Chromatin-based, *in cis* and *in trans* regulatory rewiring underpins distinct oncogenic transcriptomes in multiple myeloma

Jaime Alvarez-Benayas\(^1,6\), Nikolaos Trasanidis\(^2,6\), Alexia Katsarou\(^2,3,6\), Kanagaraju Ponnusamy\(^2\), Aristeidis Chaidos\(^2,3\), Philippa C. May\(^2\), Xiaolin Xiao\(^2\), Marco Bua\(^3\), Maria Atta\(^3\), Irene A. G. Roberts\(^4\), Holger W. Auner\(^2,3\), Evdoxia Hatjiharissi\(^5\), Maria Papaioannou\(^5\), Valentina S. Caputo\(^2,6\), Ian M. Sudbery\(^1\) \& Anastasios Karadimitris\(^2,3\)

Multiple myeloma is a genetically heterogeneous cancer of the bone marrow plasma cells (PC). Distinct myeloma transcriptome profiles are primarily driven by myeloma initiating events (MIE) and converge into a mutually exclusive overexpression of the *CCND1* and *CCND2* oncogenes. Here, with reference to their normal counterparts, we find that myeloma PC enhanced chromatin accessibility combined with paired transcriptome profiling can classify MIE-defined genetic subgroups. Across and within different MM genetic subgroups, we ascribe regulation of genes and pathways critical for myeloma biology to unique or shared, developmentally activated or de novo formed candidate enhancers. Such enhancers co-opt recruitment of existing transcription factors, which although not transcriptionally deregulated per se, organise aberrant gene regulatory networks that help identify myeloma cell dependencies with prognostic impact. Finally, we identify and validate the critical super-enhancer that regulates ectopic expression of *CCND2* in a subset of patients with MM and in chronic lymphocytic leukemia.
Multiple myeloma (MM) is a common, genetically heterogeneous and incurable cancer of the bone marrow (BM) plasma cells (PC), the terminally differentiated, immunoglobulin-secreting B lineage cells. The first level of genetic heterogeneity in MM is imparted by well-defined myeloma-initiating events (MIE) that are associated with distinct transcriptome profiles. In nearly half of MM cases, MIE include overexpression of the oncogenes CCND1, MAF and MMSET by their juxtaposition to the immunoglobulin heavy chain (IgH) enhancer, thus defining the t(11;14), t(14;16) and t(4;14) cytogenic subgroups, respectively. Hyperdiploidy (HD), a functional heterogeneous subgroup characterised by additional odd number chromosome, is the MIE in the rest of MM cases. Secondary events comprising copy number aberrations, single nucleotide variants and indels generate additional genetic heterogeneity and further shape the distinct impact of the MIE on oncogenic transcriptomes. This heterogeneity converges, in most cases, to a functionally dichotomous, mutually exclusive over-expression of the cell cycle regulators CCND1 and CCND2 to which myeloma PC remain addicted, irrespective of primary or secondary genetic events. The transcriptional mechanisms that result in CCND2 over-expression, seen in nearly 50% of MM cases and spanning all genetic subgroups except t(11;14), are not fully known. A previous study of a single myeloma cell line identified but not validated a super-enhancer that spans the promoter of CCND2, leaving the possibility of a distal enhancer/super-enhancer regulating transcription of CCND2 unexplored.

Chromatin accessibility profiling by ATAC-seq has been used to characterise the regulatory landscape of hundreds of different solid tumour and blood cancers, such as chronic lymphocytic leukaemia (CLL). Further, by means of transcription factor (TF) footprinting, ATAC-seq allows inference of TF binding profiles. This, in combination with paired transcriptome profiles, enables the construction of gene regulatory networks. Such networks may identify TFs with previously unrecognised roles in the biology of a given cancer. In previous studies of chromatin-based regulatory changes of gene expression, myeloma was treated as a homogeneous cancer. However, how heterogeneity links chromatin accessibility and regulatory status with distinct oncogenic gene expression profiles has not been elucidated, while global insights into the interplay between in cis and in trans regulatory factors of gene transcription in MM are limited.

Here, by integrating chromatin accessibility dynamics of myeloma PC with their respective transcriptome profiles and other epigenetic datasets we resolve the chromatin changes that regulate distinct oncogenic transcriptomes and biological pathways. Through this process we discover cis- and trans-regulators of myeloma biology, including those involved in the regulation and aberrant expression of CCND2 in MM.

Results

Enhanced accessibility of distal chromatin elements is associated with gene over-expression in myeloma plasma cells. We isolated fresh, highly purified BM (CD19+/−) PC from three healthy normal donors (ND) and myeloma PC from 30 MM patients, covering the main MIE subgroups as defined by fluorescent in situ hybridization (Fig. 1a, Supplementary Fig. 1a and Supplementary Data 1) and spanning both diagnostic and relapsed stages of the disease. For each sample, we obtained paired chromatin accessibility and transcriptome profiles by ATAC-seq and RNA-seq, respectively.

ATAC-seq analysis generated 295,238 chromatin accessibility peaks from all samples (Supplementary Fig. 1b). These peaks were non-randomly distributed on promoters, coding and intergenic regions, and were found to be unique or shared between two or more genetic subgroups (Supplementary Fig. 1b and c). In each genetic subgroup and overall, chromatin accessibility of myeloma PC was enhanced in comparison to ND PC (Fig. 1b and Supplementary Data 2).

Transcriptome profiling identified 3036 differentially expressed genes (DEG) between myeloma and ND PC, including over-expression of known MIE-driven oncogenes in myeloma PC (Supplementary Fig. 1d and Supplementary Data 3). In addition, using genesets previously identified as classifiers for each myeloma genetic subgroup and the corresponding MIE, we clustered the transcriptome profiles of our samples along with those of 892 myeloma PC samples from the MMRF CoMMpass study (Supplementary Fig. 1e). This analysis confirmed the accuracy of genetic subgroup annotation of our cohort samples by FISH, and assisted in the classification of samples with unknown cytogenetics status.

Malignant vs normal state analysis showed that over-expressed genes were non-randomly distributed over non-TSS peaks with increased accessibility than non-overexpressed genes (Supplementary Fig. 1f), and the distance from over-expressed genes to the closest such peak was shorter (Supplementary Fig. 1g).

Combined logistic regression analysis revealed that the number of more accessible non-TSS regions was predictive of over-expression (p < 2 × 10−16) but a more open promoter was not (p = 0.5; Fig. 1c), although gain in TSS accessibility was mildly predictive when ignoring distal peaks (over expression odds ratio of 1.3 for TSS accessibility gain vs non, p = 0.01).

Using the 3D genome architecture of GM12878 B cells as reference, we found a significant enrichment for differentially accessible regions (DARs) correlating with DEG within the same topologically associating domain (TAD) as compared to permuted DEG-DAR associations (p < 0.001, see Fig. 1d). We also found that the strength of correlations, i.e., mean DAR-DEG correlation coefficient and the percentage (%) of DAR-DEGs that show significant correlation (p < 0.05), is higher when DEG-DAR pairs are in the same TAD compared to distance matched controls not in the same TAD (p = 0.003, Supplementary Fig. 1h and i).

Thus, myelomagenesis leads to an overall increase in chromatin accessibility and, within the spatial limits of TADs, chromatin decondensation at distal regions rather than at promoters is associated with gene over-expression in the cancer state.

Over-accessible chromatin in myeloma PC partially distinguishes distinct myeloma transcriptomes according to MIE.

