Genetic subtypes of smoldering multiple myeloma are associated with distinct pathogenic phenotypes and clinical outcomes

Mark Bustoros1,2,15, Shankara Anand3,4,15, Romanos Sklavenitis-Pistofidis1,3, Robert Redd5, Eileen M. Boyle6, Benny Zhitomirsky3, Andrew J. Dunford3, Yu-Tzu Tai1, Selina J. Chavda7, Cody Boehner1, Carl Jannes Neuse1,8, Mahshid Rahmat1, Ankit Dutta1, Tineke Casneuf9, Raluca Verona10, Efstatios Kastritis11, Lorenzo Trippa3, Chip Stewart3, Brian A. Walker12, Faith E. Davies6, Meletios-Athanasios Dimopoulos11, P. Leif Bergsagel13, Kwee Yong7, Gareth J. Morgan6, François Aguet3, Gad Getz3,14,15 & Irene M. Ghobrial1,3,15

Smoldering multiple myeloma (SMM) is a precursor condition of multiple myeloma (MM) with significant heterogeneity in disease progression. Existing clinical models of progression risk do not fully capture this heterogeneity. Here we integrate 42 genetic alterations from 214 SMM patients using unsupervised binary matrix factorization (BMF) clustering and identify six distinct genetic subtypes. These subtypes are differentially associated with established MM-related RNA signatures, oncogenic and immune transcriptional profiles, and evolving clinical biomarkers. Three genetic subtypes are associated with increased risk of progression to active MM in both the primary and validation cohorts, indicating they can be used to better predict high and low-risk patients within the currently used clinical risk stratification models.
Multiple Myeloma (MM) is an incurable plasma cell malignancy with significant inter- and intra-patient heterogeneity. It is almost always preceded by the asymptomatic precursor stages monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM). Approximately 1.5% of MGUS patients will progress to MM per year, while SMM patients have a higher overall progression risk of 10% per year. Like MM, SMM is a heterogeneous condition — some patients have over a 50% risk of progression within two years, while others have more MGUS-like disease that grows slowly.

