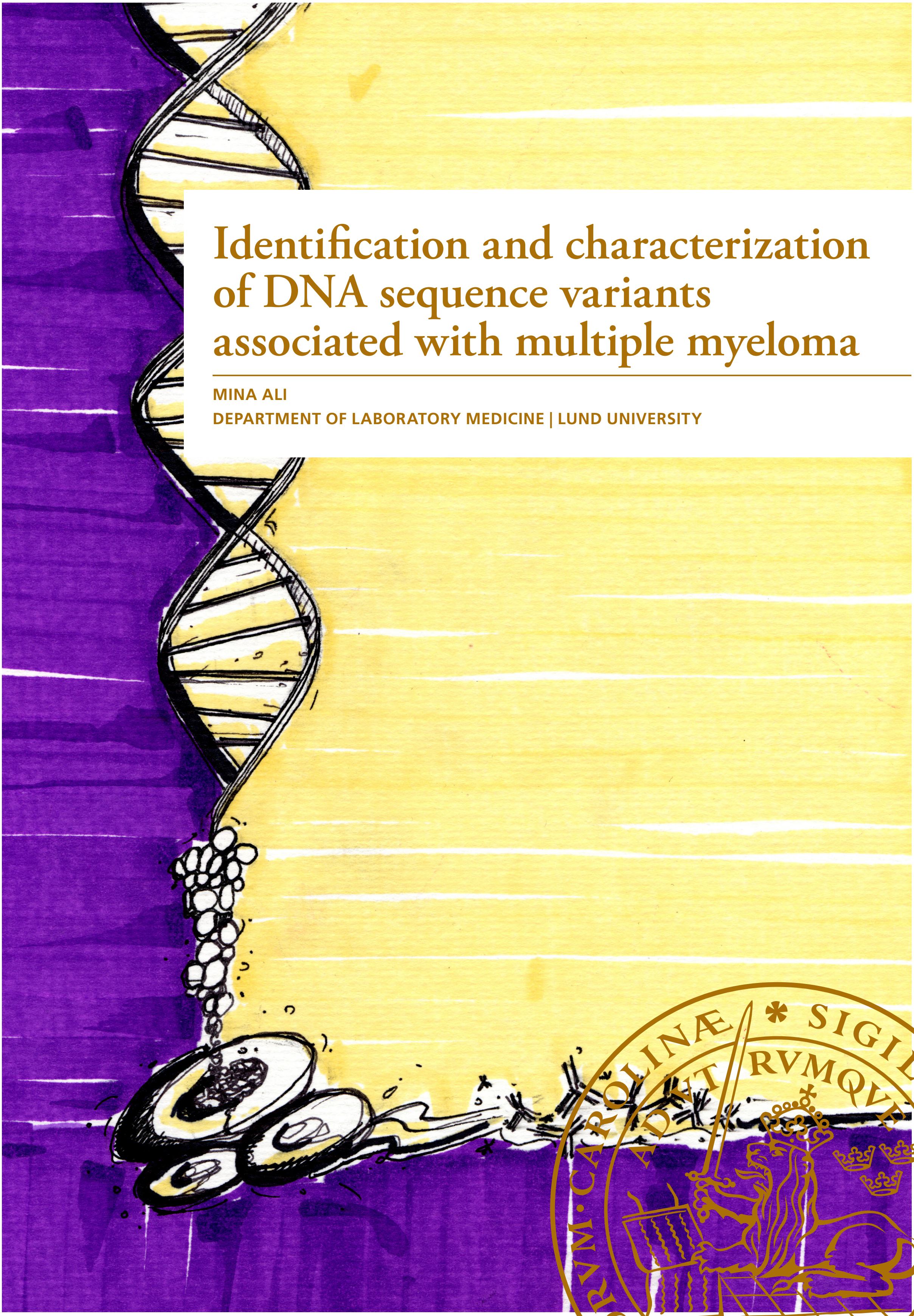

Identification and characterization of DNA sequence variants associated with multiple myeloma

MINA ALI

DEPARTMENT OF LABORATORY MEDICINE | LUND UNIVERSITY

Identification and characterization of DNA sequence variants associated with multiple myeloma

Identification and characterization of DNA sequence variants associated with multiple myeloma

Mina Ali

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Medicine, Lund University, Sweden.
To be defended on Tuesday 12th of June 2018, at 13:00 in I1345 lecture hall at Biomedical Centre.

Faculty opponent
Professor Richard Rosenquist Brandell

Organization LUND UNIVERSITY Faculty of Medicine Department of Laboratory Medicine Division of Hematology and Transfusion Medicine	**Document name** DOCTORAL DISSERTATION
	Date of disputation 2018-06-12
Author: Mina Ali	**Sponsoring organization**

Title: Identification and characterization of DNA sequence variants associated with multiple myeloma

Abstract

Multiple myeloma (MM) is the second most common hematologic malignancy. The disease is characterized by an uncontrolled growth of malignant plasma cells in the bone marrow, producing a monoclonal immunoglobulin (“M protein”) that can be detected in peripheral blood. Clinically, MM is characterized by bone marrow failure, lytic bone lesions, hypercalcemia, and kidney failure. It is preceded by monoclonal gammopathy of unknown significance (MGUS), a common condition defined as a clonal growth of plasma cells that does not yet satisfy the criteria for MM, but progresses to MM at a rate of 1% per year. Increasing evidence supports that the biology of MM is influenced by inborn genetic variation. First degree relatives of patients with MM and MGUS seem to have higher risk for MM, and a higher risk of certain other malignancies. In this Ph.D. project, we aim to find DNA sequence variants that predispose for MM, and understand how these variants contribute to MM development.

Case-control genome-wide association study (GWAS) is our approach for finding variants. Previous GWASs have identified eight genetic variants that associate with MM development. As a first step in this study, we also carried out a GWAS on a case-control data set from Sweden-Norway and Iceland (Paper I). Following statistical analysis we could identify one novel MM risk locus related to the *ELL2* (Elongation Factor for RNA Polymerase II 2) gene and a promising association with the *TOM1-HMGXB4* locus at 22q13 and we could confirm the previously reported loci. In the second part of the project, we carried out a meta-analysis of six genome-wide association studies together with the United Kingdom, Germany, Netherlands, the United States and Iceland. Previously reported loci were confirmed and eight new loci were discovered (Paper II). Understanding how the identified risk variants contribute to MM development is the next challenge. In the third part of the project we focused on the *ELL2* gene located on the locus 5q15 and tried to understand its mechanism of action in MM (Paper III). Finally, we retrieved clinical information for 871 patients diagnosed with MM from the Swedish Multiple Myeloma Registry. We tested for association between sequence variants and MM overall survival (Paper IV). Our findings provide further insights into the genetic and biological basis of MM predisposition.

Key words: Multiple myeloma, inherited susceptibility, genome-wide association study, *ELL2,* expression quantitative trait loci

Classification system and/or index terms (if any)

Supplementary bibliographical information		**Language:** English
ISSN and key title: 1652-8220		**ISBN:** 978-91-7619-650-2
Recipient’s notes	**Number of pages 49**	**Price**
	Security classification	

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature Mina Ali **Date** 2018-05-21

Identification and characterization of DNA sequence variants associated with multiple myeloma

Mina Ali

LUND UNIVERSITY

Cover: “Ig secretion and DNA in plasma cells”, Drawing by Nahid Pordel.

Lund University, Faculty of Medicine, Doctoral Dissertation Series 2018:84

Department of Laboratory Medicine, Division of Hematology and Transfusion Medicine

ISBN 978-91-7619-650-2

ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University

Lund 2018

NORDIC SWAN ECOLABEL 1234 5678

ISO 14001:2004 CERTIFICATION Intertek

MADE IN SWEDEN

Media-Tryck is an environmentally certified and ISO 14001 certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

Dedicated to my parents

کار ما نیست شناسایی راز گل سرخ،

کار ما شاید این است

که در افسون گل سرخ شناور باشیم.

پشت دانایی اردو بزنیم،

صبح ها وقتی خورشید در می آید متولد بشویم،

هیجان ها را پرواز دهیم،

روی ادراک فضا، رنگ، صدا، پنجره گل نم بزنیم.

سهراب سپهری (۱۳۵۹-۱۳۰۷)

“*We are not able to fathom the mystery of the rose.*
Perhaps our business is to float within the magic of the rose.
To camp behind wisdom,
be born again when the sun rises in the mornings
and allow our excitement to fly”

Sohrab Sepehri (1928-1980)

Content

Preface

Biology in the 21st century is merging more and more with other disciplines. Mathematics, physics, engineering and computational sciences are playing an increasingly important role in biological research. Advances in sequencing and genotyping technologies are prime examples of this integration. With these high-throughput genomic technologies, it is now possible to sequence large amount of DNA and RNA in shorter periods of time and at a reasonable cost. Today, we have the biological facilities and computational power needed to read the genome and find out the differences between the genetic make-ups of individuals.

Genome-wide association study (GWAS) serves to detect genetic variants that associate with a given disease or trait. GWA studies have proven a powerful technique for identifying genetic variations that predispose for complex diseases. Yet, the heritability of many diseases and traits remain incompletely explained. Even though our ability to read the genome is increasing, our understanding of the functional consequences of DNA sequence variants that have been identified by GWAS has not matured enough.

This thesis provides an overview of four years of research at the Division of Hematology and Transfusion medicine at Lund University. The focus of the thesis is on multiple myeloma (MM) which is the second most common hematologic malignancy. This disease is characterized by an uncontrolled growth of plasma cells in the bone marrow. By GWAS, we have identified DNA sequence variants that predispose for MM and, by combining *in vitro* and *in silico* methods, we have tried to understand how some of these variants promote MM development.

The thesis is organized as follows: Chapter 1 covers the general concepts and background of the study. Chapter 2 focuses on the identification of genetic variants through GWAS. Chapter 3 provides a brief overview of different methods used for characterization of GWAS-identified variants. Chapter 4 describes our original work on MM predisposition, and the *ELL2* risk allele. Finally, Chapter 5 summarizes the conclusions of the articles on which this thesis is built up on.

Lund, June 2018

List of papers

I. **Variants in *ELL2* influencing immunoglobulin levels associate with multiple myeloma.**

Swaminathan B*, Thorleifsson G*, Jöud M*, **Ali M***, Johnsson E, Ajore R, Sulem P, Halvarsson B-M, Eyjolfsson G, Haraldsdottir V, Hultman C, Ingelsson E, Kristinsson S.Y, Kähler A.K, Lenhoff S, Masson G, Mellqvist U-H, Månsson R, Nelander S, Olafsson I, Sigurðardottir O, Steingrimsdóttir H, Vangsted A, Vogel U, Waage A, Nahi H, Gudbjartsson D.F, Rafnar T, Turesson I, Gullberg U, Stefánsson K**, Hansson M**, Thorsteinsdóttir U**, and Nilsson B**.

Nature Communications, (2015). 6, 7213

II. **Genome-wide association study identifies multiple susceptibility loci for multiple myeloma**

Mitchell J.S*, Li N*, Weinhold N*, Försti A*, **Ali M***, Van Duin M*, Thorleifsson G, Johnson D.C, Chen B, Halvarsson B-M, Gudbjartsson D.F, Kuiper R, Stephens O.W, Bertsch U, Broderick P, Campo C, Einsele H, Gregory W.A, Gullberg U, Henrion M, Hillengass J, Hoffmann P, Jackson G.H, Johnsson E, Jöud M, Kristinsson S.Y, Lenhoff S, Lenive O, Mellqvist U-H, Migliorini G, Nahi H, Nelander S, Nickel J, Nöthen M.M, Rafnar T, Ross F.M, da Silva Filho M.I, Swaminathan B, Thomsen H, Turesson I, Vangsted A, Vogel U, Waage A, Walker B.A, Wihlborg A-K, Broyl A, Davies F.E, Thorsteinsdottir U, Langer C, Hansson M, Kaiser M, Sonneveld P, Stefansson K**, Morgan G.J**, Goldschmidt H**, Hemminki K**, Nilsson B** and Houlston R.S**.

Nature Communications, (2016). 7, 12050

III. **The multiple myeloma risk allele at 5q15 lowers *ELL2* expression and increases ribosomal gene expression**

Ali M, Ajore R, Wihlborg A-K, Niroula A, Swaminathan B, Johnsson E, Stephens O.W, Morgan G, Meissner T, Turesson I, Goldschmidt H, Mellqvist U-H, Gullberg U, Hansson M, Hemminki K, Nahi H, Waage A, Weinhold N, Nilsson B.

Nature Communications, (2018). 9, 1649

IV. **Sequence variation at the *MTHFD1L-AKAP12* and *FOPNL* loci does not influence multiple myeloma survival in Sweden**

Ali M, Lemonakis K, Wihlborg A-K, Veskovski L, Turesson I, Mellqvist U-H, Gullberg U, Hansson M and Nilsson B.

Submitted manuscript, (2018).

* denotes shared first author. ** denotes shared supervision.

Abbreviations

3C	Chromosome conformation capture
ANOVA	Analysis of variance
ATAC	Assay for transposase-accessible chromatin
Chip	Chromatin immunoprecipitation
CNV	Copy number variation
ELL2	Elongation factor for RNA polymerase II 2
ENCODE	Encyclopedia of DNA elements
eQTL	Expression quantitative trait loci
GLM	Generalized linear model
GWAS	Genome-wide association study
HWE	Hardy-Weinberg equilibrium
IBD	Identical-by-descent
Ig	Immunoglobulin
INDEL	Insertion and deletion
LD	Linkage disequilibrium
MAF	Minor allele frequency
MGUS	Monoclonal gammopathy of undetermined significance
MM	Multiple myeloma
MPRA	Massively parallel reporter assay
PCA	Principle component analysis
QC	Quality control
qPCR	Quantitative polymerase chain reaction
REMC	Roadmap epigenomics mapping consortium
SEC	Super elongation complex
SNV	Single nucleotide variation
TF	Transcription factor
WTCCC	Wellcome Trust Case Control Consortium

Chapter 1

Introduction and conceptual background

The study of genetics and biological basis of diseases is an active and productive area of research. DNA sequence variations are underlying causes of phenotypic differences in and among species. The haploid form of human genome, which is found in germ cells consists of three billion DNA base pairs, while the diploid form is found in the somatic cells and has twice the DNA content as the haploid form. Humans are identical in the 99.5 % of their genome[1]. Variation in the remaining 0.5% of the genome is what makes each of us unique. Genetic variation can divided into different forms according to the size and type of genomic variation. The frequency of variations decreased when size of variation increased[2]. The most common genetic variation in the human genome is single base-pair differences called single nucleotide variations (SNVs). In comparison to the human reference genome, around 85 million SNVs have been discovered[3]. Other types of variation are insertions and deletions (INDELs), where strings of base-pairs are inserted in or deleted from the genome of an organism. This can range from two to hundreds of base-pairs in length. The largest type of variation is structural variation which refers to changes in structure of chromosomes. The five most common types of structural variants are insertion, deletion, inversion, duplication and copy number variation (loss or gain) (CNV).

Advances in genomic technologies have made a significant impact on biomedical and biological researches. Two international scientific research projects are behind these rapid changes, the Human Genome Project and the International HapMap Project. The Human Genome Project is the world's largest collaborative project in biology to date. It was started in 1990 and was declared completed in 2003[4]. Its main goals were to determine the accurate sequence of the 3 billion DNA base pairs that make up human genome and to map all the human genes within the 23 pairs of chromosomes[5]. The International HapMap Project is another scientific effort to identify common genetic variations among people[6]. The aim of the International HapMap Project was to validate the millions of genetic variations that were identified during and after the completion of the Human Genome Project, and to characterize their correlation patterns in populations of European, Asian and

African ancestry[6,7]. Data from these projects are stored in computerized databases and are available worldwide.

Following the Human Genome and HapMap Projects, companies like Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA) developed cost-efficient high-throughput genotyping platforms with capacity of over one million SNV assays. These efforts together with reducing time of genome sequencing made it feasible to read the genome of individuals and accelerated the identification of genes and genetic variants that associated with developing diseases. By utilizing this technology, GWAS aims to scan the entire genome to detect variants that differ between a group with a particular trait and the control group.

Published GWASs have been catalogued by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/gwas/)[8]. For inclusion in the catalogue, studies must include an analysis of at least 100,000 variants with genome-wide coverage and SNV-trait associations must have a P-value $<1 \times 10^{-5}$ [8,9]. It is also possible to search for GWAS information in the Database of Genotypes and Phenotypes (dbGap, http://www.ncbi.nlm.nih.gov/gap) which is a National Institutes of Health-sponsored repository charged to archive, curate and distribute information produced by studies investigating the interaction of genotype and phenotype[10,11]. As of April 2018, the GWAS Catalogue contains 59,967 unique SNP-trait associations from 3,349 publications[8].

An example of early GWAS is the one undertaken in the British population published by Wellcome Trust Case Control Consortium (WTCCC) in 2007[12]. By using a chip that allowed for genotyping of 500,000 variants , they genotyped 2,000 patients for each of seven complex human diseases of public health importance including bipolar disorder, Crohn’s disease, coronary artery disease, hypertension, rheumatoid arthritis, type 1 and type 2 diabetes. For control, they used the genotype information of 3,000 healthy individuals. Following case-control comparisons they could identify 24 independent association signals[12]. Since then, GWAS became a popular tool and leaded to continuous identification of new susceptibility genes and variants in common, complex diseases.

Chapter 2 Identification of genetic variants through GWAS

GWAS serves to find out if any genetic variant is associated with a trait. In the context of predisposition for human diseases, GWAS employs a case-control setup that uses high-throughput genotyping technologies (typically SNP microarrays) to compare the genome of two large groups of individuals: one healthy control group and one case group affected by a disease. By examining the allele frequency of single-nucleotide variations (SNVs) across the genome, it finds variations that occur more frequently in people with a particular disease than in people without the disease[13]. Development of cost-efficient high-throughput genotyping platforms together with the development of biobanks and large-scale population-based registries (*e.g.,* in Iceland and the United Kingdom) has enabled GWASs comprising many thousands of cases and controls, and, as a result, around 60,000 disease-related genetic variants have been identified so far[8].

2.1 How to conduct a GWAS?

To carry out a case-control GWAS, germline DNA from affected patients and unaffected controls needs to be obtained, purified, and analysed[14]. Statistical tests for association are then applied to identify DNA sequence variants that are carried more frequently by cases than by controls.

Several factors need to be considered when designing a GWAS. Firstly, the sample size needs to be sufficiently large, in order for the study to have adequate power to detect. Secondly, the choice of genotyping technology will determine the number of variants analyzed in the study. Thirdly, the procedure of quality control on samples and genotypes needs to be considered. Fourthly, to increase the genomic resolution, untyped variants can be statistically predicted using a suitable set of reference haplotypes. Fifthly, a suitable statistical method needs to be selected for association testing. Finally, downstream of the discovery step, having access to a suitable replication set is essential for validating candidate variants identified in the discovery step, and if several independent studies existed, meta-analysis of results is a good strategy to increase the statistical power and to reduce false positive findings. In the rest of this chapter, each step has been discussed in more detail.

2.1.1 Quality Control

The ability of GWASs to identify true genetic associations depends on the quality of the data, sample sizes, geographical matches between case and control and minor allele frequency of variants. A number of quality control (QC) measures should be applied both on sample and SNV basis. QC procedures for GWAS have been described in many reviews[15-19].

For QC at the genotype level, SNV assays that failed on a large number of samples (>5%) are poor assays and should be removed from the study. It is also important to filter SNVs based on minor allele frequency. The suitable threshold depends on the size of the study and the expected effect sizes. At least for the sample sizes used in this study, statistical power is very low for rare SNVs (<1% MAF). Power calculation software can be used for power calculations and inform the allele frequency below which the study becomes underpowered. Checking for Hardy-Weinberg Equilibrium (HWE) is another important factor. Under Hardy-Weinberg assumptions, allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences such as selection, mutation and genetic drift. Departure from this equilibrium can be due to genotyping errors, technical batch effects, population stratification, or actual, strong association with the phenotype being studied[20].

For QC at the sample level, samples with low genotype efficiency or call rate and duplicated samples should be removed. Further, it is better to exclude Samples from closely related individuals because their genotype information is correlated and this could affect the statistical association analysis. Level of relatedness can be measured through the probabilities that two individuals share zero, one or two pairs of alleles that are identical-by-descent (IBD) [21,22]. Another source of bias can be due to population stratification which is the presence of a systematic difference in allele frequencies between subpopulations, possibly due to different ancestry[23]. Principal components analysis (PCA) is widely used to correcting for cryptic differences in population structure between cases and controls (*e.g.*, differences in geographic origin)[24]. It converts a set of correlated variables into a set of uncorrelated variables called principal components. The first component explains the most variation in the data, and each subsequent component accounts for another, smaller part of the variability. In our work, we used PCA to exclude samples of non-European origin.

2.1.2 Imputation

Imputation is the statistical prediction of genetic markers that are not directly genotyped. It increases the genomic resolution of association studies, and is useful for combining GWAS results generated from different genotyping platforms with different SNV content[25]. Imputation makes use of the fact that the human genome is arranged in blocks of high LD, separated by hotspots of recombination[26]. LD is the non-random association between alleles at different loci[27]. Two variants are said to be in LD if their genotypes show higher correlation than would be expected by chance. The correlation is measured by the r^2 statistic, which is the proportion of variation of one SNP explained by the other[28]. LD is influenced by physical genomic distance between markers, as well as by evolutionary forces such as rate of mutation, selection, genetic drift and rate of recombination between markers[29].

Imputation makes use of LD to interpolate sequence variants that detected in recent whole-genome sequencing efforts (*e.g.*, 1,000 Genomes[3] or the UK10K project[30]) into the SNP microarray data. For imputation to be effective, the reference haplotypes (*e.g.*, imputation training set) needs to be sampled in a geographic population that is at least closely related to the population from which the GWAS cases and controls come from. Algorithmically, imputation first requires phasing of the genotype data [31,32]. Phasing refers to the process where per-chromosome genotype patterns are reconstructed statistically from diploid genotype data. Genotype measures the unordered combination of alleles at each site, whereas haplotypes are the two sequences of alleles on the maternally and paternally inherited chromosomes, respectively. In essence, phasing amounts to first estimating reference haplotypes (typically done using an external data set), and then using reference haplotypes as models for decomposing the genotype data recorded in the study into per-chromosome genotype patterns (phased haplotypes). Once these patterns have been extracted, imputation algorithms essentially interpolate in-between genotypes[33,34].

2.1.3 Association testing

After quality control and imputation, the next step will be association analysis. Selection of suitable statistical test depends on the type of phenotypes. Case-control GWASs are analysed by either contingency table methods or logistic regression. In this work, we used the SNPTEST program which implements a logistic regression method that is essentially an extension of linear regression where the outcome of a linear model is transformed using a logistic function that predicts the probability of having case status given a genotype class[35,36]. For quantitative traits, generalized linear model (GLM) and the analysis of variance (ANOVA) can be used [35]. ANOVA is similar to linear regression model with categorical predictor variables (genotype classes) with the null hypothesis of there is no difference between the trait means of any genotype group. The assumptions of GLM and ANOVA are 1) the trait is normally distributed; 2) the trait variance within each group is the same and 3) the groups are independent[35].

Furthermore, the choice of test for association depends on the assumed inheritance model. Most commonly, association testing is carried out using logistic regression under the assumption of additive effects on the log-odds scale, and multiplicative effects on the odds scale. In additive model there is a uniform, linear increase in risk for each copy of the allele, whereas in multiplicative model the risk of disease is increased n-fold with each additional disease risk allele. Other models including dominant and recessive also existed[35,36].

2.1.4 Replication and meta-analysis

Because of large numbers of genotype-phenotype association tests performed in GWA studies, multiple hypothesis testing is important[35,37,38]. Bonferroni correction is a commonly used solution[39]. However, the use of Bonferroni tends to lead to overly conservative results, as it assumes independent hypotheses, but in context of GWAS variants in LD are correlated to some degree. Other criteria like weighted Bonferroni approach have been suggested. In this approach variants are divided into different classes based on their anticipated functional impact, and association testing in then done within each class[40]. For minimizing this effect, results of association testing should be tested in an additional independent sample set drawn from the same population as the GWAS[38]. Replication of results in an additional population helps ensure that a genotype-phenotype association observed in a (GWA) study is a valid association and is not a false discovery[35,41].

Meta-analysis of GWASs is a good strategy to increase the statistical power, reduce false-positive findings and increase effect size of the GWASs[42]. There are several meta-analysis methods for GWA studies. Methods based on *P* values are the simplest approaches which combine *p*-values of independent statistical tests using methods such as Fisher's combined probability test, although they ignore heterogeneity in parameters like genetic effect size and sample size[42]. Fixed effects meta-analysis is the most popular approach for combining case-control GWAS data[42,43]. It assumes that all studies have a similar genetic effect for all risk variants and differences between study findings are due to variation in sampling[44] however it is not a good approach in presence of heterogeneity. In case of heterogeneity, random effects meta-analysis can be used, as it assumes each study population has its own genetic effect size and the average effect over all potential populations is considered, however it will bias the *P*-values and is not suitable in discovery efforts[42,44].

2.2 What GWAS can do and what it cannot do?

The role of GWAS in identification of disease susceptibility genes is undeniable. Yet, this approach is not free from limitations.

Firstly, GWAS is biased towards common variants, as a result of how genotyping chips are designed. Commercial genotyping arrays primarily include variants with at least 5% minor allele frequency (MAF). This is suitable for studying diseases or traits where many common genetic variations contribute to a person's risk with low or moderate effect size[35,45]. However, GWAS is less powerful approach when it comes to identifying rare variants. Such variants are not represented on the genotyping chips, but need to be imputed based on the local patterns of common variants. This works well for some rare variants, but not for all, and the accuracy of the imputation is dependent on the quality and geographic representativeness of the reference haplotypes used with the imputation algorithms[34,46].

Secondly, SNVs in genotyping arrays are chosen based on the LD structure or pairwise correlation between variants. Therefore, it is common that variants identified by GWAS are not causal, but merely tag the real causal variant(s) [46,47]. In some cases, the detected variants can be considered causal (*e.g.*, if the association signal is represented by a single variant or if the LD block contains a variant with obvious functional impact such as a frame-shift variant). More often, however, multiple correlated variants are detected, and it is not clear which variant is causal. Thus, while LD significantly reduces the number of SNVs that needs to be genotyped directly and allows imputation of in-between variants[45], LD also makes it more complicated to interpret GWAS results, as sets of statistically inseparable variants are detected instead of single causal variant.

Thirdly, GWAS has limited ability to discover structural variants, including long insertions and deletions[48]. This is a result of difficulties in variant calling, as the detection of long INDELs is algorithmically more complicated than the detection of single-nucleotide polymorphisms or short INDELs, due to the limited read length available with Illumina technology. While other next-generation sequencing technologies allow for longer read lengths (*e.g.*, Nanopore or PacBio), these have not yet been used to generate any major sets of reference haplotype for imputation, and efforts are still needed to expand the ability of GWASs in identification of other structural variants[48].

Fourthly, evaluation of gene-environment interactions is usually ignored. Possible reasons for this can be the lack of information on environmental exposures and other (non-genetic) risk factors[13]. Having this type of information could increase our chance to detect genetic effects that only occur if the individual is exposed to a particular environmental factor.

Chapter 3

Characterization of variants discovered by GWAS

As mentioned, DNA sequence variants identified in association studies may not be the actual causal variants, because of LD and technical limitations. More follow-up studies are needed to identify causal variants and their biological consequences. Large consortia such as the 1000 Genomes Project[49] and the Encyclopedia of DNA Elements (ENCODE) [50] aim to survey the entire human genome and build a comprehensive description of common genetic variation in the human genome.

3.1 Fine mapping of associated loci

The SNVs represented on genotyping microarrays are chosen based on the LD structure or pairwise correlation. Consequently, the probability that the GWAS-identified variants be in LD with the real causal variant is high[46,47,51]. Identification of causal variants at risk loci is an active area of research[51,52]. The aim of fine mapping is to identify the truly functional variants underlying observed association signals. Imputation methods together with the reference panels such as 1,000 Genomes[3] and UK10K[30], can fill the gaps for variants that are not existed in the initial genotyping platform[47], but are limited by allele frequency, and by the quality and geographic representativeness of the reference haplotypes[53]. Another approach for fine-mapping is to sequence the locus in better detail in risk allele carriers in order to detect underlying causal variants that are not detected on the genotyping microarrays, or by imputation (*c.f.*, ref[54]).

3.2 Analysis of expression quantitative trait loci

The vast majority of variants identified by GWAS (~93%) map to non-coding regions, and likely exert their effect by altering gene expression rather than amino acid sequences. For example, variations in the regulatory sequences can affect the timing, tissue specificity and level of gene expression, or transcript structure. CNVs can alter the level of gene expression by changing the number of copies of a gene that is present in the genome or they can alter the regulatory elements like deletion or duplication of an activator or deactivation of a repressor[55].Variants that influence gene expression are termed to have Expression Quantitative Trait Loci (eQTLs) effect. These variants are usually located close to the gene they regulate (typically in the promoter or clusters of regulatory elements inside the gene), but may also be located farther away (e.g., in long-range enhancer elements) [56].

To perform eQTL analyses, the common approach is to use microarray or RNA sequencing data from a relevant tissue, collected in individuals who have also been genotyped for the variants of interest[57]. Statistical methods are then used to associate genotypes with gene expression[58]. By now comprehensive sets of eQTL data have been made publicly available in online databases, including Genotype-Tissue Expression (GTEx)[59], Blood eQTL[60], GEUVADIS project[61], MuTHER studies[62] and SNP and CNV Annotation (SCAN)[63].

3.3 Analysis of regulatory variants

Variations that map to regulatory regions (*e.g.*, promoters, enhancers) are likely to explain eQTLs. Regulatory regions are typically associated with some characteristics such as open chromatin accessibility (as evidenced by DNAase hypersensitivity or ATAC-seq), clustering of transcription factor binding sites and specific histone modifications (detected by chromatin immunoprecipitation). Consortia such as the ENCODE[50] and the Roadmap Epigenomics Mapping Consortium (REMC)[64] have used a variety of genome-wide methods to study the chromatin state of non-coding regions in the human genome in hundreds of different cell types.

Chromatin immunoprecipitation (ChIP) with high-throughput sequencing is used to identify protein binding sites along the chromosomal DNA[65]. Immunoprecipitation is a method of isolating a specific protein from a complex mixture such as a cell lysate or blood sample. To perform ChIP-seq chromatin is isolated from cells and fragmented. Immunoprecipitation is done by using an antibody that is specific to a transcription factor (TF) or other DNA-binding protein of interest (e.g., histones with specific modifications). The DNA is recovered, sequenced and aligned to a reference genome to determine specific protein binding loci[66].

To identify genomic regions with open chromatin, hence more likely regulatory, DNA hypersensivity or ATAC-seq can be used[67,68]. In ATAC-seq, a hyperactive Tn5 transposase is utilized to insert sequencing adapters into open chromatin regions. Transposases are enzymes catalyzing the movement of transposons to other parts in the genome and naturally they have a low level of activity. The mutated transposase employed in ATAC-seq has high activity which allows for efficient cutting of exposed DNA and simultaneous ligation of adapters[69]. Adapter-ligated DNA fragments are then isolated, amplified by PCR and used for High-throughput sequencing[67].

