nature medicine

ARTICLES

https://doi.org/10.1038/s41591-018-0269-2

Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma

Guy Ledergor[1,2,22], Assaf Weiner[1,22], Mor Zada[1], Shuang-Yin Wang[1], Yael C. Cohen[3,4], Moshe E. Gatt[5,6], Nimrod Snir[4,7], Hila Magen[4,8], Maya Koren-Michowitz[4,9], Katrin Herzog-Tzarfati[4,9], Hadas Keren-Shaul[1,10], Chamutal Bornstein[1], Ron Rotkopf[10], Ido Yofe[1], Eyal David[1], Venkata Yellapantula[11,12], Sigalit Kay[3], Moshe Salai[4,7], Dina Ben Yehuda[5,6], Arnon Nagler[4,8], Lev Shvidel[6,13], Avi Orr-Urtreger[4,14], Keren Bahar Halpern[15], Shalev Itzkovitz[15], Ola Landgren[16], Jesus San-Miguel[17], Bruno Paiva[17], Jonathan J. Keats[18], Elli Papaemmanuil[12], Irit Avivi[3,4], Gabriel I. Barbash[19], Amos Tanay[20,21] and Ido Amit[1*]

Multiple myeloma, a plasma cell malignancy, is the second most common blood cancer. Despite extensive research, disease heterogeneity is poorly characterized, hampering efforts for early diagnosis and improved treatments. Here, we apply single cell RNA sequencing to study the heterogeneity of 40 individuals along the multiple myeloma progression spectrum, including 11 healthy controls, demonstrating high interindividual variability that can be explained by expression of known multiple myeloma drivers and additional putative factors. We identify extensive subclonal structures for 10 of 29 individuals with multiple myeloma. In asymptomatic individuals with early disease and in those with minimal residual disease post-treatment, we detect rare tumor plasma cells with molecular characteristics similar to those of active myeloma, with possible implications for personalized therapies. Single cell analysis of rare circulating tumor cells allows for accurate liquid biopsy and detection of malignant plasma cells, which reflect bone marrow disease. Our work establishes single cell RNA sequencing for dissecting blood malignancies and devising detailed molecular characterization of tumor cells in symptomatic and asymptomatic patients.

Multiple myeloma is a neoplastic plasma cell disorder that is characterized by clonal proliferation of malignant plasma cells in the bone marrow. Despite improved survival rates in the past decade, therapy is not curative, and almost all patients relapse[1]. The clinical spectrum of the disease includes asymptomatic conditions such as monoclonal gammopathy of undetermined significance (MGUS), a condition in which limited suspected malignant plasma cells in the bone marrow produce an abnormal monoclonal antibody (M-protein) in the blood, and smoldering multiple myeloma (SMM), a more advanced stage with a higher proportion of malignant plasma cells in the bone marrow and/or M-protein in the blood[2,3]. The rates of progression from MGUS and SMM into active myeloma are approximately 1% and 10% per year, respectively[4]. The genetic landscape underlying myeloma was mapped in several foundational genomic studies[5–9], and gene expression profiling cohorts of individuals with multiple myeloma (such as the Multiple Myeloma Research Foundation's CoMMpass Study) were further shown to be effective in predicting the risk of disease progression and survival[10–12].

The progressive nature of the disease makes it essential to develop tools for risk stratification and early detection of pre-malignant states, including solutions for molecular characterization of bone marrow aspiration procedures and accurate liquid biopsies. The large plasma cell heterogeneity in early disease stages makes it difficult to evaluate precisely the state of asymptomatic patients, severely limiting the possibilities for preventive treatments and restricting clinical practice to 'watchful waiting'[4]. Yet current strategies for genomic sequencing and transcriptional analysis in cancer were developed for mapping bulk samples from primary tumors and metastases and therefore lack the resolution and accuracy for characterizing small tumorigenic subpopulations that are likely driving MGUS, SMM and multiple myeloma residual disease progression. Single cell genomic technologies are opening the way for the development of such assays[13–16].

Here, we report the first comprehensive single cell RNA profiling of newly diagnosed asymptomatic (7 MGUS and 6 SMM) and symptomatic (12 multiple myeloma and 4 primary light chain (AL) amyloidosis) individuals encompassing the different clinical spectra

[1]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Internal Medicine, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel. [3]Department of Hematology, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel. [4]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [5]Department of Hematology, Hadassah Medical Center, Jerusalem, Israel. [6]Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel. [7]Department of Orthopedics, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel. [8]Hematology Division, Chaim Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel. [9]Department of Hematology, Assaf Harofeh Medical Center, Zerifin, Israel. [10]Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot, Israel. [11]Center for Hematological Malignancies, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [12]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [13]Hematology Institute, Kaplan Medical Center, Rehovot, Israel. [14]Genetic Institute, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel. [15]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [16]Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [17]El Centro de Investigación Médica Aplicada (CIMA), CIBER-ONC number CB16/12/00369, Clinica Universidad de Navarra, Pamplona, Spain. [18]Integrated Cancer Genomics, Translational Genomics Research Institute, Phoenix, AZ, USA. [19]Bench to Bedside Program, Weizmann Institute of Science, Rehovot, Israel. [20]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [21]Department of Biological Regulation, Weizmann Institute of Science, Rehovot, Israel. [22]These authors contributed equally: Guy Ledergor, Assaf Weiner. *e-mail: ido.amit@weizmann.ac.il

of plasma cell pre-malignant and neoplastic states. We characterized 20,586 single plasma cells from the bone marrow and 3,540 single plasma cells from the blood of 11 control individuals and 29 subjects newly diagnosed with multiple myeloma. We found minimal interdonor heterogeneity for controls, showing that normal plasma cells from the bone marrow are reproducibly organized into transcriptional states that represent variable activity of genes associated with endoplasmic reticulum stress and plasma cell physiology[17]. In contrast, individuals with multiple myeloma showed highly diverse cell states, with every subject defining a unique and individual plasma cell transcriptional program. Importantly, different subjects overexpressed common known multiple myeloma oncogenic drivers: *CCND1*, *FRZB* and several uncharacterized overexpressed multiple myeloma markers that can be confirmed in the larger CoMMpass database, including the lysosome-associated membrane protein *LAMP5*, the endopeptidase inhibitor *WFDC2* and a small intronless gene located on the X chromosome, *CDR1*. Intrasubject transcriptional heterogeneity was also observed for 10 subjects, in most cases with the same immunoglobulin clonotype. Using single cell RNA sequencing (scRNA-seq) transcriptional data, we inferred copy number alterations (sciCNAs) by averaging the relative expression of a large number of genomically adjacent genes and showed that in multiple myeloma, transcriptional changes are often regulated in trans and may be associated indirectly with genomic aberrations. By profiling single plasma cells from the blood, we further identified efficient molecular markers (such as *CD52*) to enrich for circulating tumor cells and showed that, in all cases with paired circulating tumor cells and bone marrow, the circulating tumor cells in the blood reflected the molecular disease observed in the bone marrow. Furthermore, in a follow-up analysis of 5 subjects post-therapy, we detected rare malignant cells and showed that the residual malignant plasma cells shared most of their transcriptional state with the original cells at diagnosis. In summary, our work demonstrates that scRNA-seq is a powerful tool for dissecting the heterogeneity in individuals with multiple myeloma and identifies new pathways and potential targets for diagnosis and therapy in symptomatic and asymptomatic myeloma. Further sampling of a larger cohort of subjects pre- and post-therapy could help in prioritizing efficient and personalized treatment for multiple myeloma.

Table 1 | Clinical characteristics of newly diagnosed subjects

Characteristic (units)	Value
Newly diagnosed subjects (*N*)	**29**
Age	
Median (years)	65
Range (years)	40–84
Age distribution	
≤65 years (number of subjects)	15
>65 years (number of subjects)	14
Sex	
Male (number of subjects; percentage of *N*)	18; 62
Female (number of subjects; percentage of *N*)	11; 38
Non–immunoglobulin M MGUS (number of subjects; percentage of *N*)	**7; 24**
High-risk (number of subjects; percentage of MGUS)	6; 86
Low- to intermediate-risk (number of subects; percentage of MGUS)	1; 14
SMM (number of subjects; percentage of *N*)	**6; 21**
High-risk (number of subjects; percentage of SMM)	1; 17
Low- to intermediate-risk (number of subjects; percentage of SMM)	5; 83
Multiple myeloma (number of subjects; percentage of *N*)	**12; 41**
ISS stage[a]I (number of subjects; percentage of multiple myeloma)	4; 33
ISS stage[a]II (number of subjects; percentage of multiple myeloma)	4; 33
ISS stage[a]III (number of subjects; percentage of multiple myeloma)	4; 33
Cytogenetics[b]	
High-risk (number of subjects; percentage of multiple myeloma)	4; 33
Standard-risk (number of subjects; percentage of multiple myeloma)	8; 67
AL amyloidosis (number of subjects; percentage of *N*)	**4; 11**

[a]ISS, international scoring system; calculated using serum β2-microglobulin and albumin. [b]High-risk cytogenetics is defined as one or more of the following: 17p deletion, t(4; 14), t(14; 16), t(14; 20), 1q gain. Standard-risk cytogenetics is defined as all other abnormalities.

Results

Individuals with multiple myeloma display unique signatures that converge into common malignant pathways. To better understand the heterogeneity within and across individuals with multiple myeloma, we designed a protocol for single cell transcriptomic characterization of the bone marrow plasma cells as well as the circulating plasma cells from individuals with MGUS, SMM, multiple myeloma and AL amyloidosis. Our design was focused on maintaining the in situ RNA composition of the participants' samples by instant cooling in the operating room of the bone marrow and blood, followed by immediate sorting of fresh cells for massively parallel scRNA-seq (MARS-seq) analysis[18]. We calibrated a validated flow cytometry-based method for isolation of plasma cells ($CD38^+CD138^+$), linking the intensity of the markers in each cell with the cell's expression profile using an index-sorting strategy (Extended Data Fig. 1). This allows for retrospective analysis of surface marker combinations for each individual cell[19]. We obtained fresh bone marrow samples from 29 newly diagnosed subjects with plasma cell neoplasms (Table 1 and Supplementary Table 1). Profiling the normal diversity of plasma cells in an age-matched control cohort is essential to understand the heterogeneity of the normal and malignant plasma cell disease states. To obtain normal bone marrow from control individuals with ages similar to those in our cohort of diagnosed subjects, we selected 11 older adult and elderly subjects (median age, 64 years; range, 45–83 years; 5 males and 6 females) with isolated hip osteoarthritis and without other medical comorbidities or active inflammatory processes and extracted bone marrow from the proximal femur bony canal during hip replacement surgery[20].

Following removal of low-quality cells, unsupervised clustering of 20,568 bone marrow plasma cells sorted from the 29 newly diagnosed subjects—7 with MGUS (MGUS01–07), 6 with SMM (SMM01–06), 12 with multiple myeloma (MM01–12) and 4 with AL amyloidosis (AL01–04)—as well as 11 control subjects (Hip01–11) created a detailed map comprising 29 transcriptionally homogeneous subpopulations covering the spectrum of plasma cell neoplasms (Fig. 1a, Extended Data Figs. 2–4 and Supplementary Tables 2 and 3). Despite a stringent sorting scheme for bone marrow plasma cells, 3,179 contaminating (non-plasma) cells were in silico removed prior to clustering on the basis of their transcriptional signatures (Extended Data Fig. 3; see Online Methods). Plasma cell subpopulations were based on cluster-specific expression patterns of the 1,500 most variable genes, discarding immunoglobulin genes (Extended Data Fig. 4). Clusters C1 and C2 are associated with the control group of hip replacement individuals, representing normal plasma cells with minor donor-specific enrichments in these clusters (Supplementary Table 3). Cluster C2 represents long-

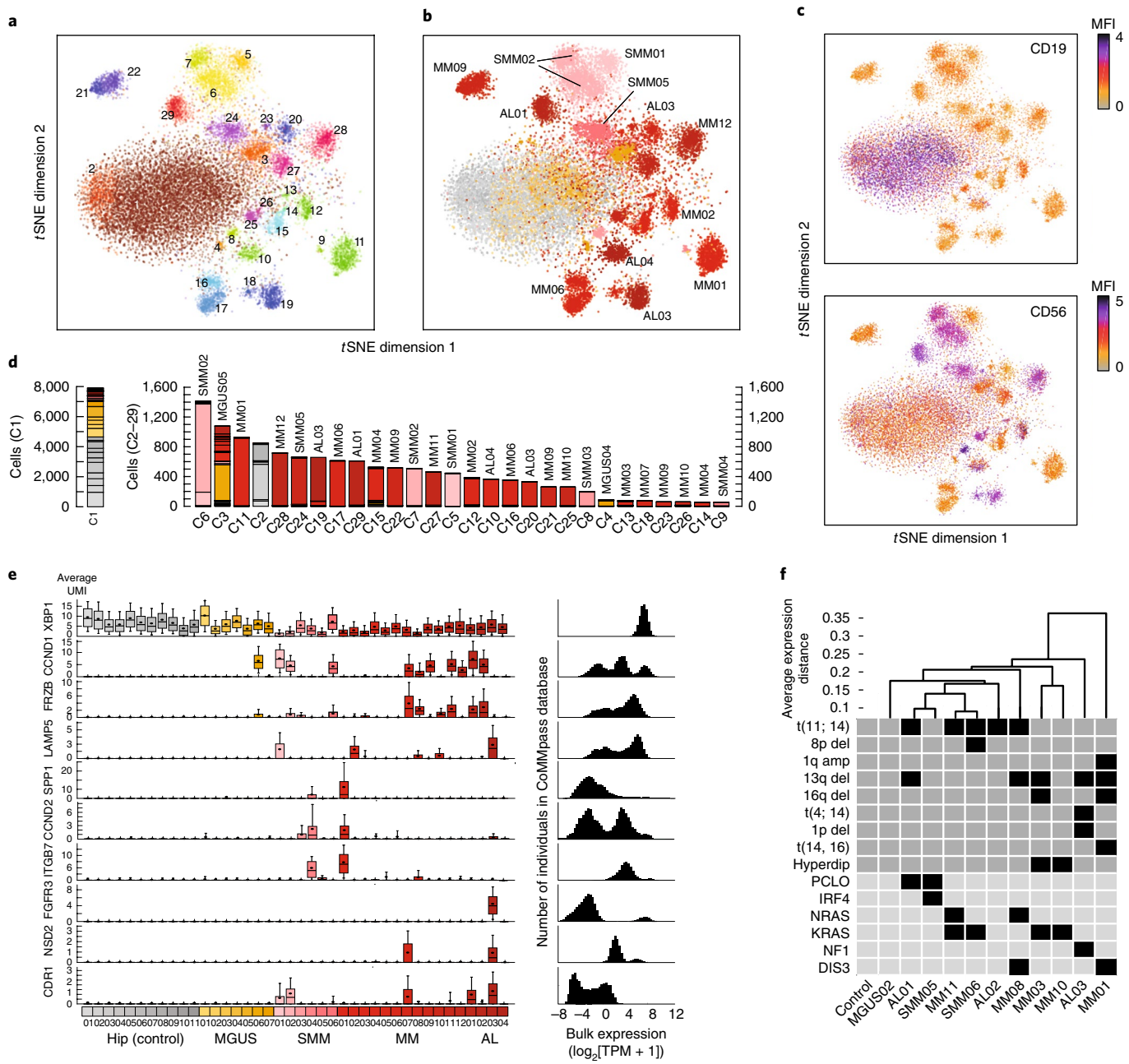


Fig. 1 | Subjects with multiple myeloma display unique transcriptional signatures that converge into defined malignant pathways. a, *t*SNE plot depicting 20,568 single bone marrow plasma cells derived from 29 newly diagnosed subjects (MGUS01–07, SMM01–06, MM1–12 and AL01–04) with plasma cell neoplasms and 11 control individuals (Hip01–11). Each cluster is represented by a specific color and number (related to the heat map in Extended Data Fig. 4). **b**, Subjects are color-coded according to a severity gradient projected on the *t*SNE (gray, control; yellow, MGUS; pink, SMM; red, multiple myeloma and amyloidosis); these colors correspond to **d** and **e**. **c**, Index-sorting flow cytometry data represented as mean fluorescence intensity (MFI; $\log_{10}$ scale) for specific surface markers, projected onto the *t*SNE (top, CD19; bottom, CD56). **d**, Bar plot showing distribution of cells from subjects with multiple myeloma and control donors across the clusters (as in **a**). Subject are color-coded according to disease severity as in **a**; names above bars correspond to the individual with the majority of cells in each cluster. **e**, Box plots of single cell gene expression for specific genes across the 29 newly diagnosed subjects and 11 control donors (left). Each box represents 0.25–0.75 percentile of UMI count with line extension to 0.1-0.9 percentile; dot represents the mean UMI count. Subjects are color-coded according to disease severity. For each gene, corresponding histograms of bulk RNA-seq expression estimates from the CoMMpass study (TPM- transcripts per kilobase million; log scale) are shown (right). **f**, Map of CNAs (dark gray background) and oncogenic mutations (light gray background) depicted in black (bottom) and dendrogram of hierarchical clustering for average RNA profile of 11 participants (top) for whom targeted bulk genomic DNA sequencing data were available.

