genetic ancestry between people in such case and control groups and differential genotyping error from different sources could hinder effective use of this approach. It is known that even if a study is restricted to self-identified "white" individuals in the United States, genotype frequency at many loci can vary based on from where in Europe ancestors came<sup>155,156</sup>. While a variety of statistical methods have been developed to identify and correct for such stratification<sup>144,157</sup>, how such correction will influence the power and type I error rate of using common controls in US-based studies remains to be seen.

In this chapter we evaluate the use of additional controls from publicly available sources in a U.S.-based GWAS. To do so, we utilize a small pancreatic cancer dataset for which we have genome-wide genotype data on 263 cases and 202 controls. We chose this dataset in part because four recently reported pancreatic cancer associated SNPs could be

used as true positives to estimate the power of this additional control approach in a real setting<sup>123,124</sup>. We found that the rank and  $p$ -value of these true disease SNPs improved significantly in our data set with additional controls, with the added benefit of more controls reaching a plateau after a control: case ratio of 10:1 is obtained. Despite a large amount of population stratification in this joint dataset, the impact of this stratification was effectively captured and corrected by principal component analysis (PCA). We demonstrate the utility of genotyping some controls at the same time as cases for comparison with the additional controls to remove SNPs that show differential allele frequencies due to disparity in data processing and technical artifacts. We thus show systematically for the first time the practical issues that concern the use of controls from different sources. This report can serve as useful guidance when using additional controls from publicly available datasets in future studies.

## **1.2 Subjects and Methods**

### **Ethics Statement**

The study was approved by the MSKCC Institutional Review Board and all participants signed informed consent.

### **Analytical power calculation**

We determined the analytical power of GWAS assuming a simple test of allelic association. We computed the power using a non-central  $\chi^2$  distribution with non-centrality parameter  $\lambda$ <sup>158</sup>. The power was computed under an additive model with the significance threshold  $\alpha=1 \times 10^{-7}$ . The genotype relative risk (GRR) was varied from 1.0-3.0 with increments of 0.1 and the disease allele frequency (DAF) was varied from 0.05 to 0.50. The number of cases used range from 100 to 3000, and the control:case ratio ranged from 1:1-50:1.

### **Simulation study for empirical power and type I error rate calculations**

The simulated genotype data for cases and controls were generated using GWASimulator<sup>159</sup>. The GWASimulator uses moving window algorithm to generate whole genome data based on a set of phased input data. As an input data, we used HapMap<sup>57</sup> individuals from European ancestry Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) and Toscan from Italy

(TSI) phased data from HapMap3. Total of 500 cases and 5000 controls were simulated to generation ratio of case: control as 1:10. The ratio of CEU to TSI in cases alone was set to 4:1. The simulated population based controls were generated separately such that the ratio of CEU: TSI controls were either 4:1 (No Stratification) or 99:1 (creating strong population stratification in the dataset). We used 300K Illumina SNP chip markers excluding markers in chromosome X to obtain the simulated genotype data. Disease SNPs were chosen with genotype relative risk 1.6 and of disease prevalence 0.05. Three categories of disease loci were chosen -1) Markers with same minor allele frequency in the two input reference population called as undifferentiated markers, 2) Markers with minor allele frequency in CEU was greater than TSI such that the difference is 0.15 called as CEU high markers and 3) markers with minor allele frequency in TSI was greater CEU and difference between MAF in TSI and CEU was 0.15 called as TSI high markers. The category 2 and 3 were differentiated SNPs. 1000 markers from each category were generated as disease locus generating 100 simulated files for each type.

To correct for population stratification, we used principle component analysis method (Eigenstrat). The independent set of markers were obtained by using LD based SNP pruning. We used  $r^2$  threshold of 0.05 to obtain 30,000 makers that were independent of each other. Using this reduce set of independent markers, PCA was calculated and top two PCs were used as covariates in the logistic regression model. The marker that have p-value less than genome wide nominal value of  $1e-07$  were considered to be genome wide successful. If the simulated disease locus has p-value less than the global significant threshold, it was considered 'success'. We computed power as the

number of times the simulated disease locus was considered as success hit out in 100 iterations. The markers that had p-value less than significant threshold and were not in LD with simulated disease locus were false positive. To determine the type I error rate, the average number of false positive was calculated divided by the total number of markers (240,000). We compared the power and type I error rate with or without PCA corrected method in our simulations.

### **Pancreatic cancer study samples and genotyping**

The pancreatic cancer study dataset was obtained from an ongoing hospital based case-control study conducted in conjunction with the Familial Pancreatic Tumor Registry (FPTR) at Memorial Sloan-Kettering Cancer Center (MSKCC). Patients were eligible if they were age 21 or over, spoke English, and had pathologically or cytologically confirmed adenocarcinoma of the pancreas. Patients were recruited from the surgical and medical oncology clinics at MSKCC when seen for initial diagnosis or follow-up. Controls were visitors accompanying patients with other diseases to MSKCC or spouses of patients. They had the same age and language eligibility requirements as the cases and were not eligible if they had a personal history of cancer (except for non-melanoma skin cancer). The 263 cases and 202 controls in this analysis were recruited between June 2003 and July 2009. The participation rate among approached and eligible individuals was 76% among cases and 56% among controls. Participants provided a blood or buccal (mouthwash or saliva) sample for DNA and completed risk factor and family history questionnaires administered by the research study assistant by telephone or in person.

Genomic DNA was isolated from buccal cells using the Puregene DNA purification kit (Qiagen, Inc; Valencia CA). DNA was also isolated from saliva samples with the Oragene saliva kits (DNA Genotek; Kanata, Ontario, Canada) or from blood using Genra Puregene blood kit (Qiagen Inc; Valencia CA). DNA samples were hydrated in 1x TE buffer. Genomic DNA was genotyped on the Illumina 370K SNP chip (either the Illumina CNV370-Duo or Illumina CNV370-Quad) at the Genomics Core Laboratory of MSKCC according to the manufacturer's protocol.

### **Additional controls from dbGaP**

Genotypes from additional controls were obtained from the NIH's Database of Genotypes and Phenotypes (dbGaP). All individuals used are controls in the underlying study and are of European ancestry. Specifically, data from six studies in dbGaP genotyped using Illumina chips were used (Table 1). These data sets provide 5485 additional controls total. Using a common set of markers present in all the datasets, we combined our MSKCC cases and controls with some or all of the additional controls to yield control: case ratios of 5:1, 10:1 or 20:1.

### **Data processing and quality control**

All genotype data was processed using PLINK <sup>69</sup>. We performed several steps of quality control (QC). First, we processed the MSKCC samples alone, without additional controls. As we could not be certain of the DNA strand the genotype calls from each study are in reference to, we removed all A/T and C/G SNPs, as strand could be confused

for these allele pairs. We removed individuals for whom less than 90% of genotypes were called and SNPs for which less than 10% of genotypes were called. We also removed SNPs with a minor allele frequency <5%, or were out of Hardy-Weinberg equilibrium in controls ( $p < 1 \times 10^{-7}$ ). A total of 314,664 markers passed the QC in the MSKCC data and were used for combining data from various sources. Similar quality control steps with the same parameters were performed on each of the additional control datasets independently. The data sets were then merged using PLINK, restricting analysis to a set of SNPs common to all datasets. We calculated genome-wide identity by descent (IBD) using PLINK (--genome) and 70 individuals with excessive IBD ( $\pi\text{-hat} > 0.4$ ) were removed from our analysis. After these steps, we applied the same thresholds for missing data, minor allele frequency, and Hardy-Weinberg equilibrium as before. We also removed 529 SNPs that showed a significant difference in rates of missing genotype calls between cases and controls ( $p < 1 \times 10^{-7}$ ) and a further 723 markers that show differential missingness ( $p < 1 \times 10^{-7}$ ) between males and females. A test for differences in missingness based on local haplotype also did not reveal any SNPs with strong evidence for differential missingness based on inferred genotype at the SNP (--test-mishap in PLINK;  $p < 1 \times 10^{-7}$ ). We compared allele frequencies and call rates between MSKCC study samples obtained from different DNA sources (buccal, saliva, or blood) and did not find any markers showing different missingness rates or genotype frequencies due to difference in DNA source ( $p < 1 \times 10^{-7}$ ).

## **Principal components analysis**

To perform principal components analysis to adjust for population substructure, we used the EIGENSTRAT software from the EIGENSOFT 2.0 package<sup>144</sup>. We first filtered the data by removing markers in high linkage disequilibrium (LD). This gave us a set of 32,619 SNPs for which pairwise  $r^2$  values within a window of 50 SNPs are all less than a specified threshold (usually 0.1; --indep-pairwise 50 5 0.1 command in PLINK). This set of markers was then used as input for EIGENSTRAT. Principal components were computed and outliers removed using default parameters. Significant principal components were determined using the Tracy-Widom statistic ( $p < 0.05$ ).

## **Additional quality control by control group comparisons**

To perform additional QC to reduce false positive findings, we tested for genotype frequency differences between each control group versus the rest of the controls. For each control group, we adjusted for the top 11 principal components and used logistic regression to test for differences in genotype frequency versus the other control groups. For the MSKCC controls, we identified 2702 SNPs that show a significant difference in genotype frequencies ( $p < 0.01$ ; Supplementary Figure 1); these SNPs were removed from further analysis. For the other control groups, we identified an additional 15 SNPs that showed significant deviation in genotype frequency in at least one control group ( $p < 1 \times 10^{-7}$ ; Supplementary Figure 1). Notably, we found that the 211 controls from the Study of Irish Amyotrophic Lateral Sclerosis (SIALS; phs000127v1) show a strong deviation from the null hypothesis on a quantile-quantile plot. Therefore,

we chose to remove these 211 controls from the final analysis. This resulted in a final dataset of 263 cases and 5416 total controls at 267,109 markers.

### **Association analysis and estimation of $\lambda$**

To test for association between disease phenotype and SNPs, we used logistic regression, as implemented in PLINK. When we do not consider population substructure, logistic regression is used without covariate adjustment; otherwise, significant principal components were used as covariates to adjust for population substructure.

We used PLINK's estimate for the genomic control parameter  $\lambda$ , which is a measure of test statistic inflation due to effects such as population stratification. PLINK reports  $\lambda$  (based on median  $\chi^2$ ) in the .log file. To test control:case ratios of 1:1, 5:1, 10:1, and 20:1, we selected appropriate subsets of the additional controls to add to the MSKCC case/control dataset.

### **TaqMan genotyping assay**

All MSKCC DNA samples were first amplified using the Illustra GenomiPhi v2 DNA Amplification Kit (GE Healthcare), following manufacturer's recommendations. The reaction was then diluted by adding 120  $\mu$ L of reduced TE buffer. Prior to use in genotyping, we performed an additional 2-fold dilution to improve assay performance. One SNP, rs2236479, was genotyped using the TaqMan allelic discrimination genotyping assay (Applied Biosystems). Genotyping was conducted according to the manufacturer's

instructions as follows: A master mix consisting of 1.375  $\mu\text{L}$  water, 2.5  $\mu\text{L}$  2X TaqMan master mix, and 0.125  $\mu\text{L}$  SNP assay (probe + primers) for each individual was prepared. 4  $\mu\text{L}$  were aliquoted into each well of a 384 well plate, and 1  $\mu\text{L}$  of amplified and diluted DNA was added. PCR was performed in an ABI Gene Amp 9700 machine under the following conditions: 95°C for 10 min followed by 48 cycles of 92°C for 15 s and 60°C for 1 min. Plates were read on an ABI Prism 7900HT fast real time PCR system, and genotype calling was performed using the ABI Sequence Detection System software version 2.3. The genotype concordance rate was computed using 346 individuals who were genotyped both with TaqMan and on the Illumina arrays.

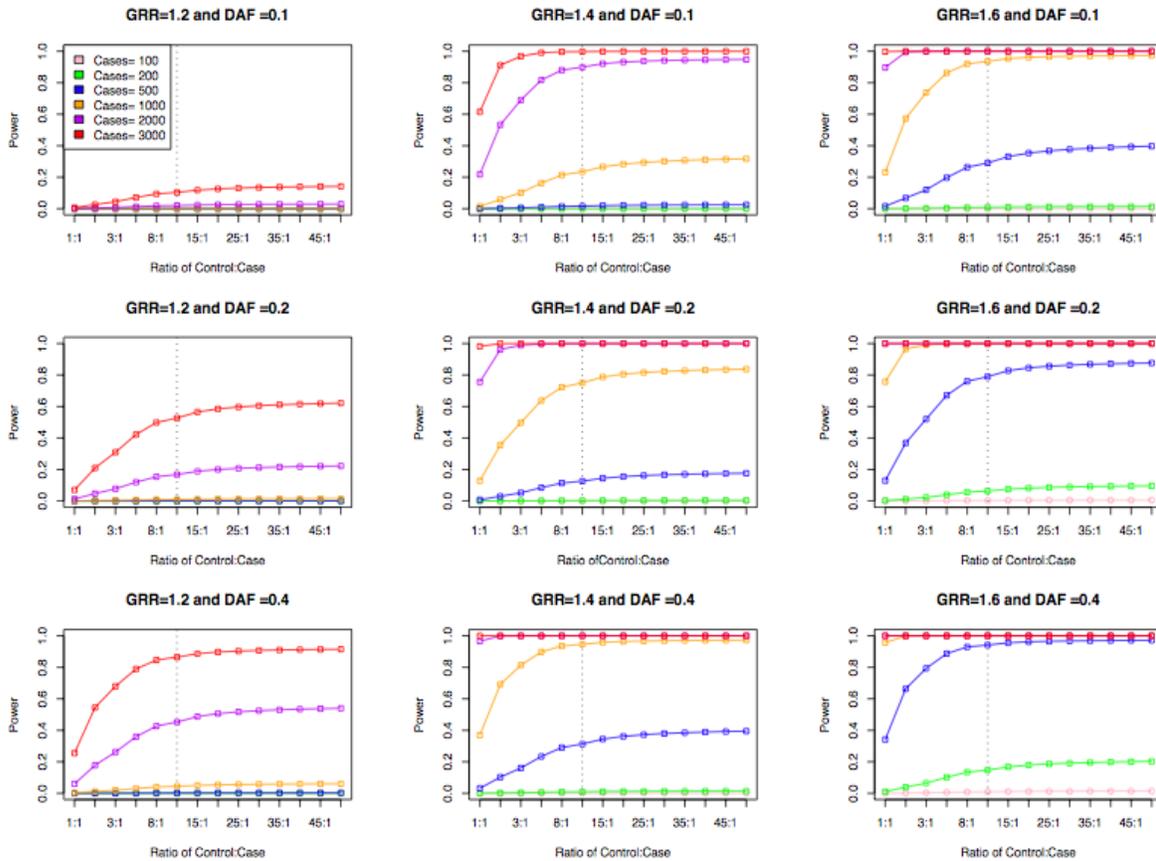
## 1.3 Results

### Analytical power

The large number of control individuals currently available in dbGaP and other databases raises the question of limiting returns. In other words, at what point is the improved power obtained through additional controls small enough that it is no longer worth adding controls? We therefore investigated the shape of the curve of power as a function of control: case ratio with a constant number of cases. As expected, the power increases with increasing number of cases, genotype relative risk and disease allele frequency. The maximum power is achieved when the control: case ratio increases to 10:1; beyond that, the power plateaus (Figure 1). For example, at a genotype relative risk of 1.6, a disease allele frequency of 20%, and significance level of  $10^{-7}$ , little increase in power is observed after a control: case ratio of 10:1. Therefore, we consider a 10:1 control: case ratio ideal for using additional controls in a GWAS.

### Figure 1 Analytical power of GWAS

All power calculations assume an additive model and significance level of  $\alpha=1 \times 10^{-7}$ . The power computed using genotype relative risk (GRR) of 1.2, 1.4, 1.6 and disease allele frequency (DAF) of 0.1, 0.2, and 0.4 were plotted.



## **Power and type I error rate from simulation studies**

The simulated genotype data for cases and controls were generated using GWASimulator<sup>159</sup> using HapMap individual from European ancestry CEU and TSI phased data from HapMap3<sup>57</sup>. Total of 500 cases and 5000 controls were simulated to generation ratio of case: control as 1:10 with no stratification or strong population stratification as described in method. The power and type I error rate were computed for the three categories of disease loci -1) undifferentiated markers, 2) CEU high markers and 3) TSI high markers as describe in method. In Table 1 for 500 cases and 500 controls with no population stratification, there was very low power (0.34) with nominal error rate. When ratio of case: control was increased to 1:10 ratio, the power increased from 0.34 to 0.87. The presence of population stratification caused an increase in type I error rate that could be successfully corrected by PCA based correction method as described in methods, even though type I error rate did not reach the level when no population stratification exist. Thus, our simulation studies motivated our desire to combine genotype data of healthy individuals from public database as common controls with data from case individuals ascertained at Memorial Sloan-Kettering Cancer Center in New York.

**Table 1 Empirical power using GWA simulation study**

The empirical power calculation based on simulated cases and controls using GWAsimulator. The population stratification was created by using CEU and TSI HapMap 3 phased data as input to GWAsimulator.

<b>Undifferentiated SNPs - with same minor allele frequencies in CEU and TSI population</b>				
<b>Case- controls parameter</b>	<b>Ancestral difference level</b>	<b>Satistical Method</b>	<b>Power</b>	<b>Type I error rate</b>
500 cases and 500 controls (1:1)	No Stratification	Association	0.39	6.90E-06
500 cases and 5000 controls (1:10)	No stratification (CEU : TSI controls = 4:1)	Association	0.87	1.28E-03
		PCA corrected Logistic regressio	0.8	1.10E-03
500 cases and 5000 controls (1:10)	Population Stratification (CEU : TSI controls = 99:1)	Association	0.89	1.23E-01
		PCA corrected Logistic regressio	0.79	1.00E-03
<b>CEU high SNPs - SNPs with minor allele frequencies in CEU is higher than TSI</b>				
<b>Case- controls parameter</b>	<b>Ancestral difference level</b>	<b>Satistical Method</b>	<b>Power</b>	<b>Type I error rate</b>
500 cases and 500 controls (1:1)	No Stratification	Association	0.29	4.88E-06
500 cases and 5000 controls (1:10)	No stratification (CEU : TSI controls = 4:1)	Association	0.586	1.26E-03
		PCA corrected Logistic regressio	0.58	1.10E-03
500 cases and 5000 controls (1:10)	Population Stratification (CEU : TSI controls = 99:1)	Association	0.62	1.23E-01
		PCA corrected Logistic regressio	0.55	9.70E-04
<b>TSI high SNPs - SNPs with minor allele frequencies in TSI is higher than CEU</b>				
<b>Case- controls parameter</b>	<b>Ancestral difference level</b>	<b>Satistical Method</b>	<b>Power</b>	<b>Type I error rate</b>
500 cases and 500 controls (1:1)	No Stratification	Association	0.23	4.36E-06
500 cases and 5000 controls (1:10)	No stratification (CEU : TSI controls = 4:1)	Association	0.97	1.43E-03
		PCA corrected Logistic regressio	0.9	1.20E-03
500 cases and 5000 controls (1:10)	Population Stratification (CEU : TSI controls = 99:1)	Association	0.98	1.23E-01
		PCA corrected Logistic regressio	0.87	1.00E-03

## Population stratification in New York based data

We were concerned that population stratification could become a significant problem in a study with controls from public data source, even if we restrict our analysis to self-identified "white" individuals, because of subtle genetic differences among different European populations<sup>156,160,161</sup>. The history of immigration to the United States suggests that a larger proportion of white Americans of Ashkenazi Jewish or southern European (*e.g.* Italian) ancestry would be found in the New York metropolitan area compared to the country as a whole. If this were the case, combining additional controls with our New York-based population would result in the detection of alleles that mark geographic ancestry within Europe rather than disease risk. To investigate whether this concern was well-founded, we performed principal component analysis (PCA) on 263 cases and 202 controls from the MSKCC pancreatic cancer study combined with 5416 individuals selected as additional controls from 6 different studies available in dbGaP (Table 2). When we examine the first and third principal components in our samples from New York, we observe many individuals along a single gradient which has been previously suggested to represent a cline extending from northwest to southeast Europe<sup>162</sup> (Figure 2). The separate cluster of individuals has been previously suggested to be individuals of Ashkenazi Jewish ancestry; all participants in our study who self-identified as Ashkenazi Jewish cluster in this group, supporting the contention that this cluster represents the Ashkenazi Jewish population (Figure 2). When we compared this PCA plot with one for the controls from dbGaP, we observe marked differences in the distribution of individuals on the plot, suggesting a different distribution of geographic

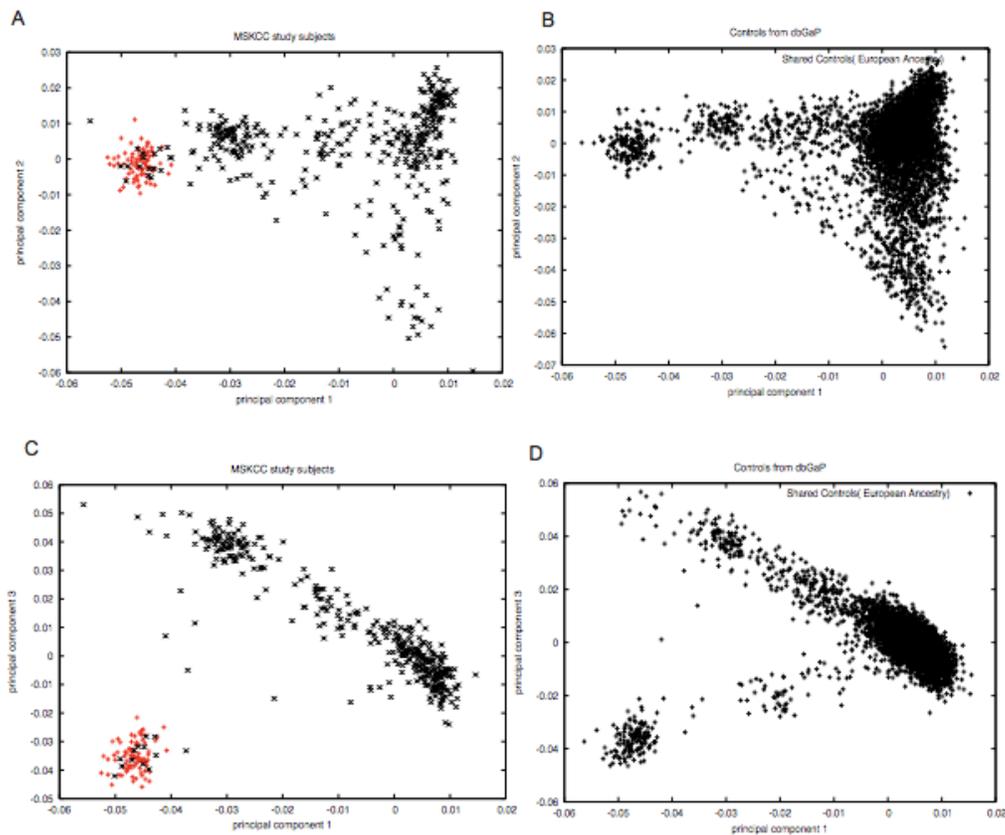
ancestry within Europe. Notably, 18% of the individuals in our study cluster in the “Ashkenazi Jewish” group, compared with 1.7% in the dbGaP control group. These differences could potentially lead to high test statistic inflation when cases and additional controls are analyzed together. Therefore, we conclude that population stratification may be a serious issue when using additional controls with a New York-based case dataset and must be addressed.