Next, we sought to explore and potentially resolve the heterogeneity present across MM patients in an unsupervised manner, by employing multi-omics factor analysis (MOFA) (Fig. 1e and f, Supplementary Fig. 2a). We found that within the combined ATAC-seq/RNA-seq model, chromatin accessibility accounted for more variance than expression (Fig. 1e). Of the top five identified MOFA latent factors (LF), the mostly chromatin accessibility-driven LF1 and the mainly transcriptome-driven LF2 distinguished ND from myeloma PC; the next three factors, LF3-5 separate different MM patients from another. To investigate possible sources of this heterogeneity, we superimposed the MIE-defined genetic subgroup characterisation of each sample. This revealed a clear separation of ND, MAF and CCND1 subgroups and less so of HD and MMSET by combined LF2, 3 and 5 (Fig. 1f and g).

By interrogating further the MOFA predictors underlying the observed segregation, we found that genes, previously identified as MIE-defined genetic subgroup classifiers, along with their linked chromatin changes delimited in the same TAD, displayed the same subgroup-segregation profile on those three LF (Fig. 1h).
We validated this using two MAF-, two CCND1- and one MMSET-translocated myeloma cell lines as a test set not used in training the model, and confirmed separation of samples according to MIE in the LF2-LF3-LF5 MOFA space (Supplementary Fig. 2b–d).

To explore the properties of MOFA clustering further, we applied silhouette width (i.e., a measurement of cluster cohesion and separation) and Discriminant Ratio (i.e., the ratio of between group variance to within group variance; see "Methods"). Silhouette width showed that the combination of accessibility and transcriptome information of LF1-5 provides a clear separation of ND, MAF and CCND1 but not HD and MMSET subgroups, and this separation was clearer than in the model built using RNA-seq alone (Supplementary Fig. 2e).

When the discriminant ratio was applied to measure the ability of each LF to discriminate subgroups, LF5 had the highest ratio at 22.7. To understand how much discriminatory power ATAC-seq data was adding to the model, we trained a second model using the transcriptome data only. The best discriminating factor in this model was LF4, with a discriminant ratio of 11.3, showing that the addition of chromatin accessibility data almost doubled the power of the most discriminatory factor to separate subgroups.
Applying linear discriminant analysis to each model (Supplementary Fig. 2), showed the subtypes are well separated in the combined ATAC-RNA model, with the exception of one MMSET sample located with the HD samples, and one HD sample located with the CCND1 samples. Conversely, for the RNA-only samples, while MAF and ND samples are well separated, the HD, MMSET and CCND1 samples are largely overlapping.

Since there was no difference between CD19+ and CD19- ND PC cells, they were merged for subsequent analyses.

Given these findings, we conclude that combined epigenome–transcriptome-based categorisation of MM maps away myeloma PC from normal PC while within the main MM genetic subgroups it clearly delineates the MAF and CCND1 subgroups and less so the HD and MMSET subgroups.

The myeloma enhancome is linked to known and novel MIE-associated genes and biological pathways. To expand our ability to delineate heterogeneity and given the importance of distal genetic elements into the regulation of MM transcriptomes, we next sought to identify the changes that together comprise the myeloma-specific enhanceome using a genetic subgroup supervised approach. We identified distal regions and genes where genetic subgroup (i.e., MAF/MMSET/CCND1/HD/ND) is a significant explanatory variable for accessibility/expression using an omnibus LRT test to maximise power compared to pairwise testing (adjusted LRT p-value <0.05 and >2-fold change between at least one MIE and ND, Supplementary Data 3). Integration of these two sets (4365 regions and 3,096 genes) gave 4199 DAR–DEG pairs within 1 Mb of one another (Supplementary Data 4), comprising 2581 unique DAR and 1354 unique DEGs. Hierarchical clustering of the DAR accessibility signal clearly separated the different myeloma MIE subgroups from each other (Fig. 2a). While this showed a group of regions where changes correlated with MIE, it also highlighted accessibility changes that are shared between different myeloma subgroups. The MAF subgroup for instance, demonstrated the highest number of distal DAR that are >2-fold different from ND, some of which were unique to MAF while others were also >2-fold more accessible in other genetic subgroups. To validate this further, we clustered myeloma PC (n = 26) from a previous study11 using these same 2581 regions, and confirmed that these samples mostly clustered with samples carrying the same MIE from this study (Supplementary Fig. 3a).

Of note, 977 of the 2581 (38%) DAR were proximal to more than one DEG, while 962 of 1354 DEG (71%) were within 1 Mb of more than one DAR (Supplementary Data 4), and often in more than one genetic subgroup. These findings suggest that diverse MIE functionally converges to aberrantly regulate the same regions of chromatin, a process consistent with chromatin accessibility-based convergence evolution.

At least 14 of the DEG linked to distal DAR have been implicated in myeloma pathogenesis, including HGF16, DKK117 and UCHL118 (Fig. 2b, Supplementary Fig. 3b and c and Supplementary Data 4). In two myeloma cell lines studied, these same regions are marked by H3K27me3 and Polycomb repressive complex components in several cell types (Supplementary Fig. 4b, Supplementary Data 5b) and enrichment for genes selectively expressed in neuronal cell types, particularly in the HD and MMSET subgroups (Supplementary Fig. 4b and c, Supplementary Data 5c).

Finally, enrichment of previously validated MIE gene set classifiers was highest in their corresponding genetic subgroups (Fig. 2d).

Together, these findings suggest that DAR with regulatory potential are to a large extend shaped by MIE, although secondary genetic events are also likely to contribute.

Developmental ‘re-commissioned’ and de novo formed enhancers in MM. Next, we sought to gain insights into the developmental origins of the MM over-accessible candidate enhancers linked to DEG. For this purpose, we tracked the candidate enhancer chromatin status, as defined by combinatorial enrichment of histone marks (ChromHMM states21), across different mature B lineage cells22 (Fig. 2e and Supplementary Data 6).

Considering that an active enhancer requires combined enrichment for H3K27ac with H3K4me123, we identified 254 (out of 832) DAR with predicted regulatory activity over 201 DEG within the same TAD that are only present in myeloma and not normal PC (Fig. 2e and Supplementary Data 6), i.e., they are de novo formed (e.g., the enhancers predicted to regulate HGF and UCHL1 shown above). TF motif enrichment analysis of these 254 DAR identified IRF and MEF families of TF amongst others as possible leading transcriptional regulators of their activity (Fig. 2f).

In addition, a smaller number of DAR are ‘re-commissioned’ in myeloma PC, i.e., active in one or more B lineage cells but inactive in ND PC (Supplementary Data 6).

Therefore, biochemical annotation of the distal over-accessible chromatin profile identifies enhancers that are myeloma PC unique or developmentally inherited.

TF ‘rewiring’ reveals myeloma dependencies with prognostic impact. Next, we employed ATAC-seq footprinting24 to identify the predicted association of DNA binding factors with chromatin across myeloma and ND PC. In total, 138 of 254 expressed TF (Fig. 3a and Supplementary Data 7a–c) displayed higher or similar predicted binding frequency in at least one myeloma subgroup compared to ND PC, and included TF such as XBP1, IRF4 and PRDM1 known to regulate myeloma transcriptomes, but also TF such as CBFB and ZNF384 which have not been previously linked to myeloma biology (Fig. 3a, b and c). Another 116 TFs were predicted to bind to chromatin in at least one MM subgroup but not in ND PC, including established, subgroup-specific oncogenic drivers (e.g., MAF). Almost one third of these 116 TF were predicted to be exclusively active in individual subgroups, with ISL2, a neural TF25 not previously linked to MM, showing activity solely in the HD subgroup (Fig. 3a–c, and Supplementary Data 7b).

CRISPR/Cas9 screens involving depletion of 247/254 TF as retrieved from the DepMap database, suggested myeloma cell dependency on 55% (137/247) of TF in at least 4/20 myeloma cell lines at a CERES score of <−0.2 (Fig. 3d and Supplementary Data 7d), and confirmed MM cell dependency on the TF CBFB and ZNF384 identified by ATAC-seq footprinting (Fig. 3c and d).

