Aus dem Deutschen Krebsforschungszentrum Heidelberg
(Geschäftsführender Direktor: Prof. Dr. Michael Baumann)

Forschungsschwerpunkt Krebsrisikofaktoren und Prävention
Abteilung für Molekulargenetische Epidemiologie
(Abteilungsleiter: Prof. Dr. Kari Hemminki)

Inherited genetic susceptibility to multiple myeloma and related diseases

Inauguraldissertation

Zur Erlangung des Doctor scientiarum humanarum (Dr.sc.hum.)

an der

Medizinischen Fakultät Heidelberg

der

Ruprecht-Karls-Universität

vorgelegt von

Subhayan Chattopadhyay

aus

Asansol, India

2018

Dekan:	Prof. Dr. Andreas Draguhn
Doktorvater:	Prof. Dr. Kari Hemminki

"*I believe that no one who is familiar, either with mathematical advances in other fields, or with the range of special biological conditions to be considered, would ever conceive that everything could be summed up in a single mathematical formula, however complex.*"

Sir Ronald Fisher

The evolutionary modification of genetic phenomena.
Proceedings of the 6th International Congress of Genetics 1, 165-72, 1932

Table of Contents

Foreword

The work in this thesis is entirely of my own unless otherwise explicitly stated. Entirety of the thesis work is described in three separate articles (see **Publications**).

Professor Hartmut Goldschmidt, Professor Kari Hemminki and Dr. Asta Försti were involved in procurement and management of genotype data on German monoclonal gammopathy of undetermined significance and multiple myeloma samples and gene expression data on selected multiple myeloma samples. Professor Gareth J Morgan and Professor Richard Houlston provided genotype data on UK multiple myeloma samples. Professor Karl-Heinz Jöckel provided genotype data on German monoclonal gammopathy of undetermined significance cases and German controls from Heinz-Nixdorf Recall study. Professor Jan Sundquist, Professor Kristina Sundquist and Professor Kari Hemminki provided access to Swedish Family Cancer Database.

Abbreviations

A2BP1	RNA Binding Fox-1 Homolog 1
AKAP12	A-Kinase Anchoring Protein 12
Akt	Protein Kinase B
AL amyloidosis	Amyloid light chain amyloidosis
ALK	Anaplastic Lymphoma Receptor Tyrosine Kinase
ALSPAC	Avon Longitudinal Study of Parents and Children
APC	Adenomatosis Polyposis Coli Tumor Suppressor
BCL6	B Cell CLL/Lymphoma 6
BMP	Bone morphogenetic protein
BNC2	Basonuclin 2
BRAF	B-Raf Proto-Oncogene, Serine/Threonine Kinase
BRCA1/2	Breast cancer early onset 1/2,
C6orf211	Chromosome 6 open reading frame 211
CALM3	Calmodulin 3
CDCA7L	Cell Division Cycle Associated 7 Like
CDH	Cadherin
CDH2/13	Cadherin 2/13
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A
CEU	Utah Residents With Northern And Western European Ancestry
CHB	Han Chinese in Beijing, China
CHEK2	Checkpoint Kinase 2
CI	Confidence interval
CMML	Chronic myelomonocytic leukemia
CRAB	Calcium, Renal insufficiency, Anemia, or Bone lesions
CRYL1	Crystallin Lambda 1
CUP	Cancer of unknown primary
ECM	Extra cellular matrix
EGFR	Epidermal growth factor receptor
eQTL	expression quantitative trait loci
FAM43A	Family With Sequence Similarity 43 Member A
FLC	Free light chain
FWER	Family-wise error rate
GALNT1	Polypeptide N-Acetylgalactosaminyltransferase 1
GCTA	Genome-wide Complex Trait Analysis
GEP5/70	Gene expression profiling 5, 70

GLCCI1	Glucocorticoid Induced 1
GNAQ	G Protein Subunit Alpha Q
GO	Gene ontology
GRB7	Growth Factor Receptor Bound Protein 7
GWAS	Genome wide association analysis
GWIS	Genome wide interaction analysis
HDAC1/2/9	Histone deacetylase 1/2/9
HER	Human epidermal receptor
HIF1-alpha	Hypoxia Inducible Factor 1 Subunit Alpha
HLA	Human leukocyte antigen
HNR	Heinz-Nixdorf Recall
HWE	Hardy-Weinberg equilibrium
IBS	Identity by state
ICD	International classification of disease
Ig	Immunoglobulin
IGFN1	Immunoglobulin-Like And Fibronectin Type III Domain Containing 1
IL6/17	Interleukin 6/17
IMWG	International Myeloma working Group
IRF8	Interferon Regulatory Factor 8
ISS	International staging system
JPT	Japanese In Tokyo, Japan
KEGG	Kyoto Encyclopedia of Genes and Genomes
KLF5	Kruppel Like Factor 5
KRAS	KRAS Proto-Oncogene, GTPase
LD	Linkage disequilibrium
LDSC	Linkage disequilibrium score
MAF	Minor allele frequency
MAPK	Mitogen-Activated Protein Kinase
MeSH	Medical subject heading
MGUS	Monoclonal gammopathy of undetermined significance
MLH1	MutL Homolog 1
MM	Multiple myeloma
MOCS	maximum of Chi-square
MP	Mammalian Protein
MR	Mendelian randomization
MSH2	MutS Homolog 2
mTOR	Mammalian target of rapamycin
NCAM2	Neural Cell Adhesion Molecule 2

NCBI	National Centre For Biotechnology Information
NF-κB	Nuclear Factor Kappa-Light-Chain-Enhancer Of Activated B Cells
NKX3-2	NK3 Homeobox 2
NRG	Neuregulin
OR	Odds ratio
PALB2	Partner And Localizer Of BRCA2
PANTHER	Protein Analysis through Evolutionary Relationships
PCA	Principal component analysis
PCGC	Phenotype Correlation-Genotype Correlation
PDE3B	Phosphodiesterase 3B
PI3K	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase
PLCB1	Phospholipase C Beta 1
PREX1	Phosphatidylinositol-3,4,5-Trisphosphate Dependent Rac Exchange Factor 1
PTP	Protein tyrosine phosphatase
PTPRD	Protein Tyrosine Phosphatase, Receptor Type D
QC	Quality control
RAB28	RAB28, Member RAS Oncogene Family
RALYL	RALY RNA Binding Protein Like
RARA	Retinoic Acid Receptor Alpha
REV-ERBα	Nuclear Receptor Subfamily 1 Group D Member 1(NR1D1)
RIP1/3	Receptor Interacting Serine/Threonine Kinase 1/3
R-ISS	Revised-International staging system
RNF41	Ring finger protein 41
RORA	RAR Related Orphan Receptor A
RORγt	RAR-related orphan receptor gamma 2
ROTI	related organ or tissue impairment
RR	Relative risk
RUNX1/2	Runt Related Transcription Factor 1/2
sAML	secondary-acute myeloid leukemia
SCC	Squamous cell carcinoma
SD	standard deviation
SERPINB9	Serpin Family B Member 9
SETBP1	SET Binding Protein 1
SFCD	Swedish Family Cancer Database
SMAD	Mothers Against DPP Homolog
SMM	Smoldering multiple myeloma
SMR	Summary-data-based Mendelian Randomization
SNOMED	Systematized Nomenclature of Medicine

SNP	Single nucleotide polymorphism
SOCS	Sum of Chi-square
SPC	Second primary cancer
STAT3	Signal Transducer And Activator Of Transcription 3
TGFβ	Transforming growth factor beta
Th17	T helper cell 17
TNC	Tenascin C
TNFRSF	Tumor necrosis factor receptor superfamily
TUT7	Terminal Uridylyl Transferase 7
UCSC	University Of California, Santa Cruz
VEGF	Vascular endothelial growth factor
WAX	Acyl-CoA Wax Alcohol Acyltransferase 2
WTCCC	Welcome Trust Case Control Consortium
W-Z	Wellek-Ziegler
YRI	Yoruba in Ibadan, Nigeria
ZCCHC6/11	Zinc Finger CCHC-Type Containing 6/11
ZNF224/229	Zinc Finger Protein 224/229

List of Figures

List of Tables

Introduction

Chapter 1: Multiple myeloma and related diseases

Multiple myeloma (MM), a rapidly progressing plasma cell dyscrasia, is a neoplastic growth in terminally differentiated plasma cells. Plasmocytes are the antibody producing cells of our immune system and in MM, monoclonal plasma cells proliferate in abundance (a state more recognized as monoclonal gammopathy) to produce plasmacytoma. These malignant plasma cells, although usually reside in host bone marrow, can also be found in peripheral blood, soft tissues and organs, predominantly towards the end stages of the disease (Gonsalves *et al.*, 2014). With the highest incidence observed in developed countries in Western Europe, northern America and Australia, MM accounts for 1.7% of all malignancies and almost 10% of all hematological cancers (Siegel *et al.*, 2016). However, the most common plasma cell dyscrasia is a benign precursor of MM termed monoclonal gammopathy of undetermined significance (MGUS). Towards a malignant progression to MM, MGUS is succeeded by yet another asymptomatic stage called smoldering multiple myeloma (SMM) and rarely MGUS also progresses to amyloid light-chain amyloidosis (AL amyloidosis). These four phases broadly inscribe the MM disease family.

Whilst the specific cause of MM is unknown, an array of environmental exposures are hypothesized to predispose to MM, such as ionizing radiation, pesticides, certain solvents, benzene, petroleum products, infectious agents and hair dye but without much precedence (Altekruse *et al.*, 1999a; Bergsagel *et al.*, 1999; Burmeister, 1981a; Khuder and Mutgi, 1997; Kuznetsova *et al.*, 2016). Genetic susceptibility to MM on the other hand is a long-proposed theory with familial studies speculating familial aggregation due to inherited MM risk (Maldonado and Kyle, 1974). However, direct compelling evidence of inherited

susceptibility to MM was, until recently, largely undescribed along with the possible mechanisms responsible for the apparently sporadic progression of MGUS to the later stages of malignancy. Historically, gammopathies in general have presented investigators with confusion in clinical, cytological and etiological characterization since as early as the early twentieth century and we are still far from understanding genetic underpinnings of this class of plasma cell dyscrasias.

1.1. History of *Gammopathy*

Gammopathies are comprised of several different conditions distinguished with clinical characterization of abnormal proliferation of cells of lymphoid lineage producing immunoglobulins (Ig). This class of plasma cell disorders was historically also known as hyperproteinemia due to abundance of Ig in blood serum. Swedish hematologist Jan G. Waldenström first hypothesized the concept of monoclonal and polyclonal gammopathy in Harvey lecture series in 1961 where he also lucidly speculated on the disease severity and possible transformation to a malignant state (Waldenstrom, 1961). Whilst monoclonal gammopathy meant an increased production of a single clone of immunoglobulin (mostly gamma globulin, the condition is also known as hypergammaglobulinemia), polyclonal gammopathies were a result of aberrant proliferation of several different immunoglobulin clones. Waldenström labeled individuals showing a fine band of hypergammaglobulinemia as harboring monoclonal protein. Even though a handful of these patients had or later developed MM, several of them initially did not show evidence of malignancy hence were described to have “essential hypergammaglobulinemia” or benign monoclonal gammopathy (Kyle and Anderson, 1997). He also later went on to coin the term “monoclonal

gammopathy of unknown etiology" while describing this unusual asymptomatic condition distinguishing it from other paraproteinemias.

A concrete case study of MGUS progressing to MM first came to light in 1966 raising more questions than it answered on the apparent *benign* status of MGUS. As professor of Medicine and Laboratory Medicine and Pathology at Mayo Clinic College of Medicine, Rochester, Minnesota, Robert A. Kyle had been studying a local cohort which he followed since 1945 who shared a MM-like electrophoretic Ig pattern (Kyle *et al.*, 1960). After 20 years of follow-up, one of the subjects developed severe MM in 1964 after undergoing a short phase of myelomatosis starting in 1963. In his seminal work citing this aberrant prognosis of MGUS to MM, Dr. Kyle first reported the notorious *M spike* (albumin – gamma globulin spike) on serum protein electrophoresis with evidence of abnormal plasma cell presence in bone-marrow aspirate (**Figure 1.1**) (Kyle and Bayrd, 1966).

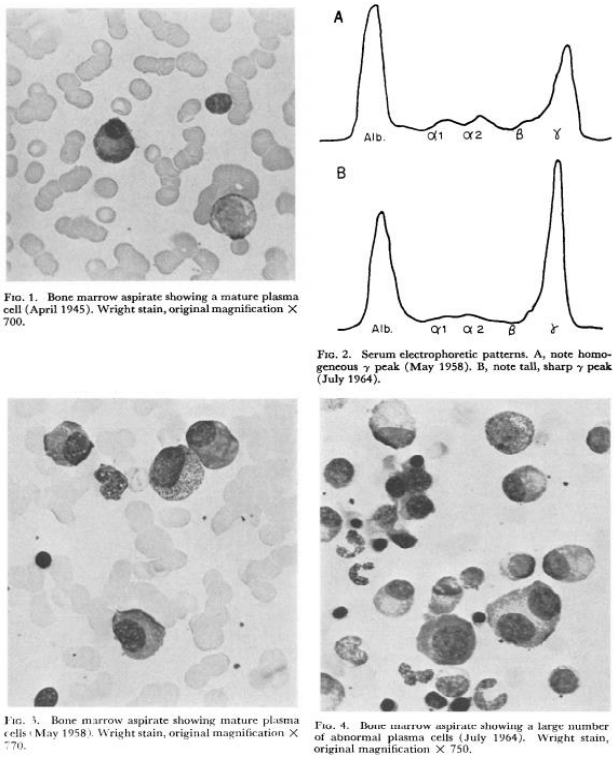


FIG. 1. Bone marrow aspirate showing a mature plasma cell (April 1945). Wright stain, original magnification × 700.

FIG. 2. Serum electrophoretic patterns. A, note homogeneous γ peak (May 1958). B, note tall, sharp γ peak (July 1964).

FIG. 3. Bone marrow aspirate showing mature plasma cells (May 1958). Wright stain, original magnification × 770.

FIG. 4. Bone marrow aspirate showing a large number of abnormal plasma cells (July 1964). Wright stain, original magnification × 750.

Figure 1. 1| Bone marrow aspirate and serum electrophoretic pattern, Kyle R. A. et al., 1966.
After 20 years of follow-up, the first thoroughly reported cases of MM subject to progression from MGUS as reported by R. A. Kyle in 1966.

In a previous observational study published in 1964, Waldenström had followed twelve patients of essential hypergammaglobulinemia from 1955 although till the conclusion of the study none of the patients developed a sustained malignancy. Only one had the ɣ-globulin related high but stable M spike signature (WaldenstrÖM, 1964). Consequently in the case report from the Mayo Clinic Dr. Kyle argued "… *this 'chronic benign process' can erupt into a progressive and serious phase at a much later date, and extends our understanding of plasma proliferative disease*."

1.2. Clinical characterization

Etiology of MGUS remains unclear to date; yet a handful of studies have established a role of genetic and environmental factors in its development (Boursi *et al.*, 2016; Korde *et al.*, 2011a; Kyle *et al.*, 2010; Landgren *et al.*, 2009). As discussed previously, contrary to the great variety of normal immunoglobulins, monoclonal gammopathy dictates a condition predominated by a single abnormal cell line. It usually yields an intact immunoglobulin free light chains but not heavy chains, however, rarely it can also produce heavy chains exclusively. Conspicuously, such abnormal cell line yields only a κ or a λ light chain, never the both. Consequently numerous discrete clinical types (IgM, non-IgM and light-chain MGUS) have arose to have established it as a clinically heterogeneous disorder. In general, MGUS is characterized with a serum M-protein <30 g/l, <10% clonal plasma cells in the bone marrow and absence of end-organ damage. Contextually, the end organ damage is a frequently observed phenotype infested in clonal plasma cell proliferative disorders.

SMM is the next stage of progression without myeloma-defining end-organ damage and is characterized by the presence of ≥ 30 g/L serum M-protein and/or 10 – 60 % bone marrow clonal plasma cell infiltration (Gao *et al.*, 2015). International Myeloma Working Group

(IMWG) defines Myeloma-defining end-organ damage with *CRAB* criteria. IMWG dictates "***r****elated* ***o****rgan or* ***t****issue* ***i****mpairment (ROTI)(end-organ damage), which is typically manifested by increased* ***c****alcium,* ***r****enal insufficiency,* ***a****naemia, or* ***b****one lesions (CRAB) attributed to the plasma cell proliferative process. Symptomatic myeloma requires evidence of ROTI*" (Anonymous, 2003).

AL amyloidosis is characterized by systemic accumulation of monoclonal Ig light chains synthesized by a bone marrow plasma cell clone in the form of misfolded amyloid protein deposits in tissues and other vital organs (heart, kidney, liver). The organ involvement pattern here is unclear and complex. A heart involvement construes to the majority followed by that of kidney, liver, peripheral nerve and soft tissues in gastrointestinal tract and lung. Detailed diagnostic criteria for the four diseases can be found in later update of IMWG definition (**Table 1.1**).

Table 1. 1| Diagnostic criteria for plasma cell disorders (taken from published IMWG definition)

Plasma cell disorder	Definition
Non-IgM (MGUS)	Serum monoclonal protein <30g/L Clonal bone marrow plasma cells <10% Absence of end-organ damage such as hypercalcemia, renal insufficiency, anemia, and bone lesions (CRAB) or amyloidosis that can be attributed to the plasma cell proliferative disorder
IgM MGUS	Serum IgM monoclonal protein <30g/L No evidence of anemia, constitutional symptoms, hyper viscosity, lymphadenopathy, hepatosplenomegaly, or other end-organ damage that can be attributed to the plasma cell proliferative disorder
Light-chain MGUS	Abnormal FLC ratio (<0.26 or >1.65) Increased level of the appropriate free light chain (increased FLC in patients with ratio >1.65 and increased FLC in patients with ratio <0.26) No immunoglobulin heavy chain expression on immunofixation Absence of end-organ damage such as hypercalcemia, renal insufficiency, anemia, and bone lesions (CRAB) or amyloidosis that can be attributed to the plasma cell proliferative disorder Clonal bone marrow plasma cells <10% Urinary monoclonal protein <500mg/24h
SMM	Both criteria must be met Serum monoclonal protein (IgG or IgA) 3 gm/dl and/or clonal bone marrow plasma cells 10%, and Absence of end-organ damage such as lytic bone lesions, anemia, hypercalcemia, or renal failure that can be attributed to a plasma cell proliferative disorder
MM	All three criteria must be met except as noted Clonal bone marrow plasma cells 10% Presence of serum and/or urinary monoclonal protein (except in patients with non-secretory multiple myeloma), and Evidence of end organ damage that can be attributed to the underlying plasma cell proliferative disorder, specifically Hypercalcemia: serum calcium 11.5 mg/dl or Renal insufficiency: serum creatinine >1.73 mmol/l) Anemia: normochromic, normocytic with a hemoglobin value of >2 g/dl below the lower limit of normal or a hemoglobin value <10 g/dl Bone lesions: lytic lesions, severe osteopenia or pathological fractures
AL amyloidosis	Presence of an amyloid-related systemic syndrome (e.g., renal, liver, heart, gastrointestinal tract, or peripheral nerve involvement) Positive amyloid staining by Congo red in any tissue (e.g., fat aspirate, bone marrow, or organ biopsy) Evidence that amyloid is light-chain-related established by direct exmination of the amyloid using mass spectrometry-based proteomic analysis or immunoeletronmicroscopy Evidence of a monoclonal plasma cell proliferative disorder (serum monoclonal protein, abnormal free light chain ratio, or clonal plasma cells in the bone marrow)

1.3. Disease progression

The cumulative probability of progression of MGUS to MM was 12% at 10 years, 25% at 20 years and 30% at 25 years (Kyle *et al.*, 2010). Individual risk of MM development was roughly 1% every year (Kyle *et al.*, 2002). At the time of recognition of MGUS, it is very difficult to predict the progression patterns to identify patients who will observe a stable condition compared to those who would observe a severely progressive disease as the underlying mechanism of prognostication is yet unclear. Nonetheless, the type of M-protein, size of the M-protein, the free light chain (FLC) ratio and the number of bone marrow clonal plasma cells present are some of the reliable indicators in identifying patients at a higher risk of further progression. At the time of recognition of MGUS, size of the M-protein is shown to be the most reliable prognosticator of progression to SMM (Kyle *et al.*, 2010). The same study estimated the risk of development of MM defining characteristic or a related condition after 20 years from MGUS diagnosis to be 49% for individuals with a 25 g/l level of M-protein, in comparison to a merely 14% for patients with an early M-spike of 5 g/l or less. Estimated risk of progression with a15 g/l M-protein abundance was two-fold in excess to that of individuals with 5 g/l. The IgM and IgA clones are in general more susceptible to progression compared to the IgG clonal MGUS. Several studies also report a monotonous proportional relation among risk of progression and abundance of clonal plasma cells in bone marrow with probable increase risk of up to 37% (Baldini *et al.*, 1996; Cesana *et al.*, 2002). Similarly progression risk is found in excess for patients with elevated FLC ratio than in those without; and this is an independent marker of progression since FLC ratio does not depend on the type or size of serum monoclonal protein (Kyle *et al.*, 2010; Rajkumar *et al.*, 2005).

Chapter 2: Population epidemiology

Proper demographic estimation of MGUS related events are difficult due to a number of reasons. Firstly, it is an asymptomatic disorder which means that there is little possibility of tracking individuals with MGUS by systematic registration at time of diagnosis. Secondly, as MGUS is associated with a rate of progression to MM of around 1% per year, additionally to SMM or AL amyloidosis with similar proportion, MGUS patients require follow-up to ascertain future events. However, the spontaneous discovery of MGUS is not uncommon and very rarely is associated to an individual's primary health-related issue. The caveats thus presented result in under-diagnosis of MGUS in routine clinical practice and hinders planning preventive strategies based on it.

2.1 Incidence, prevalence and mortality: worldwide and in Europe

MGUS infests in 3.2% of all individuals over the age of 50 years and around 5.3% of the people aged 70 years or older (Kyle *et al.*, 2006). For men, age adjusted prevalence rates for MGUS were found higher (4.0 per 100) in comparison to women (2.7 per 100). Irrespective of sex, risk of MGUS increases monotonously in comparison with age. Yearly incidence of MGUS paints a similar picture. For all men over 50 years of age, annual incidence is 120 per 100,000 which increases up to 530 for men older than 90 years of age. Whereas for women above 50 years of age incidence is 60 per 100,000 which goes up to 370 for women aged 90 or more (Therneau *et al.*, 2012). Being largely a progression free condition, the mortality patterns remains merely inflated with a death rate of 1.25 for males and 1.11 for females compared to general population.

For SMM, sex adjusted incidence is reported at 0.9 cases per 100,000 persons in United States compared to that of 0.4 in Sweden (Anonymous, 2013; Ravindran *et al.*, 2016). An estimate of newly diagnosed SMM cases is thus approximated around 4,100 annually (Ravindran *et al.*, 2016). Although the cumulative probability of progression to SMM was 73% at 15 years (Kyle *et al.*, 2007), progression rate of SMM to MM (80%–90% at 2 years) compared to that of MGUS is substantially higher affecting overall survival of SMM patients (Blum *et al.*, 2018); (Rajkumar *et al.*, 2015).

Disease burden of MM is more robustly explored over the years due to its severity in the malignant stage. An age standardized incidence rate of 2.1 per 100,000 was reported for MM in United States (Cowan *et al.*, 2018). Although predominantly more incident in western developed countries, MM commands significant cancer burden worldwide as shown below (**Figure 2.1**).

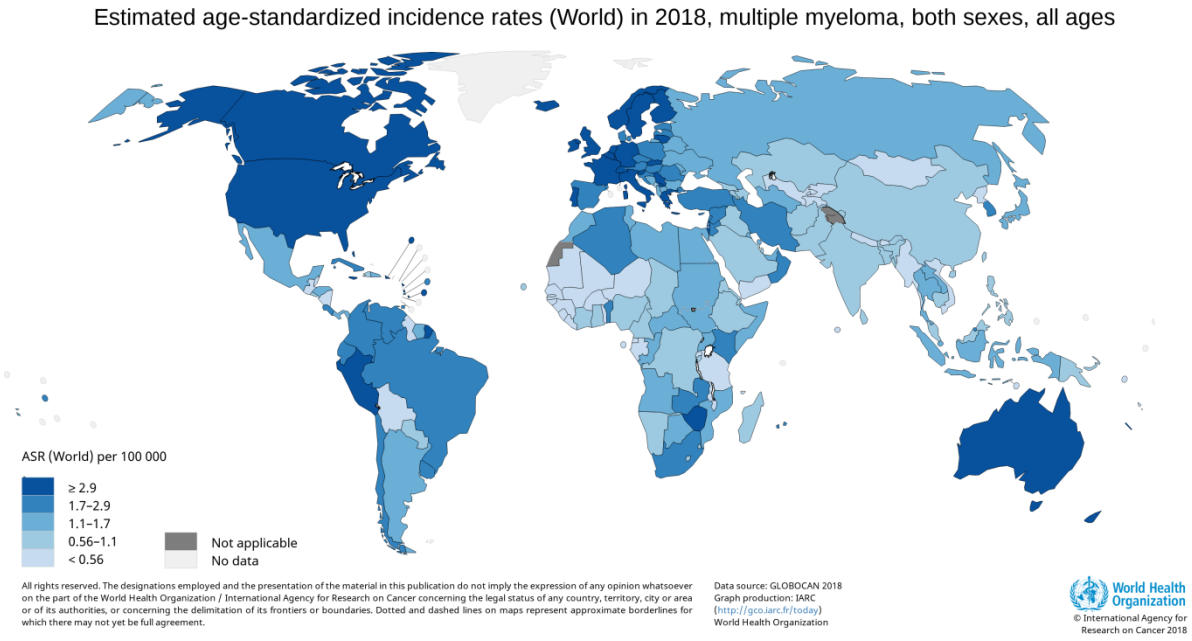


Figure 2. 1| Age standardized incidence rate per 100,000 people from MM
Country specific incidence rate of MM as calculated with the GLOBOCAN data, 2018. Well developed countries show higher incidence in general compared to that in the lesser developed ones. Produced from http://globocan.iarc.fr

In 2016 MM was responsible for approximately 98,000 deaths globally with an age-standardized death rate of 1.5 per 100,000 persons (Cowan *et al.*, 2018). The authors also report a 94% increase in MM related deaths worldwide since 1990. It is likely that population growth and ageing global population contributes to such increments in statistics, nevertheless the rate of monotonic increase in mortality is alarming. In Europe, age standardized mortality rate in MM ranges from 0.7 per 100,000 people to 2.7. Four of the five Nordic countries (Norway, Denmark, Sweden and Finland) are estimated to have mortality rate of more than 2.0 per 100,000 persons comprising some of the highest rates observed in Europe. Estimated 1-year prevalence among European nations is highest for France (10.7 per 100,000 individuals) and lowest for Albania (0.2) with Norway (10.3), Sweden (7.4) and Finland (6.9) belonging in the top 10 countries with highest prevalence rates (**Figure 2.2**).

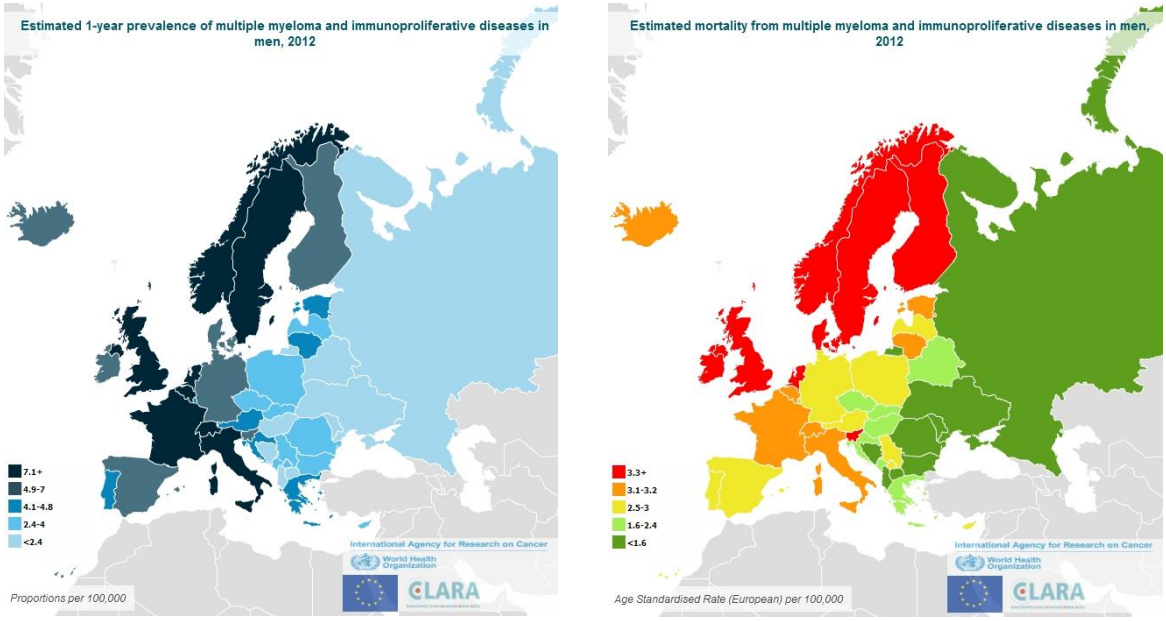


Figure 2. 2| Age standardized prevalence and mortality rate per 100,000 people from MM in Europe
Country specific prevalence and mortality rates of MM as calculated with the GLOBOCAN data, 2012. Nordic countries show a higher prevalence and mortality in general. Produced from http://globocan.iarc.fr

2.2 Subsequent primary cancers in MM survivors

Management of MM encompasses massively dynamic investigations of permuted regimens, changes in treatment modalities and improvement upon suggested therapies over time. Initiation of treatment with alkylating agents, autologous stem cell transplantation, and immunotherapy has brought incremental but dramatic changes in MM survival landscape over the last few decades (**Figure 2.3**) (Fonseca *et al.*, 2016).

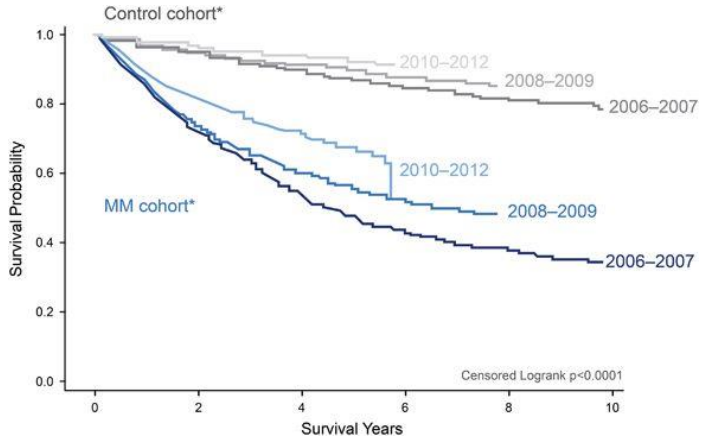


Figure 2. 3| Survival estimates of MM patients by year of diagnosis. Adapted from Fonseca *et al*., 2016
Survival estimates were presented for MM patients diagnosed and treated during 2006–2012 matched against control cohort during the same time.

This improvement in survival presented with a new problem, diagnosis of subsequent primary and therapy related cancers. Initially reports were published on frequently diagnosed second primary acute myeloid leukemia and myelodysplasias in MM patient cohorts which was later attributed to conventional chemotherapy before autologous stem-cell transplantation (Bergsagel *et al.*, 1979). From 1960s Melphalan in combination with prednisone was the standard treatment for all MM patients. With melphalan, cyclophosphamide, and carmustine, Bergsagel *et al.* conducted the first prospective clinical study that evaluated the value of a combination of 3 alkylating agents in MM treatment where they described an excess in expected incidence of several types of hematological

malignancies (Bergsagel *et al.*, 1979). However, with incorporation of high-dose melphalan followed by autologous stem cell transplantation, immunomodulatory drugs and proteasome inhibitors paved the road towards a sustained and prominent improved survival at a cost of higher numbers of subsequent cancers (Palumbo and Anderson, 2011; Palumbo *et al.*, 2014; Singhal *et al.*, 1999). Furthermore, estimations even suggested an increased number of therapy induced hematological malignancies after Ig-M MGUS (Mailankody *et al.*, 2011). Until very recently leukemias originated from myeloid cell lineage were believed to be primarily incident as second cancers and an expected increase of such cases were frequently speculated (Landgren and Mailankody, 2014; Landgren *et al.*, 2011; Thomas *et al.*, 2012). In a combined investigation of Swedish and German cancer cohorts, Chen et al reported several solid tumors occurring in MM patients aside from the hematological malignancies (Chen *et al.*, 2016). In fact prostate, colorectal and breast cancers were more frequently diagnosed than leukemia as second primaries in both the countries. This indicated to a relatively new era of MM patient management where consequences of treatment on a prolonged survival period needed to be considered.

2.3 Risk stratification and epidemiological models

Plasma cell dyscrasias including MM are genetically and biologically a heterogeneous group of disorders that present with variable disease burden depending upon their inherent characteristics which translates to variable response to treatment and outcome. Accounting for high-risk disease features, disease burden and pathogenic factors present in host, MM is categorized indicating a risk stratification of prognostic nature. The first predictive clinical staging system for MM was introduced in 1975 by assessment of A) extent of bone lesions, B) hemoglobin level, C) serum calcium level, and D) M-component levels in serum and

urine (Durie and Salmon, 1975). However, the then Durie-Salmon staging system was later criticized for being dogmatically disease-burden driven (Hari *et al.*, 2009).

IMWG introduced international staging system (ISS) for MM in 2005 incorporating a combination of serum beta2-microglobulin, serum albumin, platelet count, serum creatinine, and age (Greipp *et al.*, 2005). In a 2014 updated release of ISS, IMWG distinguished between prognostic and predictive markers to have separately refined the risk stratification criteria (Chng *et al.*, 2013). This updated ISS also takes genetic aberrations and gene expression profiles (GEP) into account. Chromosomal translocations, gains, deletions and amplifications had shown significant power in prognostic likelihood and thus inclusion of tumor cytogenetics rendered in greater accuracy in prediction. At the same time gene expression driven predictive modeling also added power in stratification. Development of GEP70 (Gene Expression Profile 70) saw the first large-scale gene expression driven classification based on which later models such as GEP5 were established (Heuck *et al.*, 2014; Shaughnessy *et al.*, 2007). Finally the revised ISS (R-ISS) was developed by pooling data from newly diagnosed MM patients enrolled on 11 international trials (**Table 2.1**) (Palumbo *et al.*, 2015). It combined the ISS with high-risk chromosomal aberrations [deletion del(17p), translocation t(4; 14) (p16; q32) or translocation t(14; 16) (q32; q23)] and serum lactate dehydrogenase to stratify patients in three risk categories. According to R-ISS, the 5-year overall survival probability of MM patients with stage I was 82%, 62% for sage II and 40% for stage III, whereas the 5-year progression-free survival for the same groups were 55%, 36% and 24%, respectively.

Table 2. 1| Standard Risk Factors for MM and the R-ISS. Adapted from Palumbo *et al.*, 2015

Prognostic Factor	Criteria
ISS stage	
I	Serum β_2-microglobulin < 3.5 mg/L, serum albumin ≥ 3.5 g/dL
II	Not ISS stage I or III
III	Serum β_2-microglobulin ≥ 5.5 mg/L
CA by iFISH	
High risk	Presence of deletion del(17p) and/or translocation t(4;14) and/or translocation t(14;16)
Standard risk	No high-risk CA
LDH	
Normal	Serum LDH < the upper limit of normal
High	Serum LDH > the upper limit of normal
A new model for risk stratification for MM	
R-ISS stage	
I	ISS stage I and standard-risk CA by iFISH and normal LDH
II	Not R-ISS stage I or III
III	ISS stage III and either high-risk CA by iFISH or high LDH

Abbreviations: CA, chromosomal abnormalities; iFISH, interphase fluorescent in situ hybridization; ISS, International Staging System; LDH, lactate dehydrogenase; MM, multiple myeloma; R-ISS, revised International Staging System.

2.4 Modifiable risk factors

Although what causes MGUS or MM is yet not definitively known, several studies have evaluated potential environmental, behavioral (externally modifiable) risk predisposing factors. One of the frequently speculated yet tantalizing risk factor is exposure to ionizing radiation. In a 1982 study, a threefold increased incidence of MM was reported with an age adjusted incidence of 0.048 cases per 100,000 person-years subject to intensity of radiation exposure to bone marrow of ≥ 0.5 Gy about 20 years after the atom bomb explosion in the cities of Hiroshima and Nagasaki (Ichimaru *et al.*, 1982). On the contrary a more recent analysis consisting of follow up data from 1950 until 1987 with 2,778,000 person-years, found that individuals with a total radiation dose exposure of < 4 Gy did not exhibit any evidence of excess risk of MM, compared to the unexposed individuals (Preston *et al.*, 1994). The authors even went on to speculate that exposure to ionizing radiation due to direct effect of atom bombs bore little to no evidence for drawing any robust conclusion on MM risk modulation. Additionally, results from investigation on effect of therapy related radiation exposure due to routine diagnostic procedure on incidence of MM has been inconclusive (Boice *et al.*, 1991; Hatcher *et al.*, 2001). Exposure to UV radiation has been shown to attribute to moderate excess risk (Boffetta *et al.*, 2008). The exact mechanism behind this is also speculated with expression regulation of established MM therapeutic target genes via irradiation but no causal inference is drawn (Shen *et al.*, 2017).

Occupational exposure to possible carcinogenic elements and associated MM risk has been studied in several populations. Farming has been systematically associated to an excess risk of MM (Burmeister, 1981b; Khuder and Mutgi, 1998; Perrotta *et al.*, 2008). Speculations are presented on possible detrimental effects of pesticide exposure, DDT exposure, and exposure to solvents such as phenoxyacetics, chlorophenols as well as exposure to farm

animals, infectious agents and other factors but all too with elusive precedence. Firefighters were also found to have an elevated susceptibility to MM (LeMasters *et al.*, 2006). The underlying mechanisms here could include recognized exposure to detrimental agents such as heavy metals (antimony, cadmium, lead), chemical constituents (formaldehyde, xylene, trichlorophenol, toluene, polycyclic aromatic hydrocarbons, methylene chloride, benzene, acrolein) along with other minerals (non-crystalline silica, crystalline and asbestos) (Brandt-Rauf *et al.*, 1988). Rather peculiarly, hairdressers were also found to be at higher risk of developing MM compared to general population with an estimated excess lifetime risk of almost 40% (Takkouche *et al.*, 2009). It's noteworthy that hairdressers admittedly have a higher risk of cancer compared to the general population primarily due to their frequent exposure to hair dye which carries a significant carcinogenic load (Altekruse *et al.*, 1999b). They are also exposed to many different chemicals including and not restricted to nitrosamines contained in hair-care products, methacrylates, formaldehyde, shampoos, hair conditioners and bleaches and propellants, aerosols from hairsprays and other volatile solvents which may contribute to the risk burden thus observed (International Agency for Research on, 1993). Additionally occupational exposure to methylene chloride, benzene, engine exhaust was also postulated to have minimum to moderate association with excess MM incidence (Liu *et al.*, 2013; Sonoda *et al.*, 2001; Vlaanderen *et al.*, 2011).

Not surprisingly, there are a multitude of lifestyle parameters and behavioral patterns that link to excess MM risk. As observed for most of the cancers, obesity and over-weight correlates with a higher proportion of both MGUS and MM (Blair *et al.*, 2005; Calle *et al.*, 2003; Samanic *et al.*, 2004). Markedly, in postmenopausal women, an elevated BMI of ≥ 36 associated with an excess in relative risk of 2.0 for MM against general population (Blair *et al.*, 2005). Effect of dietary routines has also been examined by few studies. While

investigating relationship between specific foods or food groups and MM risk, frequency of dairy (excluding yogurt) , meat and grain intake were not found to be associated (Chatenoud *et al.*, 1998; Tavani *et al.*, 2000); however, for butter consumption, positive association was found (Vlajinac *et al.*, 2003). Vegetable consumption also expectedly associated with a diminished risk (Vlajinac *et al.*, 2003). However, no significant relation was found between animal fat intake and excess risk, consumption of fish was inversely linked to MM risk (Fritschi *et al.*, 2004).

A number of studies have examined association between tobacco consumption and MM (Adami *et al.*, 1998; Mills *et al.*, 1990). There was not enough evidence to establish tobacco consumption as a major risk factor since relative risks of the exposed group (smokers) did not differ significantly to non-smokers (assuming main form of tobacco consumption is smoking) (Mills *et al.*, 1990). Even in large case-control studies, the odds ratio depicting risk effect size followed a similar trend (Brownson, 1991; Fritschi and Siemiatycki, 1996; Linet *et al.*, 1987; Williams and Horm, 1977). Contrarily, although believed to be a strong risk predisposing factor for several malignancies, epidemiological evidence for alcohol consumption in light of MM risk modulation is limited at best. Moreover, the handful numbers of studies that exist, have not found any significant excess risk of MM in relation to alcohol consumption (Brown *et al.*, 1992; Nieters *et al.*, 2005). The biological reason behind this unwavering risk has been argued with immunomodulatory effects of alcohol by inhibition of the mammalian target of rapamycin signaling via m-TOR pathway through ethanol (Hagner *et al.*, 2009). However whether these factors are causal or surrogate agencies for other socio-economic pattern related lifestyle traits is yet to be determined.

Chapter 3: Genetic epidemiology

For most of the major forms of cancers, association studies including Genome-Wide Association Studies (GWAS), Genome-Wide Interaction Studies (GWIS) and other similar study designs have demonstrated that genetic risk of cancer can be explained by the impact of co-inherited common genetic lesions. Single nucleotide polymorphisms (SNPs) are one of the major sources of genetic lesions and are presumed to be accountable, at least in part, for the singular alterations in genetic susceptibility to complex phenotypes such as cancers. It has been recently shown to hold true for Waldenström macroglobulinemia (McMaster *et al.*, 2018) and the same is also probable to be true for the disorders in MM disease family as well and several published GWAS indicate this (Broderick *et al.*, 2011; Chubb *et al.*, 2013; Mitchell *et al.*, 2016; Thomsen *et al.*, 2017). Numerous SNPs and therefore the annotated genes harboring such lesions belonging to different biological pathways have been shown to predispose to MM, although the detection strategies vary greatly to have explained MGUS and MM heritability. At the same time, high penetrance mutations, which were shown to explain a small proportion in many common cancers, has been elusive in MGUS, MM; nevertheless some somatic variations have been identified (Leich *et al.*, 2013a; Mikulasova *et al.*, 2017; Miller *et al.*, 2017).