lived plasma cells, evident from high expression levels of *CXCR4* and *TXNIP* (Extended Data Fig. 4)[17,21]. Cluster C1 shows a similar transcriptional profile, with lower expression levels of *CXCR4* and *TXNIP* and a gradient of CD81 and CD19 protein surface marker expression (Extended Data Fig. 4). Notably, we detected a variable frequency of cells with normal plasma cell phenotype in most subjects, especially in the asymptomatic subjects (Fig. 1a–c). Interestingly, in these individuals, the normal plasma cells are

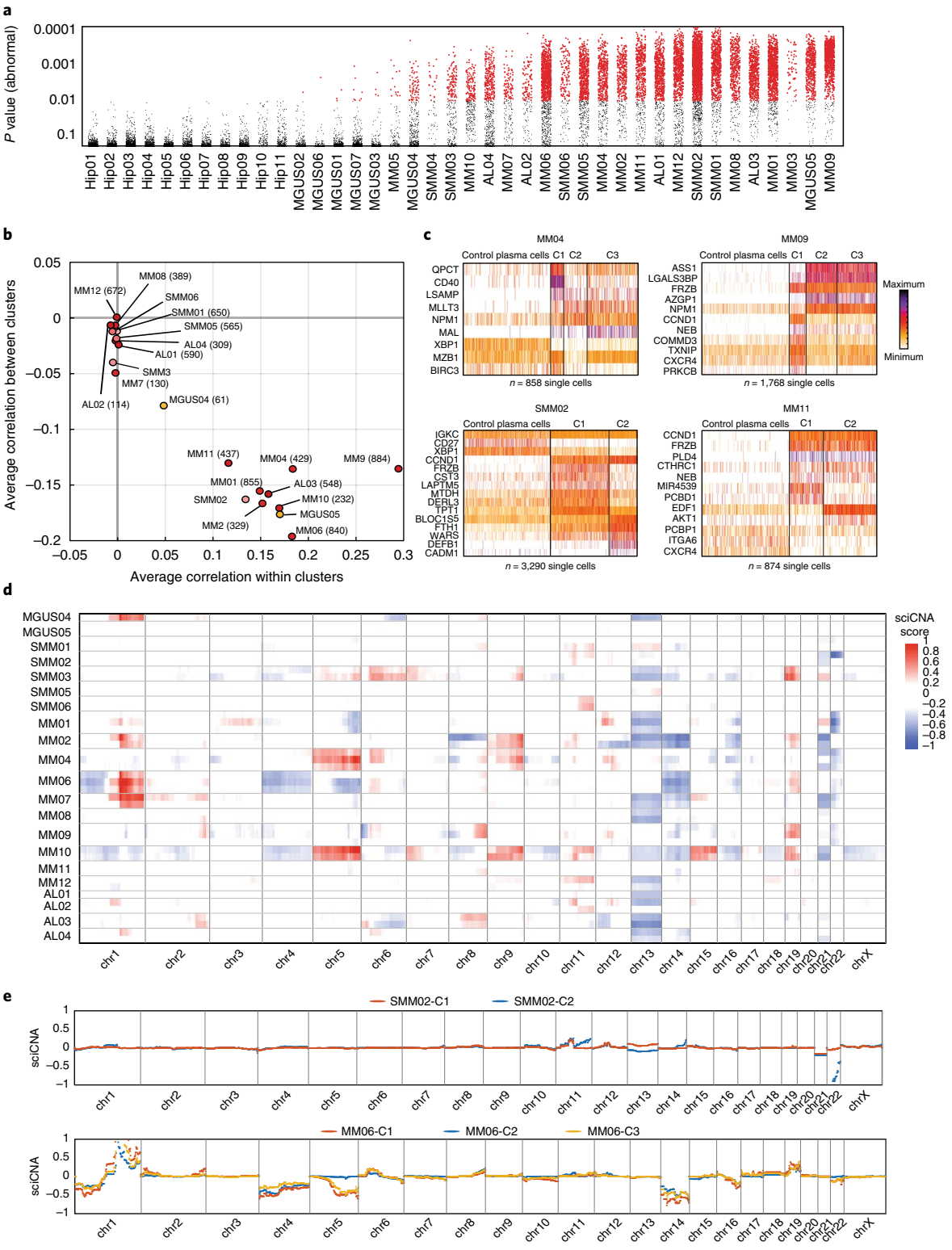


Fig. 2 | Intratumor heterogeneity in myeloma. a, Shown are *P* values to reject the null hypothesis (that a cell belongs to the control plasma cell group) for all 20,586 cells from 40 individuals, including those with multiple myeloma and controls. Dots represent individual plasma cells, classified as either normal (black) or abnormal (red; $P < 0.01$, one sided *t*-test with Bonferroni correction; see Online Methods). Individuals with multiple myeloma are ordered according to average score value from low to high. **b**, Intratumor heterogeneity measure for 21 individuals with abnormal plasma cells (as classified by the score used in **a**). Shown are Pearson correlations within and between clusters (for each individual separately). Correlation is calculated on the normalized log scale UMI count (see Online Methods). **c**, Heat maps showing clustering analysis of bone marrow plasma cells for subjects MM04 (top left; 429 cells), MM09 (top right; 884 cells), SMM02 (bottom left; 1,645 cells) and MM11 (bottom right; 437 cells), clustered with the same number of randomly sampled normal bone marrow plasma cells from control individuals. Representative variable genes are shown. Gradient shows RNA expression, row normalized (min to max value in each row). **d**, Heat map showing sciCNAs for each subject, averaged by intrasubject clustering. **e**, sciCNA profile for SMM02 (top) and MM06 (bottom); each line represents a cluster-averaged CNA profile.

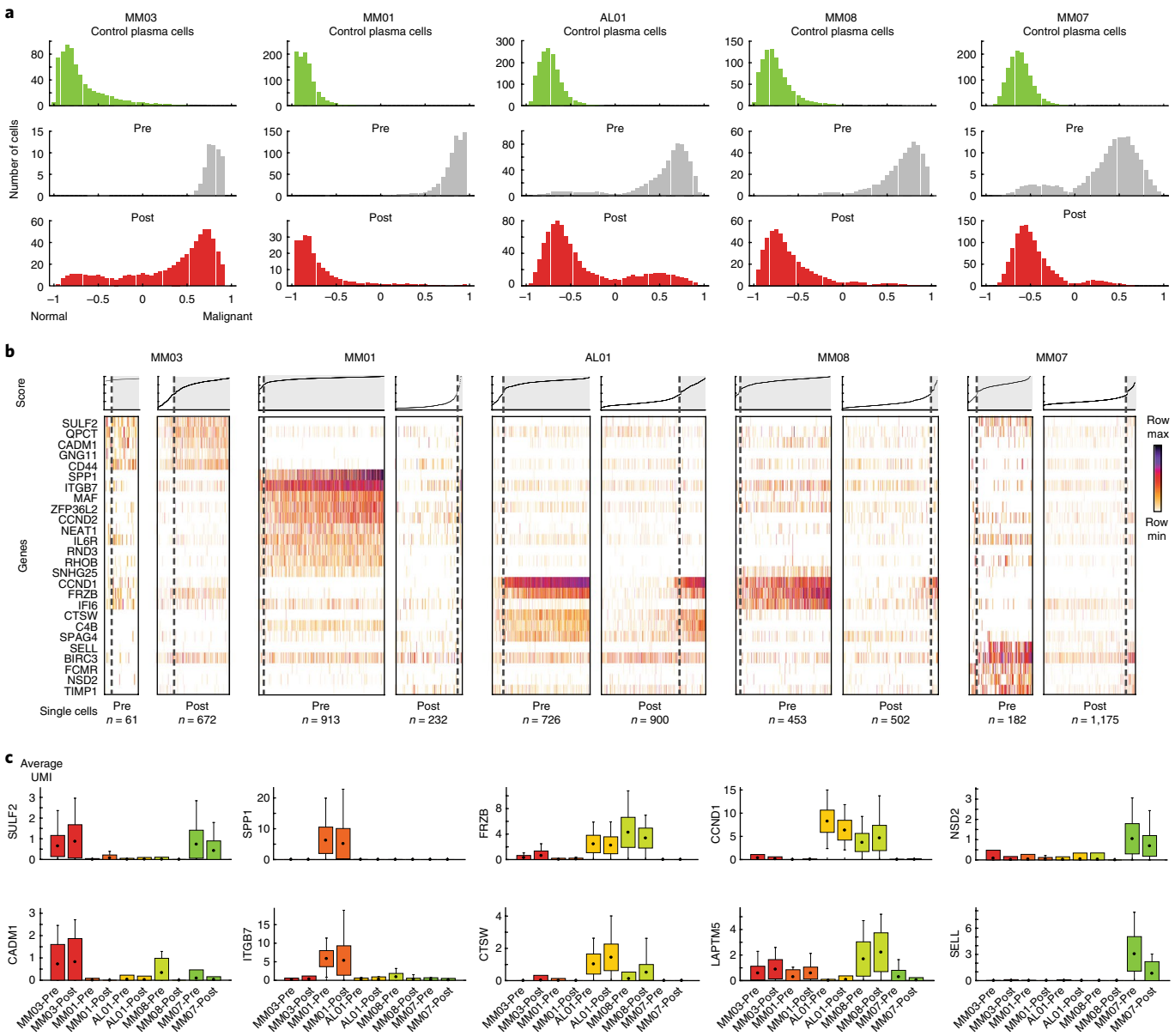


Fig. 3 | Characterization of rare residual malignant cells post-therapy. a, Histograms of normal to malignant score (Online Methods) for $n=5$ individuals with minimal residual disease, showing the distribution of scores for control cells (from donors; green, top) and for pre-treatment cells (gray, middle) and post-treatment cells (red, bottom) from subjects with minimal residual disease. **b**, Heat maps showing normalized single cell gene expression of bone marrow plasma cells pre- and post-treatment for $n=5$ individuals with minimal residual disease: subjects MM03 (61 and 672 cells, respectively), MM01 (913 and 232 cells, respectively), AL01 (726 and 900 cells, respectively), MM08 (453 and 502 cells, respectively) and MM07 (182 and 1,175 cells, respectively). Representative genes are shown. Cells are sorted by malignancy score (Online Methods), shown at the top; gray background represents the malignant cell fraction. **c**, Box plots showing gene expression of representative genes pre- and post-treatment for $n=5$ individuals with minimal residual disease. Each box represents 0.25–0.75 percentile of UMI count, with line extension to 0.1–0.9 percentile; dot represents the mean UMI count. Pre, pre-treatment; Post, post-treatment.

more similar to short-lived plasma cells expressing low levels of *CXCR4* and *TXNIP*. In contrast, clusters C3–29 represent subject-specific transcriptional state(s), with each subect characterized by an almost unique plasma cell transcriptional program (Fig. 1a–d and Extended Data Fig. 4). Although each subject is unique, we detected common overexpressed pathways shared across subgroups of participants, such as *CCND1*, *CCND2* and *NSD2-FGFR3* groups (Fig. 1e and Extended Data Fig. 3). *CCND1*-driven malignant plasma cells are observed across 11 subjects and can be found in 58.7% of the CoMMpass database (476 of 811 individuals; Fig. 1e). To validate whether these individuals harbor the *IGH* translocation, or whether they represent other overexpression mechanisms, we compared their genomic DNA interphase fluorescence in situ hybridization (iFISH) data (Online Methods). We found that 10 of 11 harbored an *IGH* translocation event. Clusters C20–22, C26 and C29 are represented by deregulation of the canonical wingless (Wnt) pathway, including overexpression of *SMAD* genes and the soluble frizzled related protein 3 (*FRZB*; $P<1\times10^{-50}$), also found in 68% of the CoMMpass database (553 of 811 individuals; Fig. 1e). The tyrosine kinase fibroblast growth factor receptor 3 (*FGFR3*, found in 75 of 811 individuals in the CoMMpass database), a known high-risk oncogene in myeloma, is featured in cluster C19

