MINA ALI

Identification and characterization of DNA sequence variants associated with multiple myeloma

MINA ALI

DEPARTMENT OF LABORATORY MEDICINE | LUND UNIVERSITY


Printed by Media-Tryck, Lund 2018

NORDIC SWAN ECOLABEL 3041 0903
Identification and characterization of DNA sequence variants associated with multiple myeloma
Identification and characterization of DNA sequence variants associated with multiple myeloma

Mina Ali

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
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

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

Date 2018-05-21
Identification and characterization of DNA sequence variants associated with multiple myeloma

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

Copyright Mina Ali

Lund University, Faculty of Medicine, Doctoral Dissertation Series 2018:84
Department of Laboratory Medicine, Division of Hematology and Transfusion Medicine
ISSN  1652-8220

Printed in Sweden by Media-Tryck, Lund University
Lund 2018
Dedicated to my parents
“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 ........................................................................................................................................... 13
List of papers .................................................................................................................................. 15
Abbreviations .................................................................................................................................. 16

### Chapter 1 Introduction and conceptual background ................................................................. 17

### Chapter 2 Identification of genetic variants through GWAS ..................................................... 19
   2.1 How to conduct a GWAS? ........................................................................................................ 19
      2.1.1 Quality Control ................................................................................................................... 20
      2.1.2 Imputation ........................................................................................................................... 20
      2.1.3 Association testing ................................................................................................................. 22
      2.1.4 Replication and meta-analysis ............................................................................................... 22
   2.2 What GWAS can do and what it cannot do? ............................................................................. 23

### Chapter 3 Characterization of variants discovered by GWAS .................................................... 25
   3.1 Fine mapping of associated loci ............................................................................................... 25
   3.2 Analysis of expression quantitative trait loci .......................................................................... 26
   3.3 Analysis of regulatory variants ................................................................................................. 26

### Chapter 4 Original work ............................................................................................................. 29
   4.1 Multiple Myeloma ...................................................................................................................... 30
   4.2 Paper I ....................................................................................................................................... 30
   4.3 Paper II ....................................................................................................................................... 33
   4.4 Paper III ..................................................................................................................................... 34
   4.5 Paper IV ..................................................................................................................................... 35

### Chapter 5 Concluding remarks .................................................................................................. 37

### Summary in English .................................................................................................................... 39

### Acknowledgements .................................................................................................................... 41

### References ................................................................................................................................... 43
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.


*Nature Communications*, (2015). 6, 7213

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


*Nature Communications*, (2016). 7, 12050

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


*Nature Communications*, (2018). 9, 1649

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


*Submitted manuscript*, (2018).

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

3C  Chromosome conformation capture
ANOV A  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. 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. 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. 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. 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. The International HapMap Project is another scientific effort to identify common genetic variations among people. 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 led 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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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) \textsuperscript{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\textsuperscript{23}. Principal components analysis (PCA) is widely used to correcting for cryptic differences in population structure between cases and controls (\textit{e.g.}, differences in geographic origin)\textsuperscript{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\textsuperscript{25}. Imputation makes use of the fact that the human genome is arranged in blocks of high LD, separated by hotspots of recombination\textsuperscript{26}. LD is the non-random association between alleles at different loci\textsuperscript{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\textsuperscript{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\textsuperscript{29}.

Imputation makes use of LD to interpolate sequence variants that detected in recent whole-genome sequencing efforts (\textit{e.g.}, 1,000 Genomes\textsuperscript{3} or the UK10K project\textsuperscript{30}) into the SNP microarray data. For imputation to be effective, the reference haplotypes (\textit{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\textsuperscript{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\textsuperscript{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\textsuperscript{35,36}. For quantitative traits, generalized linear model (GLM) and the analysis of variance (ANOVA) can be used\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{35,37,38}. Bonferroni correction is a commonly used solution\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{49} and the Encyclopedia of DNA Elements (ENCODE)\textsuperscript{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\textsuperscript{46,47,51}. Identification of causal variants at risk loci is an active area of research\textsuperscript{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\textsuperscript{3} and UK10K\textsuperscript{30}, can fill the gaps for variants that are not existed in the initial genotyping platform\textsuperscript{47}, but are limited by allele frequency, and by the quality and geographic representativeness of the reference haplotypes\textsuperscript{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\textsuperscript{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. 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).

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. Statistical methods are then used to associate genotypes with gene expression. By now comprehensive sets of eQTL data have been made publicly available in online databases, including Genotype-Tissue Expression (GTEx), Blood eQTL, GEUVADIS project, MuTHER studies and SNP and CNV Annotation (SCAN).

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 DNase hypersensitivity or ATAC-seq), clustering of transcription factor binding sites and specific histone modifications (detected by chromatin immunoprecipitation). Consortia such as the ENCODE and the Roadmap Epigenomics Mapping Consortium (REMC) 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. 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\textsuperscript{66}.

To identify genomic regions with open chromatin, hence more likely regulatory, DNA hypersensitivity or ATAC-seq can be used\textsuperscript{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\textsuperscript{69}. Adapter-ligated DNA fragments are then isolated, amplified by PCR and used for High-throughput sequencing\textsuperscript{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\textsuperscript{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)\textsuperscript{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”)\(^7\). 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)\(^7\), which is a common pre-malignant condition (3% of ≥50 year olds)\(^8\) 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%\(^8,81\). 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)\(^9\) 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\(^4\). 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 \(P < 5 \times 10^{-8}\) (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

<table>
<thead>
<tr>
<th>Locus</th>
<th>Variant</th>
<th>A1/A2*</th>
<th>RAF</th>
<th>MM+MGUS</th>
<th>MM Only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td>OR</td>
<td>P</td>
</tr>
<tr>
<td>3p22</td>
<td>rs73071352</td>
<td>A/G</td>
<td>0.13</td>
<td>3.1×10⁻⁸</td>
<td>1.32</td>
</tr>
<tr>
<td>5q15</td>
<td>rs56219066</td>
<td>C/T</td>
<td>0.72</td>
<td>1.4×10⁻⁸</td>
<td>1.23</td>
</tr>
<tr>
<td>5q31</td>
<td>rs74735889</td>
<td>C/T</td>
<td>0.003</td>
<td>5.9×10⁻¹⁰</td>
<td>4.06</td>
</tr>
<tr>
<td>6p21</td>
<td>rs6919908</td>
<td>T/C</td>
<td>0.23</td>
<td>6.3×10⁻¹⁰</td>
<td>1.19</td>
</tr>
<tr>
<td>7p15</td>
<td>rs57104699</td>
<td>A/C</td>
<td>0.66</td>
<td>2.3×10⁻⁸</td>
<td>1.38</td>
</tr>
<tr>
<td>17p11</td>
<td>rs57968458</td>
<td>A/G</td>
<td>0.098</td>
<td>2.8×10⁻¹⁰</td>
<td>1.26</td>
</tr>
<tr>
<td>22q13</td>
<td>rs138740</td>
<td>C/T</td>
<td>0.38</td>
<td>1.3×10⁻⁷</td>
<td>1.19</td>
</tr>
</tbody>
</table>

*: 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 and UK10K 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 autophagy102-104.