3.1 Inherited susceptibility

The notion of a possible inherited familial predisposition to MM was initially proposed in the 1920s. In 1925, Meyerding reported a case where a MM patient had an aunt with a bone disease with a fractured leg possibly indicative of myelomatic bone lesion (Meyerding, 1925). Later Geshickter and Copeland published a review of MM where they briefly

discussed a case where both the brothers in a family died of MM (Geschickter and Copeland, 1928). In 1954 first detailed case study was revealed where two sisters with MM were discussed in depth (Mandema and Wildervanck, 1954). More recently, Lynch reported 39 families with several family members affected by MM, MGUS, Waldenström macroglobulinemia or amyloidosis as well as another 8 African American families with multiple occurrences of MM or MGUS in 2009 (Jain *et al.*, 2008; Lynch *et al.*, 2005). To date, more than 100 families with multiple affected members either with MM, MM like or other plasma cell disorders have been reported which provide strong evidence for the existence of inherited susceptibility.

With the introduction of linkage studies in early 1980s and 1990s, a number of cancer predisposing genes have been identified in high-risk families. Rare variants in breast cancer related genes (*BRCA1*/*2*), colorectal cancer associated gene *APC* and mismatch repair genes (*MLH1*, *MSH2*), Melanoma with *CDKN2A* were shown to produce highly penetrant phenotypes but these mutations are rare and account for a very marginal proportion of the 'familial' element of a cancer (Bodmer *et al.*, 1987; Cannon-Albright *et al.*, 1992; Hall *et al.*, 1990; Lindblom *et al.*, 1993; Peltomaki *et al.*, 1993; Wooster *et al.*, 1994). The linkage and pedigree mapped familial studies in MM have not been largely as successful in discovering truly high penetrant carrier mutations. The gradual revelation of familial inclination and association of risk also argues for existence of sizable fraction of the MM susceptibility due to heritable factors. As discussed earlier, several GWAS have successfully identified a handful number of risk SNPs predisposing to MM and effect sizes exerted by these risk SNPs were meta-analytically assessed for acumen of true risk predisposition (Mitchell *et al.*, 2016).

Although these studies bring about possibility of therapeutic target discovery and development, it is yet to be known, how much of the inherited risk is explained by the already identified risk loci and what percentage of the heritable risk remain to be uncovered. Heritability estimate under the assumption of causal sentinel SNPs (along with tagged SNPs) being detected can answer this question. As the exact distribution of minor allele frequency (MAF) for MM causal SNPs is unknown any heritability estimate regarding the risk SNPs would be prone to bias. Yet taking MAF threshold of 0.5, adjusted heritability was assessed at 17.2% whereas the same with a MAF threshold of 0.1 was 27.8% (**Table 3.1**) (Mitchell *et al.*, 2015).

Table 3. 1| Heritability of multiple myeloma adjusted for incomplete LD between causal SNPs and those used to compute the genetic relationship matrix. Adapted from Mitchell *et al.*, 2015.

	Heritability	
MAF threshold	**GCTA**	**PCGC**
No adjustment	0.152 ± 0.028	0.168 ± 0.041
0.5	0.173 ± 0.032	0.192 ± 0.049
0.4	0.180 ± 0.033	0.200 ± 0.049
0.3	0.192 ± 0.035	0.212 ± 0.058
0.2	0.212 ± 0.039	0.235 ± 0.070
0.1	0.278 ± 0.051	0.307 ± 0.079

3.2 Family history of cancer and excess risk

An increase in relative risks of MM was already reported in patients with first degree relatives with cancer diagnosis almost three decades ago (Bourguet *et al.*, 1985; Brown *et al.*, 2000; Eriksson and Hallberg, 1992). In an investigation from 1989 Grufferman et al. reported that MM patients were 4.4 times more likely to have at least one first-degree relative with a prior diagnosis of degenerative or demyelinating central nervous system disease (Grufferman *et al.*, 1989). Using the Swedish cancer registry data in 2003 Hemminki et al. reported an excess risk of MM in children of parents with MM with a

standardized incidence ratio of 3.3 (Hemminki *et al.*, 2003). There have been investigations of MM risk in distantly related family members and also in spouses, although no conclusive inferences could be drawn due to the small number of reported cases in spouses (Kyle and Greipp, 1983; Kyle *et al.*, 1971; Lynch *et al.*, 2001).

As the evidence of familial clustering of MM became more pronounced, population based observational studies started investigating excess MM risk in several pockets of population associated with a history of other cancer in family. Initial studies suggested an elevated risk of MM both in males and females subject to history of MM in family (relative risks ranging up to 3.23 for females and 2.33 for males and females combined) (Ogmundsdottir *et al.*, 2005). Although, there was no excess risk of MGUS, the authors additionally claimed that irrespective of gender, there was elevated risk of hematological malignancy in individuals related to a family member diagnosed with MM. In 2006 Landgren *et al*. reported a statistically insignificant increased risk of MM among people with a first-degree relative with MGUS and speculated that the statistical insignificance was possibly due to low reporting of MGUS cases and the actual risk would probably have been far more alarming (Landgren *et al.*, 2006). Although familial clustering in MM and MGUS were previously described, these studies elucidated inherited risk predisposition to these diseases subject to existence of cancer history in family in general. As existence of strong genetic influence would become clear in the later years, these findings in principle laid the fundamental framework for investigating a true polygenic inherited susceptibility to MM.

Chapter 4: Strategies to address inherited risk

Traditionally population demography and molecular and genetic epidemiology have been the main tools to enquire inherited risk in all realms of phenotypes. The scientific apparatuses addressing the questions have themselves gone through extensive evolution. Today this dynamic metamorphosis of statistical methods and computational algorithm is happening more rapidly than ever which obviously was not always the case.

4.1 Epidemiological methods in population risk prediction

Essentially the aim of studying an association between two events is to quantify the measure of effect (of one event subject to the other). This measure of effect is usually calculated with relative risk or odds ratio. In observational framework, the relative risk is assessed by ratio of incidence proportions and the numeric estimate is often accompanied by a measure of precision, confidence intervals (confidence bands in Bayesian set up) (Tripepi *et al.*, 2007). On the other hand, the odds ratio is as the name suggests, ratio of odds of two events occurring. *Odds* are a way of presenting scaled / weighted probability. *Odds* are mostly synonymous to case-control studies where *odds* of exposure to the cases and controls are calculated as probabilistic point estimates by dividing the numbers of exposed by unexposed in each group. Similar to relative risk, the ratio of odds is also accompanied by confidence interval which in both instances is largely influenced by number of individuals contributing (sample points) and inherent heterogeneity of the data (due to confounders, non-linear effects and other parameters; not to be confused with parametric set up).

A similar notion is also employed in estimation of survival probabilities. As the rate of attrition is of prime importance in survival study, odds representing mortality is called

hazard. *Hazard* is defined in a time dependent manner as a ratio of events occurred until a specific time point and the hazard ratio can be calculated in a way similar to that of odds ratio (Clark *et al.*, 2003).

Cumulative incidence is frequently used to assess age, follow-up stratified or life-time risk of an event often with the help of bracketed survival probabilities. The added benefit of observing cumulative incidence is in consideration of competing event. Several adjustments are developed to attribute the inflation in risk due to this phenomenon (Coviello and Boggess, 2004).

Another intuitive method to assess effect of exposure in a population is demonstrated with population attributable fraction. The development of this method dates back to 1953 (Levin, 1953). It is defined as the fraction of individuals representing an outcome of interest which presumably manifests due to a certain risk factor amidst a population. A synonymous estimate is called population attributable risk and is defined as the difference in the rate or risk of disease for the population compared to the unexposed (calculated on linear scale compared to that in multiplicative scale for the former). However confusion in epidemiological studies in application of such methods due to lack in understanding is quite pronounced in literature (Zapata-Diomedi *et al.*, 2018).

It is also to be noted that all of the discussed estimation strategies can be assessed in either parametric or non-parametric fashion subject to distributional information of the underlying data pattern and conformity to inherent assumptions levied on the particular strategies to be employed.

4.2 Linkage, GWAS and GWIS

From a strictly molecular and genetic perspective, a long studied goal in explaining variation of a quantitative trait or the risk of a disease has been motivated by the identification of genes that contribute to such phenomenon. To that end the study design of choice had been linkage studies for over two decades, primarily due its viability with comparatively sparse array of genetic markers, obtaining which was technologically feasible. Linkage studies were established to investigate surplus co-segregation among sentinel alleles underlying a certain trait with the tagged alleles at a putative risk locus in family data. For years the linkage analysis had been the major instrument in interrogating the genetic mapping of both complex and Mendelian traits with familial accretion. The basic principle of linkage analysis dictates that the likelihood of meiotic recombination between two points in the genome is proportional to the distance between the physical maps of the points. Hence variations in polymorphic sites (deviant alleles) are more likely to reside in close proximity of a disease-causing locus inherited in families through generations. Therefore by studying the co-segregation of variation in polymorphic loci and inherited phenotype, certain genomic windows can be identified that are inherited with said phenotype. Formal linkage analysis has identified several risk loci related to MGUS and MM (Kristinsson *et al.*, 2009; Lynch *et al.*, 2008b). Several separate variations of this particular design have also been proposed over the years to investigate genetic co-segregation. The limitation of linkage analysis is in its detection power. Admittedly due to its nature of looking into sparse ‘candidate’ regions in context of a phenotype, linkage studies demonstrate high statistical power of detection when it comes to high penetrant alleles. Contextually, high penetrant alleles are those alleles which make largest contribution (assumed the causal loci is included) to the excess risk of expression of a phenotype or to

the regulation of a quantitative trait. Nevertheless, due to selection such high penetrant alleles tend to be rarer in nature. Contrarily for capturing signals from common alleles which tend to have a small effect size for most diseases or traits of polygenic nature, GWAS designs are adequately powered. However, until the start of last decade, performing such association studies in a genome-wide scale were not feasible due to technological caveats in obtaining dense polymorphism arrays to have acceptable detection capability.

Humoring the idea of polygenic risk, in 1974 Anderson in his investigation of familial risk in breast cancer speculated that the excess risk of cancer observed in first-degree relatives of cancer patients "*… are not indicative of a strong genetic effect. They are more suggestive of a polygenic mechanism, that is, the involvement of many genes with small effects acting in concert with environmental or nongenetic factors with larger and more important effects*" (Anderson, 1974). This reasoning was later proven to be incorrect with observational studies reporting similar inflation of relative risk in cancer-susceptible families and with the largescale linkage studies identifying cancer susceptible genes in major cancers (Cannon-Albright *et al.*, 1992; Hall *et al.*, 1990; Hemminki *et al.*, 2003; Peltomaki *et al.*, 1993). Although, the high penetrant rare variants only explained a very moderate amount of the estimated heritability which indicated the presence of aggregated risk exerted by common SNPs with comparatively lower effect sizes (**Figure 4.1**) (Sud *et al.*, 2017a).

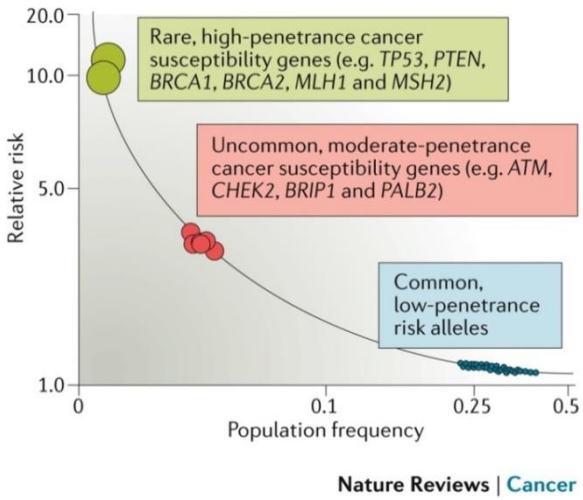


Figure 4. 1| Genetic architecture of cancer risk. Adapted from Sud *et al*., 2017a
This graph depicts the low relative risks associated with common, low-penetrance genetic variants (such as single nucleotide polymorphisms identified in genome-wide association studies); moderate relative risks associated with uncommon, moderate-penetrance genetic variants (such as ataxia telangiectasia mutated (ATM) and checkpoint kinase 2 (CHEK2)); and higher relative risks associated with rare, high-penetrance genetic variants (such as pathogenic mutations in BRCA1 and BRCA2 associated with hereditary breast and ovarian cancer).

For example, only about 25% of the two-fold excess risk observed in the first degree relatives of breast cancer patients is attributed to the *BRCA1/2* deleterious mutations (Anonymous, 2000; Peto *et al.*, 1999). Similarly, almost 60% of the heritable risk for colorectal cancer still remains unaccounted for (Chubb *et al.*, 2016; Lubbe *et al.*, 2009). Ironically, although the justification in Anderson's account was unbecoming, today polygenic inheritance is acknowledged to have greatly explained the architecture of inherited genetic predisposition to cancer. In search of such risk loci, the focus in genetic epidemiology has been shifted towards GWAS since the last decade because of the affordability and availability of compact collection of arrays containing large number of markers which can be genotyped for a much greater number of people. This school of analysis examines common variants associated with disease or quantitative trait. GWAS has been largely successful in identifying large number of risk SNPs for simple to complex phenotypes including almost every cancer. However, this methodological improvement in detection algorithm is not impervious to apparent caveats. The associative relation between

disease and risk SNPs revealed by GWAS are by no means causal in nature; at least the study design of GWAS cannot make any such assertion. Secondly, most of association studies report detection of risk SNPs dichotomously differentiated by the magnitude of effect size. Elaborating on cancer susceptible genes discovered by traditional association studies (not to be confused with GWAS) Sud *et al*. demonstrates these two classes of susceptibility loci (Sud *et al.*, 2017a). One, which are the rarer and moderately penetrant variants (MAF < 2% and effect size > 2.0) identified by candidate gene study (ex. *ATM*, *CHEK2*, *PALB2* mutations for breast cancer); two, the low penetrant risk alleles that were mostly identified by GWAS. The authors then speculate that "*it is likely that the spectrum of penetrance and frequency of risk alleles for many cancers occurs on a continuum*"; meaning there is possibly a subgroup of risk alleles which are predisposed to be readily detected in certain study designs. If we are to extrapolate, the problem of missing heritability in cancers is due to the rigidity of the study design.

In attempt to explain the problem of missing heritability, several justifications were proposed. To begin with, GWAS identified SNPs are probably surrogate markers found (lacking) in linkage disequilibrium with the real causal loci. Hence such markers even when considered together will probably lack in power to completely capture the totality of the causal effects, particularly since the causal variants if present are intermittent in general populace due to selection. Furthermore, GWAS are power-compromised in distinguishing loci with moderate effects indicating that bulks of the true risk predisposing loci are left unaccounted for. This indicates that despite some of the single SNPs having moderate to poor effect on a phenotype, their (collective) impact may be of greater magnitude and measurable from the perceived genetic data. Additionally, alarming is the fact that the single-locus testing strategy is probably underpowered to observe signals at a statistically

significant level from markers which interact with other genetic (or environmental) elements as impact of such loci remains elusive except the simultaneous existence of the contributing factors. Hence, investigating gene-gene (and gene-environment) interactions is another design to observe the missing heritability of complex phenotypes (Phillips, 2008). Disease advancement is believed to be a complex process reflecting interactions within a multifaceted biological construct structured into an assortment of interactive networks via regulation of pathways. According to modern complexity theory, biological interaction can be considered to be a sensible quantification of complexity of a biological system since the complexity is accredited to the interactions among the components of a system. Therefore, the underlying hypothesis is that a disease may be caused by joint effects of multiple loci predisposing to the disease in interaction (Cordell, 2009). Additionally from an algorithmic point of view, incentive for developing design to interrogate statistical interaction in inherited genetic predisposition is to provide improved opportunity for identifying cooperatively influencing effects of loci in interaction compared to investigating merely the marginal associations arising from each individual loci (Murcray *et al.*, 2009).

4.3 Functional validation of risk loci

Several classes of functionally stratified of genetic variations are associated as the foundation of risk predisposition via markers recognized by GWAS. Depending on the physical location of the SNPs identified, they can directly influence the amino acid sequence of the expressed protein, RNA processing or DNA methylation (Michailidou *et al.*, 2013; Schulz *et al.*, 2017; Stacey *et al.*, 2011; Wang *et al.*, 2014). In addition it is perfectly plausible for coding variants to harbor subtle influences which essentially do not involve direct regulation of protein functions, instead are responsible for tagging non-coding SNPs.

In 2010, Manolio has demonstrated that most of the GWAS detected risk loci lay on the non-coding regions of the genome and are therefore likely to be involved in gene regulation (Manolio, 2010). With expression quantitative trait loci (eQTL) analysis, effect of such variants on gene expression in cell or tissue in context can be measured. eQTL analysis on malignant plasma cells extracted from German MM patients helped identify several *cis*-regulatory signals including that of *MYC*-interacting gene *CDCA7L* by rs4487645 and several HLA genes (Weinhold *et al.*, 2015). Another study from UK later reported a moderate association with *WAX* and *PREX1* with strong signals observed in methylation quantitative trait loci (meQTL) in CD138-positive MM plasma cells (Mitchell *et al.*, 2016).

Although in the above algorithm the landscape of risk loci association is well investigated, it is not optimal to identify true causal signals scaling for the noise very plausibly present in genome-wide analyses due to the number of tests performed. Mendelian randomization (MR) leverages genetic variants as instrumental variables as they are most likely to be independent of confounding factors (Hernán and Robins, 2006; Lawlor *et al.*, 2008; Smith and Ebrahim, 2003). One of the difficulties of a MR study is to identify in addition to proper exposure and effect, instrumental variable(s) unbiased by possible confounding. Interrogating genetic data as instrumental variable, Zhu *et al*. developed MR algorithm applicable to genome-wide scale that integrates phenotypes as effect and gene expression data as exposure (Zhu *et al.*, 2016; Zhu *et al.*, 2018). Although this method of MR based inference has effectively identified causal putative loci for complex traits, large scale application on malignant phenotypes is overdue (Colodro-Conde *et al.*, 2018; Hemani *et al.*, 2018; Luijk *et al.*, 2018; Qi *et al.*, 2018).

Another school of well-traversed *in silico* algorithms for functional annotation of risk SNPs is enrichment analysis. Gene set prioritization based pathway, tissue and cell enrichment

analyses are used to consolidate the effects of multiple GWAS detected variants. Several annotation tools have been developed with this specific need in mind and have helped discovery of causal pathways in complex traits (Lamparter *et al.*, 2016; Pers *et al.*, 2015). In context of MGUS and MM, several studies have shown activation or differential regulation in NF-κB, Ras/Raf/MAPK/Erk, PI3K/Akt/mTOR, Jak2/Stat3, VEGF signaling pathways (Korde *et al.*, 2011b; Ramakrishnan and D'Souza, 2016; Zingone and Kuehl, 2011). However as more of the heritability gets explained with risk loci, the landscape of pathways and underlying genetic and mechanistic links becomes clearer.

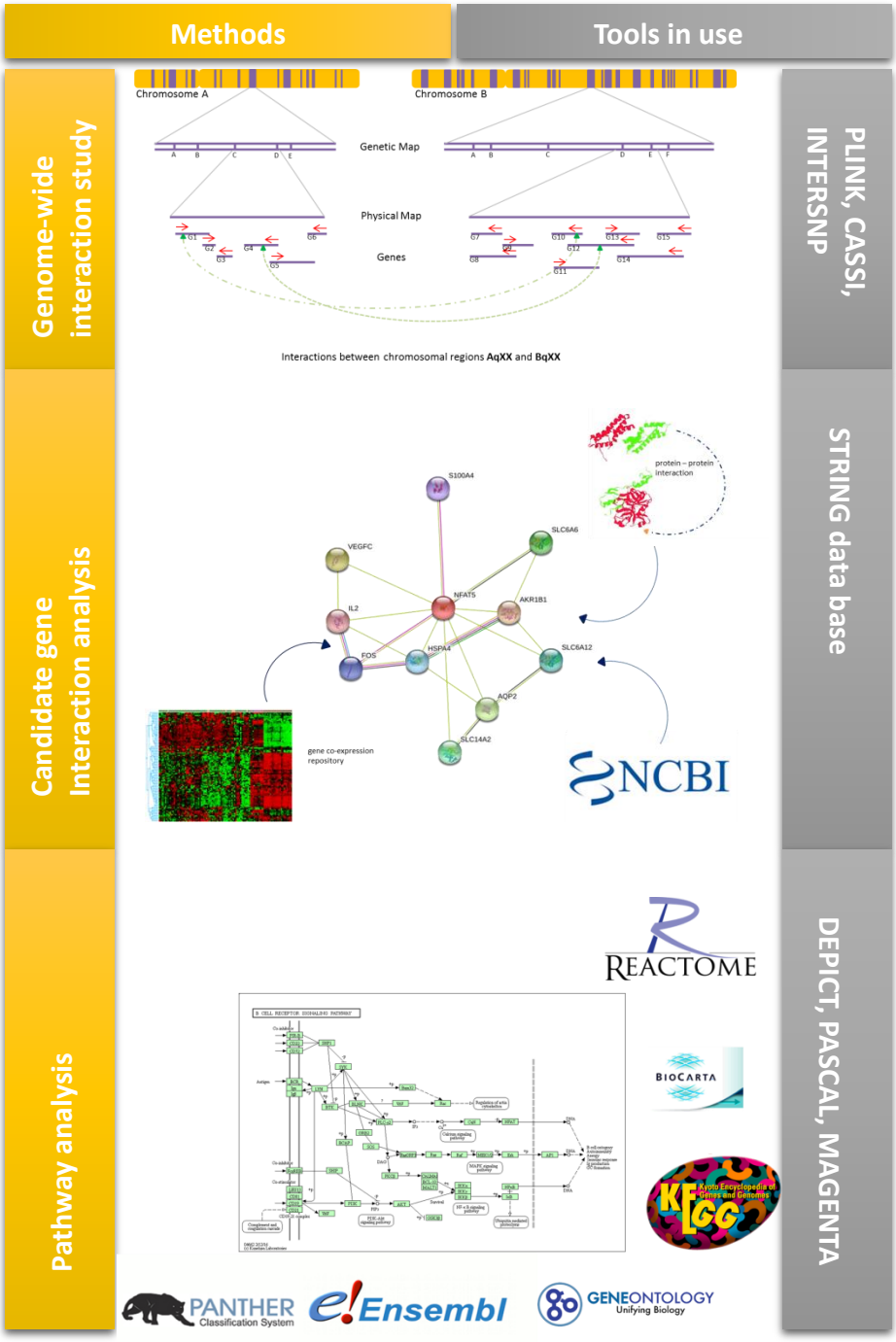
Aim of the study

As inherited genetic architecture of risk predisposition to MGUS and MM mostly involves common variants with moderate to underwhelming effect sizes, the work detailed in this thesis aims at obtaining further insight into it by interrogating genetic data with available technologies leveraging several genome-wide analyses. Additionally, it also investigates association of cancer burden exerted by family history of cancer on MM patients in developing subsequent primary cancers.

Specifically I aim at:

1) Identification of low-penetrant germline variants in interaction predisposing to MGUS risk.
2) Providing functional annotation to identified risk SNPs with genetic network construction and pathway enrichment.
3) Characterization of inherited genetic susceptibility to MM through genome-wide genetic interaction.
4) Functional annotation of discovered risk loci by interrogating eQTL, MR assessed with GWA summary statistics, gene-set prioritization, network enrichment as well as tissue and cell enrichment.
5) Investigating the role of familial susceptibility in the form of family history of cancer in MM patients in developing second primary cancers and assess causes of death.

Schematized design of the study



Materials and Methods

Chapter 5: Datasets

The molecular-genetic part of the study focuses on data obtained from genotyping. Genotyping is the process of obtaining genotype of an individual with a biological assay, frequencies of which thus obtained are compared between individuals with a certain phenotype and those without (given homogeneous ancestry) in a case-control set-up predominantly with association tests.

On the other hand, the population observational side will focus on the Swedish cancer registries with a nation-wide follow-up of complete cancer diagnoses since 1958. Detailed description on each of the relevant sources of data follows.

5.1 Genotype data

Genotype data procurement strategy and data description are directly taken from published studies without much alteration as is hereby referred to: MGUS sample data from (Chattopadhyay *et al.*, 2018c; Thomsen *et al.*, 2017); MM sample data from (Mitchell *et al.*, 2016) and eQTL sample data from (Weinhold *et al.*, 2015).

5.1.1 MGUS samples

The University Clinic of Heidelberg and the University Clinic Ulm discovered 243 MGUS cases among which 114 (47%) were males with a mean age at diagnosis of 62 years, standard deviation (SD) ± 11 years. The Ig isotype distribution was 72% IgG, 12% IgA, and 16% other Ig isotypes (Thomsen *et al.*, 2017; Weinhold *et al.*, 2014b). These MGUS cases were identified during diagnostic work-out of a different disease. Out of the

243 cases, two developed MM within three years after sampling and 46 individuals were seen only at the time of sampling. IgM MGUS cases were excluded from the Heidelberg cohort. For replication, 236/82 MGUS patients were identified for case-control/case-only replication in Essen within the Heinz-Nixdorf Recall (HNR) study (Schmermund *et al.*, 2002; Thomsen *et al.*, 2017). About 61% of the replication set were males with the mean age at diagnosis of 64 years, SD ± 9 years. Detection of MGUS was based on internationally accepted criteria (Anonymous, 2003): monoclonal protein concentration less than 30 g/l, less than 10% monoclonal plasma cells in bone marrow, normal plasma calcium and kidney function and no bone destruction or anemia. The reference population for the Heidelberg set consisted of 1285 German individuals from the HNR study with almost 50% males (Schmermund *et al.*, 2002). The reference population for the Essen set was also recruited within the HNR study, adding up to 2484 individuals (51% males) not overlapping with the reference population for the Heidelberg set.

Illumina HumanOmniExpress-12v1.1 chip arrays were used for genotyping the Heidelberg MGUS set and the corresponding control set was genotyped using the Illumina HumanOmniExpress-12v.1.0 chip array (Schmermund *et al.*, 2002). The Essen set was genotyped using six different chips: 365 (15 cases, 350 controls) were genotyped on Illumina HumanCoreExome-12v1.1 chip arrays, 1491 (82 cases, 1409 controls) on Illumina HumanCoreExome-12 v1.0 chip arrays, 133 (119 cases, 14 controls) on Illumina Human660W Quad_v1 chip arrays, 811 (45 cases, 766 controls) on Illumina Human Omni-Quad V.1 chip arrays and 1385 (82 cases, 1303 controls) on Illumina HumanOmniExpress-12v.1.0 chip arrays (**Table 5.1**).

Table 5. 1| Summary of Illumina bead chips used for genotyping different batches of cases and controls

	Genotyping chip	Number of SNPs	No. of cases	No. of controls
Chip1	Illumina HumanCoreExome-12v1.1	542,585	§ 15	§ 350
Chip2	Illumina HumanCoreExome-12v1.0	538,448	§ 82	§ 1409
Chip3	Illumina Human660W-Quad_v1	657,366	§ 119	§ 14
Chip4	Illumina HumanOmni-Quad V.1	1,140,419	§ 45	§ 766
Chip5	Illumina HumanOmniExpress 12v1.0	730,525	£ 82	¥ 1303
Chip6	Illumina HumanOmniExpress-12v1.1	730,725	¥ 243	N/A

¥ Discovery set cases and controls; £ Follow up set cases; § Replication set cases and controls

The study design was restricted to overlap of SNPs in respective chips combined for each of the analysis phase. This amount of overlaps among the SNPs genotyped between the arrays is reported in (**Table 5.2**).

Table 5. 2| Overlaps in number of SNPs prior to quality control between different genotyping arrays used. Chip numbers are defined in Table 5.1

	Chip1	Chip2	Chip3	Chip4	Chip5	Chip6
Chip1	542585	535478	128261	244172	252942	205700
Chip2		538448	128337	244385	253159	205723
Chip3			657366	392615	324520	266040
Chip4				1140419	706093	534858
Chip5					730525	534604
Chip6						730725

5.1.2 MM samples

Diagnosis of MM (International Classification of Disease (ICD)-10 C90.0) adhered to the guidelines established by World Health Organization. Samples retrieved from all subjects were either before treatment or at presentation.

The UK GWAS consisted of 2282 cases (1755 male (post quality control (QC)) recruited through the UK MRC Myeloma-IX and Myeloma-XI trials (ISRCTN68454111: Myeloma-X http://www.isrctn.com/search?q=ISRCTN68454111 and ISRCTN49407852: Myeloma- XI http://www.isrctn.com/search?q=ISRCTN49407852). DNA was extracted from EDTA-venous blood samples (90% before chemotherapy) and genotyped using

Illumina Human OmniExpress-12 v1.0 arrays (Illumina). Controls were recruited from publicly accessible data generated by the Welcome Trust Case Control Consortium (WTCCC) from the 1958 Birth Cohort (58C; also known as the National Child Development Study) and National Blood Service. The control population comprised of 5197 individuals (2628 male (post QC)). Genotyping of these controls was conducted using Illumina Human 1-2 M-Duo Custom_v1 Array chips (www.wtccc.org.uk).

The German GWAS comprised 1717 cases (981 male (post QC); mean age at diagnosis: 59 years). The cases were ascertained by the German-Speaking Multiple Myeloma Multicenter Study Group coordinated by the University Clinic, Heidelberg (ISRCTN06413384: GMMG-HD3 http://www.isrctn.com/search?q=ISRCTN06413384; ISRCTN64455289: GMMG-HD4 http://www.isrctn.com/search?q=ISRCTN64455289; ISRCTN05745813: GMMG-MM5 http://www.isrctn.com/search?q=ISRCTN05745813) . DNA was prepared from EDTA-venous blood or CD138-negative bone marrow cells (<1% tumor contamination). Genotyping of these samples was performed using Illumina Human OmniExpress-12 v1.0 arrays (Illumina). For controls, genotype data on 2,107 healthy individuals, enrolled into the HNR study was used. These samples were genotyped using either Illumina HumanOmni1-Quad_v1 or OmniExpress-12 v1.0 arrays. Out of the whole recruited control population, 2069 (1028 male) remained after QC.

5.2 Expression quantitative trait loci data

eQTL data was generated on malignant plasma cells from 665 German MM patients (389 male, mean age 59±9 years) of the Heidelberg University Clinic and the German-speaking Myeloma Multicenter Group. Plasma cells were CD138-purified from bone marrow aspirates. Gene expression profiling of CD138-purified plasma cells using Affymetrix U133

2.0 plus arrays were performed (Meißner *et al.*, 2011). Expression data have been deposited in ArrayExpress (E-MTAB-2299) (Weinhold *et al.*, 2015).

5.3 Swedish population data

The Swedish Family-Cancer Database (SFCD) includes the total population of Sweden classified in families and linked to the national cancer registry. It records a little over than 2.1 million cancer cases diagnosed in Sweden since 1958 (Chattopadhyay *et al.*, 2018d). The registry relies on distinct obligatory notifications from clinicians who diagnosed the neoplasms and from pathologists/cytologists with an estimated coverage of more than 90% of all cancer diagnoses (Ji *et al.*, 2012). The registry counts tumors not patients, except for skin and urinary tract tumors diagnosed at the same topological area (https://www.ancr.nu/dyn/resources/File/file/7/4247/1412940269/total_document_survey_optimeret.pdf). The project database is located at Center for primary health care in Malmö, Sweden.

Chapter 6: Heritable risk analysis

6.1 Quality control

Erroneous study design and faulty genotype calling introduces systematic bias in genetic cases-control studies. It leads to spurious associations increasing the amount of both false-positive and false-negative discoveries (Zondervan and Cardon, 2007). Due to the enormous number of tests to be performed, even a negligible amount of systematic error can introduce bias that can inflate or deflate signals invalidating the sensitivity and specificity of the results. This error can potentially be introduced in two stages of the study, initially in the genotypes of the study samples and then due to outlier-like markers present in original and extrapolated sample (Anderson *et al.*, 2010). Therefore, in the two-staged QC process thus applied, the second stage can also correct for loss in detection power due to removal of samples with detrimental effect on the analysis (Marchini *et al.*, 2007). These two stages of QC are described as follows:

6.1.1 Sample-based quality control

a) Sex check

Genotype data from the sex chromosome of the samples is tallied against the ascertained sex in the sample to detect discordance in sex determination. As the homozygosity rates for males (ideally 1) and females (<0.2) differ substantially, DNA sample and report concordance abnormalities are easily detectable. Errors of such kind are often due to misreporting or sample getting mixed up, although possibilities of erroneous genotyping of sex chromosome also remain.

b) Heterozygosity rate

Variation in DNA sample quality can have a major impact in determining strength of associated signals. Heterozygosity rate helps determining individual DNA sample quality of every sample. Excess heterozygote genotypes indicate contaminated DNA sample where as a low observation means highly inbred sample. Hence samples are pruned for proportion of heterozygous genotypes.

c) Relatedness

Basic assumption of case-control study design is that each individual should be distant in pedigree compared to that among second-degree family relatives. This is checked with the statistic *identity by state* (IBS).

d) Population stratification

Often based on ancestry, certain loci are shown to have undergone strong selection (Campbell *et al.*, 2005). These loci if left unchecked to their own devices, can introduce inherent population stratification resulting in instable sensitivity (Cardon and Palmer, 2003). Principal component analysis is probably the most common tool to identify (and consequently remove) people with extensive modifications in ancestry (Price *et al.*, 2006). In this analysis principal component model was assessed using genome-wide template genotype data obtained from populaces of reported ancestries using the genotype data from phase II HapMap project for Europe, CEU (60); Asia (90 CHB + 90 JPT) and Africa (60 YRI). Thus from the overlapping clusters, outliers due to dubious ancestry were detected and removed.

6.1.2 Marker-based quality control

a) Genotyping failure

Removal of substandard markers is of immense importance as they introduce large variance in data that compromises quality. Based on call-rates (< 99%) of markers, all outliers were removed

b) Hardy-Weinberg equilibrium

Markers that violate Hardy-Weinberg equilibrium (HWE) are evidence of problematic genotype distribution and are fatal to the analysis (but in *cases*, it may be indicative of selection and may be causal to phenotype). Departure from HWE in control often means genotyping errors that generate enormous type 1 error and thus are removed.

c) Differential missingness

SNPs with considerable inconsistencies in missing genotype rates among cases and controls introduce confounding due to missingness (Moskvina *et al.*, 2006). These errors usually appear if cases and controls are genotyped separately or in separate arrays. Sample qualities, array exhaustiveness, batch effect in sampling are some of the main reason for such discrepancy. These erroneous markers are hence removed with simultaneous calling of cases and controls (Plagnol *et al.*, 2007).

d) Minor allele frequency

Due to the voluminous number of tests compared to the number of samples, detection power for rare alleles is very low in genome-wide studies (Morris and Zeggini, 2009). Secondly, alleles with very low frequencies if present, introduces high rate of false positive signals in these analyses (Anderson *et al.*, 2010). Hence SNPs harboring alleles with very low frequencies are removed.

6.2 Phasing

Determination of haplotype phase estimated using computational approaches is an integral part of GWAS. Such computational methods are based on pooling information from genotype data across individuals to have estimated haplotype phase. Unrelated individuals are *phased* by assuming that common haplotype sets can explicate the probable observed genotypes (Browning and Browning, 2011). This unrelated set of individuals work as a genetic blueprint in construction of the final genetic assembly; hence the number of individuals considered in the reference panel (blueprint) is of utmost importance. The genotype data was phased against the reference panel released by the phase II HapMap project (The International HapMap Consortium *et al.*, 2010).

6.3 Imputation

Imputation is the prediction algorithm to obtain a denser genetic assembly to increase statistical power of detection. During imputation, the sample genotypes are used to be mapped on a denser platform to estimate genotypes for untyped SNPs in sample, generally obtained from published reference genotypes. This increases power and overlap between several different samples. Imputation was performed with combined phased haplotype (3,781 UK individuals) from UK10K project (http://www.uk10k.org/) and (1,092 individuals from Africa (n=246), Asia (n=286), Europe (n=379) and the Americas (n=181)) 1000 genome project (http://www.internationalgenome.org/) reference panel on NCBI build 37 (human genome 19, hg19) (Huang *et al.*, 2015; Marchini and Howie, 2010).

6.4 Association study

Association analysis evaluates the effect of individual markers (SNPs) compared between presence and absence of a given phenotype. The underlying hypothesis is that adjacent stretches of DNA are non-independently inherited through generations which lets tagged SNPs assume similar signal to that displayed by a causal SNP in such a co-inherited region (Daly *et al.*, 2001). Associations between SNPs and MGUS (and MM) phenotype were assessed by fitting logistic regression with the presumption of an additive inheritance model. Risk alleles for each of the sample population were assessed against the reference and odds ratios (95% CI) were calculated from the regression estimate and a test of association rendered the *P*-value. Adequacy of the distribution of P-values was later checked with quantile-quantile plots under the assumption of null hypothesis.

6.5 Interaction study (Epistasis)

Epistasis i.e. genetic interactions is recognized as fundamental in understanding the structure as well as functionality of biological networks and evolutionary processes of multifaceted traits for long (Phillips, 1998). Fisher introduced the idea of statistical epistasis where "*the average deviation of combinations of alleles at different loci is estimated over all other genotypes present within a population*" for studying association of the deviation and a phenotype of interest (Tong *et al.*, 2001). Genetic interaction was performed with logistic regression considering fixed effects due to any two SNPs and their joint interactive effect again with an additive inheritance model. Odds ratios and other quantitative measures were obtained in a similar fashion as was done for association analyses.

6.6 Meta-analysis

Meta-analyses were undertaken to obtain pooled estimates using the Mantel-Haenszel method to combine raw data. Joint odds ratios, 95% CIs and P values were obtained with an inverse variance weighted fixed effects model for both association and interaction obtained results.

6.7 Linkage disequilibrium score regression

Linkage disequilibrium score regression (LD score regression) assumes that due to inherited genomic widow around each GWAS detected loci, the estimated effect size of any given SNP is due to the combined effect of all SNPs in LD. Hence for polygenic complex traits, SNPs in high LD will contribute a larger χ^2test statistic compared to the SNPs with low LD (Bulik-Sullivan *et al.*, 2015b). This method tests for genetic correlation between two study populations that helps estimating dependency between two traits. Extending from the same assumption, it would mean for testing whether two traits are genetically correlated, multiplied Z scores (scale and location adjusted statistic) can be tested against χ^2 (Bulik-Sullivan *et al.*, 2015b; Yang *et al.*, 2011b). Baseline LD scores were initially calculated from genetic data supplied by the individuals of European ancestry in the UK10K and 1000 genomes project (Bulik-Sullivan *et al.*, 2015a). LD score regression was then assessed to obtain effect size of genetic correlation between meta-analyzed MM and MGUS.

6.8 Expression quantitative trait loci

As GWAS can only decipher associative genetic loci when it comes to investigating incidence of phenotype, gene expressions are widely used as quantitative trait to impose biological mechanics on the observed signals. In principle eQTL analysis links DNA

sequence variants to tissue specific differential regulations in gene expression (Clyde, 2017; Gilad *et al.*, 2008). Studies often focus on genes residing in the nearby regions of a queried SNP (*cis*-eQTL). eQTL data on malignant plasma cells of 665 German MM patients was used to observe changes in gene expression against the genetic-interaction detected sentinel SNPs (Weinhold *et al.*, 2015).

6.9 Summary data-based Mendelian randomization

In order to identify causal signals, summary data-based Mendelian randomization (SMR) analysis (http://cnsgenomics.com/software/smr/) was performed (Zhu *et al.*, 2016). Mendelian randomization (MR) is a statistical tool that uses instrumental variable to assess causal association between an exposure and its effect (Paternoster *et al.*, 2017). There has been several studies that focuses on application of MR in GWAS (Benn and Nordestgaard, 2018; Porcu *et al.*, 2018). The major problem is in determining the validation of the three major assumptions for selection of a justified instrumental variable (more precisely, the exclusion principle criteria) (Davey Smith and Hemani, 2014). In SMR gene expression is treated as the exposure, the phenotype is the effect of the exposure and the genotype-associated summary data is treated as the instrumental variable. Conformity to the underlying assumptions of the instrumental variable was examined by testing heterogeneity for independent instruments (gene expression) against multiple SNPs present in each *cis*-eQTL window. Under the null hypothesis of absence of pleiotropy, effect sizes for all the SNPs belonging to a *cis*-eQTL region would demonstrate identical effect sizes. Hence by testing for absence of heterogeneity among effect sizes between SNPs in *cis*-eQTL, the instrumental variable assumptions were also controlled for.

6.10 Selection of test statistic

The notion of statistical interaction or 'biological epistasis' if simply put is additional effect exerted by a selection of alleles on top of each of their fixed effects on expression of a quantity (or that of a phenotype). In that case the simplest form of regression structure would be modeled as a departure from single co-variate regression model by introducing join effects due to more than one variant on the phenotype odds. It'll take the following shape:

$$\log\frac{p}{1-p} = a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2$$

Where p represents probability of incitation of the chosen phenotype which in case-control set up can be replaced with the point estimate of proportion of cases in the sample population; a_0 is the location or intercept parameter (can be null depending on scenarios); a_1 and a_2 are fixed effects parameters due to marker x_1 and x_2 ; and finally a_3 is the parameter for the interaction effect due to x_1 and x_2.

Initial releases of the computational platform PLINK includes this simple logistic regression oriented model which is traditionally very computational-heavy and resource hungry. As an improvement to this design fast-epistasis is later incorporated which uses linkage dependent correlative measure as test statistic. In summary, it takes the unphased genotype data from the sample and first expands it in a 2-way cross tabular format for estimating each of the allelic pair frequencies (**Table 6.1** to **Table 6.2**)

Table 6. 1| Genotype counts from two SNPs

	BB	Bb	bb
AA	n_{22}	n_{21}	n_{20}
Aa	n_{12}	n_{11}	n_{10}
aa	n_{02}	n_{01}	n_{00}

Table 6. 2| Allele counts for alleles of two SNPs obtained from genotypes

	B	b
A	$K_1 = 4n_{22} + 2n_{21} + 2n_{12} + n_{11}$	$K_2 = 4n_{20} + 2n_{21} + 2n_{10} + n_{11}$
a	$K_3 = 4n_{02} + 2n_{01} + 2n_{12} + n_{11}$	$K_4 = 4n_{00} + 2n_{01} + 2n_{10} + n_{11}$

The log odds ratio with this table is calculated by the formula: $L_m \log \frac{K_1 K_4}{K_2 K_3}$ with a sample estimated variance $L_v = \frac{1}{K_1} + \frac{1}{K_2} + \frac{1}{K_3} + \frac{1}{K_4}$.The fast-epistasis algorithm tests under the null hypothesis that no correlation between the alleles at the two loci exists with a test on the case-control generated statistic:

$$T_{fast-epistasis} = \frac{[L_m(cases) - L_m(controls)]^2}{L_v(cases) + L_v(controls)}$$

Wellek and Ziegler later pointed out that instead of using allelic frequency proportion, if Pearsonian r is used, it can be calculated without phasing and additionally the loss in precision due to using unphased genotype data is negligible subject to the variants' conformity to HWE (Wellek and Ziegler, 2009). In this case for obtaining a Pearsonian r based measure, the statistic proposed was:

$$r = \frac{2(K_1 K_4 - K_2 K_3)}{\sqrt{(K_{1+}K_2)(K_1 + K_3)(K_2 + K_4)(K_3 + K_4)}}$$

The sample variance is also calculated in a similar estimation procedure (Wellek and Ziegler, 2009). Although their study proposes on squared difference of Z score as mean square estimate, Ueki et al. proposed a generalized efficient form if this correlation based statistic:

$$T_{WZ_{case/control}} = \frac{[r(cases) - r(controls)]^2}{Var(r(cases)) + Var(r(controls))}$$

This transformed statistic has been used for all cases-control interaction analysis discussed throughout this work.