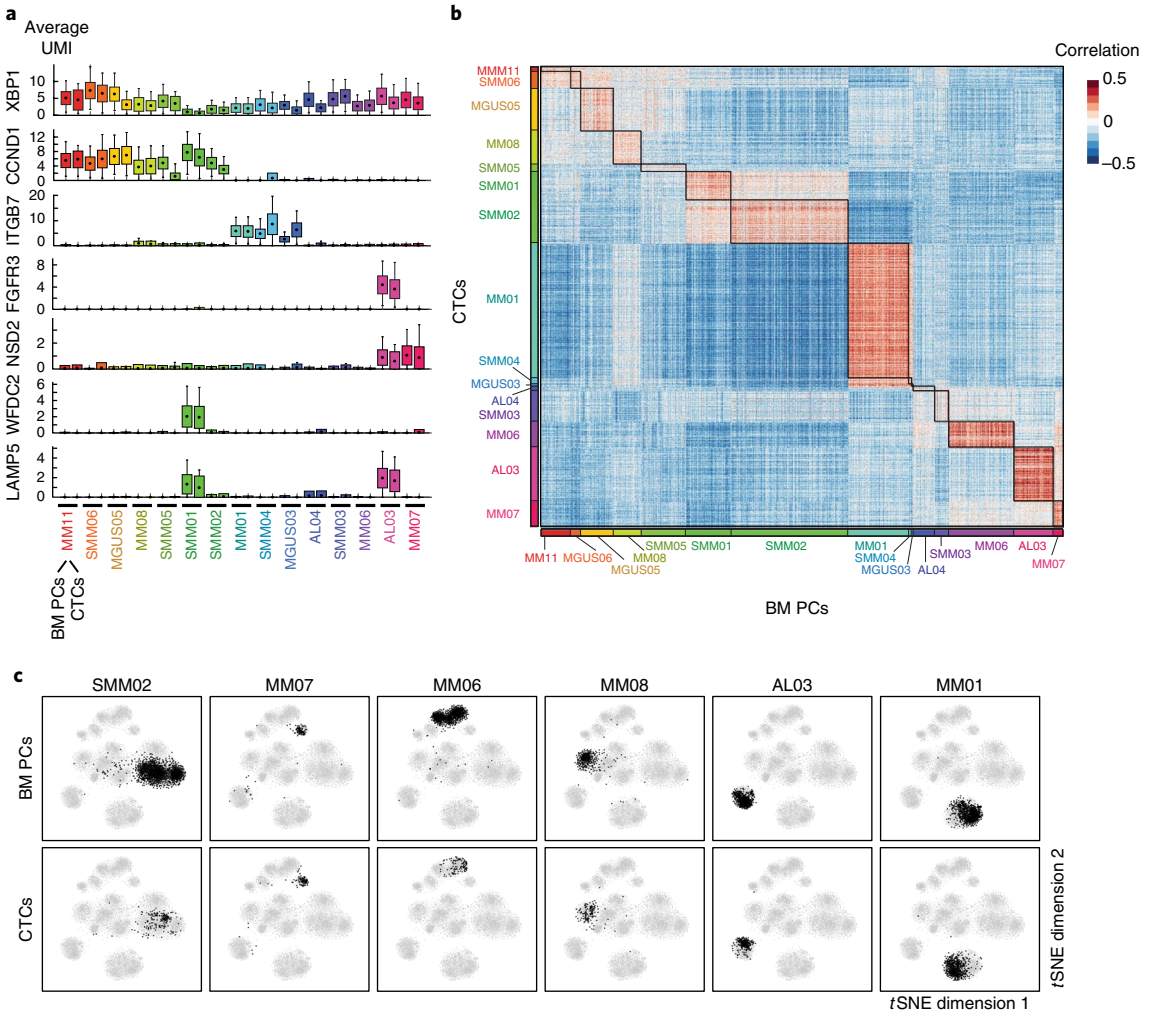


Fig. 4 | Circulating plasma cells are composed of circulating tumor cells that reflect the bone marrow disease. a, Box plots showing gene expression of representative genes from 15 subjects for whom the number of circulating tumor cells $N_{CTC} > 20$. Each subejct is represented by a different color. Shown are pairs of bone marrow plasma cells and circulating plasma cells for each subject. Each box represents 0.25–0.75 percentile of UMI count, with line extension to 0.1–0.9 percentile; dot represents the mean UMI count. **b**, Correlation matrix for bone marrow plasma cells (*x* axis; 7,969 cells) and circulating tumor cells (*y* axis; 2,299 cells) across 15 subjects. Subjects' color codes correspond to **a**. Pearson correlation is calculated on the normalized log scale UMI count. **c**, Two-dimensional *t*SNE views of paired bone marrow plasma cells (top) and circulating tumor cells in the blood (bottom) from the same subject. Projection of single cells (black) from a specific individual (subject SMM02, 1,755 bone marrow plasma cells and 216 circulating tumor cells; subject MM07, 274 bone marrow plasma cells and 138 circulating tumor cells; subject MM06, 976 bone marrow plasma cells and 128 circulating tumor cells; subject MM08, 416 bone marrow plasma cells and 164 circulating tumor cells; subject AL03, 589 bone marrow plasma cells and 267 circulating tumor cells; subject MM01, 671 bone marrow plasma cells and 141 circulating tumor cells) is shown on a gray background of all bone marrow plasma cells and circulating tumor cells (10,268 cells). BM PCs, bone marrow plasma cells; CTCs, circulating tumor cells.

(subject AL03) and was confirmed by t(4; 14) iFISH testing (Fig. 1e and Extended Data Fig. 5)[22,23]. We also identified putative multiple myeloma overexpressed genes (all with $P < 1 \times 10^{-50}$) not found in the control cohort, including lysosome-associated membrane protein-like molecule 5 (*LAMP5*), a protein localized in the endoplasmic reticulum-Golgi compartment and regulated by toll-like receptor signaling[24], in 5 of 29 subjects (overexpressed in 52% of the CoMMpass database (425 of 811 individuals)); cerebellar degeneration gene 1 (*CDR1*), a protein with a yet unknown function, in 5 of 29 subjects (overexpressed in 3.5% of the CoMMpass database (29 of 811 individuals)); and WAP four-disulfide core domain protein 2 gene (*WFDC2*), a secreted proteinase, in 2 of 29 subjects (overexpressed in 6.7% of the CoMMpass database (55 of 811 individuals)) (Fig. 1e and Extended Data Fig. 3). *LAMP5*, *CDR1* and *WFDC2* were previously implicated in plasmacytoid dendritic cells, paraneoplastic syndromes and ovarian carcinoma, respectively, but not in multiple myeloma[25–27]. Since no mutations and/or aberrations were found near the *LAMP5* locus in individuals with multiple myeloma, we first checked for overexpression of *LAMP5* in a database of myeloma cell lines and identified a large number of multiple myeloma cell lines overexpressing *LAMP5* (28 of 75). We then performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) of histone H3 lysine 4 dimethyl regions (H3K4me2), which marks promoter and enhancer regions, in a *LAMP5* multiple myeloma overexpressing cell line (KHM1B) and in a cell line negative for *LAMP5* (RPMI-8226). While genes in proximity to *LAMP5* have a similar normalized H3K4me2 signal between the two cell lines, the *LAMP5* locus reproducibly shows several active regulatory regions only in the *LAMP5*-positive KHM1B cells (Extended Data Fig. 5). These may represent regulatory regions specific to *LAMP5* that are activated in multiple myeloma in trans. In conclusion, we have comprehensively profiled malignant and normal plasma cells

using scRNA-seq and have shown that even the most careful bulk plasma cell sampling contains significant contaminants from unrelated immune cells and normal plasma cells. Using the single cell resolution data of neoplastic plasma cells, we have identified that each individual has his or her own unique transcriptional signature, and we have verified previously implicated drivers along with putative overexpressed candidate genes.

The transcriptional state of plasma cells is regulated by the interplay of the genome, epigenome and environmental contexts[28]. To test whether DNA mutations and/or copy number alterations (CNAs) can contribute to the transcriptional heterogeneity we identified in the different subjects, we used a sensitive targeted approach to sequence the DNA regions involved in multiple myeloma (Online Methods). Profiling 11 subjects from our cohort, we found DNA aberrations similar to those reported in previous multiple myeloma studies[29,30]. However, most of the transcriptional divergence in our dataset cannot be explained by the DNA mutational status alone. For example, subject MM08, who harbors both t(11; 14) translocation with *NRAS* mutated subclones and a chromosome 13 deletion, was transcriptionally homogeneous (Fig. 1f and Supplementary Table 4), whereas subject MM10, who demonstrated substantial intratumor heterogeneity, as we show below, has no subclonal mutations (Figs. 1f and 2d,e). Together, our data show that tumor heterogeneity observed in multiple myeloma cannot be intuitively explained by the DNA mutational status alone. This suggests a possible role for rare intergenic non-coding mutations and/or trans-regulated epigenetic and environmental inputs governing the full extent of transcriptional heterogeneity observed in multiple myeloma.

Characterizing the intratumor heterogeneity of individuals with multiple myeloma. Since individuals with multiple myeloma display large patient-to-patient heterogeneity, we evaluated whether intratumor heterogeneity could also be observed in our cohort[8,13]. Plasma cells originate from post-germinal center and memory B cells and thus have a limited capacity to undergo further mutations in the immunoglobulin locus[28]. In agreement with this, myeloma cells typically conserve the immunoglobulin sequence through the course of tumor evolution[31]. To gain insight into the clonotypic identity of plasma cells within each participant, we used our scRNA-seq data to annotate the immunoglobulin κ and λ light chain variable regions (*IGVK* and *IGVL*, respectively) and coupled those with the immunoglobulin heavy chain constant region (*IGHC*) for every single cell (Extended Data Fig. 6; see Online Methods). Plasma cells from control donors were composed of diverse immunoglobulin sequences (Extended Data Fig. 6). Across the different controls, the most frequently represented *IGVK*/*IGVL* sequences correlated with their prevalence in the general population[32]. Conversely, within the group of individuals with multiple myeloma, we typically observed one specific immunoglobulin clonotype. This was further validated by a microfluidic platform for single cell B cell receptor (BCR) sequencing[33] (Extended Data Fig. 6). To characterize the intratumor transcriptional heterogeneity, we developed a *k*-nearest neighbors (*k*NN)-based machine learning classifier that segregates normal from malignant plasma cells (Fig. 2a; see Online Methods). The classifier annotates plasma cells according to their similarity to the signatures of normal plasma cells in our data. In the control donors, all but a few cells (≤4 of 1,000 cells) were classified as normal plasma cells, while in individuals with multiple myeloma, plasma cells were mostly classified as abnormal (Fig. 2a). After removal of normal plasma cells, we then re-clustered the abnormal plasma cells from every individual subject separately and analyzed the intercluster versus intracluster correlations (Fig. 2b; see Online Methods). We detected substantial intratumor heterogeneity, defined by negative intracluster correlation ($r < -0.1$) and positive intercluster correlation ($r > 0.1$), in 10 subjects (Fig. 2b). For example, subject SMM02 displayed a single BCR clonotype characterized by two distinct transcriptional states: one dominated by *DEFB1* ($p < 1 \times 10^{-50}$), a gene that was not previously implicated in myeloma and is reported to be a *CCR6* ligand, and the other by the expression of *FRZB*, a gene implicated in the oncogenic Wnt pathway in multiple myeloma (Fig. 2c and Supplementary Table 5)[34]. Subject MM11 exhibited two transcriptional states, both expressing high levels of *CCND1* and *FRZB*; one state was characterized by significant overexpression of endothelial differentiation related factor 1 (*EDF1*), involved in lipid metabolism and the PPARγ pathway[35,36], while the second transcriptional clonotype overexpresses *PCBD1*, a transcriptional co-activator of *HNF1*[37]. Conversely, the plasma cells of subject SMM01 (positive serum immunofixation for immunoglobulin A light chain κ) displayed two transcriptional clonotypes: a small clonotype with immunoglobulin heavy chain class A (*IGHA*) and a larger one expressing only the κ light chain, each with a distinct transcriptional signature (Extended Data Fig. 7).

To evaluate whether these clonal structures are the product of genomic aberrations, we used the scRNA-seq data to infer sciCNAs by averaging the relative expression of a large number of genomically adjacent genes[15]. We sorted all expressed genes by their genomic locations and used a moving average of 100 adjacent genes to estimate the chromosomal CNA in each cluster (Fig. 2d; see Online Methods). The expression levels were compared with the average signature of the control donor cells. We then compared the sciCNAs to the targeted genome sequencing for the same individuals (Fig. 2d and Extended Data Fig. 7). Our sciCNA analysis showed, as expected, that individuals with MGUS had a less aberrant CNA profile compared with individuals with multiple myeloma, who showed many 13q deletions, in agreement with the frequency of this aberration in myeloma (Fig. 2d). Importantly, using sciCNA, we found that several intra-individual transcriptional clones also harbored genomic aberrations, suggesting that some of the intrapopulation transcriptional diversity is likely driven by different genome structures (Fig. 2e). For example, only one cluster from subject SMM02 showed chromosome 22 deletion, while the other showed a normal sciCNA pattern. For subject MM06, sciCNA detected 1q amplification together with chromosomes 4 and 14 deletions in all 3 transcriptional clones, while a chromosome 5 deletion was found only in 2 transcriptional clones (Fig. 2d,e). Importantly, in most subjects, the differential genes did not necessary reside in the altered chromosomes, suggesting regulation in trans. In subject SMM02, the *FRZB*, *DEFB1*, *CST3* and *WARS* genes are expressed differentially but are not related to chromosome 22 deletion in cluster 1 (Fig. 2c–e). Together, these results show that intratumor transcriptional and CNA heterogeneity are prevalent in myeloma and that they can be characterized using scRNA-seq profiling. Further, our data suggest that in multiple myeloma, transcriptional changes are often regulated in trans, associated indirectly with the observed genomic aberrations. It will be important to follow the response to therapy and the risk of disease progression in individuals with single versus multiple transcriptional clones.

Characterizing rare cancer cells in individuals with asymptomatic and minimal residual disease. Individuals with asymptomatic disease are a highly heterogeneous group with varying risk of developing multiple myeloma. Currently, limited methods exist for stratifying these individuals' molecular signature and risk of progression. We expected these individuals to have a lower tumor burden than those with active multiple myeloma. We profiled with MARS-seq 7 subjects newly diagnosed with MGUS and 6 with SMM (Table 1). The disease manifestation of subjects with SMM, as characterized by MARS-seq, was dramatically different from that of subjects with MGUS in terms of the number of malignant cells and closely resembled the profiles of the subjects with multiple myeloma. Within the group of subjects with MGUS, we detected malignant clusters for 2 individuals, subjects MGUS04 and MGUS05, with 69 of 466 and 482

of 493 plasma cells, respectively, displaying a malignant signature (Fig. 2b). Clustering the plasma cells of subject MGUS04 showed that the malignant cells (cluster 4) are characterized by a transcriptional signature overexpressing a known multiple myeloma driver (*CCND1*), along with *DPEP3* ($P < 1 \times 10^{-50}$), a dipeptidase involved in arachidonic acid metabolism that has previously been associated with triple-negative breast cancer but not with multiple myeloma[38]. These malignant plasma cells originate from a single BCR clonotype. This analysis shows that scRNA-seq can be a highly sensitive approach to molecularly characterize even a small number of malignant cells in individuals with asymptomatic disease and can potentially be used for improved patient classification and preventive treatment to halt progression into a symptomatic disease.

We applied the same sensitive approach to individuals with residual disease by performing longitudinal scRNA-seq sampling of 5 participants: subjects MM01, MM03, MM07, MM08 and AL01, who were profiled in a dynamic fashion at time of diagnosis and post-treatment. These subjects were treated with a bortezomib-based regimen, and all except for subject AL01 underwent high-dose melphalan therapy with autologous stem cell transplantation. We were able to detect rare (as little as 2%) cancer cells in 5 of 5 subjects, with abnormal serum light chain ratios (Fig. 3a–c). Importantly, 2 of these subjects (MM01 and MM08) were clinically classified as complete responders according to the International Myeloma Working Group criteria[29]. Comparing the cancer cells before and after treatment, we found that most of the tumor cells expressed transcriptional programs similar to those of the original pre-treatment neoplastic cells from the same subject; specifically, we found that the major multiple myeloma drivers in these subjects, such as *CCND1*, *NSD2*/*MMSET* and *FRZB* are expressed equivalently before and after treatment (Fig. 3c). Although most genes overexpressed in multiple myeloma do not change post-treatment, we were able to detect significantly ($P < 1 \times 10^{-50}$) differentially expressed genes for a few of the subjects. For example, the gene *ELK2AP* (a member of the ETS oncogene family) was overexpressed in subject MM07 post-treatment compared with the pre-treatment signature, and the gene lymphocyte cytosolic protein 1 (*LCP1*), involved in calcium binding and previously related to several cancers but not to multiple myeloma, was overexpressed in both subjects AL01 and MM03 post-treatment (Extended Data Fig. 8)[39]. In two additional subjects without a baseline sample, we were able to detect as little as 23 neoplastic plasma cells (1.7% of plasma cells; with a defined multiple myeloma program). For example, in subject MM13 we observed overexpression of *MAF*, *ITGB7* and *CCND2*, and in subject MM14 we observed overexpression of *NSD2*/*MMSET*, *PDIA2, AZGP1* and *MDK* (Extended Data Fig. 8 and Supplementary Table 6). This demonstrates that scRNA-seq is a powerful tool to dissect plasma cell heterogeneity and identify rare neoplastic states in the setting of low tumor burden and minimal residual disease. The relative stability of multiple myeloma driver gene expression pre- and post-treatment suggests that targeted therapy approaches in minimal residual disease settings might constitute an effective strategy and warrant further study.

