nature genetics

Article

https://doi.org/10.1038/s41588-025-02196-0

Genomic landscape of multiple myeloma and its precursor conditions

Received: 4 July 2024

Accepted: 15 April 2025

Published online: 21 May 2025

Check for updates

Jean-Baptiste Alberge[1,2,3,17], Ankit K. Dutta[1,2,3,17], Andrea Poletti[1,2,4,17], Tim H. H. Coorens[3], Elizabeth D. Lightbody[1,2,3], Rosa Toenges[1,2,3], Xavi Loinaz[3], Sofia Wallin[1], Andrew Dunford[3], Oliver Priebe[3], Johnathan Dagan[3], Cody J. Boehner[1], Erica Horowitz[1], Nang K. Su[1], Hadley Barr[1], Laura Hevenor[1], Katherine Towle[1], Rashmika Beesam[1], Jenna B. Beckwith[1], Jacqueline Perry[1], David M. Cordas dos Santos[1,2,3], Luca Bertamini[1,2,3], Patricia T. Greipp[5], Kirsten Kübler[3,6,7,8,9], Peter F. Arndt[10], Carolina Terragna[11], Elena Zamagni[4,11], Eileen M. Boyle[12], Kwee Yong[12], Gareth Morgan[13], Brian A. Walker[14], Meletios Athanasios Dimopoulos[15], Efstathios Kastritis[15], Julian Hess[3], Romanos Sklavenitis-Pistofidis[1,2,3], Chip Stewart[3], Gad Getz[3,9,16,18] ✉ & Irene M. Ghobrial[1,2,3,18] ✉

Reliable strategies to capture patients at risk of progression from precursor stages of multiple myeloma (MM) to overt disease are still missing. We assembled a comprehensive collection of MM genomic data comprising 1,030 patients (218 with precursor conditions) that we used to identify recurrent coding and non-coding candidate drivers as well as significant hotspots of structural variation. We used those drivers to define and validate a simple 'MM-like' score, which we could use to place patients' tumors on a gradual axis of progression toward active disease. Our MM precursor genomic map provides insights into the time of initiation and cell-of-origin of the disease, order of acquisition of genomic alterations and mutational processes found across the stages of transformation. Taken together, we highlight here the potential of genome sequencing to better inform risk assessment and monitoring of MM precursor conditions.

Precursor stages of MM, including monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM), progress to MM at rates of 1% to 10% per year[1–6]. Currently, patients with MGUS and SMM are monitored using variables that reflect tumor burden, such as levels of M-spike and free light chains, as well as plasma cell infiltration in the bone marrow[4,5]. Stratification models use these biomarkers to prognosticate patients and determine their risk category; however, different models can yield discordant results[6,7]. Furthermore, somatic alterations that have a role in the disease are not included in commonly used stratification systems[8,9]. Profiling of genetic biomarkers through whole-genome sequencing (WGS) offers a robust solution for precision oncology with informed risk assessment, stratification and prediction of progression[8–10], addressing a major unmet clinical need. This can improve decision-making on observation versus treatment to prevent the irreversible organ damage that manifests in MM[7,11].

Genomic studies of tumor plasma cells in patients with MGUS and SMM have shown genetic heterogeneity, similar to symptomatic MM in some cases[8,9,12–18], and prognostic value of mutations in the mitogen-activated protein kinase (MAPK) pathway (including *KRAS*), DNA repair genes and *MYC*[8,9]. Here we created a map of the genome of MM and its precursors using mostly WGS data from >1,000 patients (n = 1,030), including 218 MGUS and SMM (MGUS/SMM; n = 138 WGS data from participants of the Precursor Crowd study (PCROWD)). This map enabled the discovery of coding and non-coding drivers, identifying biomarkers of progression, modeling and timing of MGUS/SMM

A full list of affiliations appears at the end of the paper. ✉e-mail: gadgetz@broadinstitute.org; irene_ghobrial@dfci.harvard.edu

to MM transformation. These results highlight the clinical potential of genome profiling for monitoring and intercepting progression in molecularly determined high-risk patients.

Results

The MM-like genomic score captures disease progression

To characterize the somatic mutation landscape and drivers in precursor conditions of MM, we generated and analyzed deep WGS data on bone marrow samples from patients with MGUS ($n = 19$), SMM ($n = 119$) and MM ($n = 20$) and their patient-matched germline controls (median coverage: 50× in tumor samples, 35× in normal samples; Supplementary Table 1). Given that bone marrow samples contain a small fraction of MGUS/SMM cells (median, 15%), we first enriched these cells and used a low-input DNA protocol[19] to reliably extract and sequence tumor DNA (median purity, 81%; Methods). This enabled us to evenly characterize the 119 SMM patients from the three currently used risk groups (International Myeloma Working Group 2/20/20 criteria: 44 low-risk, 29 intermediate-risk and 46 high-risk patients with SMM).

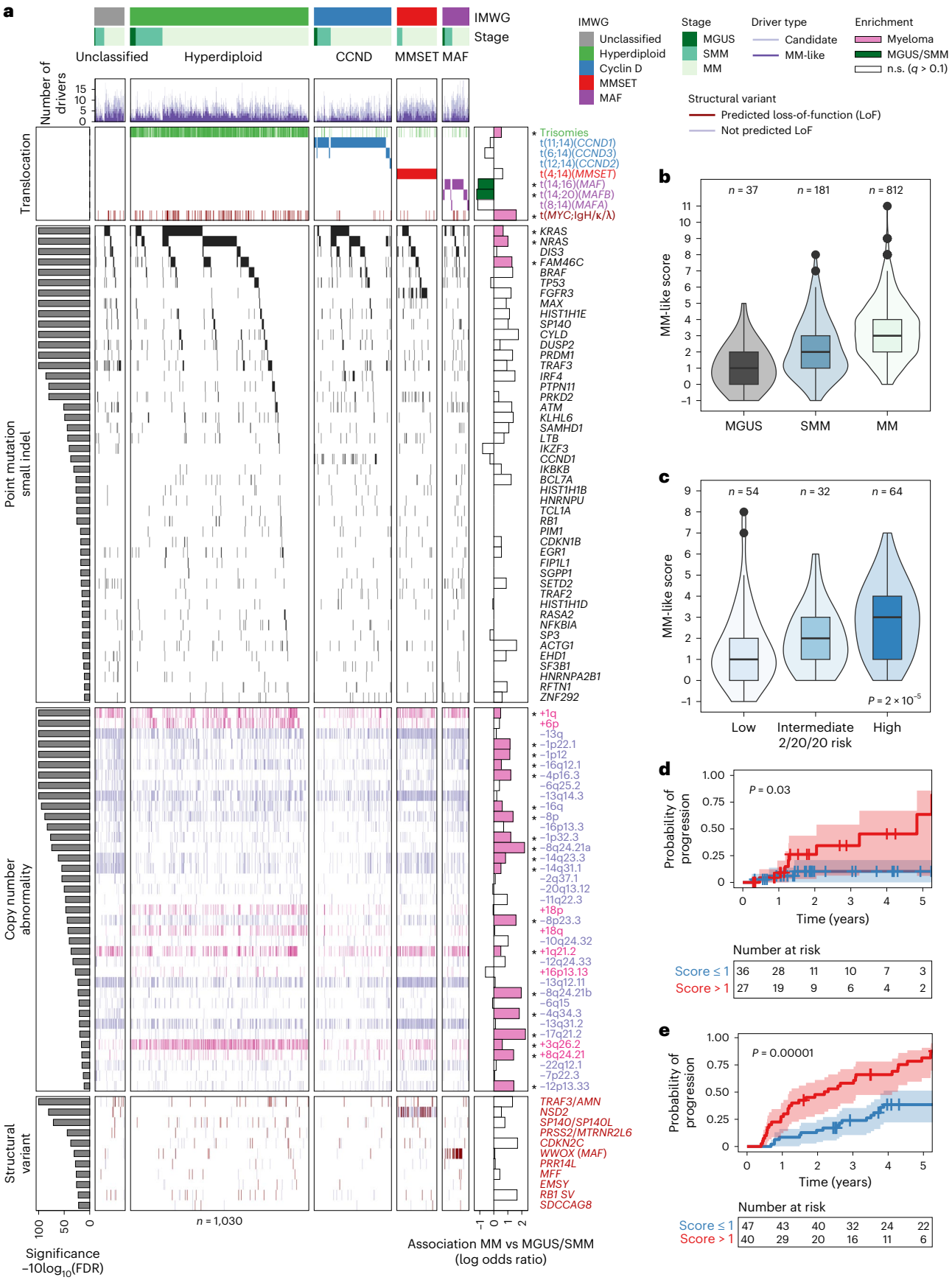
To increase cohort size and power for driver discovery, we combined our precursor-enriched dataset with genomic data from 872 additional individuals with MGUS ($n = 18$), SMM ($n = 62$) and MM ($n = 792$), comprising 810 genomes and 853 exome profiles, yielding a map of somatic mutations of MM and its precursors from 1,030 individuals (Supplementary Table 1). For each tumor, we identified somatic point mutations (single-nucleotide variants (SNVs) and small insertions and deletions (indels)), copy-number alterations and structural variants (SVs; non-coding mutations and SVs were available for 969 samples with WGS data). Our analysis is powered to discover candidate drivers mutated in 2% of patients[20] (Extended Data Fig. 1a). Indeed, analysis of random samples of the cohort confirmed that the number of drivers found in ≥2% of patients increases and plateaus at around 1,000 patients (Extended Data Fig. 1b,c).

First, we set out to discover candidate drivers of MM and precursor conditions. We identified 62 genes with significantly recurrent point mutations (MutSig2CV, $q < 0.1$), including (1) 39 already known MM drivers, such as *KRAS, NRAS, FAM46C, DIS3, BRAF* and *TP53*; (2) eight new candidate drivers (*IKZF3, IKBKB, HNRNPU, FIP1L1, SGPP1, SP3, EHD1* and *HNRNPA2B1*); and (3) 15 additional candidates mutated in <1% of patients (Fig. 1a and Supplementary Table 2). Notably, *IKZF3* (mutated in 1% of patients) is required for the generation of high-affinity long-lived plasma cells[21] and is also a known driver of chronic lymphocytic leukemia (CLL)[22] and diffuse large B cell lymphoma[23]. In individuals for whom no MM SNV driver event was found, at least one canonical initiating MM event was evident. These included either one of the known translocations, hyperdiploidy (trisomy of two or more of the odd-numbered chromosomes (3, 5, 7, 9, 11, 15, 19 or 21)), gain of 1q or deletion of 13q (Extended Data Fig. 1d). Next, we analyzed copy-number alterations and found 54 significant recurrently altered regions (21 chromosome arms, 18 gains and 3 losses; 33 focal events, 4 gains and 29 losses; GISTIC2.0 $q < 0.1$). Among these were the well-known hyperdiploidy gains, 1q, *MYC*, *BCMA*, *FAM46C*, *CDKN2C*, *CYLD* and *FGFR3/MMSET* as well as other significantly altered regions, including *TERC/GPR160*, *EVI5*, *BCL6*, *BIRC2/3* and *XBP1* (Fig. 1a and Supplementary Table 3). Finally, we discovered 11 additional candidate driver structural variations, including two known canonical drivers (*NSD2* and *MAF*; Fig. 1a; see below). When comparing the number of driver events across MM stages, we observed that precursor patients already harbor at least one MM-driver alteration (Fig. 1a). However, the number of drivers increased with MM progression. For example, point mutation drivers were rare in MGUS (median, 0; range, 0–2), increased in SMM (median, 1; range, 0–7) and further increased in MM (median, 2; range, 0–10) (Extended Data Fig. 1e).

Next, we defined a score reflecting the gradual progression from MGUS/SMM to MM by testing whether initiating translocations and each of the 103 driver alterations occurring in >1% of patients were found at different frequencies between MGUS/SMM and MM (Supplementary Table 4). Using the WGS cases, we found 26 such events ($q < 0.1$). The only events enriched in MGUS/SMM were the *MAF* and *MAFB* translocations (*MAF*: odds ratio (OR), 3.1; confidence interval (CI), 1.6–5.8; $q = 0.0026$; *MAFB*: OR, 3.4; CI, 1.1–9.8; $q = 0.04$). The other 24 alterations were more common in MM and included known drivers of the disease (*MYC*: OR, 5.0; CI, 2.2–14; $q = 5 \times 10^{-5}$; *NRAS*: OR, 2.7; CI, 1.5–5.3; $q = 0.009$; *KRAS*: OR, 1.9; CI, 1.2–3.3; $q = 0.09$; Fig. 1a and Supplementary Table 4). We then defined a genomic score by subtracting the number of MGUS/SMM-enriched drivers from the number of MM-enriched drivers for each tumor and named it the 'MM-like' score. As expected, the MM-like score increased throughout disease progression (median MM-like score of 1 in MGUS, 2 in SMM and 3 in MM; Fig. 1b). Among patients with SMM only, who can be stratified into three groups with the 2/20/20 model[4], we found a gradual increase of the MM-like score that matched the overall medians in MGUS, SMM and MM (median scores of 1 in low-risk, 2 in intermediate-risk and 3 in high-risk SMM groups, respectively; $P = 2 \times 10^{-5}$; Fig. 1c). An MM-like score above the median (score > 1) correlated with a higher chance of progression to MM in patients who were untreated for SMM with WGS data available (hazard ratio (HR), 5.3; 95% CI, 1.1–25; $P = 0.03$; Fig. 1d). We first validated this association between an increased MM-like score and higher rates of progression to MM using an internal dataset of 87 patients with SMM patients from a previous publication[8]. (HR, 3.3; 95% CI, 1.8–6.2; $P = 5 \times 10^{-5}$; Fig. 1e). We then confirmed this association using an external validation dataset of 77 patients with SMM from another publication[9] with deep targeted sequencing data, despite the fact that this dataset included only mutations and translocations (HR,1.8, 95% CI, 1.1–3.0; $P = 0.03$; Extended Data Fig. 2a). When analyzed in multivariate regression, we found that the MM-like score was independently associated with a higher chance

Fig. 1 | The MM-like score, derived from the somatic landscape of MM, places precancer stages on a gradual axis of progression toward overt disease. **a**, Waterfall plot illustrating the genomic architecture of MM and its precursors MGUS and SMM across mutation classes (SNVs, small indels, copy-number alterations, SVs and translocations) from a comprehensive collection and harmonization of tumor data ($n = 1,030$ patients). Samples are grouped by genetic subtypes from the International Myeloma Working Group (IMWG) classification: MAF (translocations of the *MAF* family of oncogenes), MMSET (translocations of the *NSD2* gene), CCND (translocations of the cyclin D family of genes), hyperdiploid (trisomies of odd-numbered chromosomes) and unclassified (other genetic subtypes). Common drivers significantly enriched in MGUS/SMM are colored in green, and those significantly enriched in MM are colored in pink in the rightmost panel. The 'MM-like' score of a tumor is then defined as the sum of MM-like drivers minus the sum of MGUS/SMM drivers. FDR, false discovery rate; IgH/k/λ, immunoglobulin heavy/kappa/lambda chain; n.s., not significant. **b**, Boxplot depicting a gradual increase in the MM-like score from MGUS ($n = 37$; median, 1) and SMM ($n = 181$; median, 2) to MM ($n = 812$; median, 3). **c**, In patients with SMM for whom IMWG classification 2/20/20 was evaluable, boxplot showing an increasing MM-like score from low-risk to high-risk categories (low, $n = 54$; median, 1; intermediate, $n = 32$; median, 2; high, $n = 64$; median, 3). Kruskal–Wallis test, d.f. = 2, $P = 1.99 \times 10^{-5}$. Two-sided Dunn's test for post-hoc comparisons: low versus intermediate risk: statistic = −1.48, $q = 0.139$; intermediate versus high: statistic = −2.41, $q = 0.03$; high versus low: statistic = −4.61, $q = 1.24 \times 10^{-5}$. **d**,**e**, Kaplan–Meier curve illustrating the probability of progression of untreated SMM to MM based on the MM-like score in WGS from this study (**d**) ($n = 63$, log-rank test statistic = 5.59, d.f. = 1, $P = 0.02$, Wald test for MM-like score strictly above median of 1, hazard ratio (HR), 5.3; 95% CI, 1.1–25; $P = 0.03$) and validated in an orthogonal dataset[8] (**e**) ($n = 87$, log-rank test statistic 21.07, d.f. = 1, $P = 4 \times 10^{-6}$, Wald test for M-like score > 1, HR, 3.3; 95% CI, 1.8–6.2, $P = 5 \times 10^{-5}$). In boxplots, the center line represents the median, the box limits represent the upper and lower quartiles and the whiskers indicate 1.5 times the interquartile range. Values exceeding whiskers appear individually as plain circles.