Interestingly only 13% (18/137) of these TF are differentially expressed in one or more myeloma subgroups compared to ND PC (Fig. 3e). This is consistent with a pattern of TF activity ‘re-wiring’ in myeloma PC that does not necessarily require transcriptional deregulation of the TF themselves.
To define this 're-wiring' of TF further, for each MM subgroup and ND PC we built TF regulatory gene networks based on the weighted frequency of binding and level of expression (Fig. 4a). In general, compared to ND PC, we observed a higher number of active TF in all myeloma subgroups. These formed higher density regulatory connections with other TF and displayed auto-regulatory loops (Supplementary Fig. 5a and Supplementary Data 7e), commensurate with increased binding frequency for >90% of TF in each myeloma subgroup (Supplementary Fig. 5b). For example, the established myeloma cell dependencies and PC lineage-defining TF IRF4, XBP1 and PRDM1, are also predicted to be connected to a higher number of other TF in HD, MMSET and CCND1 subgroups than in ND PC, while TF of the MEF family that have not been linked to myeloma biology before and
are predicted to regulate de novo formed myeloma enhancers (Fig. 2f) display higher connectivity and betweenness (centrality). A high level of myeloma cell line dependency to MEF2C in the DepMap CRISPR/Cas9 screen (Fig. 3d), highlights an important role of this TF in the biology of MM, less because of its inferred role in activating transcription of TF such as IRF4, XBPI and PRDM1 (Supplementary Data 7f).

Focusing on the MAF-translocated subgroup, we performed ChIP-seq to obtain the cistrome of oncogenic MAF in the MAF-translocated myeloma cell line MM.1S (Supplementary Fig. 5c). MAF binding was predominantly enriched at TSS/promoters and intergenic areas (Supplementary Fig. 5d). Motif analysis performed on the MAF binding sites identified significant enrichment of the MAF motif (Supplementary Fig. 5e) and motifs of IRF1-4, NRF2/NFE2L2, ATF3, BACH1 TFs (Supplementary Fig. 5f), which were also predicted to be active within the MAF subgroup gene regulatory network (Supplementary Data 9b).

Six TFs (CXXC1, BPTF, MAZ, KLF13, CBFB and RFX5) were identified as showing higher connectivity in all myeloma subgroups compared to ND PC (Supplementary Data 7e), thus exemplifying the process of existing TF ‘re-wiring’ in MM. These TF, none of which have been previously linked to MM, share another three notable features: (a) they demonstrate dependency on DepMap upon CRISPR/Cas9 and siRNA screens (Fig. 3d and Supplementary Fig. 5g), (b) display highest expression in multiple myeloma cell lines compared to >1000 other cancer cell lines; and (c) their higher expression has a significant adverse impact on survival in two independent myeloma patient cohort datasets (Fig. 4b and c). In addition, myeloma cell dependency on CXXC1 was further validated by independent shRNA knockdown experiments (Supplementary Fig. 5h). Together, this multi-layered approach reveals TF dependencies and prognostic variables in MM.

Identification and characterisation of the CCND2 super-enhancer. Furthermore, we sought to identify and characterise the regulatory mechanisms of CCND2 overexpression in MM.

The LF5 of MOFA analysis completely separated MAF-translocated from CCND1-translocated samples (Fig. 1f and g), placing extreme opposite weights on the expression of CCND2 and CCND1, respectively (Fig. 5a). LF5 also places high importance on a set of open-chromatin regions upstream of CCND2, linking them to enhanced expression of CCND2 and transcripts upstream, but not downstream, of CCND2 (Fig. 5b). Accessibility of this region correlates with CCND2 activity irrespective of MIE (Fig. 5c). Further, super-enhancer calling using H3K27ac and MED1 chromatin marks in MAF-translocated MM.1S myeloma cells identified the region of interest as a bona fide super-enhancer by both markers (Fig. 5d and Supplementary Fig. 6a).

Consistent with it being a ‘re-commissioned’ enhancer of CCND2, the region of interest is Polycomb repressed in GCB and PC but active in naïve and memory B cells (Supplementary Fig. 6b); accordingly, CCND2 is expressed in naïve and memory B cells but not GCB cells or PC (Supplementary Fig. 6c).

Using Hi-C genome data from GM12878 B cells, we identified exclusive, long-range interactions of the CCND2 promoter with the upstream accessible clusters of the putative enhancer (Fig. 5d and Supplementary Fig. 6d). In a complementary approach, we employed KRAB-dCas9 CRISPRi in MAF-translocated myeloma cells to repress the activity of four prominent constituent peaks 1–4 (Fig. 5d) which engage in high-frequency interactions with the CCND2 promoter. As expected, targeting a promoter accessibility peak resulted in a significant decrease in CCND2 expression, while in the CCND2 enhancer, the most pronounced effect, similar to that of the promoter peak, was conferred by targeting the proximal peak 4 and distal peak 1 accessibility regions (Fig. 5f). Notably, accessible peaks 1 and 4, but not others in between, are Polycomb-repressed in GCB and PC but active in naïve and memory B cells (Supplementary Fig. 6b). Thus, the relative importance of peaks 1 and 4 is also validated from a developmental perspective.

Having dissected the in cis regulatory mechanisms of CCND2 expression, we proceeded with the characterization of trans factors involved in this process.

Previous work showed that MAF binds to CCND2 promoter in vitro26, providing some insight into how CCND2 is regulated in the MAF genetic subgroup. Importantly, our ChIP-seq analysis in MM.1S myeloma cells shows that MAF binds to the enhancer of CCND2 in vivo (Fig. 5d), thus consolidating its role as a critical regulator of CCND2 over-expression in MAF-translocated MM cells. This finding provides also insights into CCND2 regulation in the MMSET genetic subgroup, in which, both CCND2 and MAF are expressed at lower levels26 (Fig. 5e and Supplementary Fig. 6e). Since chromatin accessibility signal is also lower in the MMSET than in the MAF subgroup (Fig. 5d and Supplementary Fig. 6e), together, these findings are consistent with the notion that the transcriptional activity of CCND2 enhancer in the MAF and MMSET subgroups is MAF dosage-dependent.

To identify other TF potentially regulating CCND2 expression in CCND2-expressing HD MM (which lack expression of MAF), we performed differential footprinting analysis in CCND2high vs CCND2low HD myeloma PC (Fig. 5g and Supplementary Fig. 6f and g). In addition to TF known to be implicated in MM (IRF4, PRDM1, FLI1)11,27, we also identified a potential regulatory role for TF previously not linked to MM (e.g., CXXC1, ZNF394, and IRF3).

Finally, we explored the activity of the CCND2 super-enhancer in B cell chronic lymphocytic leukemia (CLL), the most common blood cancer (Supplementary Fig. 7a and b). High CCND2 expression has been previously documented in CLL B cells residing in proliferation centers28, structures in secondary lymphoid organs where malignant B cells receive survival and proliferative signals.28 CLL malignant B cells express significantly...
higher CCND2 than CCND1 levels (Supplementary Fig. 7b) and as in CCND2-high myeloma PC, we found that the same CCND2 super-enhancer is active in CLL B cells (Supplementary Fig. 7a). This observation extends the importance of the CCND2 enhancer in a wider range of B cell lineage malignancies.