Several risk stratification models exist to help clinicians differentiate patients with high risk of progression to active myeloma from those for whom a “watchful waiting” approach is appropriate. The existing models rely solely on clinical measurements, many of which are indicators of tumor burden and universal biomarkers of MM for risk stratification. These models, however, do not fully partition progressors from non-progressors, and patients classified as low- or intermediate-risk still progress to active MM and have a 2-year progression risk of 6% and 18%, respectively (compared to 44% for high-risk patients), which warrants more accurate models that also represent the molecular heterogeneity in MM. We recently showed that genomic alterations that tend to occur together, and show that the tumors of this cluster exhibited a hyperdiploid genotype as the primary event and were significantly enriched in Nras, Traf3, and Max mutations. We named this cluster Hyperdiploid-like 1 (HL1). Cluster 2: the tumors of this cluster harbored frequent arm-level deletions, including 16q, 6p, 1p, 17p, 4q, 18q, and 20q, and the IgH translocation t(14;20), which upregulates the transcription factor MAF. Moreover, mutations in Nras, Braf, Tps53, Atm, Mafb, and Cdkn2c genes were enriched in this subgroup. Hyperdiploidy was detected in 59% of the tumors in this cluster. We named this cluster Hyperdiploid-like 2 (HL2). The tumors of this cluster were significantly enriched in deletion (16q), which involves CYLD tumor suppressor and other genes. The presence of both hyperdiploidy and t(14;20) in the same cluster could be explained by the co-occurrence of those events as described in prior studies. Indeed, half of patients with t(14;20) also had hyperdiploidy in their tumor samples. Cluster 3: Tumors of this cluster exhibited primary events such as t(4;14), which upregulates Fgfr3 and Mmset genes; t(14;16), which upregulates the transcription factor MAF; and copy number losses of 14q, 1p, 8p, 10p, 11q, 12p, and 17p. We named this cluster Translocation-like 1 (TL1). This cluster was also enriched for hypodiploid tumors, defined as having fewer than 45 chromosomes (adjusted P = 0.04). Tumors in this cluster also harbored mutations in DIS3, MAF, Fgfr3, Ptk2d, Prdm1, and Hst1h1e1. Many of these proteins and mutations in their encoding genes are essential to tumor cell survival and play roles in protein translation, secretion, and plasma cell differentiation. Indeed, differential gene expression analysis revealed that Tl1 tumors have downregulation of ribosomal proteins and the negative regulator of the MAPK pathway TRAF2. The upregulated genes included Whsc1(Mmset), Fgfr3, Klhl4, Ccnd3, and genes involved in the endoplasmic reticulum (ER) stress response (Fig. 2A). Cluster 4: this cluster comprises tumors with a hyperdiploid genotype that harbored mutations in KRAS and Nfkb1a genes, and MYC translocations as the only significant features. We named this cluster Hyperdiploid-like 3 (HL3). Cluster 5: the tumors in this cluster mainly exhibited t(11;14), Ccnd1 mutations, and gain of chromosome 11 or its long arm. We named this cluster Translocation-like 2 (TL2). Interestingly, this cluster had significantly lower M-protein levels and was enriched in light-chain disease compared to the other clusters (P < 0.001 for both). Tumors of TL2 had 243 differentially expressed genes (q < 0.1, log2FC > 1.5; 180 upregulated, 63 downregulated), including overexpression of Ccnd1, Erbb4, E2f7, E2f1, Trak2, Rbl1, and downregulation of Dusp4, Traf6, Prkd3, Ccdden, and Znfl844. Furthermore, this cluster had the highest expression of Ccnd1 compared to the other clusters (Fig. 2B). Cluster 6: this is a hyperdiploid cluster similar to HL1; however, its tumors are also enriched in Nfkb2 and Klhl6 mutations and exhibit copy gains in 2p. Interestingly, copy gains of 1q were more frequent in this cluster than HL1 and the other hyperdiploid clusters (P < 0.001 for both comparisons). We named this cluster Hyperdiploid-like 4 (HL4). Additionally, key individual genes in
myeloma pathogenesis were overexpressed in tumors of specific genetic subtypes. MCL1 was upregulated in all the genetic subgroups with the lowest expression observed in HL1 compared to the other subtypes (<i>P</i> = 0.001) (Fig. 2C). MYC oncogene was also highly expressed in the four hyperdiploid clusters (<i>P</i> = 0.009, Wilcoxon Test) (Fig. 2D). Cyclin D1 (CCND1) was significantly upregulated in TL2 tumors (<i>P</i> = 0.0001), while CCND2 was upregulated in TL1 and HL2 tumors compared to the rest of the genetic subtypes (<i>P</i> = 0.004) (Fig. 2D, Supplementary Fig. 2A, B). Moreover, in the four hyperdiploid clusters, we found that CCND2 expression was higher in samples without 11q gain, while CCND1 expression was higher in tumors with 11q gain (Supplementary Fig. 2D, G).

The genetic subgroups are enriched with specific MM expression signatures. To date, ten distinct RNA expression signatures have been defined and validated as prognostic in newly diagnosed and relapsed MM patients<sup>16,17</sup>. Each expression signature was then associated with specific primary genetic lesions identified by fluorescent in situ hybridization (FISH), including hyperdiploidy and IgH translocations that activate c-MAF and MAFB, CCND1, CCND3, or MMSET<sup>16,17</sup>. We asked whether these expression signatures were present in our SMM cohort and correlated with the six genetic subgroups. To address this, we performed a gene-set enrichment analysis of these expression signatures among the genetic subtypes (lower panel of Fig. 2G). We observed that the hyperdiploid expression signature<sup>16,17</sup>, which is seen in hyperdiploid MM patients, is upregulated in the tumors of our hyperdiploid clusters (HL1–4). The Cyclin D (CD) expression signatures, including CD-1 that highly expresses CCND1 and CD-2, which expresses the B cell markers CD20, CD79A, and CCND1, were significantly upregulated in the TL2 genetic subgroup. Moreover, the high-risk MMSET (MS) molecular signature, which is enriched in patients with t(4;14) and upregulates MMSET and FGFR3 genes, was upregulated in the TL1 cluster. The MAF (MF) signature, which has been reported in patients with t(14;16) and t(14;20) that upregulate MAF and MAFB genes, respectively, was enriched in both the TL1 and HL2 subgroups, consistent with the presence of these genetic alterations in their tumors. The low bone disease signature, which has not been previously mapped to a specific MM genetic alteration, was upregulated in the HL4, TL1, and HL2 subgroups, consistent with the presence of these genetic alterations in their tumors. The low bone disease signature, which has not been previously mapped to a specific MM genetic alteration, was upregulated in the HL4, TL1, and HL2 subgroups, suggesting it could be linked to 1q gain, which occurs frequently in these three subgroups. Interestingly, the PR signature, which is found in proliferative tumor cells, was enriched in the HL3 and TL2 subgroups. Furthermore, the NFκB signature was upregulated only in HL2, which could be explained by the high frequency of 16q deletions and CYLD mutations in this subgroup. Finally, the
PRL3 signature, which overexpresses the protein tyrosine phosphatase PTP4A3 and 27 additional genes, was upregulated only in HL4 and indicates that it could also be linked to the presence of 1q gain, which is found in all the tumors of the HL4 subgroup.