6.11 Threshold selection

The ideology behind use of *P*-value is at the least very well discussed but often poorly understood. In brief, *P*-vale of a quantity observed in sample quantifies the probability of obtaining a value at least as extreme as that observed given it is selected from the population by chance. Hence, if the *P*-value obtained from a test is smaller than a predetermined level of significance, then the null hypothesis is rejected (constraints may apply such as simple null, test statistic is uniformly minimum variance in its class etc.). Hence the two pronged problem is first, to determine a good statistic, possibly an unbiased one (which was done for both single marker and two-marker association test); and secondly, to determine the predefined threshold in a way that it controls for type I and type II errors (possibly keeps them simultaneously to a minimum). Traditionally at 95% confidence, for a single test performed to control the family wise error rate (number of type I error ≤ 1) level of significance need to be fixed at 0.05.

Genome-wide studies pose a problem with the number of tests performed. To understand this in short, if there are *N* number of tests performed at 95% confidence, it'd mean 5% of those *N* tests performed would end up being a false discovery just by chance. Now for 1 million tests performed, this number is 50,000 which compromise integrity of the design. Hence, for large number of tests to be performed, proportionally large confidence is required to restrict the number of false positive signals. Because of the LD structure present among the genomic variants, hypothetically a large amount of tests would be testing for identical association, hence correcting the level of significance with a factor of *N* (Bonferroni

correction) would result in overcorrection and subsequent deletion of true signals i.e. Type II error (Noble, 2009). As for the GWAS, a standard level of significance at $5x10^{-8}$ was employed (Fadista *et al.*, 2016). Whereas depending on final number of tests after selection, this threshold was determined at $5x10^{-10}$ for GWIS. Additional tests performed throughout the work were all corrected with Bonferroni correction unless explicitly stated otherwise.

6.12 Resources

6.12.1 *In Silico* analysis tools

a) CASSI

CASSI is a genome-wide interaction analysis software (https://www.staff.ncl.ac.uk/richard.howey/cassi/). It can implement a transformed Pearsonian *r* statistic centered genetic interaction test. Fixed effects based logistic regression was assessed with Wellek-Ziegler statistic to obtain association odds ratio, 95% CI and P-value.

In separate steps case-control and case-only studies were performed with inherent options defined in the tool for discovery, validation and replication analyses respectively. A default selection criteria for each SNPs was passed which was employed according to the single marker association test subject to passing the threshold level of significance at <0.001. This reduced the number of tests to a legible amount and conserved the effective level of significance to be employed on the output after correcting for multiple testing.

b) Genome-wide Complex Trait Analyses

Although originally developed for heritability estimation, GCTA (http://cnsgenomics.com/software/gcta/#Overview) has developed as a hub for tools for interrogating several genetic features of phenotypes using GWAS level summary statistics (Yang *et al.*, 2011a). GCTA was used to obtain results of an array of different statistical queries

PCA was used to perform principle component analysis in determining population structure and stratification.

SMR is a tool that interrogates GWAS summary data and a related exposure (eQTL data) to perform Mendelian randomization. Causal sentinel SNPs were identified with this tool.

c) IMPUTE

IMPUTE implements a haploid extrapolation algorithm (https://jmarchini.org/impute-4/) to impute untyped SNPs in association studies. It uses embedded reference genotype panel from UK Biobank (http://www.ukbiobank.ac.uk/) dataset containing around 500,000 individuals (Bycroft *et al.*, 2017). It uses weighted genotypes from sample population and extrapolates genotypes of the SNPs present in the reference panels consistent with the LD structure (genome build). It extrapolates genotypes of the said SNPs in LD from the typed SNP in sample where the signal intensity is counter-proportional to that of the genomic distance between the two. As an output it provides marginal probabilities of each projected genotype. In previous build of the program along with the projections additional information were included on information content as proxy for 'genotype quality' in a probabilistic scale of 0 to 1 which enabled user to determine imputation confidence for each variant (Marchini and Howie, 2010). Since this metric was discontinued in the latest build (IMPUTE4), the information pruning was carried out in SNPTEST.

d) INTERSNP

INTERSNP (http://intersnp.meb.uni-bonn.de) is a JAVA based alternative for computation of epistasis. The additional attribute of this package is in predictively determining genome-wide significance for each SNP with Monte Carlo simulation

(Herold *et al.*, 2009). Furthermore it is capable of considering more than two variants at a time building a complex linear interactive model for association test (gene-gene interaction or Fisher's combined test). In addition to GWIS, it includes embedded KEGG database that can be leveraged with the identified SNPs for assessing pathway enrichment with prioritized genes (Herold *et al.*, 2012). INTERSNP was used as a parallel tool to perform discovery analysis.

e) METAINTER

METAINTER is a meta-analysis tool for multiple regression analysis in GWAS and GWIS (https://metainter.meb.uni-bonn.de/) (Vaitsiakhovich *et al.*, 2015). It specializes in the file format generated with parameter provided by the INTERSNP platform. It enforces a modified 'method for the synthesis of linear regression slopes'. With an inverse variance weighted model from each population, the meta-analytic results include odds ratio, 95% CI and P value for each paired observation. It also outputs I^2 statistics as a measure of heterogeneity between the study populations.

f) METAL

METAL (http://www.sph.umich.edu/csg/abecasis/metal/) is a tool for meta analyzing GWAS which reads logistic regression output file created by PLINK (Willer *et al.*, 2010). It is a widely used tool in meta-anlyses as it can employ inverse-variance weighted linear regression on a fixed-effects model and also sample size weighted combined Z-score model (Anonymous, 2011; Lambert *et al.*, 2013; Locke *et al.*, 2015). METAL also includes a process for including genomic controlling parameter for each input population (Willer *et al.*, 2010). By comparing the median of the observed test statistic to that expected by chance, METAL estimates inflation in the test statistic.

Additional filters can also be imposed on the selection of markers obtained from each of the GWAS studies included in the analysis.

g) PLINK

PLINK v2.7 (http://zzz.bwh.harvard.edu/plink/) is an improved version of PLINK, a whole genome association computational tool designed for a range of basic large-scale analyses (Chang *et al.*, 2015; Purcell *et al.*, 2007). Several different toolkits are embedded in PLINK that can be queried from data extraction, deletion, manipulation, several steps of QC required before imputation, association analyses with different statistical models, data type conversion, pruning and so on. PLINK is typically one of the most valuable tools in any traditional GWAS design and have been used in this work in following steps:

Sample-based QC Check for sex information, heterozygosity rates and relatedness.

Marker-based QC Check for genotyping failure, differential missingness, HWE and MAF.

LD based SNP pruning.

Checking and flipping transposed alleles between reference panel and sample population.

Logistic regression based association analysis.

Fast epistasis for gene-gene interaction based association analysis.

h) QCTOOL

QCTOOL version 2 (http://www.well.ox.ac.uk/~gav/qctool_v2/) is an improvement on original release QCTOOL that performs an array of quality control related computational algorithms on genome-wide genetic data. Among many other functions, it

can annotate, merge, perform GWAS related QC steps. This tool was mainly used in this study to employ a screening based on the imputation quality in one of the final stages of QC. In some cases it was also used to prune on defined MAFs.

i) R

R v3.3.1 (https://www.R-project.org/) is a publicly owned computational platform. It is used from statistical analyses to large scale data processing, editing and for producing figures (R Development Core Team, 2018). To this end numerous packages were used that cater to specific types of statistical, bioinformatic computation or graphics processing.

j) SHAPEIT

Genotype data stored in the sample files are generally in the shape of unphased haplotypes. Hence it is ambiguous to determine which of the parental chromosomes or haplotypes an allele belong to which makes identification of co-localized alleles in shorter genomic windows impossible which is a primary requirement to test assumption of an association study design. At the same time, for imputation shared co-localized haplotypes are grouped between the study sample and the genotyped reference panels.

Segmented HAPlotype Estimation and Imputation Tool (SHAPEITv2) uses hidden Markov chain model to estimate haplotypes from genotype data (Delaneau *et al.*, 2011; Delaneau *et al.*, 2012). This method is used to create phased haplotypes from the sample genotypes that are later utilized in combination with the reference panels during imputation. In this study, the sample data and the reference haplotypes were pre-phased letting the imputation procedure run much efficiently (Delaneau *et al.*, 2014).

k) SNPTEST

SNPTEST v2.5.4 (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) is a tool for single marker association analysis in study design of GWAS (Marchini and Howie, 2010; Marchini *et al.*, 2007; The Wellcome Trust Case Control *et al.*, 2007). This tool was built to intake file formats directly generated by IMPUTE and perform association analysis by adjusting for covariates (if available). SNPs can be tested for phenotypic association under several different models such as additive, dominant, recessive, general or even heterozygote model. This was used in converting file times from IMPUTE4 output with the help of the corresponding sample information to the binary or pedigree format accessible by PLINK.

6.12.2 Web-based resources

a) 1000 Genomes project

The 1000 genome project was initiated to create a comprehensive catalogue of human genetic variation that are present with >1% (The Genomes Project *et al.*, 2012). It attains a dense map with sequencing large numbers of individuals at 4x coverage. Data from the pilot phase, phase one and phase three of the project have been made publicly available. It is currently the largest publicly available resource for genome-wide variant frequency data across different populations worldwide. Variant data from 1000 genome project was used in combination with sample data obtained for MGUS and MM separately to allow for accurate imputation of variants not directly covered in the low coverage genotyping.

b) Ensemble genome browser

The Ensemble genome browser (https://www.ensembl.org) is a genome annotation tool maintained by European bioinformatic institute. It was used to retrieve genetic information regarding genomic assembly of introns, exons and regulatory domains, known transcripts etc. (Yates *et al.*, 2016).

c) HaploReg

The HaploReg database (http://archive.broadinstitute.org/mammals/haploreg/) has a web based user interface that queries annotation of genomic variants in non-coding regions (Ward and Kellis, 2012). It also helps to visualize all other variants in user specified linkage threshold and respective chromatic state and their effect on regulatory motifs.

d) National Centre for Biotechnology Information

The NCBI web server (http://www.ncbi.nlm.nih.gov/) acts as a host for a multitude of databases and bioinformatics tools (Coordinators, 2013). Several different aspects and features of the browser were leveraged as follows:

dbSNP a database containing information on genetic variations. SNPs were queried for chromosomal position, allele frequency, and genomic build position.

Pubmed was utilized for literature search and citation gathering.

RefSeq was used for interrogating annotation.

e) The International HapMap project

The international HapMap project (http://hapmap.ncbi.nlm.nih.gov/) aims at cataloguing all common genetic variants present in human genome across several different populations (The International HapMap Consortium *et al.*, 2003). It enables retrieval of high density SNP genotype data geo-spatially stratified over populations representing different ancestries (Caucasian, Chinese, Japanese and African). Variant data obtained

from HapMap project was used to cluster sample population and detect population stratification.

f) UK10K project

The UK10K project aims at sequencing genome of 10,000 individuals at even a larger coverage (6x) (The U. K. K. Consortium *et al.*, 2015). It is mainly comprised of three cohorts: the Twins UK and the Avon Longitudinal Study of Parents and Children (ALSPAC) cohorts include 1,854 and 1,927 whole-genome sequenced individuals and a further 6,000 individuals with specific phenotypes (neurodevelopment, obesity and rare diseases) have been exome sequenced. Variant data from the whole-genome sequenced cohorts were used in combination with the 1000 genome reference panel for the imputation.

g) University of California Santa Cruz genome browser

The UCSC genome browser (http://genome.ucsc.edu/) is a virtual map of the human genome, annotated with known genes, transcripts, polymorphic variation, repeated sequences, conservation, structural variation and experimental data from external databases. These features are mapped against the physical map of chromosomes in an interactable interface. Various bioinformatic tools are embedded in the website and are utilized for visualizing genes, specific regions of DNA, introns, regulatory regions and other specific features of genomic region; in downloading tables argued by specific queries.

Chapter 7: Enrichment analysis

7.1 Gene prioritization and genetic network

Relating analytically discovered risk loci to a presumed polygenic phenotype requires identification of possibly causal sets of genes that are involved in pathway regulation or have some functional classification (Moreau and Tranchevent, 2012). As the set of possibly relevant genes are assumed to co-localize the sentinel variants discovered via analytical methods (such as GWAS, GWIS etc.), it is of immense important to identify which of those genes are responsible to have impact on the phenotype (often not only the mere expression of a disease phenotype but also disease state, stage differentiation, responsiveness to therapy and so on) and in what way (Hung *et al.*, 2012). Till date prioritization via gene set enrichment has proven involvement of several pre-associated genes in cancer and in some cases those with active mechanistic link to therapy response via gene expression modulation (Garnett *et al.*, 2012; Haibe-Kains *et al.*, 2012). Shortly, gene prioritization algorithms invoke prior knowledge of a phenotype or a process of interest either with the description of the trait with key words or by constructing a scaffold of genes already known to have been associated to the entity (Aerts *et al.*, 2006). Either of the two or both in combination are later interrogated in a biological network to identify the most closely associated candidates usually with *guilt by association* technique (Oti and Brunner, 2006; Perez-Iratxeta *et al.*, 2002). The algorithm works by querying databases containing simple networks of genes or proteins (such as protein-protein network) and associating those to the existing query genes based on a predefined enrichment index calculated depending on several quantities such as the strength of each edge, directional weights, node importance (Lage *et al.*, 2007). Several

different algorithms were used to perform gene prioritization as well as to construct protein-protein enrichment network for this study.

7.2 Pathway enrichment

In 2012, Califano *et al*. said in their introduction to discussion of strategies to leverage GWAS in building network based association models, "*the results of genome-wide association studies (GWAS) have been mostly sobering*" (Califano *et al.*, 2012). They go on to argue that association detection just identifies suggestive genomic region(s) without any prompt biological or mechanistic inheritance; arguments similar to that made by researchers before (Hardy and Singleton, 2009; Manolio *et al.*, 2009; Stranger *et al.*, 2011). But more importantly so, it was more a critique on dearth of ways to infer biological understanding rather than that of GWAS as associated loci encompasses candidate gene studies or linkage detected loci where this was a pre-existing concern voiced years ago (Altshuler *et al.*, 2008; Goldstein, 2009; Kraft and Hunter, 2009; Lyssenko *et al.*, 2008). A standalone solution to this problem is relating the prioritized genes thus detected with biological pathways combined with different data modalities (Subramanian *et al.*, 2005; Zhong *et al.*, 2010). Since genetic pathways are frequently implicated in susceptibility to phenotypes as well as progression of disease (Schadt, 2009), considering existing contextual knowledge-base of genes and pathways into account provides a superior probability of understanding underlying mechanisms enforced by genes in pathogenesis. To this end preliminary algorithms were introduced based on expression regulation of the prioritized genes as temporal aggregation of phenotypic burden was well known to be caused by moderate expression regulation in a cluster of genes (Mootha *et al.*, 2003; Subramanian *et al.*, 2007). This model of pathway identification has been largely successful and has collectively shown

that pathway discovery based approaches provide additional functional information complementary to traditional risk loci bases detection of genes (Chen *et al.*, 2010; Holmans *et al.*, 2009; Peng *et al.*, 2009; Wang *et al.*, 2009). This study implements pathway enrichment analysis in several stages with different tools to infer such biological underpinnings.

7.3 Tissue and cell enrichment

Similar to the notion of pathway enrichment, cell and tissue enrichment exerts additional implicating information on specific regulatory patterns of the association analysis detected loci. In addition to GWAS summary data, expression omnibus derived data can be used in prioritization of cell and tissue types. Although this is a new direction in including complementary evidence along with that obtained via pathway analyses, several studies reported successful validation of the sentinel loci and pathways detected which are found to be enriched in cells and tissues of contextual importance (Chan *et al.*, 2015; Geller *et al.*, 2014; Locke *et al.*, 2015; Shungin *et al.*, 2015; van der Valk *et al.*, 2015; Wood *et al.*, 2014). Following similar design, tissue and cell enrichment was performed to obtain evidence on justification of involvement of the loci discovered.

7.4 Resources

7.4.1 *In silico* analysis tools

a) MAGENTA

Meta-Analysis Gene Set Enrichment of Variant Associations (MAGENTA version 2.4) is a pathway analysis module in MATLAB compiled and maintained by Broad Institute (https://software .broadinstitute.org/mpg/magenta/) (Segrè *et al.*, 2010). It tests for

enrichment of genetic associations in functionally associated genes or predefined biological processes using SNP data obtained from association studies as input. It only requires human genome build dependent chromosomal position of the associated loci and the association test *P*-values. Literature on pathway analyses in general is fret with application of MAGENTA (Global Lipids Genetics *et al.*, 2013; Lango Allen *et al.*, 2010; Okada *et al.*, 2013; Speliotes *et al.*, 2010; the *et al.*, 2012; The International Consortium for Blood Pressure Genome-Wide Association *et al.*, 2011). For executing pathway analysis, single marker association *P*-values and chromosomal regions (hg19) were annotated to genes corresponding to a pre-existing chromosomal range and computation of gene prioritization based pathway enrichment was applied on non-confounders. False discovery rate (FDR) was inherently adjusted for multiple testing with Bonferroni correction and was provided by the software (Segrè *et al.*, 2010).

b) PASCAL

Pathway Scoring Algorithm (PASCAL) is another pathway analysis tool developed for association summary statistics for variants annotated to genes (https://www2.unil.ch/cbg/index.php?title=Pascal) (Lamparter *et al.*, 2016). PASCAL uses maximum of chi-squares (MOCS) or sum of chi-squares (SOCS) statistics with null Gamma distribution with varying degrees of freedom which has proven to be a potent estimator in several investigations (Ghosh and Bouchard, 2017; Watanabe *et al.*, 2017). To create mapping of genes and single entity gene-fusions, it considers all genes within a symmetric genomic window with an index SNP in the middle and fuses all the annotated / flanking genes together when they were found regulated in same pathway(s) to have created a single genetic entity with greater weight subject to linkage among the SNPs. With empirical sampling and subsequent supervised clustering according to the significance

levels extracted from the single marker association tests, it utilizes this idea of gene fusions i.e. clusters of correlated genes, for which the variants are in LD. Gene set prioritization and subsequent pathway enrichment analysis was performed using single marker *P*-values obtained from GWAS on MGUS and MM. Although pathway scoring was performed using both MOCS and SOCS statistics, the results were comparable and SOCS produced deflated significance levels with similar order as of that by MOCS. Sum of chi-square statistics with individual one degree of freedom was computed by summing over association statistics corresponding to each pathway. Enrichment scores of individual pathways were subsequently obtained by a test assuming chi-square distribution with degrees of freedom equal to the cardinality of fused gene sets. The enrichment FDRs against each pathway were again provided by the tool with correction due to multiple testing.

c) DEPICT

Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) is a large-scale enrichment analytical tool also maintained by Broad Institute (https://data.broadinstitute.org/mpg/depict/index.html) (Pers *et al.*, 2015). It uses predicted gene functions obtained from embedded data to prioritize the most likely causal genes derived from the sentinel SNPs and its tagged loci in LD to identify enriched pathway and highlight tissues and cells where these causal genes are most likely to have differential expression. DEPICT is built on an algorithm of large-scale gene co-expression analyses, leveraging the summary statistics of GWAS (Locke *et al.*, 2015; Wood *et al.*, 2014). Depict derives enrichment analysis viability from 77,840 gene expression datasets. DEPICT's gene set knowledge base derived from Gene ontology (GO), Ensemble, The

Mammalian Phenotype (MP), KEGG and REACTOME was employed and analogous analyses followed.

It was employed to analyze tissue and cell type enrichment that predicts differential regulation of the selected loci on any of the Medical Subject Heading (MeSH) annotations. To this end 209 such annotations were tested for 37,427 inbuilt backend human microarrays on the Affymetrix HGU133a2.0 array platform. The tissue/cell type enrichment is thus performed on the normalized expression matrix after subjecting it to user selected dimension reduction criteria. SNP pairs discovered with interaction test represented uniquely mapped regions against which the enrichment was tested against a conservative threshold of significance at negative log transformed *P*-value of 2.37 correcting for multiple testing which retained the false discovery rate at <5% (Geller *et al.*, 2014).

7.4.2 Web-based resources

a) BioCarta

BioCarta (http://www.biocarta.com/) is community-fed database featuring a collection of dynamic map of metabolic and signal transduction pathways maintained by NCI. This database is downloadable and virtually importable in any open source analysis tool to investigate pathway map, enrichment and so on. Pathway data from BioCarta was used for pathway analysis via MAGENTA and PASCAL.

b) Ensembl

Although primarily known for the Ensembl genome browser (http://www.ensembl.org/index.html), this resource maintained by European Bioinformatic Institute in collaboration with European molecular biology laboratory includes genetic libraries on

biological pathways, genomic organization of exons, introns and known regulatory domains, known transcripts, proteins, homologues and recorded variation within the gene sequence (Zerbino *et al.*, 2018). Ensembl pathway library is embedded in DEPICT and was explored in pathway enrichment analysis.

c) Gene Ontology

Gene ontology (GO) is part of genetic annotation initiative called the Open Biomedical Ontologies maintained by Gene Ontology Consortium (http://www.geneontology.org/) (Ashburner *et al.*, 2000). GO annotation tracks function of a gene with a measure of associations defined between genes and GO terms. These GO terms are often used as schematics of underlying biological processes hence relatable to biological pathways. GO database is readily embedded in MAGENTA and DEPICT for investigation on biological process enrichment.

d) Kyoto Encyclopedia of Genes and Genomes

Kyoto Encyclopedia of Genes and genomes (KEGG) is a highly diverse database of high-level biological functions and utilities of such systems (https://www.genome.jp/kegg/). It encompasses an array of datasets including those for pathway maps, process hierarchy, orthologs, compounds, biochemical reactions, enzyme, disease associated biological networks, drugs etc. KEGG pathways in particular is a library of manually drawn pathways that connects and relies on biological entities such as Metabolism, genetic information, cellular processes, interaction reaction with environmental features, organismal systems, human diseases, drug developments and so on. It is one of the most exhaustive existing libraries of biological pathways present till date. KEGG pathway data were in use by all three of the pathway analysis tool in concern.

e) Mammalian Phenotype

The Mammalian Phenotype (MP) Ontology is an annotation library that relates phenotypes present in mammals and it is part of the Mouse Genome Database (MGD) Project (http://www.informatics.jax.org/vocab/mp_ontology/) (Smith and Eppig, 2009). MP terms represent observable morphological, physiological, behavioral and other features of mammals and are used as proxy for biological phenotypes by DEPICT.

f) Protein Analysis through Evolutionary Relationships

Protein Analysis through Evolutionary Relationships (PANTHER) pathway is also a database maintained under the flagship project of GO (http://www.pantherdb.org/pathway/). It includes over 177 manually curated biological pathway maps providing individual mapping to subfamilies and protein sequences (Mi *et al.*, 2013). In the present study PANTHER was only used by MAGENTA.

g) Reactome

Reactome (https://reactome.org) is another curated database of biological pathways and reactions related to human biology. Reactome encompasses all binding, activation, translocation, degradation and classical biochemical events involving a catalyst as 'reactions' which is defined as any event that facilitates change of a biological molecule. Reactome also includes cross-references with KEGG, Ensembl, GO and other similar databases helping matching of biological functions overlapped through them. Reactome pathway definitions were imported via all three *in silico* pathway analysis tools used in this study.

h) STRING

STRING (https://string-db.org/) is a library of protein-protein interaction network maintained by Swiss Institute of Bioinformatics, CPR-NNF center for Protein Research

and European molecular Biology laboratory (Szklarczyk *et al.*, 2017). This database consolidates known and predicted protein–protein association data for a large number of organisms. In STRING protein-protein interactions are predicted from: (a) systematic co-expression analysis, (b) detection of shared selective signals across genomes, (c) automated text-mining of the scientific literature and (d) computational transfer of interaction knowledge between organisms based on gene orthology (Szklarczyk *et al.*, 2017). It also provides a dynamic user interface to visualize the interaction assembly predicted. STRING only requires protein names and referred organism for input to create a network.

The network thus constructed will retain all the proteins provided in input as nodes but only include interacting edges that have used defined threshold of confidence (interaction score). This threshold can be manually defined although the system-defined preset is at 0.4. It provides an enrichment *P*-value of the network by comparing expected and observed edges in the network. Node colors in the interaction map signify different/shared protein functionality. Colored edges convey status of predicted network edge correspondingly (a) cyan: curated database, (b) magenta: experimentally determined (c) forest green: gene neighborhood (d) red: gene fusion (e) navy blue: gene co-occurrence, (f) lawn green: text mining, (g) black: co-expression and (h) lavender indigo: protein homology. The interactive menu also lets user add first order interacting proteins to the network based on protein-protein enrichment score. There are additional options of performing K-means or MCL clustering on the network to stratify the network based on interaction score. Analysis of genetic-network with MGUS and MM GWAS/GWIS summary data in its entirety was performed with this tool.

Chapter 8: Second primary cancer risk analysis

8.1 Case identification

SFCD was used to obtain data for the present study that includes information concerning the residents of Sweden organized in families and covers more than a century and through several generations. The cancers diagnosed in the Swedish residents (first or any subsequent cancers including *in situ* tumors) are linked to the cancer registry with an individual unique proxy (Hemminki *et al.*, 2009). The database was first created in 1998 and it consists of all the cancer cases in Sweden since 1958 (Hemminki and Vaittinen, 1998). The most recent release of SFCD encompasses over 2.1 million cancer diagnosis (malignant tumors) in a little more than 16.1 million individuals until the end of 2015 (Chattopadhyay *et al.*, 2018a). For all individuals born after 1932 in Sweden, SFCD provides linkage to the same information of their biological parent(s) through the Multi-generation Register with few exceptions (lack of data for some older Swedes and immigrants) effectively making them the offspring generation (Hemminki, 2001).

The database records cancers according to the ICD-7 and also incorporates later revisions. The following system was followed to have all the cancers classified in broader categories described in **Table 8.1**.

Table 8. 1| Cancer site classification based on ICD-7

ICD-7	Cancer classification	ICD-7	Cancer classification
140, 141, 143-148, 161	Upper aerodigestive tract	177	Prostate
142	Salivary gland	178	Testis
150	Esophagus	179	Other male genitals
151	Stomach	180	Kidney
152	Small intestine	181	Urinary bladder
153, 154 (except 1541)	Colorectum	190	Melanoma
1541	Anus	191	Skin (squamous cell carcinoma)
155, 156	Liver	192	Eye
157	Pancreas	193	Nervous system
160	Nose	194	Thyroid gland
162, 163	Lung	195	Endocrine glands
170	Breast	196	Bone
171	Cervix	197	Connective tissue
172	Endometrium	200, 202	Non-Hodgkin lymphoma
173	Uterus	201	Hodgkin lymphoma
175	Ovary	203	Multiple myeloma
176	Other female genitals	204-209	Leukemia
		199	CUP

All registered cancer cases were mostly histologically verified. Although the SFCD does not publish statistics on histological verification of all first and subsequent primary cancers separately, all are reported to have been included with primary cancers for which histological verification has been around 98% from the 1970s (CentreforEpidemiology, 2013). Ad hoc study on the diagnostic accuracy of second malignancies found 98% to be correctly classified (Frodin *et al.*, 1997).

Information on causes of death were available as the SFCD is also linked to the national causes of death register and the death certificate notification database (Ji *et al.*, 2012). Causes of death are annotated with gradually updated ICD over the years in the following manner, ICD-7 (1958 –1968), ICD-8 (1969 – 1986), ICD-9 (1987 - 1996) and with ICD-10 (1997 onwards).

8.2 Study population and parameters

Out of the total 16.1 million individuals present in the database, in order to create family pedigree-based analysis people belonging to the offspring generation were only considered as index individuals. The registry assigns proxy identifier to each person registered and tags biological parents of each offspring with the same identifier corresponded to the offspring. Hence people belonging to the offspring generation can be linked as biological siblings with the combination of their parents' identifiers (unless missing). A biological family pedigree can be constructed in this way (**Figure 8.1**).

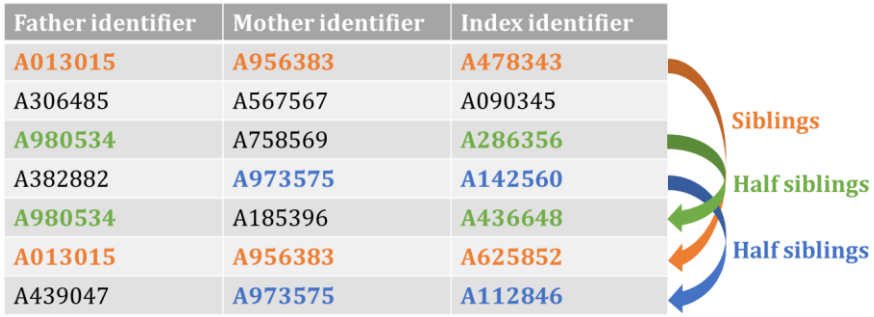


Figure 8. 1| Family identification thematized in SFCD
Siblings identified via same identifiers in both parents. Half siblings tracked via same identifier in one of the parents.

25,787 individuals were diagnosed with MM since 1958 till 2015 in Sweden out of whom 5,205 belong to the offspring generation. Among them 360 went on to have developed a second primary cancer with 4 years of median time of follow-up. Evidence of prior history of cancer in family was assessed and 246 among these 360 were found to have at least one first degree relative (either parents or siblings) with cancer.

Follow-up began from commencement of registry in 1958, with year of birth, immigration or diagnosis of MM whichever was later and was terminated in 2015 (end of registry period), on death, emigration or diagnosis of a second cancer, whichever occurred the

earliest. For further calculations regarding incidence, person years and contributing individuals were counted separately for case (people with MM diagnosis) and reference population (people without cancer) stratified over age-group (5 year age bands), calendar period (10 year bands), residential area, occupation (proxy for socio-economic status) and existence of cancer family history (binary identifier).

8.3 Familial relative risk estimation

For last couple of decades the incidence rate ratio obtained from linear regression has been a standard measure of relative risk estimation that quantifies excess risk between two populations by a calculating measure of multiplicative difference; magnitude of which lies on non-negative real line (Prentice, 1985). This method leverages generalized linear model to estimate standardized incidence risk with underlying distributional assumption of Poisson point process (inverse generalized waiting time distribution) (Nelder and Wedderburn, 1972). In short, *diagnoses of cases are* presumed to follow a waiting time distribution over the years making the *observed frequency of cases* to follow a Poisson process. As the variance of the response count variable is to be unbiasedly estimated with explanatory covariates, the regression takes the following shape:

$$f(\tilde{Y}) = X\tilde{\beta}$$

Here f is the (link) function that relates the expected value of the random variable to the linear predictor of the explanatory variables. As the underlying count process is supposed to follow a Poisson mass function, a log link of the expected occurrence odds takes the form:

$$logE\left(\frac{no.of\ cases}{contributing\ person\ years}\right)_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} \quad \ldots \qquad (1)$$

or, $\ln(\mu_j) = X\tilde{\beta}$ … (2)

Where μ_j is related to covariate matrix X via transformation with coefficients $\tilde{\beta}$ through natural logarithm link function. The estimated regression coefficients $\hat{\beta}$ are obtained by maximizing likelihood function.

Equation (2) can be extended to the following form:

$$V_{\hat{\mu}} = E(Y_j - \mu)^2 = \sum_{j=1\ to\ N}(Y_j - e^{X\hat{\beta}})^2 \ \ldots \quad (3)$$

This generalization of equidispersed variance function (as mean and variance of a Poisson variable is quantitatively equal) is then maximized first to obtain estimates of the coefficients and later to obtain confidence level. To this end, it can be extended to show:

$$\chi^2 \coloneqq \sum_{j=1\ to\ N}\frac{(Y_j - \hat{\mu}_j)}{V_{\hat{\mu}}} \quad \ldots \quad (4)$$

Statistical significance of this transformed statistics can be obtained with a test against a χ^2 distribution with degrees of freedom 1, against a pre-defined level (ex. 0.001). Following estimation of risk a test of trend was assessed with Kolmogorov-Smirnov statistic to note significance in difference in two series.

8.4 Interaction

Multiplicative interaction indexes (MIIs) and interaction contrast ratios (ICRs) were used to investigate the possible interaction between family history of a specific cancer and MM treatment (Zhang *et al.*, 2009). The formulations of interaction statistics were modified to incorporate relative risk in the context of the current study in the following way:

$$ICR = RR_{familial\ SPC_X} - RR_{cancer_X} - RR_{familial\ cancer_X} + 1$$

$$MII = \frac{RR_{familial\ SPC_X}}{RR_{cancer_X} \times RR_{familial\ cancer_X}}$$

Here $RR_{familial\ SPC_X}$is relative risk of SPC X in MM patients with history of cancer X in family; $RR_{cancer_X}$is relative risk of cancer X in general population and $RR_{familial\ cancer_X}$ is relative risk of cancer X among people with cancer X in family. $ICR > 0$ signifies a more than positive interaction and $MII > 1$ signifies greater than multiplicative interaction. Confidence intervals and *P*-values for MII and ICR were assessed with bootstrapping based on 100,000 replications on each sample and tested empirically.

Results

Chapter 9: Heritable risk in MGUS

9.1 Genetic interaction

9.1.1 Discovery set: Case-control design

The fast epistasis function provided in PLINK was Initially employed to explore genome-wide interaction on the discovery cohort of 243 German individuals diagnosed with MGUS. Among the 489,555 genotyped quality-controlled SNPs, a brute search algorithm required 1.2×10^{11} tests at a whole-genome scale to perform multivariate log-linear interaction test. PLINK operates without a prior selection criterion, hence the multiple testing adjusted level of significance to retain FWER at 5% was found at close to 4.2×10^{-13}. None of the signals thus discovered could attain this genome-wide threshold of significance (**Table 9.1**).

Table 9. 1| Top interactions from simple logistic linear interaction test (brute force epistasis with PLINK).

Gene1	Chr1	SNP1	Gene2	Chr2	SNP2	*P*-value
PRELID2	5	rs6580355	**LOC390299**	12	rs33233	3.57E-11
OR6B3	2	rs12471071	**CDC37L1**	9	rs1385453	8.67E-11
C1orf129	1	rs4656817	**hCG_1820717**	13	rs1326701	1.62E-10
ABHD5	3	rs4682696	**PPFIBP1**	12	rs7958124	2.18E-10
OR6B3	2	rs12471071	**CDC37L1**	9	rs6476893	2.72E-10
LOC730216	7	rs292661	**CSMD1**	8	rs4875730	4.37E-10
SLC13A1	7	rs4731094	**LOC390829**	18	rs2077149	6.39E-10
MYEOV2	2	rs1992307	**CDC37L1**	9	rs1385453	6.94E-10
IHPK3	6	rs568901	**TRPC6**	11	rs1230960	7.61E-10
ABHD5	3	rs4682696	**PPFIBP1**	12	rs7966058	8.04E-10
LOC729108	3	rs7652856	**LOC727862**	17	rs8074928	8.97E-10
MYEOV2	2	rs1992307	**CDC37L1**	9	rs6476893	1.52E-09
PI15	8	rs2731995	**CDKN3**	14	rs4898835	1.84E-09
CRISPLD1	8	rs2954870	**CDKN3**	14	rs4898835	2.17E-09
EFR3B	2	rs7575363	**MACROD2**	20	rs716316	3.24E-09
AGTPBP1	9	rs11141010	**ST8SIA2**	15	rs11632278	4.17E-09
BARX1	9	rs3996253	**LOC347292**	9	rs1885968	4.19E-09
MAN2A1	5	rs185088	**XRCC1**	19	rs3213356	4.43E-09
MAN2A1	5	rs185088	**ZNF575**	19	rs2030404	4.82E-09
LOC645521	2	rs13383210	**tcag7.893**	7	rs12672973	7.37E-09
GJB5	1	rs4653061	**OLFML2B**	1	rs2490431	1.18E-08
ACVR2A	2	rs12691767	**TRBV20OR9-2**	9	rs855508	1.82E-08
BARX1	9	rs4344139	**LOC347292**	9	rs1885968	2.12E-08
MCAM	11	rs2249466	**CCL23**	17	rs854666	3.97E-08
GRIA2	4	rs17246641	**ZFAND3**	6	rs6933547	6.99E-08
HABP2	10	rs4918844	**OTOR**	20	rs4814551	1.05E-07

Abbreviations:
SNP1 and SNP2 are the two SNP candidates of a pair from the discovery population belonging to chromosomes denote by Chr1 and Chr2; gene1 and gene2 are the corresponding genes annotated to SNP1 and SNP2, respectively.

Later the discovery analyses were expanded throughout three different computational platforms (**Table 9.2**).

Table 9. 2| Overview of tools and different subsequent protocols in use. Study designs enlist three stages of analysis.

Tool in use	Statistic used	Statistical model in use	Selection criteria	Study design	No. of tests performed	Bonferroni adjusted genome-wide level of significance (<1% FWER)	No. of risk loci pairs discovered
PLINK	Chi-square	log-linear model	n.a.	Discovery study	$1.2x10^{11}$	$4.2x10^{-13}$	none
CASSI	Wellek-Ziegler statistic	Logistic regression; Fixed effects weighted model	Single marker test P value < 10^{-3}	Discovery study	$2.8x10^{7}$	$5x10^{-10}$	561
				Follow up study	$4.4x10^{5}$	$5x10^{-10}$	352
				Replication study	$8.2x10^{6}$	$5x10^{-10}$	23
INTERSNP	Chi-square statistic	Full log-linear model	Top 5000 variants of single marker test	Discovery study	$1.25x10^{7}$	$8x10^{-11}$	none

Abbreviations:
FWER, family wise error rate

CASSI was later employed to explore curated genome-wide interaction on the same discovery set. It included default selection criteria (**Table 9.2**) that restricted the runs to approximately 2.8×10^7 overall tests at the system-defined single marker association test threshold of $P = 10^{-3}$. As approximately 2.8×10^7 tests were performed, Bonferroni corrected level of global threshold of significance was determined to be 5.0×10^{-10} which restricts the family-wise error rate (FWER) at less than 1% (Table 1). At this level CASSI reported 561 significant variant pairs with overall concordance with the PLINK registered results (**Table 9.3**).

Table 9. 3| Top interactions from simple logistic linear interaction test (brute force epistasis with PLINK) and its concordance with CASSI detected signals

Gene1	Chr1	SNP1	Gene2	Chr2	SNP2	PLINK *P*-value	CASSI W-Z *P*-value
PRELID2	5	rs6580355	LOC390299	12	rs33233	3.57E-11	1.65E-12
OR6B3	2	rs12471071	CDC37L1	9	rs1385453	8.67E-11	4.19E-13
C1orf129	1	rs4656817	hCG_1820717	13	rs1326701	1.62E-10	6.67E-13
ABHD5	3	rs4682696	PPFIBP1	12	rs7958124	2.18E-10	1.25E-12
OR6B3	2	rs12471071	CDC37L1	9	rs6476893	2.72E-10	2.02E-12
LOC730216	7	rs292661	CSMD1	8	rs4875730	4.37E-10	1.21E-12
SLC13A1	7	rs4731094	LOC390829	18	rs2077149	6.39E-10	1.65E-12
MYEOV2	2	rs1992307	CDC37L1	9	rs1385453	6.94E-10	2.62E-12
IHPK3	6	rs568901	TRPC6	11	rs1230960	7.61E-10	7.22E-12
ABHD5	3	rs4682696	PPFIBP1	12	rs7966058	8.04E-10	1.18E-11
LOC729108	3	rs7652856	LOC727862	17	rs8074928	8.97E-10	6.63E-13
MYEOV2	2	rs1992307	CDC37L1	9	rs6476893	1.52E-09	9.37E-12
PI15	8	rs2731995	CDKN3	14	rs4898835	1.84E-09	3.93E-12
CRISPLD1	8	rs2954870	CDKN3	14	rs4898835	2.17E-09	3.68E-12
EFR3B	2	rs7575363	MACROD2	20	rs716316	3.24E-09	7.64E-12
AGTPBP1	9	rs11141010	ST8SIA2	15	rs11632278	4.17E-09	1.34E-11
BARX1	9	rs3996253	LOC347292	9	rs1885968	4.19E-09	1.32E-11
MAN2A1	5	rs185088	XRCC1	19	rs3213356	4.43E-09	9.95E-12
MAN2A1	5	rs185088	ZNF575	19	rs2030404	4.82E-09	1.10E-11
LOC645521	2	rs13383210	tcag7.893	7	rs12672973	7.37E-09	1.10E-11
GJB5	1	rs4653061	OLFML2B	1	rs2490431	1.18E-08	8.08E-12
ACVR2A	2	rs12691767	TRBV20OR9-2	9	rs855508	1.82E-08	1.18E-12
BARX1	9	rs4344139	LOC347292	9	rs1885968	2.12E-08	2.22E-12
MCAM	11	rs2249466	CCL23	17	rs854666	3.97E-08	3.31E-12
GRIA2	4	rs17246641	ZFAND3	6	rs6933547	6.99E-08	3.05E-12
HABP2	10	rs4918844	OTOR	20	rs4814551	1.05E-07	7.00E-12

Abbreviations:
Gene1 and Gene2 are the corresponding genes annotated to SNP1 and SNP2, respectively. SNP1 and SNP2 are the two SNP candidates of a pair from the discovery population belonging to chromosomes denote by Chr1 and Chr2; W-Z signifies Wellek-Ziegler test.

INTERSNP operated with a top-down selection criteria to have a defined subpopulation of variants for the test. It selected top 5,000 variants which resulted in 5,000 × 5,000 pair-wise tests. These 2.5×10^6 tests helped the effective level of significance to restrain at 8.0×10^{-11} to obtain 99% confidence. The top signals thus observed until the level of 10^{-7} are shown in **Table 9.4**.

Table 9. 4| Top interactions from simple logistic linear interaction test in INTERSNP subject to single-marker selection criteria.