**Discussion**

Our analysis of complementary datatypes sheds insights into how changes in the regulatory genome, and in particular of the MM-specific enhanceome, shape distinct myelomagenic transcriptomes and downstream biological pathways.
Fig. 4 TF regulatory gene networks provide biological and clinical insights into MM disease. **a** TF regulatory gene networks per myeloma subgroup and normal donor PC, as inferred from footprinting analysis. TF are weighed by relative TF binding frequency (colour) and expression (node size). **b** Heatmap of ranked expression of CXXC1, BPTF, MAZ, KLF13, CBFB and RFX5 across >1000 human cancer cell lines (CCLE dataset). Multiple myeloma cell lines are highlighted in red. Bar plot depicts the number of cell lines per cancer group. **c** MM patient stratification based on CXXC1, BPTF, MAZ, KLF13, CBFB and RFX5 expression (red, high; blue, low) and analysis of overall survival using the Multiple Myeloma Arkansas (n = 414) and the MMRF Compass Dataset (n = 745). HR hazard ratio.
Increased overall chromatin accessibility in all myeloma genetic subgroups with reference to normal BM PC, the myeloma PC normal counterparts, is our first fundamental observation. Unlike previous studies, here we addressed chromatin and gene expression-derived heterogeneity in MM. Surprisingly, we found that chromatin accessibility explained more of the variance than gene expression. The main axes of variation correlated surprisingly well with ND/MM and MIE categorizations, and defining subgroups as clusters, the combined model provided better separation than either data type alone. This was possible for two of the three IgH-translocated subgroups, and less so for HD MM, likely because the latter represents an inherently biologically heterogenous group. While the small number of samples in each subgroup argues for caution in interpreting this, we were able to validate these axes with independent data from cell lines not used in training the model. While using peaks found in the primary samples to assess cell lines is likely to overemphasise the similarity of cell lines and primary samples overall, it should not make specific cell line subtypes appear more like their primary counterparts. Larger future datasets will undoubtedly be helpful for increasing further our confidence in difference between the programs associated with different MIE. We found that marker genes...
and genesets that have been previously shown to be regulated by MIE, are also correlated with distal DAR of myeloma PC chromatin. This suggests that such chromatin changes are, to a large extent, directly or indirectly, dependent on MIE. Nevertheless, some of the identified putative enhancers are predicted to regulate genes downstream of the oncogenic RAS pathway which is activated by secondary gain-of-function N- or K-RAS somatic mutations in 40–50% of MM cases\(^3,19\), thus highlighting the ability of our approach to also reveal chromatin traces of secondary driver genetic events. Future studies, specifically designed to address the impact of pre-defined, high-frequency secondary genetic events (e.g., RAS mutations and MYC structural variants), will reveal the extent to which such secondary events impact chromatin and its regulatory activity. Another notable feature, mostly restricted to HD and MMSSET MM, is the proximity of putative enhancers to genes involved in neurogenesis with ectopic expression of the TF ISL2 in HD MM being a prime example of this. While the functional significance of this observation requires further investigation, we note that recent work demonstrated ectopic activation of neural genes in different cancers with such genes engaging in functional cross-talk with the central and peripheral nervous system\(^29,30\).

By taking advantage of a comprehensive set of chromatin marks in the whole spectrum of B lineage cells, we validated the presence of H3K27ac in 832 over-accessible, candidate enhancer regions in myeloma PC. The candidate enhancers of HGF and UCHL1 exemplify de novo formed enhancers, i.e., not present at any stage of late B lineage development, while the CCND2 enhancer provides an example of ‘re-commissioned’ super-enhancer, i.e., active in myeloma but not in normal PC. In the case of CCND2, ‘recommissioning’ entails activation in myeloma PC of Polycomb-imposed ‘poised’ transcriptional states in GCB cells and normal PC.

As well as discovery of critical cis regulatory elements, ATAC-seq also affords the opportunity for inferring TF binding to chromatin through footprinting and motif analysis. A notable finding is that the majority of trans factors that are predicted to display increased binding frequency to chromatin are already active in ND PC and their expression is not deregulated in MM. Such TF engage at higher frequency interactions with other TF than in ND PC and are more likely to self-regulate, properties that have been associated with enhanced regulatory potential\(^31\). These properties, as revealed through construction of TF regulatory gene networks, led to insights with functional and prognostic implications. In the case of MEF2C, the set of its predicted target TF, includes archetypical myeloma PC dependencies, i.e., IRF4, PRDM1, thus in part explaining the high degree of myeloma cell dependency on MEF2C. Similarly, myeloma cells appear to be addicted to CBF and ZNF384 with which they have not been linked to myeloma PC biology thus far. In addition, the case of CXXC1, BPTF, MAZ, KLF13, CBF and RFX5, six TF with unknown function in MM, also highlights the strength of our functional epigenomics approach. None of these TF is differentially expressed in MM, yet they display increased predicted regulatory potential across all myeloma subgroups, likely reflecting their high expression levels in myeloma cells compared to other cancers. Moreover, all six TF demonstrated prominent myeloma cell dependency and were found to strongly predict prognosis. For CXXC1, this is in accordance with its known function in the regulation of H3K4me3 as part of the COMPASS activating complex and through its binding to CpG islands and interaction with the histone methyl-transferases SETD1A/B\(^32,33\). Although little is known about the role of CXXC1 in cancer, it has been reported that its overexpression portends adverse prognosis in all stages of gastric cancer\(^34\). Therefore, the inferred TF networks provide a firm basis for future research which will further validate and define the role of TF identified herein in myeloma biology.

Extensive genetic heterogeneity and diversification in MM poses significant therapeutic challenges. One of the striking features of myeloma biology is the early observation that a dichotomous over-expression of the cell cycle regulators CCND1 and CCND2 overarches genetic diversification\(^4\). While in the majority of MM cases overexpression of CCND1 can be explained by somatic structural variants i.e., juxtaposition to IGH enhancer or chr11q25 gain\(^31\), such aberrancies do not account for over-expression of CCND2. Our identification and functional validation by complementary approaches of the distal cis and trans regulators of CCND2 expression addresses this gap in the biology of MM and allowed further insights into how the activity of this enhancer is regulated in different myeloma genetic subgroups. While expression of MAF is highest in the MAF subgroup as a result of its juxtaposition to the powerful IgH enhancer, in MMSSET MM expression of MAF is lower and previously shown to be regulated by the TF FOS in response to activated MAPK pathway\(^35\). It is likely that these differences in MAF dosage account for the strongest chromatin accessibility signal at the MAF-bound CCND2 enhancer and a higher level of CCND2 expression in MAF-translocated MM. Interestingly, the same enhancer with the same developmental chromatin features regulates expression of CCND2 in CLL cells, a finding that tallies with the notion that CLL originates either from a naïve or memory rather GCB cell\(^36\). Overall, our dissection of the MM regulatory genome offers insights and a resource for further biological exploration. Discovery and subsequent functional dissection of the critical CCND2 enhancer is a prime example of the power of our multi-layered integrative computational approach and affords opportunity for enhancer-based therapeutic approaches.
**Methodology**

**Ethics statement.** Bone marrow aspirates were obtained in accordance with the criteria of the Declaration of Helsinki and following written informed consent and national research ethics committee (REC reference: 11/H0308/9) and Imperial College London Joint Research Office approval.

**Patient and normal donor samples.** Patient BM aspirates were subjected to red cell lysis. Multiple myeloma plasma were purified by two rounds of CD138 immunomagnetic selection (Miltenyi Biotec) following the manufacturer’s instructions. Pre- and post-selection purity was assessed by FACS analysis (BD LS-R-Forresta, using CD138, CD45, CD19, CD56 and CD38 markers (Supplementary Table 1 and Supplementary Fig. 1a), at 1:100 antibody dilution Purified cells were immediately processed for ATAC-seq and RNA-seq.