We further examined whether our genetic subtypes were enriched in specific mutational signatures for 72 samples with matched normal whole-exome sequencing. We found that the APOBEC mutational signature activity (SBS 2,13 COSMIC v3.0) differed between the genetic subtypes (P = 0.027, Kruskal-Wallis) while AID mutational signatures did not (P = 0.17) (Supplementary Fig. 2E–G). Specifically, we found APOBEC activity enriched in the HL2 & TL1 clusters vs. the rest of tumors (P = 0.006, adjusted p = 0.045) (Supplementary Fig. 2H).

Genetic subgroups have distinct transcriptional profiles. We performed GSEA on the available transcriptomic resulting data to explore which genes and biological pathways were differentially expressed among the genetic subgroups we identified. Pathways that were significantly enriched within the six genetic subtypes are described and illustrated (Fig. 2G). We found that protein secretion, unfolded protein response (UPR), glycolysis, hypoxia, and mTOR signatures were specifically enriched in the TL1 subgroup, while E2F target genes, cell cycle, heme metabolism, complement, and proliferation signatures were enriched in TL2 tumors. Genes induced by MYC were highly expressed in HL3 and HL4, consistent with MYC upregulation in these two clusters. The NFkB, cytosolic DNA sensing, and JAK-STAT signatures were enriched in the tumors of HL2. The interferon-alpha and gamma response signatures were high in TL2 but low in HL2. Interestingly, oxidative phosphorylation, WNT-beta-catenin, and TGF-beta signaling were enriched only in tumors of HL4, and the TNFa and inflammatory signatures were uniquely enriched in HL3. The ribosome biosynthesis signature was low in TL1, TL2, and HL3 but high in HL1, HL2, and HL4 subgroups.

We also looked at signatures related to the tumor immune microenvironment. Signatures of regulatory T cells and NK cells were high in HL2 and HL3, while the M2 macrophage signature was high in TL2 and HL4 tumors. The HL3 and TL2 tumors were enriched for the monocyte signature. In contrast, the signature of
plasmacytoid dendritic cells, known for their immunosuppressive effect, was only enriched in the TL1 tumors.

Genetic subtypes are differentially associated with risk of progression and evolving clinical biomarkers. To investigate the relationship between these genetic subtypes and clinical outcome, we analyzed a subset of patients (n = 87) who were followed for the natural course of their disease and did not receive any treatment in a clinical trial setting before progression to MM. Their baseline characteristics are reported in Supplementary Table 2. The median follow-up time for these patients was 7.1 years and the median time to progression (TTP) was 4 years (95% CI, 3–6). In this cohort, 57 patients (66%) have progressed, while 30 (34%) remained asymptomatic as of the last follow-up (put date of last follow-up in the methods section). The genetic subgroups had different outcomes, measured by TTP to active MM (log-rank P = 0.007) (Supplementary Fig. 3A). Median TTP for patients in HL2, TL1, and HL3 was 3.7, 2.6, and 2.2 years, respectively, while it was 4.3, 11, and 5.2 years for HL1, TL2, and HL4, respectively. The HL2, TL1, and HL3 genetic subgroups had increased hazards of progression (HR > 4.5) to active myeloma (Supplementary Fig. 3B).

We then divided the genetic subtypes based on their TTP and hazards of progression into high- (HL2, TL1, HL3), intermediate- (HL1, HL4), and low-risk (TL2) subtypes. The high- and intermediate-risk subtypes had significantly shorter TTP and increased risk of progression compared to the low-risk subtype (2.6 and 5.2 vs. 11 years, respectively, P < 0.0001) (Fig. 3A). We also stratified the patients according to the 20-2-20 model, which uses three cutoffs of M-protein > 2 g/dL, FLC ratio > 20, and bone marrow plasmacytosis > 20% to define low, intermediate, and high-risk groups based on the presence of none, one, and two or all these variables, respectively. The intermediate- and high-risk genetic subtypes and the clinically high-risk SMM group (according to the 20-2-20 model) were the only significant predictors of progression to active MM in our multivariate analysis (Fig. 3B).