https://doi.org/10.1038/s41588-025-02196-0



https://doi.org/10.1038/s41588-025-02196-0

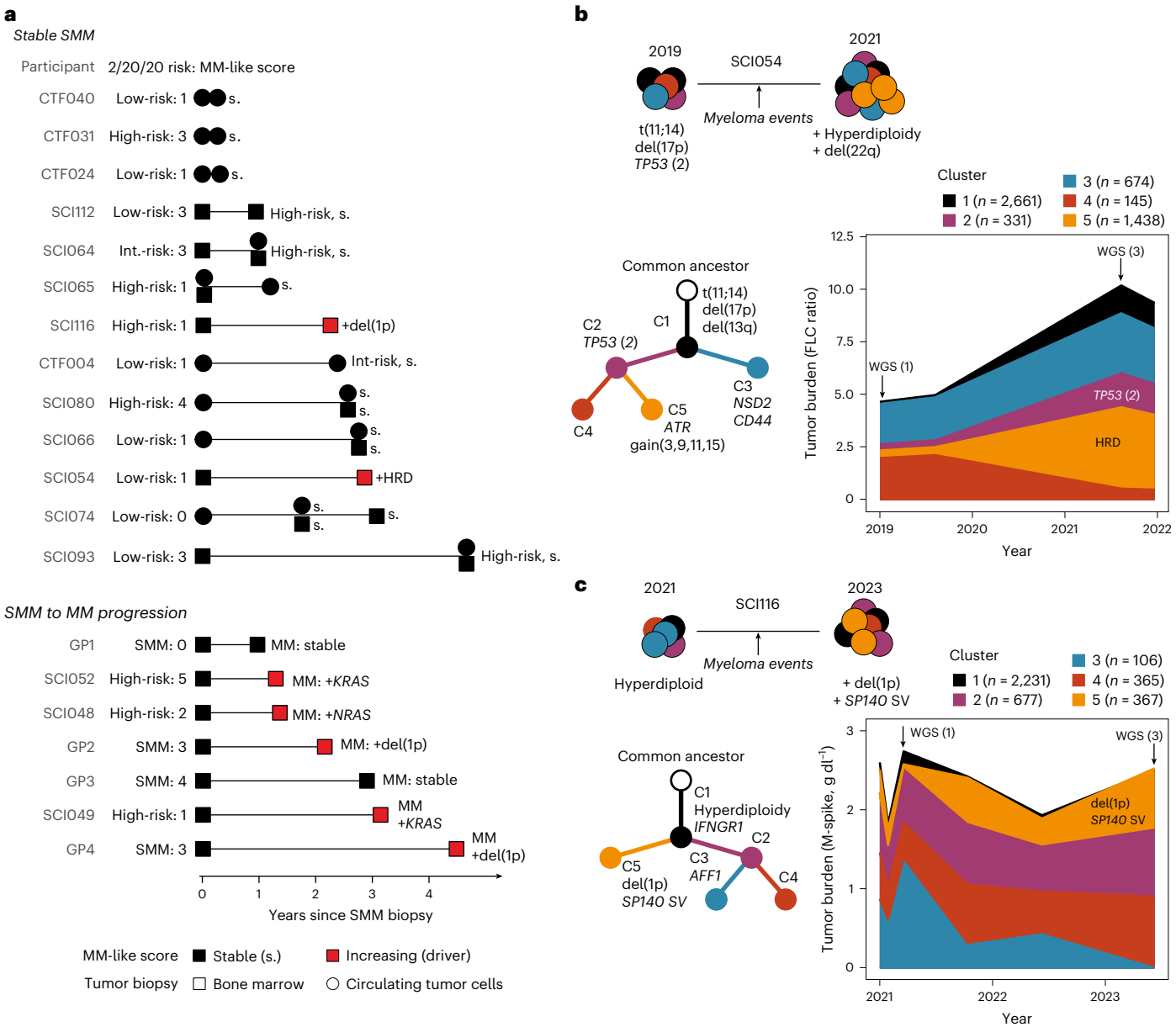


Fig. 2 | The MM-like score increases with progression to cancer and can be linked with clonal outgrowth. a, Swimmer plot showing the clinical efficacy of applying the MM-like score in a longitudinal sample setting, up to 54 months from initial biopsy. Patients characterized by either stable disease trajectory ($n = 13$, black symbols) or evolving genome ($n = 7$, red symbols) were identified, with increasing MM-like score driven by the acquisition of MM-driver mutations (*KRAS*, *NRAS*, deletion of 1p (del(1p)), hyperdiploidy (HRD)). **b**,**c**, Two examples of clinical relevance demonstrating the longitudinal impact of clinical WGS in detecting a MM-like score increase, with association with a subclonal outgrowth obtained from phylogenetic reconstruction of the disease history, and was correlated with an increase in tumor burden. The number of mutations found in each cluster is written to the right of the cluster name. MGUS monitoring over 3 years (**b**), with phylogenetic tree reconstruction to estimate clone driving growth and novel detection of hyperdiploidy within a bi-allelic *TP53* mutant context. High-risk SMM monitoring over 2 years (**c**): phylogenetic analyses revealed newly acquired recurrent MM event of deletion of 1p and *SP140* was identified and assigned to an emerging clone (cluster five (C5)). Concurrently, clinical M-spike levels show stable overall tumor burden. Clonal fractions from before the first WGS are kept equal to the first WGS timepoint for visual representation.

of progression in a model stratified by cohort (HR, 2.8, 95% CI, 1.8–4.4; $P = 7.6 \times 10^{-6}$; Extended Data Fig. 2b). The MM-like score independently correlated with progression in the two validation cohorts[8,9]. In the WGS cohort, the overall multivariate model was just above significance ($P = 0.06$), and the MM-like score had the highest HR estimate compared to the 2/20/20 risk classification variables (Extended Data Fig. 2c–e). Overall, we show that it is possible to define a simple genomic score that can be used to assess the risk of progression to MM and can be implemented in the clinic by discovering drivers of the disease and detecting those associated with the stage of the disease.

The MM-like score increases in serially sampled patients with SMM

We next assessed whether the MM-like score of serial samples from patients with SMM reflects their clinical course. To that end, we collected 47 serial tumor samples (bone marrow or circulating tumor cells) from 20 patients, taken at various time points (from diagnosis to progression) up to 54 months (Supplementary Table 5). Of these patients, 16 were originally from the PCROWD study and four were collected as an external validation cohort (patients GP1–GP4). We calculated the MM-like score for each sample and observed that the

https://doi.org/10.1038/s41588-025-02196-0

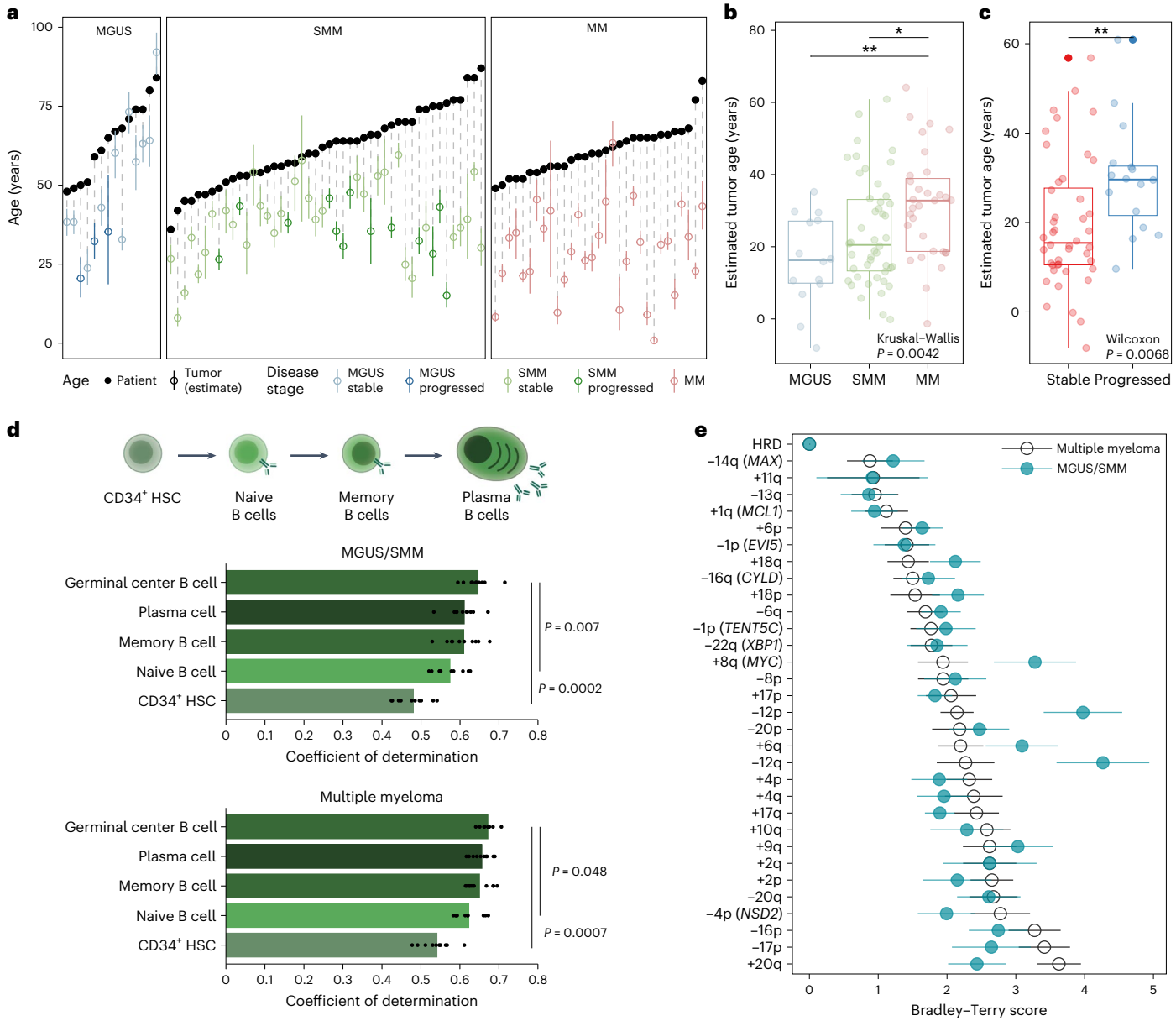


Fig. 3 | Precancer stages of MM have younger tumors at diagnosis, which evolve with late acquisition of MM-like features. a, Maximum likelihood expectation and 95% CIs obtained from binomial mixture models for patients' age at the initiation of the disease (black circles, patient age at sampling; colored circles, estimated age at the disease initiation) including MGUS ($n = 14$; median, 41 years old), SMM ($n = 46$; median, 38 years old) and MM ($n = 31$; median, 31 years old). **b**, Boxplot with point estimates and CIs for the molecular tumor age in MGUS (median, 16 years old), SMM (median, 21 years old) and MM (median, 33 years old). **c**, Boxplot showing that the tumor age estimated from early mutations was higher in patients with MGUS/SMM who progressed during the study versus those who remained stable (progressors: median, 30 years old; non-progressors: median, 15 years old). **d**, Using mutations across the genome, cell-of-origin analysis[29] shows that the mutation profile of MM and MGUS/SMM origins correlates with germinal center, memory B cells and plasma cells. HSC, hematopoietic stem cell. Statistical significance was determined between each top hit and the subsequent models using a two-tailed corrected *t*-test on the values from the tenfold cross-validation, taking into account the interdependence of tests in a cross-validation setting. Panel created in part using BioRender.com. **e**, Bradley–Terry scores estimate (bootstrap mean and standard deviation of the log abilities) from the clonality league competition in MGUS/SMM ($n = 218$; green) and MM ($n = 812$; black). Green dots on the right represent events more likely to be subclonal in MGUS/SMM than in MM and, therefore, are acquired late in the evolution of the disease (that is, a higher Bradley–Terry score). Events +8q (*MYC*), −12p and −12q are significantly more clonal in the MM Bradley–Terry model than in the MGUS/SMM one, with an absolute difference in Bradley–Terry score of >1 (Supplementary Note 6). In boxplots, the center line represents the median, the box limits represent the upper and lower quartiles and the whiskers indicate 1.5 times the interquartile range. Values exceeding whiskers appear individually as plain circles.

score indeed captured the dynamic spectrum of the disease (Fig. 2a). For 11 out of 13 patients who did not progress, the MM-like score was stable over time. These included three patients with a follow-up time less than the first time to progression (10 months), suggesting that these patients probably did not have sufficient time to acquire an additional driver event and expand that subclone to a detectable level. Seven patients clinically progressed to MM, out of whom five also had an MM-like score that increased at the time of progression. One patient (GP3) who progressed to MM with a stable MM-like score had an elevated baseline score (4), and the last one (GP1) remains unexplained with genomic data alone. For the two remaining patients (SCI054 and SCI116), the MM-like score increased but they did not clinically progress (during the measured intervals). Still, one showed a doubling of free light chain ratio indicative of an underlying

https://doi.org/10.1038/s41588-025-02196-0

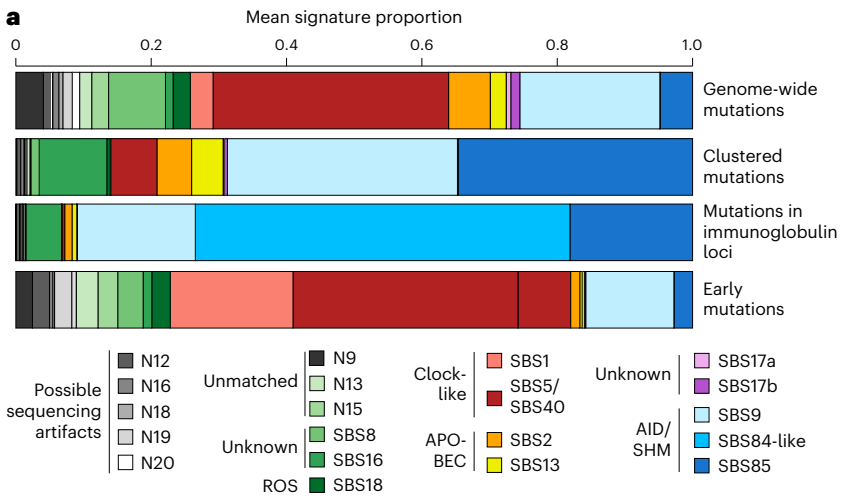


Fig. 4 | The somatic landscape of MM and its precursors is shaped by physiological and cancer-like processes, which evolve after the initiation of the disease. a, Proportion of mutational signatures for whole genomes (top row), restricted to clustered mutations only (within 1,000 bases of another; second row), which are typical of AID activity during SHM in the germinal center, and more specifically within immunoglobulin loci (third row; annotation from COSMIC reference v.3.3). The bottom row represents the proportion of mutational signatures for whole genomes using only early mutations in patients with hyperdiploidy; that is, mutations found at clonal levels on amplified chromosomes. SBS, single-base substitution; ROS, reactive oxygen species. **b**, Fraction of clonal mutations (median cancer cell fraction > 0.85) that are predicted to be caused by known mutational processes (clock-like signatures SBS5/40 and SBS1, AID/SHM signatures SBS9, SBS84 and SBS85, and APOBEC signatures SBS2 and SBS13) in all patients across the MM disease spectrum. Only deep WGS were used; samples from the CoMMpass study were excluded. Total of clonal mutations: clock-like, $n = 14{,}996$ out of 24,780, median, 69% per participant; AID/SHM, $n = 11{,}415$ out of 15,636, median, 71% per participant; APOBEC, $n = 49{,}566$ out of 101,922, median, 30% per participant. The center line represents the median, the box limits represent the upper and lower quartiles and the whiskers indicate 1.5 times the interquartile range. Values exceeding whiskers appear individually as plain circles. $^{*}P < 0.05$, $^{****}P < 0.0001$. *P* values are from Dunn's test for pairwise, two-sided comparisons and were adjusted with the Holm–Bonferroni procedure: clock-like versus AID: $P = 1.15 \times 10^{-2}$; clock-like versus APOBEC: $P = 2.6 \times 10^{-7}$; AID versus APOBEC: $P = 7.75 \times 10^{-13}$.

disease progression together with a bi-allelic *TP53* inactivation, which rarely happens early in the disease development[24]. The increases in the MM-like score were a result of different driver events, including acquiring *KRAS* (G13D), *NRAS* (G13D), deletion of 1p and gain of 1q in the progressors and hyperdiploidy and deletion of 1p in two non-progressors, respectively (Fig. 2b,c and Supplementary Note 5).

Taken together, these data show the ability of WGS data and the MM-like score to successfully capture the acquisition of MM drivers and, using clonal dynamics, identify the subclones that probably acquired and expanded the drivers.

Molecular tumor age of MGUS/SMM

To further delineate the time of initiation and cell-of-origin for MM precursor disease[25–27], we estimated an upper bound for the time of the onset of the disease based on passenger mutations that accumulate in a clock-like manner (Methods). In particular, we focused on the 91 deep WGS tumors (14 MGUS, 46 SMM and 31 MM) with a clonal hyperdiploidy event assumed to be the initiating event[24,28]. There was a wide range of estimated initiation ages, from 1 to 92 years old (median, 35 years old). Interestingly, we found that the patient's age at tumor initiation (that is, hyperdiploidy) correlated with the stage of the disease; the more advanced stages were initiated at a younger age (MGUS, 41 years old; SMM, 38 years old; MM, 31 years old; $P = 0.0035$, Fig. 3a). We could also estimate the age of each tumor by subtracting the tumor-initiating age from the patient's age at the time of biopsy and confirmed that MGUS and SMM are younger tumors than MM (MGUS, 16 years old; SMM, 21 years old; MM, 33 years old; $P = 0.004$; Fig. 3b and Extended Data Fig. 3a,b).

Next, in patients with MGUS or SMM, we wanted to test whether the tumor's age is associated with progression to MM. Indeed, we found that progressors had significantly older tumors than patients with stable disease (median, 15-year-old tumor in stable disease versus 30-year-old in progressors; $P = 0.007$; Fig. 3c). Notably, the tumor's age was independently associated with progression to MM when compared with 2/20/20 risk classification (OR, 1.95; 95% CI, 1.08–4.01; $P = 0.04$; Extended Data Fig. 3c). Together, these data show that WGS-based molecular clocks can be leveraged to differentiate disease severity and risk of progression and that the earliest oncogenic events of MM can occur in the first decades of life.

Common cell-of-origin of MGUS/SMM molecular groups

The local distribution of mutations is shaped by the epigenetic organization, which varies along the genome as well as across cell types. To elucidate the B cell differentiation stage at which MGUS/SMM acquired most of its mutations, we used the cell-of-origin prediction method, which leverages correlations between somatic mutation profiles and epigenetic tracks from healthy tissues[29,30]. We found that among healthy mature B cell types (germinal center, plasma and memory B cells), germinal center B cells were significantly more correlated with somatic mutations than early immature B cell types (naive B cells, $P = 0.007$; hematopoietic stem cells, $P = 0.0002$; corrected *t*-test). This suggests that most passenger mutations are acquired during the later stages of B cell differentiation (Fig. 3d). Notably, the cell-of-origin predictions were comparable across MM subgroups (all coefficients of determination >0.5) and stages (all coefficients of determination >0.6) and remained consistent when we used early mutations only; that is, ones obtained before genomic amplifications (Extended Data Fig. 4a). Of note, we could not find a significantly different correlation between germinal center B cells and plasma cells in MM and its precursors, unlike in non-Hodgkin lymphoma and CLL from the International Cancer Genome Consortium (Extended Data Fig. 4b,c). Overall, this suggests that most mutations in these three neoplasias with a rearranged B cell receptor are acquired during mature B cell differentiation stages and confirms that only MM also acquires mutations in the plasma cell stage.