Normal donor BM mononuclear cells from patients (BM-MNCs) were isolated using Dynal (Hamburg, Germany). The BM-MNCs were pre-cleaned of T cells and Monocytes by consecutive immunomagnetic negative selection (CD3 and CD14-EasySep StemCell Technologies) following the manufacturer’s instructions. The samples were stained (1:100 antibody dilution) and sorted for CD138+,-, CD19+,-, CD27+,-, CD45+,- and CD38+ (positive), for CD2-, CD3-, CD14-, CD16-, CD56+ (negative) and CD19+,- (positive or negative). (Supplementary Table 1 and Supplementary Fig. 1a) (FACSaria II, BD Biosciences). Sorted cells were immediately processed for ATAC-seq and RNA-seq. All FACS antibodies have been extensively validated by the manufacturer.

**Fluorescence in situ hybridisation (FISH).** FISH was undertaken using a panel of 4–7 probe sets targeting regions of common cytogenetic abnormalities in multiple myeloma (Cytogenetics, Amsterdam, The Netherlands). The cells were dropped onto a glass slide and dried briefly before fixation in situ. Hybridisation was performed according to the manufacturer’s protocols. The protocol consists of two probes (13q14 and 13qter) to detect deletion and monosity of chromosome 13, a locus specific probe to detect deletion of TP53 (17p13), two probes on chromosome 9 and 15 to detect HD and a dual colour, break-apart probe to detect rearrangements of IGH (14q32). Rearrangements of IGH were further investigated with IGH/CCND1, IGH/MAF and IGH/FGFR3 dual colour, dual fusion probes. The upper threshold for normal results is according to probe type (dual colour break apart 5%; quantitative 5%; dual colour dual fusion 2%). In all cases, a minimum of 50 interphase cells were scored by two independent analysts.

**Cell lines.** The human multiple myeloma cell line (MMCLs) MM.1S, NCI-H929 (ATCC, Manassas, VA, USA), U266, KMS12BM and OPM2 (DSMZ, Germany) were cultured in RPMI1640 media (Sigma, UK) and 20% FBS (Life Technologies). JN3 cells (DSMZ, Germany) were cultured in 40/40% DMEM/IMEM medium (Sigma, UK) and 20% FBS, HEK 293T (ATCC, Manassas, VA, USA) cells were cultured in DMEM (Sigma, UK), 10% FBS (Gibco). All cell lines were maintained at 37 °C and 5% CO2 and the growing media was supplemented with 1% penicillin/streptomycin (Sigma, UK) and 1% L-glutamine (Sigma, UK). Testing for mycoplasma presence was performed every 4 weeks.

**ATAC-seq.** ATAC-seq was performed on purified normal donor or myeloma patient samples. Briefly, 50,000 purified, myeloma plasma, myeloma or cell lines, were washed with cold PBS (Sigma, UK) at 500 x g for 4 min. The cells were resuspended in 50 μl of cold Lysis Buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPA-CA-630) and washed at 500 g at 4 °C for 10 min. The nuclei were subjected to transposable reaction for 30 min at 37 °C. 10 min NaCl, 3 mM MgCl2, 0.1% IGEPA-CA-630) and washed at 500 g at 4 °C for 10 min. The nuclei were subjected to transposable reaction for 30 min at 37 °C. The reaction and DNA purification was performed using a MiniElute Kit (Qiagen) and eluted twice with 30 μl of cold Lysis Buffer. The purified DNA was amplified as described before with NERS GPCR and 7 x 106 cells. The PCR amplified product was cleaned twice with (0.9X) AMPure beads (Beckman). The quality of the libraries was assessed with the Bioanalyzer High Sensitivity DNA kit (Agilent). The libraries were sequenced using the NEBNext Ultra II ChIP-seq library kit for Illumina (New England Biolabs) following the manufacturer’s protocols. The quality of the libraries was determined using the Qubit High Sensitivity DNA kit (Life Technologies) and library size was determined using the Bioanalyzer High Sensitivity DNA kit (Agilent). Libraries were diluted to 2 nM and sequenced using the Illumina HiSeq 4000 platform the Genomics Facility at ICL to obtain paired-end 75 bp reads.

**ChIP-seq.** For ChIP-seq, MM.1S cells were cultured using RPMI-1640 medium. 3–5 x 107 cells were pelleted by centrifugation at 300 x g, washed with PBS and crosslinked with 1% formaldehyde (Sigma, UK) for 15 min. Crosslinking was stopped by adding glycine to a final 125 mM. Cells were washed 3x with cold PBS. The cells were lysed (30 mM Tris-HCl pH8, 1% S, 10 mM EDTA (pH8), and 1% protease inhibitors (Sigma, UK) 20 min at 4 °C. Nuclear were sonicated at 4 °C in a sonicator UCD-200 (Diagenode).

Post-sonication fractions of average 500–300 bp length were confirmed on a 1.5% agarose gel. The chromatin was diluted at least 10 times in ChIP dilution buffer (0.01%SDS, 1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCL pH 8, 167 mM NaCl) containing 1x protease inhibitors (Sigma, UK). The diluted chromatin was pre-cleared by incubation with RSV-blocked magnetic beads (Dynabeads Protein A + G form Invitrogen) for 1 h at 4 °C on a rotating wheel. Pre-cleared chromatin was incubated with 2–5 μg of antibody, at an approximate dilution of 5:10000 (Supplementary Table 1), overnight at 4 °C on a rotating wheel. Protein A + G magnetic beads were added and incubated for 2–4 h at 4 °C on a rotating wheel. The immunoprecipitated complexes were washed for 5 min on a rotating wheel with 2X low salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8, 150 mM NaCl), 2X high salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8, 500 mM NaCl), 2X LiCl buffer (0.25 M LiCl, 1% IGEPAAK, 1% sodium deoxicholate, 1 mM EDTA, 10 mM Tris-HCL pH 8) and two times with TE.

Immunocomplexes were eluted by adding 150 μl elution buffer (50 mM Tris-HCl pH 8, 50 mM NaCl, 1 mM EDTA and freshly added 1% SDS and 20 mg/ml of RNaseA) at 65 °C for 1 hour to a overnight and a second time for 30 min. Both elutions were pooled and treated with proteinase K (Thermoscientific). DNA was purified using AMPure beads 1:1X ratio. ChIP and input DNA libraries were prepared using the NEBNext Ultra II ChIP-seq Library kit for Illumina (New England Biolabs) following the manufacturer’s protocols. The quality was determined using the Qubit High Sensitivity DNA kit (Life Technologies) and library size was determined using the Bioanalyzer High Sensitivity DNA kit (Agilent). The paired-end libraries were sequenced on the Illumina HiSeq 2500 platform to obtain single-end 50 bp reads. The H3K27Ac ChIP antibody has been extensively validated by the manufacturer. In addition, all ChIP antibodies are validated by ChIP qPCR, approved by the MAF antibody, cell lines not expressing the protein were used as additional negative controls.

**Cloning of dCas9-KRAB CRISPRi vectors.** Two single guide RNAs (sgRNAs) targeting each peak were designed using the online tool http://crispr.mit.edu (Supplementary Table 2). Two sgRNAs were designed for the CCND2 promoter as a positive control and the non-targeting sgRNA for GolS of the yeast S.cerevisiae as a negative control.

The vector used was the inducible Lentivector CRISPR-dCas9-KRABv2. The original Lentivector CRISPR vector (Addgene 52961, USA) was modified and kindly supplied by Dr Niklas Feldhahn (Imperial College London). The plasmid was digested with EcoRI (New England Biolabs). Removing a 2 kb stuffer.

The sgRNA oligos were phosphorylated and annealed using the T4 Ligation Buffer (New England Biolabs) and T4 P1NK (New England Biolabs). The sgRNAs were then ligated with the digested plasmid using the T4 DNA ligase and buffer (New England Biolabs). Reactions were carried out following the Zhang Lab General Cloning Protocol (Addgene).

Sanger sequencing was used to confirm the exact sequence of the cloned sgRNA (GeneWiz Ltd UK).