Moreover, the prediction performance of the combined clinical and genetic models was higher than the clinical only models (Fig. 3B). We also obtained another smaller cohort of 67 patients with targeted capture data, including common MM translocations, CNAs and SNVs, and added it to the previous cohort. In those 142 patients, HL2, TL1, HL3, and HL4 subtypes were independent predictors of progression to active myeloma (Fig. 3D) and the high-risk genetic subtypes were associated with increased risk of progression in multivariate analysis (HR: 3.4, 95% CI: 1.68–6.7). We then asked, given the small number of patients in the different cohorts, whether combining the three cohorts would provide more power and increase the significance of our genetic classification. The combined cohort contained 229 SMM patients with median follow-up and TTP of 6.9 and 5.2 years, respectively. Indeed, the genetic subtypes had a different TTP (Fig. 3F), and the high-risk genetic subtypes had significantly shorter TTP compared to the low or the intermediate-risk groups (Fig. 3F). We also found that both the individual genetic subtypes and the genetic risk groups were independent predictors of progression in the combined cohort multivariate analysis, validating our initial findings (Fig. 3G). Interestingly, within the high-risk clinical stage, patients in the low-risk genetic subgroups had significantly lower progression risk (HR 0.26, 95% CI: 0.1–0.6, P = 0.001) and median TTP of 8.7 years (log-rank P = 0.002) (Supplementary Fig. 5A). In the intermediate-risk clinical group, patients from the high-risk genetic groups had increased risk of progression to symptomatic MM (HR 4.4, 95% CI: 1.7–11.6, P = 0.002) and shorter TTP (3 vs 6.9 and 9.4 years, respectively, log-rank P = 0.001) (Supplementary Fig. 5B).

Validation of the molecular subtypes in external cohorts. To validate our findings on the clinical significance of the genetic subtypes, we developed a classifier based on the features of the clusters we identified in our primary cohort (Supplementary Fig. 4). We used an external cohort of 75 SMM patients to validate the classifier and investigate whether the genetic subtypes are predictive for progression. The patients in this cohort were enriched in the low-risk clinical stage and had a median TTP of 5 years. Similar to the primary cohort, patients in the intermediate and high-risk genetic subtypes had increased risk of progression to active MM in multivariate analysis accounting for the clinical risk stage (HR: 4.5 and 9, P = 0.039 and 0.002, respectively) (Fig. 3C). We found that adding the genetic risk groups improved the prediction of progression compared to the clinical model only (C-index: 0.76 vs 0.65, respectively) (Supplementary Table 4).

Discussion
This study modeled the genetic heterogeneity seen in SMM by identifying genetic subtypes that correspond to phenotypic attributes and clinical outcomes, providing a deeper understanding of SMM pathogenesis. We and others have previously cataloged individual driver genetic aberrations in SMM and MM cohorts. However, the present study expands on this work and identifies SMM genetic subtypes defined by multiple recurrent DNA genetic aberrations, unlike previous classification efforts that were mainly based on gene expression data. Our findings suggest that these genetic subtypes could have distinct evolutionary histories depending on the initiating genetic events (translocations or CNAs), which may influence the subsequent acquisition of cooperating genetic aberrations.

The defined genetic subtypes had distinct clinical outcomes of disease progression into symptomatic MM, which could provide us with comprehensive molecular models for predicting progression and dynamic changes in clinical biomarkers over time. They also have specific dysregulated molecular and oncogenic pathways, which could facilitate the identification of specific targets and selection of therapies for each genetic subtype.
empower precision medicine efforts, much like the specificity of venetoclax in patients with t(11;14)20,21.

We identified six clusters based on the detected genetic alterations. We divided them into three high-risk (HL2, TL1, HL3), two intermediate-risk (HL1, HL4), and one low-risk (T2) genetic group based on progression risk to active MM. We found that DNA repair aberrations were exclusive to HL2 and TL1 subgroups, which were enriched in TP53 mutations and deletions.