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

<table>
<thead>
<tr>
<th>Locus</th>
<th>Variant</th>
<th>RAFF</th>
<th>Odd ratio</th>
<th>P-value</th>
<th>Candidate Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>6p22.3</td>
<td>rs34229995</td>
<td>C/G</td>
<td>0.029</td>
<td>1.37</td>
<td>JARID2 (5' telomer)</td>
</tr>
<tr>
<td>6q21</td>
<td>rs9372120</td>
<td>G/T</td>
<td>0.218</td>
<td>1.18</td>
<td>ATG5</td>
</tr>
<tr>
<td>7q36.1</td>
<td>rs7781265</td>
<td>T/C</td>
<td>0.125</td>
<td>9.09 x 10^-15</td>
<td>SMARCD3</td>
</tr>
<tr>
<td>8q24.21</td>
<td>rs1948915</td>
<td>T/C</td>
<td>0.345</td>
<td>9.71 x 10^-9</td>
<td>CCAT1</td>
</tr>
<tr>
<td>9p21.3</td>
<td>rs2811710</td>
<td>G/A</td>
<td>0.657</td>
<td>4.20 x 10^-11</td>
<td>CDKN2A</td>
</tr>
<tr>
<td>10p12.1</td>
<td>rs2790457</td>
<td>A/G</td>
<td>0.739</td>
<td>1.72 x 10^-13</td>
<td>WAC</td>
</tr>
<tr>
<td>16q23.1</td>
<td>rs7193541</td>
<td>C/T</td>
<td>0.585</td>
<td>5.00 x 10^-12</td>
<td>RFWD3</td>
</tr>
<tr>
<td>20q13.13</td>
<td>rs6066835</td>
<td>G/T</td>
<td>0.083</td>
<td>1.36 x 10^-13</td>
<td>PREX1</td>
</tr>
</tbody>
</table>

*: 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 databases64,105 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 (rs7777189-C, rs7777185-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.\textsuperscript{106} and Ziv et al.\textsuperscript{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 MTHFD1L-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\textsuperscript{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 \textit{ELL2} MM risk allele alters the expression of \textit{ELL2} in \textit{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 \textit{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\textsuperscript{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.
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.

doustoon darom o parazhite rin darabi zandegim hestid.
References


Paper I
Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma

Bhairavi Swaminathan1,*, Guðmar Thorleifsson2,*, Magnus Jóud1,3,*, Mina Ali1,*, Ellinor Johnsson1, Ram Ajore1, Patrick Sulem2, Britt-Marie Halvarsson1, Guðmundur Eyjolfsson4, Vilhelmina Haraldsdottir5, Christina Hultman6, Erik Ingelsson7, Sigurður Y. Kristinsson8, Anna K. Kählér8, Stig Lenhoff9, Gisli Masson2, Ulf-Henrik Mellqvist10, Robert Månsson11, Sven Nelander12, Isleifur Olafsson13, Olof Sigurðardottir14, Hlif Steingrimsdóttir5, Annette Vangsted15, Ulla Vogel16, Anders Waage17, Hareth Nahi18, Daniel F. Gudbjartsson2, Thorunn Rafnar2, Ingemar Turesson9, Urban Guilberg1, Kári Stefánsson2,*, Markus Hansson1,9,*, Unnur Thorsteinsdóttir2,*, Open & Björn Nilsson1,3,18,**

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 x 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 x 10^{-9} and P = 6.4 x 10^{-3}, respectively) and, potentially, with an increased risk of bacterial meningitis (OR = 1.30; P = 0.0024).
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. According to the International Myeloma Working Group criteria, MM is defined by >10% monoclonal plasma cells in the bone marrow or >3 gM 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. It is preceded by monoclonal gammopathy of unknown significance (MGUS), a common condition (3% of ≥50 year olds) 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. 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. 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. 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 5q15 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 B cell-lineage gene 101/2. 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 OR = 0.014) and reached genome-wide significance under Bonferroni correction when the discovery and replication sets were combined (OR = 0.96 × 10^-10, Table 2). While rs138740 did not replicate, it remained borderline significant (meta-analysis OR = 5.7 × 10^-8) when the discovery and replication sets were combined and we observed effects in the same direction in a larger data set (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, which mediates rapid gene induction by suppressing transient pausing of RNA polymerase II activity along the DNA. Strikingly, ELL2 and the SEC play an important role in the differentiation of mature B cells into plasma cells. 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 into secreted Ig at a higher rate. 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. Silencing of ELL2 in mouse plasmacytoma cell lines using RNA interference decreases the ratio of secreted versus membrane-encoding 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 into secreted Ig at a high rate. 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. Silencing of ELL2 in mouse plasmacytoma cell lines using RNA interference decreases the ratio of secreted versus membrane-encoding Ig.

To characterize the ELL2 risk allele, we analysed SNP and gene expression profiles from peripheral blood (eight data sets...
Table 1 | Study populations.

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

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

Figures:

**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 12.3 million SNPs that passed quality filtering. Seven loci showed association with MM or MM + MGUS at meta-analysis \( P < 5 \times 10^{-8} \), 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 –log10 of the combined \( P \) values. The dotted line indicates the threshold for genome-wide significance of meta-analysis \( P < 5 \times 10^{-8} \). 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, hematopoietic stem cells; MEGA, megakaryocytes; MEP, megakaryocyte-erythroid progenitors; MONO, monocytes; NEU, neutrophils; NK, natural killer cells; PC, CD138+/plasma cells; PRE-B, pre-B cells; TCELL, T cells.
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 ($D^2 = 1.00$) with the sentinel SNP rs56219066 in intron 4 (Fig. 2a).