To understand how remote enhancers communicate with coding genes, techniques to map the 3D structure of chromatin are available, including Chromosome conformation capture (3C)-based methods that provide information about organization of chromatin within the three-dimensional nuclear space[70-74].

To test the impact of candidate regulatory variants on transcriptional activity, reported assays can be used (e.g., luciferase assays). Today, it is also possible to assess the transcriptional activity of thousands of variants in parallel using massively parallel reporter assays (MPRAs)[75] and by using technologies such as CRISPR-Cas9 it is possible to do genome editing such as adding, removing, modifying or replacing DNA at particular locations in the genome.

Chapter 4
Original work

The overall objective of this thesis is to identify DNA sequence variants that predispose for multiple myeloma (MM) and to understand how these variants promote MM development. The thesis is built up on four papers:

Paper I represents a GWAS on MM in Nordic populations. We used one case–control data set from Sweden and Norway, and one from Iceland. We identified one novel MM risk locus related to the *ELL2* (Elongation Factor for RNA Polymerase II 2) gene and a promising association with the *TOM1-HMGXB4* locus at 22q13. We confirmed all other loci (eight variants) that were known at the time.

Paper II represents a meta-analysis of association data on MM from our lab and five other research centres in the United Kingdom, Germany, Netherlands, the United States, and Iceland. Eight new MM risk loci were discovered and previously reported loci were confirmed.

Paper III focuses on the functional consequences of the *ELL2* MM risk allele. *ELL2* encodes a key component of the super elongation complex (SEC) and plays an important role in the production of Ig by plasma cells[76,77]. Paper III aims to understand its mechanism of action and provide mechanistic insight into MM predisposition. We show that the *ELL2* MM risk allele has a strong negative effect on *ELL2* expression as well as a positive, possibly compensatory, effect on the expression of genes involved in ribosome biogenesis.

Finally, Paper IV aims to understand whether inborn genetic variations could influence the survival of MM patients. While our data set was not large enough for a discovery GWAS, we tried to replicate two previously reported associations related to *MTHFD1L-AKAP12* and *FOPNL* loci. These loci were not robustly replicated in the original studies. In our data, which represent a population-based series of MM patients from Sweden, we did not see any evidence of association.

4.1 Multiple Myeloma

Multiple myeloma (MM) is the second most common hematologic malignancy. It originates in plasma cells, which develop from B cells and are responsible for producing immunoglobulins (Ig). In MM, malignant plasma cells accumulate in the bone marrow at the expense of normal haematopoiesis. Unlike normal plasma cells, which produce polyclonal Ig, the malignant MM plasma cells produce a monoclonal immunoglobulin ("M protein")[78]. According to the International Myeloma Working Group (IMWG) criteria, MM is defined by >10% monoclonal plasma cells in the bone marrow or >30 g/L M protein. MM is preceded by monoclonal gammopathy of undetermined significance (MGUS)[79], which is a common pre-malignant condition (3% of ≥50 year olds)[80] that is defined by the presence of the M-protein but without other signs or symptoms of MM, and progresses to MM at an annual rate of approximately 1% [80,81]. Clinically, MM is characterized by bone marrow failure, lytic bone lesions, hypercalcemia, and kidney failure. Major genetic subgroups include hyper-diploid MM and MM with translocations involving the immunoglobulin heavy chain (IgH) gene[82,83], for example t(11;14), t(4;14), t(6;14), t(14;16) and t(14;20) translocation[84]. Other common somatic genetic changes include point mutations in *BRAF*, *DIS3*, *FAM46C*, *KRAS*, *NRAS* and *TP53*[85]. The risk of developing MM is influenced by geographic origin (more common among Africans and African Americans)[86], age (more common above 65 years old) and gender (more common among men)[87]. While survival can be extended, MM remains an incurable and fatal disease[88].

Some cases of MM are thought to have an inherited background. Family studies indicate that first-degree relatives of patients with MM and MGUS have 2 to 4 times higher risk for MM, plus a higher risk of certain other malignancies[89-93]. These observations hint at the existence of DNA sequence variants that predispose for MM. At the start of my Ph.D. project, recent genome-wide association studies had identified eight common sequence variants that associate with MM[94-96].

4.2 Paper I

To identify DNA sequence variants that associate with MM in Nordic populations, we carried out a GWAS, based on one case–control data set from Sweden and Norway, and one from Iceland. For the Swedish-Norwegian discovery sample set, we obtained 1,668 and 157 samples from the Swedish National Myeloma Biobank (Skåne University Hospital, Lund, Sweden) and the Norwegian Biobank for

Myeloma (Trondheim, Norway). The samples were banked between 2003 and 2013, and we genotyped them using Illumina OmniExpress-Exome chips. For controls, we obtained SNV microarray information from previous studies on schizophrenia (n=3,754)[97] and from a population-based Swedish study of twins where we only used one individual from each pair of twins (n=9,835; TWINGENE, http://ki.se/sites/default/files/twingene_gwas_basic_info.pdf). We excluded SNVs showing >5% missing data, significant deviation from Hardy-Weinberg equilibrium, or discrepancies in allele frequency between genotyping batches. We excluded samples showing >5% missing data or excess heterozygosity, and samples from closely related individuals. Unobserved genotypes were imputed using phased haplotypes from the Phase I (b37) release of the 1,000 Genomes Project[49]. To avoid artifacts of cryptic population stratification, we included five principal components of the identity-by-state matrix. For the Icelandic discovery sample set, we used 480 patients from deCODE database diagnosed with MM from 1955 to 2013 and to increase power we expanded that with 251 cases of non-IgM MGUS patients from Landspitali University Hospital and the Icelandic Medical Center Laboratory in Mjodd. Association testing was performed using logistic regression under an additive genetic model.

For meta-analysis, we performed association testing in each discovery set separately and combined the results for variants that were shared by the Icelandic and Swedish data. The meta-analysis was done with a fixed effect model. After meta-analysis, we could identify seven MM associated loci at $P<5\times10^{-8}$ (**Table 4.1**). Four of these loci were previously known and three including 5q15 (*ELL2*), 5q31 (*ARHGAP26*), and 22q13 (*HMGXB4-TOM1*) were novel.

For replication, we obtained additional 223 MM cases from the Swedish National Myeloma Biobank and 363 MM cases from the University Hospital of Copenhagen. As controls in the replication sets, we used Swedish blood donors and randomly ascertained individuals from Denmark and Skåne County in Sweden. These samples were genotyped by qPCR for the lead variants at the novel loci. After replication, the *ELL2* remained genome-wide significant and the *TOM1* variant remained borderline significant.

Table 4.1:
MM risk loci identified through GWAS in Sweden-Norway and Iceland

Locus	Variant	A1/A2*	RAF	MM+MGUS		MM Only		Candidate Gene
				P	OR	*P*	OR	
3p22	rs73071352	A/G	0.13	3.1×10^{-8}	1.32	5.2×10^{-7}	1.32	*ULK4*
5q15	rs56219066	C/T	0.72	1.4×10^{-8}	1.23	6.5×10^{-8}	1.23	*ELL2*
5q31	rs74735889	C/T	0.003	5.9×10^{-10}	4.06	8.2×10^{-8}	3.85	*ARHGAP26*
6p21	rs6919908	T/C	0.23	6.3×10^{-10}	1.19	3.8×10^{-10}	1.19	*HLA*
7p15	rs57104699	A/C	0.66	2.3×10^{-8}	1.38	3.5×10^{-8}	1.38	*CDCA7L*
17p11	rs57968458	A/G	0.098	2.8×10^{-10}	1.26	5.7×10^{-11}	1.26	*TNFRSF13B*
22q13	rs138740	C/T	0.38	1.3×10^{-7}	1.19	1.7×10^{-8}	1.22	*TOM1*

***:** Risk alleles underlined.
RAF: risk allele frequency; ***P:*** meta-analyzed *P* values; **OR:** odd ratio
Novel Loci are shown in grey.

ELL2 is a key component of the SEC, which enhances the catalytic rate of RNA polymerase II transcription by suppressing its transient pausing activity along the DNA[98,99]. Conditional B-lineage *ELL2* knock-out mice show curtailed humoral immune responses, reduced numbers of plasma cells, and abnormal plasma cell morphology[76,100,101]. In addition to normal and malignant plasma cells, *ELL2* is also expressed in other cell types, including red blood cell precursors, salivary gland, and pancreatic islets. Similar to plasma cells, these cell types also produce large amounts of protein such as haemoglobin, amylase, and peptide hormones.

We did an eQTL study in peripheral blood and lymphoblastoid cell lines, but couldn’t detect any MM risk allele-associated effect on *ELL2* expression. We noted that one of the LD variants of the *ELL2* MM risk allele encodes a Thr298Ala missense variant in an *ELL2* domain required for transcription elongation, though we could not say for certain whether this variant was causal. However, because *ELL2* has been implicated in Ig synthesis, we tested for associations with IgA, IgG and IgM levels in 24,279, 21,981 and 20,413 Icelanders without MM or MGUS. We found that risk allele confers lower IgA and IgG levels. We saw similar effects in an independent set of 1,012 Swedish blood donors. These results indicated to us that the *ELL2* risk allele has a hypomorphic effect. Finally, we screened deCODE’s database for associations between the *ELL2* risk allele and other diseases and quantitative traits. Although we did not find any association with other B-lymphoid proliferative or malignant diseases apart from MGUS, we saw a potential association with an increased risk of bacterial meningitis (which could hypothetically be caused by the lower Ig levels).

4.3 Paper II

To increase statistical power in detecting MM risk loci, we and five other teams from the United Kingdom, Germany, Netherlands, the United States and Iceland carried out a joint meta-analysis of MM association data. After filtering, we reached a total of 7,319 discovery cases, which our lab contributed the second largest data set.

Quality control of samples and variants was as in paper I, although untyped variants were imputed using combination of 1,000 Genomes Project[2] and UK10K[30] and for Iceland, the in-house developed reference genome from deCODE Genetics was used. Association testing was done independently for each data set using logistic regression and then the results were meta-analyzed under the fixed-effects model. Promising associations were replicated using additional case–control series from the United Kingdom, Germany, Denmark and Sweden/Norway.

We confirmed nine known risk loci and discovered eight new loci with variants localize in or near *JARID2, ATG5, SMARCD3, CCAT1, CDKN2A, WAC, RFWD3* and *PREX1* genes (**Figure 4.1 and Table 4.2)**.

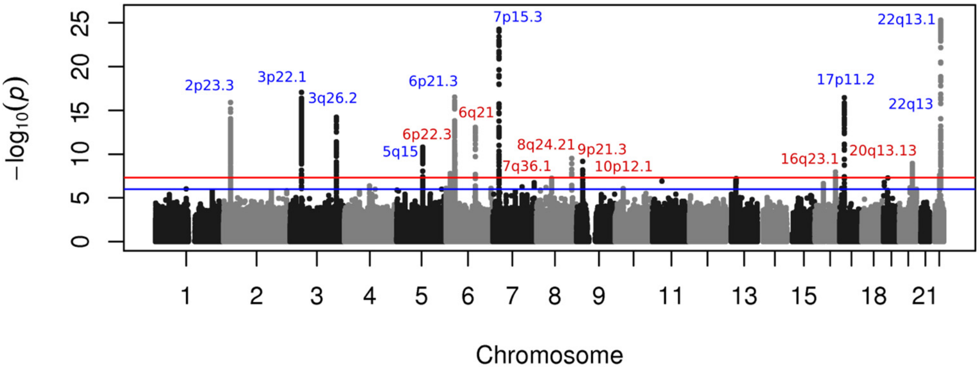

Figure 4.1: Manhattan plot from Paper II.
Genome-wide P-values (two sided) of 12.4 million successfully imputed autosomal variants in 7,319 cases and 234,385 controls from the discovery phase. Labelled in blue are known MM risk loci and labelled in red are newly identified risk loci.

Except for rs34229995, which is located 2.2 kb 5′ of *JARID2*, the other lead variants were found to map to the intragenic part of their transcribed genes. The only coding variant is rs7193541 in exon 10 of *RFWD3*. Expression quantitative trait loci (eQTL) analysis using gene expression profiles of CD138+ MM plasma cells from the United Kingdom, Germany and the United States identified significant associations between rs2790457 and decreased expression of *WAC* and between rs6066835 and increased expression of *PREX1*. *WAC* and *PREX1* also showed strong cis-methylation quantitative trait loci (meQTLs) with rs2790457 and rs6066835.We also noted that some of the identified genes like *WAC, ATG5* and *MYC* have role in autophagy[102-104].

Table 4.1:
Novel MM risk loci identified through meta-analysis

Locus	Variant	A1/A2*	RAF	Odd ratio	*P*-value	Candidate Gene
6p22.3	rs34229995	C/G	0.029	1.37	1.31×10^{-8}	*JARID2* (5' telomer)
6q21	rs9372120	G/T	0.218	1.18	9.09×10^{-15}	*ATG5*
7q36.1	rs7781265	T/C	0.125	1.19	9.71×10^{-9}	*SMARCD3*
8q24.21	rs1948915	T/C	0.345	1.13	4.20×10^{-11}	*CCAT1*
9p21.3	rs2811710	G/A	0.657	1.15	1.72×10^{-13}	*CDKN2A*
10p12.1	rs2790457	A/G	0.739	1.12	1.77×10^{-8}	*WAC*
16q23.1	rs7193541	C/T	0.585	1.13	5.00×10^{-12}	*RFWD3*
20q13.13	rs6066835	C/T	0.083	1.26	1.36×10^{-13}	*PREX1*

***:** Risk alleles underlined.
RAF: risk allele frequency

4.4 Paper III

To understand the role of *ELL2* risk alleles in developing MM, we performed eQTL analysis in CD138+ plasma cells from 1,630 MM patients from four populations. We showed that the MM risk allele lowers *ELL2* expression in these cells, but not in peripheral blood or other tissues. The *ELL2* MM risk loci is represented by about 67 single-nucleotide polymorphisms and 5 small insertions/deletions in tight LD (r^2>0.8). Hypothetically, only a few of these variants are causal, while the rest are linked tag variants with an effect of their own. To identify causal variants, we considered variants showing r^2>0.8 with rs9314162 that associate with both *ELL2* expression and MM and map to regulatory regions. To delineate regulatory regions, we used ChIP-seq data from ENCODE and Roadmap Epigenomics databases[64,105] and generated ChIP-seq data for H3K4me3 histone marks in the L363 plasma cell leukemia cell line. We identified eight candidate variants and to analyse their effect on transcriptional activity, we made luciferase vectors which were transfected into three MM plasma cell lines (L363, OPM2, and RPMI-8226) and two cell lines representing other hematologic lineages (K562 and MOLM-13). Three of the candidate variants (rs3777189-C, rs3777185-C and rs4563648-G) yielded lower luciferase activity relative to their corresponding protective variants in plasma cell lines, but not in non-plasma cell lines. Furthermore, gene set enrichment analysis with the Swedish-Norwegian mRNA sequencing data revealed that the MM risk allele associates with upregulation of gene sets related to ribosome biogenesis. To understand whether the association with ribosomal gene expression reflects a cause-effect relationship, we knocked out *ELL2* in L363 cells using CRISPR-Cas9 and analysed knockout and wild type cells by mRNA sequencing. L363-*ELL2*-KO cells showed a significant enrichment of increased expression for RPGs and other gene sets related to ribosome biogenesis and function. To exclude off-target effects, we did rescue experiments where *ELL2* expression was reconstituted in the L363-*ELL2*-KO cells, leading to downregulation of ribosomal genes in *ELL2*-transfected cells.

4.5 Paper IV

Recently, two meta-analyses by Johnson et al.[106]and Ziv et al.[107] reported associations between MM overall survival and inborn sequence variants at the *MTHFD1L-AKAP12* and *FOPNL* loci, respectively. In Johnson et al., no replication of the *MTHFD1L-AKAP12* locus was done after the initial discovery meta-analysis. In Ziv et al., replication analysis of the *FOPNL* locus was done, but the positive replication result was driven by a small sample subset (n=109) from Spain whereas six other replication subsets did not confirm the association.

Given that the two loci were not robustly replicated, we wondered if the *MTHFD1-AKAP12* and *FOPNL* associations can be detected in independent data. We therefore analysed a population-based series of 871 Swedish patients with MM, who had previously been genotyped in genome-wide association studies in Paper I and Paper II. We did not see any evidence of association between MM survival and the two reported loci.

Identifying predictive biomarkers is an important clinical goal, but is complicated by heterogeneity in treatment and other patient characteristics (*e.g.*, age). Our findings motivate the collection of larger data sets to understand the impact of inborn genetic variation on clinical outcome in MM and other hematologic malignancies.

Chapter 5
Concluding remarks

In all, the work carried out in this Ph.D. project contributed to our understanding of genetic predisposition for MM. The main conclusions are:

Paper I

- We identified a novel MM risk locus at 5q15 (*ELL2*) and a promising MM risk locus at 22q13 (*HMGXB4-TOM1*) (later validated in Paper II).
- We confirmed previously reported loci at *ULK4, HLA, CDCA7L* and *TNFRSF13B*.
- We found that the *ELL2* MM risk allele associates with decreased IgA and IgG in healthy carriers, indicating that it has a hypomorphic effect.

Paper II

- We identified eight new loci at *JARID2, ATG5, SMARCD3, CCAT1, CDKN2A, WAC, RFWD3* and *PREX1*.
- We confirmed all previously discovered risk loci plus *HMGXB4-TOM1*.

Paper III

- We showed that the MM risk allele lowers *ELL2* expression in plasma cells, consistent with its hypomorphic effect on Ig levels seen in Paper I.
- We identified candidate causal variants for the effect on *ELL2* expression.
- We identified upregulation of gene sets related to ribosome biogenesis as a downstream effect of the *ELL2* MM risk allele.

Paper IV:

- We tried to replicate two reported associations between the *MTHFD1L-AKAP12* and *FOPNL* loci and MM overall survival, as these two loci were not robustly replicated in the original studies.
- We could not see any evidence of association with survival at the two loci.

For future work, a number of questions remain to be addressed. Firstly, only a small part of the heritability of MM has been explained. This work has contributed to the identification of 10 new MM risk loci. Together with previous[94-96] and later (Went *et al.*, in revision) findings, this brings the total number of known MM risk alleles to 24. Nevertheless, these alleles only explain somewhere in the order of 20% of the estimated heritability (Went *et al.*, in revision), meaning that inherited MM susceptibility is far from completely understood. Extended association and family studies will need to be undertaken to identify additional risk alleles.

Secondly, we do not yet know how the identified MM risk variants promote MM development. This work has unveiled that the *ELL2* MM risk allele alters the expression of *ELL2* in *cis*, and has additional effects on global gene expression patterns in plasma cells. Although this provides a first glimpse into the basic molecular-genetic effects of the *ELL2* allele, our results do not explain why carriers of this allele actually have a higher risk of MM. For the remaining risk alleles, we know even less.

Thirdly, it remains to be understood how the findings from this work and other recent studies of MM predisposition can be translated into the clinic. The MM risk alleles identified so far are common variants, each of which confers a modest risk increase. Hence, looking at each one of these separately is not clinically meaningful. However, it could be the case that polygenic risk scores calculated from the combined genotype of all 24 risk loci (or a subset thereof) could provide sufficient predictive information, and recent analyses[108] indicate that somewhere in the order of 30% of familial MM cases carry a significantly larger burden of common MM risk alleles. However, the clinical usefulness of such polygenic risk scores remains to be understood.

Summary in English

Multiple myeloma (MM) is the second most common hematologic malignancy. The disease is characterized by an uncontrolled growth of malignant plasma cells in the bone marrow, producing a monoclonal immunoglobulin ("M protein") that can be detected in peripheral blood. Clinically, MM is characterized by bone marrow failure, lytic bone lesions, hypercalcemia, and kidney failure. It is preceded by monoclonal gammopathy of unknown significance (MGUS), a common condition defined as a clonal growth of plasma cells that does not yet satisfy the criteria for MM, but progresses to MM at a rate of 1% per year. Increasing evidence supports that the biology of MM is influenced by inborn genetic variation and first degree relatives of patients with MM and MGUS seem to have higher risk for MM, and a higher risk of certain other malignancies. In this Ph.D. project, we aim to find DNA sequence variants that predispose for MM, and understand how these variants contribute to MM development.

Case-control genome-wide association study (GWAS) is our approach for finding variants. Previous GWASes have identified eight genetic variants that associate with MM development. As a first step in this study, we also carried out a GWAS on a case-control data set from Sweden-Norway and Iceland (Paper I). Following statistical analysis we could identify one novel MM risk locus related to the *ELL2* (Elongation Factor for RNA Polymerase II 2) gene and a promising association with the *TOM1-HMGXB4* locus at 22q13 and we could confirm the previously reported loci. In the second part of the project, we carried out a meta-analysis of six genome-wide association studies together with the United Kingdom, Germany, Netherlands, the United States and Iceland. Previously reported loci were confirmed and eight new loci were discovered (Paper II). Understanding how the identified risk variants contribute to MM development is the next challenge. In the third part of the project we focused on the *ELL2* gene located on the locus 5q15 and tried to understand its mechanism of action in MM (Paper III). Finally, we retrieved clinical information for 871 patients diagnosed with MM from the Swedish Multiple Myeloma Registry. We tested for association between sequence variants and MM overall survival (Paper IV). Our findings provide further insights into the genetic and biological basis of MM predisposition.

Acknowledgements

I can't believe how fast time flies! It was 2010 when I was working for a management consulting company in Iran, co-operating in projects related to technology foresight. There, I realized that biology has a tremendous potential for research in 21st century. So, I became curious to learn more about it and since I have had academic background in engineering and management, Bioinformatics was the best choice for me. I applied for a Master's program in Bioinformatics and received an admission from Lund University. And… now, here I am in the charming city of Lund in Sweden, finalizing my doctoral dissertation.

Initiating and accomplishing this journey would not have been possible and enjoyable without the help and support of many people whom I would like to express my deepest gratitude. It is my honour to offer my special thank you to:

Professor **Björn Nilsson,** my main supervisor, for the full support and encouragement during these years. Being a leader of a multidisciplinary group is a complicated task. The way you manage people and projects is admirable and I appreciate your patience and understanding. You were with me from A to Z and it has been a great experience to work with you.

Markus Hansson, my co-supervisor, for providing me access to clinical data and believing in my abilities. Your feedback helped me to make sense of the clinician's point of view. Otherwise my world would be full of *P*-values and statistical graphs.

Professor **Urban Gullberg,** for his positive attitude and encouragement to my presentations and for critical comments on manuscripts.

Ingemar Turesson, the founder of the Swedish MM biobank. Without your efforts and marathon endurance, it would never have been possible to carry out large-scale studies on MM, and I am proud of being a part of this big data project.

All collaborators in Sweden, the Nordic region, and beyond. Without your help, none of this would have been possible. Particularly, I would like to thank **Unnur Thorteinsdottir**, **Gudmar Thorleifsson**, **Stefan Jonsson**, **Ingileif Jonsdottir**, **Kari Stefansson** and colleagues at deCODE, **Anders Waage** and colleagues in Trondheim, and **Annette Juul-Vangsted** and colleagues in Denmark for all your help with samples, genomics, informatics, and more.

Past and current members of BN group: **Ram, Anna-Karin, Abhishek, Ellinor, Aitzkoa, Britt-Marie, Maroulio, Jenny** and **Bhairavi,** for the great collaboration and inspiring discussions, especially in *ELL2* project, without your help we would not understand the mystery of *ELL2* to its current level and I am looking forward to hear more about it from you in near future. **Ludvig, Laura, Magnus, Mikael, Linnea, Åsa, Evelina, Caterina** and **Angelica,** you all created such a great multicultural atmosphere in the group and made my time enjoyable during fika and lunch time.

Giorgia, Annarita, Helena, Somayeh, Karina, Konstantinos, Tove, and **Jesper,** We shared our projects at some points and I learned a lot from you. I appreciate your friendship and wish you the bests in your life.

Pontus Eriksson, for teaching me R and always available for answering my questions, it doesn't matter what time of the day it is. You have a great talent in programing and you deserve the bests in your career.

Everyone in the BMC B13 crew. It is with you that I am experiencing the joyfulness of working in an international environment.

Susan and **Mina**, You deserve a super special thank you. Living far from home is challenging but I am very lucky to have you here in Sweden. Your advice, experiences and wisdom mean a lot to me. I had a great time with you talking about everything, from daily activities to economy, politics and psychology.

My twin brother, **Nima,** my dear sister, **Bita,** my brother in law, **Amir-Hossein,** my lovely two years old niece, **Vania** and my dearest mom and dad, **Manzar** and **Nasser**. Yes, I've finally accomplished that and we can spend more time together! This journey had its ups and downs. Thanks a lot for your invaluable support which helped me to deal with the downs and to dream bigger.

دوستتون دارم و باارزشترین دارایی زندگیم هستید.

References

1 Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human genome sequencing in health and disease. *Annu Rev Med* **63**, 35-61, doi:10.1146/annurev-med-051010-162644 (2012).

2 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

3 The Genomes Project, C. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393, http://www.nature.com/nature/journal/v526/n7571/abs/nature15393.html#supplementary-information (2015).

4 Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300**, 286-290, doi:10.1126/science.1084564 (2003).

5 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

6 International HapMap, C. A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).

7 International HapMap, C. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861, doi:10.1038/nature06258 (2007).

8 MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896-D901, doi:10.1093/nar/gkw1133 (2017).

9 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-1006, doi:10.1093/nar/gkt1229 (2014).

10 Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* **42**, D975-979, doi:10.1093/nar/gkt1211 (2014).

11 Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* **39**, 1181-1186, doi:10.1038/ng1007-1181 (2007).

12 Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).

13 Pearson, T. A. & Manolio, T. A. How to interpret a genome-wide association study. *JAMA* **299**, 1335-1344, doi:299/11/1335 [pii],10.1001/jama.299.11.1335 (2008).

14 Ding, C. & Jin, S. High-throughput methods for SNP genotyping. *Methods Mol Biol* **578**, 245-254, doi:10.1007/978-1-60327-411-1_16 (2009).

15 Turner, S. *et al.* Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* **Chapter 1**, Unit1 19, doi:10.1002/0471142905.hg0119s68 (2011).

16 Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol* **34**, 591-602, doi:10.1002/gepi.20516 (2010).

17 Weale, M. E. Quality control for genome-wide association studies. *Methods Mol Biol* **628**, 341-372, doi:10.1007/978-1-60327-367-1_19 (2010).

18 Teo, Y. Y. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr Opin Lipidol* **19**, 133-143, doi:10.1097/MOL.0b013e3282f5dd77 (2008).

19 Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564-1573, doi:10.1038/nprot.2010.116 (2010).

20 Wittke-Thompson, J. K., Pluzhnikov, A. & Cox, N. J. Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 967-986, doi:10.1086/430507 (2005).

21 Weir, B. S., Anderson, A. D. & Hepler, A. B. Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet* **7**, 771-780, doi:10.1038/nrg1960 (2006).

22 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575, doi:10.1086/519795 (2007).

23 Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997-1004 (1999).

24 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909, doi:10.1038/ng1847 (2006).

25 Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406, doi:10.1146/annurev.genom.9.081307.164242 (2009).

26 Morton, N. E. Linkage disequilibrium maps and association mapping. *J Clin Invest* **115**, 1425-1430, doi:10.1172/JCI25032 (2005).

27 Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**, 477-485, doi:10.1038/nrg2361 (2008).

28 VanLiere, J. M. & Rosenberg, N. A. Mathematical properties of the r2 measure of linkage disequilibrium. *Theor Popul Biol* **74**, 130-137, doi:10.1016/j.tpb.2008.05.006 (2008).

29 Schmidt, H. D. Principles of Population-Genetics - Hartl,Dl, Clark,Ag. *Homo* **41**, 289-289 (1991).

30 Consortium, U. K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90, doi:10.1038/nature14962 (2015).

31 Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511, doi:10.1038/nrg2796 (2010).

32 Howie, B. N., Donelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* **5**, e1000529, doi:10.1371/journal.pgen.1000529.g001 (2009).

33 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2012).

34 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6, doi:10.1038/nmeth.2307 (2013).

35 Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**, e1002822, doi:10.1371/journal.pcbi.1002822 (2012).

36 Lewis, C. M. Genetic association studies: design, analysis and interpretation. *Brief Bioinform* **3**, 146-153 (2002).

37 Stranger, B. E., Stahl, E. A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**, 367-383, doi:10.1534/genetics.110.120907 (2011).

38 Sale, M. M., Mychaleckyj, J. C. & Chen, W. M. Planning and executing a genome wide association study (GWAS). *Methods Mol Biol* **590**, 403-418, doi:10.1007/978-1-60327-378-7_25 (2009).

39 Johnson, R. C. *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* **11**, 724, doi:10.1186/1471-2164-11-724 (2010).

40 Sveinbjornsson, G. *et al.* Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet* **48**, 314-317, doi:10.1038/ng.3507 (2016).

41 Kraft, P., Zeggini, E. & Ioannidis, J. P. Replication in genome-wide association studies. *Stat Sci* **24**, 561-573, doi:10.1214/09-STS290 (2009).

42 Evangelou, E. & Ioannidis, J. P. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**, 379-389, doi:10.1038/nrg3472 (2013).

43 Pfeiffer, R. M., Gail, M. H. & Pee, D. On Combining Data From Genome-Wide Association Studies to Discover Disease-Associated SNPs. *Statistical Science* **24**, 547-560, doi:10.1214/09-Sts286 (2009).

44 Thompson, J. R., Attia, J. & Minelli, C. The meta-analysis of genome-wide association studies. *Brief Bioinform* **12**, 259-269, doi:10.1093/bib/bbr020 (2011).

45 Ku, C. S., Loy, E. Y., Pawitan, Y. & Chia, K. S. The pursuit of genome-wide association studies: where are we now? *J Hum Genet* **55**, 195-206, doi:10.1038/jhg.2010.19 (2010).

46 Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, 7-24, doi:10.1016/j.ajhg.2011.11.029 (2012).

47 Freedman, M. L. *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* **43**, 513-518, doi:10.1038/ng.840 (2011).

48 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

49 Consortium, T. G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

50 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

51 Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum Mol Genet* **24**, R111-119, doi:10.1093/hmg/ddv260 (2015).

52 Hormozdiari, F., Kichaev, G., Yang, W. Y., Pasaniuc, B. & Eskin, E. Identification of causal genes for complex traits. *Bioinformatics* **31**, i206-213, doi:10.1093/bioinformatics/btv240 (2015).

53 Wang, Z. & Chatterjee, N. Increasing mapping precision of genome-wide association studies: to genotype and impute, sequence, or both? *Genome Biol* **18**, 118, doi:10.1186/s13059-017-1255-6 (2017).

54 Storry, J. R. *et al.* Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype. *Nat Genet* **45**, 537-541, doi:10.1038/ng.2600 (2013).