**Cloning of pLKO.1-GFP shRNA vectors.** The pLKO.1-GFP was obtained by replacing the selection marker puromycin with sgGFP cDNA in the pLKO.1-Puro lentiviral vector (Addgene plasmid 72994). Two independent shRNA constructs (Supplementary Table 2) were annealed and cloned into AgeI/EcoRI digested pLKO.1-GFP vector.

Successful cloning of shRNA sequences was confirmed by Sanger sequencing (GeneWiz Ltd UK).

**Lentivirus production and cell transduction.** The dCas9-KRAB and pLKO.1-GFP constructs were transfected in 293T cells with the 2nd generation lentiviral packaging and envelope plasmids pSPAX2 and pMD2.g (Addgene, USA), using CaCl2, 2.5 M (Sigma, UK) and 2x HEPEs-Buffered Saline, pH = 7 (Sigma, UK). After 10 h, cells were incubated with glycerol 15% v/v) (Honeywell, USA) for 3 min. Cells were washed with PBS (Sigma, UK) and then incubated with fresh DMEM media (Sigma, UK) for 36 h. The virus was harvested, filtered (0.45 μm) and concentrated by ultracentrifugation (23,000 rpm for 1 h and 40 min, at 4 °C) 48 h and 72 h post transfection using the Thermofall Ultraconfruge, MTX
ChiPs-qseq analysis. ChiP-seq data alignment was performed with BOWTIE2 using the default settings and duplicate reads in bam files were removed with Picard MarkDuplicates. Peak calling was performed using MACS2 with the following settings: for TFs, narrow peaks were detected using the following settings: \( -b \) \(-q\) 0.01--verbose \(-v\) 4--SPMR=call-summits; for histone-marks, broad peaks were obtained using the parameters: \( -b\)-broad--broad-cutoff 0.01-\(-v\) 4--SPMR. Genome browser tracks were generated using the deeptools command set, with the signal values normalized to fold change against input. Tools from Homer package (v4.9) were used for motif analysis, super-enhancer calling and annotation of genomic regions against the hg38 human genome, following the default mode.

RNA-seq analysis. Read cleaning and filtering—Paired-end RNA-seq reads were adapter trimmed by the DNA sequencing facility at Centre for Haematology, Division of Experimental Medicine Faculty of Medicine, Imperial London College were quality controlled with FastQC version 0.11.3. RNA-seq quantification—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. RNA-seq normalization and differential expression were performed using DESeq2, version 1.18.140. Transcript read counts were summarised to per-gene counts as the total in all transcripts of that gene (based on the transcriptome table created in Salmon). For MOFA analysis and visualization, read counts were used to create the input matrix for DESeq2. Samples were quality controlled with FastQC version 0.11.3. RNA-seq normalization—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. RNA-seq normalization and differential expression were performed using DESeq2, version 1.18.140. Transcript read counts were summarised to per-gene counts as the total in all transcripts of that gene (based on the transcriptome table created in Salmon). For MOFA analysis and visualization, read counts were used to create the input matrix for DESeq2. Samples were quality controlled with FastQC version 0.11.3. RNA-seq quantification—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. Salmon was used with fragment GC bias correction, 100 bootstrap samples and using an auxiliary k-mer hash over k-mers of length 31.

Correction of batch effects, normalisation and differential expression. RNA-seq normalisation and differential expression were performed using DESeq2, version 1.18.140. Transcript read counts were summarised to per-gene counts as the total in all transcripts of that gene (based on the transcriptome table created in Salmon). For MOFA analysis and visualization, read counts were used to create the input matrix for DESeq2. Samples were quality controlled with FastQC version 0.11.3. RNA-seq quantification—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. Salmon was used with fragment GC bias correction, 100 bootstrap samples and using an auxiliary k-mer hash over k-mers of length 31.

Correction of batch effects, normalisation and differential expression. RNA-seq normalisation and differential expression were performed using DESeq2, version 1.18.140. Transcript read counts were summarised to per-gene counts as the total in all transcripts of that gene (based on the transcriptome table created in Salmon). For MOFA analysis and visualization, read counts were used to create the input matrix for DESeq2. Samples were quality controlled with FastQC version 0.11.3. RNA-seq quantification—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. Salmon was used with fragment GC bias correction, 100 bootstrap samples and using an auxiliary k-mer hash over k-mers of length 31.

Correction of batch effects, normalisation and differential expression. RNA-seq normalisation and differential expression were performed using DESeq2, version 1.18.140. Transcript read counts were summarised to per-gene counts as the total in all transcripts of that gene (based on the transcriptome table created in Salmon). For MOFA analysis and visualization, read counts were used to create the input matrix for DESeq2. Samples were quality controlled with FastQC version 0.11.3. RNA-seq quantification—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. Salmon was used with fragment GC bias correction, 100 bootstrap samples and using an auxiliary k-mer hash over k-mers of length 31.

Correction of batch effects, normalisation and differential expression. RNA-seq normalisation and differential expression were performed using DESeq2, version 1.18.140. Transcript read counts were summarised to per-gene counts as the total in all transcripts of that gene (based on the transcriptome table created in Salmon). For MOFA analysis and visualization, read counts were used to create the input matrix for DESeq2. Samples were quality controlled with FastQC version 0.11.3. RNA-seq quantification—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. Salmon was used with fragment GC bias correction, 100 bootstrap samples and using an auxiliary k-mer hash over k-mers of length 31.

Correction of batch effects, normalisation and differential expression. RNA-seq normalisation and differential expression were performed using DESeq2, version 1.18.140. Transcript read counts were summarised to per-gene counts as the total in all transcripts of that gene (based on the transcriptome table created in Salmon). For MOFA analysis and visualization, read counts were used to create the input matrix for DESeq2. Samples were quality controlled with FastQC version 0.11.3. RNA-seq quantification—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. Salmon was used with fragment GC bias correction, 100 bootstrap samples and using an auxiliary k-mer hash over k-mers of length 31.

Correction of batch effects, normalisation and differential expression. RNA-seq normalisation and differential expression were performed using DESeq2, version 1.18.140. Transcript read counts were summarised to per-gene counts as the total in all transcripts of that gene (based on the transcriptome table created in Salmon). For MOFA analysis and visualization, read counts were used to create the input matrix for DESeq2. Samples were quality controlled with FastQC version 0.11.3. RNA-seq quantification—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. Salmon was used with fragment GC bias correction, 100 bootstrap samples and using an auxiliary k-mer hash over k-mers of length 31.

Correction of batch effects, normalisation and differential expression. RNA-seq normalisation and differential expression were performed using DESeq2, version 1.18.140. Transcript read counts were summarised to per-gene counts as the total in all transcripts of that gene (based on the transcriptome table created in Salmon). For MOFA analysis and visualization, read counts were used to create the input matrix for DESeq2. Samples were quality controlled with FastQC version 0.11.3. RNA-seq quantification—Expression estimates for each sample were obtained using Salmon version 0.11.439, using the complete Ensembl 85 transcriptome. Salmon was used with fragment GC bias correction, 100 bootstrap samples and using an auxiliary k-mer hash over k-mers of length 31.
calculated in the following way:

Extended shifted tags overlapping merged sample peaks
Total extended shifted tags

Annotations of the consensus peak regions—To annotate regions, the R library Annotatr version 1.8.049 was used, with hg38 annotations from the library TxDb.HsapiENS.UCSCHG38.knownGene (https://bioconductor.org/packages/3.9/data/annotation/html/TxDB.HsapiENS.UCSCHG38.knownGene.html). Any region that overlap multiple different types of genomic annotations on both strands but for each region, particularly peaks reported on. In addition to the Peak sets described above, we also obtained annotation for a random sample of sets using the function randomize_regions from the R library Annotatr version 1.8.049. The sample is taken without allowing overlaps and with per-chromosomal regions being maintained (non-alternative, random, unknown and mitochondrial chromosomes not used).