The ELL2 risk allele associates with decreased IgA and IgG. Because of the recent mouse studies implicating ELL2 in the production of secreted Ig5, we tested for associations with blood

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

<table>
<thead>
<tr>
<th>Populations</th>
<th>EAF MM MM + MGUS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ELL2—rs56219066-T</strong></td>
<td></td>
</tr>
<tr>
<td>Discovery</td>
<td></td>
</tr>
<tr>
<td>Sweden/Norway</td>
<td>0.732</td>
</tr>
<tr>
<td>Iceland</td>
<td>0.711</td>
</tr>
<tr>
<td>Combined</td>
<td>1.23 (1.14–1.33)</td>
</tr>
<tr>
<td>Replication</td>
<td></td>
</tr>
<tr>
<td>Denmark (363/826)</td>
<td>0.732</td>
</tr>
<tr>
<td>Sweden (223/1285)</td>
<td>0.735</td>
</tr>
<tr>
<td>Combined</td>
<td>1.29 (1.08–1.54)</td>
</tr>
<tr>
<td>Combined discovery and replication</td>
<td>1.25 (1.16–1.34)</td>
</tr>
<tr>
<td><strong>TOM1—rs138740-C</strong></td>
<td></td>
</tr>
<tr>
<td>Discovery</td>
<td></td>
</tr>
<tr>
<td>Sweden/Norway</td>
<td>0.364</td>
</tr>
<tr>
<td>Iceland</td>
<td>0.415</td>
</tr>
<tr>
<td>Combined</td>
<td>1.22 (1.13–1.30)</td>
</tr>
<tr>
<td>Replication</td>
<td></td>
</tr>
<tr>
<td>Denmark (352/815)</td>
<td>0.360</td>
</tr>
<tr>
<td>Sweden (235/1285)</td>
<td>0.364</td>
</tr>
<tr>
<td>Combined</td>
<td>1.06 (0.93–1.21)</td>
</tr>
<tr>
<td>Combined discovery and replication</td>
<td>1.18 (1.11–1.25)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Populations</th>
<th>EAF MM MM + MGUS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ELL2—rs56219066-T</strong></td>
<td></td>
</tr>
<tr>
<td>Discovery</td>
<td></td>
</tr>
<tr>
<td>Sweden/Norway</td>
<td>0.732</td>
</tr>
<tr>
<td>Iceland</td>
<td>0.711</td>
</tr>
<tr>
<td>Combined</td>
<td>1.23 (1.14–1.33)</td>
</tr>
<tr>
<td>Replication</td>
<td></td>
</tr>
<tr>
<td>Denmark (363/826)</td>
<td>0.732</td>
</tr>
<tr>
<td>Sweden (223/1285)</td>
<td>0.735</td>
</tr>
<tr>
<td>Combined</td>
<td>1.29 (1.08–1.54)</td>
</tr>
<tr>
<td>Combined discovery and replication</td>
<td>1.25 (1.16–1.34)</td>
</tr>
<tr>
<td><strong>TOM1—rs138740-C</strong></td>
<td></td>
</tr>
<tr>
<td>Discovery</td>
<td></td>
</tr>
<tr>
<td>Sweden/Norway</td>
<td>0.364</td>
</tr>
<tr>
<td>Iceland</td>
<td>0.415</td>
</tr>
<tr>
<td>Combined</td>
<td>1.22 (1.13–1.30)</td>
</tr>
<tr>
<td>Replication</td>
<td></td>
</tr>
<tr>
<td>Denmark (352/815)</td>
<td>0.360</td>
</tr>
<tr>
<td>Sweden (235/1285)</td>
<td>0.364</td>
</tr>
<tr>
<td>Combined</td>
<td>1.06 (0.93–1.21)</td>
</tr>
<tr>
<td>Combined discovery and replication</td>
<td>1.18 (1.11–1.25)</td>
</tr>
</tbody>
</table>

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.
The supporting that the observe any association with IgM. Taken together, As shown, all SNPs in the PIgG and IgM levels using the Icelandic Ig data set. We overlaid the log-linear regression in addition to the sentinel SNP rs56219066, the association between 5q15 identified in the MM and MM MGUS is identical to the ELL2 haplotype that predisposes for MM and MM haplotype associating with MM and MM levels in 20,413–24,279 Icelanders without MM or MGUS. Riskallele carriers showed lower IgA (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 −log10 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 × 10−8) and IgG levels (log-linear regression P = 6.4 × 10−5; 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, 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; logistic regression P = 0.0014; β = −0.035) as previously reported for rs3777200 in ELL2 (D2/r2 = 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 hypomorph 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 meningitidis and other meningitis pathogens. Future studies will uncover the role of ELL2 in haematological malignancies and immune response.

Methods

Population stratification, we included five principal components of the identity-by-descent (https://www.1000genomes.org) and SNPTEST version 3.8 (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 schizophrenia (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 Iceland 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 who 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 Reykjavík. 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 Caucasian (CC) and GO44 for 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, country 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 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 $\chi^2$ 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 METAL2 (http://www.sph.umich.edu/csg/abecasis//meta) with a fixed effect model. Conditional association analysis with respect to rs56219066 and rs38138740 using SNPTEST38 did not reveal any other underlying, independent signals.

Gene expression analysis in haematopoietic cell types. To characterise gene expression patterns of ELL2-expressing (ELL2+/+) and ELL2 non-expressing (ELL2−/−) cells, we used gene expression microarrays. We identified gene expression patterns of ELL2-expressing MM and MGUS tissues and compared them with those of MM and MGUS tissues from patients with MM (n = 1,285; accession no. GSE16965 (ref. 44)) and MGUS, patients with unmonitored MM and healthy bone marrow donors (n = 78; accession no. GSE9900 (ref. 47)). All data were generated on Affymetrix U133A and A2v 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 (Tagma custom assay AHCTD16), ELL2 rs8157588 (Tagma assay C_22272652_30) and TOM1 rs158726 (ref. 43); plasma cells from patients with MM (n = 1,285); accession nos. GSE27839 (ref. 45); plasma cells from patients with MGUS, patients with unmonitored MM and healthy bone marrow donors (n = 78; accession no. GSE9900 (ref. 47)). All data were generated on Affymetrix U133A and A2v microarrays and quantile-normalized to a log-normal distribution.

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 population49 and 8,086 individuals of other European populations50 and lymphoblastoid cell lines (two data sets totalling 1,188 samples33,34). Gene expression in the Icelandic data set was quantified as the mean log10 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 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 and 8,086 individuals of other European populations and lymphoblastoid cell lines (two data sets totalling 1,188 samples33,34). Gene expression in the Icelandic data set was quantified as the mean log10 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 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 and 8,086 individuals of other European populations and lymphoblastoid cell lines (two data sets totalling 1,188 samples33,34). Gene expression in the Icelandic data set was quantified as the mean log10 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 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 and 8,086 individuals of other European populations and lymphoblastoid cell lines (two data sets totalling 1,188 samples33,34). Gene expression in the Icelandic data set was quantified as the mean log10 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 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 and 8,086 individuals of other European populations and lymphoblastoid cell lines (two data sets totalling 1,188 samples33,34). Gene expression in the Icelandic data set was quantified as the mean log10 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.
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, lgs, irn, vit. d, and more) were obtained from the Landsdals University Hospital, Reykjavik, and the Icelandic Medical Center Laboratory in Mödöd (Laeknasetrid), Reykjavik, in addition to more specific hormone measures (adrenal, thyroid and sex hormones). The number of independent and uncorrelated secondary traits tested for association with rs56219066 amounts to 400.