Circulating tumor cells display similar transcriptional states to the bone marrow tumor. In myeloma, and especially in its asymptomatic predecessor states, the circulating tumor cell load in the peripheral blood is low, complicating a non-invasive and accurate liquid biopsy analysis because of contamination by various immune cells and normal circulating plasma cells. A previous study utilizing whole-exome sequencing found that somatic single-nucleotide variants are shared between the blood and bone marrow in 84% of individuals with active multiple myeloma[40]. While that study used a subject-specific sorting strategy in cases with a positive aberrant surface marker, others chose a wider and less specific approach[41,42]. A prerequisite for non-invasive tumor assessment of asymptomatic states during follow-up watchful waiting, or in active disease to monitor response to treatment, is that the circulating tumor cells reflect the bone marrow disease. To test the potential of scRNA-seq applications for accurate circulating tumor cell characterization, we applied MARS-seq on plasma cells from both bone marrow and blood from 19 individuals with multiple myeloma and 2 control donors (Hip09 and Hip10; Extended Data Fig. 2). In order to develop transcriptional and protein markers for efficient purification of circulating tumor cells from the participants' blood, we initially clustered the circulating plasma cells from all 21 individuals together. In addition to the 11 subject-specific clusters, we noted a shared cluster (cPC4) of polyclonal cells with a plasmablast signature common to most individuals, including cells from the control donors (Extended Data Fig. 9). Using flow cytometry in a different cohort of individuals with relapsed multiple myeloma, we showed that circulating tumor cells with an aberrant surface profile have lower protein expression of CD52 compared with non-circulating tumor cells (Extended Data Fig. 10). In order to compare, for each subject, the circulating tumor cells with his/her bone marrow tumor cells, we first removed the normal circulating plasma cells by excluding cells with cluster cPC4 characteristics (Online Methods). We also excluded 4 subjects with fewer than 20 circulating plasma cells from further analysis. Comparing the remaining malignant circulating plasma cells with the malignant bone marrow plasma cells for each subject, we observed that in all cases (15 of 15), the circulating tumor cell signatures highly resembled the bone marrow transcriptional state(s), with a few changes likely resulting from the different environments (such as expression of *CRIP1* and *KLF6*; Fig. 4a–c). To further validate our findings, we compared the BCR clonotype of the subjects' bone marrow plasma cells to that of circulating tumor cells. The tumor load in the bone marrow and the blood differs by several orders of magnitude, affecting the confidence in our analysis of a few individuals with small circulating tumor cell clones (<20 cells; see Online Methods). Overall, in 11 of 15 subjects we found a good match in the BCR between bone marrow and blood samples (Extended Data Fig. 9). Taken together, our results suggest that circulating plasma cells in the subjects' blood are composed of clonotypic circulating tumor cells that reflect the transcriptional status of the bone marrow disease, as well as additional normal polyclonal plasmablasts. We further devised an efficient sorting strategy for circulating tumor cells by excluding contamination by circulating plasmablasts. Our approach can be applied to molecularly characterize a patient's malignant plasma cells in an iterative fashion using liquid biopsies, omitting the need for invasive bone marrow sampling.

Discussion

We report on a new methodology for sensitive characterization of the entire spectrum of clinical progression from normal plasma cells to multiple myeloma using scRNA-seq. Data on thousands of plasma cells from 11 control donors was used to characterize plasma cell heterogeneity within normal bone marrow samples, showing a polyclonal BCR repertoire and limited interindividual transcriptional variation. Based on this reference, scRNA-seq provides high sensitivity and confidence to identify and characterize neoplastic plasma cells in low-burden disease settings, such as asymptomatic MGUS, and suggests a direct molecular assay for tracking early multiple myeloma onset. We found that individuals with SMM, although asymptomatic, are indistinguishable at the molecular level from individuals with active multiple myeloma. In fully active and symptomatic disease, scRNA-seq leads to precise molecular characterization of the malignant state and to frequent identification of multi-clonal structure, offering important potential for guiding and optimizing personalized treatments and a better understanding of post-treatment resistance. Following successful treatment and remission, scRNA-seq enables sensitive and precise detection of rare residual neoplastic cells. Importantly, our

methodology is compatible with analysis of circulating tumor cells and opens the way to routine non-invasive profiling of patients who must be monitored during pre-myeloma stages or post-treatment. We further showed that scRNA-seq data can enable accurate inference of CNAs in multiple myeloma, unexpectedly revealing that trans-acting mechanisms rather than cis DNA mutations or aberrations dominate the tumor plasma cell transcriptional state, as we showed for *LAMP5*.

This study also highlights several remaining challenges. Exploring the immune microenvironment together with plasma cells from the same individual may highlight potential new targets for immunotherapy and could predict response to specific treatments. Individuals with multiple myeloma may have a patchy infiltration pattern in the bone marrow, and by sampling a single bone marrow site during a routine clinical diagnostic procedure, we may underestimate the true heterogeneity within the tumor[43]. We note that we have used a 3′-based messenger RNA sequencing method, and we are therefore limited to inferring coding sequence mutations and splice variants. This can potentially be addressed by using full-length scRNA-seq methods[44].

In the last decade, there has been an immense progress in the treatment of myeloma. Unfortunately, despite a surge of new approved drugs and treatment modalities, relapse is still the rule, and detailed understanding of the reasons for successful or failed treatments remains limited. This study introduces scRNA-seq as a key technology for precise molecular profiling of individuals with myeloma at various stages of the disease and facilitates the design of new and molecularly informed diagnosis and treatment strategies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41591-018-0269-2.

Received: 6 July 2018; Accepted: 29 October 2018;
Published online: 6 December 2018

References

1. Rajkumar, S. V. Multiple myeloma: 2016 update on diagnosis, risk-stratification, and management. *Am. J. Hematol.* **91**, 719–734 (2016).
2. Kyle, R. A. et al. Monoclonal gammopathy of undetermined significance (MGUS) and smoldering (asymptomatic) multiple myeloma: IMWG consensus perspectives risk factors for progression and guidelines for monitoring and management. *Leukemia* **24**, 1121–1127 (2010).
3. Morgan, G. J., Walker, B. A. & Davies, F. E. The genetic architecture of multiple myeloma. *Nat. Rev. Cancer* **12**, 335–348 (2012).
4. Dhodapkar, M. V. MGUS to myeloma: a mysterious gammopathy of underexplored significance. *Blood* **128**, 2599–2606 (2016).
5. Bolli, N. et al. A DNA target-enrichment approach to detect mutations, copy number changes and immunoglobulin translocations in multiple myeloma. *Blood Cancer J.* **6**, e467 (2016).
6. Chapman, M. et al. Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
7. Egan, J. et al. Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. *Blood* **120**, 1060–1066 (2012).
8. Lohr, J. G. et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell* **25**, 91–101 (2014).
9. Walker, B. et al. Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t(4; 14) and t(11; 14) myeloma. *Blood* **120**, 1077–1086 (2012).
10. Laganà, A. et al. Integrative network analysis identifies novel drivers of pathogenesis and progression in newly diagnosed multiple myeloma. *Leukemia* **32**, 120–130 (2018).
11. Shah, V. et al. Prediction of outcome in newly diagnosed myeloma: a meta-analysis of the molecular profiles of 1905 trial patients. *Leukemia* **32**, 102–110 (2018).
12. Shaughnessy, J. D. et al. A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* **109**, 2276–2284 (2007).
13. Bolli, N. et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).
14. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
15. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
16. Giladi, A. & Amit, I. Single-cell genomics: a stepping stone for future immunology discoveries. *Cell* **172**, 14–21 (2018).
17. Paiva, B. et al. Differentiation stage of myeloma plasma cells: biological and clinical significance. *Leukemia* **31**, 382–392 (2017).
18. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
19. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
20. Juneja, S., Viswanathan, S., Ganguly, M. & Veillette, C. A simplified method for the aspiration of bone marrow from patients undergoing hip and knee joint replacement for isolating mesenchymal stem cells and in vitro chondrogenesis. *Bone Marrow Res.* **2016**, 1–18 (2016).
21. Halliley, J. et al. Long-lived plasma cells are contained within the $CD19^{-}CD38^{hi}CD138^{+}$ subset in human bone marrow. *Immunity* **43**, 132–145 (2015).
22. Chesi, M. et al. Frequent translocation t(4;14)(p16.3; q32.3) in multiple myeloma is associated with increased expression and activating mutations of fibroblast growth factor receptor 3. *Nat. Genet.* **16**, 260–264 (1997).
23. Pawlyn, C. & Morgan, G. Evolutionary biology of high-risk multiple myeloma. *Nat. Rev. Cancer* **17**, 543–556 (2017).
24. Combes, A. et al. BAD-LAMP controls TLR9 trafficking and signalling in human plasmacytoid dendritic cells. *Nat. Commun.* **8**, 913 (2017).
25. Defays, A. et al. BAD-LAMP is a novel biomarker of nonactivated human plasmacytoid dendritic cells. *Blood* **118**, 609–617 (2011).
26. Fathallah-Shaykh, H., Wolf, S., Wong, E., Posner, J. B. & Furneaux, H. M. Cloning of a leucine-zipper protein recognized by the sera of patients with antibody-associated paraneoplastic cerebellar degeneration. *Proc. Natl Acad. Sci. USA* **88**, 3451–3454 (1991).
27. Hellström, I. et al. The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma. *Cancer Res.* **63**, 3695–3700 (2003).
28. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nat. Rev. Immunol.* **15**, 160–171 (2015).
29. Kumar, S. K. & Rajkumar, S. V. The multiple myelomas—current concepts in cytogenetic classification and therapy. *Nat. Rev. Clin. Oncol.* **15**, 409–421 (2018).
30. Rajan, A. M. & Rajkumar, S. V. Interpretation of cytogenetic results in multiple myeloma for clinical practice. *Blood Cancer J.* **5**, e365 (2015).
31. Puig, N. et al. The predominant myeloma clone at diagnosis, CDR3 defined, is constantly detectable across all stages of disease evolution. *Leukemia* **29**, 1435–1437 (2015).
32. Lefranc, M.-P. et al. IMGT, the international ImMunoGeneTics information system 25 years on. *Nucleic Acids Res.* **43**, D413–D422 (2015).
33. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308 (2018).
34. Tian, E. et al. The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. *N. Engl. J. Med.* **349**, 2483–2494 (2003).
35. Zhao, X.-Y. Y. et al. Long noncoding RNA licensing of obesity-linked hepatic lipogenesis and NAFLD pathogenesis. *Nat. Commun.* **9**, 2986 (2018).
36. Leidi, M., Mariotti, M. & Maier, J. Transcriptional coactivator EDF-1 is required for PPARγ-stimulated adipogenesis. *Cell. Mol. Life Sci.* **66**, 2733–2742 (2009).
37. Simaite, D. et al. Recessive mutations in PCBD1 cause a new type of early-onset diabetes. *Diabetes* **63**, 3557–3564 (2014).
38. Chen, X. et al. Prognostic value of diametrically polarized tumor-associated macrophages in multiple myeloma. *Oncotarget* **8**, 112685–112696 (2017).
39. Dubovsky, J. et al. Lymphocyte cytosolic protein 1 is a chronic lymphocytic leukemia membrane-associated antigen critical to niche homing. *Blood* **122**, 3308–3316 (2013).
40. Mishima, Y. et al. The mutational landscape of circulating tumor cells in multiple myeloma. *Cell Rep.* **19**, 218–224 (2017).
41. Lohr, J. et al. Genetic interrogation of circulating multiple myeloma cells at single-cell resolution. *Sci. Transl. Med.* **8**, 363ra147 (2016).
42. Manier et al. Whole-exome sequencing of cell-free DNA and circulating tumor cells in multiple myeloma. *Nat. Commun.* **9**, 1691 (2018).
43. Rasche et al. Spatial genomic heterogeneity in multiple myeloma revealed by multi-region sequencing. *Nat. Commun.* **8**, 268 (2017).

44. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).

Acknowledgements

We thank the participants and their families; clinical coordinators D. Yaish, I. Maman and S. Levy; N. Voskoboinik for iFISH analysis; K. Kogan for help with REDCap installation and A. Giladi for review of the manuscript. We thank Brian Fritz and Tarjei Mikkelsen from 10x Genomics for help and support with the Chromium single cell 5′ and V(D)J kits. This study was partly supported by the Benoziyo Family Fund, Clalit Health Care Services and Mrs. and Mr. Barry Lang. I.A. is supported by the Chan Zuckerberg Initiative (CZI); the Howard Hughes Medical Institute International Scholar award; the European Research Council Consolidator Grant (ERC-COG) 724471-HemTree2.0; Melanoma Research Alliance (MRA) Established Investigator Award (509044); the Israel Science Foundation (703/15); the Ernest and Bonnie Beutler Research Program of Excellence in Genomic Medicine; the Helen and Martin Kimmel award for innovative investigation; a Minerva Stiftung research grant; the Israeli Ministry of Science, Technology, and Space; the David and Fela Shapell Family Foundation; the NeuroMac DFG/Transregional Collaborative Research Center Grant,; and the Abramson Family Center for Young Scientists. A.T. is supported by the European Research Council (ERC) (scAssembly), the CZI, and the Flight Attendant Medical Research Institute. B.P. is supported by an ERC 2015 Starting Grant (MYELOMANEXT).

Author contributions

G.L. conceived and designed the study, performed and analyzed experiments, and wrote the manuscript. A.W. designed the study, performed bioinformatic analyses and wrote the manuscript. M.Z. performed flow cytometry sorting experiments. S.-Y.W performed analysis of single cell RNA sequencing. Y.C.C., M.E.G., N.S., H.M., M.K.-M., K.H.-T., M.S., D.B.Y., A.N., L.S., I.Avivi. provided input on experimental design and collected clinical data and participant samples. H.K.-S. and I.Y. performed experiments and maintained RNA-seq infrastructure. C.B. performed ChIP-seq experiments. R.R. Analyzed clinical data using REDCap. E.D. and J.J.K. performed bioinformatic analyses. V.Y., E.P., O.L. designed targeted genomic sequencing and analyzed data. A.O.-U., K.B.H. and S.I. performed FISH analysis. S.K., J.S.-M. and B.P. helped with design and FACS analysis. G.I.B. coordinated and supervised clinical data. A.T. supervised the project, designed and analyzed experiments and wrote the manuscript. I.A. supervised the project, designed and analyzed experiments and wrote the manuscript.

Competing interests

The authors declare the following competing interests: a patent application (US Provisional Patent Application No. 62/756,640) has been filed related to this work. O.L. declares receiving funding and/or honoraria from Adaptive, Amgen, Binding Site, BMS, Celgene, Cellectis, Glenmark, Janssen, Karyopharm, Pfizer, Seattle Genetics and Takeda.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-018-0269-2.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0269-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to I.A.