Quantitation, normalization and differentially accessible peaks—Shifted tags from each sample were transformed to obtain only the initial 1 bp (5’ end reflecting the TN5 DNA cleavage site) and filtered to remove tags overlapping with areas of low and high mappability defined previously. Tags were overlapped with the two peak sets (pan-myeloma peaks and sub-type peaks) using ‘bedtools intersect’ to generate tag counts in each sample for each of the two peak sets (Supplementary Fig. 8). The different tag counts in each peak set serve as the input data and are specified in each analysis type.

To account for batch effects, each sample was assigned to a batch based on sequencing run and sample preparation date. All samples from a batch that contained only a single sample passing quality control were assigned to a single batch and used as a reference level for batch effect removal (Supplementary Fig. 8). For MOFA analysis, inTAD analysis and visualization, the tag counts in pan-myeloma peaks from all primary samples (all 28 primary MM samples and the 5 ND samples Supplementary Data 1) were used. Tag counts were rlog transformed and normalized using the rlog function from DESeq2 and batch effects removed accounting for PC and MM subgroup effect (placing samples O1, O2, O3, O4 and O5 assigned to the Unknown MM subgroup and ND1–, ND1–, ND2–, ND2–, ND3+ to the ND subgroup, Supplementary Data 1) using the removeBatchEffect function from limma.

To verify the findings for the primary samples in our study from samples from derived cell lines and from Jin et al.13, we created rlog transformed and normalized matrices following the same procedure as above for our primary samples (using the pan myeloma peak set generated from our primary samples), up to the region, a part of the peak set, and then combined the new matrix with our primary sample and performed batch correction using a version of removeBatchEffects modified to use dummy contrast encoding rather than simple contrast encoding, such that our primary samples were not altered by the process and the other samples were corrected to match them. As the choice of reference point is arbitrary (grand average in simple contrast, and reference sample mean in dummy contrast), the only difference this change will make is to hold our samples constant. The relationship between samples is maintained in either contrast encoding scheme.

To identify differentially accessible peaks between normal and cancer samples, (referred to as pan myeloma differentially accessible peaks), the tag counts in pan myeloma samples were transformed and normalized using the function randomize_regions from the R library Annotatr version 1.8.049. The sample is taken without allowing overlaps and with per-chromosomal regions being maintained (non-alternative, random, unknown and mitochondrial chromosomes not used). The sample is taken without allowing overlaps and with per-chromosomal regions being maintained (non-alternative, random, unknown and mitochondrial chromosomes not used). The maximum distance allowed on the regulated genes was 1 Mb. TAD locations from the cell line GM1287813 were used. Peaks and genes were correlated using the Pearson correlation coefficient and p-values were corrected using the findCorrelation method from the inTAD package (adj.pval parameter set to TRUE). 1733 inTAD correlations were obtained between pan myeloma differentially accessible peaks and protein-coding pan myeloma DEG.

To identify likely target genes of ATAC regions selected for MOFA analysis, a similar process was followed, except that the 5000 ATAC regions were compared to all expression from all genes in the same TAD (not just those that were differentially expressed).

Modelling effect of differential promoters and candidate enhancers on gene expression. Differentially regulated promoters were obtained by intersecting the pan-myeloma DEG and pan-myeloma differentially accessible peaks set described above with the pan myeloma differentially accessible peaks set. To test this effect was limited to pairs within the same TAD, or was purely dependent on the distance between gene-region pairs, we divided all DAR-DEG pairs within 1 Mb into 10 bins (of 100 kb distance). Within each distance bin we picked an identical number of DAR-DEG pairs that were in the same TAD (inTAD) and that were not (out-TAD) to generate distance matched sets of DAR-DEG pairs either in or not in the same TAD. We then compared the average correlation of in-TAD and out-TAD DAR-DEG pairs using a Mann-Whitney U test. We repeated this process 100 times. In all 100 sets the correlation for pairs within TADs was significantly higher than those of distance matched pairs not in the same TAD (p < 0.05). We report the average p-value across these 100 matched sets.

To identify likely target genes of ATAC regions selected for MOFA analysis, a similar process was followed, except that the 5000 ATAC regions were compared to all expression from all genes in the same TAD (not just those that were differentially expressed).

Multi-omics factor analysis (MOFA). MOFA inputs—For ATAC-seq, the rLog transformed, normalized and batch corrected tag count for the pan-myeloma peak set were obtained for all samples (described above in the section Quantitation, normalization and differentially accessible peaks). Regions corresponding to sex chromosomes (chrX and chrY) and regions overlapping multi-exon TSS (see above) were removed. This yielded 273,216 remaining regions. The overall variance per peak (across the transformed counts for all 28 primary MM samples the 5 ND samples, excluding MMCL cell lines and ND3+, Supplementary Data 1) was calculated and only the 5000 peaks with the highest variance were selected.

For RNA-seq, the rLog transformed, normalized and batch corrected RNA-seq read counts for all samples were obtained (described above). Genes on sex chromosomes were removed from the table. The variance per gene was calculated and only the 5000 genes with the highest variance again using all primary patient samples were selected.

Training—Using the R library MOFAtools version 0.9.9950, the ATAC-seq and RNA-seq tables were input to MOFA. The default data and model options were
used and "Gaussian" data was selected both for ATAC-seq and RNA-seq. The training options used that were different to the defaults were: dropping factor threshold of 0.01, 10,000 maximum iterations, maximum tolerance convergence threshold of 0.01. Information on each samples' subgroup (cytogenetic MM subgroup, PC or "MM_OTHER" for MM samples with no cytogenetic information), condition (PC or MM) was not used by MOFA but was included for the analysis of the results. The variance explained was divided into 17 LF. The R script used to train the MOFA model can be found in:

```
scripts/train_MOFA.R
```

Models using different random initializations were trained and fit using only RNA-seq data, or both RNA-seq and ATAC data. Silhouette score for samples—For each model, each sample was assigned a label determined by its cytogenetic subgroup (including unknown cytogenetic samples, considered to be in the same group), the per sample silhouette score was obtained by calculating Euclidean distances between samples either all LFs as dimensions or only LF1 to LF5. This was done using the "silhouette" function from the "cluster" R package version 2.0.6. The mean silhouette score for each subgroup and model is shown.

Classification of samples using large cohort datasets. Additional molecular classification of samples used in this study was performed using previously identified, "gold standard" MM subgroup classifiers (Arkansas study, n = 414 patients). For this purpose, primary MM WGS and RNA-seq data from 892 patients were obtained from the MMRF CoMMpass study online portal (https://research.themmf.org). In addition, clinical annotation files containing the subgroup stratification of each sample based on SeqFISH genetic translocation analysis (WGS-based) and primary oncogenic markers overexpression calls (RNA-seq) were obtained from the same repository. Correlation analysis of samples used in this study and the MMRF cohort was performed based on the expression of the "gold standard" MM subgroup classifiers, using hclust (method: Pearson: linkage: complete) and visualized using pheatmap2 R packages.

Measuring discrimination of subtypes—For both the ATAC and RNA-seq consensus model and the RNA-seq only model, we calculated Fisher’s Discriminant Ratio for each LF as:

\[
S(X) = \frac{\sum_{g=1}^{G} n_g \sigma_{g,\text{within}}^2}{\sum_{g=1}^{G} n_g \sigma_{g,\text{between}}^2}
\]

where \( V_g \) is the between groups variance of \( X \) and \( V_w \) the within-group variance of \( X \) are calculated as:

\[
V_g = \frac{1}{N - G} \sum_{i=1}^{N} (X_i - \bar{X}_g)^2
\]

\[
V_w = \frac{1}{N - G} \sum_{g=1}^{G} \left[ \sum_{i=1}^{n_g} (X_i - \bar{X}_g)^2 + \sum_{i=1}^{n_g} (\bar{X}_g - \bar{X})^2 \right]
\]

where \( G \) is the number of subtypes, \( N \) the total number of samples and \( n_g, \bar{X}_g, \bar{X}_g \) are the number of samples, the mean and the variance of subgroup \( g \) respectively and \( \bar{X} \) is the grand mean. For each model we selected the LF that gave the greatest discriminant ratio. Linear discriminant analysis was performed on the first 5 LFs of each model using the 'lda' function from the R MASS package.