The order of acquisition of events across disease stages

Next, we wondered whether the order of acquisition of genetic abnormalities has a role in the progression to MM. To answer that question,

https://doi.org/10.1038/s41588-025-02196-0

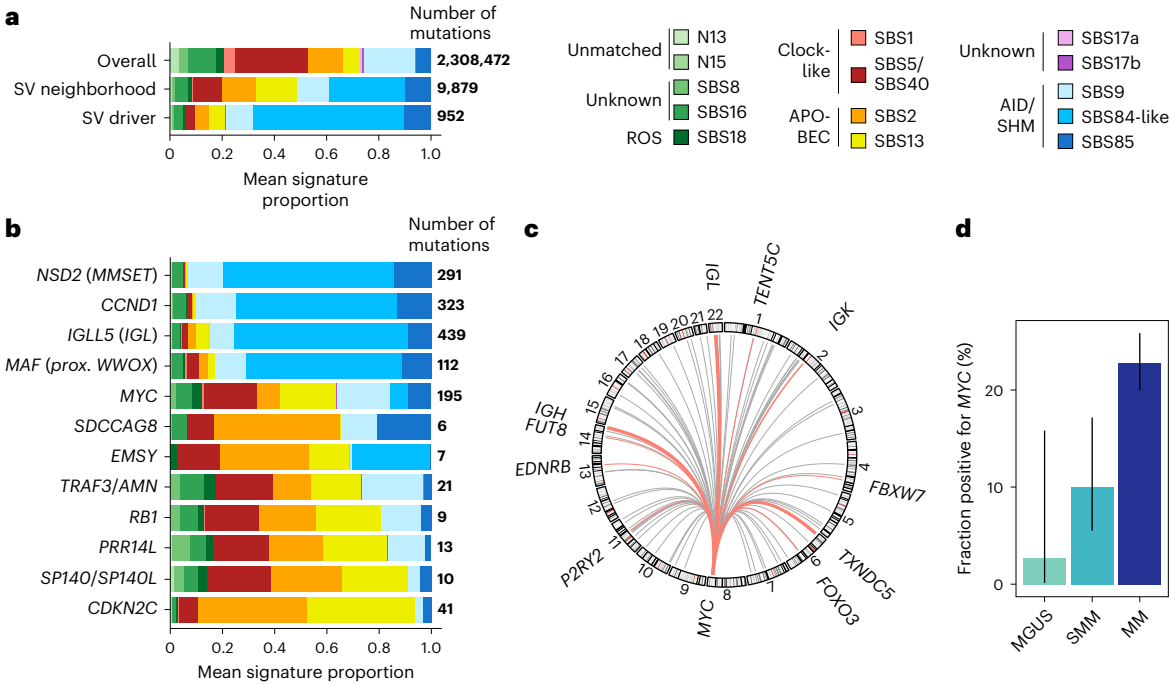


Fig. 5 | The neighborhood of initiating and secondary translocations of MM and its precursors is marked by AID and APOBEC activities. a, Relative frequency of mutational signatures in whole genomes (top row), close to SV breakpoints (<10,000 base pairs; middle row) and in significant SV drivers (bottom row). The number of mutations used for each row is indicated on the right side of the colored bars. **b**, For each SV driver and known canonical myeloma translocations (*MMSET*, *CCND1*, *WWOX/MAF*, *MYC*), frequency of mutational signatures close to SV breakpoints (<10,000 base pairs). *CCND1*, *IGLL5* and *MYC* were added to the list of SV drivers because of their known importance in MM. **c**, Circos plot depicting partners of *MYC* (chromosome 8) in MGUS/SMM, with recurrent partners (≥3 patients) shown in pink. **d**, Fraction of patients with deep WGS performed and a *MYC* translocation detected (immunoglobulin or non-immunoglobulin) across disease stages. Error bars represent binomial proportion 95% CIs for the following counts: MGUS: $n = 1$ out of 37; SMM: $n = 8$ out of 120; MM: $n = 151$ out of 812; bar height represents the frequency itself.

we developed a Bradley–Terry model based on the clonality levels of MM copy-number drivers that we termed 'clonality league comparison' (Methods). In this model, each MM-driver event is given a Bradley–Terry score that reflects the probability that such an event is found in smaller subclones than others. This allowed us to rank MM-driver events: those that are most often clonal have the lowest Bradley–Terry score, and events that are thought to happen late in cancer development have higher Bradley–Terry scores. Then, to test whether the order of acquisition of events differs between the MGUS/SMM and MM stages, we compared the distributions of Bradley–Terry scores between both models (that is, between the MGUS/SMM and MM Bradley–Terry models). We identified three events with higher Bradley–Terry scores in the MGUS/SMM league model than in the MM one; that is, events that are late in MGUS/SMM and earlier in MM ($q < 0.1$; MGUS/SMM score − MM score > 1): gain of 8q24 (*MYC*), deletion of 12p and deletion of 12q (Fig. 3e, Extended Data Fig. 5a–c, Supplementary Table 7 and Supplementary Note 6). Of those, *MYC* and deletion of 12p were also captured by our MM-like score. Deletion of 12q correlated with deletion of 12p in single chromosome deletions, but the event did not reach significance alone in the MM-like score ($q > 0.1$). No event was significantly later in MM relative to MGUS/SMM. Overall, our Bradley–Terry model further supports the role of *MYC* abnormalities and potentially other MM drivers to provide a proliferation advantage toward overt disease.

Mutational processes and signatures in the MM spectrum

In addition to the sequential acquisition of genetic abnormalities, mutational processes could influence the transformation to MM. To discover and quantify mutational processes, we first used hierarchical Dirichlet processes[31] to identify mutational signatures in the WGS data. Then, we associated them with known signatures from the Catalogue of Somatic Mutations in Cancer (COSMIC) (Fig. 4a and Methods)[32]. We found 21 signatures in our dataset, of which 13 also matched COSMIC signatures and accounted for >80% of the total mutation burden: somatic hypermutation and activation-induced cytidine deaminase (SHM/AID): SBS9, SBS84, SBS85; APOBEC: SBS2, SBS13; clock-like: SBS1, SBS5/40; reactive oxygen species signature SBS18 and signatures of unknown etiology SBS8, SBS16, SBS17a and SBS17b (Fig. 4a, Supplementary Table 8 and Supplementary Note 7). Of those, we could validate that APOBEC signatures were enriched in the MAF subgroup[33], and we report a positive association between SBS8 and the cyclin D (CCND) group independent of the disease stage (Extended Data Fig. 6).

Next, and in line with data from CLL[34], we found that three signatures associated with SHM/AID activity (SBS9, SBS84 and SBS85) were highly enriched in clustered mutations (<1,000 base pairs from each other; Fig. 4a, second row), and the association was even stronger when we looked only at the immunoglobulin loci (Fig. 4a, third row). Mutations associated with SHM/AID and clock-like signatures were mostly clonal[35] (median cancer cell fraction > 0.85: 71% and 69%, respectively; Fig. 4b). On the contrary, mutations that we associated with APOBEC activity were primarily found in subclones (median cancer cell fraction > 0.85: 30%). Furthermore, we could not find strong evidence of APOBEC activity in early mutations. Instead, early mutations were strongly enriched with clock-like signatures SBS1 and SBS5/40 in addition to SBS9 (AID), which is coherent with the phenotypically normal expansion of a B cell population before the emergence of an MGUS/SMM clone (Fig. 4a, last row). This suggests that cancer-like mutational processes (APOBEC and reactive oxygen species) increase after acquiring the initiating event of MGUS/SMM, which then gradually develops into MM.

Candidate drivers in hotspots of structural variation

We next searched for cancer driver genes that are affected by SVs. To do so, we developed a statistical method that analyzes SVs in the

https://doi.org/10.1038/s41588-025-02196-0

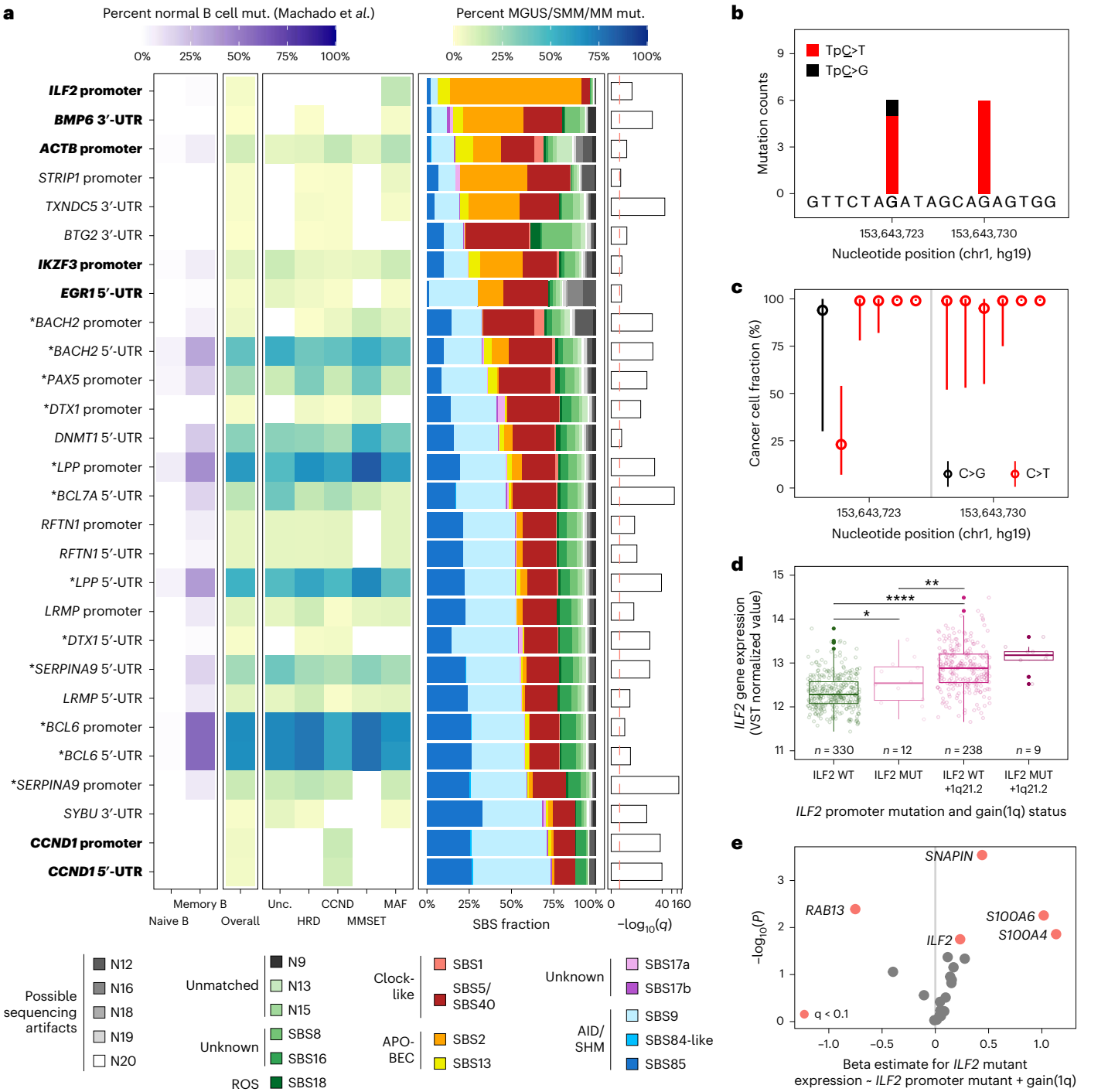


Fig. 6 | Non-coding mutation hotspots across MM stages reflect the history of the disease and candidate driver genes. a, Filtered heatmap representing significantly hypermutated non-coding genes and regulatory elements detected with the DIG algorithm on MGUS, SMM and MM participants with deep WGS (overall, $n = 177$) and molecular subgroup in yellow-green-blue shades (*CCND*, $n = 50$; HRD, $n = 68$; *MAF*, $n = 33$; *MMSET*, $n = 13$; unclassified, $n = 13$). Left panel, fraction of normal B cell expansions with mutations in the same elements (naive B cells, $n = 85$ samples; memory B cells, $n = 74$ samples; data from Machado et al.[30]). Middle-right panel, mutational signature weights in each element. Right panel, *q*-score for each element hypermutation in MGUS, SMM and MM. Bold represents non-coding elements associated with known drivers of myeloma; asterisks indicate non-coding elements that are hypermutated in normal B cells. **b**, For the *ILF2* promoter element, genomic coordinates, counts and tumor alleles found using deep WGS in this study ($n = 12$ mutations from nine participants: MGUS, $n = 2$; SMM, $n = 6$; MM, $n = 1$). **c**, Cancer cell fraction point estimates (maximum likelihood) and 95% CIs from ABSOLUTE[59] for each mutant allele with deep WGS of this study ($n = 12$, median, 99%; range, 23–99%). **d**, *ILF2* gene expression levels are grouped by status of the *ILF2* promoter mutation and of gain of 1q from the CoMMpass study ($n = 589$ covered at the *ILF2* promoter hotspots). Gene expression levels are given after variance-stabilizing transformation (DESeq2 (ref. 60); arbitrary gene expression units). *ILF2* promoter wild-type (ILF2 WT): $n = 330$, median, 12.3; *ILF2* promoter mutant (ILF2 MUT): $n = 12$, median, 12.5; ILF2 WT and gain of 1q: $n = 238$, median, 12.9; ILF2 MUT and gain of 1q: $n = 9$, median, 13.2. The center line represents the median, the box limits represent the upper and lower quartiles and the whiskers indicate 1.5 times the interquartile range. Values exceeding whiskers appear individually as plain circles. **e**, Volcano plot showing the increased expression levels of genes within the topologically associated domain surrounding *ILF2* in patients from the CoMMpass study with an *ILF2* promoter mutation versus wild-type (*P* value from the lm function in R; genes represented in red are statistically significant with $q < 0.1$, where *q* is the *P* value adjusted for false discovery rate).

neighborhood of each gene and estimates whether SVs are causing loss-of-function events more often than expected by chance (Methods). Overall, we analyzed the 969 patients with WGS data and identified 11 candidate SV drivers (Fig. 1a, Extended Data Fig. 7a and Supplementary Table 9). Of the 969 patients, 300 (31%), spanning all stages of MM, had at least one candidate SV driver event. In addition to the known immunoglobulin heavy-chain (IgH)-initiating translocations that are associated with MM subgroups (*NSD2/MMSET* in t(4;14), *WWOX*, a fragile site proximal to *MAF* in t(14;16)), our method detected nine candidates. This list includes four known drivers that are commonly affected by copy-number alterations (*CDKN2C*, *SP140*, *RB1* and *TRAF3*) as well as five new ones (*EMSY, MFF*, *PRR14L*, *PRSS2* and *SDCCAG8;* Extended Data Fig. 7b). Except for the two *IGH* translocations, SV driver mutations were less common in MGUS/SMM genomes (8%; CI, 4–13%) than in MM (27%; CI, 24–30%; $P = 4 \times 10^{-6}$).

The close neighborhood of all SVs was enriched with mutations attributed to the AID and APOBEC signatures (<10,000 base pairs; 80%; Fig. 5a, middle row), and the association was even stronger among SV drivers alone (88%; Fig. 5a, bottom row). We could differentiate between two categories of SVs: those involving immunoglobulin loci (*IGLL5*, *MMSET*, *CCND*, *MAF*) were highly enriched in AID, suggesting that they probably arose from translocations generated from double-strand break repairs in the germinal center that happened during late B cell differentiation (Fig. 5b). On the contrary, we found that all of the other seven SV drivers with mutations found in their neighborhood were enriched with APOBEC signature mutations (SBS2 and SBS13), which are not found in normal physiological B cell development and are often generated when single-stranded DNA overhangs are present during SV formation (Fig. 5b)[36].

As our algorithm to discover MM SV drivers focused on probable loss-of-function events, it did not capture the *MYC* translocation events. *MYC* translocations are known to influence MM disease progression and have been reported at various frequencies across precursor stages (from 0% to 24% (refs. 9,37,38)). In our study, we observed translocations of *MYC* in 9 out of 157 MGUS/SMM patients with WGS data, for an estimated frequency of 5.7% (CI, 2.8–11%). Similar to MM, both immunoglobulin (IgH, Igλ, Igκ) ($n = 3$) and non-immunoglobulin partner ($n = 6$) genes for *MYC* translocations were found. The most common non-immunoglobulin partner was *BMP6*/*TXNDC5* (Fig. 5c). However, in the MGUS population, *MYC* was notably rare (none in the PCROWD study and one progressor case from another study[38]; Fig. 5d), confirming the known increased risk of transformation associated with *MYC* events (Fisher's exact test, $P = 9.3 \times 10^{-5}$). Overall, this shows that structural variation data from WGS can be used to discover cancer drivers and associate them with clonal evolution and disease progression.

Non-coding candidate drivers and B cell biology

Next, we leveraged the DIG algorithm[39] to discover non-coding MM candidate drivers. We found 66 non-coding elements to be significantly recurrently mutated across MGUS/SMM and MM ($q < 0.05$; Fig. 6a). Of those, 18 were in long non-coding and micro RNA and their promoters, 15 were in the 3′ and 5′ untranslated regions (UTRs) of coding genes, 13 were in promoter regions of coding genes, one was in an enhancer region proximal to *TXNDC5*, 12 were found in immunoglobulin regions and seven were considered mapping or sequencing artifacts after manual review (Supplementary Table 10).

A total of 28 candidate elements could be directly associated with a coding gene target (3′-UTR, 5′-UTR, promoter), of which 17 were found mutated across MM subgroups and also in mature B cells[30] (*BACH2* promoter and 5′-UTR, *BCL6* promoter and 5′-UTR, *BCL7A* 5′-UTR and so on; Fig. 6a, left). The AID signatures SBS9 and SBS85 were enriched in these 17 elements, consistent with off-target AID activity that could occur during normal B cell differentiation; we postulate that these mutations are less likely to drive cancer progression. Out of the 11 remaining non-coding elements, six were located in known MM drivers or key genes (promoter and 5′-UTR of *CCND1*, promoters of *ILF2*, 5′-UTR of *EGR1*, 3′-UTRs of *BMP6* and *TXNDC5*); and five remain to be investigated (promoter and 5′-UTR of *DTX1*, promoter of *STRIP1*, 3′-UTRs of *BTG2* and *SYBU*).