References


Acknowledgements

This work was supported by research grants from the Swedish Foundation for Strategic Research (EF10-0009), 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 Adolsson, Tomas Axelsson, Anna Collin, Iláhos 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

All authors contributed to data interpretation and the final manuscript.

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


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).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit lines; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
Genome-wide association study identifies multiple susceptibility loci for multiple myeloma

Jonathan S. Mitchell1,*, Ni Li1,*, Niels Weinhold2,3,*, Asta Förstì4,5,*, Mina Alì6,*, Mark van Duin7,*, Gudmar Thorleifsson8, David C. Johnson9, Bowang Chen4, Britt-Marie Halvarsson6, Daniel F. Gudbjartsson9,10, Rowan Kuiper7, Owen W. Stephens2, Uta Bertsch3,11, Peter Broderick1, Chiara Campo4, Hermann Einsele12, Walter A. Gregory13, Urban Gullberg6, Marc Henrion1, Jens Hillengass3, Per Hoffmann14,15, Graham H. Jackson16, Ellinor Johnsson6, Magnus Jóud6,17, Sigurður Y. Kristinsson18, Stig Lenhoff19, Oleg Lenive1, Ulf-Henrik Mellqvist20, Gabriele Migliorini1, Hareth Nah21, Sven Neland22, Jolanta Nickel3, Markus M. Nöthen14,13, Thorunn Rafnar8, Fiona M. Ross24, Miguel Inacio da Silva Filho4, Bhaairavi Swaminathan5, Hauke Thomsen6, Ingemar Turesson19, Annette Vangsted25, Ulla Vogel26, Anders Waage27, Brian A. Walker2, Anna-Karin Wihlborg6, Annemiek Broyl7, Walter A. Gregory13, Urban Gullberg6, Marc Henrion1, Jens Hillengass3, Per Hoffmann14,15, Graham H. Jackson16, Ellinor Johnsson6, Magnus Jóud6,17, Sigurður Y. Kristinsson18, Stig Lenhoff19, Oleg Lenive1, Ulf-Henrik Mellqvist20, Gabriele Migliorini1, Hareth Nah21, Sven Neland22, Jolanta Nickel3, Markus M. Nöthen14,13, Thorunn Rafnar8, Fiona M. Ross24, Miguel Inacio da Silva Filho4, Bhaairavi Swaminathan5, Hauke Thomsen6, Ingemar Turesson19, Annette Vangsted25, Ulla Vogel26, Anders Waage27, Brian A. Walker2, Anna-Karin Wihlborg6, Annemiek Broyl7, Faith E. Davies2, Unnur Thorsteinsdottir8,28, Christian Langer29, Markus Hansson6,19, Martin Kaiser9, Pieter Sonneveld2, Kari Stefansson8,**, Gareth J. Morgan2,**, Hartmut Goldschmidt3,11,12, Kari Hemminki4,5,**, Björn Nilsson6,17,30,** & Richard S. Houlston1,**

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, \( P = 1.31 \times 10^{-8} \)), 6q21 (rs9372120, \( P = 9.09 \times 10^{-15} \)), 7q36.1 (rs7781265, \( P = 9.71 \times 10^{-10} \)), 8q24.21 (rs1948915, \( P = 4.20 \times 10^{-10} \)), 9p21.3 (rs2811710, \( P = 1.72 \times 10^{-16} \)), 10p12.1 (rs2790457, \( P = 1.77 \times 10^{-16} \)), 16q23.1 (rs193541, \( P = 5.00 \times 10^{-12} \)) and 20q13.13 (rs6066835, \( P = 1.36 \times 10^{-13} \)) which localize in or near to JARD2, ATGS, SMARCD3, CCAT1, CDX2A, 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 Malmö, Sweden. 6 Hematology and Transfusion Medicine, Department of Laboratory Medicine, BMC 813, SE-221 84 Lund, Sweden. 7 Department of Hematology, Erasmus MC Cancer Institute, 3075 EA Rotterdam, The Netherlands. 8 deCODE Genetics, Stuðlagil 8, 1510 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 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 V, University of Heidelberg, 69117 Heidelberg, Germany. 30 Broad Institute, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. * 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: richard.houlston@icr.ac.uk).

ARTICLE
Received 11 Dec 2015 | Accepted 24 May 2016 | Published 1 Jul 2016
DOI: 10.1038/ncomms12050
Open Access
NATURE COMMUNICATIONS | 7:12050 | DOI: 10.1038/ncomms12050 | www.nature.com/naturecommunications
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. Our understanding of MM susceptibility has been transformed by recent genome-wide association studies (GWASs), which have identified the first risk alleles for MM and its precursor condition monoclonal gammopathy of unknown significance. Although projections indicate that additional risk variants for MM can be discovered by GWAS, 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, and three previously published sample sets from Sweden/Norway, Iceland and the United States. 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 $P$-values.

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, $P = 3.13 \times 10^{-8}$), 6q21 (rs9372120, $P = 9.09 \times 10^{-10}$), 7q36.1 (rs7781265, $P = 9.71 \times 10^{-8}$), 8q24.21 (rs1948915, $P = 4.20 \times 10^{-11}$), 9p21.3 (rs2811710, $P = 1.72 \times 10^{-7}$), 10p12.1 (rs7290457, $P = 1.77 \times 10^{-8}$), 16q23.1 (rs27193541, $P = 5.00 \times 10^{-12}$) and 20q13.13 (rs666835, $P = 1.36 \times 10^{-12}$). 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 CCAAT (colon cancer-associated transcript 1; Fig. 2). The same region at 8q24.21 harbours multiple independent loci with different tumour specificities, including the B-cell malignancies diffuse B-cell lymphoma and Hodgkin’s lymphoma, and chronic lymphocytic leukaemia. 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, 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′-untranslated region of the gene encoding 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 (rs7290457) 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. The 8q24.21 (rs1948915) association mapped to intron 3 of the gene encoding ATG5 (ATG5). The 8q24.21 variant rs1948915 maps to CCAAT (colon cancer-associated transcript 1; Fig. 2). The same region at 8q24.21 harbours multiple independent loci with different tumour specificities, including the B-cell malignancies diffuse B-cell lymphoma, Hodgkin’s lymphoma, and chronic lymphocytic leukaemia. 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, 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′-untranslated region of the gene encoding 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 (rs7290457) 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. The 8q24.21 (rs1948915) association mapped to intron 3 of the gene encoding ATG5 (ATG5). The 8q24.21 variant rs1948915 maps to CCAAT (colon cancer-associated transcript 1; Fig. 2).
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 \(WAC\) \((P = 6.58 \times 10^{-14})\) and rs6066835, and increased expression of \(PREX1\) \((P = 3.85 \times 10^{-10})\) (Supplementary Fig. 2 and Supplementary Data 1). We also detected strong \(cis\)-meQTLs at \(WAC\) and \(PREX1\) with rs2790457 and rs6066835 genotypes \((P\)-values \(1.42 \times 10^{-6}\) and \(1.12 \times 10^{-11}\), 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