Also, MYC expression was higher in the hyperdiploid subgroups than the non-hyperdiploid ones, consistent with previous reports of a higher frequency of MYC alterations in hyperdiploid MM patients22. The key Cyclin D genes, CCND1 and CCND2, were highly expressed in TL2 and TL1, respectively. CCND1 and CCND2 expression patterns were previously reported to distinguish between MM patients hyperdiploid tumor samples23; indeed, in the four hyperdiploid clusters, we found the former to
be enriched in tumors with 11q gain, while the latter is highly expressed in tumors without 11q gain. However, we could not assess their prognostic impact due to the small number of samples with gene expression data in patients who were followed for their disease course.

The gene expression signatures of specific molecular and oncogenic processes also varied significantly between the genetic subgroups. For example, TL1 tumors showed specific enrichment for protein secretion, ER stress, UPR, glycolysis, and mTOR signaling. This molecular phenotype manifested clinically with patients with this genetic subtype had the highest increase in M-protein levels at six and twelve months from diagnosis. Such patients may benefit from drugs inducing cellular stress, such as proteasome inhibitors or novel molecules targeting the ER stress and UPR pathways. Alternatively, TL2 tumors were uniquely enriched with genes related to B-lymphocytes, cell cycle, heme metabolism, and complement activation signaling. Clinically, these patients had the longest TTP, lowest baseline M-protein level, and the least increase over time. We also found that the HL2 tumors were enriched for interferon-alpha response, cytosolic DNA sensing, and JAK-STAT signatures. These results underscore the phenotypic difference among the genetic subtypes and provide a conceptual framework for future functional studies that aim to validate or therapeutically target the dysregulated pathways and tumor dependencies in different genetic subtypes.

In our multicenter cohort, we found that the three high-risk subgroups (HL2, TL1, and HL3) had an increased risk of progression and were associated with evolving hemoglobin and M-protein levels, showing that these subgroups are also predictive of the dynamic changes in MM clinical biomarkers over time. The high-risk genetic subtypes were independent risk factors of progression to overt MM after accounting for the clinical risk stage by the 20-2-20 model. Moreover, among those patients considered high- and intermediate-risk by this model, those with the high-risk genetic subtypes and clinical risk stages according to the IMWG 20/2/20 model were independent risk factors of progression and were associated with evolving hemoglobin and SMM stage as we haven’t tested its prognostic significance in active or relapsed MM settings.

In conclusion, these findings move us closer to identifying the SMM patients who are truly at a high risk of disease progression through better predictive models that integrate the molecular makeup of the tumor cells and may also guide precision medicine efforts to match targeted therapies with the optimal patient groups in multiple myeloma and its asymptomatic stages.

### Methods

#### Patient samples.

We used next-generation sequencing technologies to study 214 patients with SMM at the time of diagnosis. We performed whole exome sequencing (WES) of 72 matched tumor-normal samples (mean target coverage 174×), and targeted deep sequencing on 48 samples (mean target coverage 774×). FISH data were used to determine the presence of 1q42 translocations. Samples were collected at Dana-Farber Cancer Institute, University College London, and the University of Athens in Greece, in addition to diagnostic samples from patients participating in phase II clinical trial for treating patients with SMM (NCT02316096). Patients who presented with MM symptoms at diagnosis, including hypercalcemia, renal impairment, anemia, or bone lytic lesions (CRAB), or had any myeloma-defining event were excluded from the analysis.

All samples were obtained after approval of the study protocols by the institutional review boards and ethics committees of the participating institutions including Dana-Farber Cancer Institute, University College London, and the University of Athens in Greece, and participating institutions of the above-mentioned clinical trial, and written informed consent from patients. All relevant ethical regulations were followed, and all the research was conducted in accordance with the Declaration of Helsinki.

#### Whole exome sequencing.

Tumor DNA was extracted from CD138+ cells from patients’ bone marrow samples. For germline control (normal), DNA was obtained from buccal mucosa (saliva), or peripheral blood mononuclear cells. Genomic DNA was extracted using QIAamp DNA mini kit (QIAGEN) according to the manufacturer’s protocols, and double-stranded DNA concentration was quantified using PicoGreen dsDNA Assay kit (Life Technologies). Libraries were prepared by Agilent SureSelect XT2 Target Enrichment kit. To capture the coding regions, we used the SureSelect XT2 V5 + UTR capture probes (Agilent). All sequencing was performed on the Illumina HiSeq 4000 platform at the Broad Institute. For tumor
only samples \((n = 94)\), libraries were prepared and hybridized using Agilent SureSelect XT2 V5 capture probes (Agilent) and sequenced on Illumina HiSeq 2500 platform.