55 Haraksingh, R. R. & Snyder, M. P. Impacts of variation in the human genome on gene regulation. *J Mol Biol* **425**, 3970-3977, doi:10.1016/j.jmb.2013.07.015 (2013).

56 Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120362, doi:10.1098/rstb.2012.0362 (2013).

57 Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* **27**, 72-79, doi:10.1016/j.tig.2010.10.006 (2011).

58 Gibson, G., Powell, J. E. & Marigorta, U. M. Expression quantitative trait locus analysis for translational medicine. *Genome Med* **7**, 60, doi:10.1186/s13073-015-0186-7 (2015).

59 Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank* **13**, 307-308, doi:10.1089/bio.2015.29031.hmm (2015).

60 Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-1243, doi:10.1038/ng.2756 (2013).

61 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511, doi:10.1038/nature12531 (2013).

62 Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* **7**, e1002003, doi:10.1371/journal.pgen.1002003 (2011).

63 Gamazon, E. R. *et al.* SCAN: SNP and copy number annotation. *Bioinformatics* **26**, 259-262, doi:10.1093/bioinformatics/btp644 (2010).

64 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

65 Raha, D., Hong, M. & Snyder, M. ChIP-Seq: a method for global identification of regulatory elements in the genome. *Curr Protoc Mol Biol* **Chapter 21**, Unit 21 19 21-14, doi:10.1002/0471142727.mb2119s91 (2010).

66 Liu, E. T., Pott, S. & Huss, M. Q&A: ChIP-seq technologies and the study of gene regulation. *BMC Biol* **8**, 56, doi:10.1186/1741-7007-8-56 (2010).

67 Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, 21 29 21-29, doi:10.1002/0471142727.mb2129s109 (2015).

68 John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43**, 264-268, doi:10.1038/ng.759 (2011).

69 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

70 Sati, S. & Cavalli, G. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma* **126**, 33-44, doi:10.1007/s00412-016-0593-6 (2017).

71 de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **26**, 11-24, doi:10.1101/gad.179804.111 (2012).

72 Barutcu, A. R. *et al.* C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization. *J Cell Physiol* **231**, 31-35, doi:10.1002/jcp.25062 (2016).

73 Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods* **3**, 17-21, doi:10.1038/nmeth823 (2006).

74 Grob, S. & Cavalli, G. Technical Review: A Hitchhiker's Guide to Chromosome Conformation Capture. *Methods Mol Biol* **1675**, 233-246, doi:10.1007/978-1-4939-7318-7_14 (2018).

75 Melnikov, A., Zhang, X., Rogov, P., Wang, L. & Mikkelsen, T. S. Massively parallel reporter assays in cultured mammalian cells. *J Vis Exp*, doi:10.3791/51719 (2014).

76 Benson, M. J. *et al.* Heterogeneous nuclear ribonucleoprotein L-like (hnRNPLL) and elongation factor, RNA polymerase II, 2 (ELL2) are regulators ofmRNA processing in plasma cells. *proc. Natl. Acad. Sci. USA* **109**, 16252–16257 (2012).

77 Santos, P., Arumemi, F., Park, K. S., Borghesi, L. & Milcarek, C. Transcriptional and epigenetic regulation of B cell development. *Immunol Res* **50**, 105-112, doi:10.1007/s12026-011-8225-y (2011).

78 Rajkumar, S. V. *et al.* International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol* **15**, e538-e548, doi:10.1016/S1470-2045(14)70442-5 (2014).

79 Weiss, B. M., Abadie, J., Verma, P., Howard, R. S. & Kuehl, W. M. A monoclonal gammopathy precedes multiple myeloma in most patients. *Blood* **113**, 5418-5422, doi:10.1182/blood-2008-12-195008 (2009).

80 Kyle, R. A. *et al.* Prevalence of Monoclonal Gammopathy of Undetermined Significance. *N Engl J Med* **354**, 1362-1369 (2006).

81 Landgren, O. *et al.* Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* **113**, 5412-5417, doi:10.1182/blood-2008-12-194241 (2009).

82 Dewald, G. W., Kyle, R. A., Hicks, G. A. & Greipp, P. R. The clinical significance of cytogenetic studies in 100 patients with multiple myeloma, plasma cell leukemia, or amyloidosis. *Blood* **66**, 380-390 (1985).

83 Debes-Marun, C. S. *et al.* Chromosome abnormalities clustering and its implications for pathogenesis and prognosis in myeloma. *Leukemia* **17**, 427-436, doi:10.1038/sj.leu.2402797 (2003).

84 Fonseca, R. *et al.* The recurrent IgH translocations are highly associated with nonhyperdiploid variant multiple myeloma. *Blood* **102**, 2562-2567, doi:10.1182/blood-2003-02-0493 (2003).

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

86 Landgren, O. *et al.* Risk of monoclonal gammopathy of undetermined significance (MGUS) and subsequent multiple myeloma among African American and white veterans in the United States. *Blood* **107**, 904-906, doi:10.1182/blood-2005-08-3449 (2006).

87 Kyle, R. A. *et al.* Review of 1027 patients with newly diagnosed multiple myeloma. *Mayo Clin Proc* **78**, 21-33, doi:10.4065/78.1.21 (2003).

88 Laubach, J., Richardson, P. & Anderson, K. Multiple myeloma. *Annu Rev Med* **62**, 249-264, doi:10.1146/annurev-med-070209-175325 (2011).

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

90 Landgren, O. *et al.* Risk of plasma cell and lymphoproliferative disorders among 14 621 first-degree relatives of 4458 patients with monoclonal gammopathy of undetermined significance in Sweden. *Blood* **114**, 791-795, doi:10.1182/blood-2008-12-191676,10.1182/blood-2008-12191676 (2009).

91 Altieri, A., Chen, B., Bermejo, J. L., Castro, F. & Hemminki, K. Familial risks and temporal incidence trends of multiple myeloma. *Eur J Cancer* **42**, 1661-1670, doi:10.1016/j.ejca.2005.11.033 (2006).

92 Morgan, G. J. *et al.* Inherited genetic susceptibility to multiple myeloma. *Leukemia* **28**, 518-524, doi:10.1038/leu.2013.344 (2014).

93 Vachon, C. M. *et al.* Increased risk of monoclonal gammopathy in first-degree relatives of patients with multiple myeloma or monoclonal gammopathy of undetermined significance. *Blood* **114**, 785-790, doi:10.1182/blood-2008-12-192575,10.1182/blood-2008-12192575 (2009).

94 Weinhold, N. *et al.* The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nat Genet* **45**, 522-525, doi:10.1038/ng.2583 (2013).

95 Broderick, P. *et al.* Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nat Genet* **44**, 58-61, doi:10.1038/ng.993 (2012).

96 Chubb, D. *et al.* 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 (2013).

97 Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**, 1150-1159, doi:10.1038/ng.2742 (2013).

98 Luo, Z., Lin, C. & Shilatifard, A. The super elongation complex (SEC) family in transcriptional control. *Nat Rev Mol Cell Biol* **13**, 543-547, doi:10.1038/nrm3417 (2012).

99 Liu, M., Hsu, J., Chan, C., Li, Z. & Zhou, Q. The ubiquitin ligase Siah1 controls ELL2 stability and formation of super elongation complexes to modulate gene transcription. *Mol Cell* **46**, 325-334, doi:10.1016/j.molcel.2012.03.007 (2012).

100 Milcarek, C., Albring, M., Langer, C. & Park, K. S. The eleven-nineteen lysine-rich leukemia gene (ELL2) influences the histone H3 protein modifications accompanying the shift to secretory immunoglobulin heavy chain mRNA production. *J Biol Chem* **286**, 33795-33803, doi:10.1074/jbc.M111.272096 (2011).

101 Park, K. S. *et al.* Transcription Elongation Factor ELL2 Drives Ig Secretory-Specific mRNA Production and the Unfolded Protein Response. *J Immunol*, doi:10.4049/jimmunol.1401608 (2014).

102 Joachim, J., Wirth, M., McKnight, N. C. & Tooze, S. A. Coiling up with SCOC and WAC: two new regulators of starvation-induced autophagy. *Autophagy* **8**, 1397-1400, doi:10.4161/auto.21043 (2012).

103 Cenci, S. Autophagy, a new determinant of plasma cell differentiation and antibody responses. *Mol Immunol* **62**, 289-295, doi:10.1016/j.molimm.2014.02.008 (2014).

104 Toh, P. P. *et al.* Myc inhibition impairs autophagosome formation. *Hum Mol Genet* **22**, 5237-5248, doi:10.1093/hmg/ddt381 (2013).

105 Rosenbloom, K. R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res* **41**, D56-63, doi:10.1093/nar/gks1172 (2013).

106 Johnson, D. C. *et al.* Genome-wide association study identifies variation at 6q25.1 associated with survival in multiple myeloma. *Nat Commun* **7**, 10290, doi:10.1038/ncomms10290 (2016).

107 Ziv, E. *et al.* Genome-wide association study identifies variants at 16p13 associated with survival in multiple myeloma patients. *Nat Commun* **6**, 7539, doi:10.1038/ncomms8539 (2015).

108 Halvarsson, B.-M. *et al.* Direct evidence for a polygenic etiology in familial multiple myeloma. *Blood Advances* **1**, 619-623 (2017).

Paper I

nature COMMUNICATIONS

ARTICLE

Received 18 Dec 2014 | Accepted 20 Apr 2015 | Published 26 May 2015

DOI: 10.1038/ncomms8213

OPEN

Variants in *ELL2* influencing immunoglobulin levels associate with multiple myeloma

Bhairavi Swaminathan[1,*], Guðmar Thorleifsson[2,*], Magnus Jöud[1,3,*], Mina Ali[1,*], Ellinor Johnsson[1], Ram Ajore[1], Patrick Sulem[2], Britt-Marie Halvarsson[1], Guðmundur Eyjolfsson[4], Vilhelmina Haraldsdottir[5], Christina Hultman[6], Erik Ingelsson[7], Sigurður Y. Kristinsson[8], Anna K. Kähler[6], Stig Lenhoff[9], Gisli Masson[2], Ulf-Henrik Mellqvist[10], Robert Månsson[11], Sven Nelander[12], Isleifur Olafsson[13], Olof Sigurðardottir[14], Hlif Steingrimsdóttir[5], Annette Vangsted[15], Ulla Vogel[16], Anders Waage[17], Hareth Nahi[11], Daniel F. Gudbjartsson[2], Thorunn Rafnar[2], Ingemar Turesson[9], Urban Gullberg[1], Kári Stefánsson[2,**], Markus Hansson[1,9,**], Unnur Thorsteinsdóttir[2,**] & Björn Nilsson[1,3,18,**]

Multiple myeloma (MM) is characterized by an uninhibited, clonal growth of plasma cells. While first-degree relatives of patients with MM show an increased risk of MM, the genetic basis of inherited MM susceptibility is incompletely understood. Here we report a genome-wide association study in the Nordic region identifying a novel MM risk locus at *ELL2* (rs56219066T; odds ratio (OR) = 1.25; $P = 9.6 \times 10^{-10}$). This gene encodes a stoichiometrically limiting component of the super-elongation complex that drives secretory-specific immunoglobulin mRNA production and transcriptional regulation in plasma cells. We find that the MM risk allele harbours a Thr298Ala missense variant in an *ELL2* domain required for transcription elongation. Consistent with a hypomorphic effect, we find that the MM risk allele also associates with reduced levels of immunoglobulin A (IgA) and G (IgG) in healthy subjects ($P = 8.6 \times 10^{-9}$ and $P = 6.4 \times 10^{-3}$, respectively) and, potentially, with an increased risk of bacterial meningitis (OR = 1.30; $P = 0.0024$).

[1] Hematology and Transfusion Medicine, Department of Laboratory Medicine, Lund University, BMC B13, SE-221 84 Lund, Sweden. [2] deCODE genetics, Sturlugata 8, IS-101 Reykjavik, Iceland. [3] Clinical Immunology and Transfusion Medicine, Laboratory Medicine, Office of Medical Services, Akutgatan 8, SE-221 85 Lund, Sweden. [4] The Laboratory in Mjodd, IS-109 Reykjavik, Iceland. [5] Department of Hematology, Landspitali, The National University Hospital of Iceland, IS-101 Reykjavik, Iceland. [6] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [7] Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, SE-751 85 Uppsala, Sweden. [8] Faculty of Medicine, University of Iceland, IS-101 Reykjavik, Iceland. [9] Hematology Clinic, Skåne University Hospital, SE-221 85 Lund, Sweden. [10] Section of Hematology, Sahlgrenska University Hospital, SE-413 45 Gothenburg, Sweden. [11] Center for Hematology and Regenerative Medicine, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [12] Department of Immunology, Pathology and Genetics, Uppsala University, Rudbeck Laboratory, SE-751 05 Uppsala, Sweden. [13] Department of Clinical Biochemistry, Landspitali, The National University Hospital of Iceland, IS-101 Reykjavik, Iceland. [14] Department of Clinical Biochemistry, Akureyri Hospital, IS-600 Akureyri, Iceland. [15] Department of Haematology, University Hospital of Copenhagen at Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. [16] National Research Centre for the Working Environment, Lersø Parkallé 105, DK-2100 Copenhagen, Denmark. [17] Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Box 8905, N-7491 Trondheim, Norway. [18] Broad Institute, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. * These authors contributed equally to this work. ** These authors jointly supervised this work. Correspondence and requests for materials should be addressed to U.T. (email: unnur.thorsteinsdottir@decode.is) or to B.N. (email: bjorn.nilsson@med.lu.se.).

Multiple myeloma (MM) is characterized by an uninhibited, clonal growth of plasma cells in the bone marrow, producing a monoclonal immunoglobulin ('M protein') that can be detected in peripheral blood[1]. According to the International Myeloma Working Group criteria, MM is defined by >10% monoclonal plasma cells in the bone marrow or >3 g M protein per 100 ml plasma. Characteristic symptoms include calcium elevation, renal insufficiency, anaemia and lytic bone lesions or osteoporosis. While survival can be extended, MM remains an incurable and fatal disease[2]. It is preceded by monoclonal gammopathy of unknown significance (MGUS)[3,4], a common condition (3% of ≥50 year olds)[5] defined as a clonal growth of plasma cells that does not yet satisfy the criteria for MM, but progresses to MM at a rate of ~1% per year.

Since the 1970s, several authors have reported families with multiple cases of MM, including pedigrees suggesting Mendelian inheritance[6,7]. This century, systematic family-based studies, including in population-based registries, confirmed that first-degree relatives of patients with MM and MGUS have 2–4 times higher risk for MM, and a higher risk of certain other malignancies[8–12]. These data support the existence of MM risk alleles. Recent genome-wide association studies have identified eight common sequence variants that associate with MM, and account for an estimated 13% of the familial risk[13–15]. The molecular basis of inherited MM susceptibility is thus incompletely understood.

Here we report a genome-wide association study aimed at identifying DNA sequence variants that predispose for MM in Nordic populations. We identify a novel risk locus at the *ELL2* gene at 5q31 that encodes a key component of the super-elongation complex (SEC) that drives secretory-specific Ig mRNA production and transcriptional regulation in plasma cells. We also identify a promising association with the *TOM1-HMGXB4* locus at 22q13. We find that the *ELL2* MM risk allele harbours a Thr298Ala missense variant in an *ELL2* domain required for transcription elongation. Consistent with a hypomorphic effect, we find that the MM risk allele also associates with reduced levels of IgA and IgG in healthy subjects and, potentially, with an increased risk of bacterial meningitis.

Results

Genome-wide association study. To identify MM risk loci, we carried out a genome-wide association study based on one case–control data set from Sweden and Norway, and one from Iceland (Table 1). For the Swedish–Norwegian data set, variants identified by the 1,000 Genomes consortium were imputed into genotype data generated on Illumina single-nucleotide polymorphism (SNP) microarrays. For the Icelandic data set, variants were identified by whole-genome sequencing of 2,636 Icelanders[16], and imputed into 104,220 Icelanders genotyped with Illumina SNP chips[17,18]. Using the Icelandic genealogy, we additionally calculated genotype probabilities for 294,212 relatives of the chip-typed individuals[16].

We performed association testing in the Swedish–Norwegian and Icelandic data sets, and combined the results for 12.1 million variants that passed quality filtering. Two versions of the Icelandic case–control data were used for meta-analysis: one with MM patients only, and one that was expanded with non-IgM MGUS patients to increase power (Table 1). The latter is motivated because MM evolves from MGUS[3,19], relatives of MGUS patients have increased MM risk[9,12] and known MM risk alleles tend to associate with MGUS[11]. We replicated all known MM risk loci in both meta-analyses (Supplementary Fig. 1 and Supplementary Table 1; refs 13–15). Quantile–quantile analysis showed minimal P value inflation (genomic inflation factor $\lambda = 1.005–1.020$; Supplementary Fig. 2).

Seven loci associated with MM or MM+MGUS at $P<5\times10^{-8}$ (calculated using logistic regression as described in Methods section). These included four known MM risk loci (Fig. 1a and Supplementary Table 1; refs 14,15) and three previously unknown loci at 5q15 (*ELL2*), 5q31 (*ARHGAP26*) and 22q13 (*HMGXB4-TOM1*; Supplementary Table 2). Inclusion of the Icelandic MGUS cases strengthened the associations with 5q15 and 5q31. The signals at 5q15 (rs56219066; risk allele frequency (RAF) 71.1–73.2%) and 22q13 (rs138740; RAF 36.4–41.5%) were represented by common variants with moderate effects (odds ratio (OR) = 1.20–1.39; Fig. 1a,b and Table 2), whereas the 5q31 signal (rs74735889; RAF ~0.3%) was represented by an imputed rare variant that lost significance (OR = 1.69; logistic regression $P=0.014$) when genotyped directly and was not investigated further. Examining the expression patterns of *ELL2*, *TOM1* and *HMGXB4* across different types of blood cells, we noted that *ELL2* and *TOM1* are preferentially expressed in normal and malignant plasma cells (Fig. 1c). Conditional analysis did not reveal any underlying independent association signals at the *ELL2* or *HMGXB4-TOM1* loci.

To validate the 5q15 and 22q13 loci, we genotyped an additional 586 MM cases and 2,111 controls from Sweden and Denmark for rs56219066 and rs138740 (Table 1). The rs56219066 SNP replicated in these samples (logistic regression $P=0.0046$) and reached genome-wide significance under Bonferroni correction when the discovery and replication sets were combined (meta-analysis $P=9.6\times10^{-10}$; Table 2). While rs138740 did not replicate, it remained borderline significant (meta-analysis $P=5.7\times10^{-8}$) when the discovery and replication sets were combined and we observed effects in the same direction as in the meta-analysis (OR = 1.04–1.08; Table 2). Further validation in larger data sets is needed to confirm the 22q13 locus.

ELL2 regulates RNA processing in plasma cells. The association with 5q15 was captured by numerous markers in strong linkage disequilibrium distributed across a ~40-kb haplotype block in *ELL2* (elongation factor, RNA polymerase II, 2; previously eleven-nineteen lysine-rich leukaemia gene 2) (Fig. 1b). This gene encodes a stoichiometrically limiting component of the SEC[20], which mediates rapid gene induction by suppressing transient pausing of RNA polymerase II activity along the DNA[21]. Strikingly, ELL2 and the SEC play an important role in the differentiation of mature B cells into plasma cells[22,23]. In mature and memory B cells, which express *ELL2* at a low level, *IGH*-mRNA is translated to membrane-bound Ig. In plasma cells, *ELL2* is highly expressed and helps RNA polymerase II find a promoter-proximal weak poly(A)-site that is essentially hidden in B cells. This causes *IGH*-mRNA to be translated to secreted Ig at a high rate[24,25]. B cell-lineage *ELL2* conditional knockout mice exhibit curtailed humoral responses to immunization, reduced numbers of plasma cells in the spleen and fewer antibody-producing cells in the bone marrow. Plasma cells isolated from these mice show a paucity of secreted IgH and a distended endoplasmic reticulum[26]. Silencing of *ELL2* in mouse plasmacytoma cell lines using RNA interference decreases the ratio of secreted versus membrane-encoding Ighg2b transcripts[27]. RNA sequencing studies suggest that, in addition to the *IGH*-mRNA, *ELL2* influences the processing of ~12% of transcripts expressed in plasma cells, including those of the plasma cell survival receptor *Tnfrsf17* (B-cell maturation antigen (BCMA))[22,26] and the *MYC* oncogene[28].

To characterize the *ELL2* risk allele, we analysed SNP and gene expression profiles from peripheral blood (eight data sets

Table 1 | Study populations.

	N	Per cent male	Age at diagnosis (years ± s.d.)	Genotyping method
Discovery samples				
Sweden and Norway				
Cases	1,714	57.1%	67.5 ± 11.4	Illumina OmniExpress-Exome[1]
Controls	10,391	51.5%	—	Illumina OmniExpress[1]
Iceland (MM)				
Cases	480	47.9%	71.2 ± 10.0	Illumina microarrays (n = 174); familially imputed (n = 306)[2]
Controls	212,164	48.9%	—	Illumina microarrays (n = 82,742); familially imputed (n = 129,422)[2]
Iceland (MM + MGUS)				
Cases	731	49.5%	71.0 ± 12.4	Illumina microarrays (n = 332); familially imputed (n = 399)[2]
Controls	283,999	48.8%	—	Illumina microarrays (n = 90,568); familially imputed (n = 193,431)[2]
Replication samples				
Sweden				
Cases	223	—	—	Selected SNPs
Controls	1,285	—	—	Selected SNPs
Denmark				
Cases	363	—	—	Selected SNPs
Controls	826	—	—	Selected SNPs

MGUS, monoclonal gammopathy of unknown significance; MM, multiple myeloma; SNP, single-nucleotide polymorphism.

[1] Imputed using whole-genome sequence data from 1,000 Genomes.

[2] Imputed using whole-genome sequence data from 2,636 Icelanders.

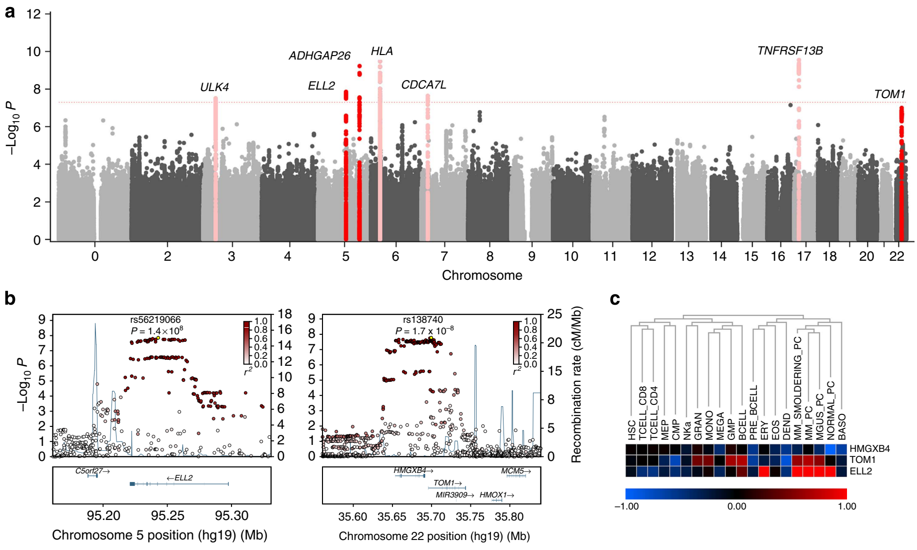

Figure 1 | Identification of *ELL2* at 5q15 as a novel MM risk locus and *HMGXB4-TOM1* at 22q13 as a candidate MM risk locus. (**a**) Manhattan plot for the meta-analysis of the Swedish-Norwegian and Icelandic MM data sets for 12.1 million SNPs that passed quality filtering. Seven loci showed association with MM or MM + MGUS at meta-analysis $P<5\times10^{-8}$, including four known MM risk loci (pink) and three previously unknown loci at 5q15 (*ELL2*), 5q35 (*ARHGAP26*) and 22q13 (*HMGXB4* and *TOM1*) (red). The *x* axis indicates genomic position of the SNPs. The *y* axis indicates the $-\log_{10}$ of the combined *P* values. The dotted line indicates the threshold for genome-wide significance of meta-analysis $P<5\times10^{-8}$. The results shown were obtained with the MM + MGUS version of the Icelandic data. Similar results were obtained with the MM version (not shown). (**b**) Regional association plots of the novel risk locus at *ELL2* and the tentative risk locus at *HMGXB4-TOM1*. Positions and *P* values of SNPs indicated on the *x* and *y* axes, respectively. Degree of linkage disequilibrium with sentinel SNPs indicated in shades of red. Blue background curves indicate meiotic recombination rates. The signal at *ARHGAP26* was represented by an imputed rare variant that lost significance when genotyped directly, and was not investigated further. (**c**) Expression of *ELL2*, *TOM1* and *HMGXB4* in 20 different types of blood cells (Affymetrix microarrays). *ELL2* and *TOM1* are preferentially expressed in plasma cells. BASO, basophils; BCELL, B cells; CMP, common myeloid progenitors; EOS, eosinophils; ERY, erythroid progenitors; GMP, granulocyte–monocyte progenitors; HSCs, haematopoietic stem cells; MEGA, megakaryocytes; MEP, megakaryocyte–erythrocyte progenitors; MONO, monocytes; NEU, neutrophils; NK, natural killer cells; PC, CD138^{+} plasma cells; PRE-B, pre-B cells; TCELL, T cells.

Table 2 | Association of sequence variants in or near *ELL2* and *TOM1*.

Populations	EAF	MM		MM + MGUS	
		OR (95% CI)	*P* value	OR (95% CI)	*P* value
ELL2—rs56219066-T					
Discovery					
Sweden/Norway	0.732	1.20 (1.11–1.32)	3.8×10^{-5}	1.20 (1.11–1.32)	3.8×10^{-5}
Iceland	0.711	1.39 (1.17–1.64)	1.1×10^{-4}	1.32 (1.15–1.51)	3.9×10^{-5}
Combined		1.23 (1.14–1.33)	6.5×10^{-8}	1.23 (1.15–1.33)	1.4×10^{-8}
Replication					
Denmark (363/826)	0.732	1.28 (1.04–1.57)	0.017	1.28 (1.04–1.57)	0.017
Sweden (223/1285)	0.735	1.30 (1.03–1.64)	0.030	1.30 (1.03–1.64)	0.030
Combined		1.29 (1.08–1.54)	0.0046	1.29 (1.08–1.54)	0.0046
Combined discovery and replication		1.25 (1.16–1.34)	9.6×10^{-10}	1.24 (1.16–1.33)	2.2×10^{-10}
TOM1—rs138740-C					
Discovery					
Sweden/Norway	0.364	1.20 (1.11–1.30)	2.4×10^{-6}	1.20 (1.11–1.30)	2.4×10^{-6}
Iceland	0.415	1.26 (1.09–1.46)	0.0017	1.15 (1.03–1.30)	0.015
Combined		1.22 (1.13–1.30)	1.7×10^{-8}	1.19 (1.11–1.27)	1.3×10^{-7}
Replication					
Denmark (352/815)	0.360	1.08 (0.90–1.30)	0.39	1.08 (0.90–1.30)	0.39
Sweden (235/1285)	0.364	1.04 (0.85–1.26)	0.72	1.04 (0.85–1.26)	0.72
Combined		1.06 (0.93–1.21)	0.38	1.06 (0.93–1.21)	0.38
Combined discovery and replication		1.18 (1.11–1.25)	5.7×10^{-8}	1.16 (1.10–1.23)	2.7×10^{-7}

Abbreviations: CI, confidence interval; EAF, effect allele frequency; MGUS, monoclonal gammopathy of unknown significance; MM, multiple myeloma; OR, odds ratio.
Association results for *ELL2* rs56219066 and *TOM1* rs138740 in the discovery samples from Sweden, Iceland and Norway and in the replication samples from Sweden and Denmark. Logistic regression and meta-analysis *P* values indicated.

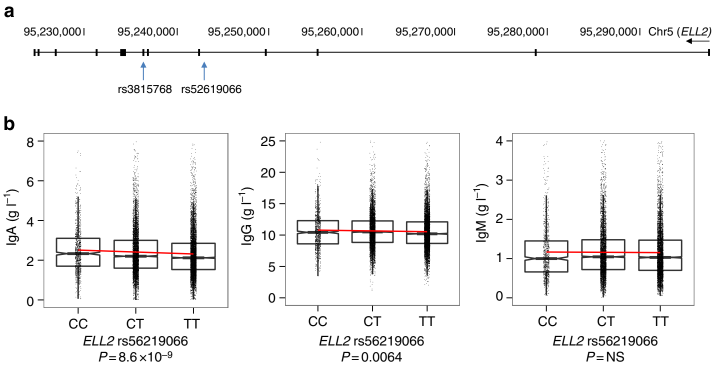

Figure 2 | The *ELL2* MM risk allele harbours a Thr298Ala missense variant, and the sentinel SNP rs56219066 is associated with reduced Ig levels. (**a**) Schematic representation of the *ELL2* gene showing the location of the sentinel SNP rs56219066 in intron 4 and the correlated variant rs3815768 in exon 7, which causes a Thr298Ala substitution in an *ELL2* domain required for transcription elongation. (**b**) We analysed blood IgA, IgG and IgM levels from 24,279, 21,981 and 20,413 Icelandic individuals without MM or MGUS. We found a significant association between IgA and IgG levels and the *ELL2* risk allele (log-linear regression *P* values shown). Compared with rs56219066C homozygotes, rs56219066T heterozygotes and homozygotes show 5.2 and 10.1% lower IgA and 2.6 and 5.1% lower IgG, respectively. We observed similar effects for IgA and IgG in an independent set of 1,012 Swedish blood donors (Supplementary Fig. 3). Boxes indicate medians and the first and third quartiles. Whiskers indicate first and third quartiles 1.5 times the interquartile range or the minimum/maximum values. Notches indicate confidence intervals around the median. NS, not significant.

totalling 9,087 samples) and lymphoblastoid cell lines (two data sets totalling 1,188 samples). We did not detect any risk allele-associated effect on *ELL2* expression (not shown). Because we did not have access to expression data from plasma cells from genotyped individuals, we could not exclude a plasma cell-specific effect on *ELL2* expression. Among the associated variants, however, we identified a Thr298Ala missense variant in *ELL2* exon 7 (rs3815768) in tight linkage disequilibrium ($D'/r^2 = 1.00/0.957$) with the sentinel SNP rs56219066 in intron 4 (Fig. 2a). The missense variant is located at the end of a ELL2 domain required for transcription elongation[29].