Next, we associated the candidate drivers with mutational signatures and MM subtypes. For instance, the *CCND1* promoter and 5′-UTR were hypermutated only in the CCND subtype and are marked by SHM/AID activity (Fig. 6a, bottom rows). This suggests that the translocation occurred before the SHM and class-switch recombination in some cases. Conversely, the *ILF2* promoter on chromosome 1q21.2 was mutated exclusively in tumors from the MAF subgroup ($n = 9$ out of 33 of MAF (27%) among the 177 cases with deep WGS; Fisher's exact test $P = 1.0 \times 10^{-7}$) and was enriched with the APOBEC signature (Fig. 6b) in a CTCF binding peak region (Extended Data Fig. 8). The majority of *ILF2* promoter mutations were clonal (11 out of 12 mutations found in nine patients; Fig. 6c) and spanned all disease stages (two MGUS, six SMM and one MM). Finally, we found that tumors with *ILF2* promoter mutations have a higher expression of *ILF2* and of three other genes in the same topologically associated domain (*SNAPIN*, *S100A4* and *S100A6*) compared to wild-type cases, independent of 1q gains, and using 589 samples from the CoMMpass study with *ILF2* promoter coverage[40] (*ILF2*, $P = 0.018$; Fig. 6d,e, Extended Data Fig. 9a,b and Supplementary Table 12). Altogether, we provide a list of non-coding candidate drivers for MM and a dictionary of B cell-specific off-target mutations reflecting the B cell to plasma cell differentiation.

Discussion

Over the last two decades, the genome of overt MM has been extensively characterized. However, a comprehensive map of the transformation from premalignancy to overt disease was lacking. In this study, we analyzed a large cohort of untreated MGUS and SMM using WGS to understand clonal evolution from the earliest genetic abnormalities found in MGUS to active cancer.

Notably, we reveal somatic coding and non-coding driver mutations, structural variations, signature activities and genetic heterogeneity already present at the earliest precancer stages of MM. Building on these observations, we developed the MM-like genomic score, combining 17 genomic events, which can be used to place patients along a trajectory of disease progression and to identify high and low-risk patients in a data-driven manner. This score, based on point mutations, copy-number alterations and canonical translocations, can be measured using sequencing of targeted panels as well as whole exomes and whole genomes, and can be applied to existing cohorts with treatment outcomes.

Our data suggests that the MM-like score can change within 1–2 years in untreated patients, influencing the choice of frequency of measurement. In treated patients, this frequency could be adjusted to detect and monitor emerging subclones[41,42]. As bone marrow biopsies are often not repeated in the same patient, particularly in the precursor stages, circulating tumor cells or cell-free DNA could potentially replace the need for such biopsies[18,43–49].

Like the classical trajectory of CLL[50] and head and neck cancers[51], we propose an order of acquisition of events for MM and its precursor conditions, where deletions of 12p and 12q, *MYC* abnormalities and MAPK pathway mutations occur late. In our model, which we named 'ladder climbing', these events are acquired stepwise and push tumor cells closer to a cancerous state.

Our refined tumor age estimates suggest that the first oncogenic events occur in the second or third decades of life for both MGUS and SMM, aligning with findings in other hematological malignancies[52]. This tumor age also indicates that by the time MGUS and SMM patients are diagnosed clinically, the tumor lifetime would already be 20–30 years old. The clinical diagnosis of MGUS/SMM in our study, rather than the screened diagnosis, could explain this late tumor initiation compared to MM[53]. This tumor age may be a better

predictor of progression time and could eventually be included in the MM-like genomic score.

In 2014, the International Myeloma Working Group reclassified some SMM as MM based on the 'nearly inevitable' progression to symptomatic disease[54]. Whether some MGUS/SMM should be categorized into MM based on their genomic features and whether treatments will be effective in this population remain open questions under investigation (for example, ClinicalTrials.gov IDs: NCT04775550, NCT05469893 and NCT05767359)[55].

Our study also describes significant non-coding drivers in MM and its precursor stages, which were mostly found in or near established myeloma drivers. These findings suggest that non-coding mutations, as well as non-genetic factors such as epigenetic changes, could have a critical role in pushing myeloma precancerous lesions toward malignancy[55,56].

Although to our knowledge, our study provides one of the largest analyses of whole-genome data for MM and its precursor conditions, more longitudinal samples per patient could further refine our models. Eventually, data from a clinical-grade WGS would replace fluorescence in situ hybridization (FISH) and reliably detect emerging high-risk subclones, including in treatment settings[42,57,58].

In summary, we provide a model of clonal evolution for MM precursor conditions that indicates a 'ladder-like' acquisition of events over time, which we combined into a genomic score available as a biomarker tool to improve risk assessment and stratification.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-025-02196-0.

References

1. Kyle, R. A. et al. Prevalence of monoclonal gammopathy of undetermined significance. *N. Engl. J. Med.* **354**, 1362–1369 (2006).
2. Kyle, R. A. et al. Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma. *N. Engl. J. Med.* **356**, 2582–2590 (2007).
3. Kyle, R. A. et al. Long-term follow-up of monoclonal gammopathy of undetermined significance. *N. Engl. J. Med.* **378**, 241–249 (2018).
4. Lakshman, A. et al. Risk stratification of smoldering multiple myeloma incorporating revised IMWG diagnostic criteria. *Blood Cancer J.* **8**, 59 (2018).
5. Mateos, M.-V. et al. International Myeloma Working Group risk stratification model for smoldering multiple myeloma (SMM). *Blood Cancer J.* **10**, 102 (2020).
6. Cowan, A. et al. Personalised progression prediction in patients with monoclonal gammopathy of undetermined significance or smouldering multiple myeloma (PANGEA): a retrospective, multicohort study. *Lancet Haematol.* **10**, e203–e212 (2023).
7. Kazandjian, D. et al. Genomic profiling to contextualize the results of intervention for smoldering multiple myeloma. *Clin. Cancer Res.* **30**, 4482–4490 (2024).
8. Bustoros, M. et al. Genomic profiling of smoldering multiple myeloma identifies patients at a high risk of disease progression. *J. Clin. Oncol.* **38**, 2380–2390 (2020).
9. Boyle, E. M. et al. The molecular make up of smoldering myeloma highlights the evolutionary pathways leading to multiple myeloma. *Nat. Commun.* **12**, 293 (2021).
10. Maura, F. et al. Genomic classification and individualized prognosis in multiple myeloma. *J. Clin. Oncol.* **42**, 1229–1240 (2024).
11. Abdallah, N. H. et al. Mode of progression in smoldering multiple myeloma: a study of 406 patients. *Blood Cancer J.* **14**, 9 (2024).
12. Walker, B. A. et al. Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. *Leukemia* **28**, 384–390 (2014).
13. Bolli, N. et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).
14. Mikulasova, A. et al. The spectrum of somatic mutations in monoclonal gammopathy of undetermined significance indicates a less complex genomic landscape than that in multiple myeloma. *Haematologica* **102**, 1617–1625 (2017).
15. Bolli, N. et al. Genomic patterns of progression in smoldering multiple myeloma. *Nat. Commun.* **9**, 3363 (2018).
16. Dutta, A. K. et al. Subclonal evolution in disease progression from MGUS/SMM to multiple myeloma is characterised by clonal stability. *Leukemia* **33**, 457–468 (2019).
17. Maura, F. et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nat. Commun.* **10**, 3835 (2019).
18. Dutta, A. K. et al. MinimuMM-seq: genome sequencing of circulating tumor cells for minimally invasive molecular characterization of multiple myeloma pathology. *Cancer Discov.* **13**, 348–363 (2023).
19. Ellis, P. et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
20. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
21. Cortes, M. & Georgopoulos, K. Aiolos is required for the generation of high affinity bone marrow plasma cells responsible for long-term immunity. *J. Exp. Med.* **199**, 209–219 (2004).
22. Lazarian, G. et al. A hotspot mutation in transcription factor IKZF3 drives B cell neoplasia via transcriptional dysregulation. *Cancer Cell* **39**, 380–393.e8 (2021).
23. Chapuy, B. et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* **24**, 679–690 (2018).
24. Manier, S. et al. Genomic complexity of multiple myeloma and its clinical implications. *Nat. Rev. Clin. Oncol.* **14**, 100–113 (2017).
25. Rustad, E. H. et al. Timing the initiation of multiple myeloma. *Nat. Commun.* **11**, 1917 (2020).
26. Barwick, B. G., Gupta, V. A., Vertino, P. M. & Boise, L. H. Cell of origin and genetic alterations in the pathogenesis of multiple myeloma. *Front. Immunol.* **10**, 1121 (2019).
27. Walker, B. A. et al. Characterization of IGH locus breakpoints in multiple myeloma indicates a subset of translocations appear to occur in pregerminal center B cells. *Blood* **121**, 3413–3419 (2013).
28. Kumar, S. K. et al. Multiple myeloma. *Nat. Rev. Dis. Prim.* **3**, 17046 (2017).
29. Kübler, K. et al. Tumor mutational landscape is a record of the pre-malignant state. Preprint at *bioRxiv* https://doi.org/10.1101/517565 (2019).
30. Machado, H. E. et al. Diverse mutational landscapes in human lymphocytes. *Nature* **608**, 724–732 (2022).
31. Teh, Y., Jordan, M., Beal, M. & Blei, D. Sharing clusters among related groups: hierarchical Dirichlet processes. *Adv. Neural Inf. Process. Syst.* **17**, 1623–1634 (2004).
32. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
33. Walker, B. A. et al. APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nat. Commun.* **6**, 6997 (2015).
34. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).

35. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
36. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
37. Misund, K. et al. MYC dysregulation in the progression of multiple myeloma. *Leukemia* **34**, 322–326 (2020).
38. Oben, B. et al. Whole-genome sequencing reveals progressive versus stable myeloma precursor conditions as two distinct entities. *Nat. Commun.* **12**, 1861 (2021).
39. Sherman, M. A. et al. Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nat. Biotechnol.* **40**, 1634–1643 (2022).
40. Skerget, S. et al. Comprehensive molecular profiling of multiple myeloma identifies refined copy number and expression subtypes. *Nat. Genet.* **56**, 1878–1889 (2024).
41. Derrien, J. et al. Acquired resistance to a GPRC5D-directed T-cell engager in multiple myeloma is mediated by genetic or epigenetic target inactivation. *Nat. Cancer* **4**, 1536–1543 (2023).
42. Lee, H. et al. Mechanisms of antigen escape from BCMA- or GPRC5D-targeted immunotherapies in multiple myeloma. *Nat. Med.* **29**, 2295–2306 (2023).
43. Manier, S. et al. Whole-exome sequencing of cell-free DNA and circulating tumor cells in multiple myeloma. *Nat. Commun.* **9**, 1691 (2018).
44. Mishima, Y. et al. The mutational landscape of circulating tumor cells in multiple myeloma. *Cell Rep.* **19**, 218–224 (2017).
45. Garcés, J.-J. et al. Circulating tumor cells for comprehensive and multiregional non-invasive genetic characterization of multiple myeloma. *Leukemia* **34**, 3007–3018 (2020).
46. Guo, G. et al. Genomic discovery and clonal tracking in multiple myeloma by cell-free DNA sequencing. *Leukemia* **32**, 1838–1841 (2018).
47. Lohr, J. G. et al. Genetic interrogation of circulating multiple myeloma cells at single-cell resolution. *Sci. Transl. Med.* **8**, 363ra147 (2016).
48. Deshpande, S. et al. Monitoring treatment response and disease progression in myeloma with circulating cell-free DNA. *Eur. J. Haematol.* **106**, 230–240 (2021).
49. Waldschmidt, J. M. et al. Cell-free DNA for the detection of emerging treatment failure in relapsed/refractory multiple myeloma. *Leukemia* **36**, 1078–1087 (2022).
50. Knisbacher, B. A. et al. Molecular map of chronic lymphocytic leukemia and its impact on outcome. *Nat. Genet.* **54**, 1664–1674 (2022).
51. Leshchiner, I. et al. Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. Preprint at *bioRxiv* https://doi.org/10.1101/508127 (2019).
52. Williams, N. et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).
53. Sigurbergsdóttir, A. Ý. et al. Disease associations with monoclonal gammopathy of undetermined significance can only be evaluated using screened cohorts: results from the population-based iStopMM study. *Haematologica* **108**, 3392 (2023).
54. Rajkumar, S. V. et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol.* **15**, e538–e548 (2014).
55. Cordas dos Santos, D. M. et al. A systematic review and meta-analysis of nonrelapse mortality after CAR T cell therapy. *Nat. Med.* **30**, 2667–2678 (2024).
56. Parreno, V. et al. Transient loss of Polycomb components induces an epigenetic cancer fate. *Nature* **629**, 688–696 (2024).
57. Cordas dos Santos, D. M., Toenges, R., Bertamini, L., Alberge, J.-B. & Ghobrial, I. M. New horizons in our understanding of precursor multiple myeloma and early interception. *Nat. Rev. Cancer* **24**, 867–886 (2024).
58. Duncavage, E. J. et al. Genome sequencing as an alternative to cytogenetic analysis in myeloid cancers. *N. Engl. J. Med.* **384**, 924–935 (2021).
59. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
60. 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).

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Department of Medicine, Harvard Medical School, Boston, MA, USA. [3]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Department of Medical and Surgical Sciences (DIMEC), University of Bologna, Bologna, Italy. [5]Department of Laboratory Medicine and Pathology, Mayo Clinic Comprehensive Cancer Center, Rochester, MN, USA. [6]Center of Functional Genomics, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany. [7]Department of Hematology, Oncology and Cancer Immunology, Charité – Universitätsmedizin Berlin, Berlin, Germany. [8]German Cancer Consortium (DKTK), Partner Site Berlin, and German Cancer Research Center (DKFZ), Heidelberg, Germany. [9]Krantz Family Center for Cancer Research and Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [10]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany. [11]IRCCS Azienda Ospedaliero-Universitaria di Bologna, Istituto di Ematologia "Seràgnoli", Bologna, Italy. [12]Department of Haematology, Cancer Institute, University College London Cancer Institute, London, UK. [13]Myeloma Research Program, Perlmutter Cancer Center, NYU Langone Health, New York, NY, USA. [14]Melvin and Bren Simon Comprehensive Cancer Center, Division of Hematology and Oncology, School of Medicine, Indiana University, Indianapolis, IN, USA. [15]Department of Clinical Therapeutics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece. [16]Department of Pathology, Harvard Medical School, Boston, MA, USA. [17]These authors contributed equally: Jean-Baptiste Alberge, Ankit K. Dutta, Andrea Poletti. [18]These authors jointly supervised this work: Gad Getz, Irene M. Ghobrial. ✉e-mail: gadgetz@broadinstitute.org; irene_ghobrial@dfci.harvard.edu

Methods

Patient sample collection and plasma cell enrichment

Bone marrow samples were collected from the PCROWD study at the Dana-Farber Cancer Institute (IRB number 14-174; NCT02269592). All patients provided written informed consent for the collection and research use of bone marrow samples. Research studies were carried out in accordance with the Declaration of Helsinki. The Research Electronic Data Capture (RedCap) solution was used to aggregate sample-level and participant-level study data[61,62]. Bone marrow samples were drawn into heparinized EDTA tubes before mononuclear cell isolation (SepMate, STEMCELL Technologies) and subsequent plasma cell selection using CD138 magnetic-activated cell sorting (MACS; Miltenyi Biotec). Enriched plasma cell counts and viability of MACS-sorted bone marrow samples were assessed by trypan blue exclusion on an automated cell counter (Invitrogen Countess 3, Thermo Fisher Scientific). In patients with low tumor burden, we also used a standardized CellSearch platform for tumor cell enrichment followed by FACS for the $CD38^{+}CD138^{+}CD19^{-}CD45^{-}$ immunophenotype (Supplementary Note 2)[18]. Each patient also had peripheral blood processed for isolation of peripheral blood mononuclear cells. Samples were cryopreserved for subsequent downstream molecular analyses. Patients with bone marrow biopsies had FISH studies performed at clinical pathology laboratories.

DNA library construction and sequencing

Genomic DNA isolation, library construction and sequencing were performed as previously described[18,63]. For DNA isolation, we used the Monarch Genomic DNA Purification Kit (New England Biolabs (NEB)). Yields from DNA isolation from both tumor plasma cells and normal peripheral blood mononuclear cells were quantified using a Qubit 3.0 fluorometer (Thermo Fisher Scientific). Then, DNA sequencing libraries were prepared using the NEBNext Ultra II FS DNA Library Prep kit (NEB) with unique dual index adaptors (NEBNext Multiplex Oligos) with 50–100 ng of input DNA material. Final library fragment sizes were measured with the BioAnalyzer 2100 (Agilent Technologies), and concentrations were quantified by a Qubit 3.0 fluorometer (Thermo Fisher Scientific) and qPCR (KAPA Library Quantification Kit) before pooling. Final sample libraries were normalized and pooled for matched tumor and normal WGS. WGS was performed at the Genomics Platform of the Broad Institute of MIT and Harvard on an Illumina NovaSeq 6000 instrument. Flowcells 'S4' were used in a 2 × 150 bp paired-end read sequencing configuration.

WGS analyses

After sequencing and library demultiplexing, reads were aligned to the GRCh38 reference genome with BWA MEM (v.0.7.15) (-K 100000000 -p -v 3 -t 16 -Y parameters)[64]. After marking duplicates, base quality scores were recalibrated with GATK 4.2 BQSR without base quality quantization. The analysis of aligned sequences was performed with the Cancer Genome Analysis workflow from the Cancer Program at Broad Institute of MIT and Harvard on an in-house cloud-based HPC system named wolF as follows:

Mutations were detected with MuTect[65] for SNVs, and with Strelka2[66] for small indels. Candidate mutants were filtered through a Panel of Normals consisting of peripheral blood mononuclear cells processed with the same library preparation protocol[20], for orientation bias inducing artifact (8-oxoguanine and formalin-fixed, paraffin-embedded)[67], for mapping artifacts with the BLAT realignment[68] and for enzymatic shearing of cruciform variant artifacts[19]. deTiN[69] was used to estimate the tumor-in-normal contamination rate and rescue candidate mutations accordingly under 10%. Fingerprinting was used to confirm matching tumor–normal pairing[70]. ConEst[71] was used to detect exogenous contamination. Candidate mutations were also matched against the bovine genome with BLAT to ensure samples were not contaminated from other species[72]. Finally, candidate mutations were annotated with GATK Funcotator (v.1.7).