<table>
<thead>
<tr>
<th>Location</th>
<th>SNP</th>
<th>Position (bp)</th>
<th>Risk allele</th>
<th>RAF</th>
<th>Data set</th>
<th>OR</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>6p22.3</td>
<td>rs34229995</td>
<td>15,244,018</td>
<td>G</td>
<td>0.029</td>
<td>Discovery</td>
<td>1.40</td>
<td>(1.76 \times 10^{-8})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1.19</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.37</td>
<td>(1.21 \times 10^{-8})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P_{\text{het}} = 0.50)</td>
<td>(P = 0)</td>
<td></td>
</tr>
<tr>
<td>6q21</td>
<td>rs9372120</td>
<td>106,667,535</td>
<td>G</td>
<td>0.218</td>
<td>Discovery</td>
<td>1.20</td>
<td>(8.72 \times 10^{-14})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1.12</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.18</td>
<td>(9.09 \times 10^{-15})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P_{\text{het}} = 0.93)</td>
<td>(P = 0)</td>
<td></td>
</tr>
<tr>
<td>7q36.1</td>
<td>rs7781265</td>
<td>150,950,940</td>
<td>T</td>
<td>0.125</td>
<td>Discovery</td>
<td>1.20</td>
<td>(1.82 \times 10^{-7})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1.15</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.19</td>
<td>(9.71 \times 10^{-9})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P_{\text{het}} = 0.24)</td>
<td>(P = 23%)</td>
<td></td>
</tr>
<tr>
<td>8q24.21</td>
<td>rs1948915</td>
<td>128,222,421</td>
<td>C</td>
<td>0.345</td>
<td>Discovery</td>
<td>1.14</td>
<td>(3.14 \times 10^{-10})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1.09</td>
<td>0.0283</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.13</td>
<td>(4.20 \times 10^{-11})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P_{\text{het}} = 0.34)</td>
<td>(P = 11%)</td>
<td></td>
</tr>
<tr>
<td>9p21.3</td>
<td>rs2811710</td>
<td>21,991,923</td>
<td>G</td>
<td>0.657</td>
<td>Discovery</td>
<td>1.14</td>
<td>(6.50 \times 10^{-10})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1.18</td>
<td>(4.02 \times 10^{-5})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.15</td>
<td>(1.72 \times 10^{-13})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P_{\text{het}} = 0.97)</td>
<td>(P = 0)</td>
<td></td>
</tr>
<tr>
<td>10p12.1</td>
<td>rs2790457</td>
<td>28,856,819</td>
<td>G</td>
<td>0.739</td>
<td>Discovery</td>
<td>1.12</td>
<td>(8.44 \times 10^{-7})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1.13</td>
<td>(6.18 \times 10^{-3})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.12</td>
<td>(1.77 \times 10^{-8})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P_{\text{het}} = 0.94)</td>
<td>(P = 0)</td>
<td></td>
</tr>
<tr>
<td>16q23.1</td>
<td>rs7193541</td>
<td>74,664,743</td>
<td>T</td>
<td>0.585</td>
<td>Discovery</td>
<td>1.12</td>
<td>(1.14 \times 10^{-8})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1.17</td>
<td>(4.79 \times 10^{-4})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.13</td>
<td>(5.00 \times 10^{-12})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P_{\text{het}} = 0.15)</td>
<td>(P = 33%)</td>
<td></td>
</tr>
<tr>
<td>20q13.13</td>
<td>rs6066835</td>
<td>47,355,009</td>
<td>C</td>
<td>0.083</td>
<td>Discovery</td>
<td>1.24</td>
<td>(1.16 \times 10^{-9})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1.35</td>
<td>(1.36 \times 10^{-5})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.26</td>
<td>(3.6 \times 10^{-11})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P_{\text{het}} = 0.072)</td>
<td>(P = 43%)</td>
<td></td>
</tr>
</tbody>
</table>

\(P\), proportion of the total variation due to heterogeneity; OR, odds ratio; \(P_{\text{het}}\), \(P\)-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.
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), 7q26.1 (rs7781265), 8q24.21 (rs1948915), 9p21.3 (rs2811710), 10p12.1 (rs2790457), 16q23.1 (rs7193541) and 20q13.13 (rs53422995). Plots (using visPig70) show association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates. –Log10 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 rs719541 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 looping 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 (rs6066855), 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 biomRNA sequencing data that gene expression within the 6q21 and 9p21.3 TADs were tightly correlated ($P < 2 \times 10^{-7}$), which is consistent with their co-regulation (Supplementary Table 17). Moreover, the region at 6q21 (rs9372120, ATG5) participates in intra-chromosomal 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 FAIRE-seq. Although there is little to no evidence across cell lines, we computed chromatin state data for multiple cell lines. HaploReg showed that the majority of MM-related SNPs were observed in regions of DNA hypersensitivity common across multiple cell lines. The protein motifs at these sites are for known TFs such as nuclear factor-kB, c-MYC, GATA, TCF4, POL2, CDCA7L, and 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 data18. Although of borderline significance and hypothesis generating, after correction for the 90 TFs assayed, there was evidence for enrichment of SP1 (alias PU.1), ($P = 0.0007$, $P_{\text{adjusted}} = 0.063$), which regulates PRDM1 and its downregulation is required for MM cell growth19. 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 MYC20,21. 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 i-isof orm of CCAT1 has been shown to interact with the MYC promoter and distal upstream enhancer elements regulating MYC transcription22. 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 CDK2NA/CDKN2B, the 9p21.3 region encompassing SNP rs2811710 interacts with the genomic region containing MTAP (methyldihydroxadenosine 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 survival24. 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 survival25. 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 polyubiquitination26, Variation at 16q23.1 defined with the correlated SNP rs4888262 (pairwise LD with rs7193541, $r^2 = 0.68, D^2 = 1.0$) has previously been shown to influence testicular cancer risk27, suggesting a common genetic and biological basis to both associations.

**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 02/895, IRBCTN64454111), the Myeloma-IX trial by the Oxfordshire Research Ethics Committee (MREC 17/09/09, IRBCTN49047382), HOVON/ GMMG-HD4 (ISRCTN 64455289; MREC 13/01/2015), HOVON/NSMG18
Genetic Markers of Susceptibility studies served as a source of controls. For example, in one study, DNA was extracted from EDTA-blood samples and genotyped using Illumina HumanOmniExpress-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 OmniExpress Exome-8v1-1 arrays.