The *ELL2* risk allele associates with decreased IgA and IgG. Because of the recent mouse studies implicating *ELL2* in the production of secreted Igs, we tested for associations with blood

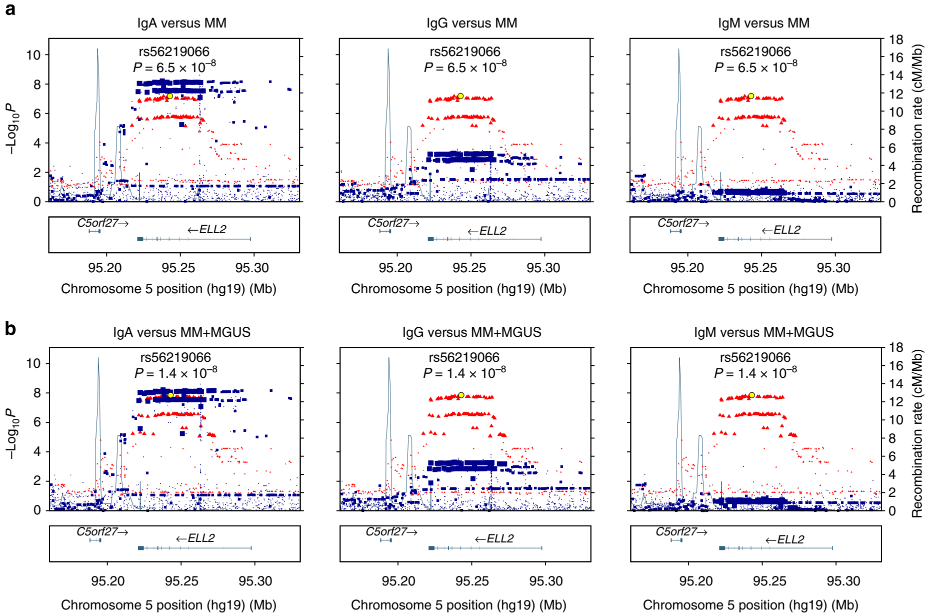

Figure 3 | The *ELL2* haplotype that predisposes for MM is identical to the *ELL2* haplotype that influences IgA and IgG levels in healthy subjects. In addition to the sentinel SNP rs56219066, the association between 5q15 identified in the MM and MM + MGUS meta-analyses was captured by numerous markers in strong linkage disequilibrium located in a ~40-kb haplotype block in *ELL2*. To verify that the *ELL2* haplotype associating with MM and MM + MGUS is identical to the *ELL2* haplotype associating with Ig levels, we tested for association between each available SNP in the *ELL2* region and IgA, IgG and IgM levels using the Icelandic Ig data set. We overlaid the log-linear regression *P* values for association with Ig levels with the meta-analysis *P* values obtained for the same SNPs for association with MM and MM + MGUS: (**a**) log-linear regression *P* values for association with IgA (red; left), IgG (red; middle) and IgM (red; right) overlaid on logistic regression *P* values for association with MM (blue); (**b**) corresponding results for MM + MGUS. The *x* axes indicate chromosomal positions. The *y* axes indicate $-\log_{10} P$ values. Sizes of markers reflect degree of association with MM or MM + MGUS. As shown, all SNPs in the ~40-kb haplotype block associating with MM or MM + MGUS associate with IgA and, to a lesser extent, with IgG. We did not observe any association with IgM. Taken together, *ELL2* SNPs associating with MM and MM + MGUS associate with IgA and IgG and vice versa, further supporting that the *ELL2* haplotype that predisposes for MM also influences Ig levels.

Ig levels in 20,413–24,279 Icelanders without MM or MGUS. Risk allele carriers showed lower IgA (log-linear regression $P = 8.6 \times 10^{-9}$) and IgG levels (log-linear regression $P = 6.4 \times 10^{-3}$; Fig. 2b). The *ELL2* haplotype associating with IgA and IgG levels was identical to the haplotype associating with MM (Fig. 3). Compared with rs56219066C homozygotes, rs56219066T heterozygotes and homozygotes showed 5.2 and 10.1% lower IgA and 2.6 and 5.1% lower IgG, respectively (Fig. 2b). We observed similar effects in an independent set of 1,012 Swedish blood donors (Supplementary Fig. 3). The risk allele does not associate with IgM levels. These results, together with the reduced Ig levels in the *ELL2* conditional knockout mice[26], suggest that the MM risk variant reduces, rather than enhances, the function of ELL2 in plasma cells.

The *ELL2* risk allele associates with bacterial meningitis. Finally, to test whether the *ELL2* risk allele affects the risk of other diseases and traits, we screened deCODE's databases harbouring about 400 independent and uncorrelated diseases and quantitative traits. While we did not find any association with other B-lymphoid proliferative or malignant diseases apart from MGUS (OR = 1.19; logistic regression $P = 0.0018$), we observed associations between rs56219066T and lower total serum protein levels ($n = 20{,}100$; log-linear regression $P = 0.0014$; $\beta = -0.035$) as previously reported for rs3777200 in *ELL2* ($D'/r^2 = 1.00/0.96$ with rs56219066; ref. 30), and an increased risk of bacterial meningitis ($n = 512$; OR = 1.30; logistic regression $P = 0.0024$). The meningitis risk could be mediated through the reduced IgA and IgG levels.

Discussion

We have identified a previously unknown MM risk locus at 5q15 (*ELL2*) and a promising MM risk locus at 22q13 (*HMGXB4-TOM1*). Neither of these loci has been previously associated with MM or other lymphoid malignancies. The identified risk variants are common and their estimated effect sizes are similar to those of previously identified MM risk variants[13–15].

While the mechanisms that promote development of MM await further exploration, our findings indicate that the *ELL2* risk allele affects plasma cell function. The fact that *ELL2* regulates mRNA processing in plasma cells is compelling, as is the reduction of IgA and IgG levels associated with the risk allele.

While the altered Ig levels as such are unlikely to be the MM-predisposing event (as other variants in the Icelandic data that alter Ig levels do not predispose for MM; not shown), these changes could reflect a hypomorphic effect on the SEC that affects mRNA processing broadly, which could predispose for malignant transformation.

Furthermore, the lower Ig levels could make *ELL2* risk allele carriers susceptible to infections. The potential association with meningitis is therefore intriguing. While these carriers are certainly not severely immunodeficient (because the allele is common), it is well known that various types of limited Ig deficiency (for example, IgG2 and IgG3 subclass deficiency) confer an increased incidence of infections, including with *Neisseria meningitidi*s and other meningitis pathogens. Future studies will uncover the role of *ELL2* in haematological malignancies and immune response.

Methods

Study populations. For the Swedish-Norwegian discovery sample set, we obtained 1,668 and 157 samples from the Swedish National Myeloma Biobank (Skåne University Hospital, Lund, Sweden) and the Norwegian Biobank for Myeloma (Trondheim, Norway), respectively. The samples were banked between 2003 and 2013. In addition, we obtained SNP microarray profiles of population-based controls from previous studies of twins ($n = 9835$; TWINGENE, http://ki.se/sites/default/files/twingene_gwas_basic_info.pdf) and schizophrenia[31] ($n = 3,754$). From TWINGENE, we only used one individual from each pair of twins. After this filtering, a total of 10,704 controls were available for further analysis.

For the Icelandic discovery sample set, we identified from the nationwide Icelandic Cancer Registry all patients diagnosed with MM (ICD-10 code C90) in Iceland from 1955 to 2013 were identified and used in the association studies ($n = 480$). To identify MGUS cases, information on the detection of an M protein on serum protein electrophoresis was gathered from 1955 to 2005 from Landspitali University Hospital and the Icelandic Medical Center Laboratory in Mjodd. A total of 251 cases of non-IgM MGUS were identified and used in the analysis.

For replication, we obtained 223 MM cases from the Swedish National Myeloma Biobank and 363 MM cases from the University Hospital of Copenhagen. As controls for the respective replication sets, we used 1,285 randomly ascertained Swedish blood donors and 826 randomly ascertained individuals from Denmark and Skåne county, Sweden (the southernmost part of Sweden next to Denmark). All samples were collected subject to ethical approval (Lund University Ethical Review Board, 2013/54; Icelandic Data Protection Authority, 2001010157; and National Bioethics Committee 01/015) and informed consent. No individuals were approached solely for the purpose of this study.

Analysis of Swedish and Norwegian samples. Samples were genotyped on Illumina OmniExpress-Exome and OmniExpress microarrays. For analysis, we used the OmniExpress SNPs, which are recorded by both array types. We excluded SNPs showing $>5\%$ missing data, significant deviation from Hardy–Weinberg equilibrium ($P<1 \times 10^{-1}$ in controls; $P<1 \times 10^{-1\times}$ in cases), or discrepancies in allele frequency between genotyping batches ($P<5 \times 10^{-8}$; χ^2-test). We excluded samples showing $>5\%$ missing data or excess heterozygosity (>3 s.d.'s), and samples from closely related individuals (proportion identity-by-descent $\hat{\pi}>0.2$; calculated using SNPs with pairwise $r^2<0.2$ using PLINK, after removing regions of high linkage disequilibrium[32,33]). After filtering, 542,599 SNPs, 1,714 cases and 10,391 controls remained. Unobserved genotypes were imputed using phased haplotypes from the Phase I (b37) release of the 1,000 Genomes Project[34] (http://www.1000genomes.org). Association testing was performed using logistic regression under an additive genetic model. To avoid artefacts of cryptic population stratification, we included five principal components of the identity-by-state matrix that were found to increase the genomic inflation factor λ (ref. 35) in the regression. The analyses were done with SHAPEIT2 (ref. 36; (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html), IMPUTE2 (ref. 37; https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) and SNPTEST[38] (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html).

Analysis of Icelandic samples. Samples were genotyped using Illumina microarrays. The whole genomes of 2,636 Icelanders were sequenced using Illumina technology to a mean depth of at least 10× (median 20×), including 909 to a mean depth of at least 30× (ref. 39). A total of 35.5 million autosomal SNPs and indels were identified using the Genome Analysis Toolkit version 2.3.9 (ref. 40). We used information about haplotype sharing to improve variant genotyping, taking advantage of the fact that all sequenced individuals had also been chip-typed and long-range phased[17]. Variants were annotated using Ensembl release 72 and Variant Effect Predictor (VEP) version 2.8 (ref. 41). The 35.5 million sequence variants found and genotyped by whole-genome sequencing were then imputed into 104,220 Icelanders who had been genotyped using Illumina chips. In addition, using the Icelandic genealogy, we calculated genotype probabilities for 294,212 untyped individuals who are close relatives of the chip-typed individuals born after 1880 (Gudbjartsson *et al.* [39]). Including this increases the power to detect associations with all diseases where ungenotyped cases are available. Logistic regression was used to test for association between SNPs and disease, treating disease status as the response and genotype counts as covariates. Other available individual characteristics that correlate with disease status were also included in the model as nuisance variables. These characteristics were as follows: sex, county of birth, current age or age at death (first and second order terms included), blood sample availability for the individual and an indicator function for the overlap of the lifetime of the individual with the time span of phenotype collection (described in detail below). The control set selected for each case group can thus be different after matching for the nuisance variables (Gudbjartsson *et al.* [39]). Correction for familial relatedness was carried out using the method of genomic control by dividing the corresponding χ^2-statistic by 1.04 and 1.02 for MM and MM + MGUS, respectively.

Meta-analysis. We performed association testing in each discovery set separately and combined the results for 12.1 million variants that were shared by the Icelandic and 1,000 Genomes whole-genome sequencing data. These variants passed the quality thresholds applied: minor allele frequency $>0.1\%$, imputation information value >0.8 and consistent frequency between the two sample sets. The meta-analysis was done using METAL[42] (http://www.sph.umich.edu/csg/abecasis/metal) with a fixed effect model. Conditional association analysis with respect to rs56219066 and rs138740 using SNPTEST[38] did not reveal any other underlying, independent signals.

Gene expression analysis in haematopoietic cell types. To characterize gene expression patterns of *ELL2*, *TOM1* and *HMGXB4*, we used microarray data from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). These data included gene expression profiles of different types of blood cells from normal haematopoiesis ($n = 211$; accession no. GSE24759 (ref. 43)), plasma cells from patients with MM ($n = 1,285$; accession nos. GSE15695 (ref. 44) GSE4581, GSE19784 (ref. 45) and GSE26760 (ref. 46)), and plasma cells from patients with MGUS, patients with smouldering MM and healthy bone marrow donors ($n = 78$; accession no. GSE5900 (ref. 47)). All data were generated on Affymetrix U133A and Av2 microarrays and quantile-normalized to a log-normal distribution.

Genotyping and association analysis of replication samples. The Swedish and Danish replication samples were genotyped by quantitative PCR for *ELL2* rs56219066 (Taqman custom assay AHCTDL6), *ELL2* rs3815768 (Taqman assay C_22272652_30) and *TOM1* rs138726 ($D'/r^2 = 1/0.997$ with rs138740; Fluidigm SNP type assay GTA0072445). Association analysis for the replication sets was done using NEMO[48] assuming a multiplicative risk model. Results for the discovery and replication cohorts were combined using a Mantel–Haenszel fixed effect model. Heterogeneity in the effect estimate was tested assuming that the estimated ORs for different groups followed a log-normal distribution using a likelihood ratio χ^2-test with degrees of freedom equal to number of groups compared minus one.

Association of the risk alleles with gene expression. To test for associations between identified risk variants and the expression of nearby genes, we analysed SNP and gene expression microarray data generated from peripheral blood samples (eight data sets totalling 973 individuals from the Icelandic population[49] and 8,086 individuals of other European populations[50]) and lymphoblastoid cell lines (two data sets totalling 1,188 samples[51,52]). Gene expression in the Icelandic data set was quantified as the mean $\log_{10}$ expression ratio compared with pooled reference RNA samples, and regressed against the number of risk alleles carried, age, gender, relatedness and differential white blood cell counts.

Association of *ELL2* allele with Ig levels. To screen for associations between the identified MM risk allele at *ELL2* and Ig levels, we used IgA, IgG and IgM measurements from 24,279, 21,981 and 20,413 individuals from the Icelandic population, respectively. Subjects diagnosed with MM or MGUS were not included in this data set. Ig levels adjusted for age, sex and site were tested for association with imputed genotypes using generalized log-linear regression[16]. Individuals diagnosed with MM or MGUS were excluded. In addition, we used IgA, IgG and IgM data from 1,012 Swedish blood donors (Clinical Immunology and Transfusion Medicine, Lund, Sweden) previously genotyped for *ELL2* rs17085249 ($D'/r^2 = 1/0.957$ with rs56219066 and $D'/r^2 = 1/1$ with rs3815768, Fluidigm SNP type assay GTA0072447). For the latter samples, we used Pearson correlation for association testing. All Ig measurements were done at certified clinical laboratories in Iceland and Sweden.

Association of the *ELL2* risk allele with other traits. The deCODE Genetics phenotype database contains medical information on diseases and traits obtained through collaboration with specialists in each field. This includes information on cardiovascular diseases (myocardial infarction, coronary arterial disease, peripheral arterial disease, atrial fibrillation, sick sinus syndrome and stroke), metabolic disorders (obesity, diabetes and metabolic syndrome), psychiatric disorders (schizophrenia, bipolar disorder, anxiety and depression), addictions (nicotine and alcohol), inflammatory diseases (rheumatoid arthritis, lupus and asthma), musculoskeletal disorders (osteoarthritis, osteoporosis), eye diseases (glaucoma), kidney diseases (kidney stones and kidney failure) and 29 types of cancer. Anthropometric measures have also been collected through several of these projects. Routinely measured traits from patient workups (sodium, potassium, bicarbonate, calcium, phosphate, creatinine, blood cell counts, haemoglobin, haematocrit, Igs, iron, vitamins, lipids and more) were obtained from the Landspitali University Hospital, Reykjavik, and the Icelandic Medical Center Laboratory in Mjodd (Laeknasetrid), Reykjavik, in addition to more specific hormonal measures (adrenal, thyroid and sex hormones). The number of independent and uncorrelated secondary traits tested for association with rs56219066 amounts to 400.

References

1. Rajkumar, S. V. *et al.* International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol.* **15,** e538–e548 (2014).
2. Laubach, J., Richardson, P. & Anderson, K. Multiple myeloma. *Annu. Rev. Med.* **62,** 249–264 (2011).
3. Weiss, B. M., Abadie, J., Verma, P., Howard, R. S. & Kuehl, W. M. A monoclonal gammopathy precedes multiple myeloma in most patients. *Blood* **113,** 5418–5422 (2009).
4. Kristinsson, S. Y. *et al.* Patterns of survival and causes of death following a diagnosis of monoclonal gammopathy of undetermined significance: a population-based study. *Haematologica* **94,** 1714–1720 (2009).
5. Kyle, R. A. *et al.* Prevalence of monoclonal gammopathy of undetermined significance. *N. Engl. J. Med.* **354,** 1362–1369 (2006).
6. Lynch, H. T., Sanger, W. G., Pirruccello, S., Quinn-Laquer, B. & Weisenburger, D. D. Familial multiple myeloma: a family study and review of the literature. *J. Natl Cancer Inst.* **93,** 1479–1483 (2001).
7. Lynch, H. T. *et al.* Familial myeloma. *N. Engl. J. Med.* **359,** 152–157 (2008).
8. Kristinsson, S. Y. *et al.* Patterns of hematologic malignancies and solid tumors among 37,838 first-degree relatives of 13,896 patients with multiple myeloma in Sweden. *Int. J. Cancer* **125,** 2147–2150 (2009).
9. Landgren, O. *et al.* Risk of plasma cell and lymphoproliferative disorders among 14 621 first-degree relatives of 4458 patients with monoclonal gammopathy of undetermined significance in Sweden. *Blood* **114,** 791–795 (2009).
10. Altieri, A., Chen, B., Bermejo, J. L., Castro, F. & Hemminki, K. Familial risks and temporal incidence trends of multiple myeloma. *Eur. J. Cancer* **42,** 1661–1670 (2006).
11. Morgan, G. J. *et al.* Inherited genetic susceptibility to multiple myeloma. *Leukemia* **28,** 518–524 (2014).
12. Vachon, C. M. *et al.* Increased risk of monoclonal gammopathy in first-degree relatives of patients with multiple myeloma or monoclonal gammopathy of undetermined significance. *Blood* **114,** 785–790 (2009).
13. Weinhold, N. *et al.* The CCND1 c.870G > A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nat. Genet.* **45,** 522–525 (2013).
14. Broderick, P. *et al.* Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nat. Genet.* **44,** 58–61 (2012).
15. Chubb, D. *et al.* Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nat. Genet.* **45,** 1221–1225 (2013).
16. Styrkarsdottir, U. *et al.* Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. *Nature* **497,** 517–520 (2013).
17. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40,** 1068–1075 (2008).
18. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462,** 868–874 (2009).
19. Landgren, O. *et al.* Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* **113,** 5412–5417 (2009).
20. Liu, M., Hsu, J., Chan, C., Li, Z. & Zhou, Q. The ubiquitin ligase Siah1 controls ELL2 stability and formation of super elongation complexes to modulate gene transcription. *Mol. Cell* **46,** 325–334 (2012).
21. Luo, Z., Lin, C. & Shilatifard, A. The super elongation complex (SEC) family in transcriptional control. *Nat. Rev. Mol. Cell Biol.* **13,** 543–547 (2012).
22. Benson, M. J. *et al.* Heterogeneous nuclear ribonucleoprotein L-like (hnRNPLL) and elongation factor, RNA polymerase II, 2 (ELL2) are regulators ofmRNA processing in plasma cells. *Proc. Natl Acad. Sci. USA* **109,** 16252–16257 (2012).
23. Santos, P., Arumemi, F., Park, K. S., Borghesi, L. & Milcarek, C. Transcriptional and epigenetic regulation of B cell development. *Immunol. Res.* **50,** 105–112 (2011).
24. Martincic, K., Alkan, S. A., Cheatle, A., Borghesi, L. & Milcarek, C. Transcription elongation factor ELL2 directs immunoglobulin secretion in plasma cells by stimulating altered RNA processing. *Nat. Immunol.* **10,** 1102–1109 (2009).
25. Shell, S. A., Martincic, K., Tran, J. & Milcarek, C. Increased phosphorylation of the carboxyl-terminal domain of RNA polymerase II and loading of polyadenylation and cotranscriptional factors contribute to regulation of the Ig heavy chain mRNA in plasma cells. *J. Immunol.* **179,** 7663–7673 (2007).
26. Park, K. S. *et al.* transcription elongation factor ELL2 drives Ig secretory-specific mRNA production and the unfolded protein response. *J. Immunol.* **193,** 4663–4674 (2014).
27. Benson, M. J. *et al.* Heterogeneous nuclear ribonucleoprotein L-like (hnRNPLL) and elongation factor, RNA polymerase II, 2 (ELL2) are regulators of mRNA processing in plasma cells. *Proc. Natl Acad. Sci. USA* **109,** 16252–16257 (2012).
28. Fowler, T. *et al.* Regulation of MYC expression and differential JQ1 sensitivity in cancer cells. *PloS ONE* **9,** e87003 (2014).
29. Shilatifard, A. *et al.* ELL2, a new member of an ELL family of RNA polymerase II elongation factors. *Proc. Natl Acad. Sci. USA* **94,** 3639–3643 (1997).
30. Franceschini, N. *et al.* Discovery and fine mapping of serum protein loci through transethnic meta-analysis. *Am. J. Hum. Genet.* **91,** 744–753 (2012).
31. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45,** 1150–1159 (2013).
32. Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. & Zondervan, K. T. Data quality control in genetic case-control association studies. *Nat. Protoc.* **5,** 1564–1573 (2010).
33. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
34. Consortium TGP. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
35. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
36. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10,** 5–6 (2013).
37. Howie, B. N., Donelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
38. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
39. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–444 (2015).
40. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
41. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. & Cunningham, F. Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics* **26,** 2069–2070 (2010).
42. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).
43. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144,** 296–309 (2011).
44. Boyd, K. D. *et al.* Mapping of chromosome 1p deletions in myeloma identifies FAM46C at 1p12 and CDKN2C at 1p32.3 as being genes in regions associated with adverse survival. *Clin. Cancer Res.* **17,** 7776–7784 (2011).
45. Broyl, A. *et al.* Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. *Blood* **116,** 2543–2553 (2010).
46. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471,** 467–472 (2011).
47. Zhan, F. *et al.* Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis. *Blood* **109,** 1692–1700 (2007).
48. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.* **35,** 131–138 (2003).
49. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
50. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
51. Consortium TIH. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
52. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

Acknowledgements

This work was supported by research grants from the Swedish Foundation for Strategic Research (KF10-0009), the Marianne and Marcus Wallenberg Foundation (2010.0112), the Knut and Alice Wallenberg Foundation (2012.0193), the Swedish Research Council (2012-1753), the Royal Swedish Academy of Science, ALF grants to the University and

Regional Laboratories (Labmedicin Skåne), the Siv-Inger and Per-Erik Andersson Foundation, the Medical Faculty at Lund University and the Swedish Society of Medicine. We thank Jörgen Adolfsson, Tomas Axelsson, Anna Collin, Ildikó Frigyesi, Patrik Magnusson, Bertil Johansson, Jan Westin and Helga Ögmundsdóttir for their assistance. We are indebted to the clinicians who contributed samples to Swedish, Icelandic, Norwegian and Danish biobanks. We are indebted to the patients and other individuals who participated in the project.

Author contributions

B.N., M.H. and U.T. designed the project. E.I., I.T., K.S., S.N. and U.G. contributed to project design. Subject recruitment and biological material collection were organized and carried out by B.N., E.J., I.T., M.H., R.A., S.L. and U.-H.M. in Sweden; I.O., O.S., S.Y.K and V.H. in Iceland; A.W. in Norway; and A.V. and U.V. in Denmark. E.I., A.V., A.W., C.H., H.N., K.S., M.H., S.L., U.-H.M., U.T. and U.V. contributed additional samples and/or genotyping data. B-M.H., B.N., B.S., D.F.G., E.J., G.T., G.M., M.A., M.J. and P.S. and carried out sequence data analysis, imputation and association analysis. B.-M.H., B.N., B.S., G.T., K.S., M.A., M.J., T.R. U.G., and U.T. Analysed data and interpreted results. E.I., R.M., S.N. contributed to data analysis. B.S, E.J., and R.A. performed experiments. B.N., E.I., I.T., M.H., S.L., S.N., U.G., and U.-H.M. obtained critical financial support. B.S., B.N., G.T., K.S., M.A., M.H., M.J. and U.T. drafted the manuscript. All authors contributed to data interpretation and the final manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: G.T., P.S., G.M., D.F.G., T.R., K.S. and U.T. are employed by deCODE Genetics/Amgen Inc. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Swaminathan, B. *et al.* Variants in *ELL2* influencing immunoglobulin levels associate with multiple myeloma. *Nat. Commun.* 6:7213 doi: 10.1038/ncomms8213 (2015).

Paper II

nature
COMMUNICATIONS

ARTICLE

Received 11 Dec 2015 | Accepted 24 May 2016 | Published 1 Jul 2016

DOI: 10.1038/ncomms12050 OPEN

Genome-wide association study identifies multiple susceptibility loci for multiple myeloma

Jonathan S. Mitchell[1,*], Ni Li[1,*], Niels Weinhold[2,3,*], Asta Försti[4,5,*], Mina Ali[6,*], Mark van Duin[7,*], Gudmar Thorleifsson[8], David C. Johnson[9], Bowang Chen[4], Britt-Marie Halvarsson[6], Daniel F. Gudbjartsson[8,10], Rowan Kuiper[7], Owen W. Stephens[2], Uta Bertsch[3,11], Peter Broderick[1], Chiara Campo[4], Hermann Einsele[12], Walter A. Gregory[13], Urban Gullberg[6], Marc Henrion[1], Jens Hillengass[3], Per Hoffmann[14,15], Graham H. Jackson[16], Ellinor Johnsson[6], Magnus Jöud[6,17], Sigurður Y. Kristinsson[18], Stig Lenhoff[19], Oleg Lenive[1], Ulf-Henrik Mellqvist[20], Gabriele Migliorini[1], Hareth Nahi[21], Sven Nelander[22], Jolanta Nickel[3], Markus M. Nöthen[14,23], Thorunn Rafnar[8], Fiona M. Ross[24], Miguel Inacio da Silva Filho[4], Bhairavi Swaminathan[6], Hauke Thomsen[4], Ingemar Turesson[19], Annette Vangsted[25], Ulla Vogel[26], Anders Waage[27], Brian A. Walker[2], Anna-Karin Wihlborg[6], Annemiek Broyl[7], Faith E. Davies[2], Unnur Thorsteinsdottir[8,28], Christian Langer[29], Markus Hansson[6,19], Martin Kaiser[9], Pieter Sonneveld[7], Kari Stefansson[8,**], Gareth J. Morgan[2,**], Hartmut Goldschmidt[3,11,**], Kari Hemminki[4,5,**], Björn Nilsson[6,17,30,**] & Richard S. Houlston[1,**]

Multiple myeloma (MM) is a plasma cell malignancy with a significant heritable basis. Genome-wide association studies have transformed our understanding of MM predisposition, but individual studies have had limited power to discover risk loci. Here we perform a meta-analysis of these GWAS, add a new GWAS and perform replication analyses resulting in 9,866 cases and 239,188 controls. We confirm all nine known risk loci and discover eight new loci at 6p22.3 (rs34229995, $P=1.31\times10^{-8}$), 6q21 (rs9372120, $P=9.09\times10^{-15}$), 7q36.1 (rs7781265, $P=9.71\times10^{-9}$), 8q24.21 (rs1948915, $P=4.20\times10^{-11}$), 9p21.3 (rs2811710, $P=1.72\times10^{-13}$), 10p12.1 (rs2790457, $P=1.77\times10^{-8}$), 16q23.1 (rs7193541, $P=5.00\times10^{-12}$) and 20q13.13 (rs6066835, $P=1.36\times10^{-13}$), which localize in or near to *JARID2*, *ATG5*, *SMARCD3*, *CCAT1*, *CDKN2A*, *WAC*, *RFWD3* and *PREX1*. These findings provide additional support for a polygenic model of MM and insight into the biological basis of tumour development.

[1] Division of Genetics and Epidemiology, The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK. [2] Myeloma Institute for Research and Therapy, University of Arkansas for Medical Sciences, Little Rock, Arkansas 72205, USA. [3] Department of Internal Medicine V, University of Heidelberg, 69117 Heidelberg, Germany. [4] German Cancer Research Center, 69120 Heidelberg, Germany. [5] Center for Primary Health Care Research, Lund University, SE-205 02 Malmo, Sweden. [6] Hematology and Transfusion Medicine, Department of Laboratory Medicine, BMC B13, SE-221 84 Lund, Sweden. [7] Department of Hematology, Erasmus MC Cancer Institute, 3075 EA Rotterdam, The Netherlands. [8] deCODE Genetics, Sturlugata 8, IS-101 Reykjavik, Iceland. [9] Division of Molecular Pathology, The Institute of Cancer Research, Surrey SM2 5NG, UK. [10] School of Engineering and Natural Sciences, University of Iceland, IS-101 Reykjavik, Iceland. [11] National Centre of Tumor Diseases, 69120 Heidelberg, Germany. [12] University Clinic of Würzburg, 97080 Würzburg, Germany. [13] Clinical Trials Research Unit, University of Leeds, Leeds LS2 9PH, UK. [14] Institute of Human Genetics, University of Bonn, D-53127 Bonn, Germany. [15] Division of Medical Genetics, Department of Biomedicine, University of Basel, 4003 Basel, Switzerland. [16] Royal Victoria Infirmary, Newcastle upon Tyne, NE1 4LP, UK. [17] Clinical Immunology and Transfusion Medicine, Laboratory Medicine, Office of Medical Services, SE-221 85 Lund, Sweden. [18] Department of Hematology, Landspitali, National University Hospital of Iceland, IS-101 Reykjavik, Iceland. [19] Hematology Clinic, Skåne University Hospital, SE-221 85 Lund, Sweden. [20] Section of Hematology, Sahlgrenska University Hospital, Gothenburg 413 45, Sweden. [21] Center for Hematology and Regenerative Medicine, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [22] Rudbeck Laboratory, Department of Immunology, Pathology and Genetics, Uppsala University, SE-751 05 Uppsala, Sweden. [23] Department of Genomics, Life & Brain Center, University of Bonn, D-53127 Bonn, Germany. [24] Wessex Regional Genetics Laboratory, University of Southampton, Salisbury SP2 8BJ, UK. [25] Department of Haematology, University Hospital of Copenhagen at Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. [26] National Research Centre for the Working Environment, DK-2100 Copenhagen, Denmark. [27] Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Box 8905, N-7491 Trondheim, Norway. [28] Faculty of Medicine, University of Iceland, IS-101 Reykjavik, Iceland. [29] Department of Internal Medicine III, University of Ulm, D-89081 Ulm, Germany. [30] Broad Institute, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. * These authors contributed equally to this work. ** These authors jointly supervised the work. Correspondence and requests for materials should be addressed to K.H. (email: K.Hemminki@dkfz-heidelberg.de) or to B.N. (email: bjorn.nilsson@med.lu.se) or to R.S.H. (email: richard.houlston@icr.ac.uk).

Multiple myeloma (MM) is a malignancy of plasma cells that has a significant genetic component as evidenced by the two- to fourfold increased risk shown in relatives of MM patients[1]. Our understanding of MM susceptibility has been transformed by recent genome-wide association studies (GWASs), which have identified the first risk alleles for MM[2–5] and its precursor condition monoclonal gammopathy of unknown significance[5]. Although projections indicate that additional risk variants for MM can be discovered by GWAS[6], the statistical power of these individual studies is limited.