**Table 2 Controls from dbGaP used in the present study**

<b>Abbreviation</b>	<b>Study</b>	<b>Number of controls</b>	<b>dbGaP accession number</b>	<b>Reference</b>
SAGE	Study of Addiction: Genetics and Environment	1285	phs000092v1	
CGEMS Breast Cancer	CGEMS Breast Cancer GWAS - Stage 1 - NHS	1142	phs000147v1	98
CGEMS Prostate Cancer	CGEMS Prostate Cancer GWAS - Stage 1 - PLCO	1148	phs000207v1	104
CIDR PD	CIDR: Genome Wide Association Study in Familial Parkinson Disease	863	phs000126v1	
SIALS	Study of Irish Amyotrophic Lateral Sclerosis	211	phs000127v1	[26]
A Genome Wide Scan of Lung Cancer and Smoking	A Genome Wide Scan of Lung Cancer and Smoking	844	phs000093v2	163

## Figure 2 Population substructure of MSKCC pancreatic cancer cases and additional controls

Principal components were computed for the MSKCC and additional control samples combined, and plotted separately. (A, C) Principal components of the 263 cases and 202 controls from the MSKCC (New York) pancreatic cancer study. The first principal component is plotted against the second (A) or third (C). Individuals in red self-identified as Ashkenazi Jewish in the study questionnaire. (B, D) Principal components of the additional controls from dbGaP. The first principal component is plotted against the second (B) and third (D) principal components.

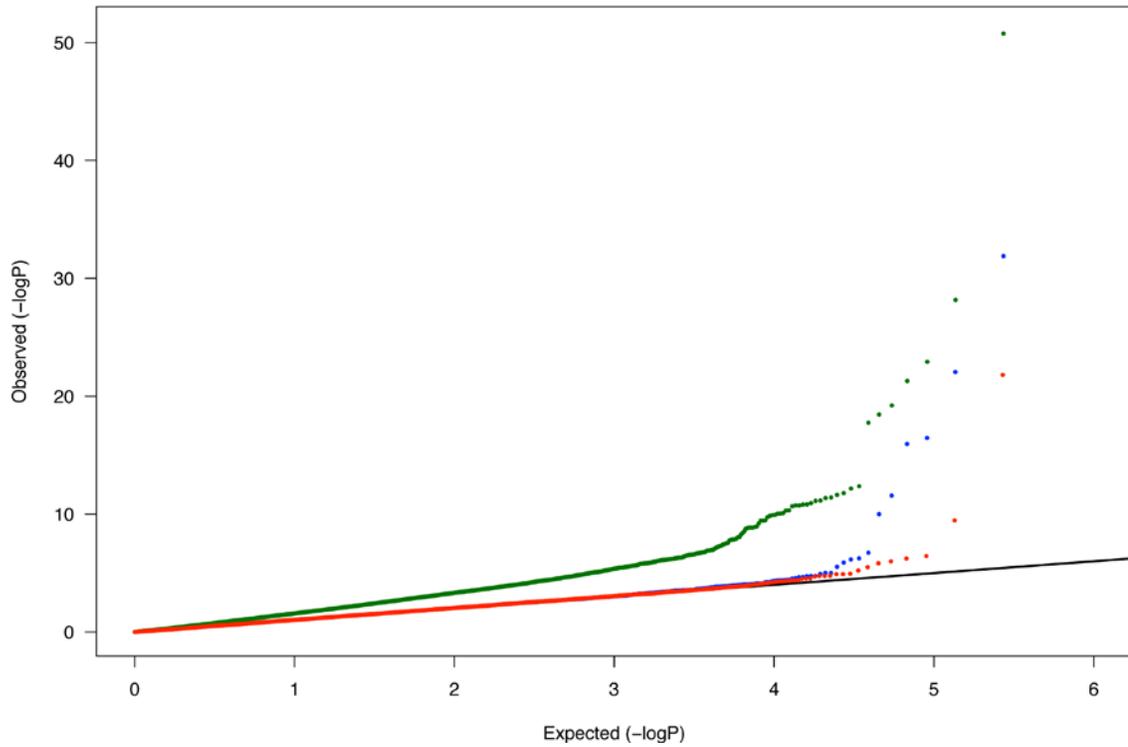


## PCA based correction method using additional controls

We next asked if stratification between our New York-based case dataset and controls from dbGaP results in false positives and if PCA can properly correct for it. We limited the data to those SNPs in common among all studies. As all studies were conducted using the Illumina platform, there were 272,796 overlapping SNPs. The full dataset results in a control:case ratio of 20:1, twice as much as we would recommend based on the analytical power calculations. Using an independent set of markers (all pairwise LD  $r^2 < 0.1$ ), we determined the significant principal components using EIGENSTRAT<sup>144</sup>. The top principal components were used as covariates in a logistic regression model. As can be seen on the quantile-quantile plot, there is an immense inflation of the test statistic when we do not correct for population structure; we interpret this to be due to stratification rather than any true positive finding (Figure 3). When we correct for population structure by adjusting for the top 21 eigenvectors, the quantile-quantile plot follows the distribution expected for the null hypothesis much more closely (Figure 3), even though there is a little inflation near the tail. Therefore simple adjustment for principal components can largely correct for population stratification introduced when using additional controls.

**Figure 3 Quantile -quantile plot of GWAS of pancreatic cancer cases with additional controls**

At a 20:1 control:case ratio, this plot compares the association statistics without any population stratification correction (green), after correction with principal components analysis (red), or with both PCA and removal of SNPs that show differences between the MSKCC controls and additional controls (blue). The black line shows the expected result under the null hypothesis of no association.



### **Additional quality control through comparison of control groups**

The presence of six SNPs at the genome-wide significance threshold of  $10^{-7}$  concerned us as such highly significant associations should have been found in the previously reported pancreatic cancer GWAS. When we examined the previously reported GWAS of pancreatic cancer in dbGaP, none of these six SNPs were significant (all  $p > 0.05$ ) (Table 3). This failure to replicate raises the possibility that the significant results in our study may represent false positives even after following QC steps used in regular case-control GWAS. We next asked if SNPs that lead to false positives could be detected by comparing the MSKCC controls with the additional controls from dbGaP using logistic regression. The quantile-quantile plot of this comparison shows no inflation of test statistics when correcting for 11 principal components (genomic inflation factor  $\lambda = 1.01$ ). Five out of six potential false positive SNPs showed a nominally significant difference ( $p < 0.01$ ) in allele frequency between control groups (Table 3). We then examined the normalized intensity plots for the sixth SNP, rs1975920, in the data we generated (Figure 4). While the plot shows distinct clusters, we noticed that this SNP was monomorphic in the samples we genotyped on the Illumina CNV370-Quad array, while it was polymorphic in the larger number of samples genotyped using the Illumina CNV370-Duo array. As only 20 controls were genotyped using the Illumina CNV370-Quad array, we were not able to detect this artifact through the control group comparison. However, 84 out of 263 cases were genotyped on the CNV370-Quad, presumably driving the signal seen in the case-control analysis. Thus, we introduce an additional QC step by removing 2863 SNPs that show significant difference ( $p < 0.01$ ) in allele frequencies between MSKCC controls group and additional controls. We extended this analysis to the

other control groups, comparing each group with all other control groups. We excluded 15 markers with significant differences in genotype frequency ( $p < 1 \times 10^{-7}$ ). We also visually inspected the quantile-quantile plot of each test for excess test statistic inflation. Notably, we found that the 211 controls from the Study of Irish Amyotrophic Lateral Sclerosis (SIALS; phs000127v1) show deviation from the null hypothesis in the Q-Q plot. Thus, we removed these 211 controls from the final analysis. We reanalyzed 263 pancreatic cancer cases with 5416 additional controls after performing this additional QC step and found that most of the SNPs with extremely low p-value were removed except one (rs2236479). We genotyped rs2236479 in our cohort using a different technology (TaqMan). The concordance rate between the two technologies (TaqMan and Illumina) for rs2236479 was 85%, suggesting that false positives may still be present due to genotyping error. Therefore, we conclude that careful quality control by using a small control group genotyped simultaneously with cases can effectively reduce false positive findings when using additional controls by identifying SNPs that show different genotype frequencies between control groups.

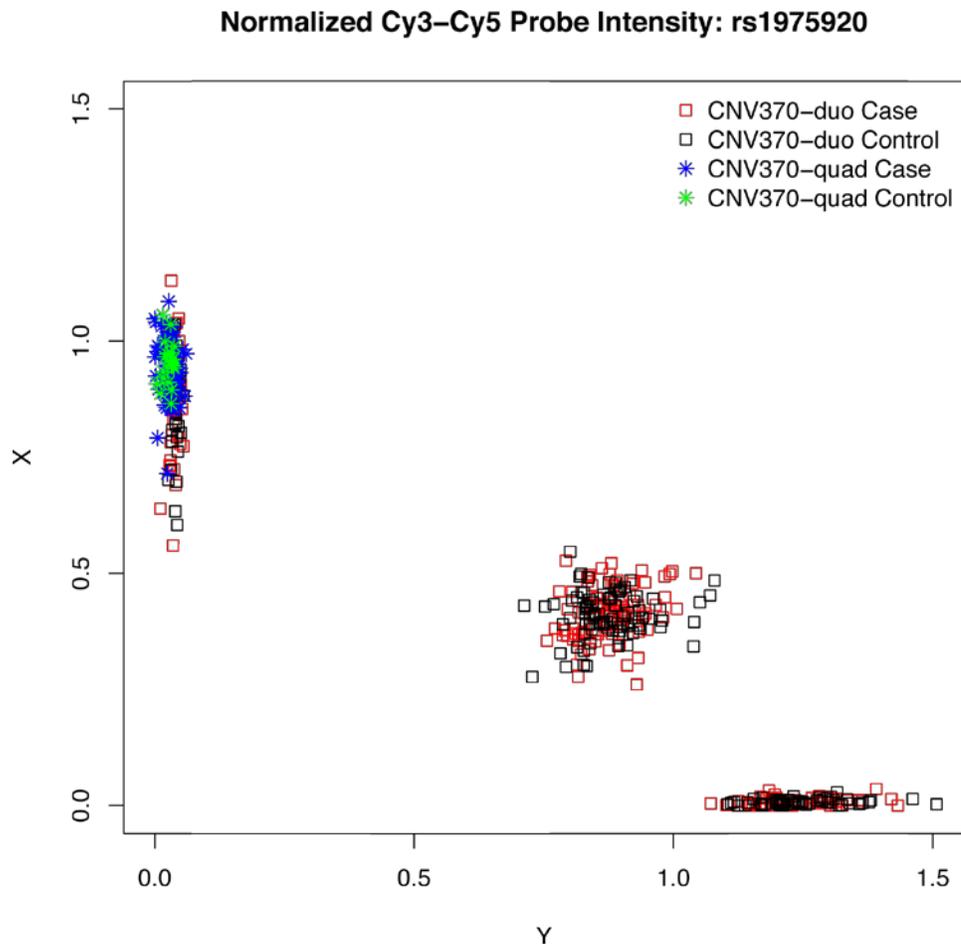
**Table 3 SNPs associated with pancreatic cancer**

SNPs associated with pancreatic cancer at genome-wide significance ( $p < 1 \times 10^{-7}$ ) before additional quality control. All additional controls (control:case = 20:1) were used. Differential missingness is measured by a test for differences in the missing data frequency between the two groups ( $p$ -value). The PanScan analysis  $p$ -value is obtained from published data. The control versus control analysis compared MSKCC controls with additional controls, correcting for population structure. Chr.=Chromosome.

SNP	Chr.	Analysis using additional controls ( $p$ )	Differential missingness ( $p$ )	PanScan analysis ( $p$ )	Additional controls vs. MSKCC controls ( $p$ )
rs7503953	17	$2.7 \times 10^{-12}$	$7.8 \times 10^{-5}$	0.5273	$8.2 \times 10^{-5}$
rs2236479	21	$8.9 \times 10^{-23}$	0.08729	0.7827	0.003
rs1975920	12	$1 \times 10^{-10}$	0.448	0.5081	0.55
rs1455311	4	$1.3 \times 10^{-32}$	1	0.2184	$3.5 \times 10^{-15}$
rs1810636	20	$3.4 \times 10^{-17}$	1	0.4524	$1.5 \times 10^{-5}$
rs1447826	3	$1.1 \times 10^{-16}$	1	0.2049	0.0014

**Figure 4 Normalized signal intensity plot for rs1975920**

The normalized signal intensity for different SNP chips (Illumina CNV370-duo and CNV370-quad) used in our study



## **Effect of data source on inflation factor**

We next analyzed how test statistic inflation is influenced by the number and choice of sets of additional controls. We used the genomic control parameter  $\lambda$  as an estimate of the test statistic inflation<sup>163</sup>. We measured  $\lambda$  in both the original case-control dataset (no additional controls) and with the addition of various additional controls from dbGaP. We observe that  $\lambda$  is near 1 when no additional controls are used (Table 4), indicative of no test statistic inflation. As the control:case ratio is increased by adding data from different sources,  $\lambda$  increases, suggesting the existence of population stratification and/or other technical artifacts. In this analysis,  $\lambda$  is maximal at 1.81 when data from all six different studies are added for a control:case ratio of 20:1 (Table 4). When all significant principal components from PCA were used to correct for population stratification,  $\lambda$  reduces to nearly 1 (range 1.01-1.03; Table 4). Thus, as expected from our quantile-quantile plot analysis, PCA based correction can properly account for the population stratification that results when using additional controls.

**Table 4 Genomic inflation factor for analysis with various datasets**

Control: case Ratio	Controls used	Number of controls	Significant PCs		
				Without PCA Correction	With PCA Correction
1:01	MSKCC pancreatic cancer study controls	202	3	1.009	1.005
5:01	SAGE , MSKCC pancreatic cancer study controls	1488	5	1.5	1.014
5:01	CGEMS Breast Cancer, MSKCC pancreatic cancer study controls	1344	6	1.52	1.018
5:01	CGEMS Prostate Cancer, MSKCC pancreatic cancer study controls	1350	5	1.64	1.019
5:01	CIDR PD, MSKCC pancreatic cancer study controls	1276	5	1.53	1.008
10:01	SAGE, A Genome Wide Scan of Lung Cancer and Smoking , SIALS, MSKCC pancreatic cancer study controls	2522	7	1.71	1.015
20:01	SAGE, A Genome Wide Scan of Lung Cancer and Smoking, CIDR PD, SIALS, CGEMS Breast Cancer CGEMS Prostate Cancer, MSKCC pancreatic cancer study controls	5628	20	1.81	1.03

## Performance of known pancreatic cancer associated SNPs

We next turned to the question of whether the use of additional controls in GWAS will enable new discoveries. To investigate this question, we asked whether we would have been able to discover the four recently reported pancreatic cancer susceptibility SNPs in our data combined with additional controls. We asked what rank and  $p$ -value are observed for each of these four SNPs both in our original cohort and as we add more additional controls. Theoretically, the power to detect each of these SNPs doubles as the control:case ratio increases from 1:1 to 20:1 (Table 5). We found that rank and  $p$ -value of the four pancreatic cancer associated SNPs improved after adding additional controls in a manner that appears to correlate with the computed power. There is a two-fold increase in power for each of the four SNPs when the control: case ratio is increased from 1:1 to 20:1. SNP rs9543325 has the highest increase in power and largest improvement in rank and  $p$ -value. There is some fluctuation in rank and  $p$ -value for all four SNPs when we compare control:case ratios of 10:1 and 20:1. We assume this is due to sampling variability rather than a difference in power as power plateaus out beyond a 10:1 control:case ratio. These results demonstrate that using additional controls in GWAS can help bring true positive hits towards the top of the list, though in this case none of the true positives reached genome-wide significance. These powers should be compared to the power of the original PanScan study, which had 99% power to detect these 4 SNPs at  $\alpha=0.05$ , and reasonable power at  $\alpha=10^{-7}$ , suggesting that our inability to find these true positive at genome-wide significance was to be expected.

We also asked if, for a given number of additional controls, the choice of dataset(s) from which the additional controls are taken influences our ability to detect association with these four SNPs. Using additional controls from four different studies of approximately equal size, we asked what rank and  $p$ -value are observed for each of the four known pancreatic cancer risk SNPs. We observed variability in both the rank and  $p$ -value for each of these four SNPs depending on the choice of control samples. As no control group is consistently the best for all four SNPs, we attribute this variability to sampling variation rather than intrinsic factors in any of the control groups (Table 6).

**Table 5 Rank and p-value of four pancreatic cancer-associated SNPs**

This analysis is done with varying number of additional controls. Correction for population stratification is performed in all analyses. Analytical power is computed assuming an additive model with  $\alpha=0.05$ .

SNP		Control:case ratio			
		Control:case ratio			
		1:01	5:01	10:01	20:01
rs 505922	Rank	105668	6769	5302	216
1.2	<i>p</i> -value	0.393	0.02	0.01	0.0007
0.358	Power	0.2	0.33	0.349	0.364
rs9543325	Rank	477	21	72	52
1.26	<i>p</i> -value	0.0019	$8,2 \times 10^{-5}$	$2,5 \times 10^{-4}$	$1,6 \times 10^{-4}$
0.317	Power	0.29	0.48	0.5	0.53
rs3790844	Rank	102024	7645	1977	1357
0.77	<i>p</i> -value	0.38	0.02	0.007	0.004
0.21	Power	0.265	0.49	0.51	0.53
rs401681	Rank	265649	239819	152561	157875
1.19	<i>p</i> -value	0.99	0.91	0.57	0.58
0.434	Power	0.198	0.313	0.32	0.347

**Table 6 Effect of choice of controls on association statistics for known pancreatic cancer risk SNPs.**

Analytical power is computed assuming a additive model with  $\alpha= 0.05$ .

SNP		Control data sets			
Odds Ratio		SAGE	CGEMS Prostate Cancer	CGEMS Breast Cancer	CIDR PD and SIALS
Minor Allele Frequency	Number of controls	1487	1350	1344	1065
rs505922	Rank	6769	2866	1131	481
1.2	<i>p</i> -value	0.02	0.01	0.004	0.0018
0.358	Power	0.333	0.328	0.32	0.315
rs9543325	Rank	21	101	133	445
1.26	<i>p</i> -value	$8.2 \times 10^{-5}$	0.0004	0.0004	0.001
0.317	Power	0.483	0.477	0.476	0.459
rs3790844	Rank	7645	84087	20488	92396
0.77	<i>p</i> -value	0.02	0.31	0.075	0.34
0.21	Power	0.49	0.491	0.491	0.476
rs401681	Rank	239819	244531	77059	173589
1.19	<i>p</i> -value	0.91	0.94	0.28	0.64
0.434	Power	0.313	0.308	0.308	0.297

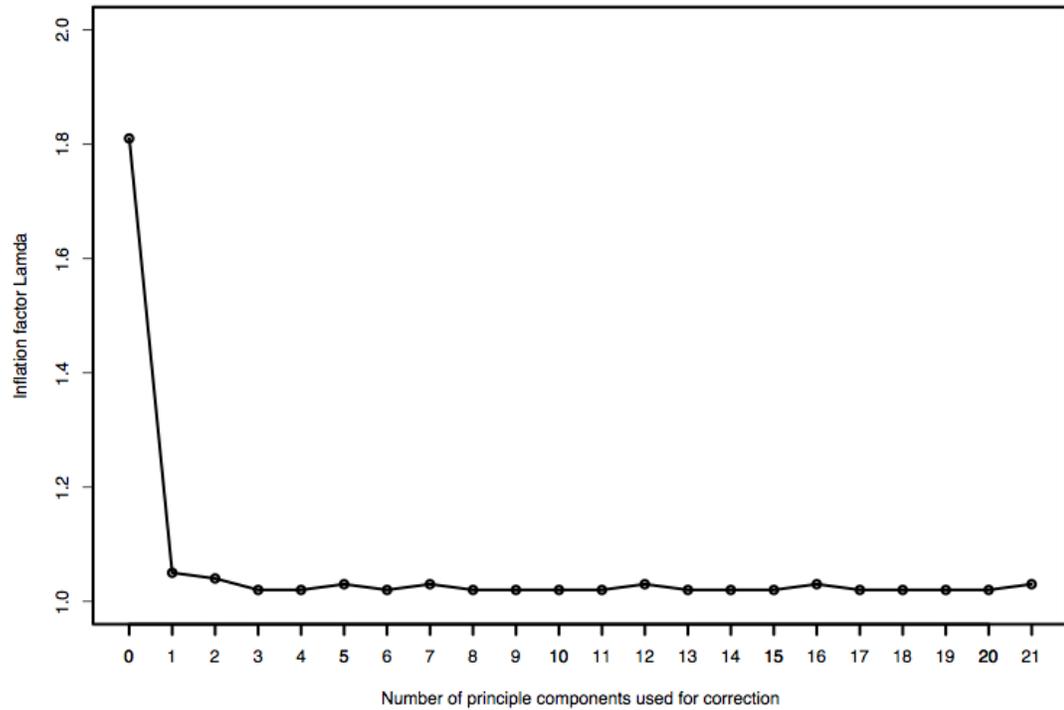
## Number of significant principal components

One choice that must be made is how many principal components are included as covariates in the model. If one simply asks which principal components are significant using Tracy-Widom statistics<sup>144</sup>, the number of covariates to use increases as additional sources of control individuals are added (Table 4). For instance, in our example with a 20:1 control:case ratio there are 21 significant principal components to include. To ask whether these many covariates are necessary, we varied the number of top principal components used as covariates and measured test statistic inflation using the genomic control parameter  $\lambda$  (Figure 5). We find that  $\lambda$  decreases drastically with the first principal component and decreases somewhat more as the next three are added (Figure 5). While this suggests that all 21 principal components are not needed as covariates, it does not tell us whether including extra principal components as covariates decreases the power of the test. When we examine the 4 known pancreatic risk SNPs, we find that the ranks of the 4 SNPs do not change dramatically as more principal components are added after the first few (Table 7). This suggests that while only 4 principal components may be needed in this situation to correct for population stratification, the risk of decreased power through adding additional principal component covariates is minimal. To address the question of what these 21 significant principal components may represent, we first asked if any of the PCs appear to associate with membership in specific studies. Visual inspection of plots of two principal components at a time, with studies color-coded, does not reveal any striking correlation between principal components and study membership. Regression analysis revealed that only the top 4 principal components, for which we

recommend adjusting in the GWAS, are associated with study membership (data not shown). We next repeated the PCA analysis with a more stringent  $r^2$  threshold for LD-based SNP pruning. When the  $r^2$  threshold for pruning is lowered from 0.1 to 0.05, the number of significant eigenvectors (Tracy-Widom  $p < 0.05$ ) drops from 21 to 11. Therefore, we conclude that using additional controls can increase the power of relatively small GWAS after strict QC steps and properly correcting population stratification.