The UK GWAS comprised 2,329 cases (1,060 male) and 1,76 female controls (post quality control (QC)), mean age at diagnosis: 64 years (ranging from 0 to 99) recruited through the UK MRC Myeloma IX and Myeloma X (EudraCTnr 2007-004007-34, METC 20/11/2008, HOVON/95/EMN02 https://www.clinicaltrialsregister.eu/ctr-search/trial/2009-019703-28/AT). DNA was extracted from EDTA-venous blood samples and genotyped using Illumina HumanOmniExpress-12 v1.0 arrays (Illumina). For controls, we used publicly available data generated by the Wellcome Trust Case Control Consortium from the 1980s Birth Cohort (SOC, 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 Swedish/German GWAS 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 HumanOmniExpress-Exome arrays (Illumina). Control genotypes on 10,704 individuals were obtained from previously published studies of schizophrenia and TWINGENE5.

The USA GWAS comprised 1,512 cases (867 male post QC), mean age at diagnosis: 59 years (ranging from 0 to 99) recruited by the German-Speaking Multiple Myeloma Multicenter Study Group (GMMG) coordinated by the University Clinic, Heidelberg (SBIRCTN0613384; GMMG-HD3 http://www.isrec.org/). DNA was extracted from EDTA-venous blood samples and genotyped using Illumina HumanOmniExpress-12 v1.0 arrays (Illumina). For controls, we used genotype data on 2,107 healthy individuals, enrolled into the Heinz Nixdorf Recall (HNR) study genotyped using Illumina HumanOmniExpress-12 v1.0 arrays (Illumina). For controls, we used publicly available data generated by the Wellcome Trust Case Control Consortium from the 1980s Birth Cohort (SOC, also known as the National Child Development Study) and National Blood Service. Genotyping of controls was conducted using Illumina Human 1-2M-Duo Custom_v1 Array chips (www.wtccc.org.uk).

The German GWAS comprised 1,512 cases (867 male post QC), mean age at diagnosis: 59 years (ranging from 0 to 99) recruited by the German-Speaking Multiple Myeloma Multicenter Study Group (GMMG) coordinated by the University Clinic, Heidelberg (SBIRCTN0613384; GMMG-HD3 http://www.isrec.org/). DNA was extracted from EDTA-venous blood samples and genotyped using Illumina HumanOmniExpress-12 v1.0 arrays (Illumina). For controls, we used genotype data on 2,107 healthy individuals, enrolled into the Heinz Nixdorf Recall (HNR) study genotyped using Illumina HumanOmniExpress-12 v1.0 arrays (Illumina). For controls, we used publicly available data generated by the Wellcome Trust Case Control Consortium from the 1980s Birth Cohort (SOC, also known as the National Child Development Study) and National Blood Service. Genotyping of controls was conducted using Illumina Human 1-2M-Duo Custom_v1 Array chips (www.wtccc.org.uk).

The Swedish/German GWAS 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 HumanOmniExpress-Exome arrays (Illumina). Control genotypes on 10,704 individuals were obtained from previously published studies of schizophrenia and TWINGENE5.

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 within 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 = 571) (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. 9.4). 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 (LGCS, 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. Karyotyping was used for cytogenetic studies of MM cells and standard criteria for the definition of a clone were applied. Fluorescence in situ hybridization and pancytology classification of UK samples was conducted using the methodologies previously described44. Fluorescence in situ hybridization and pancytology classification of German samples was performed as previously described45. The XL IGH Break Apart probe (MetaSystems, Altlussheim, Germany) was used to detect t(11;14) and 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 460 Myeloma XI trial patients46.
Association between genotype and patient outcome. To examine the relationship between SNP genotype and patient outcome, we analysed GWAS data on four patient cohorts 2–4, specifically (i) 1,165 cases from the UK MRC Myeloma IX trial (UK-GWAS); (ii) 877 MM cases from the UK MRC Myeloma IX 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)27. Clinical trial information on these patients has been previously reported 25–27. The primary analysis end point was myeloma-specific overall survival and analysis was performed as previously described 27. 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).

Heritability analysis. We used Genome-wide Complex Trait Analysis to estimate the polygenic variance ascribable to all genotyped and imputed GWAS SNPs in the patient cohorts 2–4, specifically the UK and German GWAS 26. 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 6,6,4, 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, GSE189784, GSE248090, GSE2638 and GSE15695; in the European Genome-phenome Archive (EGA) with accession code EGAS000000000001; in the European Bioinformatics Institute (Part of the European Molecular Biology Laboratory) (EMBL-EBI) with accession code E-MTAB-1138; and in the database of Genotypes and Phenotypes (dbGaP) with accession code phs002087.v1.p1.

Expression data that support the findings of this study have been deposited in GEO with accession codes GSE21349, GSE2658, GSE3161 and EMBL-EBI with accession code E-MTAB-2299.

Whole-exome sequencing data that support the findings of this study have been deposited in EMBL-EBI with accession code 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
18. Hnisz, J. et al. Genome U133 2.0 Plus Array data for plasma cells from 183 MRC Myeloma IX trial patients. NATURE COMMUNICATIONS| 7:12050 | DOI: 10.1038/ncomms12050 | www.nature.com/naturecommunications
19. Plotkin, J. B. et al. Heritability analysis of UK and German GWAS data sets was performed under a standard fixed-effects model.

Note: The table and figures are not included in the natural text representation.

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 from the 1000 Genomes Project. This study is supported by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) supported at the University College London Hospitals NHS Foundation Trust and the海洋血液学研究所在英国的血液学研究中心。
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 Adelsson, Tomas Axelsson, Anna Collin, Ildiko Frigyesi, Patrik Magnusson, Bertil Johansson, Jan Westin and Helga Ogmundsdóttir 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

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


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).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material.

To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
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.
Multiple myeloma (MM) is the second most common hematologic malignancy. It is defined by an unhindered, clonal growth of 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 premalignant condition that progresses to MM at a rate of about 1% per year.

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. 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 other malignancies. Recently, genome-wide association studies have identified DNA sequence variants at 18 independent loci that associate with MM risk. The MM risk allele is represented by ~70 sequence variants in tight linkage disequilibrium (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.

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.

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 ($n = 604$). In all these datasets, rs3815768-C associates with lower ELL2 expression (Fig. 1b; Fisher’s inverse $\chi^2$ test combined $P = 2.5 \times 10^{-22}$ 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, respectively (average across exons 1–6 and 9–11) than samples homozygous for the protective allele.

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), 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 reduced 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.

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...
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). 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.

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 erythroblastic 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, 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.