To gain comprehensive insight into MM predisposition, we performed a meta-analysis of these GWAS, new GWAS and replication comprising 9,866 cases and 239,188 controls. We confirmed all nine known risk loci and discovered eight new risk loci for MM. Our findings provide further insights into the genetic and biological basis of MM predisposition.

Results

Association analysis. To identify new MM susceptibility loci, we analysed genome-wide association data from six populations of European ancestry (Supplementary Tables 1 and 2): a new sample set from the Netherlands, two previously reported sample sets from United Kingdom and Germany, to which we added additional cases[2], and three previously published sample sets from Sweden/Norway, Iceland and the Unites States[5,7]. After filtering, the six studies provided single-nucleotide polymorphism (SNP) microarray genotypes on 7,319 cases and 234,385 controls (Supplementary Tables 1 and 2). To increase genomic resolution, we imputed >10 million SNPs using either the 1,000 Genomes Project[8] combined with UK10K[9] (MM data sets from the Netherlands, United Kingdom, Germany, Sweden/Norway and the United States) or deCODE Genetics (MM data set from Iceland[10]) as reference. Quantile–quantile plots for SNPs with minor allele frequency (MAF) > 0.5% post imputation did not show evidence of substantive overdispersion ($\lambda = 1.00$–1.06; Supplementary Fig. 1). Pooling association testing results from the six sample sets, we derived joint odds ratios and 95% confidence intervals under a fixed-effects model for each SNP and associated per allele P-value. In this analysis, associations for all nine established risk loci showed a consistent direction of effect with previously reported studies and have $P < 5.0 \times 10^{-8}$ (Fig. 1 and Supplementary Table 3).

We identified 315 SNPs at 16 loci that showed evidence of association ($P < 1.0 \times 10^{-6}$) not previously implicated in the risk of developing MM (Fig. 1 and Supplementary Tables 4 and 5). For 13 of the 16 loci, the strongest signal was provided by an imputed SNP. We confirmed the fidelity of imputation for 12 of the 13 imputed SNPs in multiple series (Supplementary Tables 6 and 7; rs78311596 imputation unconfirmed). Using allele-specific PCR, we genotyped the 15 substantiated SNPs in additional UK, Germany, Sweden/Norway and Denmark sample series totalling 2,547 cases and 4,803 controls. Meta-analysing the discovery and replication samples, we identified genome-wide significant associations for MM with eight previously unreported loci (Table 1 and Supplementary Tables 8 and 9) at 6p22.3 (rs34229995, $P = 1.31 \times 10^{-8}$), 6q21 (rs9372120, $P = 9.09 \times 10^{-15}$), 7q36.1 (rs7781265, $P = 9.71 \times 10^{-9}$), 8q24.21 (rs1948915, $P = 4.20 \times 10^{-11}$), 9p21.3 (rs2811710, $P = 1.72 \times 10^{-13}$), 10p12.1 (rs2790457, $P = 1.77 \times 10^{-8}$), 16q23.1 (rs7193541, $P = 5.00 \times 10^{-12}$) and 20q13.13 (rs6066835, $P = 1.36 \times 10^{-13}$). We also observed two promising associations (that is, $P < 5.0 \times 10^{-7}$) at 6q27 (rs1034447) and at 7q22.3 (rs17507636) (Supplementary Tables 8 and 9). Conditional analysis of GWAS data showed no evidence for additional independent signals at the loci.

The 6q21 association marked by rs9372120 (Fig. 2) maps to intron 6 of *ATG5* (*Homo sapiens* autophagy related 5). The 8q24.21 variant rs1948915 maps to *CCAT1* (colon cancer-associated transcript 1; Fig. 2). The same region at 8q24.21 harbours multiple independent loci with different tumour specificities[11], including the B-cell malignancies diffuse B-cell lymphoma[12], Hodgkin's lymphoma[13] and chronic lymphocytic leukaemia[14]. With the possible exception of chronic lymphocytic leukaemia, the linkage disequilibrium (LD) blocks defining these identified cancer risk loci are distinct from the 8q24.21 MM association signal (pairwise LD metrics $r^2 < 0.03$; Supplementary Table 10). The 9p21.3 variant rs2811710 maps to intron 1 of *CDKN2A/p16INK4A* (cyclin-dependent kinase inhibitor 2A, Fig. 2). Although the 9p21.3 region is a susceptibility locus for multiple tumour types including breast and lung cancer, glioma and acute lymphoblastic leukaemia[15], the rs2811710 association for MM is distinct (Supplementary Table 11). The 16q23.1 (rs7193541) association is a non-synonymous SNP I564V of *RFWD3* (encoding ring finger WD domain 3; Fig. 2). 6p22.3 (rs34229995) and 7q36.1 (rs7781265) associations mark chromatin-regulating genes; rs34229995 is 2.2-kb telomeric to the 5′ of *JARID2* (jumonji, AT-rich interactive domain 2; Fig. 2) and rs7781265 localizing to intron 2 of *SMARCD3* (swi/snf-related, matrix-associated, actin-dependent regulator of chromatin, subfamily d, member 3; Fig. 2). The 10p12.1 (rs2790457) association localizes to intron 3 of the gene encoding *WAC* (ww domain-containing adaptor with coiled-coil region), which has recently been shown to be part of an extended autophagy network[16]. The 20q13.13 (rs6066835) association mapped to intron 3 of *PREX1* (phosphatidylinositol-3, 4, 5-trisphosphate-dependent Rac exchange factor 1) (Fig. 2).

Relationship between the new MM SNPs and phenotype. We tested for associations between sex or age at diagnosis and genotype for each of the eight risk SNPs by case-only analysis

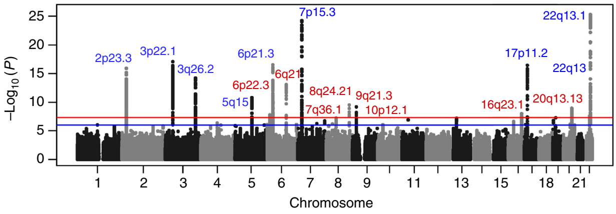

Figure 1 | Manhattan plot of association *P*-values. Shown are the genome-wide *P*-values (two sided) of 12.4 million successfully imputed autosomal SNPs in 7,319 cases and 234,385 controls from the discovery phase. Labelled in blue are previously identified risk loci and labelled in red are newly identified risk loci. The red horizontal line represents the genome-wide significance threshold of $P = 5.0 \times 10^{-8}$ and the blue horizontal line represents the threshold of $P = 1.0 \times 10^{-6}$ used to define promising SNPs.

Table 1 | Summary results for SNPs associated with multiple myeloma risk.

Location	SNP	Position (bp)	Risk allele	RAF	Data set	OR	*P*-value
6p22.3	rs34229995	15,244,018	G	0.029	Discovery	1.40	1.76×10^{-8}
					Replication	1.19	0.214
					Combined	**1.37**	$\mathbf{1.31 \times 10^{-8}}$
						$P_{\text{het}} = 0.50$	$I^2 = 0\%$
6q21	rs9372120	106,667,535	G	0.218	Discovery	1.20	8.72×10^{-14}
					Replication	1.12	0.0147
					Combined	**1.18**	$\mathbf{9.09 \times 10^{-15}}$
						$P_{\text{het}} = 0.93$	$I^2 = 0\%$
7q36.1	rs7781265	150,950,940	T	0.125	Discovery	1.20	1.82×10^{-7}
					Replication	1.15	0.0136
					Combined	**1.19**	$\mathbf{9.71 \times 10^{-9}}$
						$P_{\text{het}} = 0.24$	$I^2 = 23\%$
8q24.21	rs1948915	128,222,421	C	0.345	Discovery	1.14	3.14×10^{-10}
					Replication	1.09	0.0283
					Combined	**1.13**	$\mathbf{4.20 \times 10^{-11}}$
						$P_{\text{het}} = 0.34$	$I^2 = 11\%$
9p21.3	rs2811710	21,991,923	G	0.657	Discovery	1.14	6.50×10^{-10}
					Replication	1.18	4.02×10^{-5}
					Combined	**1.15**	$\mathbf{1.72 \times 10^{-13}}$
						$P_{\text{het}} = 0.97$	$I^2 = 0\%$
10p12.1	rs2790457	28,856,819	G	0.739	Discovery	1.12	8.44×10^{-7}
					Replication	1.13	6.18×10^{-3}
					Combined	**1.12**	$\mathbf{1.77 \times 10^{-8}}$
						$P_{\text{het}} = 0.94$	$I^2 = 0\%$
16q23.1	rs7193541	74,664,743	T	0.585	Discovery	1.12	1.14×10^{-8}
					Replication	1.17	4.79×10^{-4}
					Combined	**1.13**	$\mathbf{5.00 \times 10^{-12}}$
						$P_{\text{het}} = 0.15$	$I^2 = 35\%$
20q13.13	rs6066835	47,355,009	C	0.083	Discovery	1.24	1.16×10^{-9}
					Replication	1.35	1.36×10^{-5}
					Combined	**1.26**	$\mathbf{1.36 \times 10^{-13}}$
						$P_{\text{het}} = 0.072$	$I^2 = 43\%$

I^2, proportion of the total variation due to heterogeneity; OR, odds ratio; P_{het}, *P*-value for heterogeneity; RAF, risk allele frequency; SNP, single-nucleotide polymorphism.
RAF is risk allele frequency across all cases and controls in the discovery set, where the risk allele is the allele corresponding to the estimated OR. Positions are based on NCBI build 37 of the human genome.

using all individuals in five of the six sample sets and observed no such relationships (Supplementary Tables 12 and 13). In addition, case-only analysis provided no evidence for associations between risk SNPs and cytogenetic MM subtype (Supplementary Table 14) or MM-specific overall survival (Supplementary Table 15). Collectively, these data are compatible with the risk variants having generic effects on MM development rather than tumour progression.

Biological inference. To the extent that they have been deciphered, many of the GWAS loci map to non-coding regions of the genome and influence gene regulation. In this respect, it is perhaps not surprising that none of the genes annotated by the GWAS signals we identify are somatically mutated in MM (Supplementary Table 16). Hence, to gain insight into the biological mechanisms for the associations at the eight newly identified risk SNPs, we first performed expression quantitative trait loci (eQTL) analysis using gene expression profiles of CD138-positive MM plasma cells from the United Kingdom ($n = 183$), Germany ($n = 658$) and the United States ($n = 608$) cases (Affymetrix Human Genome U133 2.0 Plus Array; NCBI GEO Data sets GSE21349, GSE31161, GSE2658 and EBI ArrayExpress E-MTAB-2299). In addition, we interrogated publicly accessible expression data on whole blood, adipocytes, skin cells and lymphoblastoid cell lines (LCLs). To explore methylation QTL (meQTLs) at each risk locus, we analysed Illumina Infinium HumanMethylation450 BeadChip data on CD138-positive MM plasma cells from 365 UK patients. In MM plasma cells, we identified significant associations between rs2790457 and decreased expression of *WAC* ($P = 6.58 \times 10^{-24}$) and rs6066835, and increased expression of *PREX1* ($P = 3.85 \times 10^{-5}$) (Supplementary Fig. 2 and Supplementary Data 1). We also detected strong *cis*-meQTLs at *WAC* and *PREX1* with rs2790457 and rs6066835 genotypes (*P*-values 1.42×10^{-6} and 1.12×10^{-4}, respectively; Supplementary Data 1). The direction of these eQTLs and meQTLs is compatible with the 10p12.1 signal encompassing an active promotor for *WAC*, whereas the 20q13.13 signal does not capture an active promotor in the gene body of *PREX1* (Fig. 2).

DNA methylation plays a central role in epigenetic regulation of gene expression; however, meQTLs and *cis*-acting eQTLs do

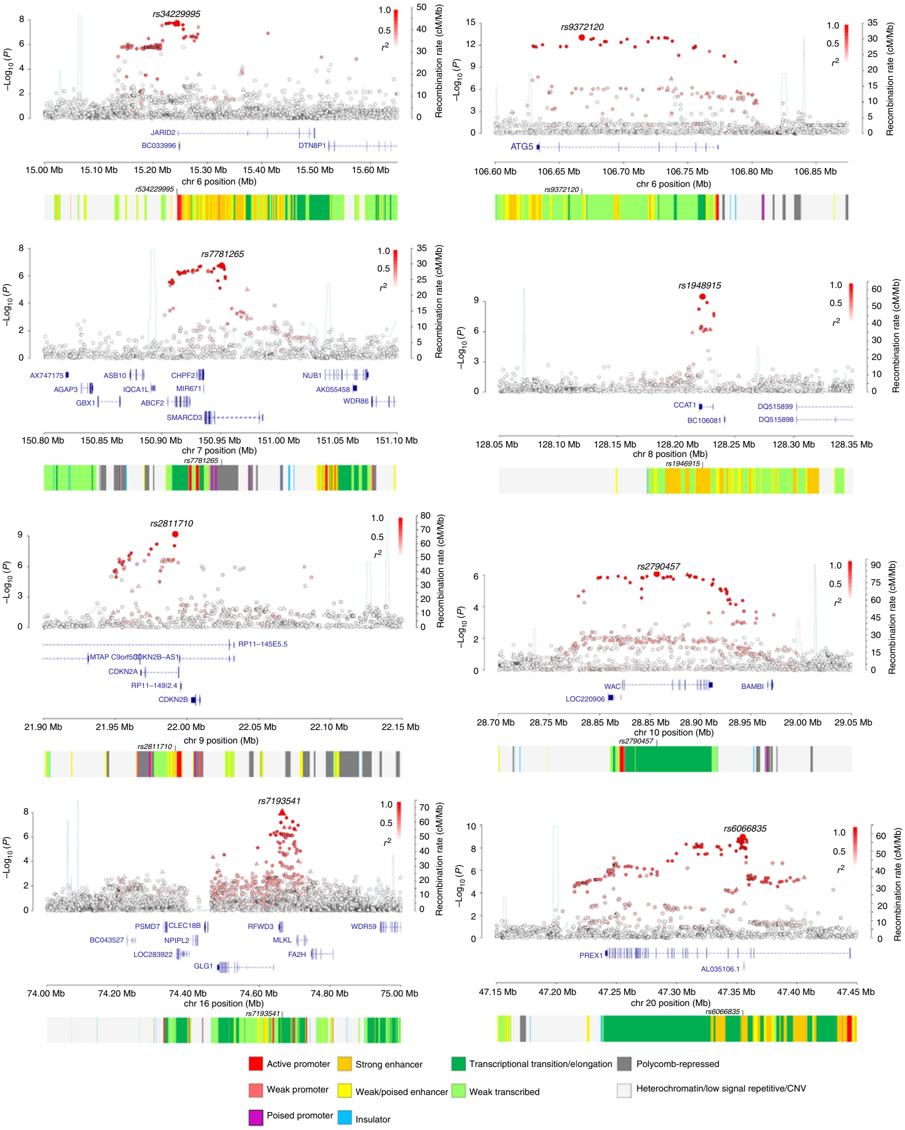

Figure 2 | Regional plots of association results and recombination rates for the newly identified risk loci for multiple myeloma. Results for 6p22.3 (rs34229995), 6q21 (rs9372120), 7q36.1 (rs7781265), 8q24.21 (rs1948915), 9p21.3 (rs2811710), 10p12.1 (rs2790457), 16q23.1 (rs7193541) and 20q13.13 (rs6066835). Plots (using visPig[70]) show association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates. $-\text{Log}_{10}$ *P*-values (*y* axes) of the SNPs are shown according to their chromosomal positions (*x* axes). The sentinel SNP in each combined analysis is shown as a large circle or triangle and is labelled by its rsID. The colour intensity of each symbol reflects the extent of LD with the top SNP, white ($r^2=0$) through to dark red ($r^2=1.0$). Genetic recombination rates, estimated using 1,000 Genomes Project samples, are shown with a light blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to the region of association. Genes have been redrawn to show their relative positions; therefore, maps are not to physical scale. On the bottom is the chromatin-state segmentation track (ChromHMM) for lymphoblastoid cells using data from the HapMap ENCODE Project.

not always overlap. Thus, although rs7193541 showed a strong meQTL for *RFWD3* methylation and reduced expression of *RFWD3* in whole blood, no eQTL was shown in MM plasma cells (Supplementary Data 1).

Various lines of evidence indicate that chromatin lopping interactions formed between enhancer elements and genes that they regulate map within distinct chromosomal topological associating domains (TADs)[17]. To map candidate causal SNPs to TADs and identify patterns of local chromatin patterns, we analysed Hi-C data on the LCL cell line GM12878 (ref. 17), as a source of B-cell information (Supplementary Fig. 3). Looping chromatin interactions and TADs were shown at 6q21 (rs9372120), 8q24.21 (rs1948915), 9p21.3 (rs2811710) and 20q13.13 (rs6066835), involving a number of genes with biological relevance to MM development. With the limitations of cell line data from LCL, which may not fully reflect MM biology, we demonstrated with MM RNA-sequencing data that gene expression within the 6q21 and 9p21.3 TADs were tightly correlated ($P<2.0\times10^{-5}$), which is consistent with their co-regulation (Supplementary Table 17). Moreover, the region at 6q21 (rs9372120, *ATG5*) participates in intra-chromosome looping with the transcriptional repressor *PRDM1* (Supplementary Fig. 3b). Similarly, the 8q24.21 region of association defined by rs1948915, which contains *CCAT1* (colon cancer-associated transcript 1), interacts with *MYC* and distal upstream enhancer elements (Supplementary Fig. 3d).

To explore the epigenetic profile of association signals at each of the new MM risk loci, we used HaploReg and RegulomeDB to examine whether the sentinel SNPs and those in high LD (that is, $r^2>0.8$ in the 1,000 Genomes EUR reference panel) annotate putative transcription factor (TF) binding or enhancer elements. We also assessed B-cell-specific chromatin dynamics using FANTOM5, which uses the pre-computed chromatin state data for multiple cell lines. HaploReg showed that the majority of MM-related SNPs were observed in regions of DNase hypersensitivity common across multiple cell lines. The protein motifs at these sites are for known TFs such as nuclear factor-κB, c-MYC, GATA, TCF4, POL24H8, CEBPB or POL2 (Supplementary Data 2). We examined for statistical evidence of enrichment in specific TF binding across the eight new and nine established risk loci using GM12878 data[18]. Although of borderline significance and hypothesis generating, after correction for the 90 TFs assayed, there was evidence for enrichment of SPI1 (alias PU.1), ($P=0.0007$, $P_{\mathrm{adjusted}}=0.063$), which regulates *PRDM1* and its downregulation is required for MM cell growth[19]. Collectively, these observations are compatible with the identified risk SNPs mapping within regions of active chromatin state, which have a role in the B-cell *cis*-regulatory network.

Discussion

We have performed the largest GWAS of MM to date. We identified eight novel MM risk loci taking the total count to 17. Fully deciphering the functional impact of these SNP associations on MM development requires additional analyses. However, seven of the SNPs map intragenic to transcribed genes, which are relevant to MM or B-cell biology. Although a number of SNPs displayed an eQTL/meQTL in MM plasma cells, the absence of a relationship does not preclude the possibility of a subtle cumulative long-term relationship intrinsic to plasma cells or a predisposition through altered gene function in other cell types.

Studies in other cancers have shown that the multiple risk loci at 8q24.21 are enhancers interacting with *MYC*[20,21]. As deregulation of *MYC* is a feature of MM, it is plausible that the susceptibility to MM has a similar mechanistic basis. Indeed, *MYC* promotes *CCAT1* transcription by binding to its promoter, and in colorectal cancer the L-isoform of *CCAT1* has been shown to interact with the *MYC* promoter and distal upstream enhancer elements regulating *MYC* transcription[22]. We have previously shown the MM risk SNP at 7p15.3 influences expression of *CDCA7L*, a binding partner of p75 potentiating *MYC*-mediated transformation. In addition to local interactions with *CDKN2A/CDKN2B*, the 9p21.3 region encompassing SNP rs2811710 interacts with the genomic region containing *MTAP* (methylthioadenosine phosphorylase). *MTAP* plays a major role in polyamine metabolism and deletion of *MTAP* is common in cancer, being closely linked to homozygous deletion of *p16* (ref. 23).

ATG5 at 6q21 is highly expressed in plasma cells and essential for autophagy and plasma cell survival[24]. Strikingly, the same locus also contains the transcriptional repressor *PRDM1* (formerly *BLIMP1*), which is key to the development of plasma cells from B cells and a determinant of plasma cell survival[25]. The RFWD3 protein is an E3 ubiquitin ligase that positively regulates p53 stability by forming an RFWD3–MDM2–p53 complex, thereby protecting p53 from degradation by MDM2-mediated polyubiquitination[26]. Variation at 16q23.1 defined with the correlated SNP rs4888262 (pairwise LD with rs7193541, $r^2=0.68$, $D'=1.0$) has previously been shown to influence testicular cancer risk[27], suggesting a common genetic and biological basis to both associations.

JARID2 functions as a transcriptional repressor through recruitment of Polycomb repressive complex 2 and has recently been identified as a regulator of haematopoietic stem cell function[28], and the 6p22.3-p21.31 region is commonly gained in MM tumours[29]. Inhibition of *JARID2* leads to loss of Polycomb binding and a reduction of histone H3 lysine-27 trimethylation levels on target genes. *SMARCD3* recruits BAF chromatin remodelling complexes to specific enhancers. Although there is currently no evidence to implicate the transcriptional repressors *JARID2* or *SMARCD3* in terms of somatic mutation in MM, multiple genes including *CDKN2A* and *TP53* are silenced by methylation in MM. Overexpression of histone methyltransferase and inactivating mutations in histone demethylase (*UTX*) typifies a subset of MM[30] and our findings add to the impact of chromatin remodelling genes on MM.

We have previously shown an association for MM at *ULK4*, a key regulator of mammalian target of rapamycin-mediated autophagy[4]. We now suggest a more extensive set of associations involving *ATG5* and *WAC*, and by virtue of the role of MYC in autophagy[31], *CCAT1*, *CDCA7L*, *DNMT3A* and *CBX7*. Collectively, these data invoke deregulation of DNA methylation, telomere length, differentiation and autophagy, and immunoglobulin production as determinants of MM susceptibility.

Our findings provide further evidence for an inherited genetic susceptibility to MM. However, further studies are necessary to understand the biology behind these risk variants. We estimate that the currently identified risk SNPs for MM account for 20% of the heritable risk attributable to all common variation; hence, further GWAS-based studies in concert with functional analyses should lead to additional insights into MM biology. Importantly, such studies may inform the development of new therapeutic agents[32,33].

Methods

Ethics. Collection of patient samples and associated clinico-pathological information was undertaken with written informed consent and relevant ethical review board approval at respective study centres in accordance with the tenets of the Declaration of Helsinki, specifically for the Myeloma-IX trial by the Medical Research Council (MRC) Leukaemia Data Monitoring and Ethics committee (MREC 02/8/95, ISRCTN68454111), the Myeloma-XI trial by the Oxfordshire Research Ethics Committee (MREC 17/09/09, ISRCTN49407852), HOVON65/GMMG-HD4 (ISRCTN 644552890; METC 13/01/2015), HOVON87/NMSG18

(EudraCTnr 2007-004007-34, METC 20/11/2008), HOVON95/EMN02 (EudraCTnr 2009-017903-28, METC 04/11/10), University of Heidelberg Ethical Commission (229/2003, S-337/2009, AFmu-119/2010), University of Arkansas for Medical Sciences Institutional Review Board (IRB 202077), Lund University Ethical Review Board (2013/54) and Icelandic Data Protection Authority (2,001,010,157 and National Bioethics Committee 01/015).

Genome-wide association studies. The diagnosis of MM (ICD-10 C90.0) was established in accordance with World Health Organization guidelines. All samples from patients for genotyping were obtained before treatment or at presentation. The meta-analysis was based on GWAS conducted in the Netherlands, the United Kingdom, Germany, Sweden/Norway, the United States and Iceland (Supplementary Tables 1 and 2).

The Dutch GWAS consisted of 608 cases (316 male). The cases were ascertained from three clinical trials: HOVON65/GMMG-HD4 ISRCTN64455289 (restricted to Dutch cases; $n = 158$), HOVON87/NMSG18 ($n = 292$) and HOVON95/EMN02 ($n = 105$) (ISRCTN64455289: GMMG-HD4 http://www.isrctn.com/search?q=ISRCTN64455289, HOVON87/NMSG18; HOVON87/NMSG18 https://www.clinicaltrialsregister.eu/ctr-search/trial/2007-004007-34/BE and HOVON95/EMN02 https://www.clinicaltrialsregister.eu/ctr-search/trial/2009-017903-28/AT). DNA was extracted from venous blood samples and genotyped using Illumina Human OmniExpress-12 v1.0 arrays (Illumina, San Diego, USA). For controls, we used the B-PROOF data set (B-vitamins for the prevention of osteoporotic fractures). Controls were genotyped using Illumina OmniEpress Exome-8v1-1 arrays[34].

The UK GWAS[2] comprised 2,329 cases (1,060 male (post quality control (QC)); mean age at diagnosis: 64 years) recruited through the UK MRC Myeloma-IX and Myeloma-XI trials (ISRCTN68454111: Myeloma IX 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). For controls, we used publicly accessible data generated by the Wellcome Trust Case Control Consortium from the 1958 Birth Cohort (58C; also known as the National Child Development Study) and National Blood Service. Genotyping of controls was conducted using Illumina Human 1-2M-Duo Custon_v1 Array chips (www.wtccc.org.uk).

The German GWAS[2] comprised 1,512 cases (867 male (post QC); mean age at diagnosis: 59 years) recruited by the German-Speaking Multiple Myeloma Multicenter Study Group (GMMG) 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; and ISRCTN05745813: GMMG-HD5 http://www.isrctn.com/search?q=ISRCTN05745813). DNA was prepared from EDTA-venous blood or CD138-negative bone marrow cells ($<1\%$ tumour contamination). Genotyping was performed using Illumina Human OmniExpress-12 v1.0 arrays (Illumina). For controls, we used genotype data on 2,107 healthy individuals, enroled into the Heinz Nixdorf Recall (HNR) study genotyped using either Illumina HumanOmni1-Quad_v1 or 1428 OmniExpress-12 v1.0 arrays.

The Swedish/Norwegian GWAS[5] was based on 1,668 and 157 MM patients from the Swedish National Myeloma Biobank (Skåne University Hospital, Lund, Sweden) and the Norwegian Biobank for Myeloma (Trondheim, Norway), respectively. Genotyping was performed using Illumina Human OmniExpress-Exome arrays (Illumina). Control genotypes on 10,704 individuals were obtained from previously published studies of schizophrenia and TWINGENE[5].

The USA GWAS[7] comprised 1,076 newly diagnosed patients treated at the UAMS Myeloma Institute for Research and Therapy (NCT00083551: Total therapy II https://clinicaltrials.gov/ct2/show/NCT00083551; NCT00081939: Total therapy III https://clinicaltrials.gov/ct2/show/NCT00081939; NCT00572169: Total therapy 3B https://clinicaltrials.gov/ct2/show/NCT00572169; and NCT00734877: Total therapy 4 https://clinicaltrials.gov/ct2/show/NCT00734877). DNA was isolated from peripheral blood samples collected from patients after granulocyte–colony-stimulating factor mobilization of stem cells. Genotyping was performed using Illumina Human OmniExpress-12 v1.0 arrays and OmniExpress arrays (Illumina)[7]. Genotype data from 2,234 healthy individuals enroled into the Cancer Genetic Markers of Susceptibility studies served as a source of controls.

The Icelandic GWAS comprised 480 MM cases identified from the nationwide Icelandic Cancer Registry[5]. Samples were genotyped using Illumina microarrays[5].

Analysis of GWAS. The Swedish/Norwegian GWAS has been previously published in its entirety with a full description of QC[5]. Adopting the same standard, quality-control measures were applied to the UK, German, US and the Netherlands GWAS. Specifically, we excluded individuals with low call rate ($<95\%$) and those found to have non-European ancestry on the basis of HapMap version 2 CEU, JPT/CHB and YRI population reference data (Supplementary Fig. 4). For first-degree relative pairs, we excluded the control or the individual with the lower call rate. SNPs with a call rate $<95\%$ were excluded as were those with a MAF<0.01 or displaying significant deviation from Hardy–Weinberg equilibrium (that is, $P<10^{-5}$). Post QC, the 5 GWAS provided genotype data on 6,839 cases and 22,221 controls. GWAS data were imputed for all scans for >10 million SNPs using 1,000 Genomes Project (phase 1 integrated release 3, March 2012)[8] and UK10K data (ALSPAC, EGAS00001000090/EGAD00001000195 and TwinsUK EGAS00001000108/EGAS00001000194 studies only)[9] as reference in conjunction with IMPUTE2 v2.3 software[35] (Supplementary Tables 1 and 2). Imputation was conducted separately for each scan and each GWAS was pruned to a common set of SNPs between cases and controls. We pre-set thresholds for imputation quality, to retain potential risk variants with MAF>0.005 for validation. Specifically, we excluded poorly imputed SNPs (that is, information measure Is <0.80). Test of association between imputed SNPs and MM was performed using logistic regression using SNPTESTv2.5.2 (ref. 36). The adequacy of the case–control matching was formally evaluated using quantile–quantile plots of test statistics (Supplementary Fig. 1). The inflation factor λ was based on the 90% least-significant SNPs[37]. Where appropriate, principle components (zero for UK, five for Sweden/Norway, two for Germany, zero for USA and zero for the Netherlands), generated using common SNPs, were included to limit the effects of cryptic population stratification. Eigenvectors for the GWAS data sets were inferred using smartpca (part of EIGENSOFT[38]) by merging cases and controls with Phase II HapMap samples.

For the Icelandic GWAS, SNP genotypes were phased using a long-range method based on whole genome sequence data on 2,636 Icelanders. Sequence variants (35.5 million) were then imputed into 104,220 Icelanders, which had been genotyped using Illumina chips. We corrected for familial relatedness by genomic control dividing the χ^2-statistic by 1.04.

Meta-analysis. We performed association testing in the discovery sets separately and then combined the results for 12.4 million variants. We assessed the fidelity of imputation through the concordance between imputed and directly genotyped SNPs in a subset of GWAS samples (Supplementary Tables 6 and 7). Meta-analysis was undertaken using the inverse-variance approach under a fixed-effects model implemented in META v1.6 (ref. 39). Cochran's Q-statistic was calculated, to test for heterogeneity, and the I^2 statistic measured, to quantify the proportion of the total variation due to heterogeneity[40]. Meta-analysis summary statistics and LD correlations from a reference panel of 1,000 Genomes Project combined with UK10K, we used GCTA[41] to perform conditional association analysis. Association statistics were calculated for all SNPs conditioning on the top SNP in each loci showing genome-wide significance. This is performed in a step-wise manner.

Replication genotyping. To validate promising associations, we analysed four case–control series from the United Kingdom, Germany, Denmark and Sweden/Norway.

The UK replication comprised 812 MM cases (412 male) ascertained through the UK MRC Myeloma-IX ($n = 95$) and XI trials ($n = 717$). Controls comprised 1,110 healthy individuals with self-reported European ancestry (420 male, aged 18–69 years) with no personal history of malignancy ascertained through GEnetic Lung CAncer Predisposition Study ($n = 536$) (ref. 42) and National Study of Colorectal Cancer Genetics ($n = 574$) (ref. 43). All cases and controls were UK residents.