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

Methods

Collection of bone marrow plasma cells during hip replacement surgery. Individuals with isolated hip osteoarthritis who were otherwise healthy were recruited by the orthopedic department at Tel Aviv Medical Center. The protocol was approved by the institutional review board committees of the Weizmann Institute of Science and Tel Aviv Medical Center. The procedure was performed in the operating room, as described previously[20], with several modifications. Briefly, after informed consent and general anesthesia that did not include corticosteroid use, the femoral canal was probed with a metal suction device following femoral neck removal. Bone marrow cells were suctioned into a sterile tube that contained heparin sodium (Pfizer) diluted with saline to 1,000 international units. Bony fragments were removed by forcing cells through a metal sieve, diluted 1:1 with ice-cold FACS buffer (2 mM EDTA pH 8.0, 0.5% BSA in PBS), placed on ice and immediately transported to the lab.

Obtaining subjects' plasma cells from iliac crest aspirates and peripheral blood. Individuals suspected for plasma cell neoplasm were recruited to the study from hematology departments at five medical centers in Israel. The protocol was approved by the Weizmann Institute of Science institutional review board committee, as well as those at each medical center. After informed consent, bone marrow aspirates and peripheral blood samples (20 ml) were placed in EDTA-containing tubes (Becton Dickinson). Tubes were mixed, placed on ice and immediately transported to the lab.

Participants' clinical and demographic data. Clinical study data were collected and managed using Research Electronic Data Capture (REDCap) tools hosted at the Weizmann Institute of Science. REDCap is a secure, web-based application designed to support data capture for research studies, providing an intuitive interface for validated data entry, audit trails and automated export procedures[45].

Single cell sorting. Bone marrow cells were diluted 2:1 in ice-cold FACS buffer, washed and strained with a 100 µm strainer. Peripheral blood cells were diluted 1:1 in ice cold FACS buffer. Mononuclear cell separation was performed by density centrifugation media (Ficoll-Paque; GE Healthcare Life Sciences) in a 1:1 ratio with diluted blood or marrow cells. Centrifugation (460 *g*, 25 min) was performed at 10 °C, and the mononuclear cells were carefully aspirated and washed with ice-cold FACS buffer. After red blood cell lysis (Sigma-Aldrich) for 5 min at 4 °C and washing, peripheral blood cells were enriched for CD38 with magnetic beads (Miltenyi), washed and stained with antibodies against CD38 (Cytognos, multi-epitope, cat. no. CYT38-F2), CD138 (BD, clone MI15, cat. no. 562935, CD56 (Cytognos, clone: C5.9, cat. no. CYT-56PE), CD19 (Beckman Coulter, clone J3.119, cat. no. IM3628U), CD117 (BD, clone 104D2, cat. no. 341096), CD27 (Biolegend, clone O323, cat. no. 302836), CD45 (PerCP-Cy5.5, Biolegend, clone EO1, cat. no. 368504), CD81(Cytognos, clone M38, cat. no. CYT-81AC750) or CD52 (BD, clone 4C8, cat. no. 563611). Bone marrow cells were stained without prior magnetic bead enrichment. Samples were filtered through a 40 µm strainer before sorting commenced. Single cell sorting was performed using either FACSAria II SORP or FACSAria Fusion (BD Biosciences). After doublet exclusion, isolated single cells were index-sorted into 384-well cell-capture plates containing 2 µl lysis solution and barcoded poly(T) reverse transcription primers for scRNA-seq. In each 384-well plate, 4 emply wells were kept as a no-cell control for data analysis. Immediately after sorting, each plate was spun down to ensure cell immersion in the lysis solution, snap-frozen on dry ice and stored at −80 °C until processed. To record surface marker levels of each single cell, the FACSDiva v.8 'index sorting' function was activated during single cell sorting. Following the sequencing and analysis of the single cells, each surface marker was linked to the genome-wide expression profile.

MARS-seq library preparation. Single cell libraries were prepared as previously described[18,19,46]. Briefly, mRNA from cells sorted into cell capture plates was barcoded and converted into complementary DNA and then pooled using an automated pipeline. The pooled sample was then linearly amplified by T7 in vitro transcription, and the resulting RNA was fragmented and converted into a sequencing-ready library by tagging the samples with pool barcodes and Illumina sequences during ligation, reverse transcription and PCR. Each pool of cells was tested for library quality, and concentration was assessed as described previously[18,19,46]. Overall, barcoding was done on three levels: cell barcodes allow attribution of each sequence read to its cell of origin, thus enabling pooling; unique molecular identifiers (UMIs) allow tagging each original molecule in order to avoid amplification bias; and plate barcodes allow elimination of the batch effect.

Analysis of scRNA-seq data. MARS-seq libraries, pooled at equimolar concentrations, were sequenced using an Illumina NextSeq 500 sequencer, at a sequencing depth of 50,000–100,000 reads per cell. Reads are condensed into original molecules by counting the same UMI. We used statistics on empty-well spurious UMI detection to ensure that the batches we used for analysis showed a low level of cross–single cell contamination (less than 3%). MARS-seq reads were processed as previously described[46]. Mapping of reads was done using HISAT (v.0.1.6); reads with multiple mapping positions were excluded. Reads were associated with genes if they were mapped to an exon, using the University of California, Santa Cruz (UCSC) Genome Browser (https://genome.ucsc.edu/) for reference. Exons of different genes that shared genomic position on the same strand were considered a single gene with a concatenated gene symbol. Cells with fewer than 500 UMIs were discarded from the analysis. Genes with mean expression les than 0.001 UMIs per cell or with above-average expression and low coefficient of variance (<1.2) were also discarded. Plasma cells were filtered based on immunoglobulin gene expression (sum over all annotated immunoglobulin genes) using a cut-off of 100 UMIs per cell. This cutoff was selected to discriminate the two modes of the dataset as fitted by two-Gaussian mixture model (plasma cells and non-plasma cells).

Graph-based clustering analysis. In order to assign cells to homogeneous clusters, we used the PhenoGraph clustering algorithm[47]. Low-level processing of MARS-seq reads results in a matrix U with n rows and m columns, where rows represent genes and columns represent cells. Entry U_{ij} contains the number of UMIs from gene i that were found in cell j. PhenoGraph first builds a kNN graph using the Euclidean distance ($k=30$) and then refines this graph with the Jaccard similarity coefficient, where the edge weight between each pair of nodes is the number of neighbors they share divided by the total number of neighbors they have[47]. To partition the graph into modules/communities, PhenoGraph uses the Louvain method. P values for differential expression analysis between different clusters were calculated using the Mann–Whitney U test with false discovery rate correction (Matlab R2016a *ranksum* function). In order to evaluate the robustness of our clustering analysis, we performed clustering with a more sensitive analysis package, MetaCell[48]. Briefly, informative genes were identified and used to compute cell-to-cell similarity to build a kNN graph that groups cells into cohesive groups (or meta-cells). Then, the algorithm uses bootstrapping to derive strongly separated clusters, as previously described[48]. We also compared the results with clustering using Seurat[49]. Our PhenoGraph clustering analysis shows great agreement with the analysis using MetaCell and Surat.

Two-dimensional projection. Cells were visualized in two dimensions using tSNE (Matlab R2017a *tsne* function).

Myeloma cell lines. RPMI-8226 and KHM1B myeloma cell lines were purchased from the American Type Culture Collection and the Japan Cell Repository Bank, respectively. Cells were cultured using an aseptic technique in RPMI medium (Gibco) supplemented with 10% heat-inactivated FBS, 1 mM sodium pyruvate, 2 mM L-glutamine, 1% penicillin–streptomycin (Thermo Fisher Scientific). Cells were stored in 10–50 ml flasks (Corning) in an incubator (Thermo Fisher Scientific) with humidified air and 5% CO_2 at 37 °C, at a concentration of 0.5–1 million cells per ml. Cell lines were validated for lack of mycoplasma infection using primers for mycoplasma-specific 16S rRNA gene region (EZ-PCR Mycoplasma Kit; Biological Industries). For flow cytometry intracellular staining, cells were stained with Live/Dead Violet (Invitrogen) and then washed and fixed/permeabilized with the Foxp3/Transcription Factor Staining Buffer Set (eBioscience), followed by staining of either control REA(I)-PE or anti-human LAMP5-PE (Miltenyi Biotech, clone REA590, cat. no. 130-109-203) and subsequent FACS analysis (FlowJo; BD Biosciences).

ChIP-seq library construction. ChIP and library preparation were performed as described previously, with a few modifications[50]. Briefly, viable cells (negative for Live/Dead Violet; Invitrogen) were sorted into FACS buffer, fixed for 10 min with 1% formaldehyde (Sigma-Aldrich) at room temperature, quenched with 0.125 M glycine and washed with ice-cold PBS. Cross-linked cells were resuspended in lysis buffer (12 mM Tris-HCl pH 8.0, 0.1× PBS, 6 mM EDTA) supplemented with protease inhibitor (Roche). Chromatin was sheared using an NGS Bioruptor Sonicator (Diagenode). The sonicated cell lysate (whole cell extract) was incubated with 2.5 µg of antibody against H3K4me2 (Abcam, EPR17707, ChIP Grade cat. no. ab176878) at 4 °C for 5 h, and then for an additional 1 h with protein G magnetic beads (Invitrogen). A 96-well magnet (Invitrogen) was used in all further steps. Cell lysate was removed, and samples were washed five times with cold RIPA buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 14 mM NaCl, 1% Triton X-100, 0.1% SDS, 0.1% DOC; 200 µl per wash), twice with RIPA buffer supplemented with 500 mM NaCl (200 µl per wash), twice with LiCl buffer (10 mM TE, 250 mM LiCl, 0.5% NP-40, 0.5% DOC) and once with TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and were then eluted in elution buffer (0.5% SDS, 300 mM NaCl, 5 mM EDTA, 10 mM Tris-HCl pH 8.0). The eluate was treated sequentially with 2 µl ribonuclease A (Roche) for 30 min and with 2.5 µl proteinase K (New England Biolabs) for 2 h and then reverse cross-linked overnight at 65 °C. DNA was purified by mixing reverse-cross-linked samples with paramagnetic solid-phase reversible immobilization beads (Agencourt AMPure XP; Beckman Coulter), incubated for 4 min. Beads were washed on the magnet with 70% ethanol and then air-dried for 4 min. The DNA was eluted in EB buffer (10 mM Tris-HCl pH 8.0). For the remainder of the library construction process (DNA end-repair, adenosine base addition, adaptor ligation and enrichment), the same solid-phase reversible immobilization bead cleanup was used. DNA ends were first repaired by T4 polymerase (New England Biolabs). Next, T4 polynucleotide kinase (New England

Biolabs) added a phosphate group at the 5′ ends. An adenosine base was then added to the blunt-ended fragments with Klenow enzyme (New England Biolabs), and a barcode Illumina compatible adaptor (Integrated DNA Technologies) was ligated to each fragment with T4 quick ligase (New England Biolabs). DNA fragments were amplified by 12 cycles of PCR (Kapa HiFi HotStart PCR ReadyMix; Kapa Biosystems) using specific primers (Integrated DNA Technologies) to the ligated adaptors. The quality of each library was analyzed by Tapestation (Agilent).

ChIP-seq data processing and analysis. All H3K4me2 libraries were sequenced using Illumina's NextSeq 500. Reads were aligned to the human reference genome (hg38) using the Bowtie 2 aligner v.2.3.4.1 (Johns Hopkins University; http://bowtie-bio.sourceforge.net/bowtie2/) with default parameters. The Picard tool 'MarkDuplicates' (Broad Institute; http://broadinstitute.github.io/picard/) was used to remove PCR duplicates. To identify regions of enrichment (peaks) from H3K4me2 reads, we used the Homer (University of California, San Diego; http://homer.ucsd.edu/homer/) package 'makeTagDirectory' followed by the 'findPeaks' command with the histone parameter using the appropriate whole cell extract control. Peaks from all samples were merged using 'mergePeaks' from the Homer package. Reads from all samples were counted using 'annotatePeaks' from Homer with the default Homer genome data hg38, merged peaks area file and the parameter -raw so as not to normalize by the default read count. Normalization of peaks was done by dividing reads inside peaks with the average of all reads. Data was visualized using the Interative Genomics Viewer (IGV, http://software.broadinstitute.org/software/igv/) with a 1-megabase window around the *LAMP5* gene.

Genomic iFISH. Bone marrow cells were enriched for CD138 using magnetic beads (Miltenyi), fixated in methanol and glacial acetic acid (3:1), placed on slides and hybridized with the following DNA probes: CKS1B/CDKN2C (P18) 1q21.3/1p32.3 Amplification/Deletion; IGH Plus 14q32.33 Breakapart; IGH/FGFR3 Plus, Dual Fusion 14q32.33/4p16.3 Translocation; IGH/MYEOV Plus, Dual Fusion 14q32.33/11q13.3 Translocation; P53 (TP53) 17P13.1 Deletion (Cytocell) according to the manufacturer instructions. For analysis, 50 nuclei were counted per slide. Karyotyping (G-banding) was performed using the non-enriched bone marrow fraction. Images were taken with a Nikon Ti-E inverted fluorescence microscope equipped with a ×100 oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software (Molecular Devices). The image-plane pixel dimension was 0.13 µm. Images were done on stacks of 15 optical sections with *Z* spacing of 0.3 µm.

Genomic DNA extraction. Bulk sorting of 100,000–500,000 plasma cells ($CD38^+CD138^+$; Extended Data Fig. 1) into PBS was performed using either FACSAria II SORP or FACSAria Fusion (BD Biosciences). After centrifugation (300 *g*, 10 min), the supernatant was aspirated and the pellet was snap-frozen. DNA extraction was performed with the Universal Quick-DNA Miniprep Kit (Zymo Research), and extracted DNA was quantified with a NanoDrop One spectrophotometer (Thermo Fisher Scientific).

Preparation of genomic DNA libraries for targeted sequencing. All 11 tumor samples and 16 unmatched bone marrow control samples (magnetically enriched for CD138) were subjected to a targeted custom sequencing approach using myTYPE. myTYPE is a custom capture panel designed to capture 120 recurrently mutated genes implicated in myeloma pathogenesis as well as *IGH* rearrangements and arm-level CNAs. The target-enrichment design was based on DNA pull-down by cRNA baits (SureSelect, Agilent Technologies, Santa Clara, CA). A total of 11 subject samples were pooled, and target DNA was subsequently enriched using one reaction tube, each from the SureSelect kit. All 22 samples were sequenced on a HiSeq2500 with a 100–base pair (bp) paired-end protocol.

Targeted genomic DNA sequencing analysis. *Alignment.* Short insert paired-end reads were aligned to the GRCh37 reference human genome with 1,000 genome decoy contigs using Burrows–Wheeler Aligner (BWA)-mem. After sequencing, we obtained a median of 21.2 million 100-bp paired-end reads per sample. After alignment, we obtained a median mean bait coverage of 758.6× per sample

Somatic mutation calling. Single-base substitutions were called using CaVEMan (Cancer Genome Project; http://cancerit.github.io/CaVEMan/). The algorithm compares sequence data from each tumor sample, albeit with an unmatched non-cancerous sample, and calculates a mutation probability at each genomic locus. To improve specificity, a number of post-processing filters were applied as follows: at least one-third of the alleles containing the mutant must have base quality of ≥25; if mutant allele coverage is ≥10×, there must be a mutant allele of at least base quality 20 in the middle third of a read; if mutant allele coverage is <10×, a mutant allele of at least base quality 20 in the first two-thirds of a read is acceptable; the mutation position is marked by <3 reads in any sample in the unmatched normal panel; the mutant allele proportion must be >5 times that in the unmatched normal sample (or it is zero in the unmatched normal); if the mean base quality is <20, then less than 96% of mutations carrying reads are in one direction. Mutations within simple repeats, centromeric repeats and regions of excessive depth (UCSC Genome Broswer; https://genome.ucsc.edu/) and low mapping quality were excluded. Additional unmatched normal filtering was performed using a set of unmatched normal samples. Mutations that were detected in >5% of the unmatched normal panel at ≥5% of the mutant allele burden were excluded. Variant annotation was done in Ensembl v.74 using VAGrENT.