SNP1	Chr1	Allele1	BP1	Single marker *P*-value	SNP2	Chr2	Allele2	BP2	Single marker *P*-value	Logistic model Interaction *P*-value	Standard error
rs10099120	8	T	85101510	3.17E-03	rs3738270	1	C	201195119	1.38E-02	9.0E-08	0.207567
rs6604717	1	A	223647907	5.85E-03	rs10236139	7	T	9667148	8.62E-03	1.9E-07	0.199375
rs10991722	9	A	93718266	5.98E-03	rs882937	11	A	93908872	8.52E-03	2.8E-07	0.139297
rs10266589	7	C	150182338	6.47E-04	rs2171529	3	C	46047032	2.38E-02	2.9E-07	0.24014
rs2651148	3	A	193561684	1.79E-02	rs6857709	4	A	65784245	2.44E-02	3.8E-07	0.435404
rs10991722	9	A	93718266	5.98E-03	rs1792634	11	A	93929510	1.90E-02	4.6E-07	0.140359
rs177040	9	A	93714388	6.28E-03	rs882937	11	A	93908872	8.52E-03	5.0E-07	0.139257
rs3853275	9	T	30443903	2.01E-02	rs6470796	8	T	131064299	2.03E-02	5.4E-07	0.18791
rs6976643	7	C	77841529	1.17E-02	rs10505385	8	A	121982031	1.93E-02	6.2E-07	0.235587
rs10986270	9	T	126967977	1.52E-02	rs269554	5	T	140324431	1.66E-02	7.0E-07	0.285327
rs177040	9	A	93714388	6.28E-03	rs1792634	11	A	93929510	1.90E-02	7.2E-07	0.139967
rs7201659	16	A	17273019	1.30E-02	rs1795734	4	T	89933630	2.36E-02	7.5E-07	0.154884
rs11002693	10	T	80584333	1.78E-02	rs11663706	18	A	11219792	2.42E-02	9.5E-07	0.273795
rs9472446	6	A	12734891	1.32E-02	rs8124695	20	A	39028436	2.40E-02	9.8E-07	0.372317
rs3853275	9	T	30443903	2.01E-02	rs7833007	8	T	131108193	2.27E-02	9.9E-07	0.186747

Abbreviations:
SNP1 and SNP2 are the two SNP candidates of a pair from the discovery population belonging to chromosomes denote by Chr1 and Chr2 represented by two corresponding effect alleles Allele1 and Allele2; BP1 and BP2 are the base pair position of the SNPs. Single marker *P*-values are calculated by INTERSNP to prune through and select the top SNPs for interaction test P-value for ehich is demonstrated in the penultimate column. Standard error is calculated as a measure of heterogeneity.

Subsequently, 693 unique interaction pairs were identified at 5.0×10^{-5} significance level where none of the observations reached genome-wide threshold of approximately 8.0×10^{-11}. The top interaction was found to be between rs10099120 and rs3738270 with $P = 9.0 \times 10^{-8}$ and conspicuously, rs10099120 is located in the intronic region of *RALYL* and rs3738270 corresponds to a missense mutation on *IGFN1*.

Contrastingly the top ranked overlapped interaction (rs12471071 [2q37] - rs1385453 [9p24]) from the discovery set had a CASSI Wellek-Ziegler (W-Z) $P = 4.2 \times 10^{-13}$ and a simple logistic regression PLINK-$P = 8.7 \times 10^{-11}$ (**Table 9.3**). Although the CASSI algorithm detected 561 common variant SNP pairs to be genome-wide significant, previous researches demonstrate that such findings were often subject to false discovery. Hence the investigation was extended with exhaustive search using other algorithms. Overall 52 common variant pairs were co-discovered with varied level of confidence for both INTERSNP and CASSI. Top signals from CASSI discovery analysis is summarized in **Table 9.5**.

Table 9. 5| Top interaction signals from logistic regression defined Wellek-Ziegler test using CASSI

Gene1	Chr1	SNP1	BP1	Gene2	Chr2	SNP2	BP2	W-Z *P*-value
FBXL17	5	rs1799011	107325759	MAGI2	7	rs967489	78529329	2.59E-36
LOC728394	4	rs11090644	92431729	FBLN1	22	rs7677659	45982997	1.52E-35
NELL1	11	rs11640925	21308199	A2BP1	16	rs7127622	8156523	1.73E-35
DTNBP1	6	rs10266202	15625808	NXPH1	7	rs2743868	8957041	1.13E-34
NULL	4	rs8181443	8541869	RAB11FIP2	10	rs6447879	119670998	3.07E-32
IRX1	5	rs12051446	3845683	A2BP1	16	rs9687393	7409031	5.21E-32
PLXDC2	10	rs2015847	20405138	CDRT4	17	rs2461941	15318855	7.52E-30
ERBB4	2	rs2144066	212974828	DIO3OS	14	rs17416172	101938855	4.42E-29
LOC646538	1	rs7159563	81177525	LOC730105	14	rs841666	82783075	2.93E-28
TSNARE1	8	rs10904319	143233312	LOC338588	10	rs10110636	4741842	3.03E-28
SOX11	2	rs214742	5361397	TMEM135	11	rs10181393	86985581	4.05E-28
LOC646538	1	rs12588076	81177525	LOC730105	14	rs841666	82746477	1.42E-27
TCERG1L	10	rs8045250	132963432	A2BP1	16	rs4751335	6879583	1.62E-27
LOC344371	2	rs2502294	34575895	RCADH5	6	rs7577875	67792729	1.69E-27
FHIT	3	rs909876	61177577	DHX35	20	rs7617424	38025235	1.78E-27
COL9A1	6	rs509333	71064295	MN1	22	rs7772055	27640747	2.71E-27
NKAIN2	6	rs10759037	124200765	PTPRD	9	rs9388287	9064330	4.52E-27
NSUN2	5	rs11070218	6599222	LOC644779	15	rs6876835	39823058	8.56E-27
RAP2B	3	rs10267303	152994772	AUTS2	7	rs7355869	70082913	9.61E-27
UHRF2	9	rs7983829	6459274	MYO16	13	rs524888	109736676	1.25E-26
ROBO2	3	rs6575656	77720924	C14orf177	14	rs9883373	98761422	1.49E-26
PTPRD	9	rs4755435	8227101	LDLRAD3	11	rs10976860	36136580	1.94E-26
RIMS1	6	rs4820127	72908366	LOC730062	22	rs1852702	34395171	2.74E-26
LOC390419	13	rs3859840	91448197	ISX	22	rs2152310	35359702	3.04E-26
LOC388474	18	rs134794	36232589	MN1	22	rs1540018	27668370	3.05E-26
HTR1B	6	rs12193281	78298165	SNAP91	6	rs2252216	84325184	5.75E-26

Abbreviations:
Gene1 and Gene2 are the corresponding genes annotated to SNP1 and SNP2, respectively. SNP1 and SNP2 are the two SNP candidates of a pair from the discovery population belonging to chromosomes denote by Chr1 and Chr2 and their chromosomal locations described in base pairs by BP1 and BP2; W-Z signifies Wellek-Ziegler test.

9.1.2 Confirmation set: Case-only design

As a confirmational study to the detection sensitivity provided by CASSI, a follow-up case-only analysis on the 82 cases rendered approximately 4.4×10^5 overall tests after initial single marker test shrinkage similar to that described above. The most significant interaction (rs4433825 [16p13] - rs2295179 [20p12]) showed a case-only $P = 3.3 \times 10^{-24}$ against W-Z case-control $P = 1.5 \times 10^{-12}$ with a statistically significant odds ratio of 2.05 with interactions between two unique sets of variants of *A2BP1* and *PLCB1*. At 5×10^{-10} level, 352 variant pairs replicated in the discovery set with varying levels of significance (**Table 9.6**). The magnitude of test *P*-values observed in the follow-up analysis is consistent with that of a case-only design that usually demonstrates higher detection power due to the inherent mathematical assumptions. One needs to be careful in interpretation of the results as an evidence of functional relation between variants because of the constraint due to restriction of analysis within SNP pairs found in overlap that were tested with the discovery set. As the pre-selection criteria is in place according to the single marker test which are performed with different designs in the two different set-ups, the observed overlaps are ensured to have higher individual fixed effects in case-only design if there is bias due to linkage, a caveat in this design. Whilst the case-only study does not accurately identify sentinel interacting pairs; nonetheless it confirmed viability of the case-control replication study in larger cohort as the highest signals were observed from overlapping chromosomal regions in interaction.

Table 9. 6| Overlapped top interactions from simple logistic linear interaction test in case-only and cases-control analysis in separate cohorts observed in CASSI

				Case-only analysis					Case-control analysis				
Gene1	**Gene2**	**Chr1**	**Chr2**	**SNP1**	**BP1**	**SNP2**	**BP2**	**Case-only *P*-value**	**SNP1**	**BP1**	**SNP2**	**BP2**	**Case-control *P*-value**
A2BP1	PLCB1	16	20	rs4433825	7378192	rs2295179	8678446	3.35E-24	rs11860241	7294639	rs6055995	8662222	1.50E-12
LOC644624	RUNX1	4	21	rs12509315	124592367	rs2242901	36456189	7.25E-20	rs1433218	124707000	rs2834757	36467070	2.90E-10
CNTNAP2	DDHD1	7	14	rs1496547	146923861	rs1959843	53859798	3.43E-19	rs3194	148114265	rs1954308	53940700	7.07E-12
CMYA5	ASTN2	5	9	rs259103	79094685	rs4240422	120163466	6.17E-17	rs13159668	79047407	rs10481683	119803933	1.43E-10
BUB3	DSEL	10	18	rs6599673	125008718	rs12966710	65286267	8.16E-17	rs6599689	125106194	rs9319727	64967509	1.00E-11
RAPGEF2	CSMD1	4	8	rs3846243	161292814	rs4875703	3224561	1.23E-16	rs4591547	160357877	rs11136609	3214873	3.55E-12
ERC2	SOX2OT	3	3	rs1795648	55571760	rs9845058	181342415	5.29E-15	rs9878600	56017598	rs4855056	181638250	4.65E-11
GBE1	LOC729993	3	16	rs9877327	81643583	rs4781415	13217905	1.46E-14	rs7618878	83089596	rs8056716	13643974	1.60E-10
LOC391470	MAGI2	2	7	rs7566549	196442418	rs2714676	79293845	3.17E-14	rs1849068	195967027	rs1829989	77830751	1.83E-10
CNTNAP5	CAMK1D	2	10	rs12616423	124279775	rs3802570	12805430	9.70E-14	rs1543901	125189434	rs7897059	12591712	4.13E-10
NRXN1	A2BP1	2	16	rs2194388	50871821	rs7203146	7597995	1.11E-13	rs9679539	51065644	rs1424125	6062553	6.29E-12
IQGAP2	CNTN5	5	11	rs4704346	75908982	rs7947488	99198902	1.33E-13	rs4235701	75964280	rs7947002	98626907	3.94E-10
CTNND2	LOC283584	5	14	rs26001	11303908	rs8016687	86636646	1.42E-13	rs31897	11428523	rs12436912	86821763	3.22E-11
KLHL29	CNTN5	2	11	rs7598792	23444263	rs2514231	98699317	9.23E-12	rs1653763	23724991	rs1815913	97852498	3.23E-10
DNM3	ADRA2A	1	10	rs9425291	172312769	rs1537769	112871088	1.42E-11	rs633995	172186729	rs7098615	113139169	5.45E-11

Abbreviations:

Results are described in two different panels for case-only and case-control designed analysis respectively. SNP1 and SNP2 are the two SNP candidates of a pair from the discovery population belonging to chromosomes denote by Chr1 and Chr2; BP1 and BP2 are the base pair position of the SNPs.

9.1.3 Replication set: Case-control design

Signals from discovery sets were compared against a replication set to obtain culminating evidence. Replication set was evaluated with W-Z interactions ordained by CASSI in the consisting 8.2×10^6 test pairs with an inflation factor of 1.02. This set was able to replicate 23 out of all 561 genome-wide significant variant pairs of the discovery set which are annotated to same chromosomal regions with varying degrees of significance (**Table 9.7**, **Figure 9.1**). The top interaction was found among variants annotated to *TNC* and *CRYL1* corresponding to 9q33.1 and 13q12.11 (rs10118040 – rs7337130, W-Z $P = 6.9 \times 10^{-11}$ and rs1330368 – rs7337231, W-Z $P = 2.4 \times 10^{-8}$ respectively). Among the 23 replications, 14 were unique regions and there were 5 regions with multiple unique interactions. Interestingly, *SETBP1* and *PREX1* interaction at 18q12.3 and 20q13.13 were represented by 6 SNP-SNP pairwise interactions with LD coefficient of $r^2 < 0.2$ between SNPs belonging to each of the corresponding regions. The 20q13.13 locus is speculated as a risk predisposing locus for MM and as an expression and methylation QTL at *PREX1* without having any direct impact on an active promoter site (Mitchell *et al.*, 2016). SETBP1 on the other hand has been reported to harbor somatic mutations and play a role in oncogenesis. It is often found harboring aberrations in various myeloid malignancies including secondary acute myeloid leukemia (sAML) and chronic myelomonocytic leukemia (CMML) although germ-line mutations are reported (Makishima *et al.*, 2013). The first GWAS on MGUS reported 10 sentinel variants with moderately strong signals, among those two SNPs were yet again identified with moderate interactions: rs10251201 (7p21.3, *GLCCI1*) with rs1104869 (2p23.2-p23.1,

ALK), W-Z $P = 8.7 \times 10^{-7}$ and rs16966921 (18q12.2, *GALNT1*) with rs8092870 (18q12.1, *CDH2*), W-Z $P = 1.7 \times 10^{-7}$.

Table 9. 7| Overlapped top interactions from simple logistic linear interaction test in discovery cases-control analysis found replicated in replication cohort observed in CASSI

				Discovery set								Replication set						
Gene1	**Chr1**	**Gene2**	**Chr2**	**SNP1 (Risk allele)**	**Position (hg19,bp)**	**MAF**	**SNP2 (Risk allele)**	**Position (hg19,bp)**	**MAF**	**WZ P value**	**OR (95% CI)**	**SNP1 (Risk allele)**	**Position (hg19,bp)**	**MAF**	**SNP2 (Risk allele)**	**Position (hg19,bp)**	**MAF**	**WZ P value**	**OR (95% CI)**
TNC	9q33.1	CRYL1	13q12.11	rs10118040 (T)	117879414	0.40	rs7337130 (C)	21021343	0.31	**6.91E-11**	2.64 (1.91-3.65)	rs1330368 (A)	117821026	0.48	rs7337231 (G)	20896618	0.49	2.48E-08	1.05 (0.96 – 1.14)
SETBP1	18q12.3	PREX1	20q13.13	rs12959213 (C)	42769020	0.41	rs6066791 (T)	47251687	0.26	**7.07E-11**	2.39 (1.75-3.25)	rs11082429 (G)	42743790	0.44	rs170536 (A)	46878722	0.32	4.25E-08	1.01 (0.93 – 1.09)
SETBP1	18q12.3	PREX1	20q13.13	rs12959213 (C)	42769020	0.41	rs6066791 (T)	47251687	0.26	**7.07E-11**	2.39 (1.75-3.25)	rs1376230 (T)	42703052	0.35	rs6063251 (C)	47015157	0.43	6.37E-07	1.03 (0.94 – 1.11)
ERBB4	2q34	RORA	15q22.2	rs1546717 (G)	212902339	0.10	rs1159814 (A)	61431996	0.41	**9.07E-11**	5.03 (2.89-8.77)	rs6745249 (G)	213130571	0.48	rs974065 (A)	60952440	0.34	**1.06E-10**	1.13 (1.04 – 1.22)
PARK2	6q26	C14orf177	14q32.2	rs6455744 (T)	162060468	0.38	rs7359146 (C)	99084602	0.14	**1.12E-10**	2.92 (1.96-4.33)	rs6927285 (G)	162010329	0.43	rs8022922 (A)	98987292	0.44	**1.23E-14**	1.06 (0.98 – 1.14)
ETNK1	12p12.1	TMC2	20p13	rs2467112 (C)	23071644	0.19	rs1028441 (T)	2600186	0.24	**1.20E-10**	3.24 (2.12-4.95)	rs7313039 (C)	23091130	0.47	rs6050256 (T)	2554907	0.48	2.04E-07	1.05 (0.97 – 1.14)
HFM1	1p22.2	LOC647259	13q21.1	rs674135 (G)	91675675	0.26	rs4146191 (A)	62872965	0.47	**1.44E-10**	2.61 (1.89-3.60)	rs7416823 (T)	157386394	0.31	rs428328 (C)	63110606	0.41	2.24E-09	1.05 (0.96 – 1.13)
ERBB4	2q34	PTPRD	9p23	rs1437919 (A)	212110840	0.23	rs10978043 (G)	9860402	0.19	**2.64E-10**	3.38 (2.19-5.22)	rs6747637 (G)	212406789	0.45	rs4427223 (A)	10663815	0.48	**7.35E-14**	1.01 (0.93 – 1.09)
AUTS2	7p11.22	HS6ST3	13q32.1	rs1011780 (A)	70124648	0.28	rs9556582 (G)	97040531	0.46	**2.68E-10**	2.40 (1.75-3.29)	rs10267303 (T)	70082913	0.47	rs12876541 (C)	97304003	0.44	3.33E-08	1.06 (0.97 – 1.14)
SETBP1	18q12.3	PREX1	20q13.13	rs12959213 (C)	42769020	0.41	rs4810836 (T)	47228931	0.25	**3.04E-10**	2.40 (1.75-3.25)	rs11082429 (G)	42743790	0.44	rs170536 (A)	46878722	0.32	4.25E-08	0.98 (0.90 – 1.06)
SETBP1	18q12.3	PREX1	20q13.13	rs12959213 (C)	42769020	0.41	rs4810836 (T)	47228931	0.25	**3.04E-10**	2.40 (1.75-3.25)	rs1376230 (T)	42703052	0.35	rs6063251 (C)	47015157	0.43	6.37E-07	0.97 (0.89 – 1.05)
CNTN4	3p26.3	FAM19A1	3p14.1	rs2619566 (C)	2624938	0.12	rs1032376 (A)	68317975	0.19	**3.28E-10**	4.66 (2.71-8.02)	rs1499133 (C)	2952214	0.41	rs7610023 (T)	68123731	0.40	4.14E-09	1.05 (0.97 – 1.14)
CNTN4	3p26.3	FAM19A1	3p14.1	rs2619566 (G)	2624938	0.12	rs1032376 (A)	68317975	0.19	**3.28E-10**	4.66 (2.71-8.02)	rs1178491 (G)	2342825	0.36	rs6549098 (A)	68323280	0.40	2.83E-08	0.98 (0.90 – 1.06)
TNC	9q33.1	CRYL1	9q33.1	rs2071520 (T)	117880792	0.32	rs7337130 (C)	21021343	0.31	**3.50E-10**	2.80 (1.99-3.15)	rs1330368 (A)	117821026	0.48	rs7337231 (G)	20896618	0.49	2.48E-08	0.96 (0.89 – 1.04)
CSMD1	8p23.2	LOC392301	9q13	rs1700112 (G)	4097418	0.41	rs410684 (A)	31673588	0.42	**3.84E-10**	2.16 (1.62-2.87)	rs2740939 (C)	3872513	0.48	rs7853053 (T)	32211402	0.49	**2.04E-16**	1.04 (0.95 – 1.12)

Table 9.7 continued| Overlapped top interactions from simple logistic linear interaction test in discovery cases-control analysis found replicated in replication cohort observed in CASSI

				Discovery set								Replication set						
Gene1	**Chr1**	**Gene2**	**Chr2**	**SNP1 (Risk allele)**	**Position (hg19,bp)**	**MAF**	**SNP2 (Risk allele)**	**Position (hg19,bp)**	**MAF**	**WZ P value**	**OR (95% CI)**	**SNP1 (Risk allele)**	**Position (hg19,bp)**	**MAF**	**SNP2 (Risk allele)**	**Position (hg19,bp)**	**MAF**	**WZ P value**	**OR (95% CI)**
CSMD1	8p23.2	LOC392301	9q13	rs1700112 (G)	4097418	0.41	rs410684 (A)	31673588	0.42	**3.84E-10**	2.16 (1.62-2.87)	rs2740929 (C)	3879918	0.49	rs7853053 (T)	32211402	0.49	**3.81E-10**	1.04 (0.95 – 1.12)
ERBB4	2q34	PTPRD	9p23	rs1437919 (A)	212110840	0.23	rs7851513 (G)	9842176	0.19	**3.92E-10**	3.10 (2.05-4.69)	rs6747637 (G)	212406789	0.45	rs4427223 (A)	10663815	0.48	**7.35E-14**	0.92 (0.85 – 0.99)
KHDRBS3	8q24.23	KSR2	12q24.23	rs4909494 (C)	136646548	0.46	rs10774941 (T)	118037655	0.27	**4.22E-10**	2.51 (1.83-3.45)	rs16905387 (G)	136539132	0.42	rs7972142 (A)	118211046	0.44	**3.97E-13**	1.05 (0.96 – 1.13)
SETBP1	18q12.3	PREX1	20q13.13	rs12959213 (C)	42769020	0.41	rs6095212 (T)	47233383	0.25	**4.25E-10**	2.39 (1.75-3.25)	rs11082429 (G)	42743790	0.44	rs170536 (A)	46878722	0.32	4.25E-08	1.04 (0.96 – 1.12)
SETBP1	18q12.3	PREX1	20q13.13	rs12959213 (C)	42769020	0.41	rs6095212 (T)	47233383	0.25	**4.25E-10**	2.39 (1.75-3.25)	rs1376230 (T)	42703052	0.35	rs6063251 (C)	47015157	0.43	6.37E-07	1.03 (0.95 – 1.11)
MAN1A1	6q22.31	FRMD4A	10p13	rs808034 (A)	119467743	0.39	rs789761 (C)	14137678	0.48	**4.72E-10**	0.46 (0.34-0.61)	rs1295392 (G)	119676177	0.45	rs751498 (A)	13929130	0.47	1.25E-09	1.05 (0.96 – 1.14)
BNC2	9p22.3	CDH13	16q23.3	rs7867771 (T)	16314909	0.28	rs11149564 (C)	83441027	0.44	**4.82E-10**	2.20 (1.62-3.00)	rs1415471 (A)	16656653	0.44	rs7194615 (G)	82769498	0.44	1.07E-08	1.02 (0.94 – 1.09)
DAOA	13q33.2	TOM1L1	17q22	rs5012127 (G)	105119100	0.17	rs4793773 (A)	52646414	0.27	**4.89E-10**	2.81 (1.88-4.02)	rs3015345 (A)	105860621	0.45	rs8070668 (G)	52991636	0.38	6.06E-10	1.05 (0.97 – 1.14)

Risk allele is the allele corresponding to which the test is performed and the odds ratio is calculated. Frequency of risk allele pair is tested against controls. SNP1 and SNP2 are the two SNP candidates of a pair from each population; gene1 and gene2 are the corresponding genes annotated to SNP1 and SNP2, respectively. WZ p-value is Wellek Ziegler test p-value. OR, interaction odds ratio; MAF, minor allele frequency; bp, base pair; hg19, human genome build 19; CI, confidence interval.

Bolding indicate genome wide significant observation at 99% level of significance. Underline indicate significant odds ratio.

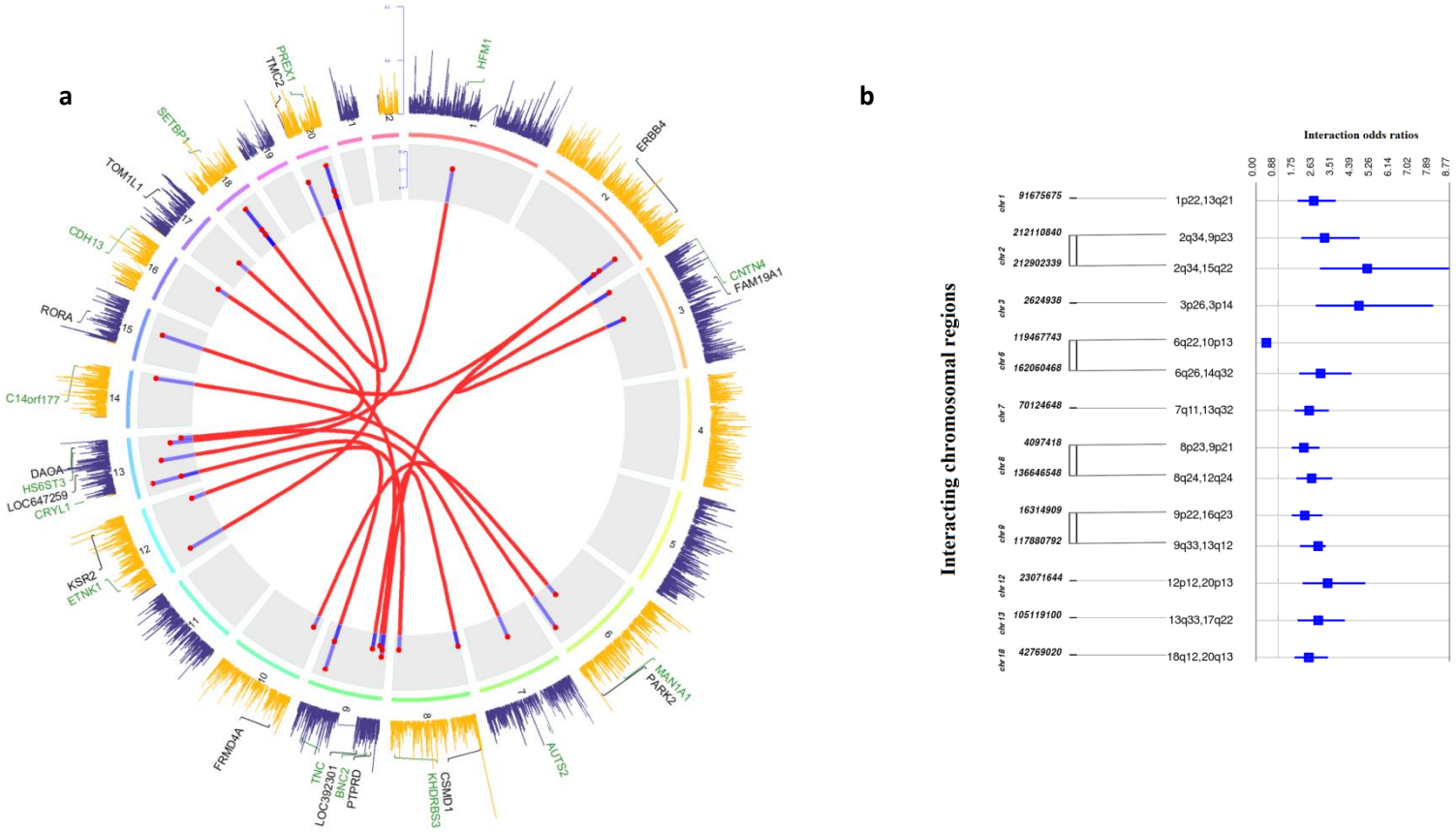


Figure 9. 1| Interaction Analysis identifies unique risk loci pairs.

a, Circos plot of genome-wide association and significant interaction results for the identified paired risk loci. The second outer most panel displays results from genome-wide association study on a Manhattan plot for autosomal variants on a log transformed scale (-log). Negative log transformed interaction *P*-values corresponding to each of the interaction pair is calculated from log linear transformed regression on the discovery set and is represented on an adjusted inflated scale of 9.3 to 10.2 in the second inner most panel. More than one unique variant pair combinations are present in the same interacting regions which are marked with their corresponding odds in this panel. Genome-wide significant paired loci are line-joined in the inner most panel based on their chromosomal positions (NCBI build 19 human genome). Annotations of single nucleotide polymorphisms to gene ids are displayed at the outer most panels.

b, Forest plot with embedded confidence intervals for each of the identified interaction pairs. Each pair indicates two interacting chromosomal locations with base pair information for the indexing loci. Paired variants annotated to the same indexing chromosomes are line joined

Abbreviations: chr, chromosome; BP, base pair; OR, odds ratio; CI, confidence interval.

9.2 Genetic interaction-based network

A partnership dependence structure of functional network was constructed with the risk variants from the final overlapping set subject to identifiable annotation from the interaction analyses. 26 such reconstituted genes were used as nodes which together with first order interacting genes created scaffolding for further enrichment analysis (**Figure 9.2**). 36 potentially differentially regulated pathways were identified (**Table 9.8**). Among them were 18 enriched pathways at 0.01 level of significance, with as many as 5 gene nodes downstream to KEGG ErbB signaling pathway ($P = 7.1 \times 10^{-5}$) and 3 gene nodes downstream to KEGG B cell receptor signaling pathway ($P = 5.3 \times 10^{-3}$) were found to be the two most significant pathways.

Table 9. 8| Gene set enrichment analysis in genetic network with STRING.

Pathway	Gene count	*P*-value	Pathway	Gene count	*P*-value
ErbB signaling pathway	5	7.09E-05	Neurotrophin signaling pathway	3	1.32E-02
B cell receptor signaling pathway	3	5.32E-03	Thyroid cancer	2	1.32E-02
Fc epsilon RI signaling pathway	3	5.32E-03	FoxO signaling pathway	3	1.38E-02
Prolactin signaling pathway	3	5.32E-03	Natural killer cell mediated cytotoxicity	3	1.38E-02
MicroRNAs in cancer	4	5.32E-03	Circadian rhythm	2	1.38E-02
Renal cell carcinoma	3	5.32E-03	Insulin signaling pathway	3	1.62E-02
Endometrial cancer	3	5.32E-03	Parkinson s disease	3	1.73E-02
Glioma	3	5.32E-03	Prion diseases	2	1.73E-02
Chronic myeloid leukemia	3	5.32E-03	Hepatitis B	3	1.73E-02
Acute myeloid leukemia	3	5.32E-03	Bladder cancer	2	1.76E-02
Non-small cell lung cancer	3	5.32E-03	PI3K-Akt signaling pathway	4	2.22E-02
Gap junction	3	8.29E-03	Chemokine signaling pathway	3	3.04E-02
GnRH signaling pathway	3	8.29E-03	Long-term depression	2	3.87E-02
Prostate cancer	3	8.29E-03	VEGF signaling pathway	2	3.88E-02
Proteoglycans in cancer	4	8.66E-03	Focal adhesion	3	3.94E-02
Estrogen signaling pathway	3	8.82E-03	Long-term potentiation	2	4.15E-02
Dorso-ventral axis formation	2	9.67E-03	Ras signaling pathway	3	4.65E-02
T cell receptor signaling pathway	3	9.67E-03	Melanoma	2	4.68E-02

Based on the indexing nodes and the additional predicted first order interacting nodes, STRING performs enrichment analysis on several molecular, biological, cellular process related pathway analysis with Gene Onltology (GO) and KEGG database.
All tests are performed with guilt by association assumption and P values are corrected for multiple testing. A protein-protein enrichment index is reported with analysis depicting level of confidence in the detected enriched processes which is reported to be 0.0039 (significant at 5% level)

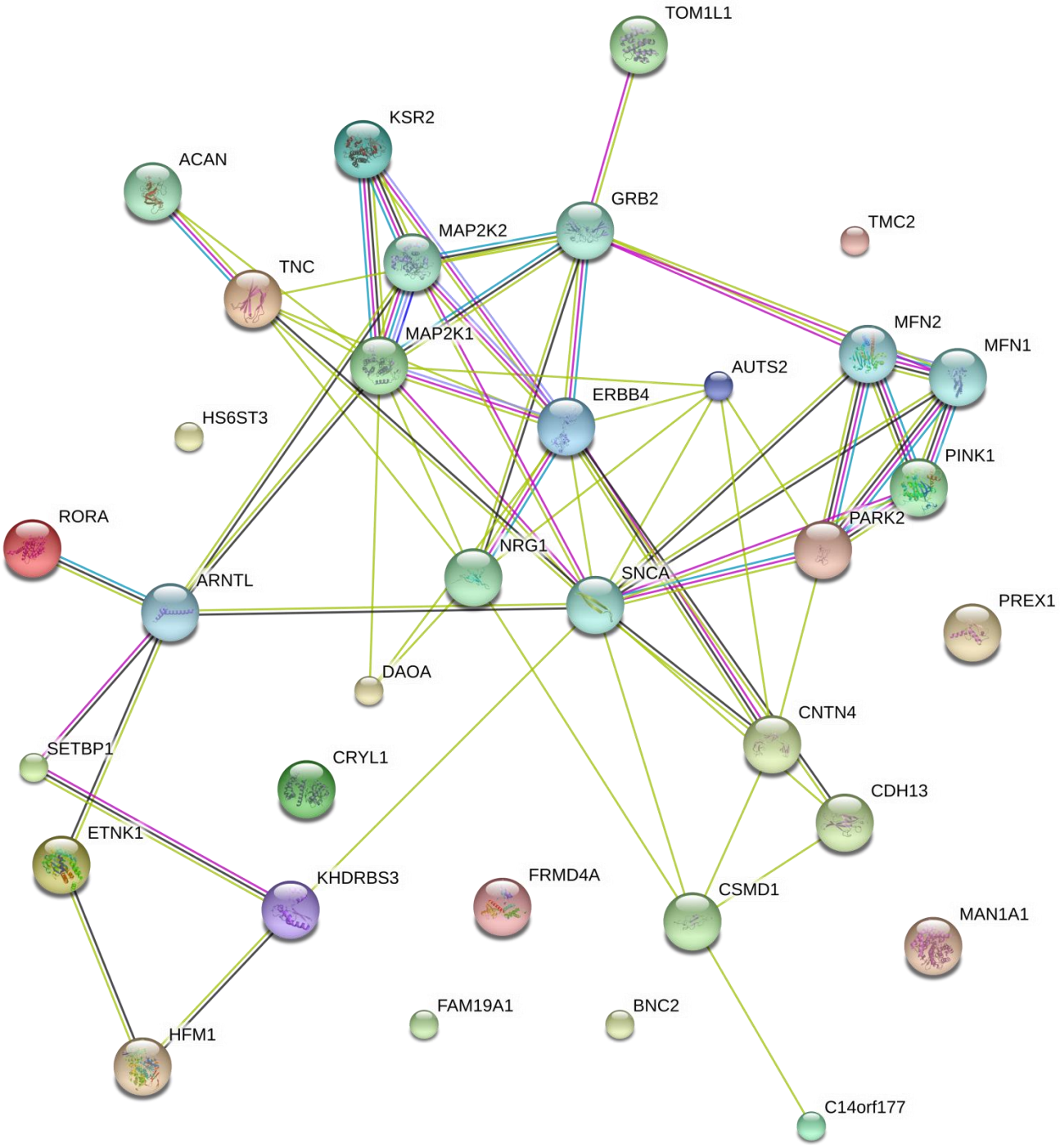


Figure 9. 2| Genetic interaction network constructed with STRING.
A network of 26 identified genes annotated to risk loci with added predicted genes in interaction. All nodes represent first order interaction. Colored edges convey status of predicted network edge correspondingly cyan, curated database; magenta, experimentally determined; forest green, gene neighborhood; red, gene fusion; navy blue, gene co-occurrence; lawn green, text mining; black, co-expression; lavender indigo, protein homology. Node color signifies protein functionality. Additional nodes are considered based on prediction score ≥ 0.9 (for more details, refer to STRING data base).

9.3 Pathway analysis

Gene-set enrichment and subsequent pathway analysis was interrogated with three different *in silico* approaches. Employing MAGENTA on the discovery set GWAS summary results 111 functionally enriched pathways significant at 5% Bonferroni corrected level of significance (22 at 99%) were detected (**Table 9.9**). For confirmation of the results the same summary statistics were assessed with PASCAL and 65 enriched pathways were identified at 5% adjusted level of significance (19 at 1%, **Table 9.10**). Although 28 overlapping pathways between these two algorithms used were discovered at a combined simultaneous testing corrected $P < 2.5 \times 10^{-3}$ (**Table 9.11**), to extend the search and impart functional information, further enrichment analysis was carried out utilizing curated microarray data. Expression data-based gene set enrichment and pathway enrichment was carried out with DEPICT to have identified 99 pathways at a genome-wide suggestive threshold of 1.0×10^{-5}) and 4 at genome-wide threshold of $P = 5.0 \times 10^{-8}$ (**Table 9.12**). Whilst demonstrating varied significance throughout out different platforms, a combined pooled analysis with the summary statistics from all three algorithms, 9 pathways with multiple-test adjusted 5% levels of significance were identified (**Table 9.13**). Among the overlapping pathways, KEGG allograft rejection pathway (combined $P = 5.6 \times 10^{-4}$) and KEGG autoimmune thyroid disease pathway (combined $P = 9.3 \times 10^{-4}$), both downstream to B cell receptor signaling pathway, the most significant pathway detected interrogating interaction detected loci. EGFR downregulation, a signaling cascade upstream to ERBB signaling pathway was another observation with a moderate signal (combined $P = 2.4 \times 10^{-2}$). Thus, the B-cell receptor signaling pathway and EGFR regulatory network are found in both the interaction and GWAS-driven analyses implicating a role in MGUS.

Table 9. 9| MAGENTA gene set enrichment analysis results at 1% level of significance.

Data base	Pathway	*P*-Value
Ingenuity	T Cell Receptor Signaling	1.10E-03
REACTOME	CD28_DEPENDENT_VAV1_PATHWAY	1.30E-03
PANTHER_BIOLOGICAL_PROCESS	Blood_clotting	1.80E-03
PANTHER_BIOLOGICAL_PROCESS	Purine_metabolism	2.00E-03
GOTERM	nucleotide binding	2.50E-03
GOTERM	cilium assembly	3.20E-03
REACTOME	EGFR_DOWNREGULATION	3.20E-03
REACTOME	INTEGRIN_ALPHAIIBBETA3_SIGNALING	3.50E-03
GOTERM	odontogenesis	4.10E-03
GOTERM	small GTPase mediated signal transduction	4.60E-03
GOTERM	regulation of cell shape	4.70E-03
KEGG	KEGG_ALLOGRAFT_REJECTION	5.10E-03
GOTERM	lipopolysaccharide binding	5.40E-03
REACTOME	PLATELET_AGGREGATION_PLUG_FORMATION	5.50E-03
GOTERM	intracellular protein transport	6.30E-03
PANTHER_MOLECULAR_FUNCTION	Interleukin	7.20E-03
GOTERM	positive regulation of interleukin-8 production	7.50E-03
BIOCARTA	ASBCELL_PATHWAY	8.30E-03
PANTHER_MOLECULAR_FUNCTION	Non-receptor_tyrosine_protein_kinase	8.40E-03
BIOCARTA	DC_PATHWAY	8.40E-03
PANTHER_BIOLOGICAL_PROCESS	Cell_cycle	9.20E-03
PANTHER_BIOLOGICAL_PROCESS	Cytoskeletal_regulation_by_Rho_GTPase	1.00E-02

Table 9. 10| PASCAL gene set enrichment analysis results at 1% level of significance

Data base	Pathway	*P*-value
REACTOME	METABOLISM_OF_POLYAMINES	3.77E-05
REACTOME	THROMBIN_SIGNALLING_THROUGH_PROTEINASE_ACTIVATED_RECEPTORS_PARS	6.17E-04
REACTOME	GPCR_DOWNSTREAM_SIGNALING	8.98E-04
REACTOME	THROMBOXANE_SIGNALLING_THROUGH_TP_RECEPTOR	9.35E-04
REACTOME	G_ALPHA_Z_SIGNALLING_EVENTS	1.24E-03
REACTOME	KERATAN_SULFATE_BIOSYNTHESIS	1.31E-03
REACTOME	KERATAN_SULFATE_KERATIN_METABOLISM	2.38E-03
REACTOME	SIGNAL_AMPLIFICATION	2.62E-03
REACTOME	ADP_SIGNALLING_THROUGH_P2RY1	3.44E-03
REACTOME	G_ALPHA_I_SIGNALLING_EVENTS	4.07E-03
REACTOME	AMINE_LIGAND_BINDING_RECEPTORS	4.54E-03
BIOCARTA	TCRA_PATHWAY	4.76E-03
REACTOME	AQUAPORIN_MEDIATED_TRANSPORT	5.40E-03
REACTOME	PROSTACYCLIN_SIGNALLING_THROUGH_PROSTACYCLIN_RECEPTOR	5.87E-03
KEGG	KEGG_PARKINSONS_DISEASE	8.18E-03
REACTOME	G_ALPHA_Q_SIGNALLING_EVENTS	8.43E-03
KEGG	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_KERATAN_SULFATE	8.45E-03
REACTOME	GLUCAGON_TYPE_LIGAND_RECEPTORS	9.74E-03
REACTOME	GPCR_LIGAND_BINDING	9.85E-03

Table 9. 11| All detected pathways mutually discovered in both MAGENTA and PASCAL at a 5% level of combined significance.

		PASCAL	MAGENTA	
Data base	**Pathway**	**P value**	**P value**	**Combined P value**
REACTOME	CD28_DEPENDENT_VAV1_PATHWAY	4.55E-02	1.30E-03	5.92E-05
REACTOME	PLATELET_AGGREGATION_PLUG_FORMATION	4.40E-02	5.50E-03	2.42E-04
KEGG	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_KERATAN_SULFATE	8.45E-03	3.65E-02	3.08E-04
KEGG	KEGG_ALLOGRAFT_REJECTION	8.38E-02	5.10E-03	4.27E-04
REACTOME	G_PROTEIN_ACTIVATION	1.40E-02	3.74E-02	5.22E-04
KEGG	KEGG_TYPE_1_DIABETES_MELLITUS	4.48E-02	1.94E-02	8.69E-04
BIOCARTA	ASBCELL_PATHWAY	1.10E-01	8.30E-03	9.15E-04
KEGG	KEGG_AUTOIMMUNE_THYROID_DISEASE	4.32E-02	2.24E-02	9.68E-04
REACTOME	P130CAS_LINKAGE_TO_MAPK_SIGNALING_FOR_INTEGRINS	3.60E-02	3.23E-02	1.16E-03
REACTOME	INTEGRIN_CELL_SURFACE_INTERACTIONS	1.61E-01	1.11E-02	1.78E-03
REACTOME	EGFR_DOWNREGULATION	7.77E-01	3.20E-03	2.49E-03
BIOCARTA	DC_PATHWAY	3.50E-01	8.40E-03	2.94E-03
BIOCARTA	INTEGRIN_PATHWAY	8.65E-02	3.62E-02	3.13E-03
KEGG	KEGG_DORSO_VENTRAL_AXIS_FORMATION	1.52E-01	2.20E-02	3.34E-03
REACTOME	MRNA_3_END_PROCESSING	2.18E-01	1.90E-02	4.14E-03
REACTOME	TOLL_RECEPTOR_CASCADES	3.90E-01	1.18E-02	4.60E-03
BIOCARTA	TH1TH2_PATHWAY	1.20E-01	4.40E-02	5.26E-03
BIOCARTA	ACH_PATHWAY	1.56E-01	4.03E-02	6.28E-03
BIOCARTA	CTLA4_PATHWAY	2.03E-01	4.25E-02	8.64E-03
REACTOME	SEMA3A_PAK_DEPENDENT_AXON_REPULSION	2.96E-01	2.93E-02	8.66E-03
REACTOME	MAPK_TARGETS_NUCLEAR_EVENTS_MEDIATED_BY_MAP_KINASES	1.98E-01	4.52E-02	8.97E-03
REACTOME	FGFR_LIGAND_BINDING_AND_ACTIVATION	2.95E-01	3.10E-02	9.14E-03
REACTOME	SIGNALING_BY_ROBO_RECEPTOR	4.78E-01	1.92E-02	9.19E-03
REACTOME	CD28_CO_STIMULATION	1.91E-01	4.83E-02	9.25E-03
KEGG	KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUCTION	5.99E-01	2.14E-02	1.28E-02
BIOCARTA	CYTOKINE_PATHWAY	3.29E-01	4.55E-02	1.50E-02
REACTOME	CDT1_ASSOCIATION_WITH_THE_CDC6_ORC_ORIGIN_COMPLEX	6.42E-01	3.38E-02	2.17E-02
BIOCARTA	MAPK_PATHWAY	4.66E-01	4.67E-02	2.17E-02

Table 9. 12| DEPICT gene set prioritization analysis utilized enriched pathways at Bonferroni corrected genome wide 5% significance level.