### Table 1 Association testing results

<table>
<thead>
<tr>
<th>Exon</th>
<th>Chr.</th>
<th>Position (hg38)</th>
<th>Effect size ($\beta$)</th>
<th>$r^2$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>95885097–95888967</td>
<td>-0.15</td>
<td>0.02</td>
<td>0.036</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>95890605–95889130</td>
<td>-0.16</td>
<td>0.03</td>
<td>0.026</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>95891102–95892074</td>
<td>-0.17</td>
<td>0.03</td>
<td>0.018</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>95895627–95895691</td>
<td>-0.16</td>
<td>0.03</td>
<td>0.018</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>95898239–95898810</td>
<td>-0.15</td>
<td>0.02</td>
<td>0.040</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>95900692–95900780</td>
<td>-0.14</td>
<td>0.02</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>95900955–95901080</td>
<td>-0.04</td>
<td>0.00</td>
<td>0.601</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>95906522–95906782</td>
<td>-0.11</td>
<td>0.01</td>
<td>0.133</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>95917770–95919394</td>
<td>-0.20</td>
<td>0.04</td>
<td>0.007</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>95919423–95919545</td>
<td>-0.20</td>
<td>0.04</td>
<td>0.007</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>95913001–95913049</td>
<td>-0.19</td>
<td>0.04</td>
<td>0.010</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>9591574–95912071</td>
<td>-0.12</td>
<td>0.02</td>
<td>0.093</td>
</tr>
</tbody>
</table>

Association testing results for the Swedish-Norwegian mRNA-sequencing dataset ($n = 185$). Effect size ($\beta$), squared Pearson regression coefficient ($r^2$), and $P$ values indicated.
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 $\chi^2$ 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 indiected 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 (rs14232669), 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 extracts, about 20 of which are expressed in MM plasma cells (Supplementary Tables 4 and 5). Electrophoretic mobility shift assays (EMSSAs) 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 EMSSAs 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 MAFF and MAFG expression (Supplementary Table 5), and rs3777189 maps to an annotated MAFK ChIP-seq peak in lymphoid cells (Supplementary Table 1). The MAFF protein family (MAF, MAFA, MAFF, MAPK, 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 cells22,25,40.

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 plasmacytoma 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.
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 plasmacytoma 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 plasmacytoma cell lines. Two of these (rs3777185 and rs3777189) exhibited altered binding of nuclear proteins, and rs3777189-C diminishes binding of MAFF/G/K family transcription factors. In addition, we identified 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 efficient13,19,21–23. 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-occurrence 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 cancer32.

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

---

**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
Icelandic samples (deCODE Genetics, Reykjavik; unpublished). The sample collection was done subject to informed consent and ethical approval (Lund University Ethical Review Board, 2013/34; Icelandic Data Protection Authority, 200100157; 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 USA29.

The German sample set consists of 658 MM patients from the Heidelberg University Clinic and the German-speaking Myeloma Multicenter Group53. The British sample set comprises 183 MM patients enrolled in the UK Medical Research Council Myeloma IX trial31. The US sample set comprises 604 samples from newly diagnosed patients treated at the UAMS Myeloma Institute for Research and Therapy32. The three validation datasets were generated using Affymetrix U133A 2.0 plus microarrays and custom chip definition files ("http://brainarray.mgh.harvard.edu/Brainarray/Databases/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. #76235 or #762174). The RNA integrity (RIN) was assessed using Agilent 2100 Bioanalyzer (v.6.7.1.1), or LabChip GX (Perkin-Elmer, Waltham, MA, USA). Index 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 indices, PCR amplification, and QC were 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 were done using HCSL3.8-1.4 and RNAI 10.36–12.2 4 analysis packages. Demultiplexing and generation of FASTQ files was performed using scripts from Illumina (h2clspace v1.8). Sequence alignment and fragment counts was done with TopHat2 and HtSeq-count, respectively57. 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 (v1.17)52.

Genotyping. The Swedish-Norwegian sample set was genotyped at two levels: first, all samples (n = 183) were genotyped for the ELL3 MA risk allele using the coding variant rs17158768, which could be robustly typed manually from the RNA sequencing data using Integrative Genomics Viewer (Supplementary Fig. 1a). In addition, a subsample of the Swedish-Norwegian sample set was genotyped on Illumina Human OmniExpress microarrays (n = 118). To increase the genomic resolution, these data were haplotype- phased using SHAPEIT2 (v.2.790)58 and imputed by IMPUTE2 (v.2.3.5)59 with the 1000 Genomes Phase 3 3 complement reference data (October 2014 release)57. The German, UK, and US myeloma sample sets were genotyped previously on Illumina Human OmniExpress-12 v1.1 arrays60,61 and imputed using the UK10K compendium62. For the Icelandic blood samples, genotypes were obtained by imputing variants identified by whole-genome sequencing of 8433 Icelanders into 150,656 chip genotyped individuals using long-range phasing based imputation52,53. Probabilities of genotypes were also predicted from newly diagnosed patients treated at the UAMS Myeloma Institute for Research and Therapy32.

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 mon level, in order to allow detection of expression-specific effects and to avoid signal dilution due to alignment bias caused by variant codings (Supplementary Fig. 1b). For association testing, we used Pearson’s correlation as implemented in R (v.3.4.0). Effect sizes (beta, ρ) and standard errors (SE) of eQTLs were calculated using R (v.3.3.0). The coefficient of linkage disequilibrium (D’) and r-squared (r²) were calculated using the Central European part of the 1000 Genomes compendium available via Haploking 4.1. To estimate the risk allele ratios in rs185767-CT heterozygotes, we counted the two corresponding to DNA sequences TCTGGTAAGTCTCGAGCGCCCGG (clone #16) and TCTGGTAAGTCTCGAGCGGCCCGG (clone #6). All probes were biotin-labeled at the 5’ end and hybridized to the raw RNA-seq data from the CRISPR design tool (http://www.crispr.mit.edu-tool) and target ELL2 exon 1, were synthesized and ligated into lentCRISPRv2 vector (AddGene, Cambridge, MA, USA; cat. no. 592961) using published protocols63. An aliquot of

Chromatin immunoprecipitation sequencing. L363 cells were crosslinked with 1% paraformaldehyde (ThermoFisher, #28908) at 37 °C in water bath for 11 min. After crosslinking, chromosomes were sonicated using BioUltra (USA) sonicator for 24 samples/100ul at 6-8% intensity at 100 °C for 1.5 minutes. The crosslink was reversed by heating at 65 °C for 2 hours. The chromatin was diluted 10-fold with PBS and rotated overnight at 4 °C. After centrifugation at 16,000 g for 10 min, the supernatant was used for immunoprecipitation with an antibody against 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 concatemer assay kit (Zymogen, #89205). Concentration was measured using Qubit 2.0 fluorometer. The ChIP-Seq library was prepared using ThruPLEX DNA-seq kit (Rubicon Genomics, #R804004). 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 sequencers 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 (h2clspace v1.8). FASTQ files were used to assess read quality, GC content, the presence of adapters, over-represented k-mers and duplicated reads. Bases with low quality score were removed using Trimmmomatic program (v.0.36)64. Trimmed reads were aligned using Bowtie2 (v.2.3.0)58.