Allelic copy-number ratios were calculated with the AllelicCapSeg[59] method, which clusters together copy-ratio segments and allelic imbalance obtained from the variant allele frequencies of heterozygous SNPs from the matched normal. ABSOLUTE[59] was used to estimate purity, ploidy and cancer cell fraction of mutations and copy-number ratios, and all solutions were manually reviewed. When several samples were available for the same participant, subclonal structures were deciphered with the PhylogicNDT suite of tools[51,73].

SVs were detected with Manta[74], SvABA[75] and dRanger/BreakPointer[76]. Consensus lists of candidate SVs were then established as previously described[77] with the exception of the 10% variant allele frequency filter, which was set to 0% to increase sensitivity to detect immunoglobulin rearrangements. Finally, all candidate SVs from the deep WGS cohorts were manually reviewed with IGV[78].

Discovery of driver candidates (except SVs)

To detect candidate coding drivers, we first selected one sample per patient to avoid double-counting of events (the earliest sample was chosen). The list of candidate drivers from point mutations and indels was established with the MutSig2CV method[20]. Copy-number hotspots and associated genes were discovered with the GISTIC2.0 method[79]. After repeatedly resampling and running MutSig2CV and GISTIC2.0 on subsets of the cohort (with as few as ten patients), we modeled the number of drivers for each category of drivers (those found in <1%, 1–2%, 2–3%, 3–4%, 4–5% and >5% of patients) with logistic regression on the number of samples to test whether the number of drivers discovered plateaus at around 1,000 unique samples (Extended Data Fig. 1). We used the DIG method to discover non-coding hotspots of mutations (candidate drivers)[39]. In brief, the DIG algorithm models constant mutation rates along the genome and identifies non-coding, annotated elements from the PCAWG dataset[80] that are more mutated than expected by chance after adjusting for their sequence context and the overall mutation rate of the cohort. Then, candidate drivers were flagged if they were found recurrently mutated (≥2 samples) in WGS data from normal lymphocytes[30].

Discovery of SV candidate drivers

Candidate SV drivers were determined using SVelfie[81], a computational algorithm that seeks to discover enrichment of probable loss-of-function SV events on a per-gene basis. SVelfie creates a contingency table estimating the probabilities of all possible types of SVs for a given gene (that is, whether each breakpoint of the SV putatively ends up in an exonic, intronic, 5′-UTR, 3′-UTR, promoter or intergenic region, and whether both breakpoints are local to that gene) based on how SV breakpoints would be annotated to that gene using the annotator of dRanger (see above). Such a contingency table is computed for cases in which at least one breakpoint of an SV is within one megabase of the given gene, and it assumes a uniform distribution of breakpoints local to the gene. The probability of both breakpoints being local to a gene is based empirically on how frequently an SV is deemed to have both breakpoints within 1 Mb of a gene for the set of genes being tested by SVelfie. From this information, a subset of the events in the contingency table are categorized as probable loss-of-function to the gene, based on specific breakpoint combinations of an SV event that would putatively disrupt the coding region for a gene and lead to loss of its function. The corresponding probabilities of such loss-of-function events from the contingency table are summed, giving an estimated probability of a given gene losing its function for having an SV with at least one breakpoint local to it. Then, using this probability, a binomial test is performed to determine whether there is significant enrichment of probable loss-of-function events for the gene, and multiple test correction is applied among a starting list of genes being tested by SVelfie, using the Benjamini–Hochberg false discovery rate procedure.

To quantify mutational signatures present specifically in the vicinity of candidate SV drivers (see below), we extracted point mutations

that were found within 10 kb of SV breakpoints. Then, for each candidate SV driver, we calculated the proportions of mutational signature loadings among all such mutations for each gene by averaging their respective probabilities.

Mutational signature extraction and deconvolution

Mutational signatures were extracted from SNVs across all donors using a Bayesian hierarchical Dirichlet process (*hdp* package, https://github.com/nicolaroberts/hdp)[31]. If a given donor had multiple different tumor samples sequenced, SNVs were grouped into pseudo-samples based on their presence or absence across samples, such that a given SNV was only counted once. SNVs were further divided into three categories, each forming a further pseudo-sample: SNVs within immunoglobulin genes, clustered SNVs with an inter-mutational distance of less than 1,000 bases and other SNVs. SNVs were collapsed into their 96 counts of base change within their trinucleotide context. The hierarchy used by hdp consisted of (pseudo-)samples grouped by donor. The hdp algorithm was run in 20 different chains, for 40,000 iterations with a burn-in of 20,000 iterations, as previously described[82].

The resulting extracted signatures were further compared to the reference COSMIC signatures (v.3.3)[32]. Each signature extracted by hdp was modeled as a linear combination of reference signatures previously described in lymphoid cancers[32] or normal lymphocytes[30] using an expectation-maximization algorithm, as previously described[82]. If the cosine similarity of the extracted signature and reconstructed signature exceeded 0.85, the extracted hdp signature was replaced with its constituent COSMIC reference signature. If not, the hdp signature was kept as an unmatched mutational signature (Supplementary Note).

Timing of hyperdiploidy

The acquisition of (sub-)chromosomal duplications can be timed using the relative ratio of duplicated and non-duplicated SNVs[83]. In brief, this ratio can be estimated using a binomial mixture model[84] on the counts of variant supporting reads and total depth on the chromosomal regions affected by duplications, taking into account the purity of the tumor and the local copy-number profile. From these ratios, the relative timing can be calculated:

$$T = \frac{C_{\mathrm{M}} + C_{\mathrm{m}}}{C_{\mathrm{M}} + \frac{P_{ND}}{P_{\mathrm{D}}}},$$

where C_{M} and C_{m} are the major and minor copy numbers, respectively, and P_{D} and P_{ND} are the proportion of mutations assigned to the duplicated or non-duplicated copy number. The timepoint of duplication will be a number between 0 and 1, the former representing the zygote and the latter the most recent common ancestor to the clone; that is, MGUS, SMM or MM. For purposes of practicality, the 1 is often taken to represent the time of sampling, which will give an upper bound of the timing of the hyperdiploidy. Then, this fraction and its confidence interval are converted to actual age at initiation[52,83,84] by multiplying them by the patient's age at the time of biopsy, assuming this time is conflated with the time at which the most recent common ancestor clone existed. Note that this approach assumes a constant mutation rate before and after duplications. Therefore, only SNVs that were most likely to be attributed to SBS1 and SBS5 (both clock-like signatures) were used for the timing analysis.

Ordering of events with a Bradley–Terry model

To accommodate low rates of whole-genome duplications and a high fraction of low-depth WGS from the CoMMpass data, we adapted a method to estimate the relative ordering of acquisition of events based on their relative clonality. As copy-number abnormalities from GISTIC2.0 peaks are often found in different patients, we used a Bradley–Terry model[85] (League model) to allow comparison of events, even if they are not found together in the same sample. In particular, the input matrix of the Bradley–Terry model was modified based on known abnormalities in MM (that is, hyperdiploidy is a single event) and on the precision of WGS (that is, points are assigned to 'winner' copy-number alterations only if the difference in clonality between two samples is significant; see Supplementary Note 6). Then, the standard Bradley–Terry formulas were used to compute log abilities using the BradleyTerry2 R package[86]. In brief, genomic abnormalities $i \in \{1, \ldots, K\}$ are described with their log ability (or Bradley–Terry score) $\beta_i = \log(\alpha_i)$. The probability p_{ij} that i has higher clonality levels than j in the model is expressed as the ratio of their abilities $\frac{\alpha_i}{\alpha_j}$, or alternatively, in its logarithmic form:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{i-}\beta_j.$$

The log abilities β_i are estimated by maximum likelihood estimation:

$$lik(\beta_1, \ldots, \beta_K) = \prod_{m=1}^{n} p_{i_m j_m}^{Y_m} \left(1 - p_{i_m j_m}\right)^{1-Y_m},$$

where n comparisons of genetic abnormalities $(i_1, j_1), \ldots, (i_n, j_n)$ are played and Y_m is the outcome of the mth comparison $Y_m \in (0,1)$.

Then, we ran the same Bradley–Terry model in the MGUS/SMM 'League', versus in the MM 'League', and therefore termed the method 'clonality league comparison'. Bradley–Terry estimates (log abilities) were compared with a two-tailed t-test. The leagues include MM-specific initiating events (hyperdiploidy), arm-level copy numbers and significant GISTIC2.0 peaks. Events with an absolute difference in Bradley–Terry estimate of 1 and adjusted $q < 0.1$ were considered significant.

Cell-of-origin prediction analysis

We followed a previously described[29,30] methodology with some modifications. In brief, aggregated genome-wide somatic tumor mutation density was modeled from chromatin immunoprecipitation followed by sequencing (ChIP–seq) reads of normal B cell types using random forest regression with tenfold cross-validation. For each tumor mutational profile, we trained a separate model using all ChIP–seq profiles associated with each B cell type, combining the two replicates. Clonal and subclonal tumor mutations were counted in 2,125 1 Mb genomic windows, and, in addition to the previously reported excluded regions, we now also excluded the windows spanning the heavy (IgH) and light (Igκ and Igλ) chains. B cell type epigenetic data with replicates were collected from IHEC and EnCODE (ERS663722, ERS663732, ERS643304, ERS728717, ERS663728, ERS728718, CEMT0032.gDNA, ENCDO412QUR, CEMT0135.gDNA, CEMT0134.gDNA) using the following ChIP–seq histone modification tracks: Control, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3. Statistical significance was determined between each top hit and the subsequent models using a two-tailed corrected t-test on the values from the tenfold cross-validation, accounting for the interdependence of tests in a cross-validation setting.

Genomic data collection

Next-generation sequencing data from previous MM studies were analyzed with the same methods as above to ensure harmonized collection of data for comparative analyses, including WGS of the Multiple Myeloma Research Foundation (MMRF) CoMMpass database[40] (dbGAP accession number phs000748.v1.p1, $n = 792$), of our group[18] (phs003084.v1) and others[38,87] (EGAD00001006363, EGAS00001004467). For WGS from the MMRF CoMMpass, which has been sequenced with three different library construction methods, we built one Panel of Normals for each method using matched normal samples with no tumor-in-normal contamination and excluding the

5% lowest fragment size and depth of coverage (Illumina TruSeq, Kapa and KapaHyper). Whole-exome sequencing data matched to the shallow WGS from the MMRF CoMMpass study were analyzed with the Cancer Genome Analysis pipeline as previously described[8] and used to detect coding mutations in these samples. Data analyzed from a previous study were obtained from the original publication[8]: translocations with FISH, mutations and copy-number profiles from whole-exome sequencing[8]. For the molecular timing of MM, processed mutation counts and copy-number estimates were obtained from additional hyperdiploid genomes in public repositories EGAD00001003309 and phs000348.v2.p1.

Statistical analyses

Statistical analyses have been conducted as previously described[18]: median and interquartile ranges are reported to describe quantitative variables unless otherwise stated. The significant difference in average between multiple groups was tested with the Kruskal–Wallis test, followed by Dunn's post-hoc tests. When only two groups were compared, the Wilcoxon test was used. Means are given with their 95% CIs unless stated otherwise, when they are estimated from random processes. The significant difference in event frequency between multiple groups was tested with a χ^2 Pearson test or Fisher's exact test, where appropriate. In survival analyses, patients were stratified per the median of quantitative variables, and time-to-event was calculated from the sampling date to the date of clinical progression to MM or death of any cause, whichever happened first. Patients were censored at the date of last clinical follow-up, or at the start of treatment for MGUS/SMM, where appropriate. Cox regression models were tested with the log-rank test. *P* values were adjusted with the false discovery rate control procedure from Benjamini and Hochberg and are provided as *q*-values[88]; $q < 0.05$ was considered statistically significant (or <0.1 for genomic driver discovery; see above). For boxplots, we used the default median for center line, upper and lower quartiles for box limits, 1.5× interquartile range for whiskers and points for remaining outliers.

Reporting summary

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

Data availability

The sequencing data presented in this publication are available on the dbGAP archive database for sequencing data (dbGAP phs003846.v1). Data from the previously published CTC cohort are publicly available (dbGAP phs003084.v1). Data from the MMRF CoMMpass database were downloaded from the National Cancer Institute Genomics Data Commons website linked to dbGAP (accession number phs000748). Additional WGS from the literature were downloaded from the European Genome–Phenome Archive and dbGAP (EGAD00001006363, EGAS00001004467, EGAD00001003309 and phs000348.v2.p1). Mutation data from CLL and non-Hodgkin lymphoma were obtained from the study of Kübler and colleagues[29]. B cell type epigenetic data with replicates were collected from IHEC, Blueprint and EnCODE and accessed through IHEC Data Portal (https://epigenomesportal.ca/ihec/grid.html; sample names ERS663722, ERS663732, ERS643304, ERS728717, ERS663728, ERS728718, CEMT0032.gDNA, ENCDO412QUR, CEMT0135.gDNA and CEMT0134.gDNA) using the following ChIP–seq histone modification tracks: control, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3. ChIP tracks from the ENCODE Project[89] were downloaded from the ENCODE website (https://www.encodeproject.org; ENCODE experiment numbers GM12878: CTCF: ENCSR000DZN, H3K27Ac: ENCSR000AKC. NCI-H929: CTCF: ENCSR634OAQ, H3K27Ac: ENCSR368CYW. MM.1S: CTCF: ENCSR402IDP, H3K27Ac: ENCSR758OEC. KMS-11: CTCF: ENCSR430YRJ, H3K27Ac: ENCSR955IXZ).

Code availability

The code used to generate analyses and figures for this paper is stored on GitHub and Zenodo in the following repositories: for main figures and data collection, https://github.com/jalberge/ms-wgs ref. 90. Code for the clonality league comparison and Bradley–Terry models is available at https://github.com/andrea-poletti-unibo/MS_TiMMing ref. 91. The method to discover significant SV drivers and probable loss-of-function events, SVelfie, is available at https://github.com/getzlab/SVelfie/releases/tag/v0.0.1 (ref. 92).

References

61. Harris, P. A. et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).
62. Harris, P. A. et al. The REDCap consortium: building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
63. Nadeem, O. et al. Deeper response predicts better outcomes in high-risk-smoldering-myeloma: results of the I-PRISM phase II clinical trial. *Nat. Commun.* **16**, 358 (2025).
64. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
65. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
66. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
67. Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
68. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
69. Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018).
70. Javed, N. et al. Detecting sample swaps in diverse NGS data types using linkage disequilibrium. *Nat. Commun.* **11**, 3697 (2020).
71. Cibulskis, K. et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601–2602 (2011).
72. Wienand, K. et al. Genomic analyses of flow-sorted Hodgkin Reed–Sternberg cells reveal complementary mechanisms of immune evasion. *Blood Adv.* **3**, 4065–4080 (2019).
73. Leshchiner, I. et al. Inferring early genetic progression in cancers with unobtainable premalignant disease. *Nat. Cancer* **4**, 550–563 (2023).
74. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
75. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
76. Drier, Y. et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* **23**, 228–235 (2013).
77. Morton, L. M. et al. Radiation-related genomic profile of papillary thyroid carcinoma after the Chernobyl accident. *Science* **372**, eabg2538 (2021).
78. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

79. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
80. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
81. Loinaz, X. et al. SVelfie: a method for discovering cancer drivers based on enrichment of likely functional structural variants. Preprint at *bioRxiv* https://doi.org/10.1101/2025.03.09.642284 (2025).
82. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
83. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
84. Coorens, T. H. et al. Lineage-independent tumors in bilateral neuroblastoma. *N. Engl. J. Med.* **383**, 1860–1865 (2020).
85. Bradley, R. A. & Terry, M. E. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345 (1952).
86. Turner, H. & Firth, D. Bradley–Terry models in R: the BradleyTerry2 package. *J. Stat. Softw.* **48**, 1–21 (2012).
87. Maura, F. et al. Initial whole-genome sequencing of plasma cell neoplasms in first responders and recovery workers exposed to the World Trade Center attack of September 11, 2001. *Clin. Cancer Res.* **27**, 2111–2118 (2021).
88. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
89. Abascal, F. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
90. Alberge, J. B. jalberge/ms-wgs (v0.9.0). *Zenodo* https://doi.org/10.5281/zenodo.14774594 (2025).
91. Poletti, A. andrea-poletti-unibo/MS_TiMMing: TiMMing (clonality League Model). *Zenodo* https://doi.org/10.5281/zenodo.14813516 (2025).
92. Loinaz, X. getzlab/SVelfie: SVelfie v0.0.1. *Zenodo* https://doi.org/10.5281/zenodo.15132973 (2025).
93. Alberge, J.-B. et al. Novel non-coding candidate drivers of multiple myeloma linked to plasma cell tumor biology and B cell development. *Blood* **144**, 1894 (2024).
94. Alberge, J.-B. et al. A 1,000-genome map for multiple myeloma and its precursor stages, and its impact on clinical outcome. *Blood* **144**, 1019 (2024).

Acknowledgements

We thank A. V. Justis and M. Miller for providing writing and editing support during the preparation of the paper. We acknowledge funding support from the Multiple Myeloma Research Foundation, Adelson Medical Research Foundation, International Myeloma Society Career Development Award (J.-B.A.), Dana-Farber Cancer Institute Medical Oncology Grant Award (A.K.D.), Krouse Family Foundation (I.M.G.) and National Institutes of Health R35CA263817 (I.M.G.). This research was also supported by a Stand Up To Cancer Dream Team Research Grant (grant number SU2C-AACR-DT-28-18). Stand Up To Cancer is a program of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the scientific partner of Stand Up To Cancer. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by Stand Up To Cancer, the Entertainment Industry Foundation or the American Association for Cancer Research. Preliminary results of this study were presented at the 66th Annual Meeting of the American Society of Hematology[93,94].