Data base	Functional pathway	Chi-square *P*-value
ENSEMBLE	BCL2A1 subnetwork	7.09E-11
GO	T cell activation	1.87E-09
MP	decreased T cell apoptosis	2.61E-09
MP	abnormal pro-erythroblast morphology	9.89E-09

GO, Gene Ontology; MP, Mammalian Protein

Table 9. 13| Combined results of gene set enrichment analysis from MAGENTA, PASCAL and DEPICT.

		PASCAL	DEPICT	MAGENTA	
Data base	**Pathway**	***P*-value**	***P*-value**	***P*-value**	***Combined *P*-value**
KEGG	Allograft rejection	0.083	0.001	0.005	5.62E-04
KEGG	Autoimmune thyroid disease	0.043	0.001	0.022	9.30E-04
KEGG	Glycosaminoglycan biosynthesis keratan sulfate	0.008	0.171	0.036	9.89E-03
REACTOME	Platelet aggregation plug formation	0.044	0.952	0.005	2.28E-02
REACTOME	EGFR downregulation	0.776	0.951	0.003	2.45E-02
REACTOME	Integrin cell surface interactions	0.16	0.396	0.011	4.51E-02
KEGG	Dorso ventral axis formation	0.151	0.233	0.022	4.78E-02
REACTOME	P130CAS linkage to MAPK signaling for integrins	0.035	0.988	0.032	4.85E-02

Pathways are pooled from several repositories which are enlisted with data base. *P*-values from PASCAL, DEPICT and MAGENTA are corrected for multiple testing.
*Pooled p values are combined using empirical Brown's method assuming dependency across test hypotheses.

Chapter 10: Heritable risk in MM

10.1 Genetic interaction

Two separate cohorts of MM patients consisting 2,282 cases (and 5,197 controls obtained from WTCCC) from the UK and 1,717 cases from Heidelberg, Germany (and 2,069 HNR controls) were subjected to interaction analysis each consisting of genotype data on approximately 430,000 and 520,000 common SNPs respectively. The W-Z statistics was employed as described previously. Subsequent meta-analysis of the linear interaction summary statistics subject to controlling for variance in each set rendered 16 unique SNP pairs belonging to 16 exclusive chromosomal regions that attained genome-wide significant threshold of 5.0×10^{-10} (**Table 10.1**, **Figure 10.1**).

The strongest meta-analyzed signal was observed for an interaction between rs7048811 at 9q21.31 (associated gene *GNAQ*) and rs7204305 at 16q24.1 (*IRF8*) (OR-Meta = 1.22; 95% CI = 1.12 – 1.32; $P = 1.3 \times 10^{-10}$). This interaction was consistent in both cohorts with a conservative level of significance (UK cohort: OR= 1.20, 95% CI = 1.08 – 1.33, $P = 7.0 \times 10^{-6}$; German cohort: OR = 1.24, 95% CI = 1.09 – 1.41, $P = 7.6 \times 10^{-6}$). The highest statistically significant effect size was observed for the second most strong interaction signal between rs2167453 at 11p15.2 (*PDE3B*) and rs2734459 at 19q13.31 (*ZNF229*) (OR-Meta = 1.52, 95% CI = 1.33 – 1.73, $P = 1.3 \times 10^{-10}$).

Table 10. 1| Genome-wide interaction analysis of the UK and the German MM samples and their meta-analysis.

								UK samples		German samples		Meta-analysis	
Gene1	**SNP1 (Risk allele)**	**Chr1**	**Position (hg19,bp**	**Gene2**	**SNP2 (Risk allele)**	**Chr2**	**Position (hg19,bp**	**OR (95% CI)**	P_{UK}	**OR (95% CI)**	P_{German}	**OR (95% CI)**	$P_{Meta\text{-}analysis}$
GNAQ	rs7048811 (G)	9q21.2	80469747	IRF8	rs7204305 (A)	16q24.1	86068776	1.20 (1.08 - 1.33)	7.0E-06	1.24 (1.09 - 1.41)	7.6E-06	1.22 (1.12 - 1.32)	1.3E-10
PDE3B	rs2167453 (G)	11p15.2	14865666	ZNF229	rs2734459 (C)	19q13.31	44928885	1.48 (1.21 - 1.80)	3.0E-05	1.55 (1.30 - 1.85)	1.7E-07	1.52 (1.33 - 1.73)	1.3E-10
LRRC15	rs923934 (G)	3q29	194081900	RMND1	rs13201167 (C)	6q25.1	151773504	1.32 (1.11 - 1.56)	9.8E-06	1.25 (1.12 - 1.39)	6.6E-06	1.27 (1.16 - 1.39)	1.5E-10
CACNA1C	rs2238087 (T)	12p13.33	2613716	KCNA5	rs17777157 (T)	12p13.32	5221668	1.35 (1.17 - 1.57)	8.7E-05	1.39 (1.16 - 1.67)	6.9E-08	1.37 (1.22 - 1.53)	1.6E-10
PGCP	rs6990629 (A)	8q22.1	98180213	NELL1	rs10766743 (T)	11p15.1	20925039	0.64 (0.57 - 0.72)	9.4E-08	0.93 (0.84 - 1.01)	8.0E-05	0.81 (0.75 - 0.87)	2.0E-10
CNR1	rs806366 (C)	6q15	88847589	FABP5L1	rs17089906 (C)	13q22.1	73686332	1.68 (1.52 - 1.85)	1.1E-07	0.83 (0.72 - 0.96)	8.2E-05	1.34 (1.24 - 1.46)	2.3E-10
ACTL8	rs4141983 (C)	1p36.13	18122009	CSMD2	rs3131529 (T)	1p35.1	34514486	1.17 (1.07 - 1.27)	1.6E-05	1.31 (1.18 - 1.45)	5.7E-07	1.23 (1.15 - 1.31)	2.3E-10
TUT7	rs2860107 (T)	9q21.33	89212523	RUNX1	rs2834882 (T)	21q22.12	36666340	1.03 (0.92 - 1.23)	7.3E-07	0.90 (0.82 - 1.09)	1.4E-05	0.96 (0.87 - 1.06)	2.7E-10
DISC1	rs1888601 (C)	1q42.2	232266922	TTC5	rs10130942 (C)	14q11.2	20756405	0.82 (0.80 - 1.05)	2.0E-04	1.26 (1.13 - 1.40)	8.0E-08	1.07 (0.98 - 1.16)	4.2E-10
PRKD1	rs12436395 (T)	14q12	30706026	TMEPAI	rs427278 (T)	20q13.31	56235119	1.42 (1.28 - 1.50)	2.0E-07	1.22 (1.11 - 1.35)	9.2E-05	1.34 (1.26 - 1.42)	4.6E-10
HDAC9	rs7788833 (C)	7p21.1	19034191	NCAM2	rs2408239 (T)	21q21.1	23332626	0.64 (0.57 - 0.72)	1.9E-07	0.96 (0.89 - 1.04)	9.4E-05	0.85 (0.79 - 0.90)	4.6E-10
C6orf195	rs6918808 (A)	6p25.2	2608995	TGDS	rs17181808 (A)	13q32.1	95186815	1.18 (1.08 - 1.29)	2.4E-07	0.83 (0.72 - 0.96)	7.7E-05	1.07 (0.99 - 1.15)	4.7E-10
HSP90AA4P	rs1496937 (G)	4q35.2	190004805	LOC730121	rs1365524 (T)	14q31.3	87770671	1.25 (1.12 - 1.39)	8.3E-07	1.61 (1.45 - 1.78)	2.2E-05	1.43 (1.33 - 1.54)	4.7E-10
THSD7B	rs719790 (C)	2q22.1	138207269	SLC8A2	rs4802363 (C)	19q13.32	47940364	1.13 (1.06 - 1.21)	2.9E-05	1.28 (1.13 - 1.46)	6.4E-07	1.16 (1.09 - 1.23)	4.7E-10
HSP90AB2P	rs17362130 (G)	4p15.33	12613974	LOC642681	rs4706511 (G)	6q13	73448086	1.24 (1.09 - 1.41)	1.4E-05	1.24 (1.12 - 1.38)	1.3E-06	1.24 (1.14 - 1.34)	4.8E-10
SORCS1	rs7095427 (C)	10q25.1	108426206	LOC646801	rs7130727 (T)	11p11.12	50232757	1.47 (1.21 - 1.80)	1.0E-05	1.19 (1.08 - 1.30)	1.9E-06	1.23 (1.13 - 1.34)	4.9E-10

Abbreviations:
SNP, single nucleotide polymorphism; Chr, Chromosomal band; Position, base pair; OR, odds ratio; CI, confidence interval, P_x, P value obtained from X

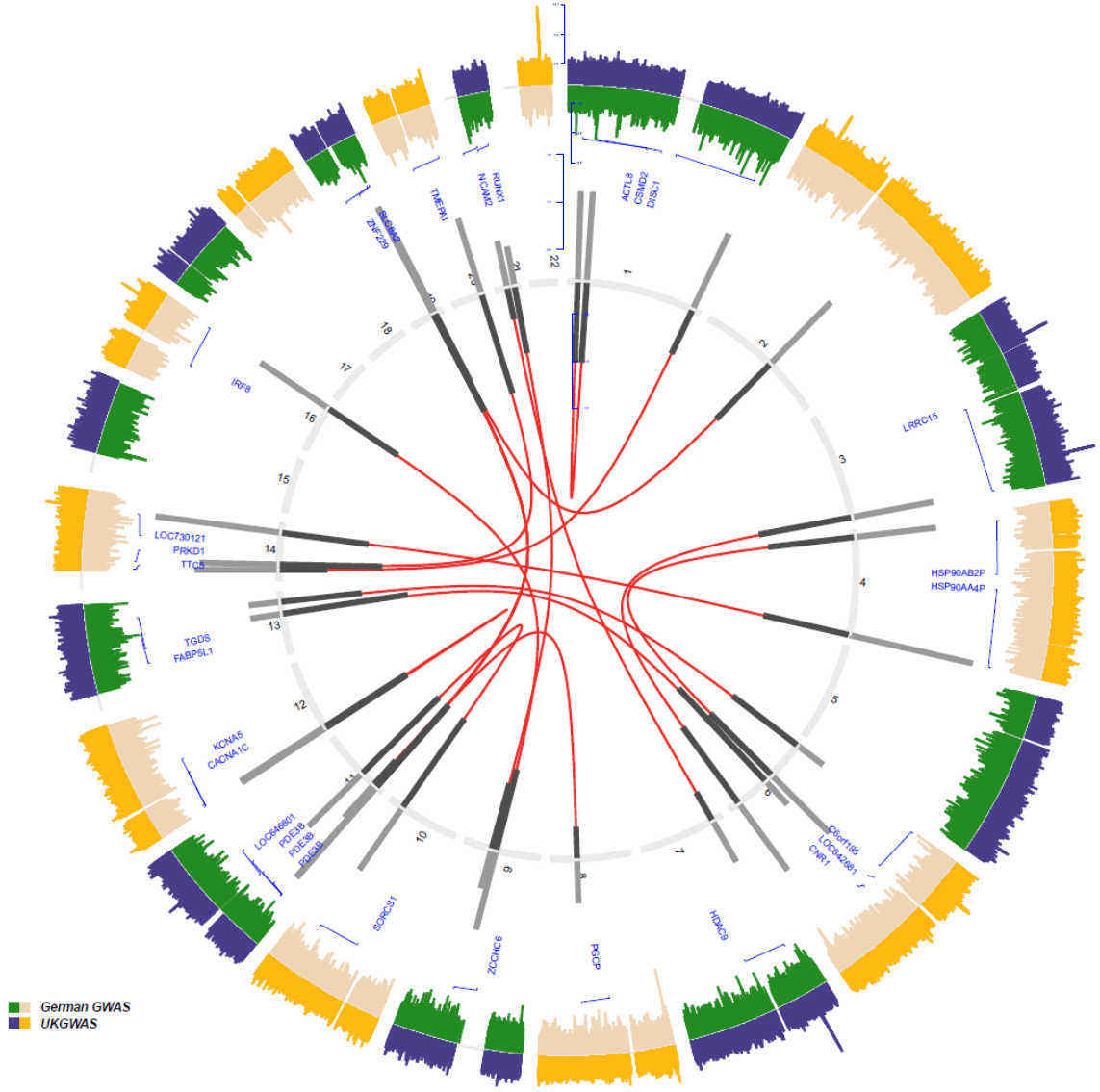


Figure 10. 1| Interaction Analysis identifies 16 unique risk loci pairs.
Circos plot of genome-wide association and significant interaction results for the identified paired risk loci. The two outer most panels display results from genome-wide association study on a Manhattan plot for autosomal variants on a negative log transformed scale. Inner numbered panel represents the chromosomes and effect-sizes of significant interacting pairs are plotted on bar charts from both samples (dark: German sample; light: UK sample). Interacting pairs are line joined in the inner most panels based on their chromosomal positions (NCBI build 19, human genome). Annotations of single nucleotide polymorphisms to gene ids are displayed on the inner manhattan plot.

10.2 Biological inference of the interacting chromosomal loci

Most of the risk SNPs identified, although showing promising signals for genotypic interactions, are mapped to non-coding regions of the genome and possibly contribute to MM etiology by affecting gene expression via differential regulation. Association of variations in quantitative traits with respect to the identified risk loci can shed light into such mechanisms. Hence eQTL data generated on malignant plasma cells obtained from MM patients of the German MM trials was interrogated. The most prominent eQTL signals were observed by rs2167453 at 11p15.2 for cytochrome P450, family 2, subfamily R, polypeptide 1 (*CYP2R1*) and by rs923934 at 3q29 for family with sequence similarity 43, member A (*FAM43A*), both with $P_{eQTL} = 4.4 \times 10^{-5}$ (**Table 10.2**). The interacting partners of these SNPs, rs2734459 and rs13201167 served as eQTLs with a moderate signal for other genes (rs2734459 for *CLASRP*, *ZNF224* and *APOE* and rs13201167 for *AKAP12* and *C6orf211*).

Summary-data-based Mendelian randomization addresses pleiotropic heritable effects observed between a trait and genetic exposure via gene expression regulation and the observed genetic component (usually genotypes). SMR was used to analyze pleiotropic effects between the GWAS signal and the cis-eQTL for genes residing within 1 Mb window of the sentinel loci in interaction to assess causal association between SNPs and disease phenotype via instrumentation of gene regulation. The strongest pleiotropic signal was observed at 4p15.33 by rs17362130 for RAS oncogene family member 28, *RAB28* ($P_{SMR} = 4.8 \times 10^{-3}$) and at 6p25.2 by rs6918808 for receptor (*TNFRSF*)-interacting serine/threonine kinase 1, *RIP1* ($P_{SMR} = 5.4 \times 10^{-3}$, **Table 10.2**, **Figure 10.2**), respectively. Contextually, it is well known that oncogenic ras family members are frequently mutated in MM (Aronson *et al.*, 2014). RIP1 interacts with RIP3 to activate the necrosome complex responsible for

instigation of several death receptors that induces apoptosis, necroptosis or cell proliferation. Additionally rs17362130 is found to be an eQTL for NK3 Homeobox 2 (*NKX3-2*) with a moderate signal ($P_{eQTL} = 2.1 \times 10^{-3}$) and rs6918808 was an eQTL for Serpin Family B Member 9 (*SERPINB9*).

Table 10. 2| GWAS summary data-based Mendelian randomization.

probe	Gene name	Gene ID	SNP ID	eQTL *P*-value	GWAS *P*-value	SMR *P*-value
9364_at	RAB28, member RAS oncogene family	RAB28	rs17362130	1.14E-03	3.68E-05	4.84E-03
8737_at	receptor (TNFRSF)-interacting serine-threonine kinase 1	RIP1	rs6918808	1.23E-03	4.01E-05	5.04E-03
7289_at	tubby like protein 3	TULP3	rs2238087	1.14E-03	2.58E-04	1.27E-02
808_at	calmodulin 3 (phosphorylase kinase, delta)	CALM3	rs4802363	1.76E-03	1.99E-03	1.28E-02
11133_at	kaptin (actin binding protein)	KPTN	rs4802363	1.98E-03	2.91E-03	1.30E-02
8605_at	phospholipase A2, group IVC (cytosolic, calcium-independent)	PLA2G4C	rs4802363	1.72E-03	4.62E-03	1.33E-02
120227_at	cytochrome P450, family 2, subfamily R, polypeptide 1	CYP2R1	rs10832312	4.40E-05	2.53E-02	1.55E-02
120227_at	cytochrome P450, family 2, subfamily R, polypeptide 1	CYP2R1	rs11023346	4.40E-05	2.56E-02	1.72E-02
120227_at	cytochrome P450, family 2, subfamily R, polypeptide 1	CYP2R1	rs11821380	4.40E-05	2.56E-02	1.72E-02
57820_at	cyclin B1 interacting protein 1, E3 ubiquitin protein ligase	CCNB1lP1	rs10130942	1.41E-03	3.98E-03	1.86E-02
10082_at	glypican 6	GPC6	rs17181808	1.06E-03	6.41E-04	1.86E-02
1690_at	coagulation factor C homolog, cochlin (Limulus polyphemus)	COCH	rs12436395	3.52E-04	1.88E-02	2.03E-02
120227_at	cytochrome P450, family 2, subfamily R, polypeptide 1	CYP2R1	rs2167453	4.40E-05	2.56E-02	2.10E-02
579_at	NK3 homeobox 2	NKX3-2	rs17362130	2.11E-03	1.18E-03	2.72E-02
80759_at	KH homology domain containing 1	KHDC1	rs4706511	1.01E-03	5.49E-03	3.47E-02
10553_at	HIV-1 Tat interactive protein 2, 30kDa	HTATIP2	rs10766743	1.85E-03	2.11E-03	3.60E-02
79624_at	chromosome 6 open reading frame 211	C6orf211	rs13201167	4.40E-04	2.47E-03	3.65E-02
160897_at	G protein-coupled receptor 180	GPR180	rs17181808	4.40E-04	5.04E-03	3.66E-02
5272_at	serpin peptidase inhibitor, clade B (ovalbumin), member 9	SERPINB9	rs6918808	1.01E-03	3.04E-03	3.80E-02
23483_at	TDP-glucose 4,6-dehydratase	TGDS	rs17181808	4.84E-04	6.32E-03	3.82E-02
440145_at	mitotic spindle organizing protein 1	MZT1	rs17089906	2.64E-04	9.85E-03	4.20E-02
9590_at	A kinase (PRKA) anchor protein 12	AKAP12	rs13201167	2.16E-03	3.63E-03	4.27E-02
688_at	Kruppel-like factor 5 (intestinal)	KLF5	rs17089906	1.76E-04	2.01E-02	4.54E-02
7767_at	zinc finger protein 224	ZNF224	rs2734459	7.04E-04	3.59E-03	4.66E-02
348_at	apolipoprotein E	APOE	rs2734459	1.23E-03	5.19E-03	5.55E-02
81029_at	wingless-type MMTV integration site family, member 5B	WNT5B	rs2238087	1.98E-03	6.45E-03	5.65E-02
404550_at	chromosome 16 open reading frame 74	C16orf74	rs7204305	1.98E-03	6.74E-03	5.67E-02
11129_at	CLK4-associating serine/arginine rich protein	CLASRP	rs2734459	2.64E-04	1.32E-02	8.43E-02
131583_at	family with sequence similarity 43, member A	FAM43A	rs923934	4.40E-05	1.90E-02	8.89E-02

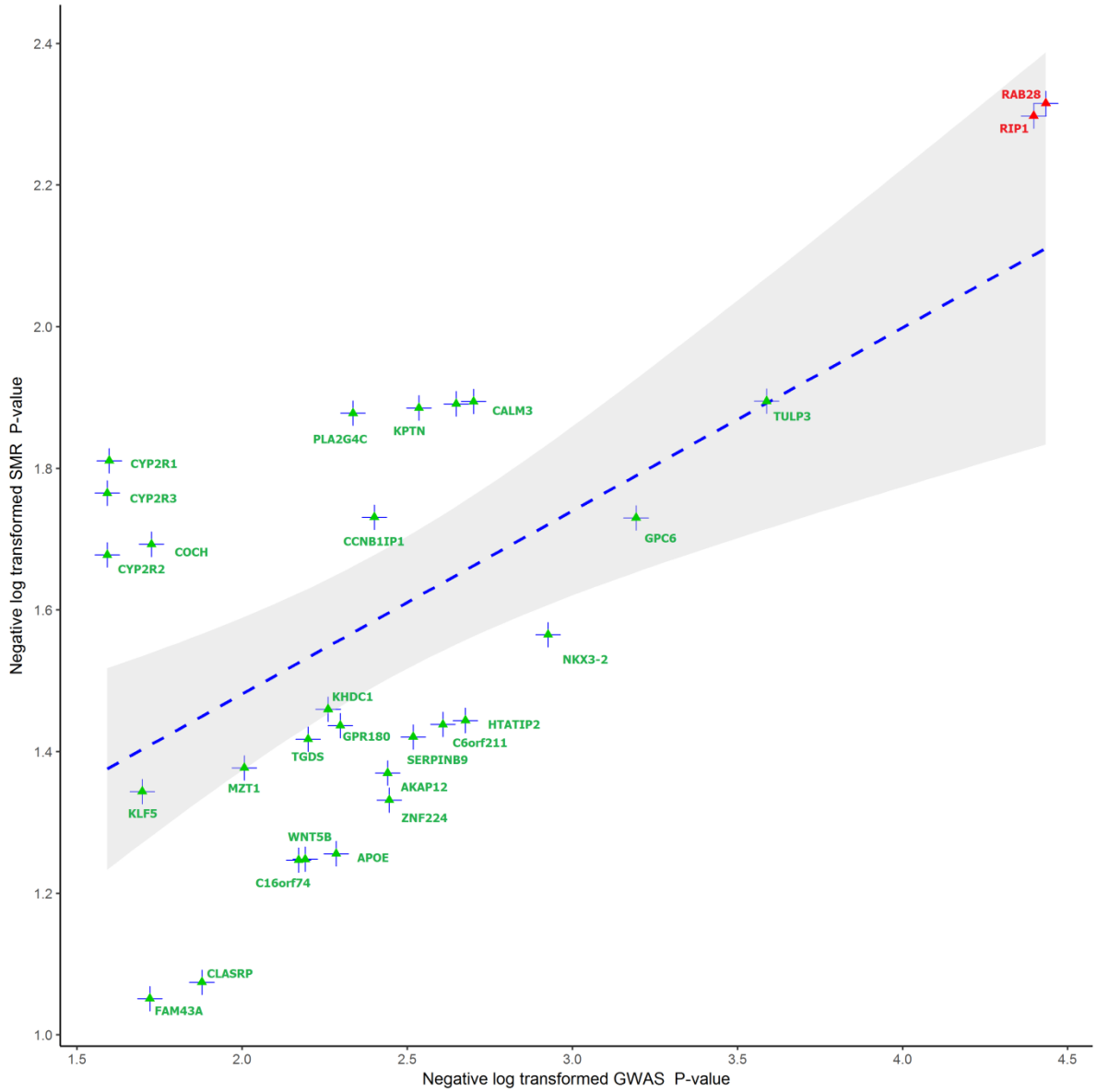


Figure 10. 2| Summary-data-based Mendelian randomization analysis of interaction detected multiple myeloma risk loci and gene expression in plasma cell

Negative log transformed P-values are plotted from GWAS against that of SMR identified causal cis-eQTLs at suggestive level. Top two significant elements are annotated in red. The blue line represents fitted liner regression representing linear association and the shaded region encompasses 95% confidence interval

10.3 Genetic interaction based Network

The hypothesis behind genetic interaction relies on deregulation of an array of genes that have an impact on a biological process leveraging expression of a certain phenotype. To investigate shared biological reciprocity as well as information driven connection between genes annotated to the variants identified via interaction study, a genetic network map was constructed. Unique annotations from the 16 interaction-identified variants along with the SMR-identified causally related genes were thus assessed with network enrichment and first order interacting genes based on data-mined enrichment index (protein-protein interaction index >0.95 on a scale of 0 to 1) were additionally added to increase confidence of the network.

The network thus created had a statistically significant enrichment *P*-value of $P_{network} = 2.7 \times 10^{-5}$. The top most enriched nodes were three genes (*ZNF224*, *ZNF229* and *KLF5*) heavily involved in transcription regulation along with enzyme modulators like *CALM3* and genes with direct involvement in B-cell selection and survival such as *GNAQ*. The disconnected nodes were disregarded from the network and interacting edges with minimum enrichment of 0.7 (on a scale of 0 to 1) were included (**Figure 10.3**).

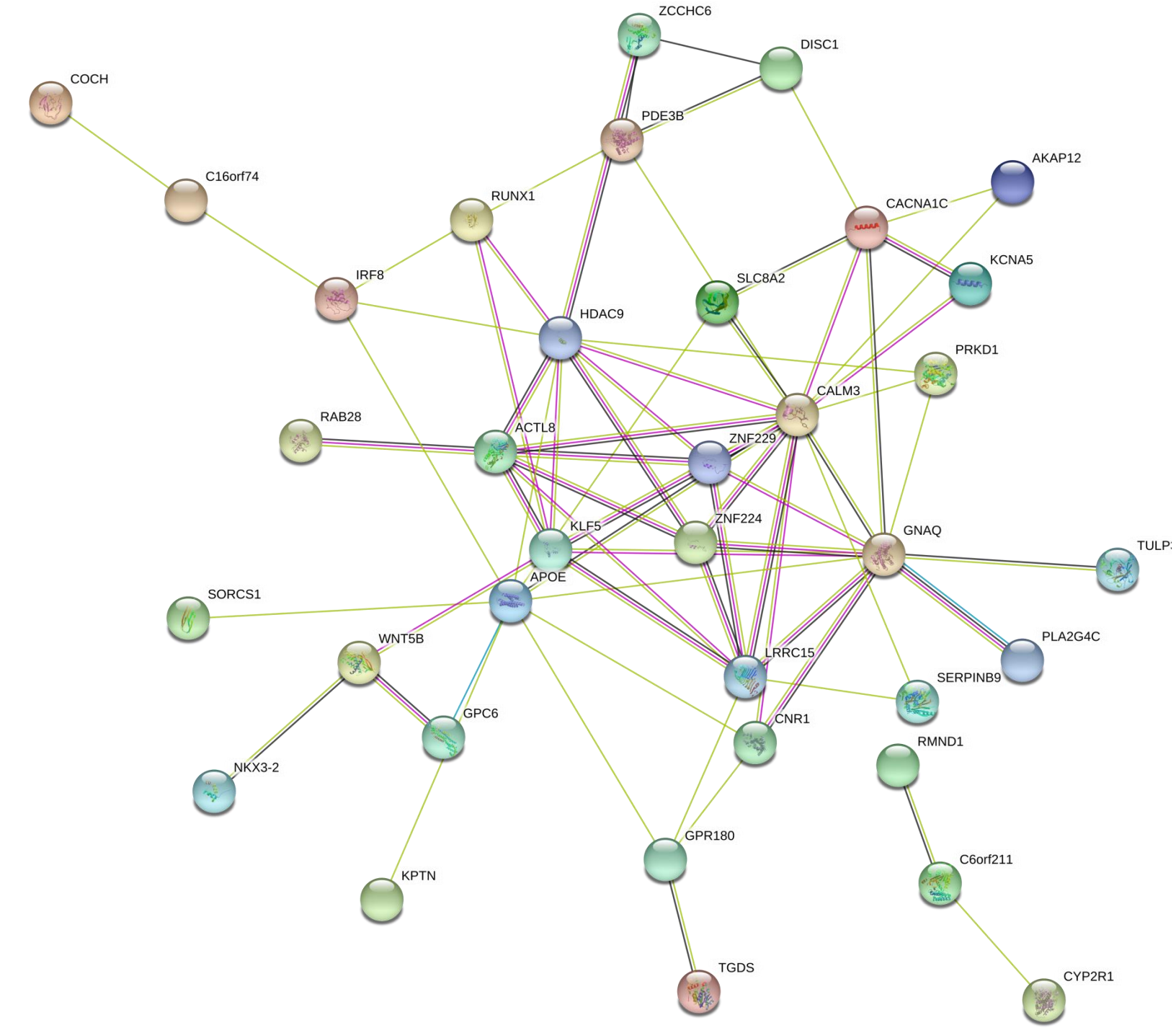


Figure 10. 3| Genetic network enrichment with STRING
All nodes represent direct annotations of interaction-identified elements or first order interaction. Colored edges convey status of predicted network edge correspondingly cyan, curated database; magenta, experimentally determined; forest green, gene neighborhood; red, gene fusion; navy blue, gene co-occurrence; lawn green, text mining; black, co-expression; lavender indigo, protein homology. Node color signifies different/shared protein functionality. Additional nodes are considered based on prediction score ≥ 0.99

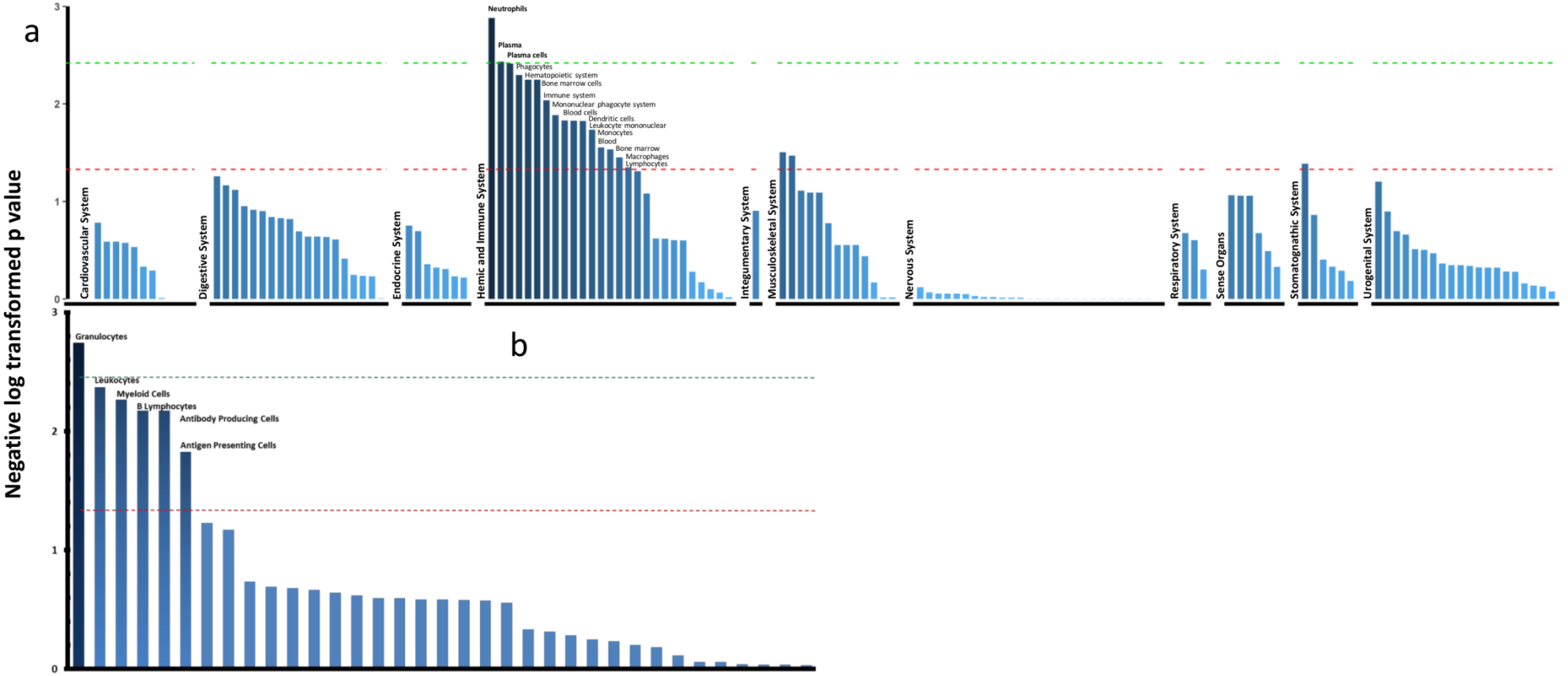


Figure 10. 4| Tissue and cell-type enrichment of interaction identified loci with DEPICT
a. Tissue enrichment identifies significant tissue types mostly affected with interaction-identified genes. b. Cell type enrichment analysis identifies cells with observed expression regulation of the same candidates.

10.4 Tissue and cell type enrichment

In silico detection of gene expression enrichment in tissues and cell types for the genes annotated to the interaction-associated loci were performed with DEPICT. Sentinel SNPs were prioritized based on the annotation and backend expression data predicted functional relevance and were subsequently clustered into 12 unique loci. These fused loci were tested for significant deregulated expression of the corresponding genes in 209 MeSH annotations against 37,427 microarrays procured in backend. A total of 27 tissue or cell type annotations were found significant at a suggestive level ($P < 0.05$). Among the enriched tissue annotations, 16 were pertinent to the hemic and immune system, two belonged the musculoskeletal system and one to the stomatognathic system (**Figure 10.4a**) and additionally six cell types were found enriched that were related to hematopoietic system (**Figure 10.4b**).

10.5 Biological inference of the GWAS-identified loci with Pathway analysis

Next, relationships amongst the previous GWAS-identified loci in the context of regulation of pathways using the pathway analysis tool PASCAL were interrogated. To avoid possible complications arising from unconformity to statistical convergence of the test statistic, sum of chi-square method was used to test for functional association against pathway annotations extracted from REACTOME, KEGG and BIOCARTA libraries. A total of 12 enriched pathways reached a global threshold of 0.0025 for the combined pooled *P*-value (**Table 10.3**). Among all the pathways thus detected, three were signaling cascades reflecting the activation status of the SMAD family proteins, as signal transducers for receptors of the cytokine Transforming Growth Factor β. They were represented with the following

pathways, “SMAD2 SMAD3 SMAD4 heterotrimer regulates transcription”, $P_{combined} = 6.9 \times 10^{-4}$, “TGFβ receptor signaling activates SMADs”, $P_{combined} = 1.1 \times 10^{-3}$ and “Transcriptional activity of SMAD2 SMAD3 SMAD4 heterotrimer”, $P_{combined} = 2.8 \times 10^{-3}$. “Circadian repression of expression by REV-ERBA”, $P_{combined} = 5.5 \times 10^{-4}$ was the top signal; and an additional pathway “RORA activates circadian expression", $P_{combined} = 2.1 \times 10^{-3}$ was also found related to the regulation of circadian rhythm which was mediated by two nuclear receptor proteins RORA and REV-ERBA. Furthermore, modulation of *ALK* receptor tyrosine kinase activity was indicated to be enriched with ALK pathway, $P_{combined} = 2.8 \times 10^{-3}$.

Table 10. 3| Pathway enrichment analysis with PASCAL detects 12 putative pathways related to MM.

Data base	Pathway	P_{Ger}	P_{UK}	P_{Meta}	$^{\ddagger}P_{Combined}$
REACTOME	Circadian repression of expression by REV-ERBA	3.50E-04	1.45E-01	4.16E-03	5.52E-04
REACTOME	APOBEC3G mediated resistance to HIV infection	5.79E-02	1.74E-03	2.09E-03	1.02E-03
REACTOME	RORA activates circadian expression	1.24E-03	1.83E-01	1.20E-02	2.13E-03
REACTOME	Deposition of new CENP-A containing nucleosomes as the centromere	7.00E-02	7.49E-03	3.82E-03	4.48E-03
REACTOME	SMAD2 SMAD3 SMAD4 heterotrimer regulates transcription	8.83E-02	7.81E-03	1.88E-02	5.70E-03
REACTOME	TGFβ receptor signaling activates SMADs	1.73E-02	6.39E-02	4.38E-03	8.60E-03
REACTOME	GABAA receptor activation	2.36E-02	6.27E-02	1.62E-02	1.11E-02
REACTOME	Iron uptake and transport	4.84E-02	4.20E-02	8.91E-03	1.46E-02
REACTOME	Transcriptional activity of SMAD2 SMAD3 SMAD4 heterotrimer	9.53E-02	2.18E-02	4.15E-02	1.49E-02
REACTOME	Purine salvage	8.82E-02	2.51E-02	3.71E-02	1.57E-02
REACTOME	Apoptosis induced DNA fragmentation	1.76E-02	1.29E-01	2.32E-02	1.60E-02
BIOCARTA	ALK pathway	9.49E-03	3.28E-02	3.12E-02	2.82E-03

Abbreviations:
‡combined with Brown's method for dependent *P*-values
$\mathbf{P_X}$: *P*-value obtained from interaction analysis of set ***X***

Chapter 11: Risk of second primary cancer in MM patients

11.1 Rationale

As is the case with almost every other cancer, advancement in treatment of MM has resulted in increasingly prolonged survival and in turn has observed rising incidence of second primary cancers (SPCs). The reasons behind increased risk of second cancers are of several multitudes and have been discussed in details in section 2.2. Over the last two decades MM risk has been extensively established to be moderately carried by an inherited/shared familial component (Lynch *et al.*, 2008a; Lynch *et al.*, 2005). The first investigation on risk of MM in people with family history of a cancer was performed more than three decades ago (Bourguet *et al.*, 1985), and since then a number of confirmational studies have shown a family history of cancer greatly influences the risk of MM itself (Alexander *et al.*, 2007; McDuffie *et al.*, 2009). On the other hand starting from the same period, an existing history of cancer in family has been shown to harbor detrimental effects towards developing subsequent cancers in patients of several different cancers (Bernstein *et al.*, 1992; Kony *et al.*, 1997). First documented case of development of second cancers in MM patients dates back to 1982 and a plethora of investigations have shown almost all type of cancers arising in MM patients ever since (Chen *et al.*, 2016; Razavi *et al.*, 2013; Thomas *et al.*, 2012; Zalcberg *et al.*, 1982).

It has long been postulated that a subset of patients with cancer display a high sensitivity to mutational agents because of genetic predisposition. The extent of impact due to family history of cancer in context of MM is previously investigated by means of the 2004th update of the FCD which demonstrated pertinent familial clustering of MM with several different types of leukemia as well as with a number of solid tumors (Altieri *et al.*, 2006). Family

history is a prominent surrogate for heritable genetic and environmental constituents and the impact of family history of cancer in predisposition of excess heritable risk of second cancer has not been addressed yet.

To gain insight on relationship of family history of cancer and SPC in concordant sites, a cohort of 5,205 Swedish MM patients was analyzed. Also influence of second cancers on the cause of death was investigated to understand severity of outcome in patients with MM.

11.2 Patients

Starting from 1958 in Sweden there were 5,205 MM patients identified via the family cancer registry who had full parental information mapped and were diagnosed before the end of 2015, marked by the end of study follow-up. Among them 360 (6.9%) developed a subsequent SPC. Familial SPCs were compared to non-familial cases where analysis of all SPCs was restricted subject to availability of at least two cases having the same tumor (concordant) in a parent or sibling. Family history was treated as a dichotomous outcome without quantification of number of affected family members present. Without consideration of the overlapping impact of more than one cancer in family, prostate cancer was the major contributor to the family history (20%) followed by colorectal (14%), breast (10%), bladder (5%),and lung cancer (4%) and skin SCC (4%).

11.3 Familial risk of second cancer in patients with MM

In patients without a family history of cancer, the risk of SPC was increased for skin cancer (squamous cell carcinoma, SCC, RR = 2.58, 95% CI = 1.81 - 3.67, **Table 11.1**) and leukemia (RR = 4.55, 95% CI = 3.11 - 6.24). For patients with a family history of cancer,

even though case numbers were low, familial risks were found with significant excess with a trend test for colorectal ($RR_{familial}$, 95% CI: 2.10 [1.00 - 4.41] vs. $RR_{non\text{-}familial}$, 95% CI: 1.01 [0.69 - 1.47]), prostate ($RR_{familial}$, 95% CI: 1.60 [1.03 - 2.48] vs. $RR_{non\text{-}familial}$, 95% CI: 0.56 [0.41 - 0.77]) and skin SCC ($RR_{familial}$, 95% CI: 8.82 [3.31 - 23.52] vs. $RR_{non\text{-}familial}$, 95% CI: 2.58 [1.81 - 3.67]). Although high excess familial risk was observed for lung cancer, the trend test was not significant ($P = 0.061$) possibly indicative of weak confidence due to inadequate sample size. The highest familial SPC risk was observed for MM patients with a family history of leukemia ($RR_{familial}$, 95% CI: 9.14 [2.29 - 36.55], only 2 cases) although again with insignificant trend *P*-value. Overall patients with any cancer history in family (N = 246) were 68.3% of all SPCs and the RR was in significant excess; $RR_{familial}$, 95% CI: 1.38 [1.22 - 1.57] vs. $RR_{non\text{-}familial}$, 95% CI: 1.13 [1.17 - 1.43] respectively (trend test $P < 0.001$).

11.4 Population drift and temporal effect on incidence

Population drift overtime is a major source of bias in epidemiological studies especially when the sample size is expanding in a non-linear trend over the follow-up period (Carstensen, 2006). Sensitivity of the analysis needed to be tested for possible skewed patient reporting based on the multiple applied conditions and to this end case frequencies were plotted to observe the patient accrual over the study period (**Figure 11.1**). The figure shows MM patients with SPC and with or without family history (246 and 114 patients) plotted with temporal stratification by 5-year intervals of MM diagnosis. No skewed drift was observed to suspect incremental bias.

11.5 Cause of death

By the end of 2015 a total of 2872 (55.2%) among 5,205 MM patients were declared deceased; and the total number of deaths among 360 patients with SPC was 228 (60.6%). The proportion was equally high among 246 patients with familial SPC, of whom 146 (59.3%) had died by then. Kolmogorov-Smirnov test on proportion difference found no evidence of statistical difference between the familial and non-familial groups ($P > 0.05$).

MM itself is characterized with moderate to poor prognosis in most cases, and unsurprisingly MM was the most common cause of death in patients without SPC (83%, 2194/2644), with 17% of deaths due to other causes. For MM patients with a SPC, the distribution of causes of death is shown in **Table 11.2**. Here also MM was found to be the foremost cause of death with 38.7% of demises, followed by SPCs accountable for 35.8% and other causes the rest 25.5% of all death; among other causes the majority of deaths (62.9%) were due to non-neoplastic causes. The mortality of SPC varied between second cancer types in proportion to the severity of detrimental survival. For second pancreatic cancer, all 7 patients died of this cancer; more than half of MM patients died of SPC when it was lung or nervous system cancer or leukemia. Other causes were important for CUP as SPC. There were 82 deaths observed in patients with SPC in absence of a cancer family history, again majority of death was due to MM (36.6%), followed by SPCs (34.2%).