Small insertions and deletions. Small somatic insertions and deletions (indels) were identified using a modified version of Pindel (Cancer Genome Project; https://github.com/cancerit/cgpPindel). To improve specificity, a number of post-processing filters were applied that required the following: for regions with sequencing depth <200×, the mutant variant must be present in at least 8% of total reads; for regions with sequencing depth ≥200×, the mutant variant must be present in at least 4% of total reads; the region with the variant should have ≤9 small (<4 nucleotides) repeats; the variant is not seen in any reads in the unmatched normal sample or the unmatched normal panel; the number of Pindel calls in the tumor sample is greater than 4; and either (1) the number of mutant reads mapped by BWA in the tumor sample is greater than zero or (2) the number of mutant reads mapped by BWA in the tumor sample is equal to zero, but there are no repeats in the variant region, and there are reads mapped by Pindel in the tumor sample on both the positive and negative strand; the Pindel 'SUM-MS' score (sum of the mapping scores of the reads used as anchors) must be ≥150. Additional unmatched normal filtering was performed using a set of unmatched normal samples ($n = 221$). Mutations that were detected in >1% of the unmatched normal panel at ≥1% mutant allele burden were excluded. Variant annotation was done in Ensembl v.74 using VAGrENT. For both substitutions and indels, variants that may have failed post-processing filtering criteria but mapped to recurrent oncogenic mutations in COSMCI were retained for manual curation.

Secondary pipelines for substitution and indel discovery and post-call. To identify subclonal variants present at very low frequencies in the tumor samples, mutation calling using secondary pipelines was done. Strelka2 (v.2.8.3)[51] was used to call point mutations and indels using tumor sample and matched normal. All 'PASS' calls were examined for their presence in CaVEMan and Pindel outputs. Calls uniquely identified by Strelka2 were retained for downstream analysis. We additionally examined the unfiltered calls from CaVEMan and Pindel that failed the criteria defined above and retained then for downstream analysis.

The following filters were applied on calls identified by primary and secondary pipelines: filter calls with >3% minor allele frequency in Exac (v.0.3) or 1000 Genomes; filter calls with >0.5% minor allele frequency in Exac or 1000 Genomes unless present in COSMIC (v.81); filter calls present in the panel of unmatched normal unless present in COSMIC; filter calls within the *IGH* locus and synonymous variants.

Cross-referencing with known myeloma datasets. Calls retained after applying the above filters were additionally annotated with variants from the Multiple Myeloma Research Foundation's CoMMpass Interim Analysis 9 exomes ($n = 889$). Calls were annotated if present at the exact genomic position with the exact mutation of if present in close proximity toa mutation (±9 bp). All calls retained were manually curated.

Structural rearrangements. Given the smaller fragment insert sizes in targeted capture, the 100-bp paired-end reads were trimmed to 50 bp from the 3′ end of the read for better discovery of intrastructural rearrangements. Alignment on the trimmed reads was performed as previously described, and structural rearrangements were detected by an in-house algorithm, BRASS (https://github.com/cancerit/BRASS), which first groups discordant read pairs that span the same breakpoint and then, using Velvet de novo assembler, performs local assembly within the vicinity to reconstruct and determine the exact position of the breakpoint to nucleotide precision. All calls having supported by fewer than 5 reads were excluded. Additionally, translocations in which either of the breakpoints is involved with the *IGH* locus and all deletions, inversions and tandem-duplications involving the *IGH* locus were excluded for downstream analysis. Additionally, an orthogonal pipeline using Delly (v.0.7.6) was used to identify structural rearrangements. Delly was run on each tumor sample using an unmatched control sample, and only those calls classified as 'PASS' by Delly were retained. All calls identified in the unmatched normal were also filtered. Additionally, for translocations, only those calls having at least 6 spanning reads and 2 junction reads or at least 30 spanning reads were retained.

Deletions, duplication and inversions. Passing thresholds were 6 spanning reads and 2 junction reads. As previously described for BRASS, translocations in which either of the breakpoints is involved with the *IGH* locus and all deletions, inversions and duplications involving the *IGH* locus were excluded for downstream analysis. The resulting calls retained after the described filters were manually curated.

CNAs. CNVKit (https://github.com/etal/cnvkit) was used to identify somatic CNAs in the data. To negate sample-specific biases in CNA analysis, all 16 control samples were combined into a pooled reference. Each tumor sample was then compared with the pooled reference to identify somatic CNAs in each sample.

CNVKit corrects for biases in regional coverage and GC content, according to the given reference, before calculating the log ratios between the built pooled reference and tumor. Subsequently, circular binary segmentation algorithm was applied to obtain the $\log_2$[fold change] values.

BCR variable region annotation from scRNA-seq data. In order to accurately extract BCR sequence annotation, we realigned the raw fastq reads using blastall (v.2.2.26, with *e*-value bound of 1×10^{-10}; https://blast.ncbi.nlm.nih.gov/Blast.cgi) to the the international ImMunoGeneTics information system reference sequences[32] (updated August 2016). For each cell we chose heavy chain constant region and light chain constant and variable region based on the highest coverage and longest coverage.

Single cell 5′ transcriptome and BCR sequencing (Chromium's 10×). For one individual with multiple myeloma (subect MM02) and one control individual (Hip13), the bone marrow mononuclear cell fraction was split; half proceeded to antibody staining and single cell sorting by MARS-seq, and the other half was enriched for CD138 using magnetic beads (Miltenyi), counted using light microscopy and trypan blue stain and then loaded onto a 10× Chromium microfluidics system according to the manufacturer's guidelines. Two sets of libraries were prepared from the 10× loaded samples: a 5′ mRNA library and a single cell BCR sequencing library, using custom primers for BCR amplification according to the manufacturer's instructions. The 5′ mRNA library was sequenced with Illumina's NextSeq 500 using 75 paired-end reads at a coverage of 48,105 mean reads per cell. The single cell BCR sequencing library was sequenced with Illumina's NextSeq 500 using 150 paired-end reads, at a coverage of 6,711 reads per cell. Data was analyzed using Chromium's Cell Ranger pipeline with default parameters. For scRNA-seq data to sequence BCR, we used Chromium's V(D)J-Loupe for analysis of BCR clonotypes and for visualization, with default parameters.

***k*NN classifier for normal and abnormal plasma cells.** In order to classify cells into normal/abnormal phenotype, we used a *k*NN-based classifier. Our method was based on the observation that the transcriptome profiles of malignant cells are very different from those of normal plasma cells, as can be clearly observed in our dataset. We first calculated the similarity between a given plasma cell and all normal plasma cells using Spearman correlation, and to prevent potential batch effects, we excluded cells sharing the same source (individual). We then selected the top $k=100$ most similar cells. We chose $k=100$ to ensure that the selected cells were from the same subpopuation, although we only observed two very close subpopulations in our deep analysis on the normal plasma cells. The distribution of the average Spearman correlation to the $k=100$ most similar cells was normally distributed. After estimating the normal distribution parameters, we used the one-sample *t*-test with Bonferroni correction for multiple tests to calculate the *P* value of each cell as normal. The lower the *P* value, the more likely cell is to be 'abnormal' (malignant).

Intratumor heterogeneity score. To detect heterogeneity within each subect, we first determined the number of clusters per subject (k) and decided whether the differences between clusters was significant enough to define two or more transcriptional clones. We combined supervised and unsupervised analyses to determine the number of different tumor clones per subject. We clustered each subject separately after in silico-removing normal plasma cells (based on the *k*NN classifier described above), allowing higher sensitivity in detecting changes of a relatively smaller number of genes. For each subject we calculated a heterogeneity score by comparing the average cell-to-cell correlation within and between clusters. Correlation is calculated on the normalized log UMI count (X):

$$X_i^g = \log_2\left(\frac{N\times U_i^g}{\sum_{j=1}^{N} U_j^g}\right)$$

where U_i^g is the UMI counts of gene g in cell i and N is the total number of cells. For individuals with substantial transcriptional heterogeneity, we would expect a negative intercluster correlation and a positive within-cluster correlation. Those with a uniform transcriptional state would have near-zero intercluster correlation. Finally, after devising this score, we have manually inspected each subject's clustering results, which confirmed our analysis and showed that indeed the subjects with the most substantial intratumor transcriptional heterogeneity showed negative intercluster correlation.

Gene ontology analysis. To gain insight into gene functions, we performed gene ontology analysis using Metascape (http://metascape.org/). We extracted the upregulated genes for each cluster with $\log_2$[fold change] > 1.5 and $-\log_{10}[P] > 10$ compared with cluster C1. The upregulated genes for each cluster were then provided as the input for Metascape to obtain enriched gene ontology terms and pathways.

sciCNA. CNAs were inferred from scRNA-seq as previously described[15] with modifications to MetaCell[48]. Briefly, we calculated the $\log_2$[fold change] for each cluster relative to the average expression profile of the control donors. Average expression was calculated using the log transformed data ($\log_2[1+\text{UMI}]$), and absolute values of fold change were bound by 3. For this analysis we used only genes with more than 100 UMIs for the control donor group. Finally, genes were sorted by their genomic location, and fold change was smoothed for each chromosome using a moving average over 100 adjacent genes.

Detecting tumor cells in longitudinal minimal residual disease samples. To detect rare malignant cells in longitudinal samples, we created a normal/malignant score based on the similarity of each post-treatment cell to the pre-treatment cells and to an equivalently sized group of normal plasma cells (sampled from the healthy cohort). For each subject i we defined the group G^i to contain N pre-treatment malignant cells and N normal control plasma cells (total of $2N$ cells). For each cell c from subject i post-treatment, we calculated the correlation to all cells in group G^i. Next, we sorte the vector of correlations and saved the order of the malignant and normal cells. For example, if we have 5 malignant cells and 5 control cells, the order of similarity (their rank order, from first to last) is as follows, writing H for a control cell and M for a malignant cell: MMMMHMHHHH.

Next, we assign numeric ranks to all cells and add up the ranks of cells, which come from the malignant pool. We calculate the statistic U as $U=R-\frac{N(N+1)}{2}$, where N is the number of malignant cells (pre-treatment) and R is the sum of the ranks of the malignant cells. In our example above, $R=1+2+3+4+6=16$ and $U=16-\frac{5\times6}{2}=1$. U is approximately normally distributed; we standardize the values of U and get a score between −1 and 1, where 1 represents all malignant cells exceed the healthy cells in our rank vector (MMMMMHHHHH) and −1 the opposite.

Flow cytometry analysis of circulating tumor cells from individuals with relapsed myeloma. Bone marrow and blood from 3 individuals with relapsed myeloma were analyzed, as previously described[52]. Aberrant plasma cells were identified either by antigen underexpression (CD19, CD27, CD38, CD45, CD52, CD81) or antigen overexpression (CD56, CD138). Data acquisition was performed in a FACSCantoII flow cytometer (BD Biosciences) using the FACSDiva 6.1 software (BD Biosciences). Data analysis was performed using the Infinicyt software (Cytognos). Principal component analysis quantifies the significance (contribution to principal component 1) of each surface marker to separate between 'ciculating tumor cell' and 'normal plasma cell'. Each row represents a different subject.

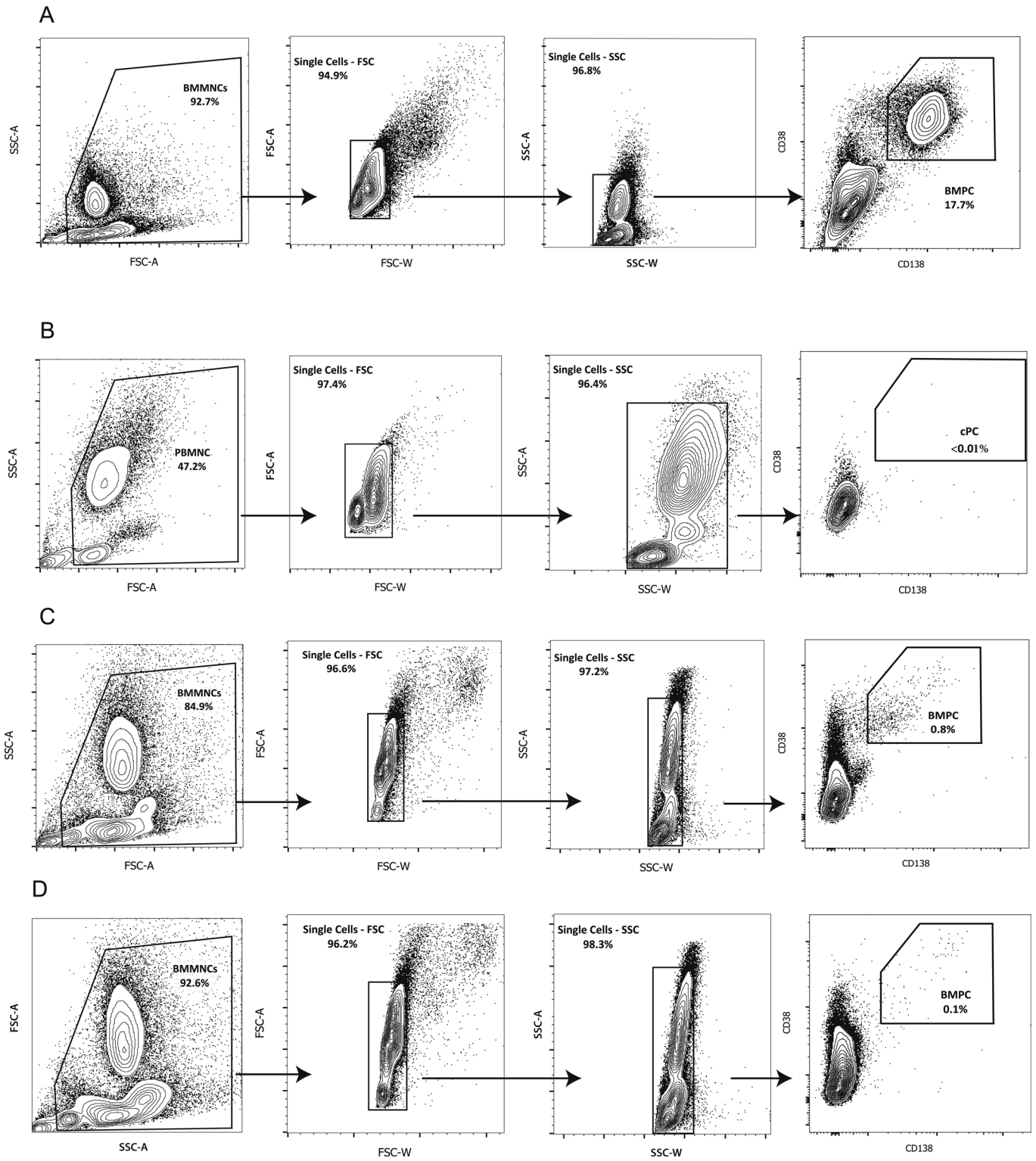
Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw and processed sc RNA-seq data and ChIP-seq data were deposited to the National Center for Biotechnology Information (NCBI)'s Gene Expression Omnibus with accession number GSE117156. Raw and processed genomic DNA targeted sequencing data were deposited to NCBI's Sequence Read Archive with accession number SRP165705. Source code used for scRNA-seq analysis can be found at https://bitbucket.org/amitlab/multiple-myeloma-2018/. The participats' clinical information is available in Supplementary Table 1.