The German replication series comprised 1,149 cases collected by the German Myeloma Study Group (Deutsche Studiengruppe Multiples Myelom (DSMM)), GMMG, University Clinic, Heidelberg, and University Clinic, Ulm (676 male, mean age at diagnosis 57.6 years, s.d. 9.8). Controls comprised of 1,582 healthy German blood donors recruited between 2004 and 2007 by the Institute of Transfusion Medicine and Immunology, University of Mannheim, Germany (885 male, mean age 55.8 years, s.d. 10.0).

The Swedish/Norway and Danish replication series comprised 223 MM cases from the Swedish National Myeloma Biobank and 363 MM cases from the University Hospital of Copenhagen. As controls for these respective replication sets, we analysed 1,285 Swedish blood donors and 826 individuals from Denmark and Skåne County, Sweden (the southernmost part of Sweden adjacent to Denmark).

Replication genotyping was performed using allele-specific PCR KASPar chemistry (LGC, Hertfordshire, UK; UK replication series). Primers, probes and conditions used are available on request. Call rates for SNP genotypes were $>95\%$ in each of the replication series. The quality of genotyping in all assays was assessed by measuring 1–10% duplicates (showing a concordance of $>99\%$) and at least two negative controls for each centre. Technical artefacts were excluded by cross-platform validation of 96 samples and sequencing of a set of 96 randomly selected samples from each case and control series confirmed genotyping accuracy. Concordance of $>99\%$ demonstrated robust performance.

Translocation detection and mutation analysis. Karotyping was used for cytogenetic studies of MM cells and standard criteria for the definition of a clone were applied. Fluorescence *in situ* hybridization and ploidy classification of UK samples was conducted using the methodologies previously described[44]. Fluorescence *in situ* hybridization and ploidy classification of German samples was performed as previously described[45]. The XL IGH Break Apart probe (MetaSystems, Altlussheim Germany) was used to detect any IGH translocation in German samples. Logistic regression in case-only analyses was used to assess tumour karyotype . The frequency of somatic mutation in genes annotated by GWAS signals was derived from tumour whole-exome sequencing of 463 Myeloma XI trial patients[46].

Association between genotype and patient outcome. To examine the relationship between SNP genotype and patient outcome, we analysed GWAS data on four of the patient cohorts[2–4,7], specifically (i) 1,165 cases from the UK MRC Myeloma-IX trial (UK-GWAS); (ii) 877 MM cases from the UK MRC Myeloma-XI trial (UK-GWAS); (iii) 511 of the patients recruited to the German GWAS; and (iv) 703 MM cases in the UAMS Myeloma Institute for Research and Therapy GWAS (USA GWAS)[7]. Clinical trial information on these patients has been previously reported[47–50]. The primary analysis end point was myeloma-specific overall survival and analysis was performed as previously described[51]. Cox regression analysis was used to derive genotype-specific hazard ratio and associated 95% confidence intervals. Meta-analysis was performed under a fixed-effects model (Supplementary Table 15).

eQTL analysis. We performed an eQTL analyses using Affymetrix Human Genome U133 2.0 Plus Array data for plasma cells from 183 MRC Myeloma IX trial patients[29], 658 Heidelberg patients and 608 US patients as recently described. Briefly, GER, UK and US data were separately pre-processed and analysed using a Bayesian approach to probabilistic estimation of expression residuals to infer broad variance components, thus accounting for hidden determinants influencing global expression such as copy number, translocation status and batch effects[52]. The association between genotype of the sentinel variant and gene expression of genes within 500 Kb either side was evaluated based on the significance of linear regression coefficients. We pooled data from each study under a fixed-effects model controlling for false discovery rate (FDR) calling significant associations with a FDR$\leq$0.05. In addition, we queried publicly available eQTL messenger RNA expression data using MuTHER and the Blood eQTL browser. MuTHER contains expression data on LCLs, skin and adipose tissue from 856 healthy twins[53]. The Blood eQTL browser contains expression data from 5,311 non-transformed peripheral blood samples[54].

meQTL analysis. We performed *cis*-meQTL analysis using Illumina 450K methylation array data on plasma cells from 384 MRC Myeloma XI trial patients. As with analysis of MM expression (eQTL) data, we inferred hidden determinants influencing global methylation. The genetic association was tested under an additive model between each SNP and each normalized methylation probe, adjusting for plate and methylation-based principal component analysis score. Controlling for a FDR of 0.05 across the 338,456 methylation traits required a *P*-value for association to be $<4.0\times10^{-5}$.

ENCODE and chromatin state dynamics. Risk SNPs and their proxies (that is, $r^2>0.8$ in the 1,000 Genomes EUR reference panel) were annotated for putative functional effect using HaploReg v3 (ref. 55), RegulomeDB[56] and SeattleSeq[57] annotation. These servers make use of data from ENCODE[58], genomic evolutionary rate profiling[59] conservation metrics, combined annotation dependent depletion scores[60] and PolyPhen scores[61]. We examined for an overlap of associated SNPs with predicted enhancers using the FANTOM5 enhancer atlas[62]and searched for overlap with 'super-enhancer' regions using data from Hnisz *et al.*[63], restricting our analysis to GM12878.

To formally examine for enrichment in specific TF binding across risk loci, we adopted the method of Gaulton *et al.*[18] Briefly, for each risk locus we derived a credible set of SNPs with a 99% probability of containing the causal SNP; posterior probability for each SNP being computed from its Bayes factor. SNPs were ranked by their posterior probability and included so that the cumulative posterior probability for association was >0.99. Binding sites for 90 TF in GM12878 were obtained from ENCODE. For each TF the total posterior probability over all credible set SNPs overlapping all binding sites was calculated. A null distribution was generated by randomly relocating each binding site up to 100 kb from its original location. For these perturbed sites, the total posterior probability over all overlapping SNPs was calculated. This process was repeated 10,000 times and enrichment *P*-values calculated as the fraction of permutations where the total posterior probability was greater than for the unperturbed binding sites.

Hi-C data and definition of topological domains at risk loci. Hi-C data was used to map the candidate causal SNPs to chromosomal TADs and identify patterns of relevant, local chromatin interactions. We made use of publicly available raw Hi-C data on GM12878 cells[17]. Valid Hi-C pairs were generated aligning raw reads to the reference genome using Burrows-Wheeler alignment (BWA), matching pairs of reads and filtering for biases. *Bona fide* Hi-C ditags were allocated to a contact matrix, with a predefined, uniform resolution of 5 kb. We corrected for experimental bias using the matrix balancing approach[64]. We inferred TADs from the contact matrix by means of the arrowhead algorithm for domain detection as previously proposed.

To investigate whether genes within TADs are co-regulated, we obtained RNAseq transcript counts from 66 MM cell lines from the Keat's lab Data Repository (http://www.keatslab.org/data-repository)[65]. We performed pairwise correlation by calculating the Pearson's product–moment correlation coefficient of the transcript counts for all pairs of genes within respective TADs.

Heritability analysis. We used Genome-wide Complex Trait Analysis to estimate the polygenic variance ascribable to all genotyped and imputed GWAS SNPs simultaneously for the UK and German GWAS[41,66,67]. SNPs were excluded based on low MAF, poor imputation and poor HWE. Principal components were included as covariates in the heritability analysis of the German data. As previously advocated when calculating the heritability of a disease such as cancer we used the lifetime risk[68,69], which for MM is estimated to be 0.007 for the UK population (http://www.cancerresearchuk.org/cancer-info/cancerstats/types/myeloma/incidence/uk-multiple-myeloma-incidence-statistics#Lifetime) and 0.006 for the German population. We estimated the heritability explained by risk SNPs identified by GWAS as located within regions associated with MM. Meta-analysis of heritability estimates from UK and German GWAS data sets was performed under a standard fixed-effects model.

Data availability. SNP genotyping data that support the findings of this study have been deposited in Gene Expression Omnibus with accession codes GSE21349, GSE19784, GSE24080, GSE2658 and GSE15695; in the European Genome-phenome Archive (EGA) with accession code EGAS00000000001; in the European Bioinformatics Institute (Part of the European Molecular Biology Laboratory) (EMBL-EBI) with accession code E-MTAB-362 and E-TABM-1138; and in the database of Genotypes and Phenotypes (dbGaP) with accession code phs000207.v1.p1.

Expression data that support the findings of this study have been deposited in GEO with accession codes GSE21349, GSE2658, GSE31161 and EMBL-EBI with accession code E-MTAB-2299.

Whole-exome sequence data that support the findings of this study have been deposited in EGA with accession code EGAS00001001147.

Transcription profiling data from MuTHer studies that support the findings of this study have been deposited in EMBL-EBI with accession code E-TABM-1140. Data from Blood eQTL have been deposited in EMBL-EBI with accession codes E-TABM-1036, E-MTAB-945 and E-MTAB-1708.

The remaining data are contained within the paper and Supplementary Files or available from the author upon request.

References

1. Morgan, G. J. *et al.* Inherited genetic susceptibility to multiple myeloma. *Leukemia* **28,** 518–524 (2014).
2. Chubb, D. *et al.* Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nat. Genet.* **45,** 1221–1225 (2013).
3. Weinhold, N. *et al.* The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nat. Genet.* **45,** 522–525 (2013).
4. Broderick, P. *et al.* Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nat. Genet.* **44,** 58–61 (2012).
5. Swaminathan, B. *et al.* Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. *Nat. Commun.* **6,** 7213 (2015).
6. Mitchell, J. S. *et al.* Implementation of genome-wide complex trait analysis to quantify the heritability in multiple myeloma. *Sci. Rep.* **5,** 12473 (2015).
7. Erickson, S. W. *et al.* Genome-wide scan identifies variant in 2q12.3 associated with risk for multiple myeloma. *Blood* **124,** 2001–2003 (2014).
8. Genomes Project, C *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
9. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6,** 8111 (2015).
10. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–444 (2015).
11. Fletcher, O. & Houlston, R. S. Architecture of inherited susceptibility to common cancer. *Nat. Rev. Cancer* **10,** 353–361 (2010).
12. Cerhan, J. R. *et al.* Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nat. Genet.* **46,** 1233–1238 (2014).
13. Enciso-Mora, V. *et al.* A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). *Nat. Genet.* **42,** 1126–1130 (2010).
14. Crowther-Swanepoel, D. *et al.* Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat. Genet.* **42,** 132–136 (2010).
15. Sherborne, A. L. *et al.* Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nat. Genet.* **42,** 492–494 (2010).
16. Joachim, J., Wirth, M., McKnight, N. C. & Tooze, S. A. Coiling up with SCOC and WAC: two new regulators of starvation-induced autophagy. *Autophagy* **8,** 1397–1400 (2012).
17. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
18. Gaulton, K. J. *et al.* Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47,** 1415–1425 (2015).

19. Tatetsu, H. *et al.* Down-regulation of PU.1 by methylation of distal regulatory elements and the promoter is required for myeloma cell growth. *Cancer Res.* **67,** 5328–5336 (2007).
20. Jia, L. *et al.* Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.* **5,** e1000597 (2009).
21. Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc, Natl Acad. Sci. USA* **107,** 9742–9746 (2010).
22. Xiang, J. F. *et al.* Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Res.* **24,** 513–531 (2014).
23. Kryukov, G. V. *et al.* MTAP deletion confers enhanced dependency on the PRMT5 arginine methyltransferase in cancer cells. *Science* **351,** 1214–1218 (2016).
24. Cenci, S. Autophagy a new determinant of plasma cell differentiation and antibody responses. *Mol. Immunol.* **62,** 289–295 (2014).
25. Pengo, N. *et al.* Plasma cells require autophagy for sustainable immunoglobulin production. *Nat. Immunol.* **14,** 298–305 (2013).
26. Fu, X. *et al.* RFWD3-Mdm2 ubiquitin ligase complex positively regulates p53 stability in response to DNA damage. *Proc. Natl Acad. Sci. USA* **107,** 4579–4584 (2010).
27. Chung, C. C. *et al.* Meta-analysis identifies four new loci associated with testicular germ cell tumor. *Nat. Genet.* **45,** 680–685 (2013).
28. Kinkel, S. A. *et al.* Jarid2 regulates hematopoietic stem cell function by acting with polycomb repressive complex 2. *Blood* **125,** 1890–1900 (2015).
29. Walker, B. A. *et al.* A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood* **116,** e56–e65 (2010).
30. Pawlyn, C., Kaiser, M. F., Davies, F. E. & Morgan, G. J. Current and potential epigenetic targets in multiple myeloma. *Epigenomics* **6,** 215–228 (2014).
31. Toh, P. P. *et al.* Myc inhibition impairs autophagosome formation. *Hum. Mol. Genet.* **22,** 5237–5248 (2013).
32. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47,** 856–860 (2015).
33. Ocio, E. M., Mateos, M. V., Maiso, P., Pandiella, A. & San-Miguel, J. F. New drugs in multiple myeloma: mechanisms of action and phase I/II clinical findings. *Lancet Oncol.* **9,** 1157–1165 (2008).
34. Zheng, H. F. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526,** 112–117 (2015).
35. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
36. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39,** 906–913 (2007).
37. Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37,** 1243–1246 (2005).
38. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
39. Liu, J. Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.* **42,** 436–440 (2010).
40. Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21,** 1539–1558 (2002).
41. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
42. Eisen, T., Matakidou, A., Houlston, R. & Consortium, G. Identification of low penetrance alleles for lung cancer: the GEnetic Lung CAncer Predisposition Study (GELCAPS). *BMC Cancer* **8,** 244 (2008).
43. Penegar, S. *et al.* National study of colorectal cancer genetics. *Br. J. Cancer* **97,** 1305–1309 (2007).
44. Chiecchio, L. *et al.* Deletion of chromosome 13 detected by conventional cytogenetics is a critical prognostic factor in myeloma. *Leukemia* **20,** 1610–1617 (2006).
45. Neben, K. *et al.* Combining information regarding chromosomal aberrations t(4;14) and del(17p13) with the International Staging System classification allows stratification of myeloma patients undergoing autologous stem cell transplantation. *Haematologica* **95,** 1150–1157 (2010).
46. Walker, B. A. *et al.* APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nat. Commun.* **6,** 6997 (2015).
47. Goldschmidt, H. *et al.* Joint HOVON-50/GMMG-HD3 randomized trial on the effect of thalidomide as part of a high-dose therapy regimen and as maintenance treatment for newly diagnosed myeloma patients. *Ann. Hematol.* **82,** 654–659 (2003).
48. Merz, M. *et al.* Subcutaneous versus intravenous bortezomib in two different induction therapies for newly diagnosed multiple myeloma: an interim analysis from the prospective GMMG-MM5 trial. *Haematologica* **100,** 964–969 (2015).
49. Morgan, G. J. *et al.* Cyclophosphamide, thalidomide, and dexamethasone as induction therapy for newly diagnosed multiple myeloma patients destined for autologous stem-cell transplantation: MRC Myeloma IX randomized trial results. *Haematologica* **97,** 442–450 (2012).
50. Morgan, G. J. *et al.* Long-term follow-up of MRC Myeloma IX trial: survival outcomes with bisphosphonate and thalidomide treatment. *Clin. Cancer Res.* **19,** 6030–6038 (2013).
51. Johnson, D. C. *et al.* Genome-wide association study identifies variation at 6q25.1 associated with survival in multiple myeloma. *Nat. Commun.* **7,** 10290 (2016).
52. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7,** 500–507 (2012).
53. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44,** 1084–1089 (2012).
54. Westra, H. J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
55. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40,** D930–D934 (2012).
56. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–1797 (2012).
57. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).
58. de Souza, N. The ENCODE project. *Nat. Methods* **9,** 1046 (2012).
59. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
60. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
61. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
62. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
63. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
64. Jager, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* **6,** 6178 (2015).
65. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471,** 467–472 (2011).
66. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol. Biol.* **1019,** 215–236 (2013).
67. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43,** 519–525 (2011).
68. Lu, Y. *et al.* Most common 'sporadic' cancers have a significant germline genetic component. *Hum. Mol. Genet.* **23,** 6112–6118 (2014).
69. Lee, S. H. *et al.* Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum. Mol. Genet.* **22,** 832–841 (2013).
70. Scales, M., Jager, R., Migliorini, G., Houlston, R. S. & Henrion, M. Y. visPIG--a web tool for producing multi-region, multi-track, multi-scale plots of genetic data. *PLoS ONE* **9,** e107497 (2014).

Acknowledgements

In the United Kingdom, Myeloma UK and Bloodwise provided principal funding. Additional funding was provided by Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund) and The Rosetrees Trust. This study made use of genotyping data on the 1958 Birth Cohort generated by the Wellcome Trust Sanger Institute (http://www.wtccc.org.uk). We are grateful to all investigators who contributed to NSCCG and GELCAPS, from which controls in the replication were drawn. We also thank the staff of the CTRU University of Leeds and the NCRI haematology Clinical Studies Group. The US GWAS was supported by a grant from the National Institutes of Health (P01CA055819). The German study was supported by the Dietmar-Hopp-Stiftung, Germany, the German Cancer Aid (110,131), the German Ministry of Education and Science (CLIOMMICS 01ZX1309), the German Research Council (DFG; Project SI 236/8-1, SI236/9-1, ER 155/6-1 and the DFG CRU 216) and the Multiple Myeloma Research Foundation. The patients were collected by the GMMG and DSMM studies. The German GWAS made use of genotyping data from the population-based HNR study, which is supported by the Heinz Nixdorf Foundation (Germany). The genotyping of the Illumina HumanOmni-1 Quad BeadChips of the HNR subjects was financed by the German Center for Neurodegenerative Disorders (DZNE), Bonn. We are grateful to all investigators who contributed to the generation of this data set. The German replication controls were collected by Peter Bugert, Institute of Transfusion Medicine and Immunology, Medical Faculty Mannheim, Heidelberg University, German Red Cross Blood Service of Baden-Württemberg-Hessen, Mannheim, Germany. This work was supported by research grants from the Swedish Foundation for Strategic Research (KF10-0009), the Marianne and Marcus Wallenberg Foundation (2010.0112), the Knut and Alice Wallenberg Foundation (2012.0193), the Swedish Research Council (2012–1753), the Royal

Swedish Academy of Science, ALF grants to the University and Regional Laboratories (Labmedicin Skåne), the Siv-Inger and Per-Erik Andersson Foundation, the Medical Faculty at Lund University and the Swedish Society of Medicine. We thank Jörgen Adolfsson, Tomas Axelsson, Anna Collin, Ildikó Frigyesi, Patrik Magnusson, Bertil Johansson, Jan Westin and Helga Ögmundsdóttir for their assistance. We are indebted to the clinicians who contributed samples to Swedish, Icelandic, Norwegian and Danish biobanks. We are indebted to the patients and other individuals who participated in the project.

Author contributions

R.S.H. designed the study. R.S.H. drafted the manuscript with contributions from K.H., G.J.M., B.N., N.W., J.S.M. and M.D. J.S.M. performed principal statistical and bioinformatic analyses. N.L., D.C.J., M.H., G.M. and O.L. performed additional bioinformatics analyses. P.B. coordinated UK laboratory analyses. N.L. performed genotyping and sequencing of UK samples. D.C.J. managed and prepared Myeloma IX and Myeloma XI Case Study DNA samples. M.K., G.J.M., F.E.D., W.A.G. and G.H.J. performed ascertainment and collection of Case Study samples. B.A.W. performed UK expression analyses. F.M.R. performed UK fluorescence *in situ* hybridization analyses. H.G., U.B., J.H., J.N., and N.W. coordinated and managed Heidelberg samples. C.L. and H.E. coordinated and managed Ulm/Wurzburg samples. A.F. coordinated German genotyping. C.C. performed German genotyping. P.H. and M.M.N. performed GWAS of German cases and controls. B.C. and M.I.d.S.F. carried out statistical analysis. K.H. coordinated the German part of the project. M.M.N. generated genotype data from the Heinz-Nixdorf recall study. M.H. and B.N. coordinated the Swedish/Norwegian part of the project. M.A. and B.-M.H. performed data analysis. B.S., M.J., E.J., S.L., C.H., A.-K.W., U.-H.M., H.N., S.N., A.V., U.V., A.W., I.T. and U.G. performed sample acquisition, sample preparation, clinical data acquisition and additional data analyses of Sweden/Norway samples. In Iceland, G.T. and D.F.G performed statistical analysis. S.Y.K. provided clinical information. T.R. performed additional analyses. U.T. and K.S. performed project oversight. M.v.D., P.S., A.B. and R.K. coordinated and prepared HOVON65/GMMG-HD4, HOVON87/NMSG18 and HOVON95/EMN02 studies for participating in this study, and coordinated genotyping and pre-processing. At the Myeloma Institute, University of Arkansas for Medical Sciences, N.W. coordinated the US part of the project and performed statistical and eQTL analyses. O.W.S. and N.W managed Case Study samples and performed confirmation genotyping. G.J.M. and F.E.D. performed ascertainment and collection of Case Study samples.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: G.T., P.S., G.M., D.F.G., T.R., K.S. and U.T. are employed by deCode Genetics/Amgen Inc. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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

Paper III

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04082-2 OPEN

The multiple myeloma risk allele at 5q15 lowers *ELL2* expression and increases ribosomal gene expression

Mina Ali[1], Ram Ajore[1], Anna-Karin Wihlborg[1], Abhishek Niroula [1], Bhairavi Swaminathan[1], Ellinor Johnsson[1], Owen W Stephens[2], Gareth Morgan[2], Tobias Meissner[3], Ingemar Turesson[1], Hartmut Goldschmidt[4,5], Ulf-Henrik Mellqvist[6], Urban Gullberg[1], Markus Hansson [1,7], Kari Hemminki[8,9], Hareth Nahi[10], Anders Waage[11], Niels Weinhold[2] & Björn Nilsson[1,12]

Recently, we identified *ELL2* as a susceptibility gene for multiple myeloma (MM). To understand its mechanism of action, we performed expression quantitative trait locus analysis in CD138$^+$ plasma cells from 1630 MM patients from four populations. We show that the MM risk allele lowers *ELL2* expression in these cells ($P_{\text{combined}} = 2.5 \times 10^{-27}$; $\beta_{\text{combined}} = -0.24$ SD), but not in peripheral blood or other tissues. Consistent with this, several variants representing the MM risk allele map to regulatory genomic regions, and three yield reduced transcriptional activity in plasmocytoma cell lines. One of these (rs3777189-C) co-locates with the best-supported lead variants for *ELL2* expression and MM risk, and reduces binding of MAFF/G/K family transcription factors. Moreover, further analysis reveals that the MM risk allele associates with upregulation of gene sets related to ribosome biogenesis, and knockout/knockdown and rescue experiments in plasmocytoma cell lines support a cause–effect relationship. Our results provide mechanistic insight into MM predisposition.

[1] Department of Laboratory Medicine, Hematology and Transfusion Medicine, SE 221 84 Lund, Sweden. [2] Myeloma Institute for Research and Therapy, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA. [3] Department of Molecular and Experimental Medicine, Avera Cancer Institute, Sioux Falls, SD 57105, USA. [4] Department of Internal Medicine V, University of Heidelberg, 69117 Heidelberg, Germany. [5] National Center for Tumor Diseases, Ulm 69120 Heidelberg, Germany. [6] Section of Hematology, South Elvsborg Hospital, SE 501 83 Borås, Sweden. [7] Hematology Clinic, Skåne University Hospital, SE 221 85 Lund, Sweden. [8] German Cancer Research Center, 69120 Heidelberg, Germany. [9] Center for Primary Health Care Research, Lund University, SE 205 02 Malmö, Sweden. [10] Center for Hematology and Regenerative Medicine, Karolinska Institutet, SE 171 77 Stockholm, Sweden. [11] Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway. [12] Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA. These authors contributed equally: Ram Ajore, Anna-Karin Wihlborg. Correspondence and requests for materials should be addressed to B.N. (email: bjorn.nilsson@med.lu.se)

Multiple myeloma (MM) is the second most common hematologic malignancy. It is defined by an uninhibited, clonal growth of plasma cells in the bone marrow, producing a monoclonal immunoglobulin ("M protein") that can be detected in peripheral blood[1]. Clinically, MM is characterized by bone marrow failure, lytic bone lesions, hypercalcemia, and kidney failure. It is preceded by monoclonal gammopathy of unknown significance (MGUS)[2,3], a common premalignant condition that progresses to MM at a rate of about 1% per year[4].

Several lines of evidence support that heritable factors contribute to the development of MM. Since the 1920s, several authors have reported families with multiple cases of MM and MGUS[5,6]. Systematic family studies show that first-degree relatives of patients with MM and MGUS have two to four times higher risk of MM, and a higher risk of certain other malignancies[6–11]. Recently, genome-wide association studies have identified DNA sequence variants at 18 independent loci that associate with MM risk[12–15], and show further polygenic etiology in a subset of familial MM cases[16].

One of the MM susceptibility genes is *ELL2* (elongation factor for RNA polymerase II 2)[12,13] at chromosome 5q15. This gene encodes a key component of the super-elongation complex (SEC), which enhances the catalytic rate of RNA polymerase II[17,18]. *ELL2* is highly expressed in normal and MM plasma cells, and helps RNA polymerase II find a promoter-proximal weak poly (A)-site in the immunoglobulin (Ig) heavy gene that is hidden in B cells, allowing Ig heavy chain messenger RNA (mRNA) to be translated to secreted Ig at a high rate[13,19,20]. Conditional B-lineage *Ell2* knockout mice show curtailed humoral immune responses, reduced numbers of plasma cells, and abnormal plasma cell morphology[21–23]. The *ELL2* MM risk allele is represented by ~70 sequence variants in tight linkage disequilibrium ($r^2 > 0.8$ with the first reported lead variant rs56219066[13] or the lead variant from a subsequent multi-center analysis, rs1423269; $r^2/D' = 0.96/0.98$ with rs56219066)[12]. Interestingly, the same allele that predisposes for MM also associates with lower Ig levels[13], altered Ig glycosylation[24], lower total serum protein levels[25], and an increased risk of MGUS[13], salivary gland carcinoma[26], and possibly bacterial meningitis[13].

Here we investigate the effects of the *ELL2* MM risk allele. Since this allele is represented by non-coding variants (apart from one missense variant of unclear relevance[13]), we hypothesize that its effects are due to changes in *ELL2* expression. Using expression quantitative locus (eQTL) analysis, we detect a negative effect of the MM risk allele on *ELL2* expression in MM plasma cells. This finding is further supported by data showing that several of the risk variants map to regulatory chromosomal regions, including three that yielded reduce transcriptional activity. Interestingly, one of these (rs3777189-C) is located only 514 bp from the lead variant for *ELL2* expression (rs9314162) and 2616 bp from the best-supported lead variant for MM risk (rs1423269), and diminishes binding of MAFF/G/K family transcription factors. In addition to the effect on *ELL2* itself, we find that the MM risk allele perturbs the expression of genes involved in ribosome biogenesis and function.

Results

The MM risk allele lowers *ELL2* expression in MM plasma cells. To identify effects of the *ELL2* MM risk allele on gene expression, we generated mRNA-sequencing data for CD138$^+$ plasma cells from bone marrow samples from 185 MM patients from Sweden and Norway, and genotyped these samples for one of the linked MM risk variants at the *ELL2* locus (rs3815768; Supplementary Fig. 1a). In addition, 158 of the samples were genotyped using Illumina OmniExpress™ single-nucleotide polymorphism (SNP) microarrays, and imputed using phased haplotypes from the 1000 Genomes compendium[27].

In our mRNA sequence data, we found that the MM risk allele lowers *ELL2* expression. While this effect was clearest across the distal part of the gene (exons 9–11; Pearson correlation $P = 0.007–0.01$, $\beta = -0.19$ to -0.20), we saw significant associations with all exons (Fig. 1a and Table 1), except with exons 7 and 8, which could not be quantified reliably for technical reasons (Supplementary Fig. 1b), and the last exon, which could not be quantified accurately because of uneven coverage in the 3′ untranslated region. Samples heterozygous and homozygous for the risk allele showed 34% and 43% lower *ELL2* expression, respectively (average across exons 1–6 and 9–11) than samples homozygous for the protective allele. We also observed an allelic imbalance in expression for heterozygous individuals among rs3815768-TC heterozygotes (54.5% for T-allele vs 45.4% C-allele; $P < 0.005$). No differences in *ELL2* splicing patterns were detected between the T- and C-allele using replicate multivariate analysis of transcript splicing[28].

For further validation of the observed effect, we used gene expression microarray data for CD138$^+$ plasma cells from MM patients from Germany ($n = 658$), the United Kingdom ($n = 183$), and the USA ($n = 604$)[12,29]. In all these datasets, rs3815768-C associates with lower *ELL2* expression (Fig. 1b; Fisher's inverse χ^2 test combined $P = 2.5 \times 10^{-27}$ and $\beta = -0.24$ for the four datasets). Moreover, regional analysis of these data and the Swedish-Norwegian samples genotyped on SNP microarrays showed that the set of variants that most strongly influence MM risk are those that have the largest effect on *ELL2* expression (Fig. 2a, b). Additionally, we observed slightly more significant *P* values across the second half of intron 2 and across intron 3, including both the lead variant for *ELL2* expression (rs9314162) and MM risk (rs1423269). These data demonstrate a concordance between the effects of sequence variants on *ELL2* expression and MM risk, and indicate that the same sequence variations at this locus affect both.

Effect on *ELL2* expression in other cell types. While *ELL2* is highly expressed in normal and malignant plasma cells, the gene is also expressed in other cell types, including red blood cell precursors, salivary gland, and pancreatic islets (Supplementary Fig. 2)[13,30,31]. Curiously, these cell types resemble plasma cells in that they produce large amounts of protein (hemoglobin, amylase, and peptide hormones), and the same allele that predisposes to MM also predisposes to salivary gland carcinoma (rs3777204; $r^2/D' = 0.96/0.98$ with rs1423269)[32]. Yet, unlike the highly reproducible effect on *ELL2* expression in MM plasma cells, we could not detect any effect on *ELL2* expression in mRNA-sequencing data from peripheral blood from 2515 Icelanders (Supplementary Fig. 3), nor in eQTL data from 8086 Europeans in the Blood eQTL database[33] or any of the 44 tissues represented in GTEx[34]. Although some tissues, including salivary gland, could not be studied because of lack of data, these results indicate that the effects of the MM risk allele on *ELL2* expression are restricted to certain cell types.