**Targeted deep sequencing.** Genomic DNA was extracted using QIAamp DNA micro kit (QIAGEN) according to the manufacturer’s protocol. The libraries for targeted sequencing were prepared using SureSelect XT Reagent Kits (Agilent), and an in-house bait set targeting 117 genes, including pan-cancer driver genes and frequently mutated genes in MM. The libraries were quantified using Agilent Tapestation, then pooled and loaded onto the Illumina HiSeq 4000 sequencer.

**Computational analysis.** The output from Illumina software was processed by the Picard data processing pipeline to yield BAM files containing well-calibrated, aligned reads. We have utilized the Broad Institute and the Getz Lab CGA WES Characterization pipeline (https://getzlab.slack.com/archives/DHC9613KQ/p164754517601073289) to call, filter, and annotate somatic mutations and copy number variation. The pipeline employs the following tools: MuTect2,26 ContEst27,8 Orientation Bias Filter28 DeTN29 AllelicCapSeg29, MAFPoNFilt29, ABSOLUTE30,31 GATK32, PicardTools33, Variant Effect Predictor34, Oncotator35. We applied ABSOLUTE to estimate sample purity, ploidy, and absolute somatic copy numbers. These were used to infer the cancer cell fraction (CCF) of point mutations from the WES data, following the framework previously described35. Mutations were thereafter classified as clonal based on the posterior probability that the CCF exceeded 0.90 and subclonal if otherwise.

**Germline filtering of tumor-only cohort.** For each SNP or indel that passed all standard filters, ploidy, and_copy number VAR filter, we used the Germline Somatic Log odds filter for common germline mutations and artifacts and filters for OxoG and FFPE damage36.

**Paired tumor-normal cohort.** We applied ABSOLUTE to estimate sample purity, ploidy, and absolute somatic copy numbers. These were used to infer the CCFs of point mutations from the WES data. We excluded 13 samples from this group due to low tumor fraction (\(\leq 20\%\)) and inconclusive FISH results. Bleed-through error associated with sequencing was observed and cleaned using a custom PoN filter run through the same sequencing pipeline, as described previously36. Two more artifacts were identified in this cohort, primarily characterized by C>A and C>T substitutions, respectively, henceforth referred to as A1 and A2. Artifact A1 was shown to represent reference bias, with a preponderance of C>A over G>T substitutions, while artifact A2 occurred using DNA library preparation, as previously described4. To address this, we developed a tool that removes C>A SNPs with a low number of reads supporting the alternate allele from a sample, until the p-value of a binomial test assuming a probability of 0.5 exceeds 0.1. Of unidentified origin, artifact A2 was characterized by a preponderance of C>T SNPs in the GCC trinucleotide context over COSMIC signature 5 and was addressed by removing C>T SNPs in the GCC context with low alternate allele counts until they matched the number of C>T SNPs in the CCG context, assuming reference COSMIC signature 5 as a null. Of note, these artifacts did not affect any of the SNPs reported in genes that are frequently mutated in MM.

**Tumor-only cohort.** In this cohort, we observed an artifact of unidentified origin that was primarily characterized by T>G substitutions (Supplementary figure, panel C) and hotspot mutations in genes not reported before in multiple myeloma, henceforth referred to as A3. We addressed it by removing all hotspot mutations with less than 2 occurrences in COSMIC and do not affect genes reported to be recurrently mutated in multiple myeloma. Of note, this artifact did not affect any of the SNPs reported infrequently mutated genes in MM.

**Bulk RNA-sequencing.** Out of the 214 unique patient tumor DNA samples, there were 89 matching tumor samples material for RNA sequencing. These samples were isolated using Qiagen RNA kit. Libraries were prepared using Illumina Total mRNA kit and submitted for sequencing on HiSeq 2500 machines. We computationally processed these RNA samples using the GTEX V8 pipeline and aligned them to Hg19 Genome v19.39. Data quality control metrics are provided in Supplementary figure 6.
patients. We applied the classifier to these cohorts to identify the genetic subtypes that each patient sample belongs to and perform subsequent analysis.