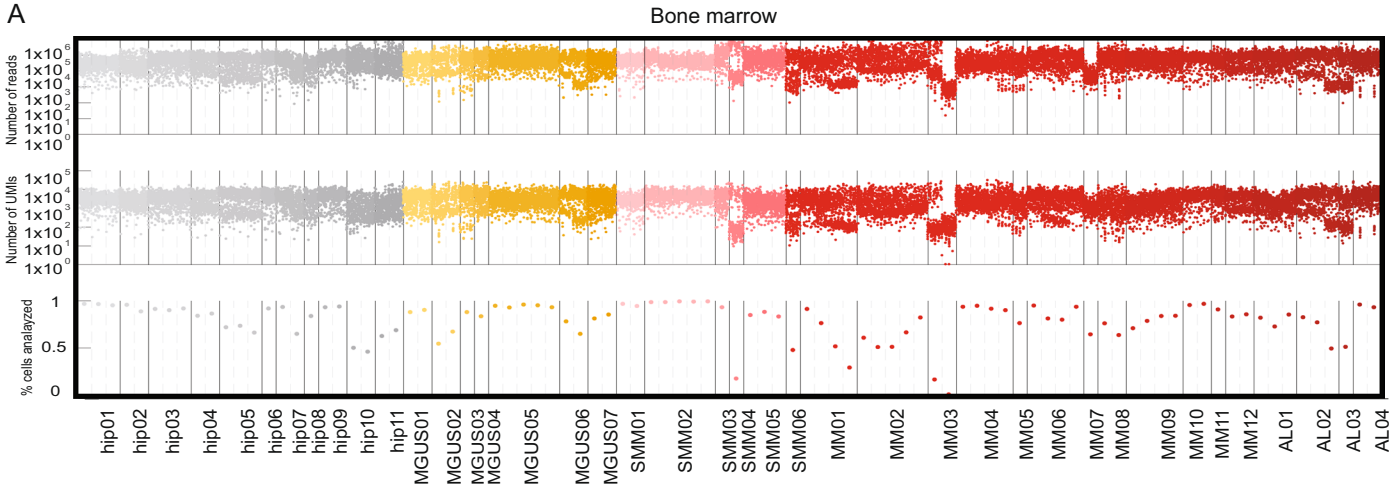
References

45. 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).
46. Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* **169**, 1276–1290 (2017).
47. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
48. Baran, Y. et al. MetaCell: analysis of single cell RNA-seq data using k-NN graph partitions. Preprint at bioRxiv https://doi.org/10.1101/437665 (2018).
49. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
50. Blecher-Gonen, R. et al. High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nat. Protoc.* **8**, 539–554 (2013).
51. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
52. Flores-Montero, et al. Next generation flow for highly sensitive and standardized detection of minimal residual disease in multiple myeloma. *Leukemia* **31**, 2094 (2017).

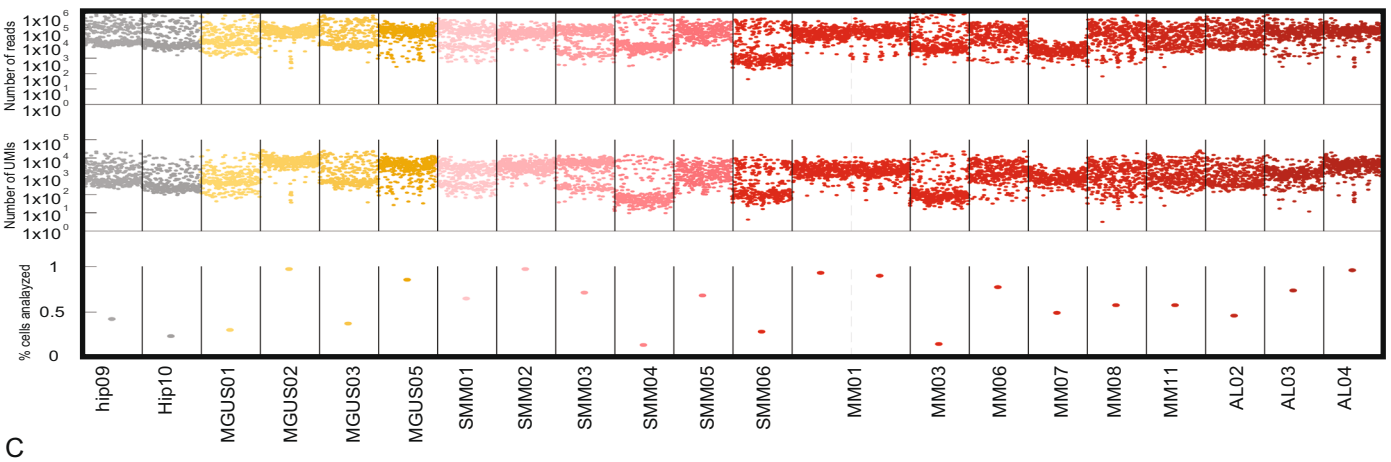


Extended Data Fig. 1 | Sorting strategy of plasma cells: representative flow cytometry plots showing sorting strategy ($CD38^+CD138^+$) for plasma cells after doublet exclusion. Sorting experiments were performed for 40 individuals with multiple myeloma and controls. For 21 subjects, additional sorting was performed for circulating plasma cells from peripheral blood (PB) (separately). **a**, Bone marrow plasma cells of subject MM08 (active myeloma). **b**, Circulating plasma cells of subject MM04 (active myeloma). **c**, Bone marrow plasma cells of subject MGUS05. **d**, Bone marrow plasma cells of subject MM13 (active myeloma with minimal residual disease). Plots were generated using FlowJo software (Online Methods). During sorting, surface marker expression for additional markers in the Euroflow panel (CD19, CD81, CD27, CD56, CD117, CD45) was recorded for each single cell (Online Methods).

A



B



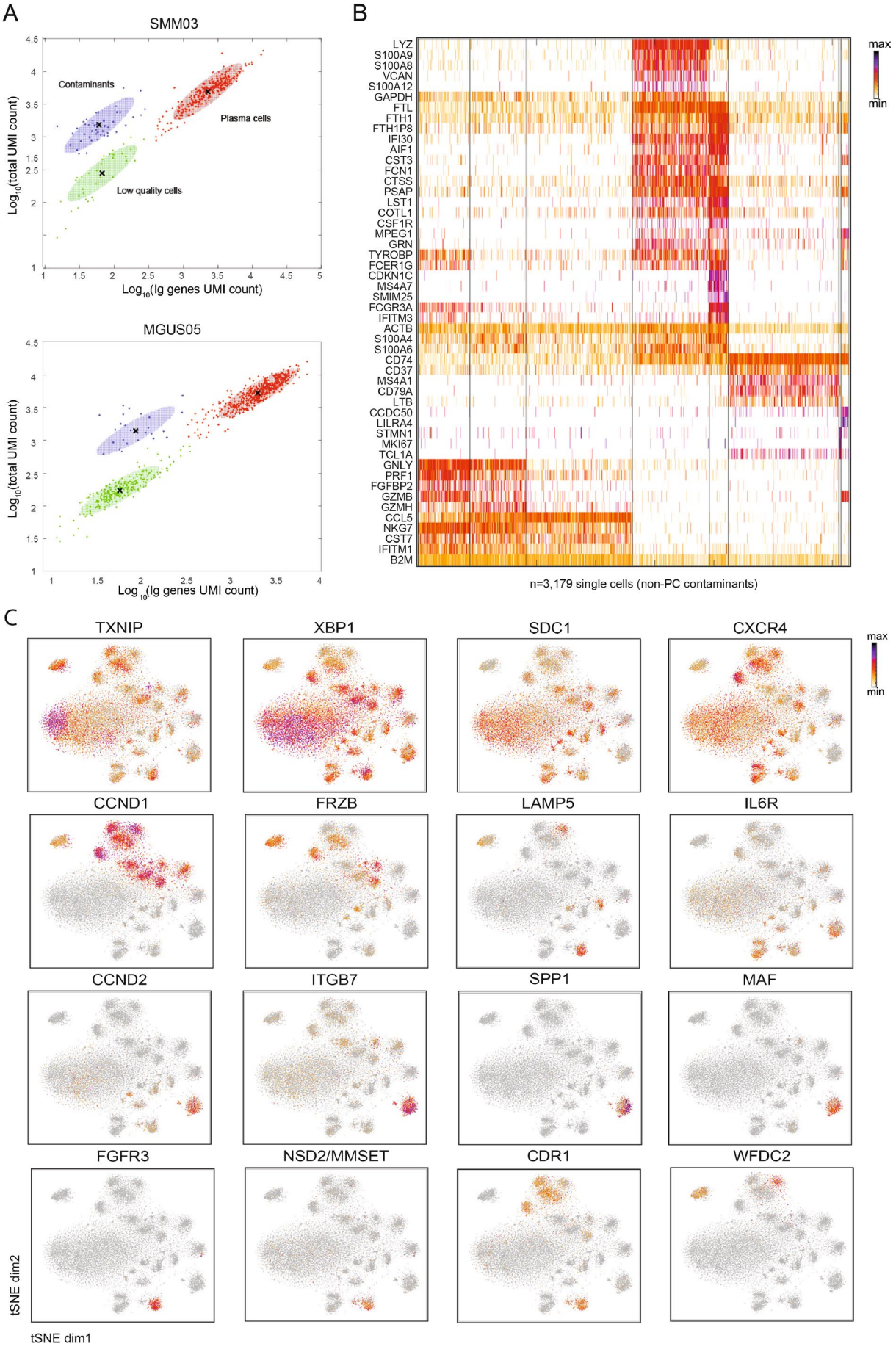
C

Patient code	Bone Marrow # of cells sequenced	# of cells passed QC	# plasma cells	Blood # of cells sequenced	# of cells passed QC	# plasma cells
hip01	1152	1102	1051			
hip02	768	649	482			
hip03	1152	1034	943			
hip04	768	625	437			
hip05	1152	807	705			
hip06	384	352	333			
hip07	768	580	373			
hip08	384	324	276			
hip09	768	718	678	384	148	21
hip10	384	202	128	384	95	44
hip11	768	388	263			
Total control	8448	6781	5669	768	243	65
MGUS01	768	679	589	384	100	32
MGUS02	768	341	273	384	369	363
MGUS03	384	338	301	384	139	66
MGUS04	768	641	466			
MGUS05	768	530	493	384	334	246
MGUS06	384	314	225			
MGUS07	1920	1818	700			
Total MGUS	5760	4661	3047	1536	942	707
SMM01	768	732	695	384	250	225
SMM02	1920	1884	1821	384	367	253
SMM03	384	355	279	384	279	244
SMM04	384	65	56	384	58	39
SMM05	768	740	703	384	228	90
SMM06	384	181	168	384	106	96
Total SMM	4608	3957	3722	2304	1288	947

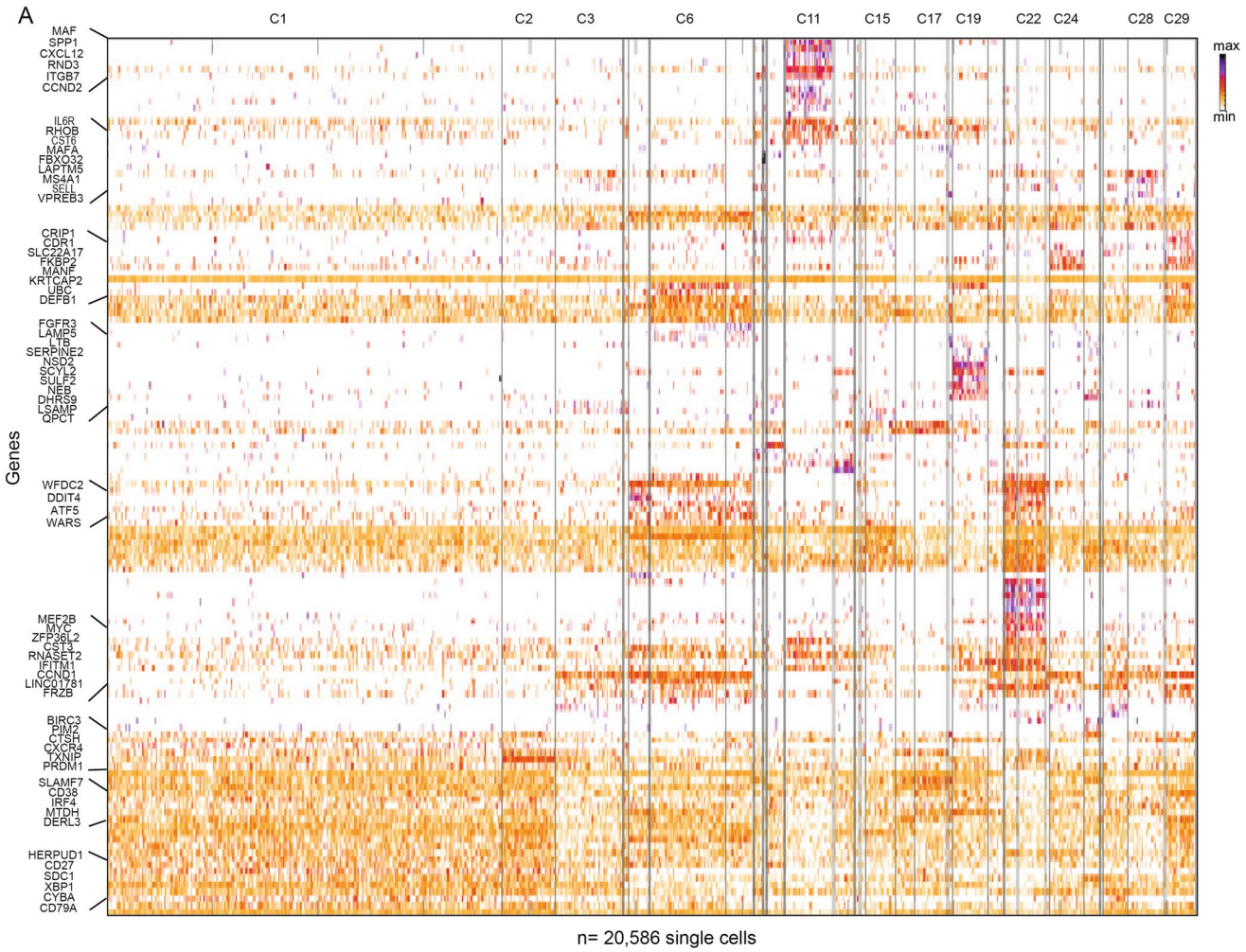
Patient code	Bone Marrow # of cells sequenced	# of cells passed QC	# plasma cells	Blood # of cells sequenced	# of cells passed QC	# plasma cells
MM01	1536	994	904	768	723	685
MM02	1920	768	362			
MM03	768	62	61	384	69	33
MM04	1536	1420	509			
MM05	384	287	135			
MM06	1536	1375	1084	384	301	167
MM07	384	279	179	384	232	135
MM08	768	605	437	384	254	192
MM09	1536	1318	896			
MM10	384	346	302			
MM11	768	617	492	384	170	71
MM12	1152	984	779			
AL01	1152	762	714			
AL02	384	203	145	384	325	29
AL03	768	709	593	384	361	336
AL04	1152	1027	538	384	281	173
Total MM+AL	16128	11756	8130	3840	2716	1821