Author contributions

J.-B.A., A.K.D., G.G. and I.M.G. conceptualized the project. J.-B.A., A.K.D., A.P., T.H.H.C., X.L., C.T., E.Z., J.H., C.S., G.G. and I.M.G. devised the methodology. J.-B.A., A.K.D., A.P., T.H.H.C., E.D.L., R.T., X.L., S.W., A.D., O.P., D.M.C.d.S., L.B., P.T.G., K.K., P.F.A., C.T., E.Z., E.M.B., K.W., G.M., B.A.W., M.A.D., E.K., J.H., R.S.P., C.S., G.G. and I.M.G. conducted the investigation. J.-B.A., A.K.D., A.P., E.D.L., R.T., C.J.B., E.H., N.K.S., H.B., L.H., K.T., R.B., J.B.B., J.P. and R.S.P. performed sample processing, sequencing and data acquisition. J.-B.A, A.K.D., A.P., T.H.H.C., X.L., S.W., O.P., K.K., P.F.A., C.S., G.G. and I.M.G. visualized the project. I.M.G. and G.G. acquired funding. G.G. and I.M.G. supervised the project. J.-B.A., A.K.D., A.P., G.G. and I.M.G. wrote the original draft of the paper. All authors contributed to discussions of content along with review and editing of the paper before submission.

Competing interests

J.-B.A., A.K.D., G.G. and I.M.G. are inventors on patent applications for MinimuMM-seq and for the MM-like score. A.K.D. and E.D.L. have consulted for and received honoraria from Menarini Silicon Biosystems. G.G. receives research funds from IBM, Pharmacyclics/Abbvie, Bayer, Genentech, Calico, Ultima Genomics, Inocras, Google, Kite and Novartis, and is also an inventor on patent applications filed by the Broad Institute related to MSMuTect, MSMutSig, POLYSOLVER, SignatureAnalyzer-GPU, MSEye, MinimuMM-seq, and DLBclass. He is a founder, consultant and holds private equity in Scorpion Therapeutics; he is also a founder of and holds privately held equity in Predicta Biosciences. He was also a consultant to Merck. I.M.G. has consulted for Bristol-Myers Squibb, AstraZeneca, Amgen, Curio Science, Sanofi, Janssen, Pfizer, Menarini Silicone Biosystems, Aptitude Health, GlaxoSmithKline, AbbVie, Adaptive Biotechnologies, Window Therapeutics and Regeneron. She has received honoraria or speaker fees from Vor Biopharma, Janssen, MJH Life Sciences, Novartis, Takeda, Amgen, Regeneron, Curio Science, Standard Biotools and Physicians' Education Resource. She is a founder and executive board member of and holds private equity in Predicta Biosciences. Her spouse is the chief medical officer of and holds private equity in Disc Medicine. R.S.P. is a founder, consultant and holds equity in Predicta Biosciences. E.Z. has received honoraria and served in advisory roles for Janssen, Bristol-Myers Squibb, Sanofi, Amgen, GlaxoSmithKline, Pfizer, Oncopeptides and Menarini–Stemline. The other authors declare no competing interests.

Additional information

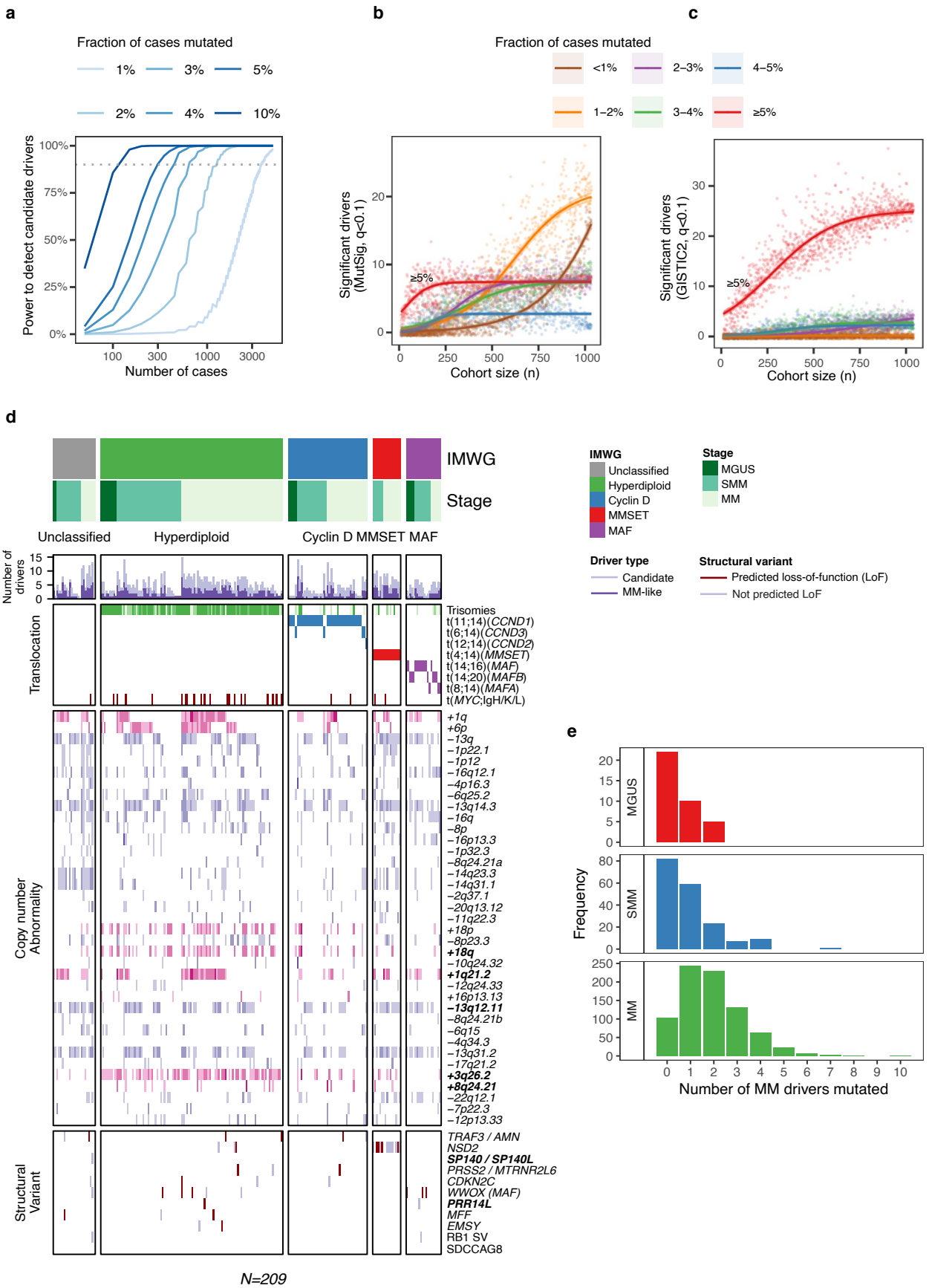
Extended data is available for this paper at https://doi.org/10.1038/s41588-025-02196-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02196-0.

Correspondence and requests for materials should be addressed to Gad Getz or Irene M. Ghobrial.

Peer review information *Nature Genetics* thanks John Jones and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-025-02196-0

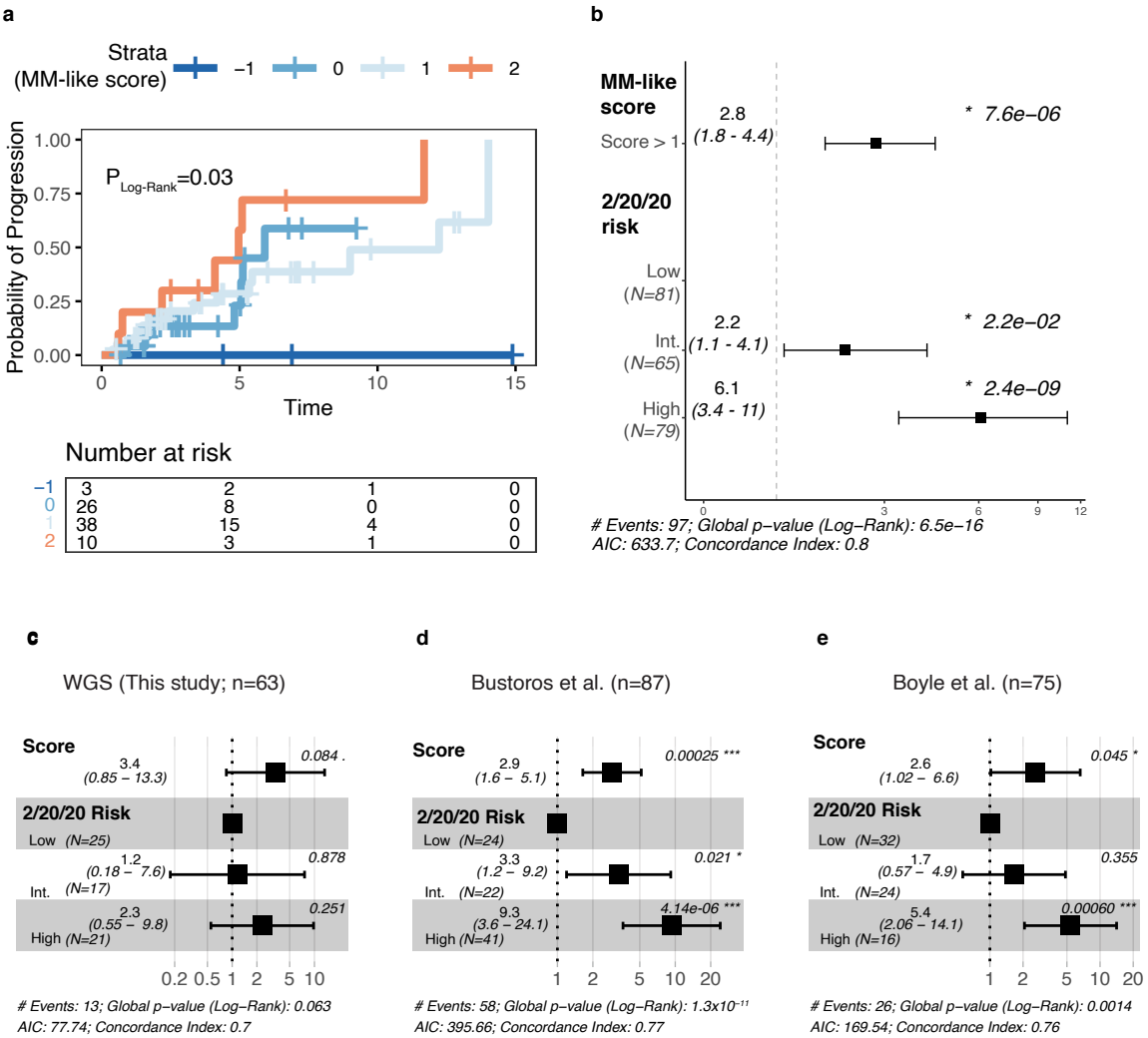


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41588-025-02196-0

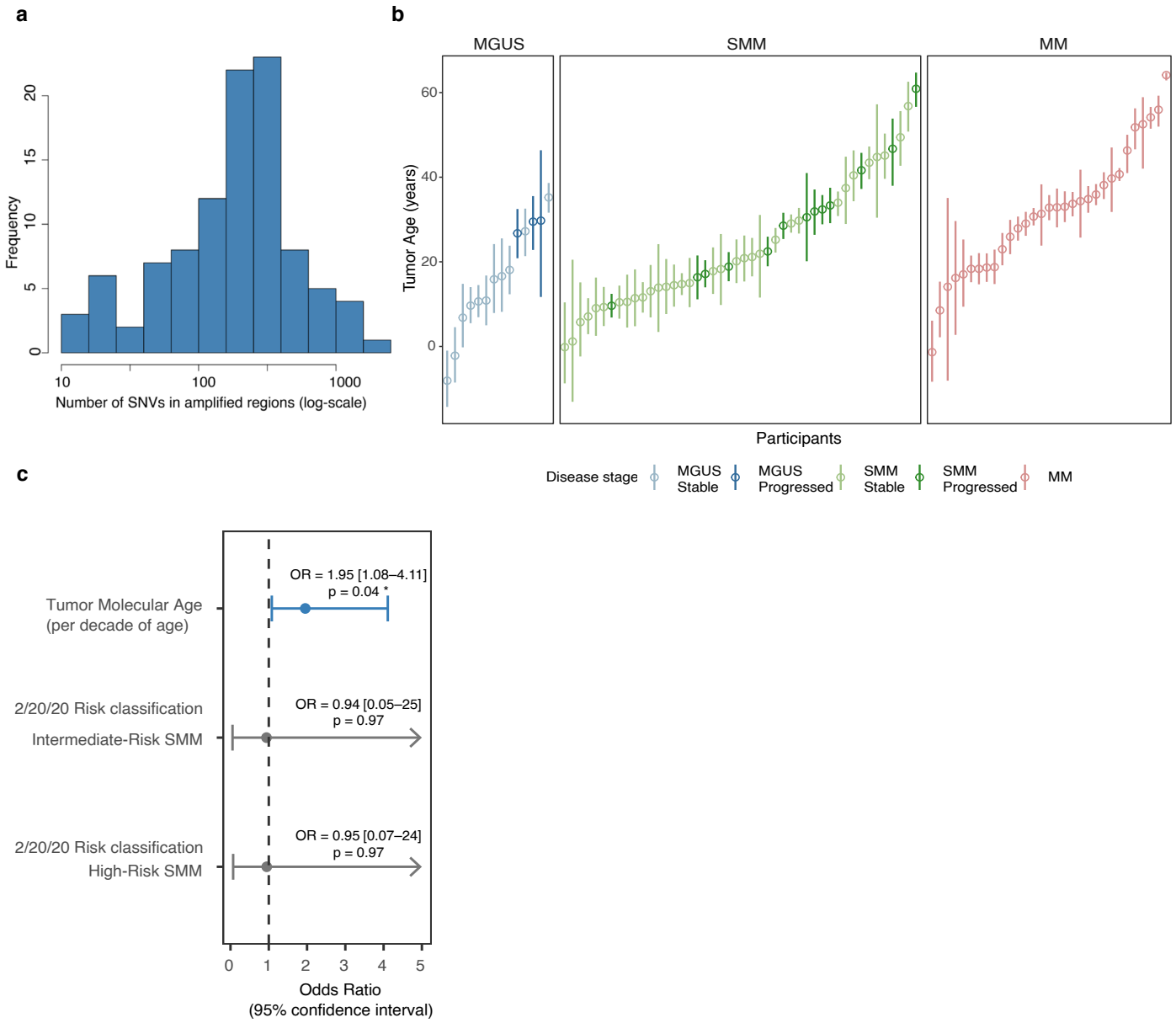
Extended Data Fig. 1 | Sensitivity, saturation analysis for driver discovery and distribution of drivers across MM stages. a) Binomial power analysis showing sensitivity to detect candidate point mutation drivers according to the number of independent tumors-normal (TN) pairs sequenced (50 to 5,000) for diverse ranges of mutation frequencies (> 1%, >2%, >3%, >4%, >5%, >10%), and following the procedure in Lawrence et al.[20]. Dashed line represents the 90% detection power. **b)** Number of candidate point mutations drivers found by MutSig2CV analysis across different frequencies of mutation (each color represents a frequency range for a driver to be mutated in the population as indicated on the graph legend). Each dot represents a random sampling of N participants of the full cohort from N = 10 to 1,030. Saturation curves are modeled by logistic regression. **c)** Saturation analysis for candidate driver discovery based on focal copy-number abnormalities by the GISTIC2 algorithm with axes and modeling similar to panel B. **d)** Copy-number and structural variants in patients without MM-driver point mutation detected (N = 209; MGUS: 21, SMM: 83, MM: 105). Two participants (MMRF_2153 and SMM_102_Tumor) had driver copy-number detected below the cutoff used in this study (MMRF_2153: del(13q, 13%), SMM_102_Tumor: del(16q), 10%, compared with 14% or 0.1 units of $\log_2$ratio). **e**) Distribution of the number of MM drivers found mutated per disease stage and medians 0, 1, 2 for MGUS, SMM, and MM, respectively. Total number of patients: MGUS: 37; SMM: 181; MM: 812.