Identification of causal variants. A total of 67 SNPs and 5 small insertions/deletions are highly correlated with the best-supported sentinel MM risk variant (rs1423269) and the strongest *ELL2* expression variant (rs9314162) ($r^2 > 0.8$; Supplementary Tables 1 and 2). Hypothetically, some of these variants may be causal in that they alter the efficiency of *ELL2* transcription, whereas others only tag the causal markers. To search for such causal variants, we considered variants

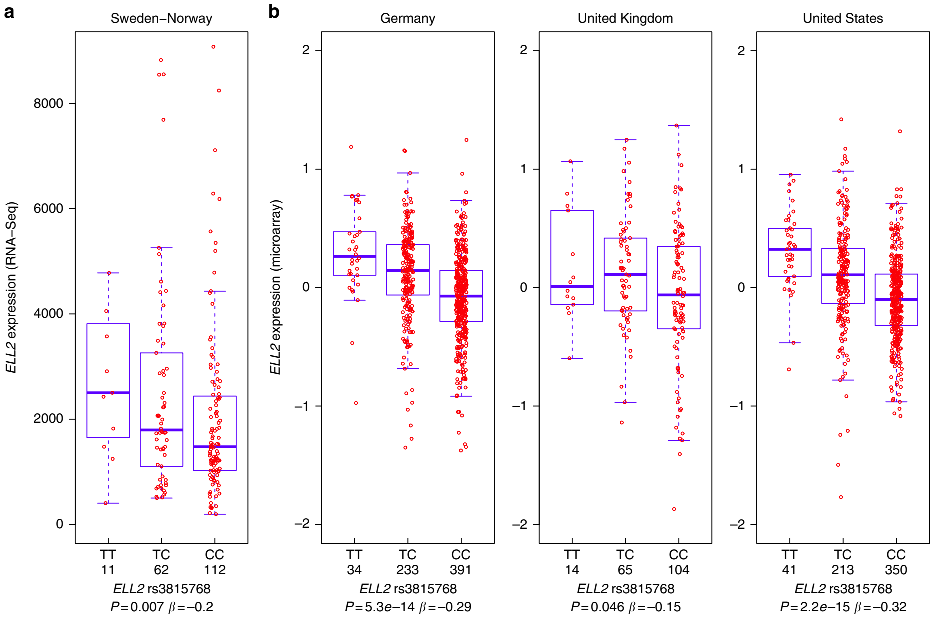

Fig. 1 The *ELL2* MM risk allele confers lower *ELL2* expression in MM plasma cells. **a** Boxplot showing decreased expression of *ELL2* exon 10 associated with the MM risk allele (represented by rs3815768-C) compared to the protective allele (rs3815768-T) in CD138$^+$ bone marrow cells from 185 Swedish and Norwegian MM patients. Expression values are fragments per kilo base of exon per million fragments mapped (FPKM) obtained by mRNA sequencing. **b** Boxplots showing a corresponding association in Affymetrix gene expression microarray data CD138$^+$ bone marrow cells from MM patients from Germany, United Kingdom, and the US. Boxes indicate medians and the first and third quartiles. Whiskers indicate first and third quartiles plus 1.5 times the interquartile range. Outliers are plotted as individual dots. Pearson correlation *P* values and effect size (*β*) indicated

Table 1 Association testing results

Exon	Chr.	Position (hg38)	Effect size (*β*)	r^2	*P*
1	5	95885097-95888987	−0.15	0.02	0.036
2	5	95889085-95889130	−0.16	0.03	0.026
3	5	95891102-95891274	−0.17	0.03	0.018
4	5	95895627-95895691	−0.16	0.03	0.025
5	5	95898239-95898810	−0.15	0.02	0.040
6	5	95900692-95900780	−0.14	0.02	0.050
7	5	95900955-95901080	−0.04	0.00	0.601
8	5	95906522-95906782	−0.11	0.01	0.133
9	5	95913770-95913934	−0.20	0.04	0.007
10	5	95919423-95919545	−0.20	0.04	0.007
11	5	95913001-95913049	−0.19	0.04	0.010
12	5	95911574-95912071	−0.12	0.02	0.093

Association testing results for the Swedish-Norwegian mRNA-sequencing dataset ($n = 185$). Effect size (*β*), squared Pearson regression coefficient (r^2), and *P* values indicated

in linkage disequilibrium ($r^2 > 0.8$) with rs9314162 that associate with both *ELL2* expression and MM (top-right clusters in Fig. 2b) and map to regulatory regions. To delineate regulatory regions, we used ChIP-seq (chromatin immunoprecipitation with next-generation sequencing) data for histone modifications representing enhancers and promoters, and for transcription factors, in GM12878 lymphoid cells from the ENCODE and Roadmap compendia (Supplementary Table 1)[35,36]. In addition, we generated ChIP-seq data for H3K4me3 histone marks in the L363 plasma cell leukemia cell line to delineate promoter regions relevant in plasma cells. Using our criteria, we identified eight candidate variants (rs1841010, rs9314162, rs3777189, rs3777185, rs4563648, rs6877329, rs3777184, and rs889302). All of these mapped near rs1423269 and rs9314162, and five (rs3777185, rs4563648, rs6877329, rs3777184, and rs889302) to an internal promoter in intron 2, as defined by the presence of the H3K4me3 histone mark (Fig. 2c).

To evaluate the candidate variants, we made luciferase vectors containing 120 bp of genome sequence with the respective risk and protective variants in the center (Supplementary Table 3). We transfected these vectors into three plasma cell lines (L363, OPM2, and RPMI-8226) and two cell lines representing other hematologic lineages (K562 and MOLM-13; acute myeloid leukemia cell lines with eryhtroblastic and monocytic differentiation, respectively). Consistent with our observation of an eQTL effect in MM plasma cells but not in peripheral blood, three risk variants (rs3777189-C, rs3777185-C, and rs4563648-G) yielded decreased luciferase activity relative to their corresponding protective variants in plasma cell lines, but not in non-plasma cell lines (Fig. 3a). Interestingly, rs3777189 is located only 514 bp from rs9314162; and rs3777185 and rs4563648 in the internal promoter in intron 2.

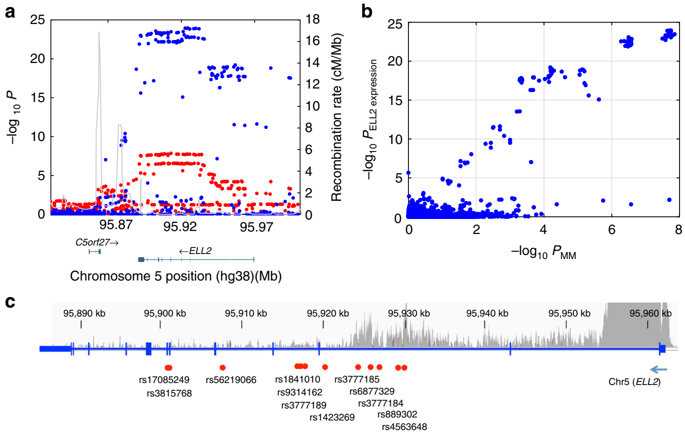

Fig. 2 *ELL2* MM risk variants coincide with *ELL2* eQTLs. **a** Regional association plot of the *ELL2* locus at chromosome 5q15. Blue dots indicate association with *ELL2* expression based on the four sample sets ($-\log_{10}$-transformed Fisher's inverse χ^2 *P* values; meta-analysis of 158 Swedish-Norwegian and 1445 SNP microarray-genotyped samples from Germany, UK, and US). Red dots indicate association with MM ($-\log_{10}$-transformed logistic regression *P* values from our previous Swedish-Norwegian-Iceland MM association study). The lead variant for the effect on *ELL2* expression is rs9314162. Meiotic recombination rates calculated from the 1000 Genomes compendium indicted by the gray curve. **b** Two-dimensional plot of the same association *P* values. The two top-right clusters contained 66 variants influencing both MM risk and *ELL2* expression. **c** Schematic representation of *ELL2*. The indicated variants are the lead variants for *ELL2* expression (rs9314162), the first reported MM lead variant (rs56219066), the best-supported MM lead variant (rs1423269), the eight variants selected for functional evaluation (rs1841010, rs9314162, rs3777189, rs3777185, rs6877329, rs3777184, rs889302, and rs4563648), and the coding variant rs3815768 used for genotyping in the mRNA-sequencing data. Gray curve indicates ChIP-seq read density for the H3K4me3 histone mark in L363 cells, and main (high peak around exon 1) and internal promoters (lower peaks across introns 1 and 2)

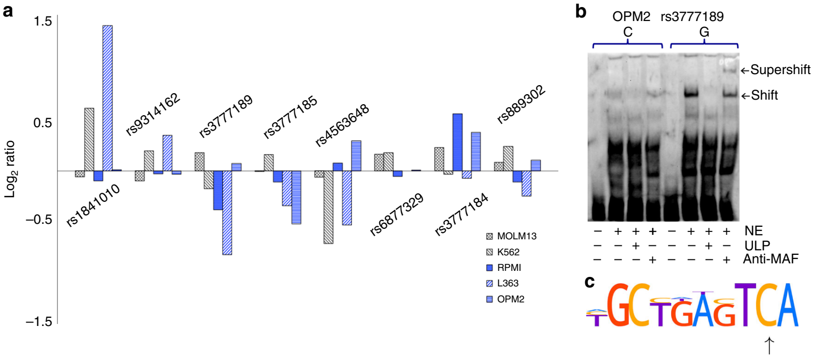

Fig. 3 Identification of causal variants. **a** We evaluated eight variants located in regulatory regions of *ELL2* (Fig. 2c and Supplementary Table 1) using luciferase assays in RPMI-8226, OPM2, and L363 plasmocytoma lines, and in K562 and MOLM-13 cells that represent other hematologic lineages. Consistent with the effect on *ELL2* expression in plasma cells but not in peripheral blood, three risk variants (rs3777189-C, rs3777185-C, and rs4563648-G) yielded reduced activity in the plasma cell lines but not in the control cell lines, whereas the remaining variants showed opposite or inconsistent effects. Plotted values represent $\log_2$-ratios of luciferase activities for the risk alleles over their corresponding protective alleles (median over tri or quadruplicates). **b** To identify the responsible transcription factors, we carried out sequence analyses and electrophoretic mobility shift assays (EMSA) (Supplementary Figs. 5 and 6). Shown here is EMSA with nuclear extract (NE) from OPM2 cells (lanes 2–4 and 6–8), and probes representing genomic sequence at rs3777189 with the risk/low-expressing allele (C) or the protective/high-expressing allele (G) in the center (lanes 1–4 and 5–8, respectively). The G allele showed an allele-specific shift (lane 6) that was outcompeted with unlabeled probe (ULP; lane 7) and supershifted with MAFF/G/K antibody (lane 8). Similar results were seen with L363 cells (Supplementary Fig. 5). **c** Sequence analysis predicted loss of a MAFF/G/K motif (Supplementary Table 4). Shown here is the MAFK motif from the HOCOMOCO-10 database. Arrow indicates G changed to C by the rs3777189-C risk variant

We screened these three variants for gain or loss of transcription factor-binding motifs. We identified numerous candidate factors, about 20 of which are expressed in MM plasma cells (Supplementary Tables 4 and 5). Electrophoretic mobility shift assays (EMSAs) with L363 and OPM2 nuclear extracts revealed allele-dependent binding of nuclear proteins for rs3777189 and rs3777185, but not for rs4563648 (Supplementary Fig. 4).

To search for differentially bound nuclear proteins, we carried out EMSA assays with antibodies against factors predicted to gain or lose a binding site at rs3777189 or rs3777185. We observed supershift with antibody against the MAFF/G/K transcription factors with probes for the protective/high-expressing allele rs3777189-G, but not with probes for the risk/low-expressing allele rs3777189-C (Fig. 3b, c and Supplementary Fig. 5). Moreover, *ELL2* expression correlated with *MAFK* and *MAFG* expression (Supplementary Table 5), and rs3777189 maps to an annotated MAFK ChIP-seq peak in lymphoid cells (Supplementary Table 1). The MAF protein family (MAF, MAFA, MAFF, MAFG, and MAFK) are paralogous basic leucine zipper (bZIP)-type transcription factors that form homo and heterodimers both with each other and certain other bZIP transcription factors (e.g., BACH1)[37–39]. MAFF/G/K are thought to be functionally redundant, and have similar binding motifs (Supplementary Table 4). Our results indicate that rs3777189-C leads to loss of a binding site for at least one of MAFF/K/G, and thereby reduced transcriptional drive. No additional supershifts were identified for rs3777189 or rs3777185 (Supplementary Fig. 6).

The *ELL2* MM risk allele upregulates ribosomal genes. ELL2 is a key component of the SEC. Accordingly, variation in *ELL2* expression could influence gene expression in a broader sense, either through modulation of RNA polymerase II or through cellular responses to altered protein synthesis. Consistent with this notion, mouse studies have shown that *Ell2* influences Ig heavy chain exon usage, and the processing of a large percentage of transcripts in plasma cells[22,23,40].

To gain insight into the downstream effect of variation in ELL2 function, we first calculated the correlation between *ELL2* and other genes expressed in MM plasma cells in the Swedish-Norwegian mRNA-sequencing data, which had high sequence coverage (about 100 million reads per sample) and allow accurate, linear estimation of transcript levels. Here, *ELL2* showed a significant correlation with a large set of genes, including 4890 genes with <5% false discovery rate (Supplementary Data 1). Interestingly, gene set enrichment analysis showed an over-representation of positive correlations among multiple gene sets related to ribosomal biogenesis and function (Supplementary Table 6), including a set of 80 genes encoding the proteins of the large and small ribosomal subunits (ribosomal protein coding genes, RPGs) and a set of seven genes encoding other members of the SEC (Fig. 4a)[41]. These results are consistent with co-regulation of cellular components required for high-rate protein synthesis, and the role of ELL2 in driving the production of secreted Ig.

Next, we correlated the *ELL2* MM risk allele with the expression of other genes in the mRNA-sequencing dataset. Compared to the signature obtained by correlating with *ELL2*

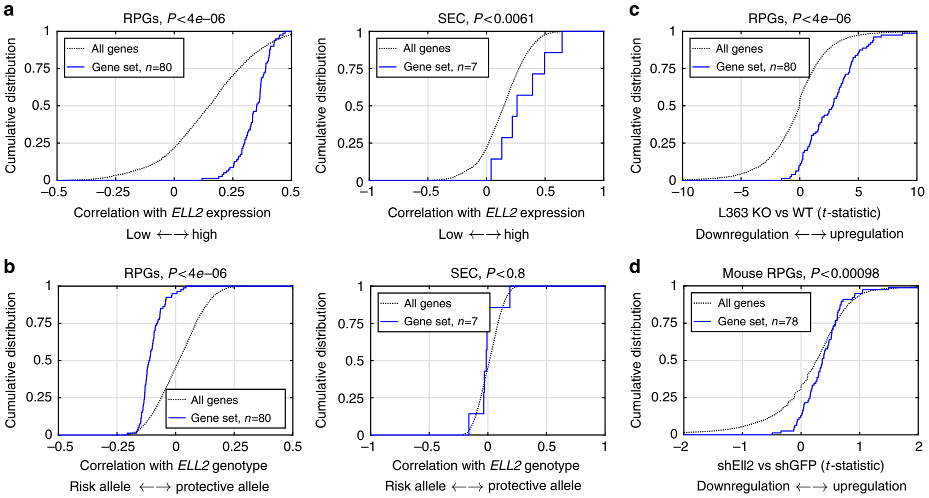

Fig. 4 The *ELL2* MM risk allele increases ribosomal gene expression. **a** To explore the downstream effects of reduced ELL2 function, we first calculated the correlation between *ELL2* and other genes in the Swedish-Norwegian mRNA-sequencing data. Here, *ELL2* showed significant correlation with a large set of genes. Enrichment analysis revealed an over-representation of positive correlations among multiple gene sets related to ribosomes biogenesis and function, including ribosomal protein coding genes (RPGs) and the SEC (see also Supplementary Table 6). **b** Enrichment analysis of correlation between the *ELL2* MM risk allele and gene expression in the same dataset identified RPGs and other gene sets related to ribosomes. This enrichment was in the direction of the *ELL2* risk allele, which confers lower *ELL2* expression (see also Supplementary Table 7). **c** Similarly, analysis of *ELL2* CRISPR-Cas9 knockout (KO) L363 cells showed an upregulation of RPGs and other gene sets related to ribosome biogenesis and function (see also Supplementary Tables 8 and 9), i.e., effects in the same direction as the *ELL2* MM risk allele. **d** Finally, similar changes were seen in mouse MPC1 plasmocytoma cells treated with shRNA against either *Ell2* vs GFP. These data support that, in addition to the effect on *ELL2* itself, the *ELL2* MM risk allele confers additional changes in gene expression, including an increased expression of genes involved in ribosomal biogenesis, possibly as a compensatory reaction

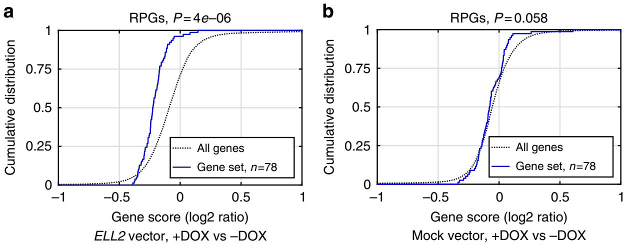

Fig. 5 Reconstitution of *ELL2* expression. For further validation of the effect of *ELL2* knockout on RPG expression, we transduced CRISPR-resistant, doxycycline-inducible *ELL2* or mock vector into our L363-*ELL2*-KO cells. Following culture with or without doxycycline (DOX), the cells were gene expression-profiled using mRNA sequencing: **a** Comparing the gene expression profiles of *ELL2*-transduced cells cultured with ($n = 3$) vs without DOX ($n = 4$), we observed an enrichment of negative gene scores (i.e., downregulation) of ribosomal gene sets, consistent with a rescue effect; **b** no similar enrichment was seen with mock-transduced control cells ($n = 4$ samples with vs 4 without DOX). These data further support a cause–effect relationship, and that the results in Fig. 4b are not due to CRISPR-Cas9 off-target effects

expression, this signature was weaker (Supplementary Data 2), which is expected as the *ELL2* MM risk allele only explains a part of the variance in *ELL2* expression. Yet, gene set enrichment analysis again identified gene sets related to ribosome biogenesis and function (Fig. 4b and Supplementary Table 7). Unexpectedly, the detected enrichment was in the direction of the *ELL2* MM risk allele, which confers lower *ELL2* expression.

To understand whether the association with ribosomal gene expression reflects a cause–effect relationship, we knocked out *ELL2* in L363 cells using lentiviral CRISPR-Cas9 (Supplementary Fig. 7), and analyzed knockout and wild-type cells by mRNA sequencing. Strikingly, L363-*ELL2*-KO cells showed a significant enrichment of increased expression for RPGs (Fig. 4c) and other gene sets related to ribosome biogenesis and function (Supplementary Tables 8 and 9). We also observed a similar trend in pre-existing mRNA-sequencing data from mouse plasmocytoma cells treated with short hairpin RNA against Ell2 or GFP (Fig. 4d). To exclude off-target effects of CRISPR-Cas9 editing or lentivirus integration, we carried out rescue experiments where *ELL2* expression was reconstituted in the L363-*ELL2*-KO cells. For this, we generated a vector containing CRISPR-resistant *ELL2* controlled by a doxycycline-inducible promoter (Supplementary Fig. 8). *ELL2*- and mock-transduced L363-KO cells were cultured with and without doxycycline, and analyzed with mRNA sequencing. Consistent with a rescue effect, we observed doxycycline-dependent downregulation of ribosomal genes in *ELL2*-transfected cells, but not in mock-transfected cells (Fig. 5). These data support that decreased *ELL2* expression/function increases ribosomal biogenesis, possibly as a compensatory reaction in response to reduced protein synthesis.

Discussion

ELL2 has been associated with MM and several other phenotypes. It has been postulated that the MM risk allele has a negative effect on *ELL2* function, yet the reason for this has been unclear. We show that the MM risk allele lowers *ELL2* expression in plasma cells, providing an explanation for the hypomorphic effect. Further, we identify three risk variants that map to regulatory regions and yield decreased transcriptional activity in plasmocytoma cell lines. Two of these (rs3777185 and rs3777189) exhibit altered binding of nuclear proteins, and rs3777189-C diminishes binding of MAFF/G/K family transcription factors. In addition, we identify increased expression of ribosomal genes as a downstream effect.

Our data are consistent with a working model where the expression of *ELL2* is co-regulated with the expression of ribosomal gene sets to allow production of secreted Ig in a coordinated manner. The MM risk allele confers lower *ELL2* expression, which makes the production of secreted Ig less efficient[13,19,21–23]. Hypothetically, plasma cells sense this and try to compensate by increasing the drive on Ig synthesis, which leads to relative upregulation of gene sets related to ribosome biogenesis and function. Such a model would explain the co-occurence of the positive correlation between *ELL2* and ribosomal gene sets, and the negative correlation between the *ELL2* MM risk allele and ribosomal gene sets.

Regarding limitations, our study is based on plasma cells from MM patients. While it seems likely that our findings extend to normal plasma cells, it remains verify this using samples from healthy individuals. However, this is hard to do in practice as it would require isolation of CD138$^+$ cells from bone marrow samples from a large number of healthy volunteers. Moreover, this isolation would need to be done by fluorescence-activated flow cytometry, instead of magnetic-bead sorting, as plasma cells are rare (<1% of nucleated cells) in samples from healthy individuals. It would also be interesting to test whether our findings extend to patients with MGUS or smoldering MM. Further, while complete testing of all the linked variants that tag the *ELL2* MM risk allele would be desirable, we focused on variants in regulatory regions for practical reasons. Similarly, our data do not exclude an effect of the missense variant rs3815768 on top of the reduced expression, and we have not been able to look for effects at the protein level due to lack of material. Finally, it would be interesting to look for effects on *ELL2* and ribosomal gene sets in salivary gland samples, in light of the recently reported association with salivary gland cancer[32].

An intriguing question is how the *ELL2* risk allele promotes MM development. Hypothetically, one possibility is that the lower Ig levels could lead to slower antigen clearance and stimulation of the B-cell system for longer periods of time, and thereby a higher risk of malignant transformation. Alternatively, it is conceivable that altered ribosome function could promote MM development owing to the connection between altered ribosome biogenesis and perturbation of oncogenic pathways (c.f., refs. [42–45] and references therein).

Methods

Study populations. To generate the Swedish-Norwegian gene expression dataset, we used CD138$^+$ cells isolated from 185 bone marrow samples obtained at diagnosis from MM patients. For 158 samples, we also obtained matching DNA from peripheral blood (Swedish National Myeloma Biobank, Lund, Sweden and Norwegian Biobank for Myeloma, Trondheim, Norway). Finally, to look for effects of *ELL2* expression in peripheral blood, we used mRNA expression data for 2515

Icelandic samples (deCODE Genetics, Reykjavik; unpublished). The sample collection was done subject to informed consent and ethical approval (Lund University Ethical Review Board, 2013/54; Icelandic Data Protection Authority, 2001010157; and National Bioethics Committee 01/015), and in accordance with the principles of the Declaration of Helsinki.

For validation, we used three sets of pre-existing gene expression profiles of CD138$^+$ plasma cells isolated from MM patients from Germany, UK, and USA[29]. The German sample set consists of 658 MM patients from the Heidelberg University Clinic and the German-speaking Myeloma Multicenter Group[29]. The British sample set comprises 183 MM patients enrolled in the UK Medical Research Council Myeloma IX trial[29]. The US sample set comprises 604 samples from newly diagnosed patients treated at the UAMS Myeloma Institute for Research and Therapy[12]. The three validation datasets were generated using Affymetrix U133 2.0 plus microarrays and custom chip definition file ("BrainArray"; http://brainarray.mhri.med.umich.edu/Brainarray/Database/CustomCDF).

Gene expression profiling of Swedish-Norwegian samples. For the Swedish-Norwegian samples, total RNA was purified from immune-magnetically isolated CD138$^+$ cells using standard methods (Macherey Nagel NucleoSpin® RNA #740955.10 or QIAamp RNA blood #52304). Icelandic blood samples were collected in PAXgene tubes (PreAnalytix, Switzerland; cat no. #762165) and RNA was isolated using the PAXgene 96 Blood RNA or the Paxgene Blood RNA Kit (PreAnalytix; cat nos. #762331 or #762174). The RNA integrity (RIN) was assessed using the BioAnalyzer (Agilent, Santa Clara, CA, USA) or LabChip GX (PerkinElmer, Waltham, MA, USA) instruments. Indexed sequencing libraries were prepared using the TruSeq RNA sample preparation v2 kit in 96-well format (Illumina, San Diego, CA, USA). Between 0.1 and 1 µg of total RNA was used for poly-A mRNA capture using oligo-dT attached magnetic beads. Complementary DNA synthesis was done using SuperScript II and random hexamer priming (ThermoFisher, Waltham, MA, USA). End-repair, 3′-adenylation, ligation of indexed adaptors and PCR amplification was performed according to Illumina protocols. Quantity and quality of each sequencing library was assessed using the LabChip GX, followed by standard dilutions and sample/plate storage at −20 °C. Further quality assessment was performed by doing pool sequencing (≤24 samples/pool) on a MiSeq instrument in order to optimize cluster densities and assess insert size, sample diversity, and so on. Primary processing and base calling was done using HCS1.3.8–1.4.8 and RTA1.10.36–1.12.4.2 analysis packages. Demultiplexing and generation of FASTQ files was performed using scripts from Illumina (bcl2fastq v.1.8). Sequence alignment and fragment counts was done with TopHat2 and HTSeq-count, respectively[46,47]. The plasma cell gene expression data will be deposited in the NCBI Gene Expression Omnibus (GEO) database when the manuscript is accepted. The German, UK, and US gene expression datasets were generated in previous studies using Affymetrix U133A 2.0 plus arrays with a custom chip definition file (v.17)[1,2].

Genotyping. The Swedish-Norwegian sample set was genotyped at two levels: first, all samples ($n = 185$) were genotyped for the ELL2 MM risk allele using the coding variant rs3815768, which could be robustly typed manually from the RNA-sequencing data using Integrative Genomics Viewer (Supplementary Fig. 1a). In addition, a subset of the Swedish-Norwegian samples was genotyped on Illumina Human OmniExpress microarrays ($n = 158$). To increase the genomic resolution, these data were haplotype-phased using SHAPEIT2 (v2.790)[48] and imputed by IMPUTE2 (v2.3.2)[49] with the 1000 Genomes Phase 3 compendium reference data (October 2014 release)[27]. The German, UK, and US myeloma sample sets were genotyped previously on Illumina Human OmniExpress-12 v.1.0 arrays[12,29] and imputed using the UK10K compendium[14,15,50]. For the Icelandic blood samples, genotypes were obtained by imputing variants identified by whole-genome sequencing of 8453 Icelanders into 150,656 chip genotyped individuals using long-range phasing based imputation[51,52]. Probabilities of genotypes were also predicted for 294,212 first and second-degree relatives of chip-typed individuals[53]. A description of the alignment to the reference genome, genotype calling, and imputation and haplotype phasing is given in a recent publication[54].

Association testing. In the Swedish-Norwegian sample set, test of association between the *ELL2* risk variants and expression values generated from the MM plasma cell mRNA-sequencing data was done at the exon level, in order to allow detection of exon-specific effects and to avoid signal dilution due to alignment bias caused by coding variants (Supplementary Fig. 1b). For association testing, we used Pearson correlation as implemented in R (v.3.3)[55]. Effect sizes (beta, β) and standard errors (SE) of eQTLs were calculated using R (v.3.3). The coefficient of linkage disequilibrium (D′) and r-squared (r^2) were calculated using the Central European part of the 1000 Genomes compendium as available via HaploReg 4.1. To estimate risk allele ratios in rs3815768-CT heterozygotes, we counted the two allelic sequences ([C/T]AGCATTCTGAGACGGATTTAGTTTTC, representing the site of rs3815768) in the raw RNA-sequencing reads using BBTools (http://jgi.doe.gov/data-and-tools/bbtools). Exact matches of the variant sequence and its complement were counted. In the German, UK, and US sample sets, the association was done using MatrixEQTL under a linear model[12,29]. In the Icelandic mRNA-sequencing dataset, we used generalized linear regression to test for association on rank-transformed expression estimates. To account for family structure, an estimate of the inverted kinship matrix was incorporated into the test[52]. Effect sizes (beta, β) and SE of eQTLs were calculated using R (v.2.8). Meta-analysis of P values for eQTL associations was performed using the Fisher's inverse χ^2 test in MATLAB.

Chromatin immunoprecipitation sequencing. L363 cells were cross-linked with 1% paraformaldehyde (ThermoFisher, #28908) at 37 °C in water bath for 11 min. Shearing and immunoprecipitation was done according to manufacturer's instructions (Millipore, #17-10085). The DNA was sonicated between 200–400 bp fragment length on Bioruptor Pico Sonication System (Diagenode) at 4 °C for 30 s/30 s and 13 cycles. To pull down fragments, we used 1–10 µg of H3K4me3 (Millipore, #04-745) and isotype control antibodies (normal rabbit IgG, #sc-2027, Santa Cruz Biotechnology). Fragments were de-cross-linked and purified using ChIP clean and concentrate kit (Zymogen, #D5205). Concentration was measured using Qubit 2.0 fluorometer. The ChIP-Seq library was prepared using ThruPLEX DNA-seq Kit (RUBICON GENOMICS, #R400406). Following amplification, samples were run on bioanalyzer to verify amplification and fragment size. The library was purified using AMPure XP protocol described in ThruPLEX DNA-seq Kit instruction manual. The library was diluted with nuclease-free water to 2 nM concentration. Dual-indexed libraries were sequenced on Illumina HiSeq 2500 sequencer using the TruSeq v4 cluster and SBS sequencing kits, respectively (paired-end; 2 × 125 cycles). Demultiplexing and generation of FASTQ files was performed using scripts from Illumina (bcl2fastq v.1.8). FastQC (v0.11.5)[56] was used to assess read quality, GC content, the presence of adaptors, over-represented k-mers and duplicated reads. Bases with low quality score were removed using Trimmomatic program (v.0.36)[57]. Trimmed reads were aligned using Bowtie2 (v.2.3.0)[58].

Luciferase assays. Ten double-stranded nucleotide sequences of 120 bp each, including with KpnI and BglII restriction sites at terminal ends, were commercially synthesized (Integrated DNA Technologies, USA). The sequences correspond to rs1841010, rs9314162, rs3777189, rs3777185, rs6877329, rs3777184, rs889302, and rs4563648 (Supplementary Table 3). Sequences were directionally cloned into a pGL3-Basic plasmid (Promega) upstream of a luciferase reporter gene[59]. Sanger sequencing confirmed the inserts. Renilla luciferase was used as internal transfection control. L363, OPM2, RPMI-8226, MOLM-13, and K562 cells were cultured at 37 °C and 5% CO_2 in RPMI 1640 medium (Gibco, Life Technologies) supplemented with 10% fetal bovine serum (Gibco). These cells were transfected with each of the ten clones using Neon system (ThermoFisher). Post 24 h transfection, cells were harvested and lysed in lysis buffer. An aliquot of 20 µl of the lysed cells was used for luciferase measurement following manufacturer's protocol (dual-luciferase reporter assay system, Promega). Measurements were performed at GLOMAX 20/20 Luminometer using Run Promega Protocol (DLR-0-INJ). Effects were quantified as $\log_2$ ratios of renilla-normalized luminiscence values for the risk alleles divided by the corresponding values for the protective alleles (median over three to seven replicates per sequence and cell line).