**Figure 5 Genomic inflation factor lamda versus number of principal components (PCs) used for correction**

There are 21 principal components that are significant using Tacy-Widom test statistics when the control:case ratio is 20:1.



**Table 7 Rank of known pancreatic cancer-associated SNPs**

Analysis done was after correcting for the specified number of principal components (PC). In total, 267,785 markers were analyzed.

<b>Number of PCs for correction</b>	<b>rs9543325</b>	<b>rs505922</b>	<b>rs3790844</b>	<b>rs401681</b>
0	585	103084	103905	264098
1	197	1795	2692	133722
2	162	821	2859	140725
4	76	302	1382	162016
6	77	294	1382	161914
10	65	290	1676	156465
16	56	220	1651	153981
21	52	216	1357	157875

## 1.4 Discussion

In this chapter, we have performed a practical evaluation of using additional controls from publicly available databases to conduct GWAS. This approach can result in improved power by increasing the number of controls without any extra cost of genotyping. By using data from our small pancreatic cancer GWAS, we evaluated this approach through comparison with results from the recently published PanScan GWAS. When we analyzed our pancreatic cancer data with additional controls and properly accounted for population stratification, we found improvement in the rank and  $p$ -value for all four known pancreatic cancer SNPs relative to an analysis of our case-control dataset alone. However, while three of the four SNPs were significantly associated with pancreatic cancer in our analysis with  $p < 0.05$ , these results cannot be considered an independent replication of the PanScan results as a large subset of our cases and controls were included in PanScan.

While statistical theory argues that the power of a GWAS increases as the control: case ratio increases for a fixed number of cases, no clear guidelines exist to determine the maximum number of added controls after which there is little or no added benefit. Using analytical power calculations, we show that power increases rapidly as the control: case ratio moves from 1:1 to 10:1 and then plateaus out. Through our analysis of the pancreatic cancer data, we see improved power with a 20:1 control: case ratio relative to a 10:1 ratio. Based on these data, it appears that when designing a GWAS using additional controls, obtaining at least 10 controls for every case is extremely important, though additional benefit could be had by obtaining up to 20 controls for every case.

It is apparent that the QC steps of GWAS in the context of additional controls obtained from public data sources is different from conducting typical case-control GWAS. Recently, Pluzhnokov al. reported a method to estimate genotyping error from raw signal-intensity data when using GWAS control samples from existing public database<sup>164</sup>. This method can only be used when the raw signal intensity data is available, which is not always the case. As an alternative approach to deal with errors introduced from genotype data with different origins, we propose including some controls to be genotyped along with the cases. By removing SNPs that show different frequencies between our controls and the additional controls, we effectively reduced the false positive findings. We consider this step crucial in controlling false positives, especially when raw intensity data is not available.

Beside genotyping error caused due to different data sources, our results illustrates that population stratification is also a potential problem with additional controls. If there is different underlying genetic ancestry in the populations from which cases and controls are taken, an inflated type I error will result. This is clearly observed in our example, where disproportionately more self-reported white cases from the New York metropolitan area are of southern European or Ashkenazi Jewish ancestry than self-reported white controls from other parts of the U.S. This stratification results in artificially high test statistics if we combine data without correcting for population structure. Using simulation studies, it has been demonstrated that correction for population stratification can be achieved successfully by using various methods like multi-dimensional scaling

(MDS) or principal component analysis. We used the popular PCA software EIGENSTRAT to identify principal components in our data and then corrected for these components in logistic regression. Adjusting for the significant principal components substantially reduces the genomic inflation factor in every additional control dataset we tested.

The proper number of principal components to consider in correcting for population substructure remains unclear. Notably, the number of significant principal components computed using the Tracy-Widom test statistic<sup>144</sup> increased when we increased the control:case ratio by adding data from different sources. With a control:case ratio of 20:1, 21 significant PCs were identified. We explored the effect of including different numbers of principal components in our analysis and found that after 4 principal components are included, no additional benefit is gained by including more principal components. Intriguingly, in a GWAS of Alzheimer's disease, Harold *et al.* similarly found no additional improvement in  $\lambda$  after accounting for 4 principal components. As we found a reduced number of significant principal components upon lowering the  $r^2$  threshold to obtain independent markers for the PCA calculation, we hypothesize that many of the 21 PCs may be picking up local linkage disequilibrium patterns in the data rather than population substructure. Therefore, including these additional principal components is not necessary for the analysis.

We acknowledge that the additional control approach is limited by choice of genotyping platform, as it requires the same SNP to be genotyped in all samples. To

maximize overlap between SNPs, we restricted our analysis to projects that used Illumina chips for genotyping and further restricted analysis to only SNPs in common among all studies. Alternatively, imputation techniques have been used to integrate genotype data from different platforms, though how such an approach will perform when different platforms are used to genotype the cases and controls remains unclear.

Besides these technical issues, there are conceptual limitations to this approach. Using additional controls works best in consideration of genetic effects alone. While in theory gene-environment interaction can be considered if appropriate environmental data is present in dbGaP, in practice this information is often found in only some datasets and details of the collection of this data likely varies between studies.

Based on these results, it appears that using this approach with only several hundred cases to study a disease typical of the common diseases studied with GWAS will result in the true disease loci rising to the top of the list of SNPs but not reaching genome-wide significance. Therefore, we propose that the use of additional controls will work best in the context of a large case/control study. In this context, a subset of cases and controls would be selected for genome-wide genotyping. These data would be combined with additional controls. The top  $10^3$ - $10^4$  SNPs from this analysis would then be genotyped in the full case/control study both to increase power and remove false positives. In other words, additional controls may work best when included in stage 1 of a two-stage GWAS design. Standard downstream analyses including independent replication and fine mapping would then be conducted on SNPs that pass the second

stage. Thus, the use of additional controls is a promising method to increase sample sizes thus the power of the study without additional cost.

## CHAPTER 2

### Genome wide association study of myeloproliferative neoplasms

#### 2.1 Introduction

As discussed in the introduction, PV, ET, and PMF are chronic MPN which are characterized by clonal proliferation of one or more terminally differentiated myeloid elements.<sup>165</sup> The genetic basis for PV, ET, and PMF remained an enigma until 2005, when multiple groups identified a somatic activating mutation in the *JAK2* tyrosine kinase (*JAK2V617F*) in  $\approx 90\%$  of PV and in 50-60% of ET/PMF.

The majority of studies to date have focused on the role of genetic and epigenetic events that are stochastically acquired and selected during MPN pathogenesis, whereas few studies have addressed the role of germline genetic variation in MPN pathogenesis. Two recent studies, however, suggest that germline genetic context is important in these disorders. Pardanani and colleagues recently analyzed 32 single nucleotide polymorphisms (SNPs) in *JAK2*, *EPOR*, and *MPL* in affected tissue (granulocytes) from 179 patients with PV and ET, and identified three *JAK2* SNPs which were enriched in either ET or PV. Although these results suggest there are host genetic variants that influence MPN phenotype, they did not perform a genome-wide analysis for MPN predisposition alleles. More importantly, given the high rate of acquired uniparental disomy at the *JAK2* locus in PV, but not ET, their results were likely influenced by somatic loss of heterozygosity at the *JAK2* locus in the different MPN. It has also been observed that there is familial clustering in MPN cases, and in these kindreds somatic

JAK2V617F and/or JAK2 exon 12 mutations, can be identified in some, but not all, affected family members, suggesting there are inherited loci that predispose to the somatic acquisition of JAK2 mutations. In addition an epidemiologic study of 11,039 MPN cases, 43,550 controls, and 24,577 first degree relatives of MPN patients in Sweeden found that relatives of MPN patients are at  $\approx$  5-7 fold increased risk for the development of MPN, consistent with the existence of one or more MPN predisposition loci. Given these observations, we hypothesized there are unidentified germline loci relevant to MPN pathogenesis, and used genome-wide SNP array data to identify germline predisposition loci relevant to the pathogenesis of PV, ET, and PMF.

## **2.2 Materials and Methods**

### **SNP Array Analysis of MPN Samples**

MPN patient samples were obtained from the Harvard MPD Study patient cohort,<sup>17</sup> and were collected using IRB approved protocols, all patients provided informed consent. DNA was extracted from granulocytes and buccal swabs as previously described,<sup>17</sup> and RNA was extracted from patient cells stored in Trizol. 217 granulocyte DNA samples, including 113 PV patient samples and 68 ET patient samples, were chosen for SNP array analysis based on clonality studies and JAK2V617F mutational burden<sup>166</sup> in order to limit analysis to samples with >80% MPN cells. DNA samples were genotyped using Affymetrix 250K (Sty) arrays, Arrays were scanned with the GeneChip Scanner 3000, and Affymetrix Genotyping Tools Version 2.0 to ascertain genotypes (Affymetrix, Santa Clara, CA).

### **Principal Component Analysis of MPN Patients/Controls**

For principal component analysis we used genome-wide data from the 217 MPN cases and from 3000 controls from the Wellcome Trust Case Control consortium,<sup>167</sup> which were genotyped with the Affymetrix GeneChip 500k Mapping Array Set, of which the 250 K Sty chip is a subset. Before analysis, we performed quality control filtering of both samples and SNP separately for cases and controls and then merged the dataset using the common set of SNPs present in the two cohorts. To do so, we first filtered out the ambiguous SNPs (A/T or G/C alleles) to ensure we unambiguously know strand when

we merge the two datasets. 35218 ambiguous markers (out of 231786) were removed from the MPN genotype dataset, while 77934 ambiguous markers (out of 486661) were removed from the WTCCC control cohort. The quality control filters and quality assessment removed subjects with low genotype completion rates (<90%). Further data cleaning of the autosomal SNPs typed in both datasets retained SNPs that have a minor allele frequency (MAF) >5%, a rate of missing genotype <1%, and are in Hardy-Weinberg equilibrium in the WTCCC controls (exact test  $p > 10^{-7}$ ). In total, 62775 markers were identified for analysis and used in the merged case and control dataset.

To investigate potential population stratification biases that could be introduced by the shared controls we performed principal component analysis using EIGENSTRAT.<sup>144</sup> To reduce the linkage disequilibrium between markers, we first used PLINK to filter markers such that all remaining markers are in low LD ( $r^2 < 0.1$ , calculated in sliding windows 50 SNPs wide, shifted and recalculated every five SNPs). We applied the EIGENSTRAT program with default parameters and no outlier removal to infer axes of variation in the combined dataset. The case and controls that cluster together on the eigenvector plot (with the first two axes of variation) were used for the association analysis.

The main SNP of interest in JAK2, rs10974944, has G and C alleles and was therefore eliminated by our filtering for ambiguous SNPs. To see at what rank it would appear in a GWAS for MPN risk alleles, we included it in our genome-wide association analysis. Specifically, we included the germline genotypes generated using TaqMan for

the cases with the genotypes provided by the WTCCC data for the controls. A test of allelic association was performed using `–assoc` in PLINK.

### **Statistical Analysis**

The frequencies of the genotypes between cases and controls were compared using Pearson's  $X^2$  test, and when required, Fisher's exact test. The ANOVA test was used for comparison of JAK2V617G allele burden between different genotypes. SPSS version 16.0 for Windows (SPSS, Chicago, IL, USA) was used to perform all statistical tests.

### **Genotyping and Expression Analysis**

Granulocyte and buccal DNA samples were genotyped using TaqMan SNP genotyping assays for rs10974944 and rs12500918 (Applied Biosystems, Foster City, CA) assays. DNA samples from CEU HapMap founders were used as controls. Expression of *JAK2* and *HPRT1* were measured using TaqMan Gene Expression Assays (Applied Biosystems). This was done in collaboration with Levine Lab.

### **JAK2 rs10974944/Mutation Clonal Analysis**

A 3Kb PCR product containing rs10974744 and exon 14 of JAK2 was amplified from *JAK2V617F*-positive patients heterozygous for rs10974944 in the germline. PCR products were cloned using the TA cloning kit, and single colonies were sequenced using M13 and T7 primers. Sufficient colonies were sequenced from each patient to ascertain

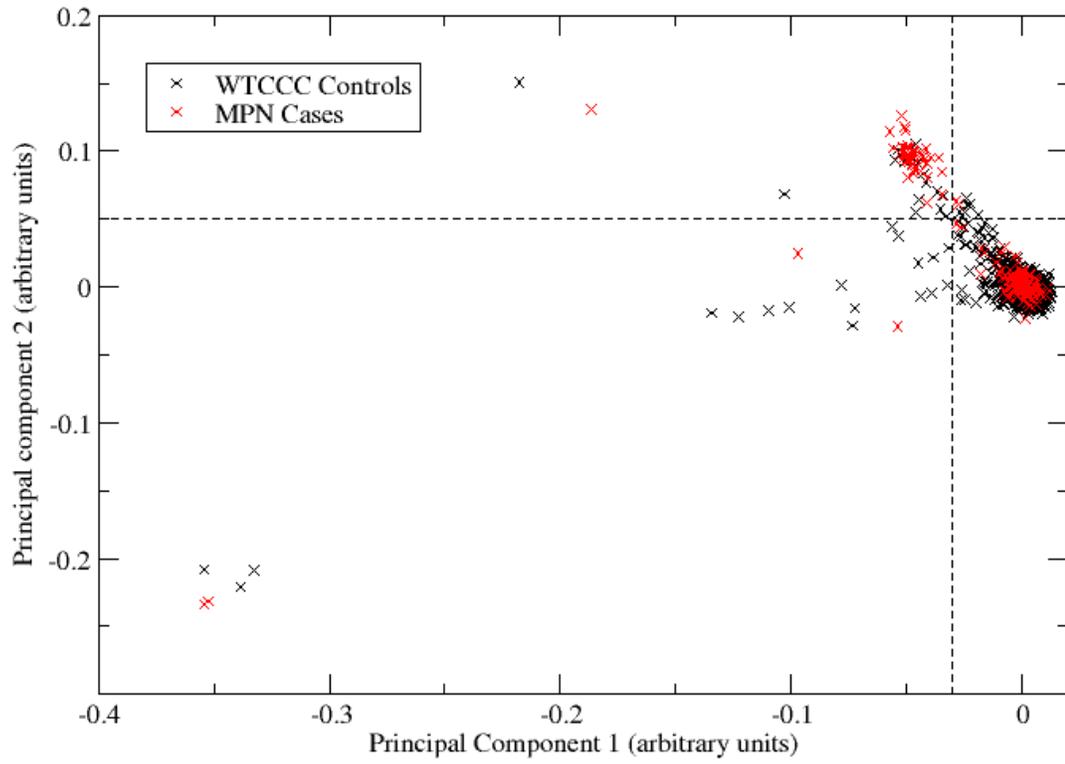
which germline genotype was in *cis* with the V617F allele in granulocyte DNA from each informative patient. This was done in collaboration with Levine Lab (Outi Kilpivaara).

## 2.3 Results

### Case-Control Analysis of Genome-Wide SNP Array Data Identifies JAK2 as a Major MPN Risk Allele

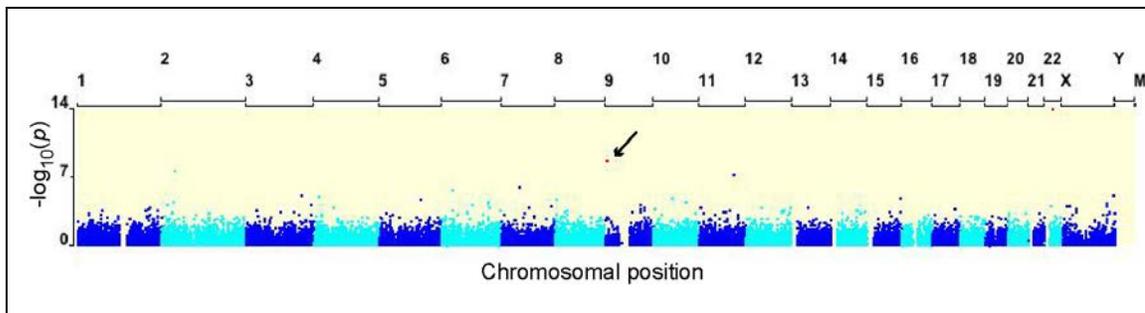
We performed a GWA study to identify genetic variants associated with MPN predisposition. We used the shared controls approach described in Chapter 1 for MPN GWA study. To do so, we combined all unambiguous SNPs genotyped in our MPN samples and in the WTCCC with our data and asked whether allele frequencies differed significantly between the two groups at any of the SNPs. In order to control for population heterogeneity, we used principal component analysis to detect population substructure in the combined cohort of MPN samples and WTCCC controls based on the genome-wide genotyping data. We selected case and control individuals who cluster on a plot of the first two principal components in a region that suggests ancestry from northern and western Europe (**Figure 6**). Four SNPs were significantly associated with MPN risk after correcting for residual population stratification and multiple testing. One of these SNPs is rs10974944, an intronic SNP in JAK2 gene in chromosome 9 that represents a MPN risk SNP (**Figure 7**).

**Figure 6 Principal component analysis of MPN cases and WTCCC controls**



**Figure 7 Genome wide SNP analysis of MPN cases and WTCCC controls**

The arrow marks the position of rs10974944.



## Germline Variation at the JAK2 Locus Influences MPN Predisposition

Genome-wide analysis of SNP array data suggested a SNP at the JAK2 locus (rs10974944) associated with MPN predisposition. However, our SNP array analysis was performed on affected (granulocyte) tissue from MPN patients, and we and others have shown that acquired uniparental disomy leading to homozygosity for the somatic JAK2V617F mutation is more common in PV than in ET<sup>14-17,26</sup>. We therefore compared the frequency of the genotypes at rs10974944 in germline DNA from 324 PV, ET, and PMF patients to published genotypes for WTCC controls (**Table 8**) and observed that the frequency of both the GG and CG genotypes is more common in cases than controls (OR=3.1,  $p=4.1 \times 10^{-20}$ ) (**Table 8**). This is consistent with the G allele at rs10974944 functioning as a dominant allele with effects in either the heterozygous or homozygous state. We observed that the minor allele (GG) was significantly more common in PV than in ET ( $p=0.01$ ). These data suggest JAK2 serves as a MPN predisposition locus, and that germline variation at *JAK2* more strongly influences MPN predisposition than MPN phenotype.

**Table 8A Germline genotype for JAK2 SNP rs10974944 and MPN predisposition**

<b>Rs10974944 Genotype</b>	<b>MPN</b>	<b>WTCCC</b>
GG	70 (21.6%)	195 (6.5%)
CG	161(49.7%)	1139 (38.0%)
CC	92 (28.7%)	1665 (55.5%)
	<b><i>p</i></b>	<b>OR (95% CI)</b>
GG vs. CG+CC	$1.5 \times 10^{-21}$	4.0 (2.9-5.4)
GG/CG vs. CC	$4.1 \times 10^{-20}$	3.1 (2.4-4.0)
GG vs. CC	$5.1 \times 10^{-32}$	6.4 (4.6-9.1)
CG vs. CC	$2.10 \times 10^{-12}$	2.5 (1.9-3.3)

**B. Germline Genotype for JAK2 SNP rs10974944 in MPN cases and Matched WTCC Controls According to Principal Component Analysis**

<b>Rs10974944 Genotype</b>	<b>MPN</b>	<b>WTCCC</b>
GG	18 (21.7%)	195 (6.6%)
CG	49(49.8%)	1121 (37.9%)
CC	26(28.5%)	1646 (55.5%)
	<b><i>p</i></b>	<b>OR (95% CI)</b>
GG vs. CG+CC	$3.5 \times 10^{-07}$	3.7 (2.0-6.4)
GG/CG vs. CC	$1.3 \times 10^{-06}$	3.0 (1.9-5.0)
GG vs. CC	$2.3 \times 10^{-10}$	6.0 (3.0-11.0)
CG vs. CC	0.0021	3.0 (1.5-4.2)

## Germline Variation at JAK2 Specifically Predisposes to the Development of JAK2V617F-Positive MPN

Given that somatic mutations at JAK2 are common in PV, ET, and PMF, we theorized that the effects of germline genetic variation on MPN predisposition might be exclusive to JAK2 mutated MPN. We assessed rs10974944 genotype in 321 MPN cases which had been genotyped for the JAK2V617F allele using a sensitive, allele-specific real time PCR assay able to detect JAK2V617F allele burden >1%,<sup>166</sup> and for JAK2 exon 12 mutations using MALDI-TOF mass spectrophotometric genotyping for all known exon 12 alleles (unpublished data done by Levine lab). We found that allelic variation at rs10974944 was strongly associated with predisposition to JAK2 positive MPN in a dominant genetic model (OR=4.0,  $p=7.7 \times 10^{-22}$ ) (**Table 9**). In contrast, allelic variation at rs10974944 was much less strongly associated with JAK2 negative MPN in a dominant genetic model (OR=1.6,  $p=0.06$ ). We also assessed whether the effects of germline genetic variation on MPN predisposition might vary with MPN phenotype. We found that allelic variation at rs10974944 was strongly associated with predisposition to PV (OR=4.3,  $p < 1.0 \times 10^{-16}$ ) and ET (O.R.=2.1,  $p=6.7 \times 10^{-5}$ ). The stronger relationship between rs10974944 and predisposition to PV is in part due to the higher incidence of JAK2 mutations in PV (95%) compared to ET (65%) in our patient cohort; we observed a higher association between rs10974944 genotype and JAK2 positive ET (O.R.=2.8  $p=2 \times 10^{-5}$ ).