https://doi.org/10.1038/s41588-025-02196-0



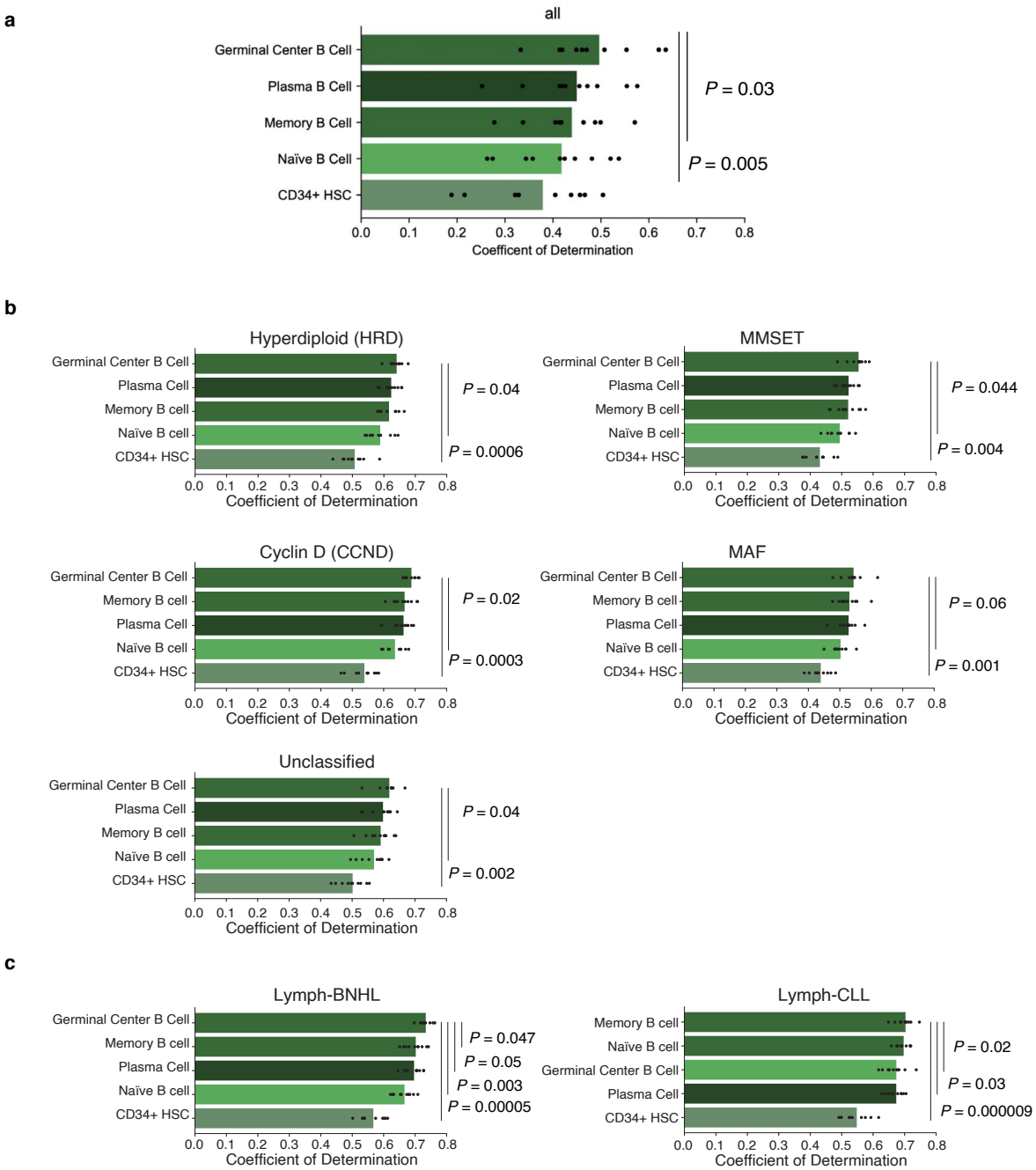
Extended Data Fig. 2 | Clinical significance of the MM-like score and time to progression from SMM to MM. a) Kaplan-Meier curves from progression-free survival in SMM from an external validation cohort (N = 77; ref. 9) with deep target sequencing panels (copy number abnormalities excluded, HR = 1.8, CI95% = [1.1-3.0], P = 0.03). **b**) Hazard ratios and 95% confidence intervals for a Cox regression to model time to progression with MM-like score > 1 and 2/20/20 risk system stratified by study of origin (n = 225). **c**-**e**) Hazard ratios and 95% confidence intervals from the Cox proportional hazard models for progression to MM with MM-like score (> 1 versus ≤1) and 2/20/20 clinical risk stratification (reference: low) as predictors, for this study (**c**), Bustoros et al. (ref. 8) (**d**), and Boyle et al. (ref. 9, minus two patients without 2/20/20 risk available) (**e**). In panels **c** to **e**, the central point and error bars represent the Hazard ratio and its 95% confidence interval from the model described. Significance levels *: P < 0.05; **: P < 0.01; ***: P < 0.001; ****: P < 0.0001.

https://doi.org/10.1038/s41588-025-02196-0



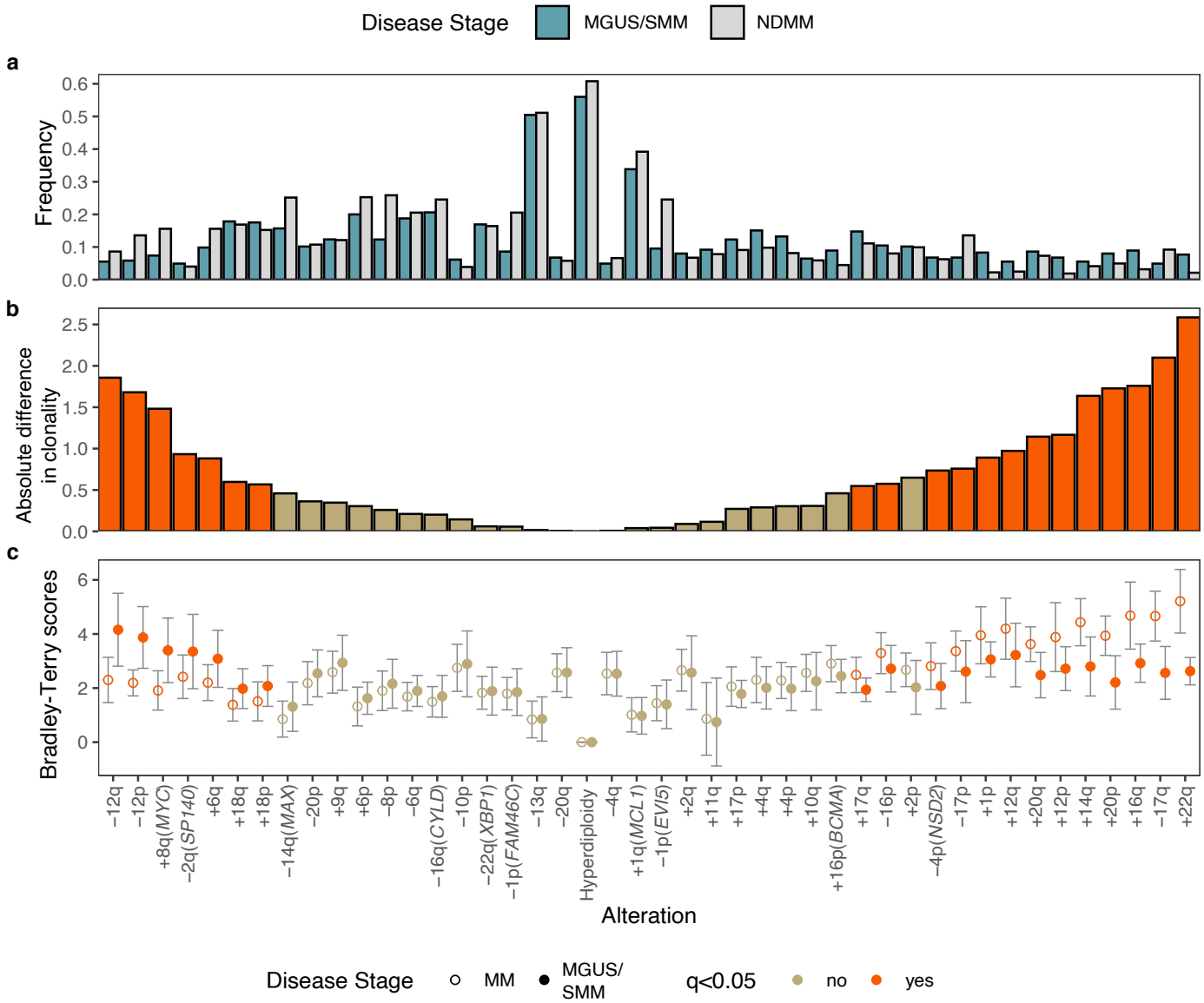
Extended Data Fig. 3 | Molecular clocks that estimate the tumor age of MGUS and SMM in hyperdiploidy patients. a) Number of SNVs used in clonally amplified odd chromosomes among 3, 5, 7, 9, 11, 15, 19, 21 to estimate the molecular age of the tumors (minimum: 10, mean: 283, 75% have more than 283). **b)** Maximum likelihood estimates and 95% confidence intervals obtained from a binomial mixture model, using mutations found on amplified chromosomes from patients with hyperdiploidy, grouped by disease stage. Darker colors represent participants who later progressed to MM (see Methods). Estimates were obtained with a binomial mixture model. MGUS (n = 14, median: 16 years old), SMM (n = 46, median: 21 years old), and MM (n = 31, median: 33 years old). **c)** Logistic regression for progression to MM with molecular tumor age (hyperdiploid clone age) and 2/20/20 risk classification as dependent variables in SMM patients with risk classification available (n = 39; 2/20/20 risk categories: low, n = 10, intermediate, n = 13, high, n = 16). Estimates and 95% confidence intervals were obtained from the "lm" function in R v4.4.1 with default parameters.

https://doi.org/10.1038/s41588-025-02196-0



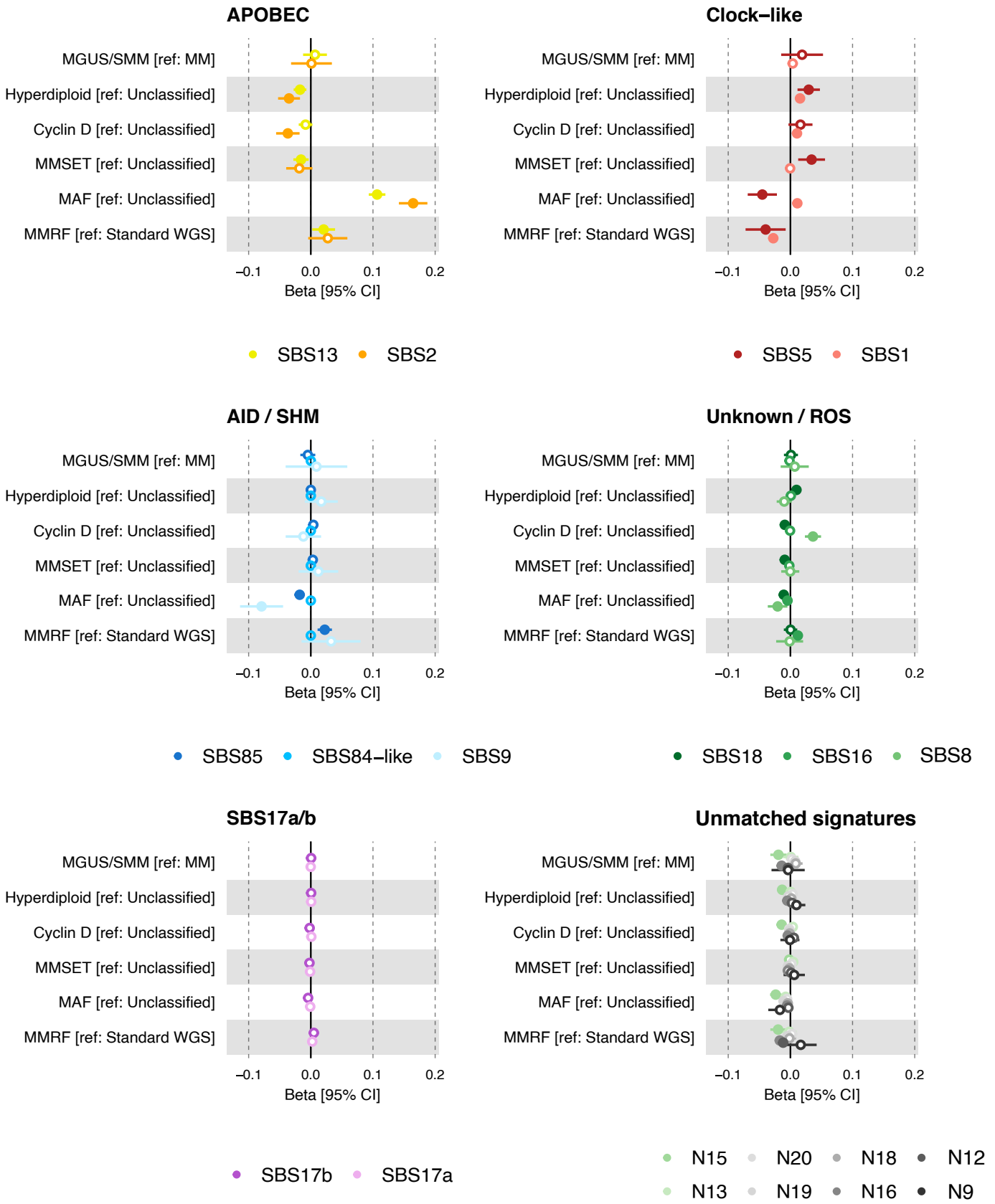
Extended Data Fig. 4 | Cell-of-origin analysis for the development of MGUS and SMM. a) Coefficients of determination obtained from 10-fold cross validation between N = 11,677 early mutations found clonal and duplicated on amplified chromosomes from hyperdiploidy patients and B cell type epigenetic tracks from IHEC and ENCODE consortia. **b)** Cell-of-origin analysis repeated as described for panel A, within each of the CCND, MMSET, MAF, HRD subgroups. **c)** Cell-of-origin analysis repeated as described for panel A for Non-Hodgkin's Lymphoma (n = 993,324 mutations) and Chronic Lymphocytic Leukemia (n = 172,791 mutations) from the ICGC consortium deep WGS. Bars represent the mean value obtained from 10-fold cross-validation (9 degrees of freedom), statistical significance was determined between each top hit and the subsequent models using a two-tailed corrected t-test on the values.

https://doi.org/10.1038/s41588-025-02196-0



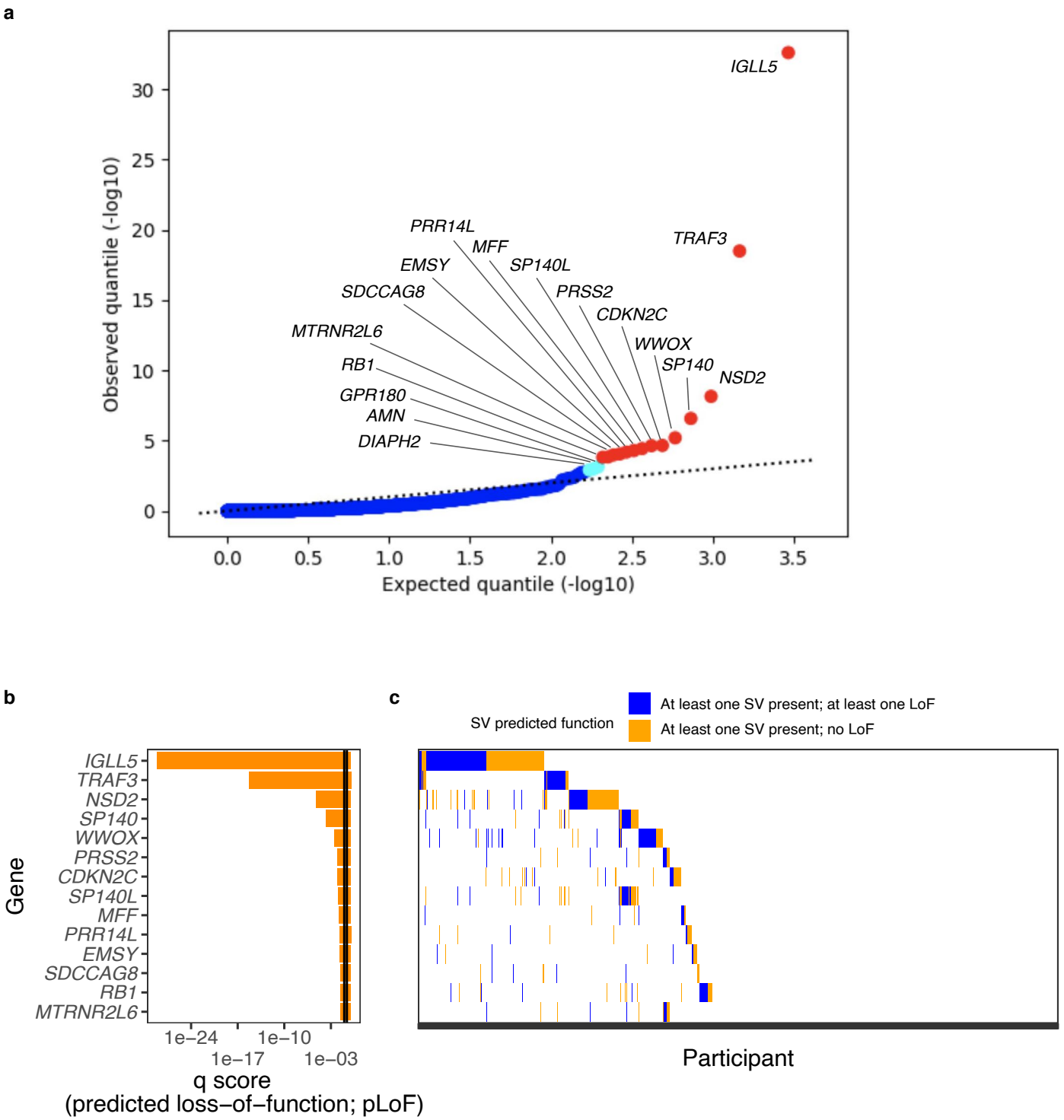
Extended Data Fig. 5 | Correlation between the frequency and the clonality of copy-number abnormalities used in the Bradley-Terry models (Clonality League Comparison). a) Overall frequency of events in the MGUS/SMM league (dark grey) versus MM league (light grey). An event is considered positive with at least a 0.1 difference in the $\log_2$ copy number space after purity adjustment. **b)** Difference of the timing of events estimated with the Bradley-Terry model (Bradley-Terry estimates; average MM minus average MGUS/SMM), in absolute values (towards left: early MM, late MGUS; towards right: early MGUS, late MM). **c)** Comparison of Bradley-Terry estimates and 95% confidence intervals obtained with bootstrapping (n = 20) across MM (empty circles) and MGUS/SMM (filled circles) leagues. Error bars represent the mean and 95% confidence interval. Events on the left are predicted to happen late in the MGUS/SMM history but early in MM (absolute difference > 1 and q < 0.05).