Electrophoretic mobility shift assays. For nuclear proteins and gel shifts[59,60], we used the following 25-bp double-stranded probes (variants in brackets): for rs3777189, ACAGTGCTGACT[G/C]AGCTCAAAATAC; rs3777185, CTCTGAAACTCT [G/A]CCTGAATGGCTC; rs4563648, GAAACTTTCTCA[C/T]CCTGACATTTGT. All probes were biotin-labeled at the 5′end of both strands; unlabeled specific competitor probes with identical sequences were used to test for specificity. For supershift assays with nuclear extracts from OPM2 and L363 cell lines (DSMZ, Braunschweig), we used these antibodies: BACH1 (#sc-271211, Santa Cruz Biotechnology), JunB (#3753S, Cell Signaling Technology), c-Fos (#4384S, Cell Signaling Technology), and MafF/G/K (D-12), #sc-166548, Santa Cruz Biotechnology. In essence, 1–2 µg antibody was added to the reaction mix and incubated 15 min at room temperature, before addition of probes and another 20 min incubation at room temperature. The cell line identities was confirmed by the supplier and mycoplasma was eliminated with ciprofloxacin, then confirmed negative in microbiological culture, RNA hybridization, and PCR assays (DSMZ, Braunschweig).

Motif analysis. To identify transcription factors whose motifs are gained or lost by sequence variants, we used PERFECTOS-APE (http://opera.autosome.ru/perfectosape) with the HOCOMOCO-10, JASPAR, HT-SELEX, Swiss Regulon and HOMER motif databases and default parameters ($P < 0.0005$ for both the reference and alternative variant; fold change >5).

Knockout using CRISPR-Cas9. To knock out *ELL2* in L363 cells, we used CRISPR-Cas9 vectors encoding two different single-guide RNAs (sgRNAs) corresponding to DNA sequences TCTGGTAAGTCTCGAGCGCCCGG (clone #6) and TGCGGGAGGAGCAGCGCTATGGG (clone #2.3). These sequences, which were designed using the CRISPR Design tool (http://www.crispr.mit.edu-tool) and target *ELL2* exon 1, were synthesized and ligated into lentiCRISPRv2 vector (AddGene, Cambridge, MA, USA; cat. no. #52961) using published protocols[61]. An aliquot of ligated mix was transformed to JM109 competent cells. sgRNA inserts were confirmed by Sanger sequencing using standard Hu6-F primer. The lentiCRISPRv2 vector containing inserts were transfected into L363 cells by electroporation and puromycin selection. Successful knockout was verified by western blot with antibodies toward ELL2 (Santa Cruz Biotechnology, cat. no. sc-376611). For this, five million cells were collected and washed with PBS. Cells were lysed using 2× Laemmli sample (100 µl) and 2-mercaptoethanol. Samples were kept on ice and sonicated on Bioruptor-pico (Diagenode) for ten cycles at 30 s/30 s on and off. Thereafter, samples were heat denatured at 96 °C for 5 min and centrifuged at full speed for 5 min. Supernatant was transferred to another vial and loaded on gel. For protein separation and blot, we used mini-protein TGX stain free gel (Bio-Rad) and trans-blot turbo transfer pack (nitrocellulose, Bio-Rad) followed by overnight incubation with ELL2 antibody (Santa Cruz Biotechnology, #37661) and development (Bio-Rad). Membranes were re-probed with GAPDH antibody after reblot treatment (Millipore, #2502).

Analysis of cell line data and gene set enrichment analysis. From wild-type and CRISPR-Cas9 *ELL2* knockout cells, we purified and sequenced mRNA using the same protocols as the primary CD138$^+$ plasma cell samples. Two replicates from wild-type cells and two replicates from each of two independent clones (clone #6 and clone #2.3) were analyzed. Differentially expressed genes were identified by comparing FPKM (fragments per kilo base of exon per million fragments mapped) values using Smyth's moderated *t*-statistic[62]. For gene set enrichment analysis, we used the RenderCat[63] tool with default parameters, Gene Ontology[64] and ABI Panther (http://panterdb.org) gene set databases, and considered genes with average FPKM >5 in the MMPC RNA-sequencing data. We also created specific gene sets comprising the ~80 genes encoding the proteins of the large and small ribosome subunits ("RPG") and 7 genes encoding other members of the super-elongation complex ("SEC"). In addition to the L363 gene expression data, we used pre-existing gene expression profiles of shEll2- vs shGFP-treated mouse MPC1 plasmocytoma cells. These data were retrieved from the NCBI Gene Expression Omnibus Omnibus (accession no. GSE40285). The MPC1 data were analyzed using the same methods as the L363 data.

Reconstitution of *ELL2* expression in L363-*ELL2*-knockout cells. To reconstitute *ELL2* expression in the L363-*ELL2*-KO cells generated using CRISPR-Cas9, we inserted *ELL2* into a Tet-ON-3G doxycycline-inducible gene expression system (Clontech). To allow the construct to escape CRISPR-Cas9 editing, we changed the sixteenth *ELL2* codon from GGG to GGC, both coding for glycine. The new codon change eliminates the PAM sequence of the sgRNA that was used to generate the L363-*ELL2*-KO cells. The coding mRNA transcript (based on NM_012081.5, 351-2273) was synthesized as gBlocks Gene Fragments from IDT. The gene fragment was cloned in pTRE3G inducible vector. The L363-*ELL2*-KO (clone #2.3) were electroporated with pTRE3G-*ELL2* and pTRE3G-EF1α (Clontech) at a ratio of 4:1 using the NEON system (Thermo-Fisher Scientific). For mock/control transfection, we used Empty pTRE3G- and pTRE3G-EF1α (Clontech). The electroporated cells were cultured with or without doxycycline (200 ng/ml) for 24 h. RNA was prepared using the RNeasy mini kit (Qiagen), quality-assessed using Nanodrop and Bioanalyzer (Agilent), and sequenced using 2 × 75-bp Illumina mRNA sequencing at the Centre for Translational Genomics facility (Lund University), yielding about 36 million paired-end reads per sample on average. Sequences were aligned to hg38 reference genome using TopHat, and expression (FPKM) values were quantified using CuffLinks[47]. Successful induction of *ELL2* expression was confirmed by western blot, and by the presence of reads containing the new glycine codon in the RNA-sequencing data in the doxycycline-treated samples. Differential gene expression was quantified using $\log_2$ ratios, and enrichment analysis was done with RenderCat[63].

Data availability. The RNA-sequencing data for wild-type and knockout L363 cells, and for L363 cells from the *ELL2* addback experiments are available via the NCBI Gene Expression Omnibus (accession nos. GSE111199 and GSE111210). The eQTL data for MM plasma cells and ChIP-seq data for L363 cells are available from the authors on a collaborative basis.

Received: 1 July 2017 Accepted: 26 March 2018
Published online: 25 April 2018

References

1. Rajkumar, S. V. et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol.* **15**, e538–e548 (2014).
2. Weiss, B. M., Abadie, J., Verma, P., Howard, R. S. & Kuehl, W. M. A monoclonal gammopathy precedes multiple myeloma in most patients. *Blood* **113**, 5418–5422 (2009).
3. Landgren, O. et al. Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* **113**, 5412–5417 (2009).
4. Kyle, R. A. et al. Prevalence of monoclonal gammopathy of undetermined significance. *N. Engl. J. Med.* **354**, 1362–1369 (2006).
5. Koura, D. T. & Langston, A. A. Inherited predisposition to multiple myeloma. *Ther. Adv. Hematol.* **4**, 291–297 (2013).
6. Morgan, G. J. et al. Inherited genetic susceptibility to multiple myeloma. *Leukemia* **28**, 518–524 (2014).
7. Frank, C. et al. Search for familial clustering of multiple myeloma with any cancer. *Leukemia* **30**, 627–632 (2016).
8. Kristinsson, S. Y. et al. Patterns of hematologic malignancies and solid tumors among 37,838 first-degree relatives of 13,896 patients with multiple myeloma in Sweden. *Int. J. Cancer* **125**, 2147–2150 (2009).
9. Landgren, O. et al. Risk of plasma cell and lymphoproliferative disorders among 14,621 first-degree relatives of 4458 patients with monoclonal gammopathy of undetermined significance in Sweden. *Blood* **114**, 791–795 (2009).
10. Altieri, A., Chen, B., Bermejo, J. L., Castro, F. & Hemminki, K. Familial risks and temporal incidence trends of multiple myeloma. *Eur. J. Cancer* **42**, 1661–1670 (2006).
11. Vachon, C. M. et al. Increased risk of monoclonal gammopathy in first-degree relatives of patients with multiple myeloma or monoclonal gammopathy of undetermined significance. *Blood* **114**, 785–790 (2009).
12. Mitchell, J. S. et al. Genome-wide association study identifies multiple susceptibility loci for multiple myeloma. *Nat. Commun.* **7**, 12050 (2016).
13. Swaminathan, B. et al. Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. *Nat. Commun.* **6**, 7213 (2015).
14. Chubb, D. et al. Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nat. Genet.* **45**, 1221–1225 (2013).
15. Broderick, P. et al. Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nat. Genet.* **44**, 58–61 (2012).
16. Halvarsson, B.-M. et al. Direct evidence for a polygenic etiology in familial multiple myeloma. *Blood Adv.* **1**, 619–623 (2017).
17. Luo, Z., Lin, C. & Shilatifard, A. The super elongation complex (SEC) family in transcriptional control. *Nat. Rev. Mol. Cell Biol.* **13**, 543–547 (2012).
18. Liu, M., Hsu, J., Chan, C., Li, Z. & Zhou, Q. The ubiquitin ligase Siah1 controls ELL2 stability and formation of super elongation complexes to modulate gene transcription. *Mol. Cell* **46**, 325–334 (2012).
19. Martincic, K., Alkan, S. A., Cheatle, A., Borghesi, L. & Milcarek, C. Transcription elongation factor ELL2 directs immunoglobulin secretion in plasma cells by stimulating altered RNA processing. *Nat. Immunol.* **10**, 1102–1109 (2009).
20. Shell, S. A., Martincic, K., Tran, J. & Milcarek, C. Increased phosphorylation of the carboxyl-terminal domain of RNA polymerase II and loading of polyadenylation and cotranscriptional factors contribute to regulation of the Ig heavy chain mRNA in plasma cells. *J. Immunol.* **179**, 7663–7673 (2007).
21. Milcarek, C., Albring, M., Langer, C. & Park, K. S. The eleven-nineteen lysine-rich leukemia gene (ELL2) influences the histone H3 protein modifications accompanying the shift to secretory immunoglobulin heavy chain mRNA production. *J. Biol. Chem.* **286**, 33795–33803 (2011).
22. Park, K. S. et al. Transcription elongation factor ELL2 drives Ig secretory-specific mRNA production and the unfolded protein response. *J. Immunol.* **193**, 4663–4674 (2014).
23. Benson, M. J. et al. Heterogeneous nuclear ribonucleoprotein L-like (hnRNPLL) and elongation factor, RNA polymerase II, 2 (ELL2) are regulators ofmRNA processing in plasma cells. *Proc. Natl Acad. Sci. USA* **109**, 16252–16257 (2012).
24. Lauc, G. et al. Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet.* **9**, e1003225 (2013).
25. Franceschini, N. et al. Discovery and fine mapping of serum protein loci through transethnic meta-analysis. *Am. J. Hum. Genet.* **91**, 744–753 (2012).
26. Boal, F. et al. TOM1 is a PI5P effector involved in the regulation of endosomal maturation. *J. Cell Sci.* **128**, 815–827 (2015).
27. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
28. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).
29. Weinhold, N. et al. 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 (2015).
30. Frezal, J. Genatlas database, genes and development defects. *C. R. Acad. Sci. III* **321**, 805–817 (1998).
31. Su, A. I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101**, 6062–6067 (2004).
32. Xu, L. et al. Genome-wide association study identifies common genetic variants associated with salivary gland carcinoma and its subtypes. *Cancer* **121**, 2367–2374 (2015).
33. Westra, H. J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).

34. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv. Biobank.* **13**, 307–308 (2015).
35. Rosenbloom, K. R. et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
36. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
37. Kurschner, C. & Morgan, J. I. USF2/FIP associates with the b-Zip transcription factor, c-Maf, via its bHLH domain and inhibits c-Maf DNA binding activity. *Biochem. Biophys. Res. Commun.* **231**, 333–339 (1997).
38. Kienast, J. & Berdel, W. E. c-maf in multiple myeloma: an oncogene enhancing tumor-stroma interactions. *Cancer Cell* **5**, 109–110 (2004).
39. Kataoka, K. Multiple mechanisms and functions of maf transcription factors in the regulation of tissue-specific genes. *J. Biochem.* **141**, 775–781 (2007).
40. Fowler, T. et al. Regulation of MYC expression and differential JQ1 sensitivity in cancer cells. *PLoS ONE* **9**, e87003 (2014).
41. Lin, C. et al. Dynamic transcriptional events in embryonic stem cells mediated by the super elongation complex (SEC). *Genes Dev.* **25**, 1486–1498 (2011).
42. Raiser, D. M., Narla, A. & Ebert, B. L. The emerging importance of ribosomal dysfunction in the pathogenesis of hematologic disorders. *Leuk. Lymphoma* **55**, 491–500 (2013).
43. Barna, M. et al. Suppression of Myc oncogenic activity by ribosomal protein haploinsufficiency. *Nature* **456**, 971–975 (2008).
44. Fumagalli, S. et al. Absence of nucleolar disruption after impairment of 40S ribosome biogenesis reveals an rpL11-translation-dependent mechanism of p53 induction. *Nat. Cell Biol.* **11**, 501–508 (2009).
45. Ajore, R. et al. Deletion of ribosomal protein genes is a common vulnerability in human cancer, especially in concert with TP53 mutations. *EMBO Mol. Med.* **9**, 498–507 (2017).
46. Kehr, B. et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588–593 (2017).
47. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
48. Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).
49. Howie, B. N., Donelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
50. Erickson, S. W. et al. Genome-wide scan identifies variant in 2q12.3 associated with risk for multiple myeloma. *Blood* **124**, 2001–2003 (2014).
51. Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
52. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
53. Styrkarsdottir, U. et al. Nonsense mutation in the *LGR4* gene is associated with several human diseases and other traits. *Nature* **497**, 517–520 (2013).
54. Benonisdottir, S. et al. Epigenetic and genetic components of height regulation. *Nat. Commun.* **7**, 13490 (2016).
55. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ (2014).
56. Andrews, S. FastQC: a quality control tool for high throughput sequence data http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).
57. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
59. Ajore, R., Dhanda, R. S., Gullberg, U. & Olsson, I. The leukemia associated ETO nuclear repressor gene is regulated by the GATA-1 transcription factor in erythroid/megakaryocytic cells. *BMC Mol. Biol.* **11**, 38 (2010).
60. Andrews, N. C. & Faller, D. V. A rapid micropreparation technique for extraction of DNA-binding proteins from limiting numbers of mammalian cells. *Nucleic Acids Res.* **19**, 2499 (1991).
61. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
62. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 1–25 (2004).
63. Nilsson, B., Hakansson, P., Johansson, M., Nelander, S. & Fioretos, T. Threshold-free high-power methods for the ontological analysis of genome-wide gene-expression studies. *Genome Biol.* **8**, R74 (2007).
64. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

Acknowledgements

This work was supported by the Swedish Foundation for Strategic Research (KF10-0009), the Knut and Alice Wallenberg Foundation (2012.0193), the Swedish Research Council (2012-1753), Cancerfonden (2017/265), ALF grants from Region Skåne, the Medical Faculty at Lund University, the Swedish Society of Medicine, the Crafoord Foundation, the Arne and Ingabritt Lundberg Foundation (2017-0055), the Borås Cancer Foundation, Deutsche Krebshilfe, the Multiple Myeloma Research Foundation, and the German Ministry of Education and Science (Cliommics, 01ZX1309B). We thank S. Jónsson, Ó. Magnússon, G. Halldórsson, D. Gudbjartsson, I. Jonsdóttir, U. Thorsteinsdóttir, and K. Stefánsson for their kind assistance with mRNA sequencing, providing access to gene expression data for peripheral blood from Icelanders, and insightful comments on the manuscript. We thank R. Houlston for providing access to gene expression data for plasma cells from patients from the United Kingdom. We thank A. Collin and M. Soller and the Swedish National Myeloma Biobank for their assistance with the sample collection. We are indebted to the patients who participated in the project.

Author contributions

M.A., A-K.W., M.H., U.G. and B.N. designed research. M.A., A-K.W., A.N., N.W. and B. N. analyzed data. I.T., M.H., H.N., A.W., O.W.S., G.M., H.G., K.H., U.-H.M. and T.M. contributed samples or data. R.A., B.S. and E.J. carried out experiments. M.A., A-K.W., N.W. and B.N. drafted the manuscript. All authors contributed to the final manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04082-2.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper IV

Correspondence: Sequence variation at the *MTHFD1L-AKAP12* and *FOPNL* loci does not influence multiple myeloma survival in Sweden

Authors

Mina Ali[1], Konstantinos Lemonakis[1,2], Anna-Karin Wihlborg[1], Ljupco Veskovski[3], Ingemar Turesson[1], Ulf-Henrik Mellqvist[3] , Urban Gullberg[1], Markus Hansson[1,2,4,*] and Björn Nilsson[1,5,*]

Affiliations

[1]Hematology and Transfusion Medicine, Department of Laboratory Medicine, BMC B13, 221 84 Lund, Sweden. [2]Hematology Clinic, Skåne University Hospital, 221 85 Lund, Sweden. [3]Section of Hematology, South Elvsborg Hospital, SE 501 83 Borås, Sweden. [4]Wallenberg Center for Molecular Medicine, 221 84 Lund, Sweden. [5]Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA. [*]These authors jointly directed the project.

Correspondence: Björn Nilsson, M.D. Ph.D.**,** Hematology and Transfusion Medicine, Department of Laboratory Medicine, BMC B13, 221 84 Lund, Sweden; e-mail: bjorn.nilsson@med.lu.se**,** tel +46-46-2220738, fax +46-46-130064.

Keywords: multiple myeloma, inherited susceptibility, survival.

No. of words: 972 in main text, 99 in abstract.

No. of display items: 1 table, 1 figure.

Abstract

Multiple myeloma (MM) is the second most common blood malignancy. Recently, two meta-analyses reported associations between MM overall survival and inborn sequence variants at the *MTHFD1L-AKAP12* and *FOPNL* loci, respectively. Here we looked for further support of these associations in a series of 871 patients with MM from Sweden, but could not detect any evidence for association with survival for either of the two loci. Our results could potentially be explained by differences in treatment or other patient characteristics, and motivate the collection of larger data sets to understand the effects of genetic variation on clinical outcome in MM.

Multiple myeloma (MM) is the second most common hematologic malignancy. The disease is defined by an uninhibited, clonal growth of plasma cells in the bone marrow[1]. It is preceded by monoclonal gammopathy of unknown significance (MGUS)[2], a common condition defined as a clonal growth of plasma cells that does not yet satisfy the criteria for MM, but progresses to MM at a rate of approximately 1% per year[3].

Increasing evidence supports that the biology of MM is influenced by inborn genetic variation. MM and MGUS show familial clustering, and genome-wide association studies have identified DNA sequence variants that influence MM risk[4-8]. Additionally, two recent studies indicate that genetic variation could also influence MM survival[9,10].

In the first of these, Johnson *et al.*[9] describe an association between overall survival in multiple myeloma (MM-OS) and rs12374648, located between the *MTHFD1L* and *AKAP12* genes at chromosome 6q25.1[9]. The protein encoded by *MTHFD1L* is involved in folate metabolism[11], and *AKAP12* is related to cell growth[12]. The association with MM-OS was detected in a meta-analysis of 3,256 cases from four clinical trials: two from the UK, one from the USA, and one from Germany (combined *P*-value=4.69×10^{-9}, hazard ratio, HR=1.34, 95% CI 1.22-1.48). The association was detected in the sample sets from the UK and USA (*P*=1.69×10^{-6} to 0.009; HR 1.06 to 1.75), but not in the one from Germany (*P*=0.55; HR=1.09, 95% CI 0.82-1.44).

In a second study, Ziv *et al.*[10] describe an association between MM-OS and rs72773978 near *FOPNL* at 16p13. The protein encoded by *FOPNL* has been implicated in centrosome function[13]. The association was detected by meta-analysis of 545 cases from two clinical trials in USA (*P*=6×10^{-10}; HR=2.65, 95% CI 1.94-3.58). The association was replicated in a meta-analysis of two other data sets (IMMEnSE, consisting six sample subsets totaling n=772 and one from Utah, n=315) (combined *P*=0.044; HR=1.34, 95% CI 1.01-1.78). Yet, the positive replication result was driven by a *P*-value of 0.004 with large effect size (HR=9.73) in a subset of 109 patients from Spain in IMMEnSE, whereas the other six subsets (Italy, Poland, Portugal, Denmark, Edmonton in IMMEnSE and the one from Utah) did not show any evidence of association (Supplementary Table 7 in ref.[10]).

Given these reports, we looked for further support of the *MTHFDL1-AKAP12* and *FOPLN* loci in a Swedish study population. We retrieved clinical data for 871 patients diagnosed with MM between 2005 to 2015 from the Swedish Multiple Myeloma Registry (Sahlgrenska Hospital, Gothenburg) (**Table 1**), which records clinical data on MM patients in Sweden and has about 90% inclusion rate compared to the Swedish Cancer Registry. The patients had been previously genotyped in genome-wide association studies using population-based samples from the Swedish National Myeloma Biobank (Skåne University Hospital, Lund)[6,7]. The clinical data and samples were obtained subject to informed consent and ethical approval (Lund University, dnr 2013/540), and in accordance with the principles of the Declaration of Helsinki. The samples were genotyped using Illumina microarrays and imputed with phased reference haplotypes from 1,000 Genomes[6,14]. To test for association between genotypes and MM-OS, we used a Cox proportional hazards model implemented in R (v.2.8) with adjustment for age, sex and International Staging System (ISS) score. Survival was calculated from the date treatment started until the date of death, or until April 5th 2016 (median follow-up time 39.5 months).

In our analysis, we did not see any evidence of association with MM-OS for either rs12374648 (P=0.7; HR=0.97, 95% CI=0.81-1.2) or rs72773978 (P=0.93; HR=0.98, 95% CI=0.7-1.4) (**Fig. 1**). For completeness, we also tested for associations between MM-OS and all variants with minor allele frequency (MAF) > 5% located within 1 Mb of *MTHFD1L-AKAP12* (6,515 variants) or *FOPNL* (3,892 variants), but could not identify any association with any of these variants (smallest *P*-value= 1.02×10^{-4}, Bonferroni threshold =7.7×10^{-6} for *MTHFD1L-AKAP12* and smallest *P*-value= 1.54×10^{-4}, Bonferroni threshold= 1.28× 10^{-5} for *FOPNL*). Thus, we could not replicate the associations between MM-OS and *MTHFD1L-AKAP12* and *FOPNL* in a population-based series, nor identify any other alleles associations with MM-OS at these loci.

We considered possible reasons why the reported associations did not replicate in our sample set. Firstly, our study is comparable in size (n=871) to the largest of the reported individual sample sets, including UK-My9 (n=1,163) and UK-My11 (n=871) where rs12374648 at *MTHFD1L* was detected, and rs72773978 at *FOPNL* was found in smaller data sets. Secondly, the absence of associations between MM-OS and rs12374648 and rs72773978 in our data is probably not due to differences in geographic origin. The two reported variants are common, both in our data (MAF 21.5% and 4.7%) and in the different populations of 1,000 Genomes[14], meaning the reported associations are unlikely to be population-specific. Thirdly, however, the absence of replication signals in our data could be explained by differences in clinical characteristics between the study populations. Here, one notable difference is that our material is population-based, whereas the studies by Johnsson *et al.*[9] and Ziv *et al.*[10] are based on patients recruited into clinical trials. As a result, our population is older (average 68 years *vs* 54 to 66 years), and has not been selected for patients without comorbidity, as is common in clinical trials. A higher incidence of comorbidity could dilute effects of DNA sequence variation on survival. Moreover, differences in age and comorbidity will carry differences in treatment. For example, some of the reported populations contain a high proportion of patients who received autologous stem cell transplantation (ASCT; 100% in the German and US sample sets in Johnson *et al.*[9]), whereas our study population contains 32.5% transplanted patients. Accordingly, it is possible that the reported effects on MM-OS are connected to a certain treatment, for example ASCT. Together, our results and refs.[9,10] motivate the collection of larger data sets to understand the impact of genetic variation on clinical outcome in MM.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

M.A., K.L., A-K.W., U.G., M.H. and B.N. designed research and analysed data. L.V., I.T., M.H., and U-H.M. contributed to the clinical data collection. M.A., M.H., and B.N. drafted the manuscript. All authors contributed to the final manuscript. We thank Anna Collin, Maria Soller, and Cecilie Blimark for their assistance. We are indebted to the patients who participated in the study.

Figure legends

Figure 1: Kaplan-Meier plots for **(a)** rs12374648 at *MTHFD1L-AKAP12* and **(b)** rs72773978 at *FOPNL*. No difference in survival between the genotype groups was observed.

References

1 Rajkumar, S. V. *et al.* International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol* **15**, e538-e548, doi:10.1016/S1470-2045(14)70442-5 (2014).

2 Weiss, B. M., Abadie, J., Verma, P., Howard, R. S. & Kuehl, W. M. A monoclonal gammopathy precedes multiple myeloma in most patients. *Blood* **113**, 5418-5422, doi:10.1182/blood-2008-12-195008 (2009).

3 Kyle, R. A. *et al.* Prevalence of Monoclonal Gammopathy of Undetermined Significance. *N Engl J Med* **354**, 1362-1369 (2006).

4 Broderick, P. *et al.* Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nat Genet* **44**, 58-61, doi:10.1038/ng.993 (2012).

5 Chubb, D. *et al.* 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 (2013).

6 Swaminathan, B. *et al.* Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. *Nat Commun* **6**, 7213, doi:10.1038/ncomms8213 (2015).

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

8 Weinhold, N. *et al.* The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nat Genet* **45**, 522-525, doi:10.1038/ng.2583 (2013).

9 Johnson, D. C. *et al.* Genome-wide association study identifies variation at 6q25.1 associated with survival in multiple myeloma. *Nat Commun* **7**, 10290, doi:10.1038/ncomms10290 (2016).

10 Ziv, E. *et al.* Genome-wide association study identifies variants at 16p13 associated with survival in multiple myeloma patients. *Nat Commun* **6**, 7539, doi:10.1038/ncomms8539 (2015).

11 Christensen, K. E. & Mackenzie, R. E. Mitochondrial methylenetetrahydrofolate dehydrogenase, methenyltetrahydrofolate cyclohydrolase, and formyltetrahydrofolate synthetases. *Vitam Horm* **79**, 393-410, doi:10.1016/S0083-6729(08)00414-7 (2008).

12 Su, B., Bu, Y., Engelberg, D. & Gelman, I. H. SSeCKS/Gravin/AKAP12 inhibits cancer cell invasiveness and chemotaxis by suppressing a protein kinase C-Raf/MEK/ERK pathway. *J Biol Chem* **285**, 4578-4586, doi:10.1074/jbc.M109.073494 (2010).

13 Sedjai, F. *et al.* Control of ciliogenesis by FOR20, a novel centrosome and pericentriolar satellite protein. *J Cell Sci* **123**, 2391-2401, doi:10.1242/jcs.065045 (2010).

14 The Genomes Project, C. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

Figure 1

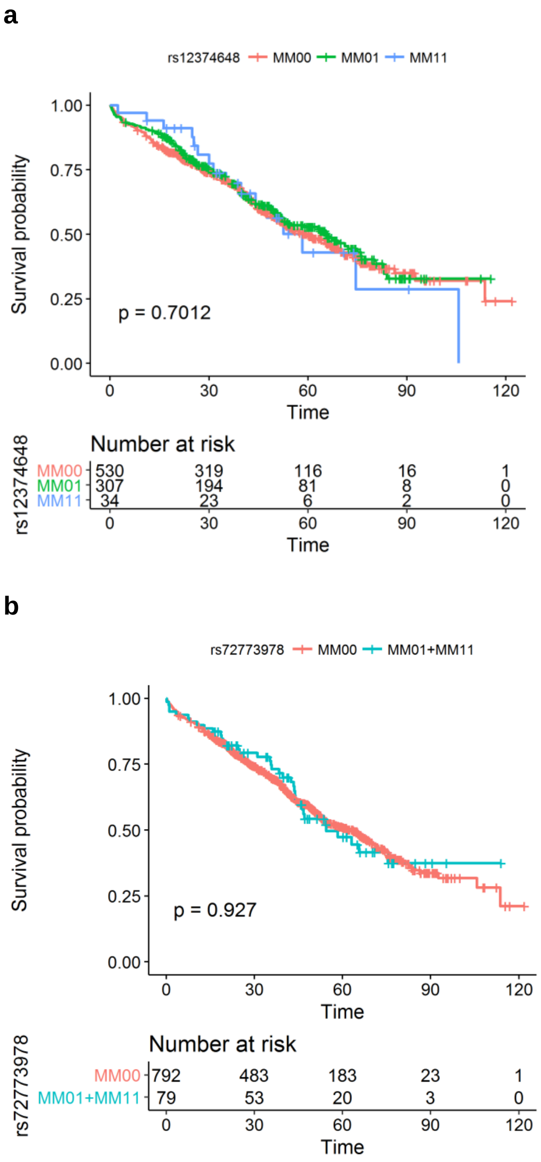
a
rs12374648 MM00 MM01 MM11
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.7012
0
30
60
90
120
Time
Number at risk
rs12374648
MM00 530 319 116 16 1
MM01 307 194 81 8 0
MM11 34 23 6 2 0
0
30
60
90
120
Time
b
rs72773978 MM00 MM01+MM11
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.927
0
30
60
90
120
Time
Number at risk
rs72773978
MM00 792 483 183 23 1
MM01+MM11 79 53 20 3 0
0
30
60
90
120
Time

Table 1: Clinical characteristics of the study population

Number of Cases	871
Gender	
Male	531
Female	340
Median age at diagnosis	68
Median follow-up (months)	39.48
Deceased during follow-up	
Yes	393
No	478
ISS	
I	179
II	339
III	234
Unknown	119
Heavy chain paraprotein	
IgA	191
IgG	536
IgD	6
IgM	6
Not detected	132
Light chain paraprotein	
Lambda	240
Kappa	446
Not detected or not done	185
Median plasma cells in bone marrow (%)	22
Treatment received	
Proteasome inhibitor	427 (49.02%)
Immunomodulatory (IMiD)	228 (26.18%)
Chemotherapy	678 (77.84%)
Autologous stem cell transplantation (ASCT)	283 (32.49%)
Other or no treatment	112 (12.86%)
Anemia (%)	26.18
Hypercalcemia (%)	8.04
Renal failure (%)	13.6

LUND UNIVERSITY

FACULTY OF MEDICINE

Division of Hematology and
Transfusion Medicine
Department of Laboratory Medicine
Lund University, Faculty of Medicine
Doctoral Dissertation Series 2018:84
ISBN 978-91-7619-650-2
ISSN 1652-8220