Statistical analysis. Binary outcomes were reported as proportions with 95% exact binomial confidence intervals. Continuous measures were summarized as median and range. Binary outcomes and other categorical variables were tested for association with continuous and other categorical variables using Wilcoxon rank-sum (or Kruskal-Wallis for three or more groups) or Fisher’s exact tests, respectively. Time-to-event endpoints are estimated using the method of Kaplan and Meier, with 95% confidence intervals calculated using Greenwood’s method of variance estimation. Differences in survival curves were assessed using log-rank tests. Median follow-up was calculated using the reverse Kaplan-Meier method. Unadjusted and adjusted Cox modeling was performed to assess the impact of the presence of a MM driver on clinical outcome measures, alone and in the presence of clinical features known to impact outcome. Time to progression (TTP) was measured from date of diagnosis to date of documented progression to MM. Clinical and laboratory parameters and genetic features were reported with hazard ratios and 95% confidence intervals with Wald p values, while genetic features were assessed for importance of association with TTP. For comparison of clinical model only vs the clinical and genetic models, we used analysis of variance test on the Cox model with the selected variables. A global assessment of each model was also assessed using a C-statistic for censored survival data. The statistic for each time-to-event model is reported with a 95% confidence interval. Values range between 0.5 to 1 indicating a poor to perfect model; nested models may be compared via overlap in the point estimates and confidence intervals. Genetic alterations that were positively associated with each cluster were identified by a one-sided Fisher test and ranked by significance of Benjamin–Hochberg adjusted p value. All other P values in the study were two-sided, and adjustment for multiple hypothesis testing was performed using the method of Benjamini and Hochberg. P and q value thresholds for significance were set at 0.05 and 0.1, respectively. Statistical analyses were performed using R version 3.6.0 (2019-04-26).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. The Genomic and transcriptomic data of the primary cohort generated in this study (including the whole exome, targeted capture, and RNA sequencing data) have been deposited in the dbGaP database under accession number phs001323.v3.p1. Access to the raw data can be obtained upon request. The other published data used as validations cohorts in this study are already deposited in public databases. For the first validation cohort, the targeted panel data are deposited in the European Genome-phenome Archive (EGA) database under accession code EGAD000000005856. The whole-exome sequencing is deposited in the EGA database under accession code EGAD000000005285. These data are available under restricted access; access can be obtained upon request. The raw data of the published second validation cohort is deposited in the NCBI Sequence Read Archive (SRA) BioProject under accession number PRJNA541307. The remaining data are available within the Article or Supplementary Information file.

Code availability. The code for the BMF consensus clustering and the subsequent analysis in the primary and validation cohorts is available through GitHub at https://github.com/getzlabs/MMM-clustering_2020.

Received: 6 May 2021; Accepted: 13 May 2022;
Published online: 15 June 2022

References.