11.6 Interaction in personal history and family history of cancer

Personal history of cancer and family history of cancer can be strong attributor of bias in this study design. Usually genetic aberrations have deleterious consequence in cancer patients starting from diagnosis, disease progression or remission. Very often such

aberrations are shared among a broad family of malignancies which the patients are already exposed to and hence the inference on association is expected to carry larger departure from causality in presence of interaction between personal and family history of cancer. Linear and non-linear interactions of significant family risks and risk of SPC with additive and multiplicative interactions were tested. A stronger than additive interaction was found for skin cancer (P = 0.04, **Table 11.3**). Although several other interactions carried large effect sizes, these were all statistically insignificant at 5% level.

Table 11. 1| Relative risks of SPCs among all multiple myeloma patients stratified over family

	At least 1 FDR with cancer			No FDR with cancer			Total			Trend test *P* value
Cancer	**N**	**RR**	**95% CI**	**N**	**RR**	**95% CI**	**N**	**RR**	**95% CI**	
Colorectum	7	**2.10**	1.00 - 4.41	27	1.01	0.69 - 1.47	34	1.13	0.81 -1.58	0.033
Lung	3	***5.40***	1.74 - 16.75	10	1.13	0.61 - 2.10	13	1.38	0.80 - 2.38	0.061
Breast	4	1.13	0.42 - 3.01	24	0.93	0.62 - 1.39	28	0.95	0.66 - 1.38	0.176
Prostate	20	**1.60**	1.03 - 2.48	38	***0.56***	0.41 - 0.77	58	**0.72**	0.56 - 0.93	0.006
Melanoma	2	**5.04**	1.26 - 20.14	18	1.46	0.92 - 2.32	20	**1.57**	1.01 - 2.44	0.087
Skin (squamous cell carcinoma)	4	***8.82***	3.31 - 23.52	31	***2.58***	1.81 - 3.67	35	***2.81***	2.01 - 3.91	0.029
Leukemia	2	***9.14***	2.29 - 36.55	32	***4.41***	3.11 - 6.24	34	***4.55***	3.25 - 6.37	0.093
All	246	***1.38***	1.22 - 1.57	114	1.13	0.94 - 1.36	360	***1.29***	1.17 - 1.43	<0.001

Abbreviation:
FDR, first degree relative; N, frequency; RR, relative risk; CI, confidence interval;
Bold, italics and underline indicate 5%, 1% and 0.1% level of significance;

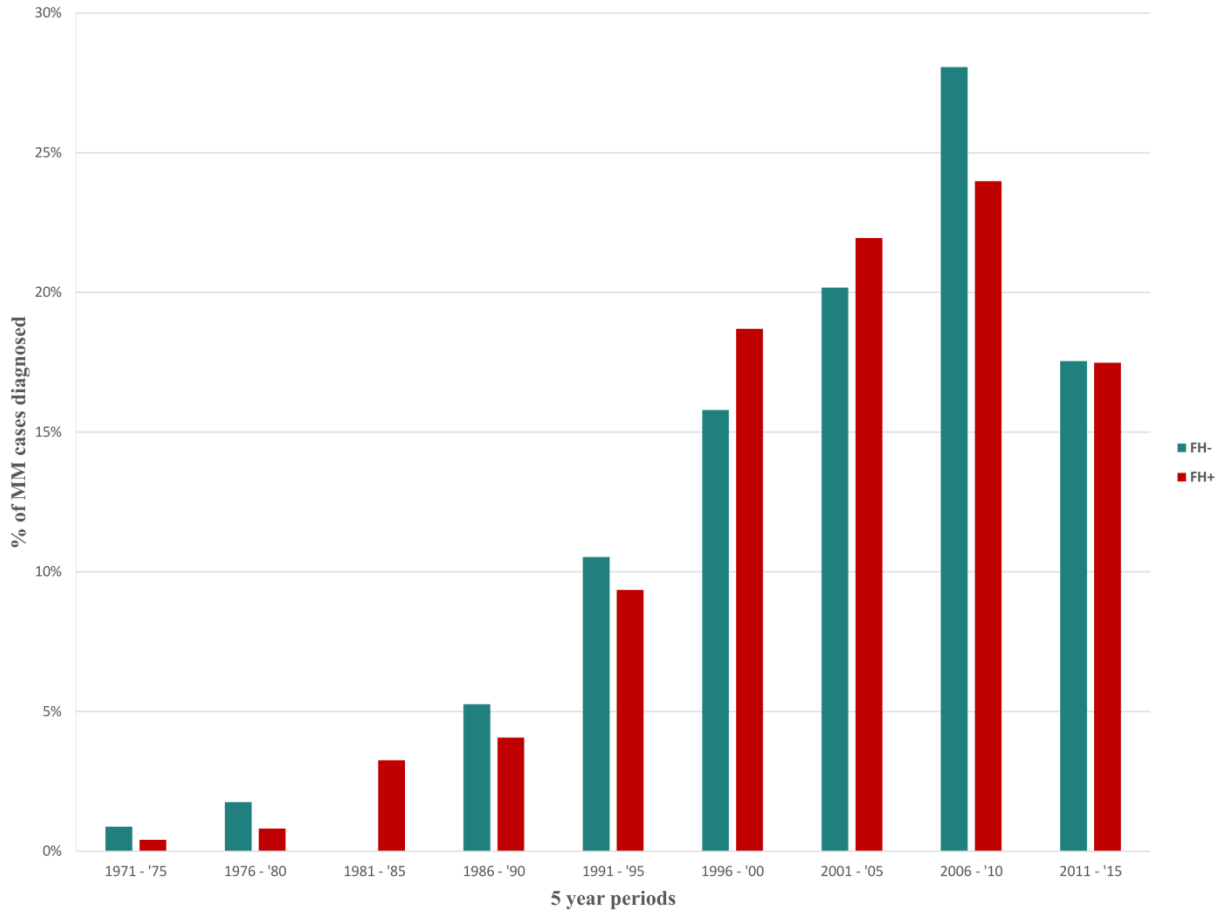


Figure 11. 1| Period overview of SPC diagnosis in MM patients
MM patients with SPC stratified with or without family history (246 and 114 patients) in 5-year intervals of MM diagnosis. Abbreviations: SPC, second primary cancer; FH-, family history negative and FH+, family history positive.

Table 11. 2| Causes of death distribution of multiple myeloma patients diagnosed with SPC

	MM		[a] Second primary cancer		Other causes	
Cancer	**N**	**%**	**N**	**%**	**N**	**%**
Upper aero-digestive tract	2	50.0	2	50.0	-	-
Stomach	-	-	4	100.0	-	-
Colorectum	8	33.3	11	45.8	5	20.9
Anus	-	-	1	100.0	-	-
Liver	2	33.3	3	50.0	1	16.7
Pancreas	-	-	7	100.0	-	-
Lung	3	13.6	15	68.2	4	18.2
Breast	6	42.9	1	7.1	7	50
Cervix	-	-	1	100.0	-	-
Ovary	1	50.0	1	50.0	-	-
Prostate	11	42.3	5	19.2	10	38.4
Kidney	3	37.5	3	37.5	2	25
Urinary bladder	5	41.7	3	25.0	4	33.3
Melanoma	7	58.3	3	25.0	2	16.7
Skin (squamous cell carcinoma)	16	72.7	1	4.5	5	22.7
Nervous system	3	42.9	4	57.1	-	-
Non-Hodgkin lymphoma	5	45.5	4	36.4	2	18.2
Hodgkin lymphoma	-	-	1	50.0	1	50
Leukemia	7	24.1	16	55.2	6	20.6
Cancer of unknown primary	3	21.4	1	7.1	10	71.4
[b] Total	94	38.7	87	35.8	62	25.5

a Cases noted only when at least one death is observed due to second cancer.
b Total includes all cancers without constraints.

Table 11. 3| Interaction between concordant cancer family history and individual history of multiple myeloma

	Risk in population with at least one FDR with concordant cancer	Risk in multiple myeloma survivors	Risk in multiple myeloma survivors with FDR diagnosed with cancer	Type of Interaction			
				Additive		Multiplicative	
Cancer site	**RR (95% CI)**	**RR (95% CI)**	**RR (95% CI)**	**ICR (95%CI)**	***P***	**MII (95% CI)**	***P***
Colorectum	1.91 (1.83 – 1.99)	1.13 (0.81 – 1.58)	2.10 (1.00 – 4.41)	0.06 (-1.22 – 5.62)	0.79	0.97 (0.71 – 2.30)	0.61
Prostate	2.49 (2.42-2.57)	0.72 (0.56 – 0.93)	1.60 (1.03 – 2.48)	-0.61 (-1.38 – 1.95)	0.66	0.89 (0.59 – 1.68)	0.57
Skin SCC	1.99 (1.80 – 2.21)	2.81 (2.01 – 3.91)	8.82 (3.31 – 23.52)	**5.02** (0.86 – 13.43)	0.04	1.58 (0.77 – 2.96)	0.39

Abbreviation:
FDR, first degree relative; RR, relative risk; CI, confidence interval; SCC, squamous cell carcinoma; ICR, interaction contrast ratio; MII, multiplicative interaction index;
Confidence intervals and P values calculated by bootstrapping 100,000 replications;
Bolding indicate statistical significance;

Discussion

Chapter 12: Inherited polygenic risk in MGUS

MGUS being an asymptomatic condition has resulted in difficulty in case ascertainment and poorly understood etiology. Since the beginning of exploration in inherited genetic landscape of MM, similar researches in MGUS have been designed to

i. Identify novel loci predisposing to MGUS

ii. Replicate risk loci identified for disease downstream to MGUS such as MM, Al amyloidosis in MGUS.

To this end initial reports were on confirmation signals observed through GWAS of MM found in MGUS which mostly reported insignificant to very moderate signals (Greenberg *et al.*, 2012; Weinhold *et al.*, 2014a). Next was the first attempt to address the architecture of polygenic predisposition in MGUS by means of GWAS on 242 people in 2017 (Thomsen *et al.*, 2017). This study discovered 10 common SNPs exerting excess MGUS risk but the signals were again mostly moderate in strength due to caveats in design.

Genetic aberrations associated with age are a major carrier of neoplastic growth burden. As MGUS is prevalent in almost 6.6 % of general "healthy" population aged 80 years or above (Wadhera *et al.*, 2011). Due to its apparent asymptomatic nature it is difficult to gauze the spectrum or enormity of its inherited genetic landscape. Notwithstanding studies depict that all MM is preceded by MGUS but the genetic makeup of progression is still elusive and so is its possible implication on survival and mortality (Bladé *et al.*, 2009; Kyle and Rajkumar, 2015; Weiss *et al.*, 2009). All of these led to investigation of the following questions:

1. Can we have a clearer picture of inherited genetic predisposition in MGUS?
2. How the identified genetic aberrations can alter biology in host?

3. How much of it is related to what we know about MM?*
4. Is the detection algorithm novel?*

To address the problem of missing heritability pertinent to cumulative aggregation of polygenic risk in MGUS the assumption of this study design was two-fold,

i. MGUS is influenced by cumulative risk of sub-par signals with no apparent deleterious impact co-inherited by the host.
ii. The susceptibility loci are truly polygenic (they predispose to MGUS with elevated risk in concert).

Keeping these in mind this investigation attempted to find answers to the previously posed questions in context of this research.

*These two points are later discussed with MM.

Can we have a clearer picture of inherited genetic predisposition in MGUS?

As the assumptions dictate departure from single marker risk, inter/inter-chromosomal risk was analyzed with genome-wide interaction studies. There were 14 unique loci confirmed with three stages of the analyses. The strongest signal was found to harbor Tenascin C (*TNC*), a protein coding gene residing in 9q33.1 in interaction with Crystallin Lambda 1 (*CRYL1*) in 13q12.11, a potent regulator of alternative glucose metabolic pathway. Expression of *TNC* is found upregulated in certain MM cell lines in the presence of mutations in insulin growth factor receptor and receptor tyrosine-protein kinase genes (Leich *et al.*, 2013b).

The second strongest signal was observed between *SETBP1* and *PREX1* interaction at 18q12.3 and 20q13.13. The locus at 20q13.13 predisposes to MM as an expression and

methylation quantitative trait locus at *PREX1* without affecting an active promoter site (Mitchell *et al.*, 2016). *PREX1* is expressed mainly in peripheral blood leukocytes and moderately in lymph nodes, and much weaker in most other tissues (Welch *et al.*, 2002). *SETBP1* is a well-established candidate gene harboring somatic mutations in various myeloid malignancies including secondary acute myeloid leukemia (sAML) and chronic myelomonocytic leukemia (CMML) (Makishima *et al.*, 2013).

Two other interactions showed a shared mutual homology. These include Erb-B2 Receptor Tyrosine Kinase 4 (*ERBB4*) at 2q34, Retinoic Acid Receptor Related Orphan Receptor A (*RORA* or alias *RORa*) at 15q22.2 and Protein Tyrosine Phosphatase, Receptor Type D (*PTPRD*) at 9p23. Both *RORA* and *PTPRD* were found to have *ERBB4* as interacting partner. All three of these genes in the context of the disease biology in concern will be discussed later. In summary, *ERBB4* is the fourth member of a tyrosine protein kinase family and is known by its alias *HER4*. It plays an important role as a cell surface receptor for neuregulins (NRGs) and EGF family members as well as in gene transcription, cell proliferation, differentiation, migration and apoptosis. *PTPRD* encodes a protein tyrosine phosphatase (PTP) family member protein. PTPs in general are implicated in several cellular processes including cell growth, differentiation, mitotic cycle and even in oncogenesis. *PTPRD* demonstrates tumor suppressing mechanism in MM; it also dephosphorylates *STAT3*, an *IL6* signaling promoter that has a major consequence in MM pathogenesis (Egan *et al.*, 2012; Kamada *et al.*, 2012; Lohr *et al.*, 2014). Lastly the protein encoded by *RORA* is a member of nuclear receptor 1 subfamily of nuclear hormone receptor. It binds to the DNA as a monomer to Retinoid-Related Orphan Receptor (ROR)

response elements and is a key regulating element in circadian clock mediation (circadian rhythm), immunity, and cellular differentiation.

How the genetic aberrations alter biology in host

Epidermal growth factor receptor, a cell membrane growth factor receptor mediated by tyrosine kinase activity, is a member of ErbB receptor family and is widely expressed in human tissues regulating important cellular processes. In both cancerous and non-cancerous cells, EGFR plays a crucial role in controlling key cellular transduction pathways influencing cell proliferation, differentiation and development and overexpression of which is associated to multiple site-specific tumors including that of breast, lung, colorectum, head and neck, pancreas and bladder (Warta and Herold-Mende, 2017; Yarden, 2001). *ERBB4* was identified to be a high risk loci interacting with 15q22.2 and 9p23 establishing its regulatory burden on EGFR downregulation pathway, one of the enriched pathways in pathway analysis. Formation of EGFR-EGFR dimers mediates the 170kDa protein functionality and is dependent on three members of human epidermal receptor (HER) family proteins; namely HER1 (ErbB1/EGFR), HER2 (ErbB2) and HER4 (ErbB4). The EGFR triggering signal transduction operated by HER1:4 includes a. RAS- and mitogen-activated protein kinase (MAPK) pathway which controls cell proliferation b. phosphatidylinositol-3 kinase (PI3K) pathway driving cell development and c. protein kinase B (Akt) pathway arbitrating apoptosis.

Also found was an interacting pair annotated to intronic anaplastic lymphoma receptor tyrosine kinase (*ALK*), an oncogene and a common variant located 36kb 5’ to *GLCCTI1* which is one of previously identified risk loci for MGUS with moderate significance. In

anaplastic large-cell lymphomas, overexpression of *ALK* shows substantial unregulated tyrosine kinase activity. Deregulation in Akt-pathway also downregulates expression of multiple members of EGFR signaling cascades which hinders cancer cell cycle arrest and cell death.

Cell adhesion is an integral part of cell surface interaction and is essential for the organization and various biological functions of multicellular organisms. There are two major types of cell adhesions; cell-to-cell and cell-to-extracellular matrix (ECM), both of which consist of transmembrane cell adhesion molecules, intracellular scaffold or signaling proteins and cytoskeletons. Cadherin (CDH) family cell adhesion molecules and their associated scaffold proteins (catenins) play important roles in the formation and functions of cell-cell adhesions. One previously identified MGUS risk locus annotated to *GALNT1* was shown in the data to have moderately significant interaction with Cadherin 2 (*CDH2*), an adhesion molecule and an important downstream target of FGFR3 signaling pathway. CDH2 has been reported to be overexpressed among a cohort with MM diagnosis having t(4,14) translocation (Dring *et al.*, 2004). The identified novel risk locus on a cadherin group gene *CDH13* was found in interaction with a tumor suppressor gene Basonuclin 2 (*BNC2*) at 9p22.3. CDH13 protects vascular endothelial cells from apoptosis due to oxidative stress and is found to be hypermethylated in myeloid leukemia, B-cell lymphomas among several other cancers (Alkebsi *et al.*, 2016; Ogama *et al.*, 2004).

Among the pathways that were found enriched in all three algorithms was dorsoventral axis formation, KEGG autoimmune thyroid disease as well as allograft rejection. Although belonging to a larger network of downstream signaling, KEGG allograft rejection pathway is regulated by differential expression of *EGFR*, *MAPK1-3*, and *NOTCH1-3*; in addition to

MM pathogenic proto-oncogenes such as *KRAS* and *BRAF*. These finding allude to underlying mechanisms that relate progression of MGUS to MM at a cellular level.

Chapter 13: Inherited polygenic risk and its implications in MM

Now as a clearer picture emerges elucidating genetic inheritance pattern and biological mechanisms pertinent to MGUS development, it burrows pathologic interpretation from available literature on MM. Genetic predisposition of MM has been broadly investigated in the recent years and so far the largest meta-analysis has identified/confirmed 23 susceptibility loci exerting excess risk (Went *et al.*, 2018). Yet all of it put together only explains a moderate portion of MM heritability (15.2%) (Mitchell *et al.*, 2015). MM is characteristically a heterogeneous disease that infests several different chromosomal abnormality profiles in different hosts and depending on that the disease progression has vastly different consequences. Hence a one model fits all type risk loci dependent linear risk estimation would always perform moderately and sensitivity of such a model to forecast pathogenic consequences would remain poor. Overcoming this caveat would require evidence implicating the risk loci to biological processes which will vary in sensitivity depending on penetrance and pathogenicity.

The problem of missing heritability still remains to be explored in MM in a similar setting as MGUS. Secondly, the risk loci thus to be discovered also require biological interpretation to MM pathogenesis and furthermore, due to the overlap of investigation design, it is also desirable to interrogate possible overlap between MGUS and MM biology (if any) that may pose as mechanistic link between the benign and malignant phases of the disease transformation which is presumably incited by myelomagenesis. Hence this phase of the study is enshrined with the following three main objectives:

1. Interrogating genetic predisposition architecture of MM form the perspective of chromosomal interaction.

2. Identifying and implicating biological mechanisms of MM predisposition
3. Investigating mechanistic overlap in biology between MM and MGUS pathogenesis.

Interferon regulatory factors and T helper cells

In the interaction study *GNAQ* at 9q21.2 and *IRF8* at 16q24.1 held the strongest meta-analyzed signal. G Protein Subunit Alpha Q (*GNAQ*) as the name suggests encodes a guanine nucleotide-binding protein that regulates B-cell development and survival (Offermanns, 2006). Mutation in this gene has been associated to aberrant platelet aggregation and activation. Interestingly, platelet aggregation and plug information was detected to be one of the most enriched pathways in MGUS (section 9.3). Involvement of *GNAQ* in MM predisposition and differential regulation of its entry pathway allude to a possible mechanistic connection between MGUS and MM. Additionally, its interacting partner *IRF8* has been implicated in a significant repertoire of MM pathogenesis literature. Most importantly *IRF8* harbors an intergenic common SNP for Ig trait modulation, a critical mechanism MM and related paraproteinemias (Jonsson *et al.*, 2017). Additionally IRF8 is responsible for critical functions in regulation of innate as well as adaptive immunity and immune cell development including B- and T-cells, dendritic cells and myeloid cells (Zhao *et al.*, 2015). In the development of B-cells IRF8 and IRF4 (another member of interferon regulatory factor family of transcription factors) function redundantly and regulate transition of pre-B-cells to matured B-cells. In germinal center development, the roles of interferon regulatory domains are complementary: where IRF8 directs early centroblast development which is later taken over by IRF4 as centrocytes mature into plasma cells. IRF8 induces

activation-induced cytidine deaminase with is a key enzyme catalyzing somatic hypermutations of plasma cell (Zhao *et al.*, 2015).

Furthermore, the underlying mechanism of IRF8 transcriptional activity in MM may also be vastly elucidated with its role in T helper cell (Th) differentiation. Elevated cytokines in bone-marrow microenvironment is a key part of MM niche. In MM, cytokines such as IL6 and TGFβ are often expressed in abundance and are important for generation of Th17 cells. Th17 cells among other interleukins produce high level of IL-17 that promotes MM cell growth and inhibits immune function. IRF8 acts as an intrinsic transcriptional inhibitor of Th17 cells, at least partly through its physical interaction with retinoic acid receptor-related orphan receptor RORγt (Ouyang *et al.*, 2011). These findings are well-aligned with previously identified MM risk SNP rs4487645 at 7p15.3, as a modulator of IRF4 binding at an enhancer element of c-Myc interacting gene *CDCA7L* and support the role of the genetic variants in *IRF8* and its interacting partner in *GNAQ* in MM susceptibility (Broderick *et al.*, 2011; Li *et al.*, 2016; Weinhold *et al.*, 2015).

Retinoic acid receptor and circadian rhythm

It is also well-known that orphan nuclear receptors (RORα, RORɣt) have indispensable role in generation and maturation of Th17 cells. As a confirmation of involvement of retinoic acid receptors in MM (also in MGUS, section 9.1), enriched function of *RORA* in circadian function in pathway analysis was observed. Contextually cytochrome P450 Family 26 Subfamily B Member 1 (*CYP2R1*) was identified to have moderate differential expression in the eQTL and SMR analysis in MM plasma cells. This gene encodes a member of the cytochrome P450 superfamily of enzyme, a vitamin D hydroxylase which converts vitamin

D3 in the vesicle membrane to 25-hydroxyvitamin D3 [25(OH)D3], an active ligand for vitamin D receptor and an inverse agonist of *RORA* reducing receptor activation (Cheng *et al.*, 2018). Additional signal implicating the same mechanism, nuclear receptor super family member REV-ERBα was consecutively found enriched in circadian expression mediation. REV-ERBs (α and β) are often co-expressed in the same tissue as RORs that bind to the same sites and co-regulate shared target genes (Solt *et al.*, 2017). As Th17 cell differentiation is also regulated by circadian clock, all of the evidence on regulation of receptor activity crucial to retinoic acid dependent mediation of circadian clock indicate it having more than the impact previously described in MM pathogenesis (Yu *et al.*, 2013).

Transforming growth factor β

A separate signaling cascade that entails major influence on immunoglobulin trait modulation, Th17 cell differentiation and bone morphogenesis is the transforming growth factor β (TGFβ) pathway (David and Massagué, 2018), and probably not so surprisingly was represented by three different enriched pathways in MM. Enhanced bone resorption in MM releases and activates TGFβ, which is a potent inhibitor of osteoblast differentiation and mineralization (Takeuchi *et al.*, 2010). Interaction analysis identified an intergenic variant rs2834882 corresponding to runt related transcription factor 1 (*RUNX1*) in interaction with rs2860107 at 9q21.33 annotated to zinc finger CCHC-type containing 6 (*ZCCHC6*, alias *TUT7*). Activities of RUNX family member transcription factors have been linked to retinoic acid signaling and TGFβ-induced IgA class switching which is involved in MM pathogenesis (Jonsson *et al.*, 2017; Takeuchi *et al.*, 2010). While *ZCCHC6* and *ZCCHC11* based TUTase inhibitors are being investigated for management lymphoid malignancies as

potential agents for targeted therapy alluding a therapeutic connection (Lin and Gregory, 2015), runt proteins demonstrate implications in vastly diverse biological processes related to MM. Transcription factors of runt domain are integral components of one of the two TGFβ family member-imposed signaling cascades including bone morphogenic proteins (BMPs). Both *RUNX1* and *RUNX2* are established regulators of BMP-2/7/9-induced osteoblast differentiation. Both of these genes are often found co-expressed in skeletal elements that regulates expression of BMP-2, 9. Mis-regulation of these induces osteogenic differentiation of mesenchymal cells and in MM, causes growth arrest and anemia (Lagler *et al.*, 2017; Ludwig, 2010). Function of these runt domain transcription factors are even more implicated by the SMR analysis. RUNX2 regulatory activity in osteoblast differentiation is regulated by transcriptional repressor protein encoded by *NKX3-2*, causal eQTL for sentinel SNP rs17362130 (Caron *et al.*, 2013). Additionally, as these runt transcription factors are also transcriptional effectors of SMAD signaling (SMADs contribute to some of the most enriched pathways along with TGFβ receptor signaling for MM), these data suggest a broader role of TGFβ family signal transduction in MM.

Differential regulation of SMAD dependent TGFβ signaling pathway has been established to be vital for cancer since its prominent regulatory role in cell growth, differentiation and migration and its mis-regulation can result in tumorigenesis. Cancer cells can circumvent tumor-suppressive actions of TGFβ in two branches, either by recruiting other stromal cell types (myofibroblast, osteoclast) facilitating tumor spread or through silencing core components of the pathway, such as TGFβ receptors (Massagué, 2008). The TGFβ cytokine receptors phosphorylate SMAD2 and SMAD3 (alias R-SMADs) which bind to SMAD4 (alias Co-SMAD) to form hetero-trimer complex that constitutes the canonical SMAD-

dependent TGF-beta signaling cascade whereas those of the other branch phosphorylate SMAD1, 5 and 8 to create other R-SMAD/Co-SMAD complexes that bind to transcription factors in order to regulate transcription of target genes. It is traditionally believed that the first group of hetero-trimer is responsible for canonical TGFβ pathway regulation (also sometimes non-canonical, SMAD-independent TGF-beta signaling pathways) and BMPs are responsible for signaling via Smad1, 5, 8-phosphorylation (Wakefield and Hill, 2013). TGFβ –activated SMADs promote growth inhibition in epithelial progenitor cells, apoptosis in pre-malignant cells and induce metastatic invasion in cancer cells (David and Massagué, 2018). Contextually in MM, TGFβ induces differentiation arrest in osteoblasts, increases osteoclast genesis enhancing MM cell growth and survival, promote angiogenesis suppressing host immunity in bone marrow microenvironment to create the so called MM niche (Takeuchi *et al.*, 2010).

Histone Deacetylase

In another context relevant to MM biology, SMADs also interact with chromatin binding proteins HDAC1 and HDAC2. *HDAC1* is a class I histone deacetylase gene and MM patients with high protein levels of HDAC1 were shown to have poor progression-free and overall survival (Mithraprabhu *et al.*, 2014). Moreover inhibition of *HDAC1* expression induces MM cell death (Mithraprabhu *et al.*, 2013). Interaction analysis identified a significant interaction pair including class II HDAC family member, *HDAC9* and *NCAM2*. Aberrant mutation and high gene expression of *HDAC9* in cells of lymphoid lineage is believed to induce B-cell lymphoproliferative disorders including Waldenström macroglobulinemia and is associated with general poor progression in cancer (Sun *et al.*,

2011), additionally deregulation of expression of *HDAC9* in B-cells upholds lymphoma and lymphoproliferative neoplastic growth (Gil *et al.*, 2016). *HDAC9* is in addition assumed to be accountable for lymphomagenesis by regulation of growth and survival related pathways and by modulation of BCL6 and p53 tumor suppressor activity (Gil *et al.*, 2016). In germinal cells *HDAC9* is often co-expressed with *BCL6*, a novel therapeutic target for MM (Hideshima *et al.*, 2009). As HDACs in general pose a vital role in cell cycle arrest induction and activation of intrinsic apoptotic mechanism, it's a fair speculation that the common variation observed in 7p21.1 may be construed to predispose to MM pathogenesis.

MGUS risk loci in context of MM

The strongest signal from the MGUS study was found between *TNC* and *CRYL1*. Expression of *TNC* is found upregulated in certain MM cell lines in the presence of mutations in insulin growth factor receptor and receptor tyrosine-protein kinase genes (Leich *et al.*, 2013b). Additionally, the recurrent significant interaction was found between several non-unique loci annotated respectively to *SETBP1* and *PREX1*. The common variation at 20q13.13 predisposes to MM as an expression and methylation quantitative trait locus at *PREX1* (Mitchell *et al.*, 2016). *PREX1* is also shown to have abundant expression in peripheral blood leukocytes and moderately in lymph nodes, and much weaker in most other tissues deregulation in which was alluded to MM pathogenesis (Welch *et al.*, 2002). *ERBB4* harbored two significant interactions and parallel to the pathway enrichment results that identified role of *EGFR* downregulation, was established to be one of the most compelling finding in this study.

Although gene amplification can usually be associated with *EGFR* expression mis-regulation, around 20% of tested glioblastomas lacks ErbB family gene amplification (Tripp *et al.*, 2005), suggesting existence of other innate mechanisms in cancer cells that promote aberrant EGFR expression. Results on gene fusion confers aggregation of HER2 and HER3 with Growth factor receptor bound protein 7 (*GRB7*), Retinoic acid receptor alpha (*RARA*) and Ring Finger Protein 41 (*RNF41*). Amplification of *RARA* has been demonstrated in hematological malignancies of myeloid lineage and *RAR*α2 overexpression is related to progression, treatment efficacy and pathogenesis in MM (Asleson *et al.*, 2010; Pedersen-Bjergaard *et al.*, 2002). Chromosomal translocation t(15:17) is hypothesized to rearrange *RARA* and give rise to aberrant *EGFR* overexpression upstream to RAS activated pathway (Pedersen-Bjergaard *et al.*, 2002). *RNF41* encodes a ubiquitin ligase and maintains steady ErbB3 levels mediating its growth factor-independent degradation (Fry *et al.*, 2011). *GRB7* is a protein coding gene which although mostly associated with *ERBB2* amplification, is influential to signal transduction in response to external growth factor. *GRB7* also promotes activation of protein kinases important to regulation of MAPK pathway such as MAPK1/3, STAT1, and AKT1. Remarkably, GRB7 is also responsible for enrichment of cell surface interaction at vascular wall which is highly regulated by KRAS and NRAS mutations.

In contrast to the cell-cell adhesion mediators, the major transmembrane proteins at cell-ECM adhesions are integrin heterodimers. The ability of integrins to dictate cellular responses to a variety of inputs lies in their capacity to differentially recognize distinct environments. Several integrins, including integrin-β1, -β7 and -α8 have been shown to play crucial a role in maintenance of MM bone marrow niche drug resistance. β1-Integrin mediated adhesion of MM cells to fibronectin provides MM cells protection against drug-

induced apoptosis, triggers nuclear factor κB-dependent transcription and secrets interleukin-6 (*IL6*), a major growth factor for MM (Damiano and Dalton, 2000). A study on integrin-β7 mediated regulation of MM cells demonstrated its critical role in MM cell adhesion, migration, invasion and bone marrow homing (Damiano and Dalton, 2000). Another report on 16 relapsed MM patients shows highly expressed integrin-α8, newly discovered from gene expression profiling indicating towards EMT-like features of MM cells, causing migration, invasion and drug resistance (Jiyeon *et al.*, 2016).

RORA has effect on MM bone marrow microenvironment and bone homeostasis via integrin channels. It resides downstream to IL6 and TGFβ protein encoding genes and synergistically enforces lineage specification to uncommitted T helper cells into Th17 cells. It is already discussed how circadian rhythm, interleukins and T helper cells are speculated to co-predispose to MM. But in the current context, *RORA* is also shown to have interaction with hypoxia-inducible factor-1 alpha (HIF1-α) in regulation of activation and transcriptional activity, a potent mediator of intergrin-β1 and therapeutic target for MM (Muz *et al.*, 2014; Perrone *et al.*, 2011). Integrin cell surface interaction pathway was enriched for MGUS. It along with discovery of related risk loci indicate interplay between cell adhesion and integrin pathways which is additionally supported by discovery of the platelet aggregation (plug information) pathway, crucial to adhesion mechanism in platelet and a regulatory agent to integrin signaling.

Although it is very desirable to be able to explain disease burden commonality between MGUS and MM, a non-linear, possibly branched heterogeneous genetic progression binds several phases of MM disease family. Dorsoventral axis formation, KEGG autoimmune thyroid disease (hsa05330) and allograft rejection (hsa05320) pathways were discovered to

be enriched in all three pathway enrichment algorithms. Whereas the allograft rejection pathway is dependent on *EGFR*, *MAPK1-3*, and *NOTCH1-3*; and regulated by proto-oncogenes such as *KRAS* and *BRAF*; both allograft rejection and autoimmune thyroid disease pathways are downstream to B cell receptor signaling pathway. Exploring pathway regulation among MGUS, SMM and MM, one study asserted KEGG allograft rejection pathway to be uniformly enriched among MGUS and SMM cell lines (Dong *et al.*, 2015). Furthermore Demchenko *et al*. reported KEGG allograft rejection pathway and autoimmune thyroid disease pathway, both of which are enriched in the data, to be differentially regulated with most significance among all the MM cell lines tested (Demchenko *et al.*, 2010). Whilst RAS and BRAF family mutations and NOTCH pathways have been well-discussed in myeloma literature, all of these hierarchy hints at a functional dependency among MGUS and MM.

Algorithm novelty and computational efficiency

By assessing genome-wide interaction with hierarchical case/control and case-only data together with subsequent follow-ups, this computational protocol reduces brute-force search to a comparably smaller genomic regions increasing efficiency and power of detection. Using the correlation-based test statistics and subsequently extending the interrogation of discovered signals against network and single-marker linear association detected signals a workflow is implemented to integrate statistical findings with biological knowledge base. Streamlining detection of risk loci with enriched protein-protein interacting networks to discover differentially regulated novel pathways facilitates understanding of disease

mechanisms and accumulates statistical evidence with biologically interpretable information.

Genome-wide interaction analysis has thoroughly faced criticism in contemporary literature because of the computational burden needed to be handled keeping in mind the loss in genomic resolution due to the high number of tests. A ‘divide and conquer algorithm’ was used to tackle this problem. Rather than testing each variant against the other throughout the genome, the data sets were partitioned into 21 different sets corresponding to each of the 22 autosomal chromosomes and comprised of all the downstream variants starting from first variant of each of the corresponding chromosome. Detection tests were parallelized in 21 different loops. For an arbitrary chromosome A, the interaction tests were performed against all the SNPs corresponding to chromosome A against all the SNPs belonging to chromosomes A to 22 where A=1, 2, 3, ⋯, 22. Parallelizing the whole test space reduced the caveat of single run on large number of tests remarkably (**Figure 12.1**). As expected, significant reduction in computation time was observed in the parallelized algorithm compared to single run. A comprehensive prediction on computational time gain with a predictive simulation was also calculated. Treating time (computational time) as a dependent quantity solely explainable by number of tests performed, a polynomial fixed effects regression with intercept conforming to linearity assumption was simulated. Computational time prediction for a gross 28,133,824 tests (total number of pairs in discovery set) with the predictive model is approximately little less than 109 days compared to mere little over than an approximate 19 days for proposed algorithm.

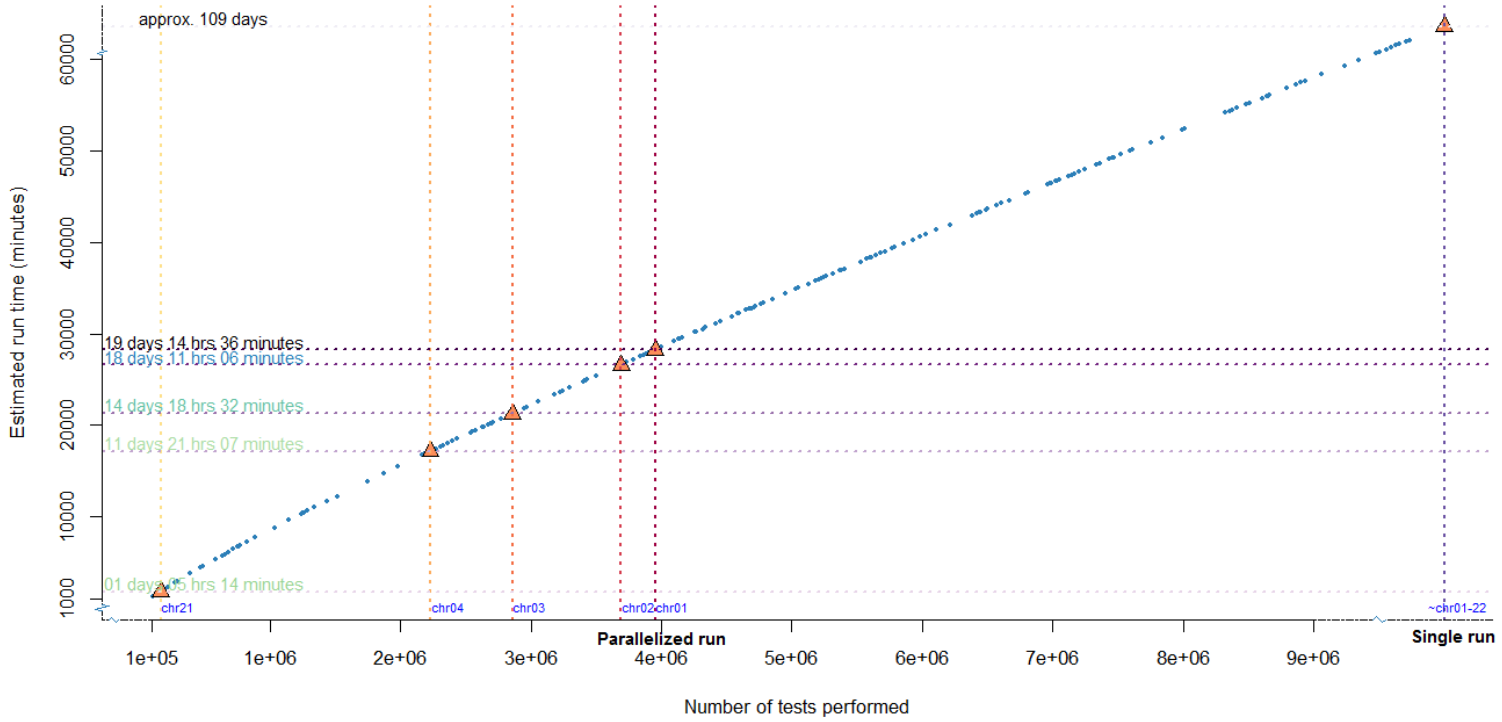


Figure 12. 1| Computational efficiency of parallelized algorithm
A single run reports a run time of approximately 109 days where as the parallelized run only take a little over than 19 days.

Limitation

The design of the analyses was developed in a way to have incorporated strict QC criteria in each individual step but certain caveats remain. Firstly, as genome wide-interaction is the main metric of investigation, the adequacy of sample size is a major concern. Although stringent QC and supervised selection mandated tests were performed, even a 0.1% level of significance would harbor a handful of false positive results. To counter this issue functional validation was of immense importance. The *in-silico* enrichment analyses provided convincing evidence referring to contextual biological processes but these results are not of causal nature and should be interpreted with caution. For better mechanistic understanding, confirmational studies need to be carried out with the repertoire of suggestive evidences provided here with before any definitive inference can be drawn. Secondly, inherited susceptibility to a phenotype is not always genetic. Pre/post-natal environment, Socio-economic environment and exposure to other environmental factors play a major role in a wide array of pathogenic development. Environmental exposure related MM predisposition is somewhat well-discussed but the same by no mean can be said true for MGUS due to its elusive asymptomatic nature. There are certain study designs for quantifying genetically heritable nature of a trait such as twin-studies, case-control studies in population with diverse ancestry. Such investigations are needed to understand (and to not overestimate) actuality of causal genetic aberrations in these diseases. Thirdly, although it is a multi-center study, the entire population is of European decent and the results need to be interpreted with this in mind. Arguably the recent development of reference panels from projects such as UK10K, HapMap consortium or 1000 Genomes provide data on densely genotyped panel primarily for western population implicating a bias towards experiment design. But none the less this is a major caveat in generalizability of the results. Additionally, association studies such as this are blind to causal regulators/enforcers that mask the genotypic effect. Impact of

metabolism, enzyme, hormonal or other biological process often act as triggers that activate downstream mis-regulation. Signals in common SNPs in encoding domain affecting such causal elements can be impossible to detect if the activation is not automatic or the phenotype can stay dormant until certain inciting event occurs. It is therefore advisable that only cautious and conservative inference be drawn from these results until further studies prove the validity of the speculations with definitive evidence.

Conclusion

In conclusion, the findings provide further evidence that MGUS and MM are primarily different phases of a family of plasma cell disorders with inherited genetic susceptibility that contribute to excess risk via regulation of an assortment of regulatory networks and pathways. Novelty of the investigation is firstly in interrogating risk predisposition mechanisms under a true polygenic assumption where low risk variants are not asymptotically nominalized as low penetrant background noise, rather is observed in pairs for shared mutual susceptibility. Secondly, in harmonizing GWAS summary statistics driven pathway, tissue, cell enrichment results with interaction detected signals. If bias due to design and sample is granted adequately controlled for, this study demonstrates that true biological signals are likely to overlap due to co-inheritance alluding to mechanistic link between risk loci and phenotype rather than elusive association signals where biological inference is left for the interpreter to speculate about. Thirdly the investigation identifies key regulators predisposing to MGUS and MM pathogenesis.

Chapter 14: Inherited risk of SPCs in MM patients

As the average survival of MM patients prolongs with change in therapeutics and management, diagnosis of SPCs has become more frequent and this study provides architecture for inherited familial risk of such SPCs. It also demonstrates that over the past few decades there has not been a population drift in diagnosis of MM and SPCs that can substantiate bias in observed effect due to age-period interaction in time of diagnosis. The Swedish study cohort of this analysis observed as high as 68.3% MM patients with a family history of cancer to have diagnosis of a SPC compared to that of 59.9% of those without such a family history. A recent study on the same cohort (with data until end of 2012) analyzed familial clustering of cancers with MM and reported almost uniform excess risk accumulation of MM with colorectal, prostate and some other cancer types while stratified over sex and type of first degree relative (parent / sibling) with the index caces (Frank *et al.*, 2015). For MM patients with family members having same cancer as was diagnosed as SPC in the patients, excess significant risk was also observed for colorectal, prostate and squamous cell skin cancer (section 11.3). This probably substantiates that genealogically inherited shared susceptibility manifests in already immunocompromised MM patients to develop subsequent primary cancer(s).

The second novel observation was influence of SPC on survival of MM patients. It was shown that 60.6% of all SPC patients died by the end of 2015 following a moderately worse survival in comparison to those without, of whom 55.2% had died. Although having a family history of cancer did not increase mortality (59.3% dead), this may be owing to the small sample size of SPC patients even in a large nation-wide cohort. Additionally, the rate of survival in MM although increasing, is still relatively poor thereby affecting the time-

window to have a diagnosis of SPC and maturity of follow-up for the patient cohort. Due to such caveats in sample size further analysis of hazard and survival or risk stratification on covariates was not carried out. In addition, the cancer register lacks data on behavioral patterns, clinical data on diagnosis and other possible variables that could have been treated as explanatory factors reducing probability of confounding as well as adjusting for inherent variation in data.