**Table 9 Germline genotype for JAK2 SNP rs10974744 in JAK2V617G -positive MPN cases and negative MPN cases compared with WTCCC**

rs10974944 genotype	JAK2-positive MPN	JAK2-negative MPN	WTCC
<b>GG</b>	60 (24.5%)	10 (13.2%)	187 (6.6%)
<b>CG</b>	127 (51.8%)	32 (42.1%)	1078 (37.9%)
<b>CC</b>	58 (23.7%)	34 (44.7%)	1578 (55.5%)
<b>Total</b>	<b>245</b>	<b>76</b>	<b>2843</b>
<b>JAK2-positive MPN</b>			
	<i>p</i>	<b>OR (95% CI)</b>	
GG vs. CG+CC	8.4x10-24	4.7 (3.4-6.5)	
GG/CG vs. CC	7.7x10-22	4.0 (3.0-5.5)	
GG vs. CC	6.9x10-37	8.8 (6.0-13.1)	
CG vs. CC	7.9x10-14	3.2 (2.3-4.4)	
<b>JAK2-negative MPN</b>			
	<i>P</i>	<b>OR (95% CI)</b>	
GG vs. CG+CC	0.017	2.3(1.1-4.4)	
GG/CG vs. CC	0.06	1.6 (1.0-2.5)	
GG vs. CC	0.008	2.6 (1.3-5.3)	
CG vs. CC	0.21	1.4 (0.8-2.3)	

## **JAK2V617F is Most Commonly Acquired in cis with JAK2 rs10974944**

We then investigated the relationship between germline variation at the *JAK2* locus and MPN risk and the high rate of somatic mutations at this same locus. Analysis of JAK2V617F-positive MPN cases revealed a strong association between germline rs10974944 genotype and JAK2V617F allele burden ( $p < 0.01$ ), and an even stronger association between granulocyte rs10974944 genotype and JAK2V617F allele burden ( $p < 0.00001$ ). We then investigated 42 patients who were heterozygous for rs10974944 in their germline and a somatic JAK2V617F allele burden  $> 50\%$  consistent with emergence of a homozygous JAK2V617F mutant clone. We found in 38 of 42 cases acquisition of a homozygous JAK2V617F mutation was associated with somatic conversion to a homozygous GG genotype at rs10974944, strongly suggesting rs10974944 favors the acquisition of JAK2V617F *in cis* with the MPN risk allele. We then performed long range PCR of a portion of the *JAK2* locus which included both rs10974944 and JAK2V617 on granulocyte DNA from 30 patients who were heterozygous for rs10974944 in their germline, and sequenced  $> 8$  individual clones in order to ascertain the strand on which *JAK2V617F* was acquired (**Figure 8B**).

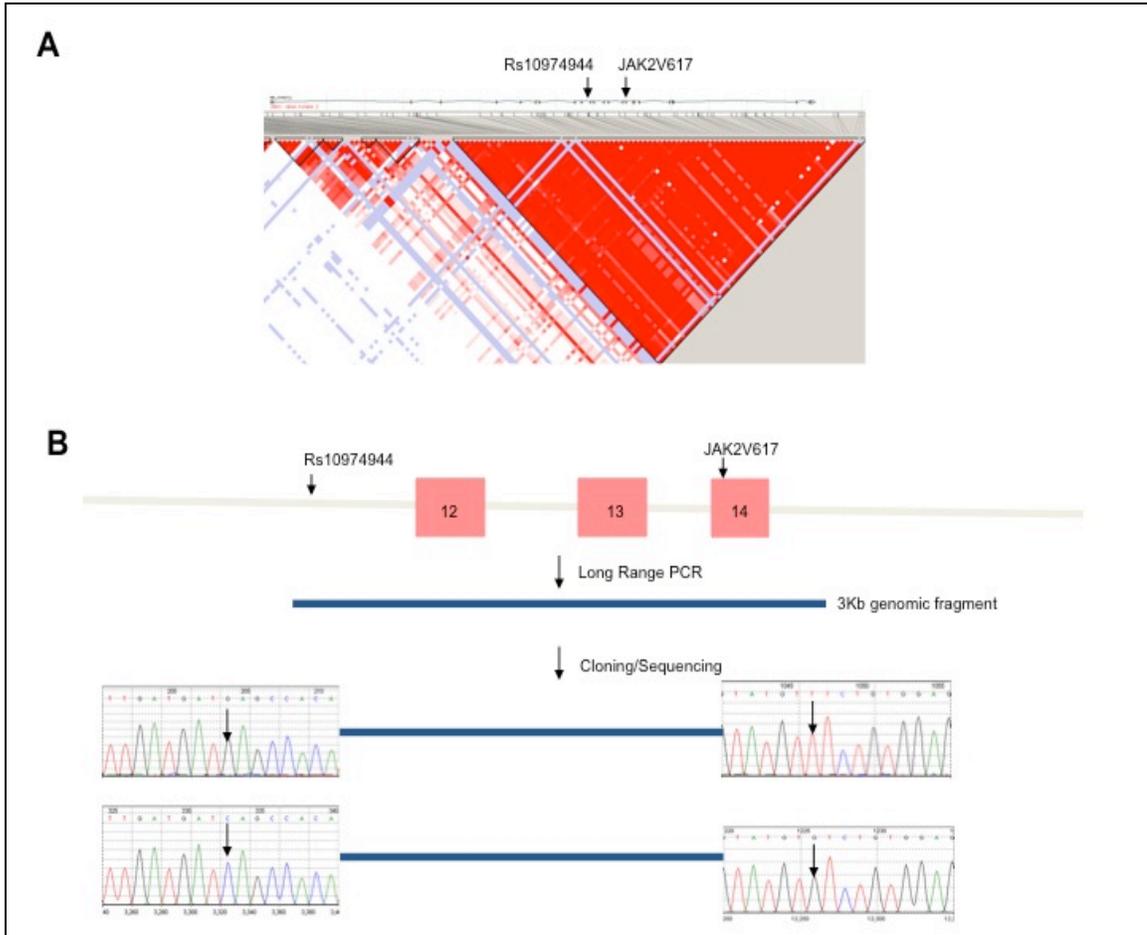
There are many different possibilities which might explain how germline variation at the *JAK2* locus might favor the acquisition of somatic *JAK2* mutations, including allele-specific expression, linkage disequilibrium (LD) with non-synonymous SNPs which alter *JAK2* function or LD with changes in the 3' untranslated region (3'UTR) which affect miRNA binding. We did not observe differences in *JAK2* expression in patients with different rs10974944 genotypes, and sequence analysis of the

entire open reading frame and of the 3'UTR in 48 MPN cases did not reveal genotype-specific non-synonymous sequence alterations or alterations in the 3'UTR. Moreover, haplotype structure of the JAK2 locus in CEPH founders (**Figure 8A**) shows that rs10974944, JAK2 exon 12, and JAK2V617 are contained in a common haplotype block distinct from the promoter and 5' exons of JAK2. These data suggest that rs10974944 favors the acquisition of JAK2 mutations *in cis* with the MPN risk allele by a heretofore-unidentified mechanism.

**Figure 8 JAK2V617F is acquired in cis with JAK2 SNP rs10974944**

**Figure 8A** shows the haplotype structure of the *JAK2* locus in CEPH HapMap founders, showing that Rs10974944 and exon 14 of *JAK2* are in a shared haplotype approximately 3Kb apart.

**Figure 8B** shows the precise location of rs10974944 in relation to exons 12, 13 and 14, and shows the result of long-range PCR and clonal sequence analysis of this 3Kb fragment in a patient who was heterozygous for rs10974944 in their germline and heterozygous for *JAK2V617F* in their affected tissue. Analysis demonstrates in this patient the G allele at rs 10974944 is in cis with the mutant T allele at *JAK2V617*, whereas, the C allele is in cis with the wild-type G allele at *JAK2V617*. The G allele was found to be in cis with the mutant T allele in 27 of 30 *JAK2V617F* positive MPN patients whom were heterozygous for rs10974944, suggesting *JAK2V617F* is almost always acquired in cis with the risk allele at rs10974944 (*experiments performed by Outi Kilpivaara*)



## 2.4 Discussion

The discovery of activating mutations in the JAK-STAT pathway in the majority of patients with PV, ET, and PMF provided important insight into the pathogenesis of these MPN; however, there remain important questions regarding the role of unknown inherited and acquired disease alleles in MPN pathogenesis. Most studies have focused on the identification of additional somatic alleles acquired during MPN pathogenesis; in contrast we searched for germline alleles that contribute to MPN predisposition and/or to MPN phenotypic pleiotropy. Genome-wide analysis allowed us to identify a germline variant in the *JAK2* gene that predisposes to the development of *JAK2*-mutant MPN that are preferentially associated with specific MPN phenotypes.

The observation that a *JAK2* germline haplotype is markedly enriched in MPN cases compared to controls suggests that germline variation at the *JAK2* locus is an important contributor to MPN predisposition. Although genome-wide association studies have identified predisposition loci for a spectrum of human diseases, in most cases the individual loci identified in these studies have a modest effect on disease predisposition. For example, a recent genome-wide association study in chronic lymphocytic leukemia identified six previously unreported CLL risk loci, each of which had an odds ratio less than 1.6 and were estimated to account for less than 3% of the excess familial risk of CLL. In contrast, in a dominant genetic model the GG/CC genotype at *JAK2* rs10974944 contributes significantly to the excess familial risk of MPN (O.R.=3.1, population attributable risk=46.0%). These effects are most evident in *JAK2*-positive MPN

(O.R.=4.0, population attributable risk=55.3%), suggesting that germline variation at JAK2 is a major determinant for the predisposition to develop JAK2-positive MPN.

The observation that germline variation in JAK2 predisposes to somatic activating mutations at the same locus is also of significant importance. We found that somatic JAK2 mutations were most commonly acquired *in cis* with the JAK2 predisposition haplotype, suggesting a direct interaction between haplotype-specific genetic variation in the JAK2 locus and secondary acquisition of somatic mutations on the same strand. We did not observe genotype-specific changes in JAK2 expression, nor did we identify non-synonymous alterations in JAK2 which were in linkage disequilibrium with the JAK2 predisposition SNP, suggesting it is unlikely the JAK2 MPN predisposition allele directly modulates JAK2 expression and/or JAK2 function. We also did not observe genotype-specific alterations in the 3'UTR of JAK2 which could influence miRNA binding, which has been delineated as the mechanism by which an alteration in the 3'UTR of the RAS locus predisposes to the development of non-small cell lung cancer. We hypothesize that genotype-specific genomic variation in the JAK2 haplotype block increases the somatic mutation rate in this locus. Although additional genetic and functional studies are needed to test this hypothesis, there is precedent for germline variation predisposing to somatic alterations at the same locus, including a previous study that found that germline variants in the *APC* gene present in the Ashkenazi Jewish population increase risk for the development of colorectal cancer by creating a hypermutable region in the *APC* gene.<sup>168</sup> It is possible that germline variation in the JAK2 locus may be specifically associated with an increase in the rate of the guanine to thymidine mutation that causes the valine to

phenylalanine substitution at codon 617. Although substitution of tryptophan, methionine, isoleucine, and leucine) for valine at codon 617 results in constitutive JAK2 activation, and alternate activating JAK2 mutations involving codon 683 are observed in Down syndrome associated ALL, JAK2V617F predominates in PV, ET, and PMF. These data suggest that germline context may be important in delineating why distinct mutations in JAK2 are acquired in different neoplasms.

Although this study demonstrates that germline genetic context is important in MPN pathogenesis, it is likely there are additional germline loci important in MPN predisposition and pathogenesis. Our data suggests that germline variation at the JAK2 locus has a minimal contribution to JAK2V617F-negative MPN predisposition. While this data must be interpreted with the caveats that our genome-wide SNP data comes from diseased tissue which may have undergone somatic changes and that we do not have complete coverage of the genome due to the large number of ambiguous SNPs removed, the idea that there are additional MPN susceptibility loci that can be identified through genome-wide association analysis is intriguing. These data suggest that germline variation is an important contributor to MPN phenotype and predisposition, and that additional genome-wide studies will identify additional germline alleles relevant to MPN pathogenesis.

## CHAPTER 3

### Mechanism for JAK2 susceptibility haplotype in MPN

#### 3.1 Introduction

Using the GWAS approach described in Chapter 2, we identified a MPN susceptibility locus (tagged by SNP rs10974944) in the JAK2 gene on chromosome 9<sup>153,169,170</sup>. Interestingly, by analyzing MPN patients with allele-specific PCR, we found that the somatic gain-of-function mutation JAK2V617F was frequently acquired in *cis* with the rs10974944 risk allele. Concurrent with the publication of our findings, two independent investigators also identified a JAK2 haplotype (referred to as “46/1” or the “GGCC” haplotype) as a major risk factor for the development of MPN. Jone et al. demonstrated that both homozygous and heterozygous JAK2V617F -positive disease is preferentially associated with 46/1 and that this haplotype seems to harbor an as-yet-uncharacterized functional variant. They estimated that 46/1 accounts for 50% of the population attributable risk of developing an MPN, but that it does not account for familial MPN<sup>169</sup>. Olcaydu et al. estimated that over 80% of all the JAK2V617F mutations in MPN occur on this specific *JAK2* haplotype<sup>171</sup>. Thus, these studies have demonstrated that the 46/1 JAK2 haplotype predisposes to JAK2V617F-positive MPN.

Although the mechanism underlying this association remains obscure, two hypotheses have been proposed. First, the 46/1 haplotype may be inherently more genetically unstable and acquire the V617F somatic mutation at a faster rate than other

haplotypes (referred to as the “hyper-mutability” hypothesis). As shown in **Figure 9A**, DNA sequence variants can define somatic mutability and could make some haplotypes more susceptible to DNA damage. A difference in mutability between two haplotypes could explain why *JAK2*<sup>V617F</sup> preferentially occurs on the 46/1 haplotype. A second hypothesis suggests that the 46/1 haplotype may carry a functional variant(s) that causes allele-specific activation or regulation of the *JAK2* gene. The V617F somatic mutation may arise on all haplotypes at equal rates, but the 46/1 haplotype may confer selective advantage to the V617F-positive clone (referred to as the “activation” hypothesis – **Figure 9B**).

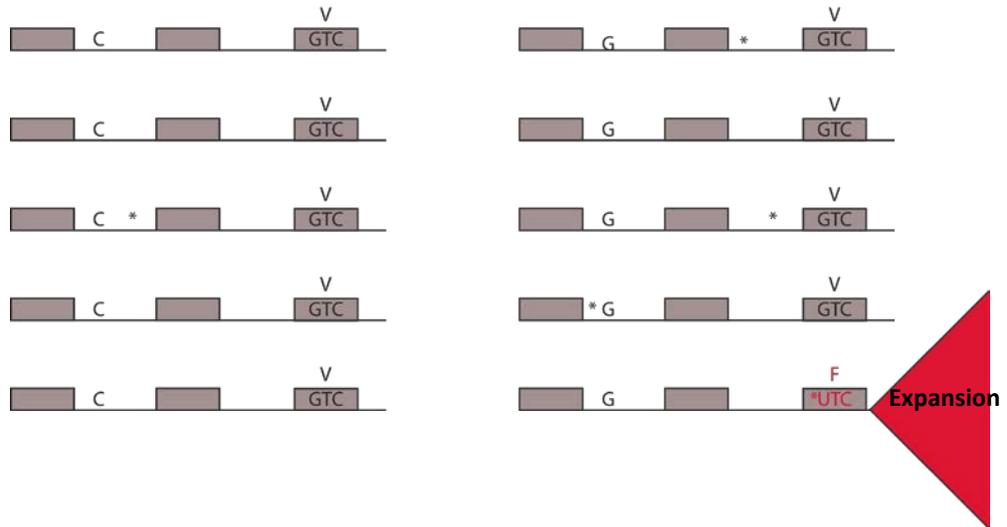
We aimed to explore these hypotheses to understand the underlying mechanism of the well-established findings of the 46/1 *JAK2* risk haplotype. We found that the tag SNP rs10974944 associated with MPN predisposition is located within 300kb extended linkage disequilibrium (LD) block that includes *JAK2*, *INSL4* and *INSL6* genes. We attempted to sequence this 300kb region in MPN cases using targeted amplification followed by next-generation sequencing. Since our previous GWAS was limited to only 60K SNPs, we expanded our study by performing a second GWAS with high-density SNP array data for 237 patients diagnosed with MPN and 1,037 shared controls. We identified 9 SNPs associated with MPN risk at genome-wide significant levels (p-value <  $1 \times 10^{-7}$ ). However, these SNP were in strong linkage disequilibrium with our previously identified MPN risk SNP (rs10974944) located in *JAK2*, thus we have replicated our previous findings.

We next explored the 46/1 haplotype in search of a functional variant that could result in allele-specific activation of *JAK2* in MPN cases. In this study, we used a combination of genotyping, imputation, sequencing, bioinformatics and functional annotation to fine-map the disease locus. Our aim was to refine the most likely disease-associated variant based on association testing at genotyped and imputed SNPs and computationally predict the causal variant associated with MPN predisposition.

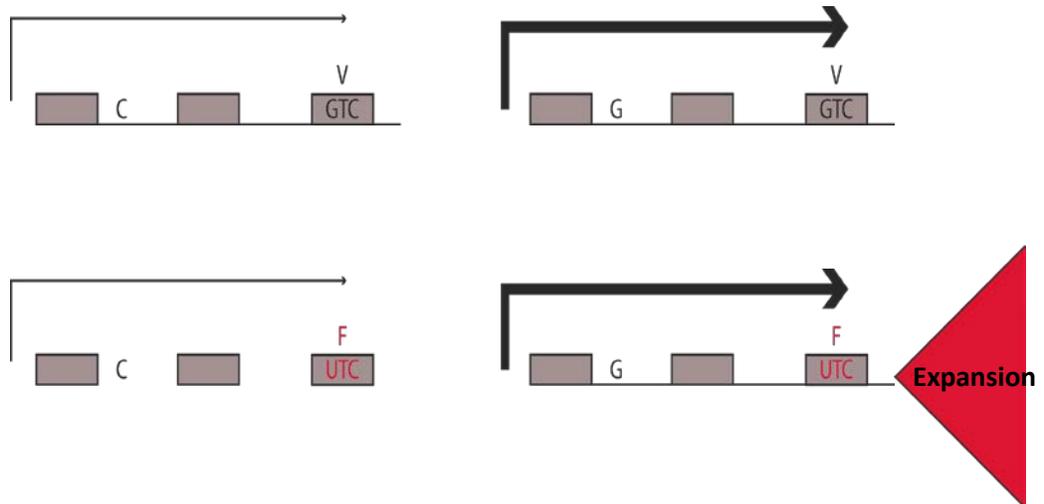
**Figure 9 The two hypotheses to explain 46/1 MPN risk haplotype**

Two hypotheses suggested to describe the observation that somatic mutation JAK2V617F occurs preferentially on 46/1 JAK2 risk haplotype.

**A) Hyper-mutability hypothesis** where 46/1 haplotype may be genetically unstable and may be susceptible to DNA damage.



**B) Activation hypothesis** where 46/1 haplotype may harbor functional variant(s) that affect allele specific JAK2 activation and provides selective



## **3.2 Methods and Materials**

### **MPN case selection**

In total, 237 patients diagnosed with MPN were recruited for this study from the Boston or New York areas. The cases were genotyped on the Illumina 1M Omni-Quad SNP genotyping array. A subset of 24 MPN cases were chosen for targeted sequencing experiments based on their genotypes at the tag SNP rs10974944. This subset included twelve MPN cases, who were homozygous G/G at rs1094744 (referred as “GG-MPN” cases) and twelve MPN cases who were homozygous C/C at rs1094744 (referred as “CC-MPN” cases). The GG-MPN cases had high JAK2V617F allele burden (greater than 90%) and acquired uniparental disomy at this locus whereas the CC-MPN cases did not acquire somatic mutation JAK2V617F. To determine both the somatic and germline genotypes at rs10974944, a Taqman genotyping assay for rs10974944 was performed for all MPN cases using DNA extracted from granulocytes (i.e disease tissue) or buccal/saliva samples, respectively.

### **JAK2 locus definition**

The MPN risk SNP rs10974944 that was identified by GWAS lies in an extended 300kb LD block as defined by HapMap CEU population using UCSC genome browser (Figure2). All analysis was done for the 300kb JAK2 locus (Chr9: 4885245 to 5269610bp).

## **Targeted amplification and next-generation sequencing**

Targeted amplification of the 300kb JAK2 risk locus was performed using the RainDance approach at the MSKCC genomics core facility. Briefly, 1284 overlapping primer pairs were designed by RainDance software with primer length ranging from 18 to 25 nucleotides. The mean size of the amplicons was 200bp and primer sets were obtained from the RainDance technology. Library preparation and sequencing run was performed using the ABI SOLiD sequencing platform at MSKCC genomic core facility according to manufacture's protocol.

## **Single nucleotide variant analysis**

ABI SOLiD sequencing reads were mapped using the Bioscope pipeline ( Corona Lite) at the MSKCC bioinformatics core facility. Using Samtools, the uniquely mapped reads from the Bam file were pileup using Human Reference sequence Mar. 2006 (NCBI36/Hg18). We used the variant calling algorithm VarScan 2.2<sup>172</sup> to identify single nucleotide variants. The filters used for variant calling were 1) minimum read depth at a position to make a call  $\geq 10X$ , 2) minimum supporting reads at a position to call variants is  $\geq 10X$  and 3) minimum variant allele frequency threshold is  $\geq 25\%$ . Homozygous calls were made where the minimum variant allele frequency threshold was greater than or equal to 90%. For heterozygous calls; we used variant allele frequency threshold between 40-60%. To minimize the number of false variant calls, we employed the

DiBayes toolset as a secondary mapping/variant pipeline. Only single nucleotide variants called by both pipelines were used downstream analyses.