Luciferase assays. Ten double-stranded nucleotide sequences of 120 bp each, including adapters and amplification primers, were commercially synthesized (Integrated DNA Technologies, USA). The sequences correspond to rs1381401, rs1914162, rs3777819, rs3777785, rs6673739, rs3777784, rs8899802, and rs1536684 (Supplementary Table 3). Sequences were directionally cloned into a pGL3-Blasticid (Promega) upstream of a luciferase reporter gene56. Sanger sequencing confirmed the inserts. Renilla luciferase was used as internal transfection control. L363, OPM2, K562, and RPMI-8226 cells were cultured at 37 °C and 5% CO2 in RPMI 1640 medium ( Gibco, Life Technologies) supplemented with 10% fetal bovine serum ( Gibco). These cells were transfected with each of the ten clone sets, using Lipofectamine 24×4 transfection reagent (Life Technologies). Post 24 h transfection, cells were harvested and lysed in lysis buffer. An aliquot of 20 µl of the lysates was used for luciferase measurement following manufacturer’s protocol (dual-luciferase reporter assay system, Promega). Measurements were measured at GLOMAX 20/20 Luminometer using Run Promega Protocol (DLR-0-INI). Effects were quantified as log2 ratios of renilla-normalized luminescence values for the risk alleles divided by the corresponding values for the protective alleles (median over three on average to seven replicates per sequence and cell line).

Electrophoretic mobility shift assays. For nuclear proteins, the gel shift assay was used the following 25 bp double-stranded probes (variations in brackets): for rs3777789, ACAGTGCTTGATGC(CT)ACCAGCTAAAAT, rs3777785, CTGGCAACTTCG(CT)ACGGAGTGCTC, rs4561648, GAAAACCTTCCTAGC(T)TCCTGGACTTGTT. 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 superfrost slices with nuclear extracts from OPM2 and L363 cell lines (DSMZ, Braunschweig, Germany), 1 µg antibody was added to the reaction mix and incubated at 4 °C for 1 hour. The cell line identities were 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://oper.austowe.com/perfectosap) with the HOCOMOCO-10, JASPAR, HT-SELEX, Swiss Regulon and PEOBases databases (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 TCTGGTAAGTCTCGAGGCCCCG (clone #6) and TCTGGTAAGTCTCGAGGCCCCG (clone #7) using the CRISPR design tool (http://www.crispr.mit.edu-tool) and target ELL2 exon 1, were synthesized and ligated into lentCRISPRv2 vector (AddGene, Cambridge, MA, USA; cat. no. 592961) using published protocols63.

An aliquot of
ligated mix was transformed to JM109 competent cells. sgRNA inserts were confirmed by Sanger sequencing using standard Hig-F primer. The lentCRISPRv2 vector containing inserts was transfected into LMD3 cells by electroporation and puroycin selection. Successful knockout was verified by western blot with anti-
 bodies 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 Bioeptune-pico (Diasogene) for ten cycles at 30°C for 1 min and off. Subsequently, samples were heat denatured at 95°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, #376611) and developmental (Bio-Rad). Membranes were re-probed with GADPH antibody after re-
 blot treatment (Millipore, #2902).

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-statistic52. For gene set enrichment analysis, we used the RendorCat tool with default parameters, Gene Ontology54 and ABI Panther (http://panterdb.org) gene sets, and considered genes with average FPKM > 5 in the MmPC RNA-seq data. We also created specific gene sets comprising the ~80 genes encoding the proteins of the large and small ribosomes subunits (RPLP7 and RPS7) 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 plasma cells. These data were retrieved from the NCBI Gene Expression 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 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_012681.5, 331–2273) was synthesized as gBlocks Gene Fragments from IDT. The gene fragment was synthesized as gBlocks Gene Fragments from IDT.

References
54. Johnsson, O., Magnusson, G., Hallldósson, D., Gudbjartsson, D., Jonsdottir, I., Thorsteinsdottir, U., Stafannson, K., Stefansson, K. & Stefansson, K. 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. Holstien for providing access to gene expression data for plasma cells from patients from the United Kingdom. We thank A. Collins 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

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018
Correspondence: Sequence variation at the MTHFD1L-AKAP12 and FOPNL loci does not influence multiple myeloma survival in Sweden

Authors
Mina Ali¹, Konstantinos Lemonakis¹,², Anna-Karin Wihlborg¹, Ljupco Veskovski³, Ingemar Turesson¹, Ulf-Henrik Mellqvist³, Urban Gullberg¹, Markus Hansson¹,²,⁴,* and Björn Nilsson¹,⁵,*

Affiliations
¹Hematology and Transfusion Medicine, Department of Laboratory Medicine, BMC B13, 221 84 Lund, Sweden. ²Hematology Clinic, Skåne University Hospital, 221 85 Lund, Sweden. ³Section of Hematology, South Elvsborg Hospital, SE 501 83 Borås, Sweden. ⁴Wallenberg Center for Molecular Medicine, 221 84 Lund, Sweden. ⁵Broad 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\(^4\text{-}^8\). Additionally, two recent studies indicate that genetic variation could also influence MM survival\(^9\text{-}^{10}\).

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 \textit{MTHFD1L} and \textit{AKAP12} genes at chromosome 6q25.1\(^9\). The protein encoded by \textit{MTHFD1L} is involved in folate metabolism\(^11\), and \textit{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\(\times\)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 (\(P=1.69\times10^{-6}\) to 0.009; HR 1.06 to 1.75), but not in the one from Germany (\(P=0.55; HR=1.09, 95\% CI 0.82\text{-}1.44\)).

In a second study, Ziv et al.\(^10\) describe an association between MM-OS and rs72773978 near \textit{FOPNL} at 16p13. The protein encoded by \textit{FOPNL} has been implicated in centrosome function\(^13\). The association was detected by meta-analysis of 545 cases from two clinical trials in USA (\(P=6\times10^{-10}; HR=2.65, 95\% CI 1.94\text{-}3.58\)). 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 \(P=0.044; HR=1.34, 95\% CI 1.01\text{-}1.78\)). 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 MTHFD1L-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 Genomes6,14. 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= $1.28 \times 10^{-5}$ 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


Figure 1

(a) rs12374648

- Survival probability vs. time
- p = 0.7012

(b) rs7273978

- Survival probability vs. time
- p = 0.927

<table>
<thead>
<tr>
<th>Number at risk</th>
<th>MM00</th>
<th>MM01</th>
<th>MM11</th>
<th>MM11+MM11</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12374648</td>
<td>530</td>
<td>319</td>
<td>116</td>
<td>16</td>
</tr>
<tr>
<td>rs7273978</td>
<td>792</td>
<td>483</td>
<td>183</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>79</td>
<td>53</td>
<td>20</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 1: Clinical characteristics of the study population

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

MINA ALI
DEPARTMENT OF LABORATORY MEDICINE | LUND UNIVERSITY