Therapy related SPCs have been largely characterized in MM in several studies many of which recorded increased risk of t/s-acute myeloid leukemia (therapy related/secondary) and this trend overlaps with the recent finding in the Swedish cohort (Chen *et al.*, 2016; Musto *et al.*, 2018). And part of the study cohort that overlaps with the cohort used in that study shows a more than four-fold increased risk of leukemia (non-familial patients). Additionally, recently effect of family history on development of SPC has been investigated for a large number of hematological malignancies where a similar familial clustering is also observed (Chattopadhyay *et al.*, 2018b; Chattopadhyay *et al.*, 2018d; Sud *et al.*, 2017b). Although therapy-induced deleterious effects are still considered weak in MM and proportion of diagnosis of SPC or secondary cancers are not in much excess (Yang *et al.*, 2012), one can speculate that this picture will change when a larger group of patients will achieve longer survival (Musto *et al.*, 2018). With continued therapeutic successes in MM management SPCs will be receiving increasing attention and this study shows family history information is crucial in evaluating strategies for long-term follow-up and possible screening of all patients with MM.

Main findings of the study

There are certain novel insights that this study provides the reader with:

1. Inherited susceptibility to MGUS and MM is possibly truly polygenic where temporal aggregation of risk due to co-inherited common variations contributes to predisposition.
2. In MGUS, common SNPs in cancer predisposing domains (*SETBP1, ALK*) were found in interaction with MM risk loci such as *PREX1*or previously discovered MGUS risk loci *GLCCTI1* to exert excess inherited risk.
3. Two major pathways are differentially regulated in MGUS; B cell receptor signaling and EGFR downregulation pathway. The identified risk loci in interaction analyses have involvement in these two pathways that may highlight further underlying mechanisms.
4. MM is also influenced by polygenic inherited risk intrinsically exerted by co-inherited common SNPs in interaction. Strongest signals were observed from genes that have key roles in cellular growth, differentiation, survival and apoptosis.
5. *GNAQ, IRF8, PDE3B, ZNF229, RUNX1, HDAC9* are some of the signals discovered in interaction that predispose to MM. These loci collectively play crucial role in MM biology via processes such as and not restricted to Ig trait modulation, osteoclast genesis, T helper cell development, interleukin secretion, bone marrow microenvironment mediation in creating MM niche.
6. Two dominant signaling cascades were identified to have shown that TGFβ signaling through its signal transducers SMADs and circadian rhythm regulation by RORA as well as REV-ERBA influence Ig class switch recombination, Th17 cell differentiation

and bone morphogenesis and may provide a mechanistic link between the predisposition markers of MGUS and intrinsic biology of progression to MM.

7. Family history of cancer makes MM patients susceptible to development of SPCs. As cancer predominantly is a disease developed due to accumulation of genetic aberrations, patients with prior history of cancer in family probably inherit certain genetic variations and/or share detrimental exposure of environmental factors which render them prone to develop subsequent primary cancers.

8. Concordant family history of leukemia, lung, squamous cell skin cancer and melanoma increases risk of SPC at the same site in MM patients by more than 5 folds compared to the patients without; whereas that for colorectal cancer is little over than 2 fold and for prostate cancer is 1.6 fold.

9. Beside squamous cell skin cancer, there was hardly any evidence found for more than additive or multiplicative interaction between individual, family history of a cancer and MM.

10. Family history of cancer in MM patients with SPC does not alter mortality patterns. As overall survival in MM is gradually improving with better-quality management, this indicates that efforts in reducing SPC diagnosis by screening with family history information will have positive impact on survival of MM patients.

Outlook

The evidence of inherited genetic susceptibility to MGUS and MM observed in this study answers some key issues but also brings up some questions. Firstly, if there is non-random mechanism to the progression of MGUS to MM and that is also true with AL amyloidosis, what is the shared genetic origin of the three diseases? Secondly, although the pathogenic consequences of the three disorders are different and heterogeneous, how far apart are they? And thirdly, Since MGUS precedes MM and AL amyloidosis, how much genetic correlation is present among the patients?

The extension of the current study will try to address these questions initially by aggregating cohorts for these three disorders and analyzing genetic correlation as well as shared heritability. Then phenotypic dependency-corrected meta-analysis and differential enrichment would be applied to investigate further into the issue.

Summary

Monoclonal gammopathy of undetermined significance is the most common plasma cell dyscrasia present in as high as 3.2% of general population below 50 years of age and up to 6.6% for population aged 80 years or older. It is a premalignant precursor of multiple myeloma, a malignant hematological neoplasia. People with monoclonal gammopathy go on to develop myeloma at a yearly rate of 0.5 - 1%. With a crude rate of incidence of 6.5 per 100,000 people, Europe is set to observe around 48,000 new multiple myeloma diagnosis in 2018. Overall prognosis of myeloma has not been very favorable throughout history nonetheless survival of myeloma patients is improving incrementally over the past few decades due to better management and improved treatment modality. This increased survival led to an increased number of second primary cancer diagnosis. Environmental factors, chemotherapy and radiotherapy induced DNA damage, wide-spread use of alkylating agents and possible induction of immunosuppressed state has been speculated to contribute to this. The fact that both the two diseases show familial clustering and all myeloma diagnoses are preceded by monoclonal gammopathy indicates that there is a certain amount of inherited susceptibility to these diseases. In the current study, the quantity under investigation is inherited genetic susceptibility to monoclonal gammopathy and multiple myeloma as well as the familial risk of second cancers.

Three sets were queried for monoclonal gammopathy consisting genotype data on 243, 82 and 326 German individuals respectively identified during routine follow-up of unrelated condition. These three sets were used to carry out separate case-control and case-only discovery, validation and replication studies. For myeloma, patients were recruited from two separate trials in Germany and UK. The German trial consisted of 1717 myeloma patients where as the one in UK recruited 2282 patients. Controls for the investigations were obtained from Heinz-Nixdorf Recall study samples and Welcome Trust Case-Control Consortium samples. For expression quantitative trait analysis, gene expression data was obtained from plasma cell samples of 665 patients enrolled in the German trial. Written consents were obtained from the trial subjects and approval for the studies was procured from respective ethics review board. For the observational study of second cancers, the Swedish Family Cancer Database was used which includes data on all cancer diagnosis in Sweden starting 1958. This database was queried for information on about 2.1 million Swedish residents with cancers matched with their biological parents (when available).

The interaction analyses with genotype data identified a number of paired susceptibility loci for monoclonal gammopathy and myeloma. These loci were found to have key roles in myeloma biology via processes such as and not restricted to Ig trait modulation, osteoclast genesis, Th cell development, interleukin secretion, bone marrow microenvironment mediation in creating myeloma niche. While subjected to enrichment analyses major biological pathways were discovered including EGFR downregulation and B cell receptor signaling pathway for monoclonal gammopathy and Circadian rhythm mediation and SMAD dependent TGFβ activation pathways in myeloma. As some of the pathways and loci were shown shared between monoclonal gammopathy and myeloma, the findings allude to shared inherited susceptibility to the two disorders. Interrogating risk of second cancer in myeloma patients stratified by history of cancer among first degree relatives, numerous cancers were noted to have excess familial risk and overall close to a 1.4-fold increased risk of second cancer was noted among people with an existing family history. Concordant family history of leukemia, lung, squamous cell skin cancer and melanoma increased risk of second cancers at the same site in myeloma patients by more than 5 folds compared to the patients without; whereas that for colorectal cancer is little over than 2-fold and for prostate cancer is 1.6-fold. Although family history was found to have a strong effect on incidence of second cancers no such effect was found in mortality pattern. No linear or multiplicative interaction was found in risks among personal, family history with history of myeloma.

All the results indicate there are certain underlying mechanistic principle relating monoclonal gammopathy to myeloma which is regulated by inherited polygenic predisposition to monoclonal gammopathy and myeloma. This study speculates about possible pathways and networks that are influenced in these diseases but conformational studies need to be carried out before any definitive conclusion can be drawn. However, in context of second cancers in myeloma patients, family history of cancer was conclusively shown to have morbid impact on incidence but lack of any such impact on patient survival was also observed which mean efforts in managing second cancer diagnosis by screening with family history information will have positive impact on survival in multiple myeloma.

Zusammenfassung

Die Monoklonale Gammopathie unklarer Signifikanz ist mit einer Prävalenz von 3,2% in der Bevölkerung unter 50 Jahren beziehungsweise 6,6% in der Bevölkerung über 80 Jahren die häufigste Plasmazellerkrankung. Diese Krankheit ist die Vorstufe des Multiplen Myeloms, einer malignen hämatologischen Neoplasie. Menschen mit MGUS entwickeln MM mit einer jährlichen Rate von 0,5 - 1%.Mit einer Inzidenz von 6,5/100.000 sind im Jahr 2018 in Europa 48.000 Neudiagnosen des Multiplen Myeloms zu erwarten. Obwohl MM immer noch eine tödliche krankheit ist, hat sich die Prognose von Patienten mit Multiplem Myelom in den letzten Jahrzehnten durch die Entwicklung neuer Behandlungsmodalitäten kontinuierlich verbessert. Die erhöhte Überlebensrate hat zu einer erhöhten Diagnoserate von zweiten Primärtumoren (second primary cancer) in Myelom-Patienten geführt. Es wird spekuliert, dass verschiedene Umweltfaktoren, Chemo- und Radiotherapie-induzierte DNA-Schäden und eine mögliche Induktion eines immunsupprimierten Zustands dazu beitragen. Die Tatsache, dass sowohl die Monoklonale Gammopathie unklarer Signifikanz als auch das Multiple Myelom familiär gehäuft auftreten und alle Myelom-Diagnosen aus einer Monoklonalen Gammopathie unklarer Signifikanz hervorgehen, weist darauf hin, dass es eine gewisse erbliche Anfälligkeit für diese Krankheiten gibt. In dieser Studie wird diese genetisch bedingte Anfälligkeit für die Monoklonale Gammopathie unklarer Signifikanz und das Multiple Myelom sowie das familiäre Risiko eines nach dem Multiplen Myelom auftretenden zweiten Primärtumors untersucht.

Für die Studie der Monoklonale Gammopathie unklarer Signifikanz wurden drei Datensätze mit Daten aus den genomweiten Assoziationsstudien analyziert, die aus 243, 82 und 328 Personen deutscher Herkunft bestanden. Die Identifikation dieser Personen erfolgte durch routinemäßige Untersuchung eines nicht zusammenhängenden Zustands. Diese drei Datensätze wurden angewandt, um Fall-(Kontroll)-Beobachtungs-, Validierungs- und Replikationsstudien durchzuführen. Für das Multiple Myelom wurden Patienten aus zwei separaten Studien in Deutschland und Großbritannien rekrutiert. Die deutsche Studie bestand aus 1717 und die englische aus 2282 Patienten. Die Kontrolldaten wurden von der Heinz-Nixdorf-Recall-Studie für den deutschen Datensatz und von dem Welcome-Trust-Case-Control-Consortium für den englischen Datensatz erhalten. Für Expression Quantitative Trait Analysis wurden Gen-Expressionsdaten von Plasmazellproben von 656 Patienten, eingebunden in der deutschen Studie, verwendet. Schriftliche Einverständniserklärungen wurden von den Studienteilnehmern erhalten. Die Zulassung der Studien waren von den jeweiligen Ethikkomitees geboten. Für die Beobachtungsstudie von zweiten Primärtumoren wurde die Swedish Family Cancer Database verwendet, welche Daten über alle Krebsdiagnosen ab 1958 in Schweden umfasst. Diese Datenbank wurde für Informationen von circa 2,1

Millionen schwedischen, an Krebs erkrankten Einwohnern mit Einbindung von Informationen der biologischen Eltern (wenn vorhanden) befragt.

Die Interaktionsanalyse mit Genotyp-Informationen identifizierte gepaarte Suszeptibilitäts-Loci für die Monoklonale Gammopathie unklarer Signifikanz und das Multiple Myelom. Es wurde festgestellt, dass diese Loci eine Schlüsselrolle in der Biologie des Multiplen Myeloms unter anderem über die folgende Prozesse spielen: Immunglobulinproduktion, Osteoklastengenese, Entwicklung von T-Helferzellen, Interleukinsekretion und Knochenmark-Mikroumgebung,. Die Anreicherungsanalysen zeigten wichtige biologische Pathways wie die EGFR-Downregulation und den B-Zell-Rezeptor-Signalweg bei der Monoklonalen Gammopathie unklarer Signifikanz sowie die Mediation des zirkadianen Rhythmus und SMAD-abhängige TGFβ-Aktivierungswege beim Multiplen Myelom. Da einige dieser Pathways und Genloci sowohl bei der Monoklonalen Gammopathie unklarer Signifikanz, als auch beim Multiplen Myelom gefunden werden konnten, legen die Ergebnisse nahe, dass eine gemeinsame vererbbare Suszeptibilität beider Krankheiten besteht. Die Risikoabschätzung von zweiten Primärtumoren in Multiplen-Myelom-Patienten, stratifiziert nach der Krebsvorgeschichte unter den Verwandten ersten Grades, zeigten einige Krebsarten mit einem erhöhten familiären Risiko. Insgesamt bestand bei den Individuuen familiärer Vorbelastung ein 1,4-fach erhöhtes Risiko, einen zweiten Primärtumor zu entwickeln. Die onkolgische Familiengeschichte von Leukämie, Lungenkrebs, Plattenepithelkarzinom der Haut und vom Melanom führte zu einem jeweils fünffach-erhöhten Risiko eines konkordanten zweiten Primärtumors in Multiplen-Myelom-Patienten im Vergleich zu Patienten ohne familiäre Vorbelastung. Für Kolorektalkrebs war dieses Risiko mehr als zweifach und für Prostatakrebs 1,6-fach erhöht. Auch wenn die familiäre Krebsvorgeschichte die Inzidenz eines zweiten Primärtumors erhöht, nimmt sie keinen Einfluss auf die Mortalität. Zudem wurde keine lineare oder multiplikative Wechselwirkung zwischen individueller Krebsvorbelastung, familiärer Krebsvorgeschichte und der Diagnose des Multiplen Myeloms gefunden.

Die Ergebnisse weisen darauf hin, dass die grundlegenden Mechanismen, welche die Monoklonale Gammopathie unklarer Signifikanz mit dem Multiplen Myelom verbinden, von der entsprechenden vererbbaren, polygenetischen Prädisposition abhängen. Diese Studie spekuliert über mögliche Pathways und Netzwerke, die in diesen Krankheiten verändert sind. Bevor jedoch endgültige Schlussfolgerungen gemacht werden können, sind weitere Bestätigungsstudien notwendig. Dennoch konnte abschließend gezeigt werden, dass eine familiäre Krebsvorbelastung bei Multiplen-Myelom-Patienten mit zweiten Primärtumoren einen Einfluss auf die Inzidenz, aber nicht auf das Patientenüberleben hat. Das Ziel ist es, durch ein verbessertes Patientenscreening die Früherkennung eines zweiten Primärtumors zu gewährleisten, und dadurch die Überlebenschancen von Multiplen-Myelom-Patienten zu verbessern.

Bibliography

Adami, J.*, et al.* (1998). **Smoking and the risk of leukemia, lymphoma, and multiple myeloma (Sweden)**. Cancer Causes & Control *9*, 49-56, doi: 10.1023/A:1008897203337.

Aerts, S.*, et al.* (2006). **Gene prioritization through genomic data fusion**. Nature Biotechnology *24*, 537, doi: 10.1038/nbt1203
https://www.nature.com/articles/nbt1203#supplementary-information.

Alexander, D. D.*, et al.* (2007). **Multiple myeloma: A review of the epidemiologic literature**. International Journal of Cancer *120*, 40-61, doi: 10.1002/ijc.22718.

Alkebsi, L.*, et al.* (2016). **Chromosome 16q genes CDH1, CDH13 and ADAMTS18 are correlated and frequently methylated in human lymphoma**. Oncol Lett *12*, 3523-3530, doi: 10.3892/ol.2016.5116.

Altekruse, S. F.*, et al.* (1999a). **Deaths from hematopoietic and other cancers in relation to permanent hair dye use in a large prospective study (United States)**. Cancer Causes & Control *10*, 617-625, doi: Doi 10.1023/A:1008926027805.

Altekruse, S. F.*, et al.* (1999b). **Deaths from hematopoietic and other cancers in relation to permanent hair dye use in a large prospective study (United States)**. Cancer Causes Control *10*, 617-625.

Altieri, A.*, et al.* (2006). **Familial risks and temporal incidence trends of multiple myeloma**. European Journal of Cancer *42*, 1661-1670, doi: https://doi.org/10.1016/j.ejca.2005.11.033.

Altshuler, D.*, et al.* (2008). **Genetic Mapping in Human Disease**. Science *322*, 881.

Anderson, C. A.*, et al.* (2010). **Data quality control in genetic case-control association studies**. Nature Protocols *5*, 1564, doi: 10.1038/nprot.2010.116
https://www.nature.com/articles/nprot.2010.116#supplementary-information.

Anderson, D. E. (1974). **Genetic study of breast cancer: identification of a high risk group**. Cancer *34*, 1090-1097.

Anonymous (2000). **Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases**. British Journal Of Cancer *83*, 1301, doi: 10.1054/bjoc.2000.1407.

Anonymous (2003). **Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group**. British Journal of Haematology *121*, 749-757, doi: 10.1046/j.1365-2141.2003.04355.x.

Anonymous (2011). **Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies**. The Lancet *377*, 641-649, doi: https://doi.org/10.1016/S0140-6736(10)62345-8.

Anonymous (2013). **Treatment for High-Risk Smoldering Myeloma**. New England Journal of Medicine *369*, 1762-1765, doi: 10.1056/NEJMc1310911.

Aronson, L. I.*, et al.* (2014). **Characterization of RAS Alterations in Myeloma: Why Direct Targeting of RAS May be the Most Appropriate Therapeutic Approach**. Blood *124*, 643.

Ashburner, M.*, et al.* (2000). **Gene Ontology: tool for the unification of biology**. Nature Genetics *25*, 25, doi: 10.1038/75556.

Asleson, A. D.*, et al.* (2010). **Amplification of the RARA gene in acute myeloid leukemia: significant finding or coincidental observation?** Cancer Genetics and Cytogenetics *202*, 33-37, doi: https://doi.org/10.1016/j.cancergencyto.2010.06.003.

Baldini, L.*, et al.* (1996). **Role of different hematologic variables in defining the risk of malignant transformation in monoclonal gammopathy**. Blood *87*, 912.

Benn, M. and Nordestgaard, B. G. (2018). **From genome-wide association studies to Mendelian randomization: novel opportunities for understanding cardiovascular disease causality, pathogenesis, prevention, and treatment**. Cardiovascular Research *114*, 1192-1208, doi: 10.1093/cvr/cvy045.

Bergsagel, D. E.*, et al.* (1979). **The Chemotherapy of Plasma-Cell Myeloma and the Incidence of Acute Leukemia**. New England Journal of Medicine *301*, 743-748, doi: 10.1056/NEJM197910043011402.

Bergsagel, D. E.*, et al.* (1999). **Benzene and multiple myeloma: appraisal of the scientific evidence**. Blood *94*, 1174-1182.

Bernstein, J. L.*, et al.* (1992). **The genetic epidemiology of second primary breast cancer**. Am J Epidemiol *136*, 937-948.

Bladé, J.*, et al.* (2009). **Are all myelomas preceded by MGUS?** Blood *113*, 5370.

Blair, C. K.*, et al.* (2005). **Anthropometric Characteristics and Risk of Multiple Myeloma**. Epidemiology *16*, 691-694.

Blum, A.*, et al.* (2018). **Smoldering multiple myeloma: prevalence and current evidence guiding treatment decisions**. Blood and Lymphatic Cancer: Targets and Therapy *8*, 21-31.

Bodmer, W. F.*, et al.* (1987). **Localization of the gene for familial adenomatous polyposis on chromosome 5**. Nature *328*, 614, doi: 10.1038/328614a0.

Boffetta, P.*, et al.* (2008). **Exposure to ultraviolet radiation and risk of malignant lymphoma and multiple myeloma—a multicentre European case–control study**. International Journal of Epidemiology *37*, 1080-1094, doi: 10.1093/ije/dyn092.

Boice, J. D.*, et al.* (1991). **Diagnostic x-ray procedures and risk of leukemia, lymphoma, and multiple myeloma**. JAMA *265*, 1290-1294, doi: 10.1001/jama.1991.03460100092031.

Bourguet, C. C., *et al.* (1985). **Multiple myeloma and family history of cancer a case—control study**. Cancer *56*, 2133-2139, doi: 10.1002/1097-0142(19851015)56:8<2133::AID-CNCR2820560842>3.0.CO;2-F.

Boursi, B., *et al.* (2016). **Reappraisal of risk factors for monoclonal gammopathy of undetermined significance**. American Journal of Hematology *91*, 581-584, doi: 10.1002/ajh.24355.

Brandt-Rauf, P. W., *et al.* (1988). **Health hazards of fire fighters: exposure assessment**. British Journal of Industrial Medicine *45*, 606.

Broderick, P., *et al.* (2011). **Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk**. Nat Genet *44*, 58-61, doi: 10.1038/ng.993.

Brown, L. M., *et al.* (1992). **Alcohol consumption and risk of leukemia, non-Hodgkin's lymphoma, and multiple myeloma**. Leukemia Research *16*, 979-984, doi: 10.1016/0145-2126(92)90077-K.

Brown, L. M., *et al.* (2000). **Multiple myeloma and family history of cancer among blacks and whites in the U.S**. Cancer *85*, 2385-2390, doi: 10.1002/(SICI)1097-0142(19990601)85:11<2385::AID-CNCR13>3.0.CO;2-A.

Browning, S. R. and Browning, B. L. (2011). **Haplotype phasing: existing methods and new developments**. Nature Reviews Genetics *12*, 703, doi: 10.1038/nrg3054.

Brownson, R. C. (1991). **Cigarette Smoking and Risk of Myeloma**. JNCI: Journal of the National Cancer Institute *83*, 1036-1037, doi: 10.1093/jnci/83.14.1036.

Bulik-Sullivan, B., *et al.* (2015a). **An atlas of genetic correlations across human diseases and traits**. Nature Genetics *47*, 1236, doi: 10.1038/ng.3406 https://www.nature.com/articles/ng.3406#supplementary-information.

Bulik-Sullivan, B. K., *et al.* (2015b). **LD Score regression distinguishes confounding from polygenicity in genome-wide association studies**. Nature Genetics *47*, 291, doi: 10.1038/ng.3211 https://www.nature.com/articles/ng.3211#supplementary-information.

Burmeister, L. F. (1981a). **Cancer Mortality In Iowa Farmers, 1971–78**. JNCI: Journal of the National Cancer Institute *66*, 461-464, doi: 10.1093/jnci/66.3.461.

Burmeister, L. F. (1981b). **Cancer Mortality In Iowa Farmers, 1971–782**. JNCI: Journal of the National Cancer Institute *66*, 461-464, doi: 10.1093/jnci/66.3.461.

Bycroft, C., *et al.* (2017). **Genome-wide genetic data on ~500,000 UK Biobank participants**. bioRxiv.

Califano, A., *et al.* (2012). **Leveraging models of cell regulation and GWAS data in integrative network-based association studies**. Nature Genetics *44*, 841, doi: 10.1038/ng.2355.

Calle, E. E., *et al.* (2003). **Overweight, Obesity, and Mortality from Cancer in a Prospectively Studied Cohort of U.S. Adults**. New England Journal of Medicine *348*, 1625-1638, doi: 10.1056/NEJMoa021423.

Campbell, C. D., *et al.* (2005). **Demonstrating stratification in a European American population**. Nature Genetics *37*, 868, doi: 10.1038/ng1607.

Cannon-Albright, L. A., *et al.* (1992). **Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22**. Science *258*, 1148.

Cardon, L. R. and Palmer, L. J. (2003). **Population stratification and spurious allelic association**. The Lancet *361*, 598-604, doi: https://doi.org/10.1016/S0140-6736(03)12520-2.

Caron, M. M. J., *et al.* (2013). **Hypertrophic differentiation during chondrogenic differentiation of progenitor cells is stimulated by BMP-2 but suppressed by BMP-7**. Osteoarthritis and Cartilage *21*, 604-613, doi: https://doi.org/10.1016/j.joca.2013.01.009.

Carstensen, B. (2006). **Demography and epidemiology: Age-Period-Cohort models in the computer age**, Department of Biostatistics, University of Copenhagen Copenhagen.

CentreforEpidemiology (2013). **Cancer incidence in Sweden 2012**, The National Board of Health and Welfare, Stockholm.

Cesana, C., *et al.* (2002). **Prognostic Factors for Malignant Transformation in Monoclonal Gammopathy of Undetermined Significance and Smoldering Multiple Myeloma**. Journal of Clinical Oncology *20*, 1625-1634, doi: 10.1200/JCO.2002.20.6.1625.

Chan, Y., *et al.* (2015). **Genome-wide Analysis of Body Proportion Classifies Height-Associated Variants by Mechanism of Action and Implicates Genes Important for Skeletal Development**. The American Journal of Human Genetics *96*, 695-708, doi: https://doi.org/10.1016/j.ajhg.2015.02.018.

Chang, C. C., *et al.* (2015). **Second-generation PLINK: rising to the challenge of larger and richer datasets**. GigaScience *4*, s13742-13015-10047-13748-s13742-13015-10047-13748, doi: 10.1186/s13742-015-0047-8.

Chatenoud, L., *et al.* (1998). **Whole grain food intake and cancer risk**. International Journal of Cancer *77*, 24-28, doi: 10.1002/(SICI)1097-0215(19980703)77:1<24::AID-IJC5>3.0.CO;2-1.

Chattopadhyay, S., *et al.* (2018a). **Impact of family history of cancer on risk and mortality of second cancers in patients with prostate cancer**. Prostate Cancer and Prostatic Diseases, doi: 10.1038/s41391-018-0089-y.

Chattopadhyay, S., *et al.* (2018b). **Second primary cancers in non-Hodgkin lymphoma: Bidirectional analyses suggesting role for immune dysfunction**. International Journal of Cancer *0*, doi: 10.1002/ijc.31801.

Chattopadhyay, S., *et al.* (2018c). **Enrichment of B cell receptor signaling and epidermal growth factor receptor pathways in monoclonal gammopathy of undetermined significance: a genome-wide genetic interaction study**. Molecular Medicine *24*, 30, doi: 10.1186/s10020-018-0031-8.

Chattopadhyay, S., *et al.* (2018d). **Risk of second primary cancer following myeloid neoplasia and risk of myeloid neoplasia as second primary cancer: a nationwide, observational follow up study in Sweden**. The Lancet Haematology *5*, e368-e377, doi: https://doi.org/10.1016/S2352-3026(18)30108-X.

Chen, L. S., *et al.* (2010). **Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data**. The American Journal of Human Genetics *86*, 860-871, doi: https://doi.org/10.1016/j.ajhg.2010.04.014.

Chen, T., *et al.* (2016). **Risk of Second Primary Cancers in Multiple Myeloma Survivors in German and Swedish Cancer Registries**. Scientific Reports *6*, 22084, doi: 10.1038/srep22084.

Cheng, C. Y. S., *et al.* (2018). **Properties of purified CYP2R1 in a reconstituted membrane environment and its 25-hydroxylation of 20-hydroxyvitamin D3**. The Journal of Steroid Biochemistry and Molecular Biology *177*, 59-69, doi: https://doi.org/10.1016/j.jsbmb.2017.07.011.

Chng, W. J., *et al.* (2013). **IMWG consensus on risk stratification in multiple myeloma**. Leukemia *28*, 269, doi: 10.1038/leu.2013.247.

Chubb, D., *et al.* (2016). **Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer**. Nature Communications *7*, 11883, doi: 10.1038/ncomms11883 https://www.nature.com/articles/ncomms11883#supplementary-information.

Chubb, D., *et al.* (2013). **Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk**. Nat Genet *45*, 1221-1225, doi: 10.1038/ng.2733.

Clark, T. G., *et al.* (2003). **Survival Analysis Part I: Basic concepts and first analyses**. British Journal Of Cancer *89*, 232, doi: 10.1038/sj.bjc.6601118.

Clyde, D. (2017). **Transitioning from association to causation with eQTLs**. Nature Reviews Genetics *18*, 271, doi: 10.1038/nrg.2017.22.

Colodro-Conde, L., *et al.* (2018). **Association between population density and genetic risk for schizophrenia**. JAMA Psychiatry, doi: 10.1001/jamapsychiatry.2018.1581.

Coordinators, N. R. (2013). **Database resources of the National Center for Biotechnology Information**. Nucleic Acids Research *41*, D8-D20, doi: 10.1093/nar/gks1189.

Cordell, H. J. (2009). **Detecting gene–gene interactions that underlie human diseases**. Nature Reviews Genetics *10*, 392, doi: 10.1038/nrg2579 https://www.nature.com/articles/nrg2579#supplementary-information.

Coviello, V. and Boggess, M. (2004). **Cumulative incidence estimation in the presence of competing risks**. STATA journal *4*, 103-112.

Cowan, A. J., *et al.* (2018). **Global burden of multiple myeloma: A systematic analysis for the global burden of disease study 2016**. JAMA Oncology, doi: 10.1001/jamaoncol.2018.2128.

Daly, M. J., *et al.* (2001). **High-resolution haplotype structure in the human genome**. Nature Genetics *29*, 229, doi: 10.1038/ng1001-229.

Damiano, J. S. and Dalton, W. S. (2000). **Integrin-Mediated Drug Resistance in Multiple Myeloma**. Leukemia & Lymphoma *38*, 71-81, doi: 10.3109/10428190009060320.

Davey Smith, G. and Hemani, G. (2014). **Mendelian randomization: genetic anchors for causal inference in epidemiological studies**. Human Molecular Genetics *23*, R89-R98, doi: 10.1093/hmg/ddu328.

David, C. J. and Massagué, J. (2018). **Contextual determinants of TGFβ action in development, immunity and cancer**. Nature Reviews Molecular Cell Biology *19*, 419-435, doi: 10.1038/s41580-018-0007-0.

Delaneau, O., *et al.* (2014). **Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel**. Nature Communications *5*, 3934, doi: 10.1038/ncomms4934 https://www.nature.com/articles/ncomms4934#supplementary-information.

Delaneau, O., *et al.* (2011). **A linear complexity phasing method for thousands of genomes**. Nature Methods *9*, 179, doi: 10.1038/nmeth.1785 https://www.nature.com/articles/nmeth.1785#supplementary-information.

Delaneau, O., *et al.* (2012). **Improved whole-chromosome phasing for disease and population genetic studies**. Nature Methods *10*, 5, doi: 10.1038/nmeth.2307 https://www.nature.com/articles/nmeth.2307#supplementary-information.

Demchenko, Y. N., *et al.* (2010). **Classical and/or alternative NF-κB pathway activation in multiple myeloma**. Blood *115*, 3541.

Dong, L., *et al.* (2015). **Pathway-based network analysis of myeloma tumors: monoclonal gammopathy of unknown significance, smoldering multiple myeloma, and multiple myeloma**. Genet Mol Res *14*, 9571-9584, doi: 10.4238/2015.August.14.20.

Dring, A. M., *et al.* (2004). **A Global Expression-based Analysis of the Consequences of the t(4;14) Translocation in Myeloma**. Clinical Cancer Research *10*, 5692.

Durie, B. G. and Salmon, S. E. (1975). **A clinical staging system for multiple myeloma. Correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival**. Cancer *36*, 842-854.

Egan, J. B., *et al.* (2012). **Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides**. Blood *120*, 1060.

Eriksson, M. and Hallberg, B. (1992). **Familial occurrence of hematologic malignancies and other diseases in multiple myeloma: a case-control study**. Cancer Causes Control *3*, 63-67.

Fadista, J.*, et al.* (2016). **The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants**. European Journal Of Human Genetics *24*, 1202, doi: 10.1038/ejhg.2015.269 https://www.nature.com/articles/ejhg2015269#supplementary-information.

Fonseca, R.*, et al.* (2016). **Trends in overall survival and costs of multiple myeloma, 2000–2014**. Leukemia *31*, 1915, doi: 10.1038/leu.2016.380.

Frank, C.*, et al.* (2015). **Search for familial clustering of multiple myeloma with any cancer**. Leukemia *30*, 627, doi: 10.1038/leu.2015.279 https://www.nature.com/articles/leu2015279#supplementary-information.

Fritschi, L.*, et al.* (2004). **Dietary Fish Intake and Risk of Leukaemia, Multiple Myeloma, and Non-Hodgkin Lymphoma**. Cancer Epidemiology Biomarkers & Prevention *13*, 532.

Fritschi, L. and Siemiatycki, J. (1996). **Lymphoma, myeloma and occupation: Results of a case-control study**. International Journal of Cancer *67*, 498-503, doi: 10.1002/(SICI)1097-0215(19960807)67:4<498::AID-IJC6>3.0.CO;2-N.

Frodin, J. E.*, et al.* (1997). **Multiple primary malignant tumors in a national cancer registry--reliability of reporting**. Acta Oncol *36*, 465-469.

Fry, W. H.*, et al.* (2011). **Quantity control of the ErbB3 receptor tyrosine kinase at the endoplasmic reticulum**. Mol Cell Biol *31*, 3009-3018, doi: 10.1128/mcb.05105-11.

Gao, M.*, et al.* (2015). **Smoldering Multiple Myeloma**. BioMed Research International *2015*, 623254, doi: 10.1155/2015/623254.

Garnett, M. J.*, et al.* (2012). **Systematic identification of genomic markers of drug sensitivity in cancer cells**. Nature *483*, 570, doi: 10.1038/nature11005 https://www.nature.com/articles/nature11005#supplementary-information.

Geller, F.*, et al.* (2014). **Genome-wide association analyses identify variants in developmental genes associated with hypospadias**. Nature Genetics *46*, 957, doi: 10.1038/ng.3063 https://www.nature.com/articles/ng.3063#supplementary-information.

Geschickter, C. F. and Copeland, M. M. (1928). **Multiple myeloma**. Archives of Surgery *16*, 807-863, doi: 10.1001/archsurg.1928.01140040002001.

Ghosh, S. and Bouchard, C. (2017). **Convergence between biological, behavioural and genetic determinants of obesity**. Nature Reviews Genetics *18*, 731, doi: 10.1038/nrg.2017.72 https://www.nature.com/articles/nrg.2017.72#supplementary-information.

Gil, V. S.*, et al.* (2016). **Deregulated expression of HDAC9 in B cells promotes development of lymphoproliferative disease and lymphoma in mice**. Disease Models & Mechanisms *9*, 1483.

Gilad, Y., *et al.* (2008). **Revealing the architecture of gene regulation: the promise of eQTL studies**. Trends in Genetics *24*, 408-415, doi: 10.1016/j.tig.2008.06.001.

Global Lipids Genetics, C., *et al.* (2013). **Discovery and refinement of loci associated with lipid levels**. Nature Genetics *45*, 1274, doi: 10.1038/ng.2797 https://www.nature.com/articles/ng.2797#supplementary-information.

Goldstein, D. B. (2009). **Common Genetic Variation and Human Traits**. New England Journal of Medicine *360*, 1696-1698, doi: 10.1056/NEJMp0806284.

Gonsalves, W. I., *et al.* (2014). **Prognostic Significance of Quantifying Circulating Plasma Cells in Multiple Myeloma**. Clinical Lymphoma Myeloma and Leukemia *14*, S147, doi: 10.1016/j.clml.2014.06.087.

Greenberg, A. J., *et al.* (2012). **Single-nucleotide polymorphism rs1052501 associated with monoclonal gammopathy of undetermined significance and multiple myeloma**. Leukemia *27*, 515, doi: 10.1038/leu.2012.232.

Greipp, P. R., *et al.* (2005). **International Staging System for Multiple Myeloma**. Journal of Clinical Oncology *23*, 3412-3420, doi: 10.1200/JCO.2005.04.242.

Grufferman, S., *et al.* (1989). **Familial Aggregation of Multiple Myeloma and Central Nervous System Diseases**. Journal of the American Geriatrics Society *37*, 303-309, doi: 10.1111/j.1532-5415.1989.tb05495.x.

Hagner, P. R., *et al.* (2009). **Alcohol consumption and decreased risk of non-Hodgkin lymphoma: role of mTOR dysfunction**. Blood *113*, 5526.

Haibe-Kains, B., *et al.* (2012). **A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes**. JNCI: Journal of the National Cancer Institute *104*, 311-325, doi: 10.1093/jnci/djr545.

Hall, J. M., *et al.* (1990). **Linkage of early-onset familial breast cancer to chromosome 17q21**. Science *250*, 1684.

Hardy, J. and Singleton, A. (2009). **Genomewide Association Studies and Human Disease**. New England Journal of Medicine *360*, 1759-1768, doi: 10.1056/NEJMra0808700.

Hari, P. N., *et al.* (2009). **IS THE INTERNATIONAL STAGING SYSTEM SUPERIOR TO THE DURIE SALMON STAGING SYSTEM? A COMPARISON IN MULTIPLE MYELOMA PATIENTS UNDERGOING AUTOLOGOUS TRANSPLANT**. Leukemia *23*, 1528-1534, doi: 10.1038/leu.2009.61.

Hatcher, J. L., *et al.* (2001). **Diagnostic radiation and the risk of multiple myeloma (United States)**. Cancer Causes Control *12*, 755-761.

Hemani, G., *et al.* (2018). **The MR-Base platform supports systematic causal inference across the human phenome**. eLife *7*, e34408, doi: 10.7554/eLife.34408.

Hemminki, K., *et al.* (2009). **The Swedish Family-Cancer Database 2009: prospects for histology-specific and immigrant studies**. International Journal of Cancer *126*, 2259-2267, doi: 10.1002/ijc.24795.

Hemminki, K., *et al.* (2003). **Familial risk of cancer: Data for clinical counseling and cancer genetics**. International Journal of Cancer *108*, 109-114, doi: 10.1002/ijc.11478.

Hemminki, K. and Vaittinen, P. (1998). **National database of familial cancer in Sweden**. Genet Epidemiol *15*, 225-236, doi: 10.1002/(sici)1098-2272(1998)15:3<225::aid-gepi2>3.0.co;2-3.

Hemminki, X. L. K. P. C. G. P. V. K. (2001). **The Nation-wide Swedish Family-Cancer Database&Updated Structure and Familial Rates**. Acta Oncologica *40*, 772-777, doi: 10.1080/02841860152619214.

Hernán, M. A. and Robins, J. M. (2006). **Instruments for Causal Inference: An Epidemiologist's Dream?** Epidemiology *17*, 360-372, doi: 10.1097/01.ede.0000222409.00878.37.

Herold, C., *et al.* (2012). **Integrated Genome-Wide Pathway Association Analysis with INTERSNP**. Human Heredity *73*, 63-72.

Herold, C., *et al.* (2009). **INTERSNP: genome-wide interaction analysis guided by a priori information**. Bioinformatics *25*, 3275-3281, doi: 10.1093/bioinformatics/btp596.

Heuck, C. J., *et al.* (2014). **Five gene probes carry most of the discriminatory power of the 70-gene risk model in multiple myeloma**. Leukemia *28*, 2410, doi: 10.1038/leu.2014.232 https://www.nature.com/articles/leu2014232#supplementary-information.

Hideshima, T., *et al.* (2009). **Bcl6 as a Novel Therapeutic Target in Multiple Myeloma (MM)**. Blood *114*, 295.

Holmans, P., *et al.* (2009). **Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder**. The American Journal of Human Genetics *85*, 13-24, doi: https://doi.org/10.1016/j.ajhg.2009.05.011.

Huang, J., *et al.* (2015). **Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel**. Nature Communications *6*, 8111, doi: 10.1038/ncomms9111 https://www.nature.com/articles/ncomms9111#supplementary-information.

Hung, J.-H., *et al.* (2012). **Gene set enrichment analysis: performance evaluation and usage guidelines**. Briefings in Bioinformatics *13*, 281-291, doi: 10.1093/bib/bbr049.

Ichimaru, M., *et al.* (1982). **Multiple Myeloma Among Atomic Bornb Survivors in Hiroshima and Nagasaki, 1950–76: Relationship to Radiation Dose Absorbed by Marrow2**. JNCI: Journal of the National Cancer Institute *69*, 323-328, doi: 10.1093/jnci/69.2.323.

International Agency for Research on, C. (1993). **Occupational exposures of hairdressers and barbers and personal use of hair colourants; some hair dyes, cosmetic colourants, industrial dyestuffs and aromatic amines**. IARC monographs on the evaluation of carcinogenic risks to humans *57*.

Jain, M., *et al.* (2008). **Familial myeloma and monoclonal gammopathy: A report of eight African American families**. American Journal of Hematology *84*, 34-38, doi: 10.1002/ajh.21325.

Ji, J., *et al.* (2012). **Comparability of cancer identification among Death Registry, Cancer Registry and Hospital Discharge Registry**. International Journal of Cancer *131*, 2085-2093, doi: 10.1002/ijc.27462.

Jiyeon, R., *et al.* (2016). **Highly Expressed Integrin-α8 Induces Epithelial to Mesenchymal Transition-Like Features in Multiple Myeloma with Early Relapse**. Mol. Cells *39*, 898-908.

Jonsson, S., *et al.* (2017). **Identification of sequence variants influencing immunoglobulin levels**. Nature Genetics *49*, 1182, doi: 10.1038/ng.3897 https://www.nature.com/articles/ng.3897#supplementary-information.

Kamada, Y., *et al.* (2012). **Identification of unbalanced genome copy number abnormalities in patients with multiple myeloma by single-nucleotide polymorphism genotyping microarray analysis**. International Journal of Hematology *96*, 492-500, doi: 10.1007/s12185-012-1171-1.

Khuder, S. A. and Mutgi, A. B. (1997). **Meta-analyses of multiple myeloma and farming**. American Journal of Industrial Medicine *32*, 510-516, doi: 10.1002/(sici)1097-0274(199711)32:5<510::aid-ajim11>3.0.co;2-5.

Khuder, S. A. and Mutgi, A. B. (1998). **Meta-analyses of multiple myeloma and farming**. American Journal of Industrial Medicine *32*, 510-516, doi: 10.1002/(SICI)1097-0274(199711)32:5<510::AID-AJIM11>3.0.CO;2-5.

Kony, S. J., *et al.* (1997). **Radiation and genetic factors in the risk of second malignant neoplasms after a first cancer in childhood**. The Lancet *350*, 91-95, doi: https://doi.org/10.1016/S0140-6736(97)01116-1.

Korde, N., *et al.* (2011a). **Monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM): novel biological insights and development of early treatment strategies**. Blood *117*, 5573.

Korde, N., *et al.* (2011b). **Monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM): novel biological insights and development of early treatment strategies**. Blood *117*, 5573-5581, doi: 10.1182/blood-2011-01-270140.

Kraft, P. and Hunter, D. J. (2009). **Genetic Risk Prediction — Are We There Yet?** New England Journal of Medicine *360*, 1701-1703, doi: 10.1056/NEJMp0810107.