Chromatin state analysis. The Chromatin State Segmentations (12 states) by ChromHMM for the 173 cell types for the GRCh38 genome available to date in The DeepBlue Epigenomic Data Server (https://deepblue.mpi-inf.mpg.de/) were retrieved. Only cell types from the B-cell lineage were kept, in total, 19 samples. For the chromatin state heatmap, 1373 region–gene interactions were calculated using inTAD (see Correlation of gene-expression with peak accessibility in the same TAD section). Of these 1190 interactions were significant (adj.pval < 0.05), from 931 unique regions (i.e. some regions correlated with multiple genes). These pan-MM enhancers were divided into 200 bp windows. The chromatin state segmentations are already divided into the same 200 bp windows. Each 200 bp enhancer region was intersected with the cell states table to obtain the chromatin state for each cell type in that region using Bedtools intersect. For each candidate enhancer, a "predominant" state was calculated (i.e. the state which accounts for the largest footprint in the window). Enhancers were filtered, keeping only those where a predominant state was found in 15/19 cell chromHMM samples. For 99 of the 931 regions there was no predominant chromatin state in 16 or more samples leaving 832 useable candidate enhancer regions. The matrix of predominant states was hierarchically clustered using the Gower metric with the daisy function from the "hclust" package, using the ATACseq mode (-A). TF network visualization for each MM subgroup network was performed using Cytoscape 3.5 by weighing node size based on TF gene expression and node colour based on the adjusted relative frequency score. The NetworkAnalyzer tool, built in Cytoscape3.5 software, was used for network metrics analysis for each MM subgroup network. Auto-regulatory TF loops were defined as the cases where TFs are predicted to bind to and regulate their own gene.

Differential DNA footprinting on CND2-high versus CND2-low HD samples. Differential footprinting was performed in patients with HD myeloma with high
The molecular classification of multiple myeloma. Blood 108, 2020–2028 (2006).

Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014).

Okonechnikov, K., Erkek, S., Korbel, J. O., Pfister, S. M. & Chavez, L. InTAD: chromatin conformation guided analysis of enhancer target genes. BMC Bioinforma. 20, 60 (2019).

Angelaguet, R. et al. MOFA+; a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 21, 111 (2020).

Rampa, C. et al. Identification of the source of elevated hepatocyte growth factor levels in multiple myeloma patients. Biomark. Res. 2, 8 (2014).

Zhou, F., Meng, S., Song, H. & Claret, F. X. Dickkopf-1 is a key regulator of myeloma bone disease: opportunities and challenges for therapeutic intervention. Blood 127, 261–267 (2013).

Hussain, S., Bedekovics, T., Chesi, M., Bergsagel, P. L. & Galardy, P. J. UCHL1 is a biomarker of aggressive multiple myeloma required for disease progression. Oncotarget 6, 40704–40718 (2015).

Lohr, J. G. et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. Cancer Cell 25, 91–101 (2014).

Katoth, Y. & Katoth, M. Hedgehog signaling, epithelial-to-mesenchymal transition and miRNA (review). Int J. Mol. Med. 22, 271–275 (2008).

Hoffman, M. M. et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41, 827–841 (2013).

Stunnenberg, H. G., International Human Epigenome, C. & Hirst, M., The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell 167, 1897 (2016).

Doright, K. M. et al. MiR3 and MiR4 facilitate enhancer RNA synthesis and transcription from promoters independently of H3K4 monomethylation. Mol. Cell 66, 568–576.e564 (2017).

Buerronrostro, J. D., Giresi, P. G., Zaha, L. G., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).

Thaler, J. P. et al. A postmitotic role for Isl-class LIM homeodomain proteins in the ascertainment of visceral spinal motor neuron identity. Neuron 41, 337–350 (2004).

Hurt, E. M. et al. Overexpression of c-maf is a frequent oncogenic event in multiple myeloma that promotes proliferation and pathological interactions with bone marrow stroma. Cancer Cell 5, 191–199 (2004).

Shaffer, A. L. et al. IRF4 addiction in multiple myeloma. Nature 454, 226–231 (2008).

Igawa, T. et al. Cyclin D2 is overexpressed in proliferation centers of chronic lymphocytic leukemia/small lymphocytic lymphoma. Cancer Sci. 102, 2103–2107 (2011).

Magnon, C. et al. Autonomic nerve development contributes to prostate cancer progression. Science 341, 1236361 (2013).

Zhao, C. M. et al. Denervation suppresses gastric tumorigenesis. Sci. Transl. Med. 6, 250ra115 (2014).

Shah, V. et al. Prediction of outcome in newly diagnosed myeloma: a meta-analysis of the molecular profiles of 1905 trial patients. Leukemia 32, 102–110 (2018).

Lee, J. H. & Skalnik, D. G. Cpg-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. J. Biol. Chem. 280, 41725–41731 (2005).

Thomson, J. et al. Cpg islands influence chromatin structure via the Cpg-binding protein Cfp1. Nature 464, 1082–1086 (2010).

Sun, J. et al. The survival analysis and oncogenic effects of CFP1 and 14-3-3 expression on gastric cancer. Cancer Cell Int. 19, 225 (2019).

Annunziata, C. M. et al. A mechanistic rationale for MEK inhibitor therapy in myeloma based on blockade of MAF oncogene expression. Blood 117, 2396–2404 (2011).

Beekman, R. et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nat. Med. 24, 868–880 (2018).

Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr. Protoc. Mol. Biol. 109, 21 29–21 29 (2015).

Ponissusamy, K. et al. The innate sensor ZBP1-IRF3 axis regulates cell proliferation in multiple myeloma. Haematologica https://doi.org/10.3324/ haematol.2020.274480 (2021).

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017).

Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).

Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray data. Nucleic Acids Res. 43, e47 (2015).
Acknowledgements

N.T., V.C. and K.P. were supported by Blood Cancer UK. A.L.K was supported by Kay Kendall Leukaemia Fund and Imperial NIHR Biomedical Research Centre. I.M.S. was supported by BBSRC grant (BB/R007268/1). We also acknowledge support from Imperial NIHR Biomedical Research Centre, Cancer Research UK Experimental Cancer Medicine Centre, LMS/NIHR Flow-cytometry Facility, Imperial NIHR Biomedical Research Centre Genomics Facility.

Author contributions

J.A.B. analysed data. N.T. designed study, performed experiments and analysed data. A. Katsarou designed study, performed experiments and analysed data. A.C. analysed data, provided clinical samples. P.C.M. performed experiments. K.P. performed experiments. X.X. analysed data. M.B. provided clinical samples. M.A. provided clinical samples. I.A.G.R. wrote the manuscript. H.W.A. provided clinical samples. E.H. provided clinical samples. M.P. provided clinical samples. V.S.C. designed study, performed experiments. I.M.S. designed the study, performed and supervised analysis, wrote the manuscript. A. Karadimitris designed study, supervised experiments, analysed data, generated draft manuscript. All authors contributed to the final draft of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-25704-2.

Correspondence and requests for materials should be addressed to Valentina S. Caputo, Ian M. Sudbery or Anastasios Karadimitris.

Peer review information Nature Communications thanks Paola Neri and the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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) 2021