To determine if variants identified in the MPN cases were ancestral or derive alleles, we compared the variants to the human ancestral sequence published by the 1000 Genomes Project<sup>77</sup>. We compared the count of single nucleotide variants present in GG-MPN cases versus CC-MPN cases using the human ancestral sequence reference. The Wilcox statistical test was performed to determine if there is any significant difference in the accumulation of variants over the generations in the two groups of MPN cases. We downloaded the single nucleotide variant call dataset for the 60 healthy individuals from European ancestry (HapMap CEU population) published by the 1000 genomes project. We analyzed these healthy CEU individuals to determine if there were any haplotype specific difference in the number of variant sites.

### **Genotyping MPN cases and shared controls**

In total, 237 MPN cases were genotyped on the Illumina Omni-1 Quad SNP genotyping array at the Genomics Core Laboratory of MSKCC according to the manufacturer's protocol. We downloaded genotype data of healthy individuals from NIH's Database of Genotypes and Phenotypes (dbGaP) to use as shared controls for our present GWA study. All individuals used as controls in the underlying study are of European ancestry. Specifically, genotype data for 1037 healthy controls from the Melanoma study was used since they were genotyped in the same platform as MPN cases (Illumina Omi-1 Quad) (dbGaP accession id: phs000187.v1.p1.c1).

## **Genotype data processing and association testing**

All genotype data was processed using PLINK<sup>69</sup>. We performed several steps of quality control (QC) to the MPN case dataset and the shared control datasets separately before merging them. Firstly, ambiguous SNPs (A/T or C/G) were removed from the analysis due to strand ambiguity in the two datasets. Next, we removed individuals with more than 10% of SNPs not called and removed SNPs that had >1% missing genotypes or a minor allele frequency <5%. A total of 723,486 markers passed QC in both the MPN case and shared control datasets that passed QC. The datasets were then merged using PLINK, restricting analysis to a set of SNPs common to both datasets. Following this, a second round of QC steps was performed (mainly to remove markers that were out of Hardy-Weinberg equilibrium in controls ( $p < 1 \times 10^{-7}$ )). We also removed SNPs that showed a significant difference in missingness rates between cases and controls ( $p < 1 \times 10^{-7}$ ). Thus, final dataset included 723,016 markers for 180 MPN cases and 1037 shared controls.

## **Population stratification correction and association test**

To adjust for population substructure, we performed principle component analysis using the EIGENSTRAT program from the EIGENSOFT 2.0 package<sup>144</sup>. We first filtered the Illumina Omni 1 SNP genotype data by removing markers in high linkage disequilibrium (LD). This gave us a set of 41,636 SNPs for which pairwise  $r^2$  values within a window of 50 SNPs were all <0.1 (--indep-pairwise 50 5 0.1 command in PLINK). These markers were then used as input for EIGENSTRAT. Principal

components were computed and outliers removed using default parameters. Eight significant principal components were determined using the Tracy-Widom statistic ( $p < 0.05$ ) and were used as covariates in a logistic regression model for risk association.

### **Imputation and association tests**

After performing QC on the Illumina Omni-1 SNP genotype data, we used the IMPUTE program (version 2.1.2)<sup>78</sup>, which imputes unobserved genotypes in MPN cases and shared controls based on a set of known haplotypes derived from initial low coverage sequencing of European ancestry (CEU) samples in the 1000 Genomes Project.

Imputation was done for the 300kb *JAK2* locus (Chr 9: 4880kb-5270kb, NCBI hg18). We had genotype data for 93 SNPs in this region from the Illumina SNP array data; these genotypes were used as input for imputation. Using default parameters for IMPUTE, there were 1,034 SNPs imputed in the analysis region based on 1000 Genomes Project reference haplotypes. The output from IMPUTE was converted to ped (PLINK) format for further analysis. Association testing was performed by use of a logistic regression model (in PLINK) that included the top eight principle components of population substructure. Furthermore, the logistic regression analysis was conditioned on the initial tag SNP (rs10974944) to determine if there were independent signals present in 300kb LD block associated with MPN predisposition.

## **Functional annotation**

We examined two sources of functional annotation: 1) the ENCODE integrated regulation track published in the UCSC genome browser, and 2) Consite, a user-friendly, web-based tool for finding *cis*-regulatory elements in genomic sequences based on the TRANSFAC database<sup>173</sup>. The ENCODE integrated regulation tracks contain information relevant to the regulation of transcription based on analyses from the ENCODE project. The “Transcription” (Txn) track shows transcription levels assayed by sequencing of polyadenylated RNA from a variety of cell types. We focused on the Txn Factor ChIP-seq track, which shows DNA regions where transcription factors (proteins responsible for modulating gene transcription) bind as assayed by chromatin immunoprecipitation with antibodies specific to the transcription factor followed by sequencing of the precipitated DNA (ChIP-seq). We downloaded the ChIP-seq signal data from the UCSC browser for the 300kb *JAK2* locus data (Chr9: 4885245 to 5269610). Using this data, we next identified the set of imputed SNPs that resided in regions of elevated ChIP-seq signals. Finally, we used Consite to identify putative transcription factor binding sites within these regions whose binding efficiency could be altered due to presence of SNPs.

## **Allele-specific JAK2 expression in MPN cases**

8 MPN cases were analyzed to determine whether JAK2 is expressed in an allele-specific manner. These cases were genotyped at rs10974944 using a Taqman SNP genotyping assay and were found to be heterozygous (CG). To determine the allele-specific expression of JAK2 in MPN cases, we assayed an exonic SNP, rs2230724 that

was in perfect LD with the JAK2 risk SNP rs10974944. The Sanger sequence traces of genomic DNA and cDNA obtained from RNA at the exonic SNP rs2230724 were compared visually to check the allele-specific difference between gDNA and cDNA trace. This was done by Levine lab to understand functional difference between the risk and wild type haplotype.

### 3.3 Results

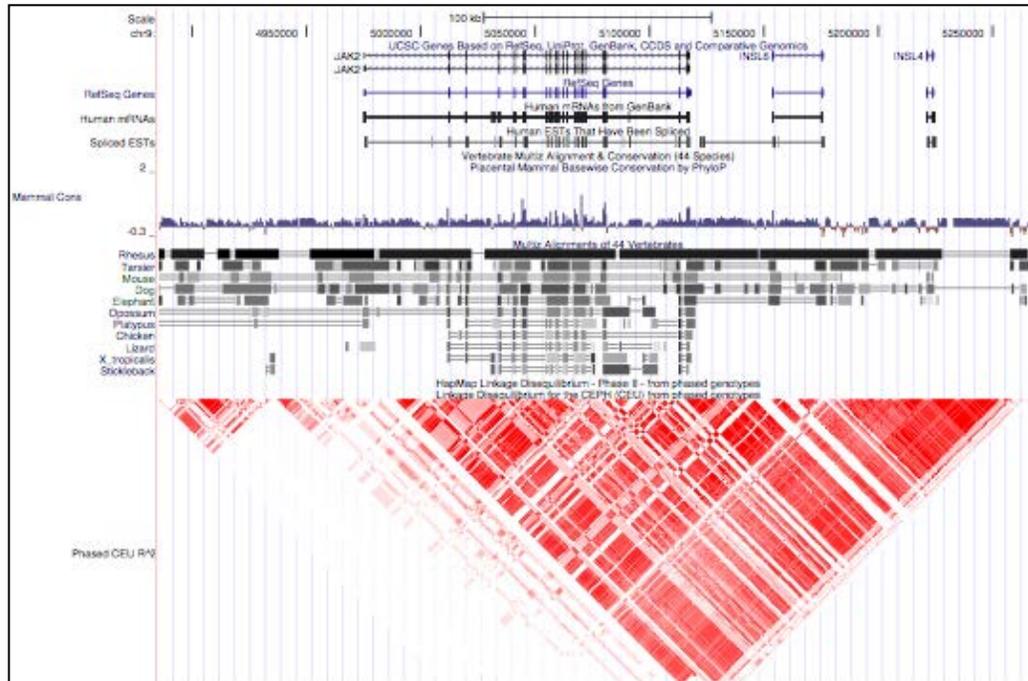
#### Targeted Sequencing of JAK2 locus

The risk-associated SNP rs10974944 is located in a 300 kb extended linkage disequilibrium block on chromosome 9 (Chr9: 4885245 to 5269610, hg18) as shown in **Figure 10**. To test whether this SNP confers hyper-mutability at the JAK2 locus, we compared two groups of MPN cases: those carrying the 46/1 MPN risk haplotype, which is tagged by rs10974944, and those not carrying the 46/1 haplotype. Specifically, twelve MPN cases homozygous for the rs10974944 risk allele (referred as GG-MPN cases) and twelve MPN cases homozygous for the protective allele (referred as CC-MPN cases) were processed for targeted amplification of the 300kb LD block followed by next generation sequencing. 80% of the targeted region was captured with a minimum of 5X read coverage. After mapping the reads from the SOLID run, we used two methods, VarScan 2.2 and diBayes, to identify single nucleotide variants carried by each of the analyzed cases. Using the NCBI36/hg18 Human genome build as a reference, we found that there was a significant difference in the number of single nucleotide variants present in GG-MPN cases when compared to CC-MPN cases (**Table 10**). Notably, the NCBI36/hg18 reference sequence contains a C allele at SNP rs10974944, which indicated the presence of the wild-type haplotype at JAK2 locus. To determine if the 46/1 haplotype acquires more single nucleotide variants over generations when compared to the wild-type haplotype, we recomputed the number of single nucleotide variant calls in both sets of MPN cases using the human ancestral sequence published by 1000 Genomes

Project as reference. As shown in **Figure 11**, no significant difference in the number of single nucleotide variants was found between the two groups of MPN.

**Figure 10 Schematic diagram of 300kb JAK2 risk locus**

Obtained from the UCSC genome browser. The LD pattern shown is based on the HapMap Phase3 CEU population.



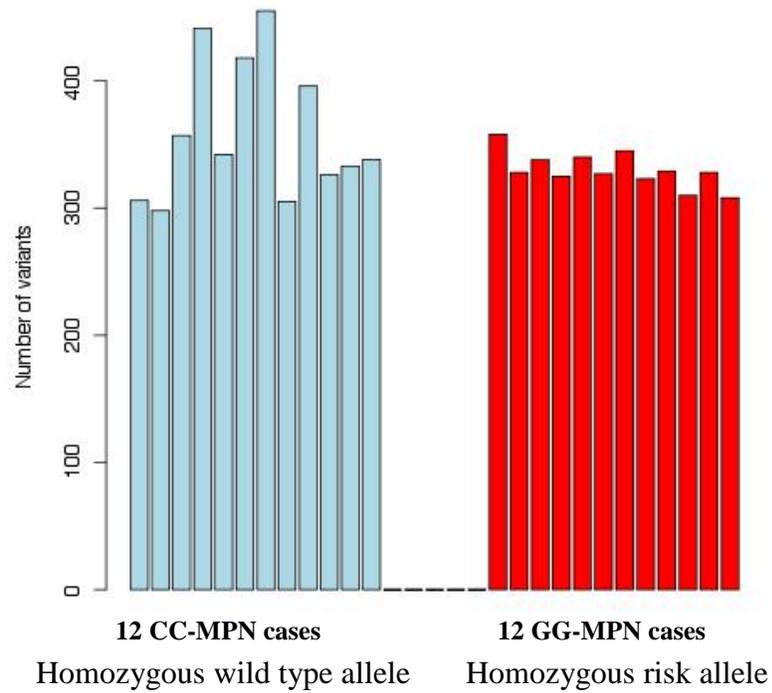
**Table 10 Single nucleotide variant count in MPN cases with or without 46/1 risk haplotype**

The right-most column shows the number of single nucleotide counts obtained from SOLID sequencing data when using the NCBI36/hg18 build as reference. Somatic genotypes at rs10974944 were assayed using blood-derived DNA (i.e. disease tissue). Germline genotypes at rs0974944 were assayed using buccal- or saliva-derived DNA. Diagnosis: polycythemia vera (PV) or essential thromocythemia (ET).

		rs10974944 genotype					Using NCBIHg18 as reference	
MPN_id	Diagnosis	Somatic genotype	germline genotype	JAK2V617F mutation burden	Age	Haplotype	Number of variants	
121	PV	CC	CC	0.08	67	C	82	
285	ET	CC	CC	0.11	61	C	86	
166	PV	CC	CC	0.18	NA	C	193	
241	ET	CC	CC	0.59	60	C	371	
265	ET	CC	CC	1.36	49	C	150	
396	ET	CC	CC	0.03	46	C	336	
164	PV	CC	NA	0.06	NA	C	382	
489	ET	CC	CC	0.1	74	C	119	
40	PV	CC	CC	0.11	54	C	293	
390	ET	CC	CC	0.12	53	C	199	
427	ET	CC	CC	0.07	49	C	222	
205	ET	CC	CC	3.45	49	C	208	
290	ET	GG	CC	74.61	50	G	419	
19	PV	GG	GG	94.13	44	G	332	
175	PV	GG	GG	58.01	64	G	424	
303	PV	GG	GG	81.18	77	G	374	
168	PV	GG	CG	94	52	G	391	
59	PV	GG	GG	94.16	59	G	396	
328	PV	GG	CG	95.11	53	G	404	
162	PV	GG	GG	58.66	60	G	440	
155	PV	GG	CG	89.19	73	G	407	
105	PV	GG	GG	88.36	50	G	427	
179	PV	GG	GG	98.43	63	G	431	
10	PV	GG	GG	99.34	67	G	422	
<b>Wilcox test p-value</b>							<b>8.8x10<sup>-6</sup></b>	

**Figure 11 Single nucleotide variant counts for MPN cases with and without 46/1 risk haplotype using human ancestral sequence as reference**

y-axis is the number of single nucleotide variants counts obtained for 12 CC-MPN cases (without 46/1 risk haplotype) and 12 GG-MPN cases (with 46/1 risk haplotype)



### **Analysis of the JAK2 risk locus in healthy individuals**

We next determined if the 46/1 haplotype is unstable in the general European population. To do so, we obtained single-nucleotide variant data published by the 1000 Genomes Project for a group of 60 healthy individuals of European ancestry (CEU) and analyzed the 300kb LD block encompassing *JAK2*. Among this group, there were four individuals homozygous for the risk allele (GG) at rs1097944, 27 individuals heterozygous (CG) for the risk allele and 29 individuals homozygous for the wild-type allele (CC). Comparing these individuals, we found there was no correlation between the number of single nucleotide variants and the genotype of individuals at rs10974944 when using a human ancestral sequence derived from the 1000 Genome Project as reference (Table 12).

**Table 11 The number of single nucleotide variants in HapMap healthy individuals from European ancestry (CEU) obtained from 1000 genomes project**

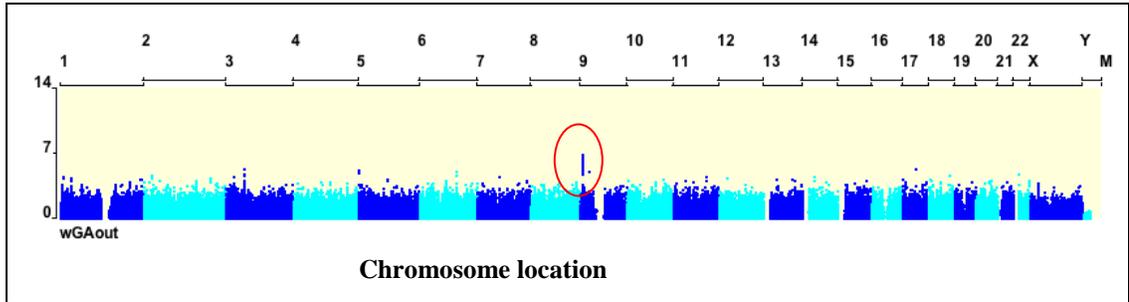
<b>CHR</b>	<b>SNP</b>	<b>VALUE</b>	<b>G11</b>	<b>G12</b>	<b>G22</b>
9	rs 10974944	GENO	G/G	G/C	C/C
9	rs 10974944	COUNTS	4	27	29
9	rs 10974944	FREQ	0.06667	0.45	0.4833
9	rs 10974944	MEAN variant Count	715.2	808.9	732
9	rs 10974944	SD	14.8	11.56	26.65

## Extended Genome Wide Association Study

We extended our MPN genome-wide association study by genotyping 237 MPN cases on a high-density SNP array (Illumina Omni-1 quad). We combined the genotype data of these MPN cases with data from 1037 shared controls genotyped using the same SNP array. After performing quality control, we performed a single-marker association test for each of 723,016 SNPs in the combined case-control dataset by use of a logistic regression model, including top nine significant principle components of population structure to adjust for stratification. Nine SNPs at the *JAK2* locus were statistically significantly ( $p$  value  $< 10^{-7}$ ) associated with MPN risk; these SNP were found to be in high linkage disequilibrium (LD) with our previously identified risk variant rs10974944 (Table 12 and Figure 12). We did not identify any novel loci associated with MPN predisposition in our study, instead replicated our previous finding. We next focused our analysis to this extended strong LD 300kb block that contains the *JAK2* risk SNPs (Figure 10). 93 SNP with genotype data for MPN cases and shared controls spanning this 300kb block was available and were used for imputation and fine mapping the MPN risk locus.

**Figure 12 Manhattan plot for extended MPN GWA study**

Manhattan plot of genome-wide association results obtained by logistic regression analysis of 723,016 SNPs in 237 MPN cases and 1,037 shared controls adjusted for population stratification. The x-axis is the chromosome location from 1 to 23 and y-axis is the negative log of the p-value for each test. The red circle is the JAK2 locus on chromosome 9.



**Table 12 List of SNPs associated with MPN risk**

The p-values, odd ratios and linkage disequilibrium measures (calculated as  $r^2$  and  $D'$ ) between each SNP and the previously identified MPN risk variant rs10974944. (LD measures were calculated based data from the 1000 Genomes Project for individuals in the CEU population.

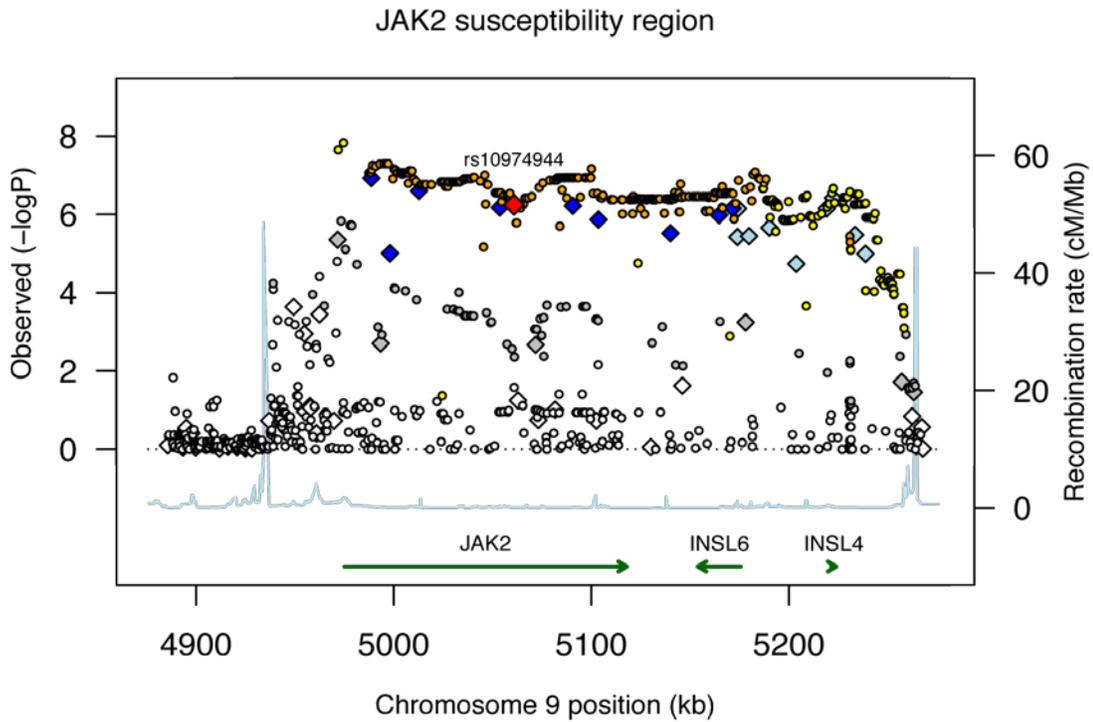
<b>Chromosome</b>	<b>SNP id</b>	<b>P-value</b>	<b>Odd Ratio</b>	<b><math>r^2</math> value</b>	<b><math>D'</math></b>
9	rs2225125	1.19E-07	2.25	0.958	1
9	rs7851556	2.51E-07	2.179	0.958	1
9	rs884132	2.84E-07	2.213	0.959	1
9	rs3780382	3.80E-07	2.181	0.715	0.903
9	rs7870694	6.02E-07	2.135	0.959	1
9	rs10815149	6.69E-07	2.111	0.959	1
9	rs7047795	7.00E-07	2.131	0.92	1
9	rs10114531	7.09E-07	2.123	0.72	0.867
9	rs12349508	7.11E-07	2.164	0.92	1

## Imputation and association test

Our next goal was to use imputation at the 300kb *JAK2* locus and conditional analysis to determine if any of the known MPN risk alleles had either (1) a better signal of association or (2) an independent, second signal of association in the associated risk locus. The imputed SNPs were tested for association with MPN risk under a logistic regression model adjusted for population stratification. The analysis yielded 450 SNPs with p-value < 1e-05 in high LD ( $r^2 > 0.8$ ) with the previously identified *JAK2* risk SNP rs10974944 (Figure 13). To refine the association signal, we conditioned the analysis of each imputed SNP on rs10974944 to look for additional statistical evidence of association. Under conditional analysis, no other SNPs showed strong evidence for association. Because of these findings, we focused solely on refining the signal of association at the *JAK2* locus. The top 10 imputed and associated SNPs are presented in Table 13 with their functional class. Of these the top 2 SNPs were located in the promoter region of *JAK2* gene namely, rs1887428 -position chr9: 4974530- pvalue = 1.48e-08 and rs36051895 position chr9: 4971866- pvalue = 2.24e-08 (Table 13). The two SNPs that are found to be associated with Crohn's disease<sup>138,174</sup> and ulcerative colitis<sup>175,176</sup> showed significant association signal in our data (rs10758669 associated with both Crohn's disease and ulcerative colitis had p-value = 4.5e-07 and rs10975003 that was found to be associated with Crohn's disease had p-value= 1.6e-5 in our analysis).