https://doi.org/10.1038/s41588-025-02196-0



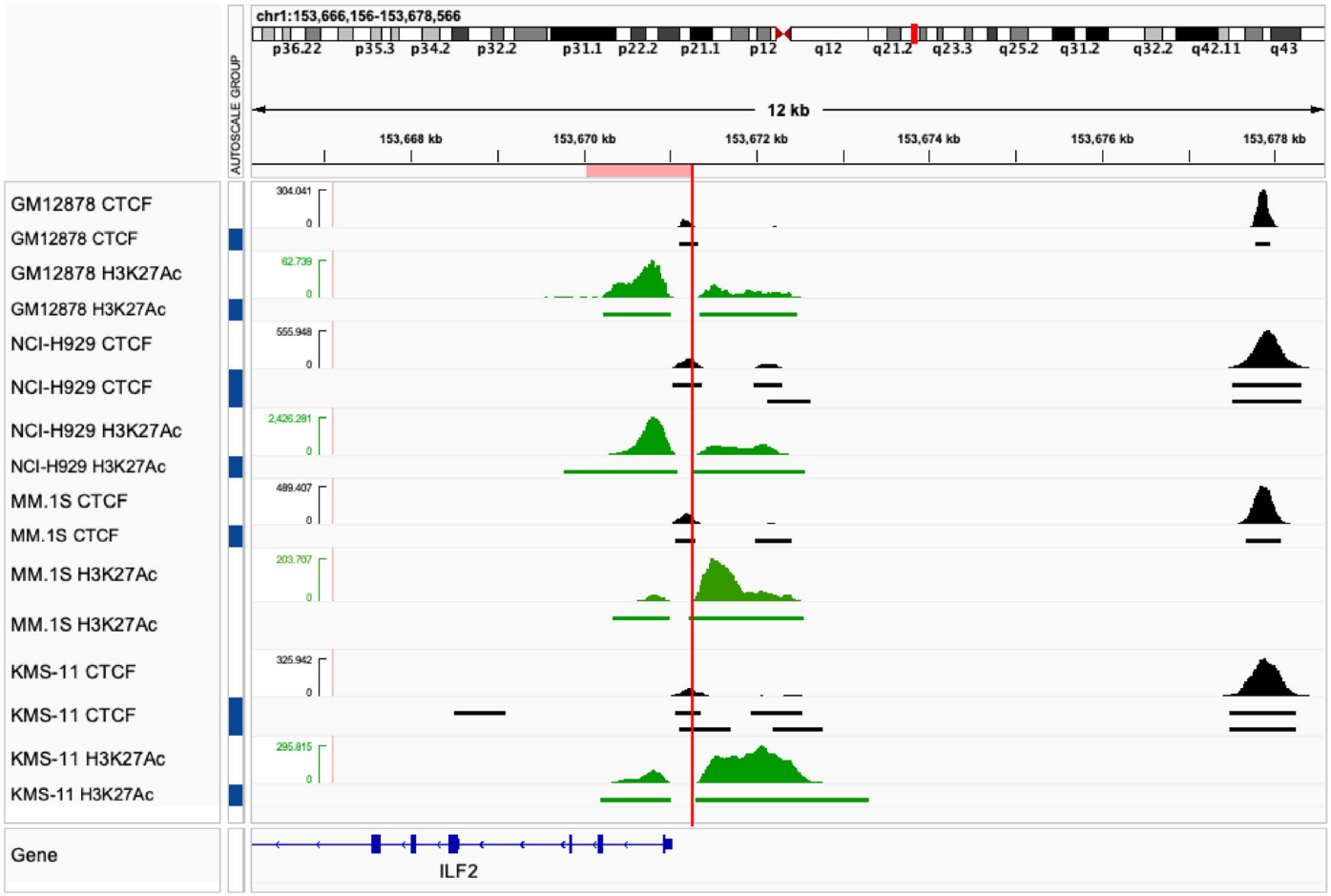
Extended Data Fig. 6 | Predictors of mutational signatures weights based on MGUS/SMM versus MM comparison and adjusted for molecular class, and WGS technique. In each panel, beta estimates and associated 95% confidence intervals for the predictor in a linear regression (SBS ~ Disease Stage + Molecular Class + WGS protocol). Mutational signatures are grouped by known or shared etiology from COSMIC catalogue v3.3. MMRF: shallow genomes from the MMRF CoMMpass database downloaded from the NCI Genomics Data Commons website. Empty dots represent associations below significance levels ($q < 0.05$). Central estimates and 95% confidence intervals were obtained from the "lm" function in R v4.4.1 with default parameters.

https://doi.org/10.1038/s41588-025-02196-0



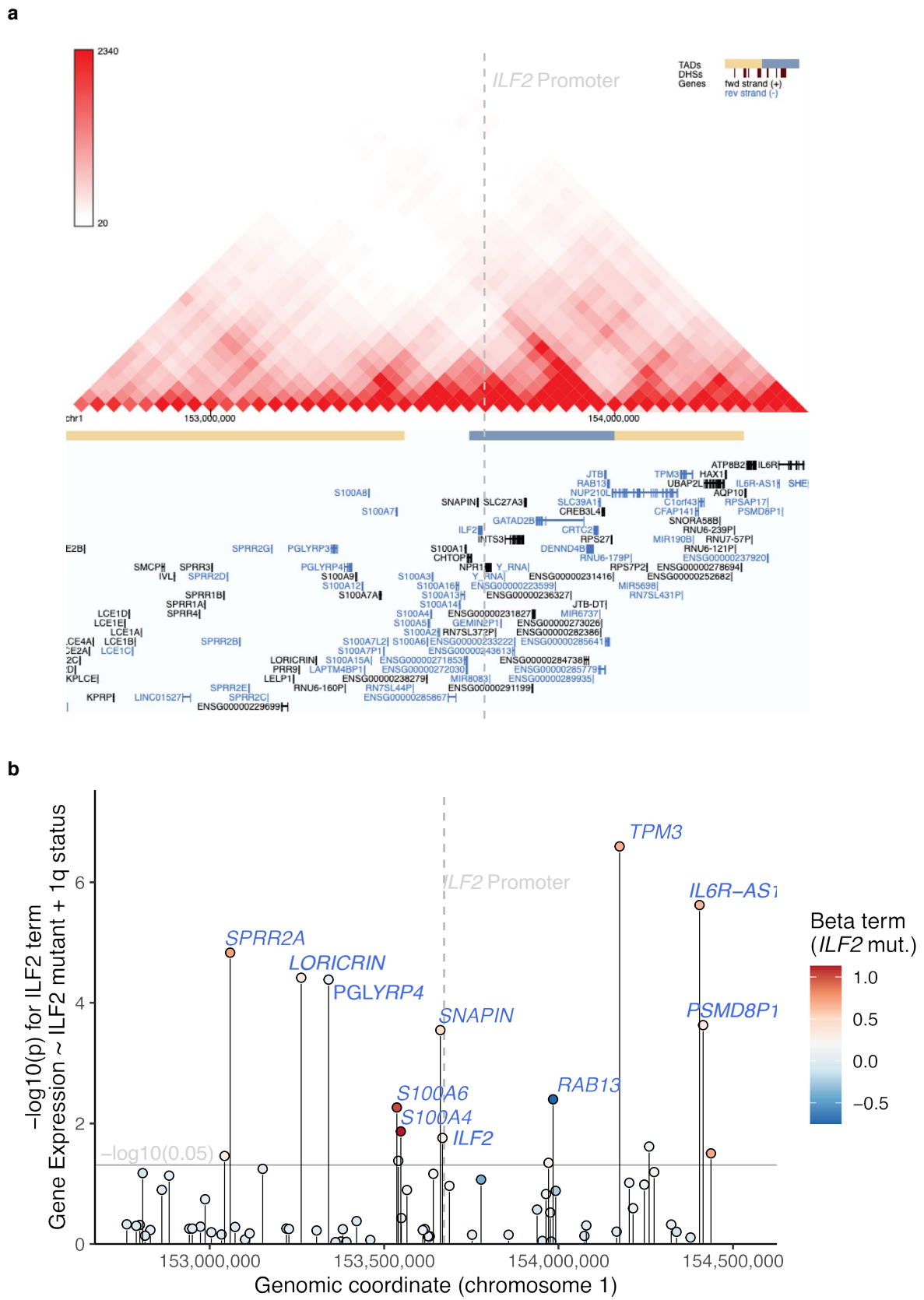
Extended Data Fig. 7 | List of significant SV drivers with predicted loss-of-function. a) Q-Q plot for SV driver discovery validating the absence of p value inflation for this statistical method SVelfie (red: significant hits q < 0.1, cyan: potentially additional hits q < 0.25). **b)** List of significant SV drivers before merging close neighbors (TRAF3/AMN, SP140/SP140L/SP100, PRSS2/MTRNR2L6), with matching q value from binomial test (See Methods). **c)** Detailed CoMut plot of predicted loss-of-function hits in candidate SV drivers for each patient (blue: loss-of-function; orange: not loss-of-function).

https://doi.org/10.1038/s41588-025-02196-0



Extended Data Fig. 8 | CTCF and H3K27Ac chromatin immunoprecipitation sequencing (ChIP-seq) tracks from the ENCODE Project around *ILF2* promoter region. Data (bigwig and BED narrow Peak files) were obtained from the encodeproject.org website, after searching for hg38 tracks for CTCF and H3K27Ac ChIP-seqs. CTCF ChIP-seq tracks are in black, while H3K27Ac are in green. *ILF2* mutation hotspot is depicted with the red line (hg38: chr1:153,671,247 and chr1:153,671,254). *ILF2* promoter region from the PCAWG dataset is depicted with the pink rectangle (hg38: chr1:153,670,032-153,671,248). Significant peaks are indicated with the colored segments below the signal tracks. ENCODE experiments numbers: GM12878: CTCF:ENCSR000DZN, H3K27Ac: ENCSR000AKC. NCI-H929: CTCF: ENCSR634OAQ, H3K27Ac: ENCSR368CYW. MM.1S: CTCF: ENCSR402IDP, H3K27Ac: ENCSR758OEC. KMS-11: CTCF: ENCSR430YRJ, H3K27Ac: ENCSR955IXZ.

https://doi.org/10.1038/s41588-025-02196-0



Extended Data Fig. 9 | Gene expression in the vicinity of the *ILF2* promoter increases within the topologically associated domain (TAD) of the GM12878 cell line and in the CoMMpass study. a) TAD boundaries and local contact-map of the GM12878 from Rao et al. 2014 (image obtained on the http://3dgenome.fsm.northwestern.edu/ website). **b)** Positive correlation between gene expression of *ILF2* and of its neighbor genes from the linear regression of dependent variable Gene expression (VST normalized with DESeq2) with predictor variables *ILF2* promoter status and gain of chromosome 1q status (multivariable linear regression). Data were obtained from the CoMMpass database with *ILF2* covered (≥ 2 reads; See Methods). Vertical axis reflects significance levels, similar to a Manhattan plot, while the color of the dots reflects the estimate itself (red: positive correlation, blue: negative correlation).

nature portfolio

Corresponding author(s): Irene M. Ghobrial

Last updated by author(s): 4/11/2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequencing data: N/A. Clinical annotation: The Research Electronic Data Capture (RedCap v14) solution was used to aggregate data from the study samples and participants (Dana-Farber Cancer Institute IRB #14-174).
Data analysis	The code used to generate analyses and figures for this manuscript is stored in the following code repository https://github.com/jalberge/ms-wgs. Code for the Clonality-League Comparison and Bradley-Terry models is available at https://github.com/andrea-poletti-unibo/MS_TiMMing/. The method to discover significant SV drivers and likely loss-of-function events, SSVelfie, is available at https://github.com/getzlab/SVelfie/tree/v0.0.1 . Other software include (see Methods): bwa mem v0.7.15, GATK v4.2 for duplicate marking, base quality calibration, and CrossCheckFingerPrints, MuTect 1.1.5, Strelka 2.9, deTiN v2.0.1, ContEst (2011), ABSOLUTE v1.5 (https://github.com/jalberge/absolute), PhylogicNDT (#6b4815e, https://github.com/jalberge/phylogicNDT) MutSig2CV (https://github.com/getzlab/MutSig2CV, 0109e27) for driver discovery in coding genes, GISTIC2 v2.0 (https://github.com/broadinstitute/gistic2) for copy-number drivers, DIG (https://github.com/maxwellsh/DIGDriver, #5bb565a) for non-coding driver discovery. hdp was used for signature extraction (https://github.com/nicolaroberts/hdp, #c78989b). Secondary analyses were performed with custom R scripts using R 4.4.1 (see github repositories).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The sequencing data presented in this publication are available on the dbGAP archive database for sequencing data (dbGAP phs003846.v1). Data from the previously published CTC cohort are publicly available (dbGAP phs003084.v1). Data from the MMRF CoMMpass database were downloaded from the National Cancer Institute Genomics Data Commons website linked to dbGAP Accession number phs000748. Additional WGS from the literature were downloaded from the European Genome-Phenome Archive and dbGAP (EGAD00001006363, EGAS00001004467, EGAD00001003309, and phs000348.v2.p1). Mutation data from CLL and NHL were obtained from the study of Kübler and colleagues (BioRXiv doi:doi: 10.1101/517565). B cell type epigenetic data with replicates were collected from IHEC, Blueprint, and EnCODE and accessed through IHEC Data Portal (https://epigenomesportal.ca/ihec/grid.html; sample names ERS663722, ERS663732, ERS643304, ERS728717, ERS663728, ERS728718, CEMT0032.gDNA, ENCDO412QUR, CEMT0135.gDNA, CEMT0134.gDNA) using the following ChIP-sequencing histone modification tracks: Control, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3. Chromatin immunoprecipitation tracks from the ENCODE project91 were downloaded from the ENCODE website (https://www.encodeproject.org/; ENCODE experiments numbers: GM12878: CTCF: ENCSR000DZN, H3K27Ac: ENCSR000AKC. NCI-H929: CTCF: ENCSR634OAQ, H3K27Ac: ENCSR368CYW. MM.1S: CTCF: ENCSR402IDP, H3K27Ac: ENCSR758OEC. KMS-11: CTCF: ENCSR430YRJ, H3K27Ac: ENCSR955IXZ).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex was self-reported and collected per study protocol (Dana-Farber Cancer Institute #IRB 14-174)
Reporting on race, ethnicity, or other socially relevant groupings	Information on race, ethnicity, and other socially relevant groupings were collected in this study (#IRB 14-174) but was not used in the data analysis for this research paper.
Population characteristics	Population characteristics included Gender (54% Female), Age (median 64.5, range 34 to 87, mean 62.8 years old) at the initial sampling date, and clinical diagnosis for cancer or pre-cancer stage. Other population characteristics collected within the study protocol (IRB #14-174) were not used for this analysis.
Recruitment	Patients were recruited in the study if they met the inclusion criteria for the PCROWD study IRB #14-174 and were seen at Dana-Farber Cancer Institute or if they found the study online at https://clinicaltrials.gov/study/NCT02269592. Briefly, inclusion criteria include "Patients with Known or Suspected Precursor Hematological Cancer" (Monoclonal Gammopathy of Undetermined Significance, Smoldering Multiple Myeloma, etc) and were "at least 18 years of age". Complete details are available online at https://clinicaltrials.gov/study/NCT02269592. Clinical recruitment may bias selection of patients with MGUS towards an increased age compared to screened diagnosis, therefore the cohort study in this research paper may be older than a screened cohort for MGUS. (see for example the results from the iStopMM study PMID: 37439374)
Ethics oversight	Bone marrow (BM) samples were collected from the Precursor Crowd (PCROWD) study at the Dana-Farber Cancer Institute (IRB #14-174). All patients provided informed consent for the collection and research use of BM samples. Research studies were carried out in accordance with the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The total sample size of our study is n=1,030 participants. This included participants with a monoclonal gammopathy of undetermined significance (MGUS) or a smoldering multiple myeloma (SMM) (total, n=218), and participants with a newly diagnosed multiple myeloma (total, n=812). Of those, 138 MGUS/SMM originate from the PCROWD Study (#IRB 14-174). A sample size of >1,000 was required per simulation analysis with the MutSig2CV method (Extended Data Figure 1) to reach a 80-90% power to detect multiple myeloma drivers found in >2% of the study population. Saturation analysis by repeated downsampling of the study cohort confirmed that the number of drivers found in >2% of the study population plateaus at around 1,000 patients. The complete composition of the cohort is given in Supplementary Table 1.

Data exclusions	In order to detect genetic drivers selected by the natural progression of the disease, MGUS/SMM patients who were previously treated were excluded from this study. Similarly, to ensure the etiology of MGUS/SMM patients resembles that of the MM spectrum, patients with a concomitant Light Chain Amyloidosis (AL) or Waldenström's macroglobulinemia were excluded from this study. No samples were excluded from this study, except if they were redundant (to avoid double-counting of events in driver discovery) or of low-quality (no tumor population found by analysis with ABSOLUTE algorithm).
Replication	The analysis includes data from previously published studies for which our analysis replicates and confirms the detection of mutations (e.g., Bustoros et al, 2020, Journal of Clinical Oncology; Dutta, et al, 2023, Cancer Discovery; Maura, et al, 2021, Clinical Cancer Research; Oben, et al, 2021, Nature Communications). The prognostic value of our MM-like genomic score was replicated and validated across three cohorts (WGS from this study, WES and Deep Target Sequencing from Bustoros, et al, 2021, Journal of Clinical Oncology, and Deep Target Sequencing + FISH from Boyle EM, et al, 2021 Nature Communications (independent external validation). The increase of MM-like mutations was confirmed with externally collected samples (n=4, Figure 2, patients GP1 to GP4). The detection of recurrent mutations in a non-coding element (promoter) of ILF2 was replicated with patients from the CoMMpass study. (Fig 6). For other estimates were replicated was not possible due to low sample size, limited material availability, or estimates based on one sample, sampling methods were used to estimate confidence intervals around the point estimate (binomial confidence intervals, mixture model, bootstrap, etc.)
Randomization	Down sampling for saturation analyses (Extended Data Figure 1) and bootstrapping (FIgure 3e) were chosen randomly.
Blinding	Investigators were blinded to clinical outcomes (disease progression event and time to progression to MM) for WGS data analysis, driver discovery, and definition of genomic scores. Investigators were not blinded to disease status (MGUS, SMM, versus MM) and this was not relevant for driver discovery.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Antibodies for samples processed with the CELLSEARCH system were provided with the CELLSEARCH (R) Circulating Multiple Myeloma Cell Kit but were not purchased by us. CELLSEARCH protocol states: "Anti-CD138 Ferrofluid: mouse monoclonal antibo and "Staining Reagent: anti-CD38(MultiEpitope)-PE sheep polyclonal, anti-CD38-PE mouse monoclonal, anti-CD19/45-APC mous monoclonal (See Supplementary Note 2 for an example of gating).
Validation	Antibodies were not purchased for this study, they were part of the CellSearch reagents used for isolation of tumor plasma cells (See Supplementary Note 2). More details can be found at https://www.siliconbiosystems.com/en-us/Oncology/Multiple-Myeloma.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*