Kristinsson, S. Y., *et al.* (2009). **Patterns of hematologic malignancies and solid tumors among 37,838 first-degree relatives of 13,896 patients with multiple myeloma in Sweden**. International Journal of Cancer *125*, 2147-2150, doi: 10.1002/ijc.24514.

Kuznetsova, I. S., *et al.* (2016). **Radiation Risks of Leukemia, Lymphoma and Multiple Myeloma Incidence in the Mayak Cohort: 1948–2004**. PLoS ONE *11*, e0162710, doi: 10.1371/journal.pone.0162710.

Kyle, R. A. and Anderson, K. C. (1997). **A Tribute to Jan Gosta Waldenström**. Blood *89*, 4245.

Kyle, R. A. and Bayrd, E. D. (1966). **"Benign" monoclonal gammopathy: A potentially malignant condition?** The American Journal of Medicine *40*, 426-430, doi: https://doi.org/10.1016/0002-9343(66)90136-7.

Kyle, R. A.*, et al.* (1960). **Diagnostic criteria for electrophoretic patterns of serum and urinary proteins in multiple myeloma: Study of one hundred and sixty-five multiple myeloma patients and of seventy-seven nonmyeloma patients with similar electrophoretic patterns**. JAMA *174*, 245-251, doi: 10.1001/jama.1960.03030030025005.

Kyle, R. A.*, et al.* (2010). **Monoclonal gammopathy of undetermined significance (MGUS) and smoldering (asymptomatic) multiple myeloma: IMWG consensus perspectives risk factors for progression and guidelines for monitoring and management**. Leukemia *24*, 1121, doi: 10.1038/leu.2010.60.

Kyle, R. A. and Greipp, P. R. (1983). **Multiple myeloma: Houses and spouses**. Cancer *51*, 735-739, doi: 10.1002/1097-0142(19830215)51:4<735::AID-CNCR2820510430>3.0.CO;2-C.

Kyle, R. A.*, et al.* (1971). **Multiple myeloma in spouses**. Archives of Internal Medicine *127*, 944-946, doi: 10.1001/archinte.1971.00310170152022.

Kyle, R. A. and Rajkumar, S. (2015). **Monoclonal gammopathy of undetermined significance and multiple myeloma**. JAMA Oncology *1*, 174-175, doi: 10.1001/jamaoncol.2015.33.

Kyle, R. A.*, et al.* (2007). **Clinical Course and Prognosis of Smoldering (Asymptomatic) Multiple Myeloma**. New England Journal of Medicine *356*, 2582-2590, doi: 10.1056/NEJMoa070389.

Kyle, R. A.*, et al.* (2006). **Prevalence of Monoclonal Gammopathy of Undetermined Significance**. New England Journal of Medicine *354*, 1362-1369, doi: 10.1056/NEJMoa054494.

Kyle, R. A.*, et al.* (2002). **A Long-Term Study of Prognosis in Monoclonal Gammopathy of Undetermined Significance**. New England Journal of Medicine *346*, 564-569, doi: 10.1056/NEJMoa01133202.

Lage, K.*, et al.* (2007). **A human phenome-interactome network of protein complexes implicated in genetic disorders**. Nature Biotechnology *25*, 309, doi: 10.1038/nbt1295 https://www.nature.com/articles/nbt1295#supplementary-information.

Lagler, C.*, et al.* (2017). **The anti-myeloma activity of bone morphogenetic protein 2 predominantly relies on the induction of growth arrest and is apoptosis-independent**. PLOS ONE *12*, e0185720, doi: 10.1371/journal.pone.0185720.

Lambert, J.-C.*, et al.* (2013). **Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease**. Nature Genetics *45*, 1452, doi: 10.1038/ng.2802 https://www.nature.com/articles/ng.2802#supplementary-information.

Lamparter, D., *et al.* (2016). **Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics**. PLOS Computational Biology *12*, e1004714, doi: 10.1371/journal.pcbi.1004714.

Landgren, O., *et al.* (2009). **Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study**. Blood *113*, 5412.

Landgren, O., *et al.* (2006). **Familial characteristics of autoimmune and hematologic disorders in 8,406 multiple myeloma patients: A population-based case-control study**. International Journal of Cancer *118*, 3095-3098, doi: 10.1002/ijc.21745.

Landgren, O. and Mailankody, S. (2014). **Update on second primary malignancies in multiple myeloma: a focused review**. Leukemia *28*, 1423, doi: 10.1038/leu.2014.22 https://www.nature.com/articles/leu201422#supplementary-information.

Landgren, O., *et al.* (2011). **Myeloma and Second Primary Cancers**. New England Journal of Medicine *365*, 2241-2242, doi: 10.1056/NEJMc1111010.

Lango Allen, H., *et al.* (2010). **Hundreds of variants clustered in genomic loci and biological pathways affect human height**. Nature *467*, 832, doi: 10.1038/nature09410 https://www.nature.com/articles/nature09410#supplementary-information.

Lawlor, D. A., *et al.* (2008). **Mendelian randomization: using genes as instruments for making causal inferences in epidemiology**. Stat Med *27*, 1133-1163, doi: 10.1002/sim.3034.

Leich, E., *et al.* (2013a). **Multiple myeloma is affected by multiple and heterogeneous somatic mutations in adhesion- and receptor tyrosine kinase signaling molecules**. Blood Cancer Journal *3*, e102, doi: 10.1038/bcj.2012.47.

Leich, E., *et al.* (2013b). **Multiple myeloma is affected by multiple and heterogeneous somatic mutations in adhesion- and receptor tyrosine kinase signaling molecules**. Blood Cancer Journal *3*, e102, doi: 10.1038/bcj.2012.47 https://www.nature.com/articles/bcj201247#supplementary-information.

LeMasters, G. K., *et al.* (2006). **Cancer risk among firefighters: a review and meta-analysis of 32 studies**. J Occup Environ Med *48*, 1189-1202, doi: 10.1097/01.jom.0000246229.68697.90.

Levin, M. L. (1953). **The occurrence of lung cancer in man**. Acta Unio Int Contra Cancrum *9*, 531-541.

Li, N., *et al.* (2016). **Multiple myeloma risk variant at 7p15.3 creates an IRF4-binding site and interferes with CDCA7L expression**. Nature Communications *7*, 13656, doi: 10.1038/ncomms13656 https://www.nature.com/articles/ncomms13656#supplementary-information.

Lin, S. and Gregory, R. I. (2015). **Identification of small molecule inhibitors of Zcchc11 TUTase activity**. RNA Biology *12*, 792-800, doi: 10.1080/15476286.2015.1058478.

Lindblom, A., *et al.* (1993). **Genetic mapping of a second locus predisposing to hereditary non–polyposis colon cancer**. Nature Genetics *5*, 279, doi: 10.1038/ng1193-279.

Linet, M. S., *et al.* (1987). **A Case-Control Study of Multiple Myeloma in Whites: Chronic Antigenic Stimulation, Occupation, and Drug Use**. Cancer Research *47*, 2978.

Liu, T., *et al.* (2013). **Occupational exposure to methylene chloride and risk of cancer: a meta-analysis**. Cancer Causes & Control *24*, 2037-2049, doi: 10.1007/s10552-013-0283-0.

Locke, A. E., *et al.* (2015). **Genetic studies of body mass index yield new insights for obesity biology**. Nature *518*, 197, doi: 10.1038/nature14177 https://www.nature.com/articles/nature14177#supplementary-information.

Lohr, J. G., *et al.* (2014). **Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy**. Cancer cell *25*, 91-101, doi: 10.1016/j.ccr.2013.12.015.

Lubbe, S. J., *et al.* (2009). **Implications of Familial Colorectal Cancer Risk Profiles and Microsatellite Instability Status**. Journal of Clinical Oncology *27*, 2238-2244, doi: 10.1200/JCO.2008.20.3364.

Ludwig, H. (2010). **BMP-2: a culprit for anemia in myeloma**. Blood *116*, 3383.

Luijk, R., *et al.* (2018). **Genome-wide identification of directed gene networks using large-scale population genomics data**. Nature Communications *9*, 3097, doi: 10.1038/s41467-018-05452-6.

Lynch, H. T., *et al.* (2008a). **Familial Myeloma**. New England Journal of Medicine *359*, 152-157, doi: 10.1056/NEJMoa0708704.

Lynch, H. T., *et al.* (2008b). **Familial Myeloma: Study of a Unique Family**. The New England journal of medicine *359*, 152-157, doi: 10.1056/NEJMoa0708704.

Lynch, H. T., *et al.* (2001). **Familial multiple myeloma: a family study and review of the literature**. J Natl Cancer Inst *93*, 1479-1483.

Lynch, H. T., *et al.* (2005). **Phenotypic Heterogeneity in Multiple Myeloma Families**. Journal of Clinical Oncology *23*, 685-693, doi: 10.1200/JCO.2005.10.126.

Lyssenko, V., *et al.* (2008). **Clinical Risk Factors, DNA Variants, and the Development of Type 2 Diabetes**. New England Journal of Medicine *359*, 2220-2232, doi: 10.1056/NEJMoa0801869.

Mailankody, S., *et al.* (2011). **Risk of acute myeloid leukemia and myelodysplastic syndromes following multiple myeloma and its precursor disease (MGUS)**. Blood.

Makishima, H., *et al.* (2013). **Somatic SETBP1 mutations in myeloid malignancies**. Nature Genetics *45*, 942, doi: 10.1038/ng.2696 https://www.nature.com/articles/ng.2696#supplementary-information.

Maldonado, J. E. and Kyle, R. A. (1974). **Familial myeloma. Report of eight families and a study of serum proteins in their relatives**. Am J Med *57*, 875-884, doi: 10.1016/0002-9343(74)90164-8.

Mandema, E. and Wildervanck, L. S. (1954). **[Kahler's disease (multiple myeloma) in two sisters]**. J Genet Hum *3*, 170-175.

Manolio, T. A. (2010). **Genomewide Association Studies and Assessment of the Risk of Disease**. New England Journal of Medicine *363*, 166-176, doi: 10.1056/NEJMra0905980.

Manolio, T. A.*, et al.* (2009). **Finding the missing heritability of complex diseases**. Nature *461*, 747, doi: 10.1038/nature08494.

Marchini, J. and Howie, B. (2010). **Genotype imputation for genome-wide association studies**. Nature Reviews Genetics *11*, 499, doi: 10.1038/nrg2796 https://www.nature.com/articles/nrg2796#supplementary-information.

Marchini, J.*, et al.* (2007). **A new multipoint method for genome-wide association studies by imputation of genotypes**. Nature Genetics *39*, 906, doi: 10.1038/ng2088 https://www.nature.com/articles/ng2088#supplementary-information.

Massagué, J. (2008). **TGFbeta in Cancer**. Cell *134*, 215-230, doi: 10.1016/j.cell.2008.07.001.

McDuffie, H. H.*, et al.* (2009). **Clustering of cancer among families of cases with Hodgkin Lymphoma (HL), Multiple Myeloma (MM), Non-Hodgkin's Lymphoma (NHL), Soft Tissue Sarcoma (STS) and control subjects**. BMC Cancer *9*, 70, doi: 10.1186/1471-2407-9-70.

McMaster, M. L.*, et al.* (2018). **Two high-risk susceptibility loci at 6p25.3 and 14q32.13 for Waldenström macroglobulinemia**. Nature Communications *9*, 4182, doi: 10.1038/s41467-018-06541-2.

Meißner, T.*, et al.* (2011). **Gene Expression Profiling in Multiple Myeloma—Reporting of Entities, Risk, and Targets in Clinical Routine**. Clinical Cancer Research *17*, 7240.

Meyerding, H. W. (1925). **Multiple Myeloma**. Radiology *5*, 132-146, doi: 10.1148/5.2.132.

Mi, H.*, et al.* (2013). **Large-scale gene function analysis with the PANTHER classification system**. Nature Protocols *8*, 1551, doi: 10.1038/nprot.2013.092 https://www.nature.com/articles/nprot.2013.092#supplementary-information.

Michailidou, K.*, et al.* (2013). **Large-scale genotyping identifies 41 new loci associated with breast cancer risk**. Nature Genetics *45*, 353, doi: 10.1038/ng.2563 https://www.nature.com/articles/ng.2563#supplementary-information.

Mikulasova, A.*, et al.* (2017). **The spectrum of somatic mutations in monoclonal gammopathy of undetermined significance indicates a less complex genomic landscape than that in multiple myeloma**. Haematologica *102*, 1617.

Miller, A.*, et al.* (2017). **High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma**. Blood Cancer Journal *7*, e612, doi: 10.1038/bcj.2017.94 https://www.nature.com/articles/bcj201794#supplementary-information.

Mills, P. K., *et al.* (1990). **History of Cigarette Smoking and Risk of Leukemia and Myeloma: Results From the Adventist Health Study**. JNCI: Journal of the National Cancer Institute *82*, 1832-1836, doi: 10.1093/jnci/82.23.1832.

Mitchell, J. S., *et al.* (2015). **Implementation of genome-wide complex trait analysis to quantify the heritability in multiple myeloma**. Scientific Reports *5*, 12473, doi: 10.1038/srep12473 https://www.nature.com/articles/srep12473#supplementary-information.

Mitchell, J. S., *et al.* (2016). **Genome-wide association study identifies multiple susceptibility loci for multiple myeloma**. Nat Commun *7*, 12050, doi: 10.1038/ncomms12050.

Mithraprabhu, S., *et al.* (2014). **Dysregulated Class I histone deacetylases are indicators of poor prognosis in multiple myeloma**. Epigenetics *9*, 1511-1520, doi: 10.4161/15592294.2014.983367.

Mithraprabhu, S., *et al.* (2013). **Histone deacetylase (HDAC) inhibitors as single agents induce multiple myeloma cell death principally through the inhibition of class I HDAC**. British Journal of Haematology *162*, 559-562, doi: 10.1111/bjh.12388.

Mootha, V. K., *et al.* (2003). **PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. Nature Genetics *34*, 267, doi: 10.1038/ng1180 https://www.nature.com/articles/ng1180#supplementary-information.

Moreau, Y. and Tranchevent, L.-C. (2012). **Computational tools for prioritizing candidate genes: boosting disease gene discovery**. Nature Reviews Genetics *13*, 523, doi: 10.1038/nrg3253 https://www.nature.com/articles/nrg3253#supplementary-information.

Morris, A. P. and Zeggini, E. (2009). **An evaluation of statistical approaches to rare variant analysis in genetic association studies**. Genetic Epidemiology *34*, 188-193, doi: 10.1002/gepi.20450.

Moskvina, V., *et al.* (2006). **Effects of Differential Genotyping Error Rate on the Type I Error Probability of Case-Control Studies**. Human Heredity *61*, 55-64.

Murcray, C. E., *et al.* (2009). **Gene-Environment Interaction in Genome-Wide Association Studies**. American Journal of Epidemiology *169*, 219-226, doi: 10.1093/aje/kwn353.

Musto, P., *et al.* (2018). **Second primary malignancies in multiple myeloma: an overview and IMWG consensus**. Annals of Oncology *29*, 1074-1074, doi: 10.1093/annonc/mdx160.

Muz, B., *et al.* (2014). **The Role of Hypoxia and Exploitation of the Hypoxic Environment in Hematologic Malignancies**. Molecular Cancer Research *12*, 1347.

Nelder, J. A. and Wedderburn, R. W. M. (1972). **Generalized Linear Models**. Journal of the Royal Statistical Society. Series A (General) *135*, 370-384, doi: 10.2307/2344614.

Nieters, A., *et al.* (2005). **Tobacco and alcohol consumption and risk of lymphoma: Results of apopulation-based case-control study in Germany**. International Journal of Cancer *118*, 422-430, doi: 10.1002/ijc.21306.

Noble, W. S. (2009). **How does multiple testing correction work?** Nature Biotechnology *27*, 1135, doi: 10.1038/nbt1209-1135.

Offermanns, S. (2006). **Activation of platelet function through G protein-coupled receptors**. Circ Res *99*, 1293-1304, doi: 10.1161/01.res.0000251742.71301.16.

Ogama, Y., *et al.* (2004). **Prevalent hyper-methylation of the CDH13 gene promoter in malignant B cell lymphomas**. Int J Oncol *25*, 685-691.

Ogmundsdottir, H. M., *et al.* (2005). **Familiality of benign and malignant paraproteinemias. A population-based cancer-registry study of multiple myeloma families**. Haematologica *90*, 66.

Okada, Y., *et al.* (2013). **Genetics of rheumatoid arthritis contributes to biology and drug discovery**. Nature *506*, 376, doi: 10.1038/nature12873 https://www.nature.com/articles/nature12873#supplementary-information.

Oti, M. and Brunner, H. G. (2006). **The modular nature of genetic diseases**. Clinical Genetics *71*, 1-11, doi: 10.1111/j.1399-0004.2006.00708.x.

Ouyang, X., *et al.* (2011). **Transcription factor IRF8 directs a silencing programme for TH17 cell differentiation**. Nature Communications *2*, 314, doi: 10.1038/ncomms1311 https://www.nature.com/articles/ncomms1311#supplementary-information.

Palumbo, A. and Anderson, K. (2011). **Multiple Myeloma**. New England Journal of Medicine *364*, 1046-1060, doi: 10.1056/NEJMra1011442.

Palumbo, A., *et al.* (2015). **Revised International Staging System for Multiple Myeloma: A Report From International Myeloma Working Group**. Journal of Clinical Oncology *33*, 2863-2869, doi: 10.1200/JCO.2015.61.2267.

Palumbo, A., *et al.* (2014). **Second primary malignancies with lenalidomide therapy for newly diagnosed myeloma: a meta-analysis of individual patient data**. The Lancet Oncology *15*, 333-342, doi: https://doi.org/10.1016/S1470-2045(13)70609-0.

Paternoster, L., *et al.* (2017). **Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges**. PLOS Genetics *13*, e1006944, doi: 10.1371/journal.pgen.1006944.

Pedersen-Bjergaard, J., *et al.* (2002). **Genetic pathways in therapy-related myelodysplasia and acute myeloid leukemia**. Blood *99*, 1909.

Peltomaki, P., *et al.* (1993). **Genetic mapping of a locus predisposing to human colorectal cancer**. Science *260*, 810.

Peng, G., *et al.* (2009). **Gene and pathway-based second-wave analysis of genome-wide association studies**. European Journal Of Human Genetics *18*, 111, doi: 10.1038/ejhg.2009.115 https://www.nature.com/articles/ejhg2009115#supplementary-information.

Perez-Iratxeta, C., *et al.* (2002). **Association of genes to genetically inherited diseases using data mining**. Nature Genetics *31*, 316, doi: 10.1038/ng895 https://www.nature.com/articles/ng895#supplementary-information.

Perrone, G., *et al.* (2011). **HIF 1 Alpha: A Suitable Target for Multiple Myeloma**. Blood *118*, 2901.

Perrotta, C., *et al.* (2008). **Multiple myeloma and farming. A systematic review of 30 years of research. Where next?** Journal of Occupational Medicine and Toxicology *3*, 27, doi: 10.1186/1745-6673-3-27.

Pers, T. H., *et al.* (2015). **Biological interpretation of genome-wide association studies using predicted gene functions**. Nature Communications *6*, 5890, doi: 10.1038/ncomms6890 https://www.nature.com/articles/ncomms6890#supplementary-information.

Peto, J., *et al.* (1999). **Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer**. J Natl Cancer Inst *91*, 943-949.

Phillips, P. C. (1998). **The Language of Gene Interaction**. Genetics *149*, 1167.

Phillips, P. C. (2008). **Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems**. Nature Reviews Genetics *9*, 855, doi: 10.1038/nrg2452.

Plagnol, V., *et al.* (2007). **A Method to Address Differential Bias in Genotyping in Large-Scale Association Studies**. PLOS Genetics *3*, e74, doi: 10.1371/journal.pgen.0030074.

Porcu, E., *et al.* (2018). **Mendelian Randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits**. bioRxiv.

Prentice, R. L. (1985). **Relative risk regression analysis of epidemiologic data**. Environmental Health Perspectives *63*, 225-234.

Preston, D. L., *et al.* (1994). **Cancer incidence in atomic bomb survivors. Part III. Leukemia, lymphoma and multiple myeloma, 1950-1987**. Radiat Res *137*, S68-97.

Price, A. L., *et al.* (2006). **Principal components analysis corrects for stratification in genome-wide association studies**. Nature Genetics *38*, 904, doi: 10.1038/ng1847 https://www.nature.com/articles/ng1847#supplementary-information.

Purcell, S., *et al.* (2007). **PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses**. American Journal of Human Genetics *81*, 559-575.

Qi, T., *et al.* (2018). **Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood**. Nature Communications *9*, 2282, doi: 10.1038/s41467-018-04558-1.

R Development Core Team (2018). **R: A Language and Environment for Statistical Computing** (

Rajkumar, S. V.*, et al.* (2005). **Serum free light chain ratio is an independent risk factor for progression in monoclonal gammopathy of undetermined significance**. Blood *106*, 812.

Rajkumar, S. V.*, et al.* (2015). **Smoldering multiple myeloma**. Blood *125*, 3069.

Ramakrishnan, V. and D'Souza, A. (2016). **Signaling Pathways and Emerging Therapies in Multiple Myeloma**. Current Hematologic Malignancy Reports *11*, 156-164, doi: 10.1007/s11899-016-0315-4.

Ravindran, A.*, et al.* (2016). **Prevalence, incidence and survival of smoldering multiple myeloma in the United States**. Blood Cancer Journal *6*, e486, doi: 10.1038/bcj.2016.100.

Razavi, P.*, et al.* (2013). **Patterns of second primary malignancy risk in multiple myeloma patients before and after the introduction of novel therapeutics**. Blood Cancer Journal *3*, e121, doi: 10.1038/bcj.2013.19.

Samanic, C.*, et al.* (2004). **Obesity and cancer risk among white and black United States veterans**. Cancer Causes & Control *15*, 35-44, doi: 10.1023/B:CACO.0000016573.79453.ba.

Schadt, E. E. (2009). **Molecular networks as sensors and drivers of common human diseases**. Nature *461*, 218, doi: 10.1038/nature08454.

Schmermund, A.*, et al.* (2002). **Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: Rationale and design of the Heinz Nixdorf RECALL Study**. American Heart Journal *144*, 212-218, doi: https://doi.org/10.1067/mhj.2002.123579.

Schulz, H.*, et al.* (2017). **Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus**. Nature Communications *8*, 1511, doi: 10.1038/s41467-017-01818-4.

Segrè, A. V.*, et al.* (2010). **Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits**. PLOS Genetics *6*, e1001058, doi: 10.1371/journal.pgen.1001058.

Shaughnessy, J. D.*, et al.* (2007). **A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1**. Blood *109*, 2276.

Shen, Y.*, et al.* (2017). **Epigenetic and genetic dissections of UV-induced global gene dysregulation in skin cells through multi-omics analyses**. Scientific Reports *7*, 42646, doi: 10.1038/srep42646 https://www.nature.com/articles/srep42646#supplementary-information.

Shungin, D.*, et al.* (2015). **New genetic loci link adipose and insulin biology to body fat distribution**. Nature *518*, 187, doi: 10.1038/nature14132 https://www.nature.com/articles/nature14132#supplementary-information.

Siegel, R. L., *et al.* (2016). **Cancer statistics, 2016**. CA Cancer J Clin *66*, 7-30, doi: 10.3322/caac.21332.

Singhal, S., *et al.* (1999). **Antitumor Activity of Thalidomide in Refractory Multiple Myeloma**. New England Journal of Medicine *341*, 1565-1571, doi: 10.1056/NEJM199911183412102.

Smith, C. L. and Eppig, J. T. (2009). **The mammalian phenotype ontology: enabling robust annotation and comparative analysis**. Wiley Interdisciplinary Reviews: Systems Biology and Medicine *1*, 390-399, doi: 10.1002/wsbm.44.

Smith, G. D. and Ebrahim, S. (2003). **'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?** Int J Epidemiol *32*, 1-22.

Solt, L. A., *et al.* (2017). **Negative regulation of Th17 cell development and function by the REV-ERBs**. The Journal of Immunology *198*, 127.115.

Sonoda, T., *et al.* (2001). **Meta-analysis of Multiple Myeloma and Benzene Exposure**. Journal of Epidemiology *11*, 249-254, doi: 10.2188/jea.11.249.

Speliotes, E. K., *et al.* (2010). **Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index**. Nature Genetics *42*, 937, doi: 10.1038/ng.686 https://www.nature.com/articles/ng.686#supplementary-information.

Stacey, S. N., *et al.* (2011). **A germline variant in the TP53 polyadenylation signal confers cancer susceptibility**. Nature Genetics *43*, 1098, doi: 10.1038/ng.926 https://www.nature.com/articles/ng.926#supplementary-information.

Stranger, B. E., *et al.* (2011). **Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics**. Genetics *187*, 367.

Subramanian, A., *et al.* (2007). **GSEA-P: a desktop application for Gene Set Enrichment Analysis**. Bioinformatics *23*, 3251-3253, doi: 10.1093/bioinformatics/btm369.

Subramanian, A., *et al.* (2005). **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. Proceedings of the National Academy of Sciences *102*, 15545.

Sud, A., *et al.* (2017a). **Genome-wide association studies of cancer: current insights and future perspectives**. Nature Reviews Cancer *17*, 692, doi: 10.1038/nrc.2017.82 https://www.nature.com/articles/nrc.2017.82#supplementary-information.

Sud, A., *et al.* (2017b). **Risk of Second Cancer in Hodgkin Lymphoma Survivors and Influence of Family History**. Journal of Clinical Oncology *35*, 1584-1590, doi: 10.1200/JCO.2016.70.9709.

Sun, J. Y., *et al.* (2011). **Histone Deacetylase Inhibitors Demonstrate Significant Preclinical Activity as Single Agents, and in Combination with Bortezomib in Waldenström's Macroglobulinemia**. Clinical Lymphoma, Myeloma and Leukemia *11*, 152-156, doi: 10.3816/CLML.2011.n.036.

Szklarczyk, D., *et al.* (2017). **The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible**. Nucleic Acids Research *45*, D362-D368, doi: 10.1093/nar/gkw937.

Takeuchi, K., *et al.* (2010). **TGF-β Inhibition Restores Terminal Osteoblast Differentiation to Suppress Myeloma Growth**. PLOS ONE *5*, e9870, doi: 10.1371/journal.pone.0009870.

Takkouche, B., *et al.* (2009). **Risk of cancer among hairdressers and related workers: a meta-analysis**. International Journal of Epidemiology *38*, 1512-1531, doi: 10.1093/ije/dyp283.

Tavani, A., *et al.* (2000). **Red meat intake and cancer risk: A study in Italy**. International Journal of Cancer *86*, 425-428, doi: 10.1002/(SICI)1097-0215(20000501)86:3<425::AID-IJC19>3.0.CO;2-S.

the, D. G. R., *et al.* (2012). **Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes**. Nature Genetics *44*, 981, doi: 10.1038/ng.2383
https://www.nature.com/articles/ng.2383#supplementary-information.

The Genomes Project, C., *et al.* (2012). **An integrated map of genetic variation from 1,092 human genomes**. Nature *491*, 56, doi: 10.1038/nature11632
https://www.nature.com/articles/nature11632#supplementary-information.

The International Consortium for Blood Pressure Genome-Wide Association, S., *et al.* (2011). **Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk**. Nature *478*, 103, doi: 10.1038/nature10405
https://www.nature.com/articles/nature10405#supplementary-information.

The International HapMap Consortium, *et al.* (2010). **Integrating common and rare genetic variation in diverse human populations**. Nature *467*, 52, doi: 10.1038/nature09298
https://www.nature.com/articles/nature09298#supplementary-information.

The International HapMap Consortium, *et al.* (2003). **The International HapMap Project**. Nature *426*, 789, doi: 10.1038/nature02168
https://www.nature.com/articles/nature02168#supplementary-information.

The U. K. K. Consortium, *et al.* (2015). **The UK10K project identifies rare variants in health and disease**. Nature *526*, 82, doi: 10.1038/nature14962
https://www.nature.com/articles/nature14962#supplementary-information.

The Wellcome Trust Case Control, C., *et al.* (2007). **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. Nature *447*, 661, doi: 10.1038/nature05911
https://www.nature.com/articles/nature05911#supplementary-information.

Therneau, T. M., *et al.* (2012). **Incidence of Monoclonal Gammopathy of Undetermined Significance and Estimation of Duration Before First Clinical Recognition**. Mayo Clinic Proceedings *87*, 1071-1079, doi: 10.1016/j.mayocp.2012.06.014.

Thomas, A*., et al.* (2012). **Second malignancies after multiple myeloma: from 1960s to 2010s**. Blood *119*, 2731.

Thomsen, H*., et al.* (2017). **Genomewide association study on monoclonal gammopathy of unknown significance (MGUS)**. European Journal of Haematology *99*, 70-79, doi: 10.1111/ejh.12892.

Tong, A. H. Y*., et al.* (2001). **Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants**. Science *294*, 2364.

Tripepi, G*., et al.* (2007). **Measures of effect: Relative risks, odds ratios, risk difference, and 'number needed to treat'**. Kidney International *72*, 789-791, doi: https://doi.org/10.1038/sj.ki.5002432.

Tripp, S. R*., et al.* (2005). **Relationship between EGFR overexpression and gene amplification status in central nervous system gliomas**. Anal Quant Cytol Histol *27*, 71-78.

Vaitsiakhovich, T*., et al.* (2015). **METAINTER: meta-analysis of multiple regression models in genome-wide association studies**. Bioinformatics *31*, 151-157, doi: 10.1093/bioinformatics/btu629.

van der Valk, R. J. P*., et al.* (2015). **A novel common variant in DCST2 is associated with length in early life and height in adulthood**. Human Molecular Genetics *24*, 1155-1168, doi: 10.1093/hmg/ddu510.

Vlaanderen, J*., et al.* (2011). **Occupational Benzene Exposure and the Risk of Lymphoma Subtypes: A Meta-analysis of Cohort Studies Incorporating Three Study Quality Dimensions**. Environmental Health Perspectives *119*, 159-167, doi: 10.1289/ehp.1002318.

Vlajinac, H. D*., et al.* (2003). **Case-control study of multiple myeloma with special reference to diet as risk factor**. Neoplasma *50*, 79-83.

Wadhera, R. K*., et al.* (2011). **Incidence, clinical course, and prognosis of secondary monoclonal gammopathy of undetermined significance in patients with multiple myeloma**. Blood *118*, 2985.

Wakefield, L. M. and Hill, C. S. (2013). **Beyond TGFβ: roles of other TGFβ superfamily members in cancer**. Nature Reviews Cancer *13*, 328, doi: 10.1038/nrc3500.

Waldenstrom, J. (1961). **Studies on conditions associated with disturbed gamma globulin formation (gammopathies)**. Harvey Lect *56*, 211.

WaldenstrÖM, J. A. N. (1964). **The Occurrence of Benign, Essential Monoclonal (M Type), Non-macromolecular Hyperglobulinemia and its Differential Diagnosis**. Acta Medica Scandinavica *176*, 345-365, doi: 10.1111/j.0954-6820.1964.tb00942.x.

Wang, K*., et al.* (2009). **Common genetic variants on 5p14.1 associate with autism spectrum disorders**. Nature *459*, 528, doi: 10.1038/nature07999
https://www.nature.com/articles/nature07999#supplementary-information.

Wang, Y., *et al.* (2014). **Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer**. Nature Genetics *46*, 736, doi: 10.1038/ng.3002 https://www.nature.com/articles/ng.3002#supplementary-information.

Ward, L. D. and Kellis, M. (2012). **HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants**. Nucleic Acids Research *40*, D930-D934, doi: 10.1093/nar/gkr917.

Warta, R. and Herold-Mende, C. (2017). **Helping EGFR inhibition to block cancer**. Nature Neuroscience *20*, 1035, doi: 10.1038/nn.4605.

Watanabe, K., *et al.* (2017). **Functional mapping and annotation of genetic associations with FUMA**. Nature Communications *8*, 1826, doi: 10.1038/s41467-017-01261-5.

Weinhold, N., *et al.* (2014a). **Inherited genetic susceptibility to monoclonal gammopathy of unknown significance**. Blood *123*, 2513.

Weinhold, N., *et al.* (2014b). **Inherited genetic susceptibility to monoclonal gammopathy of unknown significance**. Blood *123*, 2513-2517; quiz 2593, doi: 10.1182/blood-2013-10-532283.

Weinhold, N., *et al.* (2015). **The 7p15.3 (rs4487645) association for multiple myeloma shows strong allele-specific regulation of the MYC-interacting gene CDCA7L in malignant plasma cells**. Haematologica *100*, e110-e113, doi: 10.3324/haematol.2014.118786.

Weiss, B. M., *et al.* (2009). **A monoclonal gammopathy precedes multiple myeloma in most patients**. Blood *113*, 5418.

Welch, H. C. E., *et al.* (2002). **P-Rex1, a PtdIns(3,4,5)P3- and G-beta-gamma-Regulated Guanine-Nucleotide Exchange Factor for Rac**. Cell *108*, 809-821, doi: 10.1016/S0092-8674(02)00663-3.

Wellek, S. and Ziegler, A. (2009). **A Genotype-Based Approach to Assessing the Association between Single Nucleotide Polymorphisms**. Human Heredity *67*, 128-139.

Went, M., *et al.* (2018). **Identification of multiple risk loci and regulatory mechanisms influencing susceptibility to multiple myeloma**. Nature Communications *9*, 3707, doi: 10.1038/s41467-018-04989-w.

Willer, C. J., *et al.* (2010). **METAL: fast and efficient meta-analysis of genomewide association scans**. Bioinformatics *26*, 2190-2191, doi: 10.1093/bioinformatics/btq340.

Williams, R. R. and Horm, J. W. (1977). **Association of Cancer Sites With Tobacco and Alcohol Consumption and Socioeconomic Status of Patients: Interview Study From the Third National Cancer Survey**. JNCI: Journal of the National Cancer Institute *58*, 525-547, doi: 10.1093/jnci/58.3.525.

Wood, A. R., *et al.* (2014). **Defining the role of common variation in the genomic and biological architecture of adult human height**. Nature Genetics *46*, 1173, doi: 10.1038/ng.3097 https://www.nature.com/articles/ng.3097#supplementary-information.

Wooster, R., *et al.* (1994). **Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13**. Science *265*, 2088.

Yang, J., *et al.* (2011a). **GCTA: A Tool for Genome-wide Complex Trait Analysis**. American Journal of Human Genetics *88*, 76-82, doi: 10.1016/j.ajhg.2010.11.011.

Yang, J., *et al.* (2012). **Secondary Primary Malignancies in Multiple Myeloma: An Old Nemesis Revisited**. Advances in Hematology *2012*, 801495, doi: 10.1155/2012/801495.

Yang, J., *et al.* (2011b). **Genomic inflation factors under polygenic inheritance**. European Journal Of Human Genetics *19*, 807, doi: 10.1038/ejhg.2011.39.

Yarden, Y. (2001). **The EGFR family and its ligands in human cancer: signalling mechanisms and therapeutic opportunities**. European Journal of Cancer *37*, 3-8, doi: https://doi.org/10.1016/S0959-8049(01)00230-1.

Yates, A., *et al.* (2016). **Ensembl 2016**. Nucleic Acids Research *44*, D710-D716, doi: 10.1093/nar/gkv1157.

Yu, X., *et al.* (2013). **TH17 Cell Differentiation Is Regulated by the Circadian Clock**. Science *342*, 727.

Zalcberg, J. R., *et al.* (1982). **Chronic lymphatic leukemia developing in a patient with multiple myeloma: immunologic demonstration of a clonally distinct second malignancy**. Cancer *50*, 594-597.

Zapata-Diomedi, B., *et al.* (2018). **Population attributable fraction: names, types and issues with incorrect interpretation of relative risks**. British Journal of Sports Medicine *52*, 212.

Zerbino, D. R., *et al.* (2018). **Ensembl 2018**. Nucleic Acids Research *46*, D754-D761, doi: 10.1093/nar/gkx1098.

Zhang, H., *et al.* (2009). **Prostate cancer as a first and second cancer: effect of family history**. British Journal Of Cancer *101*, 935, doi: 10.1038/sj.bjc.6605263.

Zhao, G.-N., *et al.* (2015). **Interferon regulatory factors: at the crossroads of immunity, metabolism, and disease**. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease *1852*, 365-378, doi: https://doi.org/10.1016/j.bbadis.2014.04.030.

Zhong, H., *et al.* (2010). **Integrating Pathway Analysis and Genetics of Gene Expression for Genome-wide Association Studies**. The American Journal of Human Genetics *86*, 581-591, doi: https://doi.org/10.1016/j.ajhg.2010.02.020.

Zhu, Z., *et al.* (2016). **Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets**. Nature Genetics *48*, 481, doi: 10.1038/ng.3538 https://www.nature.com/articles/ng.3538#supplementary-information.

Zhu, Z., *et al.* (2018). **Causal associations between risk factors and common diseases inferred from GWAS summary data**. Nature Communications *9*, 224, doi: 10.1038/s41467-017-02317-2.

Zingone, A. and Kuehl, W. M. (2011). **Pathogenesis of Monoclonal Gammopathy of Undetermined Significance (MGUS) and Progression to Multiple Myeloma**. Seminars in hematology *48*, 4-12, doi: 10.1053/j.seminhematol.2010.11.003.

Zondervan, K. T. and Cardon, L. R. (2007). **Designing candidate gene and genome-wide case–control association studies**. Nature Protocols *2*, 2492, doi: 10.1038/nprot.2007.366 https://www.nature.com/articles/nprot.2007.366#supplementary-information.

Publications

Papers published either as a direct result from or through collaborative work during this thesis:

†***Chattopadhyay S**, *Thomsen H, da Silva Filho MI, Weinhold N, Hoffmann P, Nöthen MM, et al. Enrichment of B cell receptor signaling and epidermal growth factor receptor pathways in monoclonal gammopathy of undetermined significance: a genome-wide genetic interaction study. Molecular medicine (Cambridge, Mass). 2018;24(1):30.

†**Chattopadhyay S**, Thomsen H, Yadav P, da Silva Filho MI, Weinhold N, Nöthen MM, et al. Genome-wide interaction and pathway-based identification of key regulators in multiple myeloma. Under review. 2018.

†**Chattopadhyay S**, Yu H, Sud A, Sundquist J, Försti A, Hemminki A, et al. Multiple myeloma: family history and mortality in second primary cancers. Blood Cancer Journal. 2018;8(8):75.

***Chattopadhyay S**, *Zheng G, *Sud A, Yu H, Sundquist K, Sundquist J, et al. Risk of second primary cancer following myeloid neoplasia and risk of myeloid neoplasia as second primary cancer: a nationwide, observational follow up study in Sweden. The Lancet Haematology. 2018;5(8):e368-e77.

Chattopadhyay S, Sud A, Zheng G, Yu H, Sundquist K, Sundquist J, et al. Second primary cancers in non-Hodgkin lymphoma: Bidirectional analyses suggesting role for immune dysfunction. International Journal of Cancer. 2018;0(0).

Chattopadhyay S, Hemminki O, Försti A, Sundquist K, Sundquist J, Hemminki K. Impact of family history of cancer on risk and mortality of second cancers in patients with prostate cancer. Prostate Cancer and Prostatic Diseases. 2018.

Chattopadhyay S, Zheng G, Hemminki O, Försti A, Sundquist K, Hemminki K. Prostate cancer survivors: Risk and mortality in second primary cancers. Cancer Medicine. 2018;0(0).

Chattopadhyay S, Hemminki A, Försti A, Sundquist K, Sundquist J, Hemminki K. Familial risks and mortality in second primary cancers in melanoma. JNCI Cancer Spectrum. 2018; Accepted.

Sud A, **Chattopadhyay S**, Thomsen H, Sundquist K, Sundquist J, Houlston RS, et al. Familial risks of acute myeloid leukemia, myelodysplastic syndromes, and myeloproliferative neoplasms. Blood. 2018;132(9):973.

Zheng G, **Chattopadhyay S**, Forsti A, Sundquist K, Hemminki K. Familial risks of second primary cancers and mortality in ovarian cancer patients. Clinical epidemiology. 2018;10:1457-66.

†Articles directly related to work described in the thesis
*Authors contributed equally

Curriculum Vitae

Personal information

Name:	Subhayan Chattopadhyay
Date of birth:	15th September, 1993
Place of birth:	Asansol, India
Nationality:	Indian
Marital status:	Single

School

School (1999 - 2009)	Asansol Ramakrishna mission high school, Asansol, India
High School (2009 - 2011)	Bidhan Chandra Institute, Durgapur, India

Universities

B. Sc. (2011 - 2014)	Department of statistics, Narendrapur Ramkrishna mission residential college, University of Calcutta, India
M. Sc. (2014 - 2016)	School of Mathematics and Statistics, University of Hyderabad, India
Dr. Sc. Hum (2016 -)	Department of molecular genetic epidemiology, German cancer research center, University of Heidelberg, Germany

Acknowledgements

I would like to thank Kari for providing me with the opportunity to carry out research in his group. I cannot thank you enough for your patience every time we discussed a problem, for your understanding in all those times there were hiccups in analyses, for your advices that paved the way for a successful outcome and for your overall dedication without which I'd not have made progress. I have learnt a great many things from you for which I'll remain ever grateful.

Many thanks to Asta, for your guidance and mentorship helped me realize my projects far outreaching than I would have ever been able to. Our discussions were always very enlightening; your patient and detailed criticism enriched each of our studies and had it not been for your insightful constructive supervision, this thesis would not have come to fruition.

To Hauke, I will owe my degree if I ever get around getting one. You have shown the light ahead since my day one in this lab. Your help, patience and guidance paved the way every time I was stuck be it in analyses or reading letters in German. Your guidance has always meant the world.

My special thanks to Hao, Yasmeen, Erina, Iman, Angelica, Marion, Obul, Pankaj, Hamid and the larger C050 family; you all have been my friends and never just colleagues. From barbecue to dinner, our warm welcoming discussions to debates, you are what made work fun. A big thanks to Ayushi and Diamanto in addition for your kind help with translation. Thanks to our small lunch group that kept mensa close to my heart. I would further like to thank Amit for his friendship and help; I'll always look up to you.

In addition to the people in our lab I would thank DKFZ for funding and hosting my research and thanks to all the patients, participants and individuals who helped us procure data which made this work a success.

I would like to show my sincere gratitude to Jiban Banerjee. You ignited my passion for statistics and showed me what mathematics is really all about. I want to express my veneration for Madhuchhanda Bhattacharjee and Sailu Yellaboina for you have supported, guided and quite literally shaped my career. If I have learned ever so little and contributed anything back, it is because all of your inspiration.

My deepest thank you is for my parents. Things are not bright in the night sky but you two have been my true north all along guiding me had I ever got lost. I wish time was kinder on us.

A big thank you to Guoqiao. Your passion for research and dedication to work made me realize I'm lazy. Our frequent discussions have helped us become better at research and tuned our minds to critical thinking (entirely my belief). But above all you have brought me a smile every time I was down. There was no avoiding your lack of humor, thanks for that.

Sidhu, Souvik and Bama, I'd thank you guys for nothing. You deserve better. And to all my friends, guys, this is just a thesis not a Thanksgiving skit. If I ever get a Fields medal, I'll mention all of you by full name, promise!