**Figure 13 Association plot for imputed and genotyped SNPs at JAK2 susceptibility locus**

The x-axis represents chromosome position and the y-axis is the negative of log(p-values) obtained from logistic regression analysis using eight principle components as covariates. Diamond shape are for genotyped SNPs and circles are imputed SNPs. The initial JAK2 risk SNP rs0974944 is shown as red diamond. Colors: blue = genotyped SNP with high LD with rs10974944 ( $r^2 > 0.8$ ), light blue = genotyped SNPs with moderate LD with rs10974944 ( $0.8 > r^2 > 0.5$ ), grey = genotype SNP with weak LD with rs10974944 ( $0.5 > r^2 > 0.2$ ), Orange = imputed SNP with high LD with rs10974944 ( $r^2 > 0.8$ ), yellow = imputed SNPs with moderate LD with rs10974944 ( $0.8 > r^2 > 0.5$ ), pink = imputed SNP with weak LD with rs10974944 ( $0.5 > r^2 > 0.2$ ), white = genotyped or imputed SNPs not in LD with rs10974944. Light blue lines shows the recombination rate and green lines show the three genes in this locus.



**Table 13 Association results for imputed SNPs with their functional annotation**

<b>Rank</b>	<b>SNP</b>	<b>Imputed p-value</b>	<b>Position</b>	<b>Functional class</b>	<b>Gene</b>
1	rs1887428	1.48E-08	4974519	promoter SNP	JAK2
2	rs36051895	2.24E-08	4971866	promoter SNP	JAK2
3	rs12349508	2.60E-08	5174222	intronic SNP	INSL6
4	rs2225125	4.27E-08	4988639	intronic SNP	JAK2
5	rs59384377	5.05E-08	4995034	intronic SNP	JAK2
6	rs62541529	5.05E-08	4996345	intronic SNP	JAK2
7	rs11999928	5.05E-08	4996743	intronic SNP	JAK2
8	9-4997138	5.05E-08	4997138	intronic SNP	JAK2
9	rs10120763	5.14E-08	4992911	intronic SNP	JAK2
10	rs1327494	5.51E-08	4989303	intronic SNP	JAK2

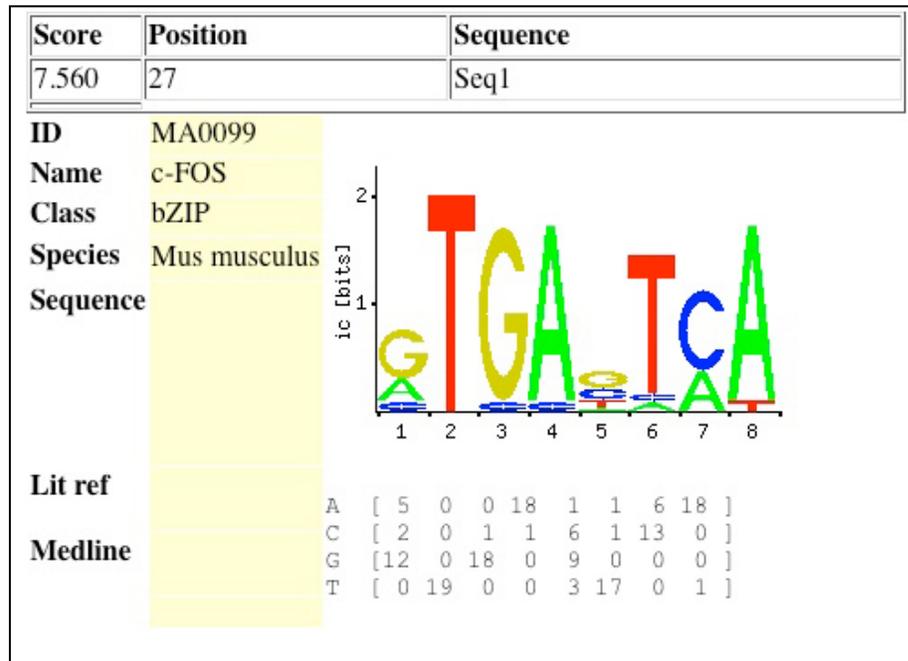
## Functional prediction of causal variant

To prioritize the imputed SNPs as functional or causal variants, we used the ENCODE integrated regulation track published in UCSC genome browser and Consite, an algorithm to predict transcription binding factor sites (TFBS). Of the 1034 SNPs present in the 300kb *JAK2* analysis region (based on 1000 Genomes Project reference haplotypes), we successfully imputed 925 SNPs. Each imputed SNP was tested for association with MPN risk. We next checked if any of these 925 SNPs lied within the Encode Txn Factor ChIP-seq signal regions. There were 60 SNPs (out of 925) located within the ChIP-seq signals or blocks, of which 18 SNPs were also present in the targeted sequencing data for 12 GG-MPN cases. SNPs that affect the efficiency of TFBS are excellent candidates for GWAS hits as they are thought to be causally involved in complex diseases. Thus, to identify putative causal variants located in ChIP-seq signal regions, we used the TF binding site prediction tool Consite. We found two imputed SNPs that showed significant association p-values and allele-specific transcription factor binding as predicted by Consite. The best predicted functional variant was rs1887428, located in the promoter region of the *JAK2* gene (position chr9: 4974530). It was the top ranked SNP in the association analysis (pvalue =  $2.9e-11$ ) and is in strong LD with the known risk SNP rs10974944 ( $r^2 = 0.59$ ). This SNP was predicted to affect the binding of transcription factor c-Fos (Figure 14). Only the risk allele (G) at rs1887428 was predicted to enable c-Fos binding. This suggested that the SNP mediates allele-specific *JAK2* activation or regulation. 11 of the 12 GG-MPN cases analyzed in the SOLID-RainDance sequencing experiment had risk allele at this locus whereas it was absent in CC-MPN cases. An additional SNP of interest, rs10815157 was found in an intronic region of *JAK2*

intron (position chr9: 5,099,021), with an association p-value =  $1.1e-7$  and strong LD with rs10974944 ( $r^2 = 0.9$ ). This SNP was predicted to affect the binding of n-Myc and was present in 10 of the 12 GG-MPN cases analyzed in the SOLID-RainDance sequencing experiment.

**Figure 14 Predicted functional SNP rs1887428**

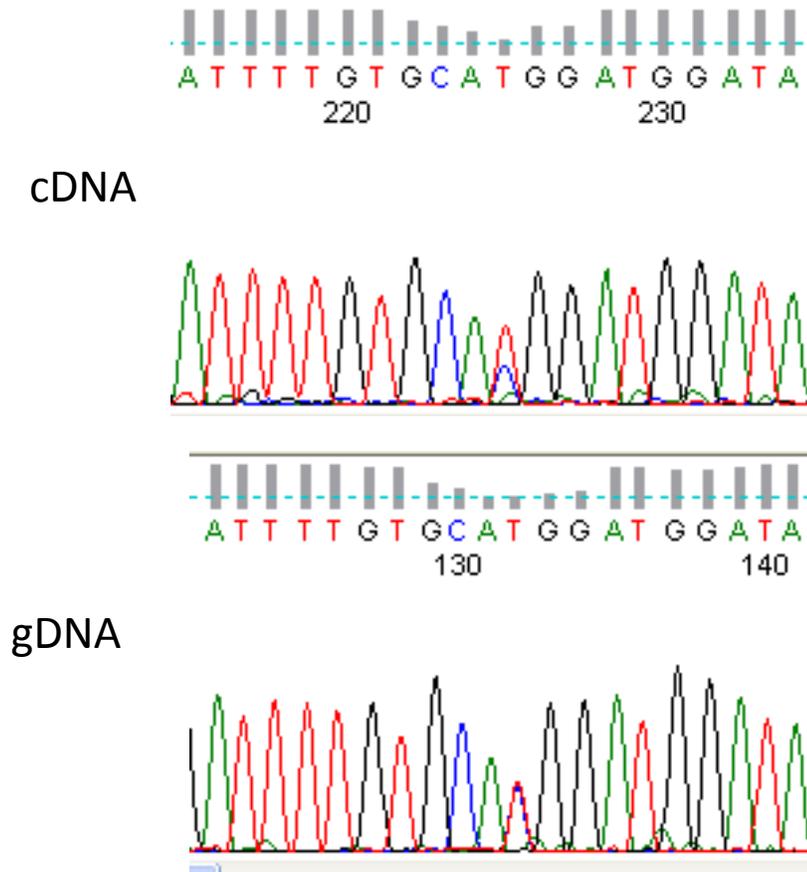
Consite output showing the position-weight matrix at SNP rs1887428 and the putative c-Fos binding site



### **Allele specific JAK2 expression in MPN cases**

To determine if allele-specific expression of *JAK2* gene in MPN cases could be observed, we analyzed the MPN patients that were heterozygous at the risk SNP rs10974944. We assayed an exonic SNP rs2230724 that is in perfect LD with the MPN risk SNP rs10974944 in these heterozygous MPN cases by Sanger sequencing of genomic and cDNA. Figure 15 shows the representative heterozygous MPN case with sequence trace at the exonic SNP for genomic DNA and cDNA. We observed minute allelic imbalance in the sequence trace of cDNA, thus suggesting a subtle allele-specific difference in expression of *JAK2* gene.

**Figure 15 Allele-specific expression of JAK2 in heterozygous MPN cases** assayed by Sanger sequencing of genomic DNA and cDNA.



### 3.4 Discussion

In the present chapter, we aimed to understand the mechanism by which the 46/1 MPN risk haplotype acquires the somatic mutation JAK2V617F in cis in MPN cases. The finding that the JAK2V617F mutation is acquired preferentially on a 46/1 *JAK2* haplotype was unexpected and the mechanism underlying this observation remains unexplained. There are two hypotheses suggested: a hyper-mutability hypothesis and an activation hypothesis. The hyper-mutability hypothesis is similar to the phenomenon observed in the *APC* and *TP53* genes<sup>168</sup>, whereby DNA sequence variations in those genes predispose them to somatic mutagenesis. Although the activation hypothesis cannot explain occurrences of JAK2 mutations directly, other acquired or inherited genetic variants on the 46/1 haplotype may predispose to the acquisition of JAK2 mutations.

To dissect the two hypotheses, we carried out targeted sequencing and fine mapping at the disease locus JAK2. The MPN risk SNP rs10974944 is located in the intron of JAK2 and tags the 46/1 haplotype. This locus has an extended 300kb linkage disequilibrium block as determined from the HapMap individuals of European ancestry (CEU). We concluded that there was no haplotype-specific difference in the number of single nucleotide variants present in MPN cases when using the human ancestral sequence published by 1000 genomes project as reference. We also verified this for healthy individuals from 1000 genomes project CEU population and confirmed that the 46/1 haplotype is neither unstable nor accumulates single nucleotide variants over

generations in MPN cases as well as in general population.

Recently, Jones et al found that 46/1 was overrepresented in *JAK2*V617F positive and negative ET cases, *MPL* exon10 mutated ET cases versus controls<sup>177</sup>. An excess of 46/1 was also seen in *JAK2* exon 12 mutated cases and these mutations preferentially arose on the 46/1 chromosome<sup>178</sup>. Thus the excess of 46/1 in *MPL* mutated cases argues against the hyper-mutability hypothesis.

On the other hand, a 46/1 tag SNP showed robust association with Crohn's disease<sup>138</sup>, a nonmalignant disorder that is believed to have an inflammatory cause. GWA studies in Crohn's disease also detected significant associations with genes encoding the IL-23 receptor and STAT3. These findings suggests the role of 46/1 *JAK2* haplotype in activation of *JAK2* gene in allele-specific manner that may cause different diseases. We observed allele-specific expression of *JAK2* in MPN cases that were heterozygous at tag SNP rs10974944. Thus, we favor the activation hypothesis and used a combination of genotyping, imputation, sequencing, bioinformatics and functional annotation to fine-map the disease locus. The discovery of functional variants is aided by a deep examination of genetic variation in the linkage disequilibrium (LD) block in which tag SNP resides. We determined the SNP(s) most likely to be functional within the fine-mapped regions based on potential functional role using various functional annotation tools. We genotyped 233 MPN cases using a dense SNP genotyping platform, the Illumina Omni-1 quad, and combined the genotype data with shared controls from a public database. 93 genotyped SNPs within the 300kb *JAK2* locus served as the basis imputation of additional untyped. We identified novel SNPs in the promoter region of

JAK2 gene to be associated with MPN predisposition. Using ENCODE project data and a transcription factor binding site prediction algorithm, we have further identified a candidate SNP, rs1887428, in the promoter region of JAK2 (position chr9: 4974530) that was predicted to affect the binding of transcription factor c-Fos. The risk allele G at rs1887428 was predicted to form a c-Fos binding site. It has been shown that c-fos is stably induced during normal hematopoietic differentiation and Jun/Fos acts as positive modulators of hematopoietic differentiation. Thus, we hypothesized that somatic mutation of the *JAK2* gene in MPN cases will lead to c-Fos activation, a downstream target gene of the JAK-STAT pathway and that c-Fos may bind to the JAK2 promoter leading allele-specific JAK2 regulation. Thus, we concluded that the 46/1 haplotype does not seem to be hyper-mutable and may harbor functional variants supporting allele-specific JAK2 activation.

## CHAPTER 4

### An Evolutionary Model for the JAK2 Susceptibility Locus

#### 4.1 Introduction

As discussed in chapters 2 and 3, we and others have identified a JAK2 haplotype (designated as 46/1 or “GGCC”) that is strongly associated with the development of JAK2V617F positive MPN<sup>153,169,171</sup>. These findings suggest a complex interplay between germline variations and somatic mutation at the JAK2 locus in MPN patients. The MPN risk SNP rs10974944 lies in an extended 300kb linkage disequilibrium (LD) block. This 300kb region exhibits a low recombination rate. The risk allele (G) of rs10974944 is an ancestral allele and the frequency of homozygous GG carriers in the European population is 5%. In this chapter, we aimed to explore the evolution of the MPN susceptibility haplotype in order gain new insights into its disease association.

There are several examples of disease-associated germline variants in which the risk allele is ancestral allele. For example, a variant in the apolipoprotein E (APOE) gene associated with increased the risk of coronary artery disease and Alzheimer’s disease<sup>65,150</sup> carries the ancestral allele<sup>179</sup>. Similarly, it has been shown that the PPARG gene harbors an ancestral variant allele that influences type 2 diabetes susceptibility<sup>66</sup>. Likewise, the ancestral allele of a germline variant in the CAPN10 gene has been shown to increase the risk of metabolic syndrome<sup>180,181</sup>. These examples and others have led Rienzo and Hudson to develop an explicit evolutionary model: the ancestral-susceptibility model<sup>147</sup>.

In the present chapter, we investigated whether the MPN associated JAK2 haplotype can be explained by the ancestral-susceptibility evolutionary model. To do so, we analyzed whole-genome SNP array data for a set MPN cases and shared controls from a public repository and focused on a broader set of SNPs within the 300kb extended linkage disequilibrium block encompassing JAK2. A haplotype-association test of SNPs in this region was able to identify the previously reported 46/1 (or “GGCC”) MPN risk haplotype<sup>169,171</sup>. We then reconstructed the phylogenetic tree of haplotypes observed in our MPN cases using chimpanzee sequence as an out-group and found that MPN risk haplotype forms a separate cluster from other haplotypes. In addition, the MPN risk haplotype showed the highest degree of sequence similarity to chimpanzee, thus indicating that it most likely represents an ancestral haplotype. Next, using HapMap Phase 3 population data, we found that the JAK2 locus, despite the lack of strong evidence of recent positive selection, has an excess derived allele frequency compared to genomic regions under neutral selection. Our findings suggest that the JAK2 MPN risk locus is consistent with the ancestral-susceptibility model.

## **4.2 Materials and Methods**

### **Study population and genotype data**

The MPN cases, controls from public database and SNP genotype data is described in detail in Chapter 3 materials and methods. We used the same dataset in the present chapter to understand the evolution model of JAK2 risk haplotype. In total, 237 MPN cases and 1037 controls from melanoma study genotyped in Illumina Omni-1 quad SNP array were used in the present study. All genotype data was processed using PLINK<sup>69</sup>. For haplotype analysis, we focused on the 300kb JAK2 risk locus.

### **Haplotype block definition and association test**

The linkage disequilibrium (LD) pattern in the analysis region was determined using Haploview version 4.2 (<http://www.broad.mit.edu/mpg/haploview/>)<sup>182</sup>. The Gabriel protocol, which is the default method for Haploview, was applied the case-control dataset with an upper D' confidence interval bound of 0.98, a lower D' confidence interval bound of 0.70, and with 5% of informative markers required to be in strong LD<sup>183</sup>. We next performed haplotype disease association tests by comparing the observed frequency of each haplotype in MPN cases and shared controls (significance was determined empirically using 1000 permutations of the case-control labels). Haploview was used to plot the observed LD pattern across 93 SNPs in the JAK2 region based on the analysis of 166 MPN cases and 1,037 shared controls.

## **Phylogenetic analysis**

Phylogenetic analysis was performed using PHYLIP (Phylogenetic Inference Package, version 3.69), a package developed by Felsenstein from the University of Washington<sup>184</sup>. We selected haplotype block 5 (which was identified by Haploview) for phylogenetic reconstruction analysis via programs available in PHYLIP. Using total of 16 DNA sequences, 12 haplotypes (Table 3B) determined from Haploview, NCBI hg18 human reference sequence and 3 primate sequences – Chimpanzee, Orangutan, and Monkey- we first determined sequence distance using the DNAdist program. Then, the output DNAdist was used as input for Dnapars, a DNA parsimony method. The SeqBoot program was used for bootstrapping with a parameter of 100. Finally, the reconstructed tree was drawn using outtree program.

## **HapMap project data**

We used Phase3 data from the HapMap project, which contains individuals from 11 different populations in various geographical locations.

## **Positive selection tests**

We assessed the *JAK2* locus for evidence of positive selection by performing several standard methods to characterize the pattern of variation within the human population. To evaluate our sensitivity to detect positive selection at the *JAK2* locus, we compared our results with those found at the *TYRP1* gene and neutral ancestral repeats. The *TYRP1* gene is a melanin biosynthesis gene present in chromosome 9 that has been

shown to be under positive selection pressure. To determine the distribution of derived allele frequencies, we extracted SNPs from the JAK2 region on chromosome 9 from positions 4885245 to 5269610 in NCBI build 36 and from the *TYRP1* gene region on chromosome 9 from positions 12499671 to 12884036 in NCBI build 36 and determined ancestral alleles and minor allele frequencies from dbSNP build 131. Similarly, we obtained the minor allele frequencies for ancestral repeat regions. We determined if the minor allele of every SNP in these regions were same as the derived allele and assigned the DAF accordingly. We determined the distribution of derived allele frequencies for each of the three regions and conducted all three pairwise comparisons via a two-sided Wilcoxon test.

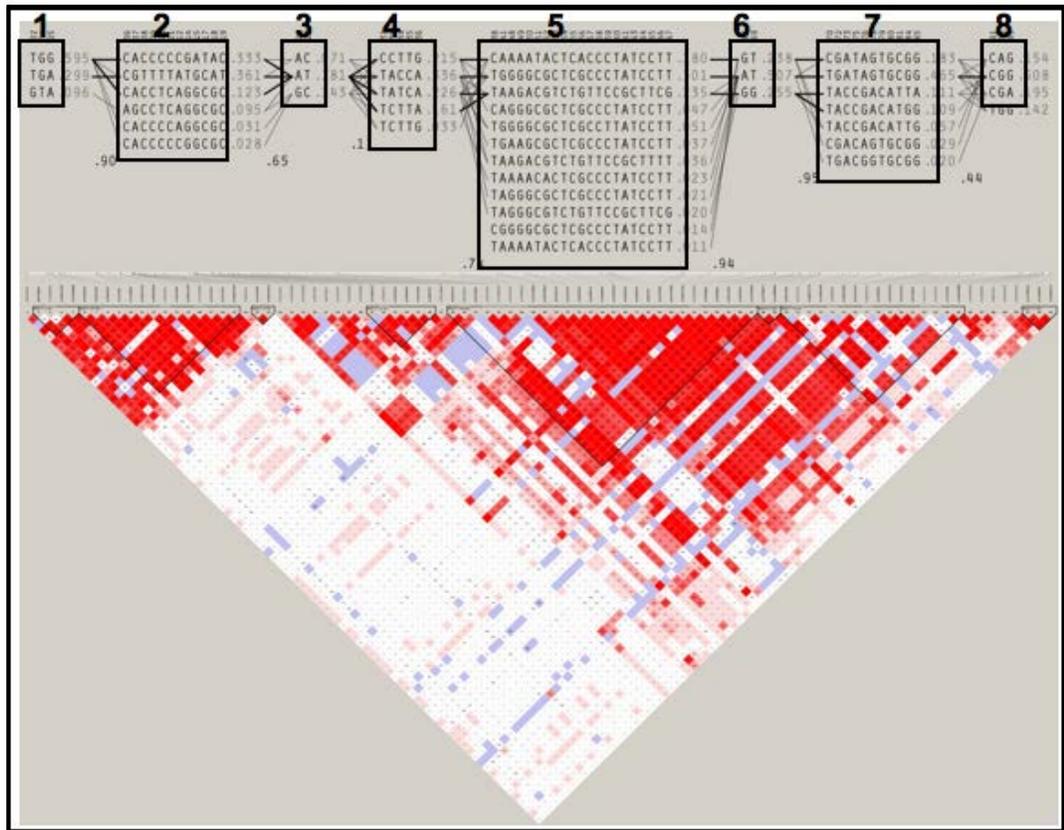
To calculate the  $F_{st}$  score between different HapMap populations, we analyzed the Hapmap Phase III dataset described above. We extracted SNPs in the JAK2 region, the TYRP1 region, and ancestral repeats to determine allele frequencies for all extracted SNPs and calculated pair-wise  $F_{st}$  for each SNP between each pair of HapMap populations. We next calculated the  $F_{st}$  score for each SNP by averaging over all pair-wise  $F_{st}$  values and compared distribution of average  $F_{st}$  for each region via a Wilcoxon test.