1. Kyle, R. A. et al. Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma. N. Engl. J. Med. 356, 2582–2590 (2007).
2. Kyle, R. A. et al. A long-term study of prognosis in monoclonal gammopathy of undetermined significance. N. Engl. J. Med. 346, 564–569 (2002).
3. Mateos, M. V. et al. International Myeloma Working Group risk stratification model for smouldering multiple myeloma (SMM). Blood Cancer J. 10, 102 (2020).
4. Bustoros, M. et al. Genomic profiling of smouldering multiple myeloma identifies patients at a high risk of disease progression. J. Clin. Oncol. 38, 2380–2389 (2020).
5. Bolli, N. et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nat. Commun. 5, 2997 (2014).
6. Walker, B. A. et al. Mutational spectrum, copy number changes, and outcome: results of a sequencing study of patients with newly diagnosed myeloma. J. Clin. Oncol. 33, 3911–3920 (2015).
7. Walker, B. A. et al. Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma. Blood 132, 587–597 (2018).
8. Lustad, E. H. et al. Revealing the impact of structural variants in multiple myeloma. Blood Cancer Discov. 1, 258–273 (2020).
9. Lobo, J. G. et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. Cancer Cell 25, 91–101 (2014).
10. Mura, F. et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. Nat. Commun. 10, 3835 (2019).
11. Boyle, E. M. et al. The molecular make up of smoldering myeloma highlights the evolutionary pathways leading to multiple myeloma. Nat. Commun. 12, 1204 (2021).
12. Misund, K. et al. MYC dysregulation in the progression of multiple myeloma. Leukemia 34, 322–326 (2020).
13. Chapuy, B. et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nat. Med. 24, 679–680 (2018).
14. Žitnik, M. & Žapan, B. NIMFA: a python library for nonnegative matrix factorization. J. Mach. Learn. Res. 13, 849–853 (2012).
15. Chapman, M. A. et al. Initial genome sequencing and analysis of multiple myeloma. Nature 471, 467–472 (2011).
16. Broyl, A. et al. Gene expression profiling for molecular classification of smouldering multiple myeloma in newly diagnosed patients. Blood 116, 2543–2553 (2010).
17. Zhan, F. et al. The molecular classification of multiple myeloma. Blood 108, 2020–2028 (2006).
18. Chauhan, D. et al. Functional interaction of plasmacytoid dendritic cells with multiple myeloma cells: a therapeutic target. Cancer Cell 16, 309–323 (2009).
19. Ravi, P. et al. Evolving changes in disease biomarkers and risk of early progression in smouldering multiple myeloma. Blood Cancer J. 6, e54 (2016).
20. Kumar, S. K. et al. Venetoclax or placebo in combination with bortezomib and dexmethasone in patients with relapsed or refractory multiple myeloma (BELLINI): a randomised, double-blind, multicentre, phase 3 trial. Lancet Oncol. 21, 1630–1642 (2020).
21. Moreau, P. et al. Promising efficacy and acceptable safety of venetoclax plus bortezomib and dexmethasone in relapsed/refractory MM. Blood 130, 2392–2400 (2017).
22. Walker, B. A. et al. APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. Nat. Commun. 6, 6997 (2015).
23. Bergpael, P. L. et al. Cyclin D dysregulation: an early and unifying pathogenic event in multiple myeloma. Blood 106, 296–303 (2005).
24. Harnoss, J. M. et al. Disruption of IRE1alpha through its kinase domain attenuates multiple myeloma. Proc. Natl Acad. Sci. USA 116, 16420–16429 (2019).
25. Rajkumar, S. V. et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. Lancet Oncol. 15, e538–e548 (2014).
26. Gulsulks, K. et al. Sensitive detection of somatic point mutations in immune and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
27. Gulsulks, K. et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics 27, 2801–2802 (2011).
28. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from whole-genome tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
29. Costello, M. et al. Discovery and characterization of artificial mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 41, e67 (2013).
30. Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in normal contamination. Nat. Methods 15, 531–534 (2018).
31. Landau, D. A. et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell 152, 714–726 (2013).
32. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495–501 (2014).
33. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. 30, 413–421 (2012).
34. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
35. Picard Tools. <https://github.com/broadinstitute/picard>.
36. McLaren, W. et al. The Ensembl variant effect predictor. Genome Biol. 17, 122 (2016).
37. Ramos, A. H. et al. Oncotator: cancer variant annotation tool. Hum. Mutat. 36, E2423–E2429 (2015).
38. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and comprehensive analysis of genomic copy number alterations. Genome Biol. 12, R41 (2011).
39. Consortium, G. T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330 (2020).
40. Hoadley, K. A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158, 929–944 (2014).

41. Nguyen, T., Taggart, R., Diaz, D. & Drapchis, S. A novel approach for data integration and disease subtyping. Genome Res. 27, 2025–2039 (2017).

42. Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 46, 10546–10562 (2018).

43. Liu, J., Wang, C., Gao, J. & Han, J. In: Proceedings of the 2013 SIAM International Conference on Data Mining (SDM) 252–260 (2013).

44. Zhang, S. et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. 40, 9379–9391 (2012).

45. Shen, R. et al. Integrative subtype discovery in glioblastoma using iCluster. PLoS One 7, e35236 (2012).

46. Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S. & Sidow, A. Multi-omic tumor data reveals diversity in molecular mechanisms that correlate with survival. Nat. Commun. 9, 4453 (2018).

47. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods 11, 333–337 (2014).

48. Benjamin, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc.: Ser. B (Methodol.) 57, 289–300 (1995).

49. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15, R29 (2014).

50. Sergushevich, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv (2016).

51. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA 102, 15545–15550 (2005).

52. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740 (2011).

53. Schmitz, R. et al. Genetics and pathogenesis of diffuse large B-cell lymphoma. N. Engl. J. Med. 378, 1396–1407 (2018).

54. Taylor-Weiner, A. et al. Scaling computational genomics to millions of individuals with GPUs. Genome Biol. 20, 228 (2019).

55. Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. 48, 600–606 (2016).

56. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. Nature 578, 94–101 (2020).

57. Pedregosa, F. et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).

58. Uno, H., Cai, T., Pencina, M. J., D’Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat. Med. 30, 1105–1117 (2011).

59. Aanand, S. getzlab/MM, clustering, 2020: Publication Release (v1.0). Zenodo, https://doi.org/10.5281/ZENODO.6348105 (2022).