## 4.3 Results

### Haplotype association test

We analyzed 93 SNPs in the 300kb region surrounding JAK2 to identify haplotype blocks present in 166 MPN cases and 1037 shared controls using the Haploview package<sup>182</sup>. A total of 10 haplotype blocks were identified (Figure 16). Of these, haplotype block number 5 harbored the previously reported MPN risk variant (rs10974944). Using 1000 permutations, we performed haplotype association tests for each of the identified haplotype blocks and found block 5 to be statistically significantly associated with disease status (Table 15). This block contained 12 different haplotypes. Notably, the haplotype in block 5 that was most significantly associated with MPN in our study (referred to as the “MPN risk haplotype”,  $p\text{-value} = 5 \times 10^{-10}$ ) is identical to the “46/1” haplotype identified by other investigators. To understand how the haplotypes in block 5 were related to each other, we next turned towards phylogenetic analysis.

**Figure 16 Haplotype plot for MPN cases and controls constructed using Haploview**  
 A plot of the haplotype structure observed in 166 MPN cases and 1,037 shared controls as constructed by Haploview. The blocks were numbered 1-8. Block 5 contains the MPN risk SNPs that were identified in our previous GWAS.



**Table 14 Haplotype Association results obtained from Haploview**

<b>Block</b>	<b>Frequencies</b>	<b>Case, Control Frequencies</b>	<b>p-value</b>
<b>Block 1</b>			
TGG	0.595	0.595, 0.595	0.992
TGA	0.299	0.293, 0.300	0.7945
GTA	0.096	0.104, 0.095	0.6122
<b>Block 2</b>			
CGTTTATGCAT	0.361	0.313, 0.370	0.0478
CACCCCCGATAC	0.333	0.354, 0.331	0.4152
CACCTCAGGCGC	0.123	0.128, 0.122	0.7914
AGCCTCAGGCGC	0.095	0.108, 0.093	0.3811
CACCCCAAGGCGC	0.031	0.031, 0.031	0.9868
CACCCCCGGCGC	0.028	0.034, 0.027	0.4686
<b>Block 3</b>			
AT	0.781	0.761, 0.784	0.3471
GC	0.143	0.162, 0.140	0.292
AC	0.071	0.067, 0.072	0.7635
<b>Block 4</b>			
TACCA	0.336	0.207, 0.356	1.08E-07
TATCA	0.226	0.323, 0.211	7.00E-06
CCTG	0.215	0.196, 0.218	0.3689
TCTTA	0.161	0.203, 0.154	0.0266
TCTG	0.033	0.040, 0.032	0.4561
<b>Block 5</b>			
TGGGGCGCTCGCCCTATCCTT	0.301	0.227, 0.314	0.0019
TAAGACGCTGTTCGCTTCG	0.235	0.375, 0.215	<b>5.03E-10</b>
CAAAATACTCACCTATCCTT	0.18	0.178, 0.181	0.8726
TGGGGCGCTCGCCTATCCTT	0.051	0.048, 0.052	0.7862
CAGGGCGCTCGCCCTATCCTT	0.047	0.054, 0.047	0.5824
TGAAGCGCTCGCCCTATCCTT	0.037	0.010, 0.042	0.0053
TAAGACGCTGTTCGCTTTT	0.036	0.041, 0.035	0.5747
TAAAACA CT CGCCCTATCCTT	0.023	0.012, 0.025	0.1736
TAGGGCGCTCGCCCTATCCTT	0.021	0.010, 0.022	0.1474
TAGGGCGTCTGTCCGCTTCG	0.02	0.022, 0.019	0.7191
CGGGCGCTCGCCCTATCCTT	0.014	0.003, 0.015	0.0913
TAAAATACTCACCTATCCTT	0.011	0.010, 0.012	0.7588
<b>Block 6</b>			
AT	0.507	0.377, 0.528	2.80E-07
GG	0.255	0.390, 0.233	1.17E-09
GT	0.238	0.233, 0.238	0.8437
<b>Block 7</b>			
TGATAGTGCGG	0.465	0.351, 0.485	8.44E-06
CGATAGTGCGG	0.183	0.163, 0.187	0.3078
TACCGACATTA	0.111	0.168, 0.102	6.00E-04
TACCGACATGG	0.109	0.158, 0.102	0.0031
TACCGACATTG	0.057	0.086, 0.053	0.0168
CGACAGTGCGG	0.029	0.031, 0.029	0.8039
TGACGGTGCGG	0.02	0.025, 0.020	0.5526
<b>Block</b>	<b>Frequencies</b>	<b>Case, Control Frequencies</b>	<b>p-value</b>
<b>Block 8</b>			
CGG	0.508	0.484, 0.511	0.3543
CGA	0.195	0.199, 0.194	0.8481
CAG	0.154	0.189, 0.149	0.0568
TGG	0.142	0.127, 0.144	0.4102

## **Reconstruction of phylogenetic tree**

Phylogenetic analysis is a powerful tool to study the relationship among the different sequences present in various haplotypes. To reconstruct a phylogenetic tree relating the haplotypes in block 5, we applied the parsimony method as described in PHYLIP version 3.6<sup>184</sup>. For this method, we used genotype data for 21 SNPs and 12 haplotype sequences present in block 5 (Table 15). The alleles for these 21 SNPs for three primates – Chimpanzee, Orangutan, and Monkey were downloaded from the UCSC genome browser. In addition, we chose Chimpanzee as the out-group since this would provide a reference by which to measure distances between the haplotypes and would help to determine the root of the phylogenetic tree when an actual ancestral sequence is not available. As shown in Figure 17, we found that the MPN risk haplotype and two other haplotypes that had higher frequency in MPN cases compared to controls were clustered together and form a separate clade (referred to as the “MPN risk haplotype group” below). We observed that the MPN risk haplotype group is an ancestral haplotype compared to the other haplotypes present in higher frequency in healthy controls than MPN cases. We also noted that the initial risk SNP rs10974944 and SNPs found in the present study have risk alleles that are ancestral alleles. Thus, our results are most consistent with the ancestral susceptibility model of disease as proposed by De Rienzo and Hudson<sup>147</sup>.

**Table 15 List of SNPs and haplotypes present in Block 5**

A) The list of 21 SNPs present in Block 5. MAF, minor allele frequency.

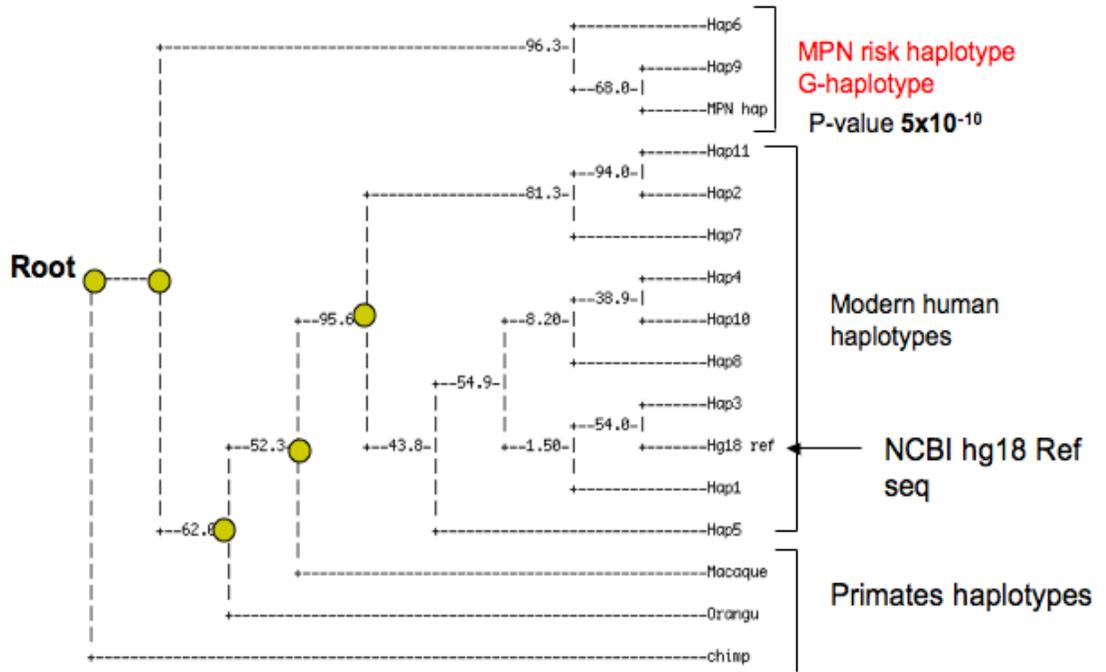
Number	Name	Position	MAF	Alleles
1	rs2274471	4975879	0.247	T:C
2	rs4372063	4993338	0.402	A:G
3	rs10119004	5061049	0.459	A:G
4	rs10974947	5062846	0.26	G:A
5	rs2230724	5071780	0.488	A:G
6	rs1410779	5073173	0.195	C:T
7	rs3824432	5081675	0.229	G:A
8	rs3780372	5087544	0.305	C:T
9	rs7870694	5090628	0.3	T:C
10	rs17425637	5100000	0.295	C:T
11	rs3780379	5102519	0.195	G:A
12	rs3824433	5103577	0.298	C:T
13	rs884132	5104522	0.295	C:T
14	rs1410780	5130115	0.051	C:T
15	rs10115962	5130841	0.298	T:C
16	rs10815167	5140058	0.299	A:G
17	rs7029084	5164638	0.298	T:C
18	rs7040922	5164829	0.298	C:T
19	rs7047795	5171467	0.298	C:T
20	rs7045342	5173892	0.261	T:C
21	rs12349508	5174222	0.263	T:G

B) 12 different haplotypes that were used for phylogenetic tree construction

Haplotype	Block 5	Case, Control Freq	P-value
Hap1	TGGGGCGCTCGCCCTATCCTT	0.227, 0.314	0.0019
<b>MPN hap</b>	TAAGACGTCTGTTCCGCTTCG	0.375, 0.215	5.03E-10
Hap2	CAAAATACTCACCCCTATCCTT	0.178, 0.181	0.8726
Hap3	TGGGGCGCTCGCCTTATCCTT	0.048, 0.052	0.7862
Hap4	CAGGGCGCTCGCCCTATCCTT	0.054, 0.047	0.5824
Hap5	TGAAGCGCTCGCCCTATCCTT	0.010, 0.042	0.0053
<b>Hap6</b>	TAAGACGTCTGTTCCGCTTTT	0.041, 0.035	0.5747
Hap7	TAAAACA CTGCCCCTATCCTT	0.012, 0.025	0.1736
Hap8	TAGGGCGCTCGCCCTATCCTT	0.010, 0.022	0.1474
<b>Hap9</b>	TAGGGCGTCTGTTCCGCTTCG	0.022, 0.019	0.7191
Hap10	CGGGGCGCTCGCCCTATCCTT	0.003, 0.015	0.0913
hap11	TAAAATACTCACCCCTATCCTT	0.010, 0.012	0.7588

**Figure 17 Phylogenetic tree of haplotypes in Block5.**

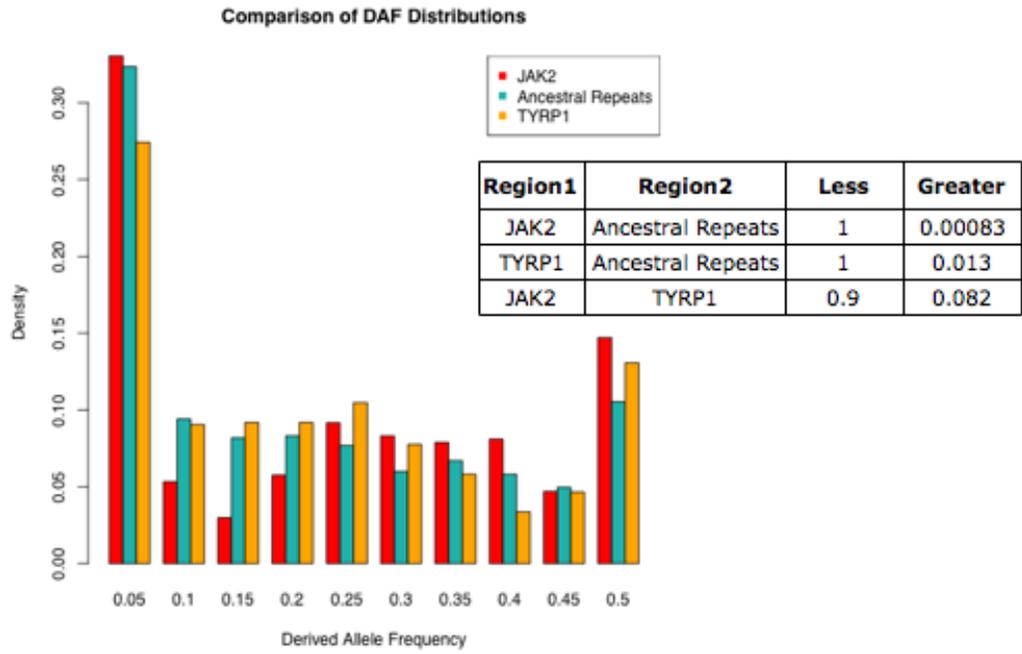
The 12 sequences (haplotypes) observed in block5, Chimpanzee, Orangutan, Monkey and NCBI36 hg18 human reference sequences obtained from the UCSC genome browser were used as input to PHYLIP to generate the phylogenetic tree.



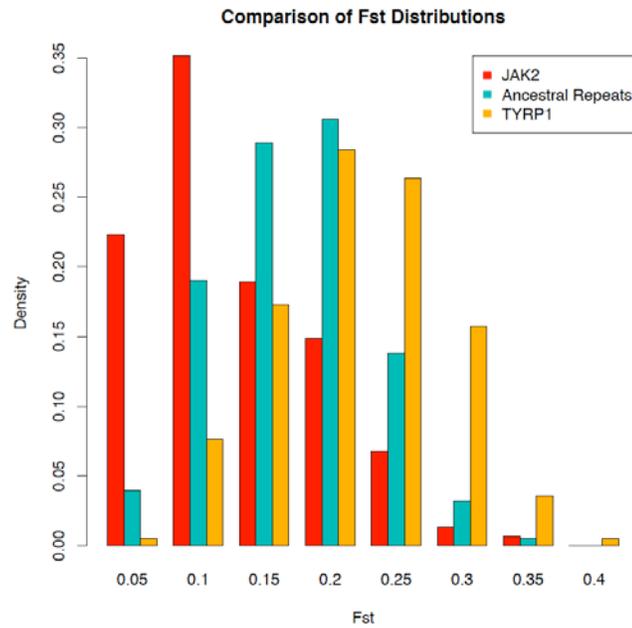
## **Selection pressure at the *JAK2* Locus**

We were next interested in assessing the *JAK2* locus for evidence of selection pressure. We compared the *JAK2* locus to the *TYRP1* gene, a positive control region on chromosome 9 that encodes a member of the melanin biosynthesis pathway and is known to be under positive selection. Additionally, we compared the *JAK2* locus to neutral ancestral repeats, which are regions of the genome that are not under selection pressure and can serve as negative controls. We first asked if the minor allele of a SNP was ancestral or derived and assigned derived allele frequency (DAF) for every SNP accordingly. We found that for every SNP in these regions the minor allele was always the derived allele. Since derived alleles are recently developed, they would more likely to be at lower frequency than the ancestral alleles unless they are under selection. We observed that the *JAK2* region exhibits a significantly different distribution of derived allele frequencies than that of both the *TYRP1* region and ancestral repeats by a two-sided Wilcoxon test (Figure 18). This result indicates that the *JAK2* region has a relatively higher proportion of derived alleles than both ancestral repeats and the *TYRP1* region. Next, using HapMap Phase III genotype data for 11 human populations, we measured population differentiation  $F_{st}$  scores in the three test regions. Figure 19 shows that the  $F_{st}$  distribution for *JAK2*, *TYRP1* and ancestral repeats is not significantly different. Hence, we can conclude that there is no evidence of the *JAK2* locus being under recent positive selection even though it exhibits an excess derived allele frequency compared to the regions under neutral selections.

**Figure 18 Distribution of derived allele frequencies at JAK2 locus, TYRP1 and ancestral repeats**



**Figure 19 Distribution of Fst calculated using 11 HapMap III population comparing JAK2 locus, TYRP1 and ancestral repeats**



<b>Region1</b>	<b>Region2</b>	<b>Two-sided</b>	<b>Less</b>	<b>Greater</b>
JAK2	Ancestral Repeats	3.55E-13	1.77E-13	1
TYRP1	Ancestral Repeats	2.22E-16	1	1.11E-16
JAK2	TYRP1	7.15E-27	3.58E-27	1

## 4.4 Discussion

Hereditary factors are known to influence susceptibility to MPN. Family members of affected individuals are themselves at high risk of MPN. Previous attempts to examine JAK2 in the familial MPN setting had led to efforts to identify such kindreds. For example, investigators in Sweden found a 5-to-7-fold increase of MPN in first-degree relatives of patients with an MPN (ref). Multiple types of MPN were observed in about 40% of these families, suggesting that they share a common predisposing genetic lesion. Three recent studies, including our previous GWAS, have provided strong evidence that JAK2 plays a critical role in MPN susceptibility and pathogenesis.

To better understand the JAK2 risk locus, we took advantage of the high-density SNP data to identify the risk-associated haplotypes in the region. Using HapMap Phase3 individuals of northern European ancestry (CEU), we found that risk SNPs in the JAK2 gene were located in an extended linkage disequilibrium spanning 300kb. We extracted 93 SNPs in this extended LD block that were genotyped in 166 MPN cases and 1037 shared controls for haplotype analysis. Upon analysis of these SNP with Haploview, we identified 10 haplotype blocks present in the 300kb region in our dataset. Notably one of these haplotypes was significantly associated with MPN risk and is identical to that discovered by Cross et al. (referred to as the “46/1”, “GGCC” or “MPN risk haplotype”)<sup>153,169,171</sup>.

To gain a better understand of the relationship among different haplotypes identified in our study, we next applied phylogenetic reconstruction. Focusing specifically on haplotypes in block 5 of the 300kb region and using chimpanzee as the out-group, we found that the MPN risk haplotype and other haplotypes with higher observed frequencies in MPN cases than healthy controls formed a separate clade from the haplotypes present at higher frequencies in healthy controls. Additionally, the MPN risk haplotype was estimated to an ancestral haplotype compared to the modern human sequence and showed highest similarity to the chimpanzee sequence. These results are consistent with the ancestral susceptibility model proposed by Rienzo and Hudson<sup>147</sup>.

Interestingly, there is growing evidence that ancestral alleles may play a role in cancer susceptibility<sup>185</sup>. In one example, the ancestral allele of SNP R72P, which is located in the TP53 gene, is associated with breast cancer<sup>186</sup>. Similarly, the ancestral alleles of SNP in MDM4 and MDM2 have been identified as breast cancer-associated risk alleles<sup>187</sup>. Notably, the MPN risk alleles identified our GWA studies were all ancestral alleles and tagged the MPN risk haplotype. It has been established by various groups that the MPN risk haplotype acquires the somatic V617F mutation, which is located in the pseudokinase domain of JAK2 and leads to constitutive activation of JAK2 and the JAK-STAT pathway. This results in aberrant cell proliferation in MPN patients. Although it remains unclear how the MPN risk haplotype and somatic mutation occur in *cis*, it may be possible that functional germline variant(s) in the haplotype interact with the somatic V617F mutation in a deleterious manner and make the development of clinically-manifested disease more likely. We can speculate that the deleterious

properties of this haplotype may have led to a decrease in its frequency over generations. Alternatively, we can envision that the ancestral human haplotype at *JAK2* once reflected ancient adaptations to previous environmental conditions and thus conferred selective advantages. However, with the onset of new environmental conditions, the ancestral alleles no may longer confer selective advantage and instead may lead to increased disease risk.

To understand the evolutionary pressure at *JAK2*, we performed tests to identify signatures of positive. In humans, several screens for positive selection based on variation within species as well as differences between chimpanzee and other primates have been performed. Using HapMap Phase3 population data, we confirmed that there was no evidence of population differentiation (as measured by  $F_{st}$ ) present in this locus, nor evidence of any significantly extended homozygosity. To evaluate the sensitivity of our methods to detect selection, we compared the *JAK2* results with those for the *TYRP1* gene region on chromosome 9 that is positively selected. Interestingly, the *JAK2* locus has an excess of derived allele frequencies compared to *TYRP1* gene.

In conclusion, we replicated the MPN risk haplotype at *JAK2* and, using phylogenetic tree analysis, found that MPN risk haplotype forms a separate, ancestral cluster that is distinct from haplotypes present in healthy individuals. Finally, although we found no strong evidence of recent positive selection at *JAK2*, we observed an excess of derived alleles compared to regions of the genome that are under positive selection.

## Implications

The use of controls from public databases as shared controls for GWA studies can result in improved power by increasing the number of controls without any extra cost of genotyping. We can adopt this approach to the next wave of genetic studies including whole genome sequencing to look for disease associated rare variant(s). Based on our study, it appears that when designing a genetic studies using shared controls, obtaining at least 10 controls for every case is extremely important. To deal with errors introduced due to data generated from different sources, we propose including some controls to be genotyped or sequenced along with the cases and compare these in-house controls with the shared controls obtained from public database to remove variants that show different frequencies between the two sets of controls.

Genome-wide analysis of MPN cases allowed us to identify a germline variant in the JAK2 gene that predisposes to the development of JAK2-mutant MPN. The JAK2 haplotype structure shows extended linkage disequilibrium in individuals from European and Asian ancestry whereas individuals from African ancestry as observed in HAPMAP data shows a lower level and distinct patterns of LD. Thus, genotyping all the associated genotyped or imputed variants at the JAK2 locus in MPN patients from African ancestry may lead to the identification of the causal variant(s). Further genetic studies in JAK2 negative MPN patients will shed light on the factors influencing the MPN phenotype.

Another theme that has emerged in the search for MPN susceptibility loci is the concept that JAK2 susceptibility variants predispose to other nonmalignant disorders like

Crohn's disease<sup>138</sup> and ulcerative colitis that is believed to have an inflammatory cause. These findings suggest the possibility of shared genetic pathways between these diseases. Further studies are needed to understand the biological importance of variation in the JAK2 gene region in relation to hematopoiesis and related disease phenotypes, including Crohn's disease<sup>138</sup>, ulcerative colitis, and MPN.

The fine mapping approach aims to narrow a region of association and pinpoint the causal variant(s) responsible. Rather than genotyping all known SNPs within the candidate region to resolve causal variant(s), 1000 Genomes data can be used to impute all the documented variants in the region for association test. Bioinformatic tools are a further refinement step for the prioritization of causal SNPs. Many tools, like the ENCYclopedia Of DNA Elements (ENCODE), which is hosted by the University of California Santa Cruz (UCSC), exist to enable identification of a candidate for the causal variant by utilizing prediction of functional effects to prioritize SNPs for downstream analysis. The aim of ENCODE is to find and document all the functional elements that exist in the genome in both coding and non-coding regions. This database essentially gathers its data from wet lab experiments. It includes data from a range of experiments in a variety of tissues and cell types including transcription factor binding sites, chromatin profiling, and histone modification. Data generated from wet lab experiments potentially offer greater evidence of putative function compared with the current predictive algorithms. Thus, an integration of genotyping, imputation, sequencing, bioinformatics and functional annotation can be used to fine-map the disease locus to prioritize the possible functional or causal variant(s) in a disease locus. The SNP(s) most likely to be

functional within the fine-mapped regions can be further followed up in laboratories using functional experiments to understand the biological implication of the disease-associated locus. Lastly, the evolutionary studies on various disease susceptibility loci may help us to understand the evolution of disease.

## References

1. SEER Cancer Statistics Review. Bethesda, MD: National Cancer Institute. (2004).
2. Vaquez, H. On a special form of cyanosis accompanied by excessive and persistent erythrocytosis. *Comp Rend Soc Biol.* 1892;12:384–388 (1892).
3. Heuck, G. Two cases of leukemia with peculiar blood and bone marrow findings, respectively. *Arch Pathol Anat.* 1879;78:475–496. (1879).
4. E Epstein , A.G. Hemorrhagic thrombocythemia with a cascular, sclerotic spleen. *Virchows Arch.* 1934;293:233–248. (1934).
5. Dameshek, W. Some speculations on the myeloproliferative syndromes. *Blood* **6**, 372-5 (1951).
6. Wasserman, L.R. The treatment of polycythemia. A panel discussion. *Blood* **32**, 483-7 (1968).
7. Murphy, S., Iland, H., Rosenthal, D. & Laszlo, J. Essential thrombocythemia: an interim report from the Polycythemia Vera Study Group. *Semin Hematol* **23**, 177-82 (1986).
8. Ruggeri, M., Tosetto, A., Frezzato, M. & Rodeghiero, F. The rate of progression to polycythemia vera or essential thrombocythemia in patients with erythrocytosis or thrombocytosis. *Ann Intern Med* **139**, 470-5 (2003).
9. Tefferi, A. & Murphy, S. Current opinion in essential thrombocythemia: pathogenesis, diagnosis, and management. *Blood Rev* **15**, 121-31 (2001).
10. Varki A, L.R., Griffith R. The syndrome of idiopathic myelofibrosis: a clinicopathologic review with emphasis on the prognostic variables predicting survival. . *Medicine (Baltimore).* 1983;62:353–371.
11. Tefferi, A. Myelofibrosis with myeloid metaplasia. *N Engl J Med* **342**, 1255-65 (2000).
12. Policitemia, G.I.S. Polycythemia vera: the natural history of 1213 patients followed for 20 years. Gruppo Italiano Studio Policitemia. *Ann Intern Med* **123**, 656-64 (1995).
13. Barosi, G. Myelofibrosis with myeloid metaplasia: diagnostic definition and prognostic classification for clinical studies and treatment guidelines. *J Clin Oncol* **17**, 2954-70 (1999).
14. James, C. et al. A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature* **434**, 1144-8 (2005).
15. Baxter, E.J. et al. Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *Lancet* **365**, 1054-61 (2005).
16. Kralovics, R. et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N Engl J Med* **352**, 1779-90 (2005).
17. Levine, R.L. et al. Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell* **7**, 387-97 (2005).
18. Ihle, J.N. & Gilliland, D.G. Jak2: normal function and role in hematopoietic disorders. *Curr Opin Genet Dev* **17**, 8-14 (2007).

19. Delhommeau, F. et al. Evidence that the JAK2 G1849T (V617F) mutation occurs in a lymphomyeloid progenitor in polycythemia vera and idiopathic myelofibrosis. *Blood* **109**, 71-7 (2007).
20. Jamieson, C.H. et al. The JAK2 V617F mutation occurs in hematopoietic stem cells in polycythemia vera and predisposes toward erythroid differentiation. *Proc Natl Acad Sci U S A* **103**, 6224-9 (2006).
21. Marty, C. et al. Myeloproliferative neoplasm induced by constitutive expression of JAK2V617F in knock-in mice. *Blood* **116**, 783-7 (2010).
22. Tefferi, A. et al. Proposals and rationale for revision of the World Health Organization diagnostic criteria for polycythemia vera, essential thrombocythemia, and primary myelofibrosis: recommendations from an ad hoc international expert panel. *Blood* **110**, 1092-7 (2007).
23. Verstovsek, S. et al. Safety and efficacy of INCB018424, a JAK1 and JAK2 inhibitor, in myelofibrosis. *N Engl J Med* **363**, 1117-27 (2010).
24. Pardanani, A. et al. JAK inhibitor therapy for myelofibrosis: critical assessment of value and limitations. *Leukemia* **25**, 218-25 (2011).
25. Kralovics, R., Guan, Y. & Prchal, J.T. Acquired uniparental disomy of chromosome 9p is a frequent stem cell defect in polycythemia vera. *Exp Hematol* **30**, 229-36 (2002).
26. Scott, L.M., Scott, M.A., Campbell, P.J. & Green, A.R. Progenitors homozygous for the V617F mutation occur in most patients with polycythemia vera, but not essential thrombocythemia. *Blood* **108**, 2435-7 (2006).
27. Wernig, G. et al. Expression of Jak2V617F causes a polycythemia vera-like disease with associated myelofibrosis in a murine bone marrow transplant model. *Blood* **107**, 4274-81 (2006).
28. Lacout, C. et al. JAK2V617F expression in murine hematopoietic cells leads to MPD mimicking human PV with secondary myelofibrosis. *Blood* **108**, 1652-60 (2006).
29. Stein, B.L. et al. Sex differences in the JAK2 V617F allele burden in chronic myeloproliferative disorders. *Haematologica* **95**, 1090-7 (2010).
30. Tiedt, R. et al. Ratio of mutant JAK2-V617F to wild-type Jak2 determines the MPD phenotypes in transgenic mice. *Blood* **111**, 3931-40 (2008).
31. Xing, S. et al. Transgenic expression of JAK2V617F causes myeloproliferative disorders in mice. *Blood* **111**, 5109-17 (2008).
32. Scott, L.M. et al. JAK2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. *N Engl J Med* **356**, 459-68 (2007).
33. Pardanani, A., Lasho, T.L., Finke, C., Hanson, C.A. & Tefferi, A. Prevalence and clinicopathologic correlates of JAK2 exon 12 mutations in JAK2V617F-negative polycythemia vera. *Leukemia* **21**, 1960-3 (2007).
34. Pietra, D. et al. Somatic mutations of JAK2 exon 12 in patients with JAK2 (V617F)-negative myeloproliferative disorders. *Blood* **111**, 1686-9 (2008).
35. Pikman, Y. et al. MPLW515L is a novel somatic activating mutation in myelofibrosis with myeloid metaplasia. *PLoS Med* **3**, e270 (2006).
36. Pardanani, A.D. et al. MPL515 mutations in myeloproliferative and other myeloid disorders: a study of 1182 patients. *Blood* **108**, 3472-6 (2006).

37. Beer, P.A. et al. MPL mutations in myeloproliferative disorders: analysis of the PT-1 cohort. *Blood* **112**, 141-9 (2008).
38. Delhommeau, F. et al. Mutation in TET2 in myeloid cancers. *N Engl J Med* **360**, 2289-301 (2009).
39. Tefferi, A. et al. TET2 mutations and their clinical correlates in polycythemia vera, essential thrombocythemia and myelofibrosis. *Leukemia* **23**, 905-11 (2009).
40. Gelsi-Boyer, V. et al. Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *Br J Haematol* **145**, 788-800 (2009).
41. Carbuccia, N. et al. Mutual exclusion of ASXL1 and NPM1 mutations in a series of acute myeloid leukemias. *Leukemia* **24**, 469-73 (2010).
42. Pardanani, A. et al. Recurrent IDH mutations in high-risk myelodysplastic syndrome or acute myeloid leukemia with isolated del(5q). *Leukemia* **24**, 1370-2 (2010).
43. Tefferi, A. et al. IDH1 and IDH2 mutation studies in 1473 patients with chronic-, fibrotic- or blast-phase essential thrombocythemia, polycythemia vera or myelofibrosis. *Leukemia* **24**, 1302-9 (2010).
44. Loh, M.L. et al. Mutations in CBL occur frequently in juvenile myelomonocytic leukemia. *Blood* **114**, 1859-63 (2009).
45. Mullighan, C.G. et al. BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453**, 110-4 (2008).
46. Jager, R. et al. Deletions of the transcription factor Ikaros in myeloproliferative neoplasms. *Leukemia* **24**, 1290-8 (2010).
47. Gery, S. et al. Lnk inhibits myeloproliferative disorder-associated JAK2 mutant, JAK2V617F. *J Leukoc Biol* **85**, 957-65 (2009).
48. Oh, S.T. et al. Novel mutations in the inhibitory adaptor protein LNK drive JAK-STAT signaling in patients with myeloproliferative neoplasms. *Blood* **116**, 988-92 (2010).
49. Ernst, T. et al. Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat Genet* **42**, 722-6 (2010).
50. Schaub, F.X. et al. Clonal analysis of TET2 and JAK2 mutations suggests that TET2 can be a late event in the progression of myeloproliferative neoplasms. *Blood* **115**, 2003-7 (2010).
51. Prchal, J.T. Classification and molecular biology of polycythemias (erythrocytoses) and thrombocytosis. *Hematol Oncol Clin North Am* **17**, 1151-8, vi (2003).
52. Arcasoy, M.O. & Karayal, A.F. Erythropoietin hypersensitivity in primary familial and congenital polycythemia: role of tyrosines Y285 and Y344 in erythropoietin receptor cytoplasmic domain. *Biochim Biophys Acta* **1740**, 17-28 (2005).
53. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends Genet* **17**, 502-10 (2001).
54. Bellanne-Chantelot, C. et al. Genetic and clinical implications of the Val617Phe JAK2 mutation in 72 families with myeloproliferative disorders. *Blood* **108**, 346-52 (2006).

55. Rumi, E. et al. Familial chronic myeloproliferative disorders: clinical phenotype and evidence of disease anticipation. *J Clin Oncol* **25**, 5630-5 (2007).
56. Landgren, O. et al. Increased risks of polycythemia vera, essential thrombocythemia, and myelofibrosis among 24,577 first-degree relatives of 11,039 patients with myeloproliferative neoplasms in Sweden. *Blood* **112**, 2199-204 (2008).
57. The International HapMap Project. *Nature* **426**, 789-96 (2003).
58. Chapman, J.M., Cooper, J.D., Todd, J.A. & Clayton, D.G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**, 18-31 (2003).
59. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-7 (1996).
60. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536-9 (1996).
61. Walker, F.O. Huntington's Disease. *Semin Neurol* **27**, 143-50 (2007).
62. Gusella, J.F. et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234-8 (1983).
63. Kerem, B. et al. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073-80 (1989).
64. Bertina, R.M. et al. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369**, 64-7 (1994).
65. Corder, E.H. et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921-3 (1993).
66. Altshuler, D. et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* **26**, 76-80 (2000).
67. Frazer, K.A. et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).
68. Moskvina, V., Craddock, N., Holmans, P., Owen, M.J. & O'Donovan, M.C. Effects of differential genotyping error rate on the type I error probability of case-control studies. *Hum Hered* **61**, 55-64 (2006).
69. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
70. Yang, Q., Cui, J., Chazaro, I., Cupples, L.A. & Demissie, S. Power and type I error rate of false discovery rate approaches in genome-wide association studies. *BMC Genet* **6 Suppl 1**, S134 (2005).
71. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat Med* **9**, 811-8 (1990).
72. Sabatti, C., Service, S. & Freimer, N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* **164**, 829-33 (2003).
73. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* **96**, 434-42 (2004).
74. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
75. Chanock, S.J. et al. Replicating genotype-phenotype associations. *Nature* **447**, 655-60 (2007).

76. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133-41 (2008).
77. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
78. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
79. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
80. Willer, C.J. et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161-9 (2008).
81. Sanna, S. et al. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet* **40**, 198-203 (2008).
82. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-23 (2009).
83. Cooper, R.S., Tayo, B. & Zhu, X. Genome-wide association studies: implications for multiethnic samples. *Hum Mol Genet* **17**, R151-5 (2008).
84. Farrer, L.A. et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* **278**, 1349-56 (1997).
85. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
86. Klein, R.J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-9 (2005).
87. Yamazaki, K. et al. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum Mol Genet* **14**, 3499-506 (2005).
88. Ozaki, K. & Tanaka, T. Genome-wide association study to identify SNPs conferring risk of myocardial infarction and their functional analyses. *Cell Mol Life Sci* **62**, 1804-13 (2005).
89. Duerr, R.H. et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461-3 (2006).
90. Sladek, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-5 (2007).
91. Kathiresan, S. et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* **40**, 189-97 (2008).
92. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-8 (2010).
93. Visscher, P.M. Sizing up human height variation. *Nat Genet* **40**, 489-90 (2008).
94. Speliotes, E.K. et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**, 937-48 (2010).
95. Heid, I.M. et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* **42**, 949-60 (2010).

96. Easton, D.F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-93 (2007).
97. Ahmed, S. et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet* **41**, 585-90 (2009).
98. Hunter, D.J. et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**, 870-4 (2007).
99. Thomas, G. et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* **41**, 579-84 (2009).
100. Stacey, S.N. et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **39**, 865-9 (2007).
101. Gold, B. et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc Natl Acad Sci U S A* **105**, 4340-5 (2008).
102. Kirchhoff, T. et al. The 6q22.33 locus and breast cancer susceptibility. *Cancer Epidemiol Biomarkers Prev* **18**, 2468-75 (2009).
103. Zheng, W. et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* **41**, 324-8 (2009).
104. Yeager, M. et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645-9 (2007).
105. Yeager, M. et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet* **41**, 1055-7 (2009).
106. Gudmundsson, J. et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39**, 631-7 (2007).
107. Eeles, R.A. et al. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* **40**, 316-21 (2008).
108. Gudmundsson, J. et al. Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet* **40**, 281-3 (2008).
109. Eeles, R.A. et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* **41**, 1116-21 (2009).
110. Amundadottir, L.T. et al. A common variant associated with prostate cancer in European and African populations. *Nat Genet* **38**, 652-8 (2006).
111. Amos, C.I. et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* **40**, 616-22 (2008).
112. Wang, Y. et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* **40**, 1407-9 (2008).
113. Hung, R.J. et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633-7 (2008).
114. Jaeger, E. et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet* **40**, 26-8 (2008).
115. Tomlinson, I. et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* **39**, 984-8 (2007).
116. Broderick, P. et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* **39**, 1315-7 (2007).

117. Zanke, B.W. et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* **39**, 989-94 (2007).
118. Houlston, R.S. et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* **40**, 1426-35 (2008).
119. Tenesa, A. et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* **40**, 631-7 (2008).
120. Tomlinson, I.P. et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* **40**, 623-30 (2008).
121. Wu, X. et al. Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat Genet* **41**, 991-5 (2009).
122. Kiemeny, L.A. et al. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat Genet* **40**, 1307-12 (2008).
123. Amundadottir, L. et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet* **41**, 986-90 (2009).
124. Petersen, G.M. et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet* **42**, 224-8 (2010).
125. Di Bernardo, M.C. et al. A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat Genet* **40**, 1204-10 (2008).
126. Crowther-Swanepoel, D. et al. Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet* **42**, 132-6 (2010).
127. Skibola, C.F. et al. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat Genet* **41**, 873-5 (2009).
128. Shete, S. et al. Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* **41**, 899-904 (2009).
129. Song, H. et al. A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nat Genet* **41**, 996-1000 (2009).
130. Hindorff, L.A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).
131. Graham, R.R. et al. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc Natl Acad Sci U S A* **104**, 6758-63 (2007).
132. Sigurdsson, S. et al. Polymorphisms in the tyrosine kinase 2 and interferon regulatory factor 5 genes are associated with systemic lupus erythematosus. *Am J Hum Genet* **76**, 528-37 (2005).
133. Graham, R.R. et al. A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nat Genet* **38**, 550-5 (2006).

134. Dideberg, V. et al. An insertion-deletion polymorphism in the interferon regulatory Factor 5 (IRF5) gene confers risk of inflammatory bowel diseases. *Hum Mol Genet* **16**, 3008-16 (2007).
135. Stahl, E.A. et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* **42**, 508-14 (2010).
136. Verlaan, D.J. et al. Targeted screening of cis-regulatory variation in human haplotypes. *Genome Res* **19**, 118-27 (2009).
137. Moffatt, M.F. et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470-3 (2007).
138. Barrett, J.C. et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-62 (2008).
139. Barrett, J.C. et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* **41**, 703-7 (2009).
140. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-9 (2010).
141. Jia, L. et al. Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet* **5**, e1000597 (2009).
142. Pomerantz, M.M. et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**, 882-4 (2009).
143. Wasserman, N.F., Aneas, I. & Nobrega, M.A. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res* **20**, 1191-7 (2010).
144. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
145. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**, 124-37 (2001).
146. Chakravarti, A. Population genetics--making sense out of sequence. *Nat Genet* **21**, 56-60 (1999).
147. Di Rienzo, A. & Hudson, R.R. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet* **21**, 596-601 (2005).
148. Stengard, J.H. et al. Apolipoprotein E polymorphism, Alzheimer's disease and vascular dementia among elderly Finnish men. *Acta Neurol Scand* **92**, 297-8 (1995).
149. de Knijff, P., van den Maagdenberg, A.M., Frants, R.R. & Havekes, L.M. Genetic heterogeneity of apolipoprotein E and its influence on plasma lipid and lipoprotein levels. *Hum Mutat* **4**, 178-94 (1994).
150. Strittmatter, W.J. et al. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A* **90**, 1977-81 (1993).
151. Neel, J.V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet* **14**, 353-62 (1962).
152. Crowther-Swanepoel, D. et al. Genetic variation in CXCR4 and risk of chronic lymphocytic leukemia. *Blood* **114**, 4843-6 (2009).
153. Kilpivaara, O. et al. A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. *Nat Genet* **41**, 455-9 (2009).

154. Zhuang, J.J. et al. Optimizing the power of genome-wide association studies by using publicly available reference samples to expand the control group. *Genet Epidemiol* **34**, 319-26.
155. Price, A.L. et al. Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet* **4**, e236 (2008).
156. Tian, C. et al. European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. *Mol Med* **15**, 371-83 (2009).
157. Li, Q. & Yu, K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol* **32**, 215-26 (2008).
158. Klein, R.J. Power analysis for genome-wide association studies. *BMC Genet* **8**, 58 (2007).
159. Li, C. & Li, M. GWAsimulator: a rapid whole-genome simulation program. *Bioinformatics* **24**, 140-2 (2008).
160. Tian, C. et al. Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet* **4**, e4 (2008).
161. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98-101 (2008).
162. Paschou, P. et al. Tracing sub-structure in the European American population with PCA-informative markers. *PLoS Genet* **4**, e1000114 (2008).
163. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
164. Pluzhnikov, A. et al. Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. *Am J Hum Genet* **87**, 123-8 (2010).
165. Campbell, P.J. & Green, A.R. The myeloproliferative disorders. *N Engl J Med* **355**, 2452-66 (2006).
166. Levine, R.L. et al. X-inactivation-based clonality analysis and quantitative JAK2V617F assessment reveal a strong association between clonality and JAK2V617F in PV but not ET/MMM, and identifies a subset of JAK2V617F-negative ET and MMM patients with clonal hematopoiesis. *Blood* **107**, 4139-41 (2006).
167. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
168. Laken, S.J. et al. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet* **17**, 79-83 (1997).
169. Jones, A.V. et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat Genet* **41**, 446-9 (2009).
170. Olcaydu, D. et al. The role of the JAK2 GGCC haplotype and the TET2 gene in familial myeloproliferative neoplasms. *Haematologica* **96**, 367-74 (2011).
171. Olcaydu, D. et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. *Nat Genet* **41**, 450-4 (2009).
172. Koboldt, D.C. et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-5 (2009).

173. Sandelin, A., Wasserman, W.W. & Lenhard, B. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res* **32**, W249-52 (2004).
174. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).
175. Anderson, C.A. et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* **43**, 246-52 (2011).
176. McGovern, D.P. et al. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat Genet* **42**, 332-7 (2010).
177. Jones, A.V. et al. The JAK2 46/1 haplotype predisposes to MPL-mutated myeloproliferative neoplasms. *Blood* **115**, 4517-23 (2010).
178. Olcaydu, D. et al. The 'GGCC' haplotype of JAK2 confers susceptibility to JAK2 exon 12 mutation-positive polycythemia vera. *Leukemia* **23**, 1924-6 (2009).
179. Fullerton, S.M. et al. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am J Hum Genet* **67**, 881-900 (2000).
180. Song, Y., Niu, T., Manson, J.E., Kwiatkowski, D.J. & Liu, S. Are variants in the CAPN10 gene related to risk of type 2 diabetes? A quantitative assessment of population and family-based association studies. *Am J Hum Genet* **74**, 208-22 (2004).
181. Weedon, M.N. et al. Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. *Am J Hum Genet* **73**, 1208-12 (2003).
182. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-5 (2005).
183. Gabriel, S.B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
184. Felsenstein, J. Mathematics vs. Evolution: Mathematical Evolutionary Theory. *Science* **246**, 941-2 (1989).
185. Puente, X.S. et al. Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC Genomics* **7**, 15 (2006).
186. Osorio, A. et al. A haplotype containing the p53 polymorphisms Ins16bp and Arg72Pro modifies cancer risk in BRCA2 mutation carriers. *Hum Mutat* **27**, 242-8 (2006).
187. Atwal, G.S. et al. Altered tumor formation and evolutionary selection of genetic variants in the human MDM4 oncogene. *Proc Natl Acad Sci U S A* **106**, 10236-41 (2009).