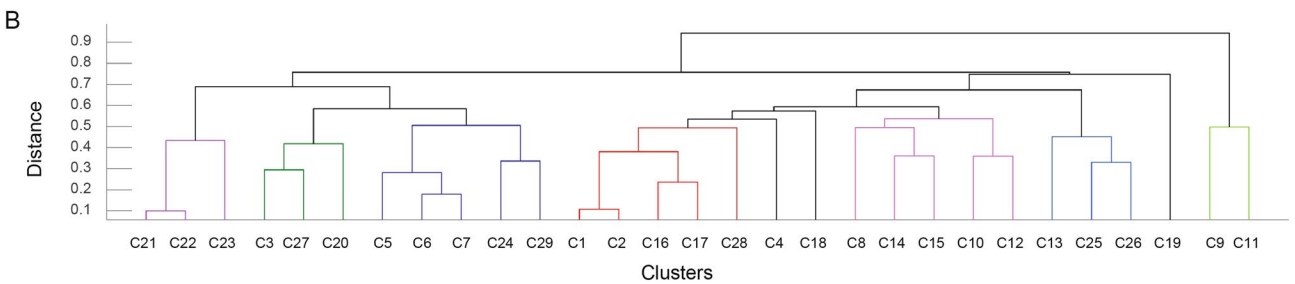
total BM cells sequenced	34944
total BM cells passed QC	27155
total BM PC analyzed	20586
total PB cells sequenced	8448
total PB cells passed QC	5189
total PB PC analyzed	3540

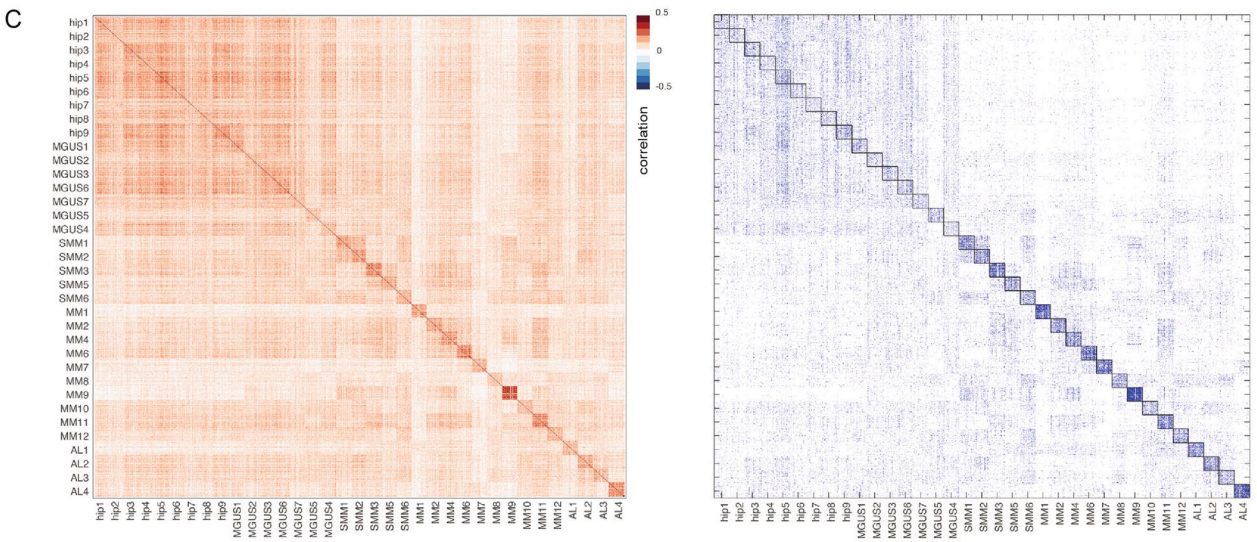
Extended Data Fig. 2 | Quality control metrics of single cell sequencing: bone marrow and peripheral blood single cell quality control metrics. a, Shown are number of reads, number of UMIs and percentage of cells analyzed per batch of 384 cells (that were pooled for library construction) for all $CD38^+CD138^+$ bone marrow single cells from 40 participants (29 newly diagnosed subjects and 11 control donors). Cells were sorted into plates, were sequenced and underwent quality control evaluation and filtering out of non-plasma cell contaminants (Online Methods). **b**, Shown are number of reads, number of UMIs and percentage of cells analyzed per batch of 384 cells (that were pooled for library construction) for all $CD38^+CD138^+$ peripheral blood single cells from 21 participants (19 newly diagnosed subjects and 2 control donors). Cells were sorted into plates, were sequenced and underwent quality control evaluation and filtering out of non-plasma cell contaminants (Online Methods). **c**, Table summarizing the total $CD38^+CD138^+$ cell numbers from bone marrow and blood of newly diagnosed individuals and control donors, that were sorted and sequenced, passed quality control and were analyzed (after filtering for non-plasma cell contaminants; see Online Methods).


A
SMM03
Contaminants
Plasma cells
Low quality cells
Log10(total UMI count)
Log10(Ig genes UMI count)
MGUS05
Log10(total UMI count)
Log10(Ig genes UMI count)
B
LYZ
S100A9
S100A8
VCAN
S100A12
GAPDH
FTL
FTH1
FTH1P8
IFI30
AIF1
CST3
FCN1
CTSS
PSAP
LST1
COTL1
CSF1R
MPEG1
GRN
TYROBP
FCER1G
CDKN1C
MS4A7
SMIM25
FCGR3A
IFITM3
ACTB
S100A4
S100A6
CD74
CD37
MS4A1
CD79A
LTB
CCDC50
LILRA4
STMN1
MKI67
TCL1A
GNLY
PRF1
FGFBP2
GZMB
GZMH
CCL5
NKG7
CST7
IFITM1
B2M
max
min
n=3,179 single cells (non-PC contaminants)
C
TXNIP
XBP1
SDC1
CXCR4
max
min
CCND1
FRZB
LAMP5
IL6R
CCND2
ITGB7
SPP1
MAF
FGFR3
NSD2/MMSET
CDR1
WFDC2
tSNE dim2
tSNE dim1

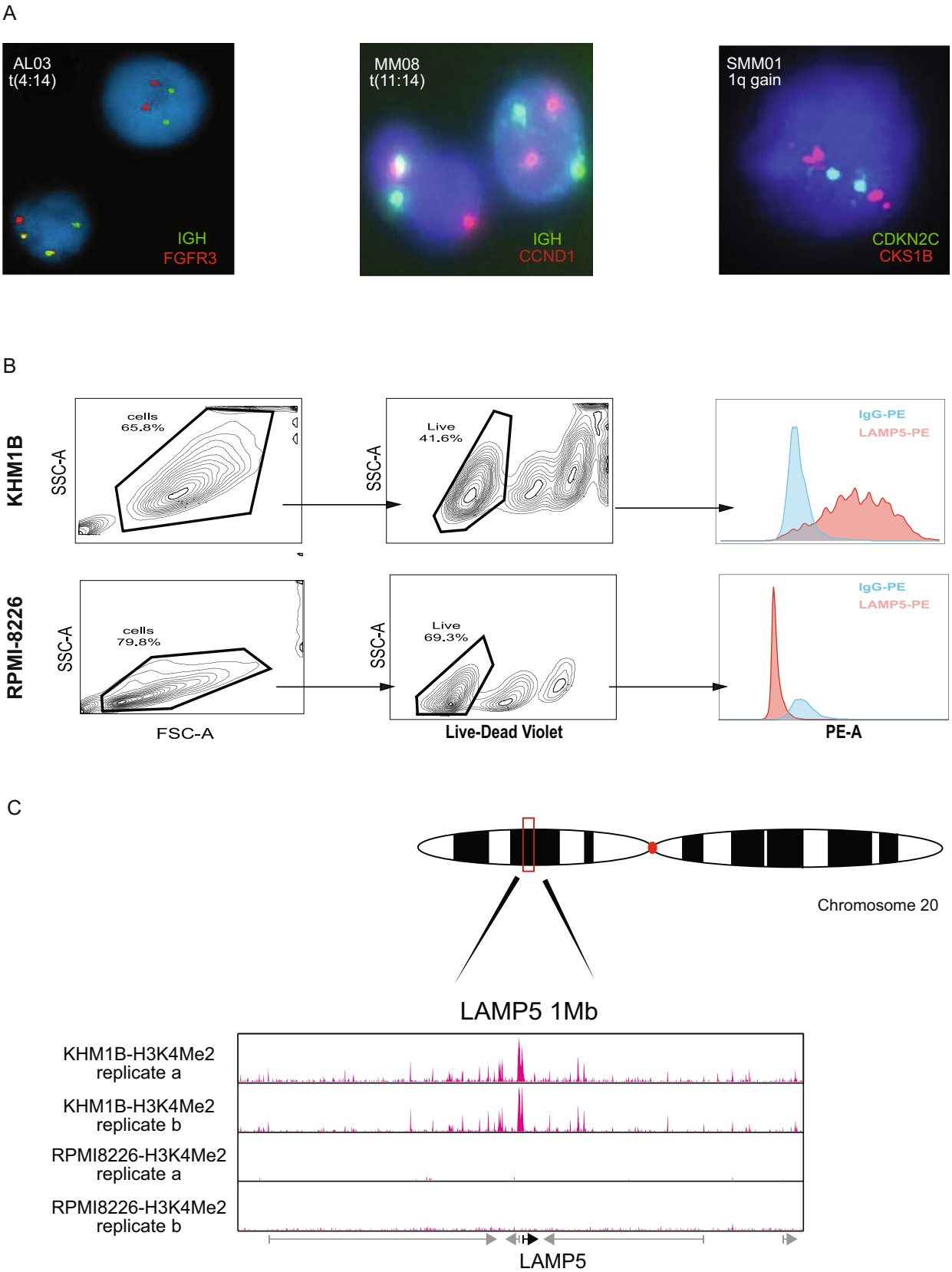
Extended Data Fig. 3 | Filtering strategy for plasma cells, and normalized single cell expression for selected genes. a, Plot depicting plasma cell in silico filtering according to immunoglobulin load and UMI count per cell (Online Methods) **b**, Heat map showing clustering analysis of 3,179 'contaminating' cells that pass quality control but do not express immunoglobulin genes above the cut-off (100 UMIs per cell). Representative genes (mostly unrelated to the plasma cell program) are shown. **c**, Normalized single cell expression (UMI count, log scale) of 16 representative genes, projected onto a *t*SNE map of all 20,586 bone marrow plasma cells from 40 participants (29 newly diagnosed individuals and 11 control donors).


A
C1
C2
C3
C6
C11
C15
C17
C19
C22
C24
C28
C29
max
min
Genes
MAF
SPP1
CXCL12
RND3
ITGB7
CCND2
IL6R
RHOB
CST6
MAFA
FBXO32
LAPTM5
MS4A1
SELL
VPREB3
CRIP1
CDR1
SLC22A17
FKBP2
MANF
KRTCAP2
UBC
DEFB1
FGFR3
LAMP5
LTB
SERPINE2
NSD2
SCYL3
SULF2
NEB
DHRS9
LSAMP
QPCT
WFDC2
DDIT4
ATF5
WARS
MEF2B
MYC
ZFP36L2
CST3
RNASET2
IFITM1
CCND1
LINC01781
FRZB
BIRC3
PIM2
CTSH
CXCR4
TXNIP
PRDM1
SLAMF7
CD38
IRF4
MTDH
DERL3
HERPUD1
CD27
SDC1
XBP1
CYBA
CD79A
n= 20,586 single cells

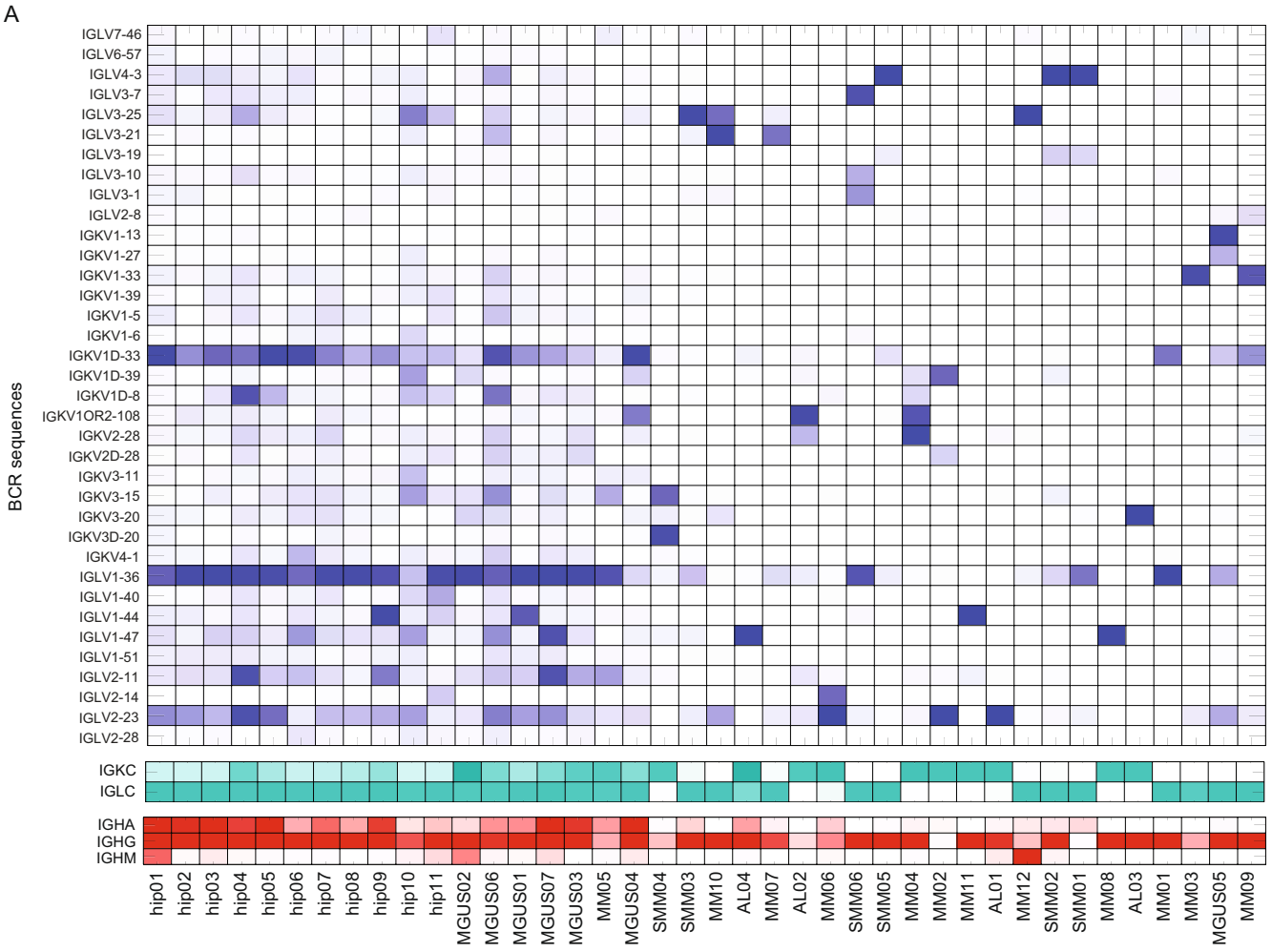

B
Distance
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
C21 C22 C23 C3 C27 C20 C5 C6 C7 C24 C29 C1 C2 C16 C17 C28 C4 C18 C8 C14 C15 C10 C12 C13 C25 C26 C19 C9 C11
Clusters

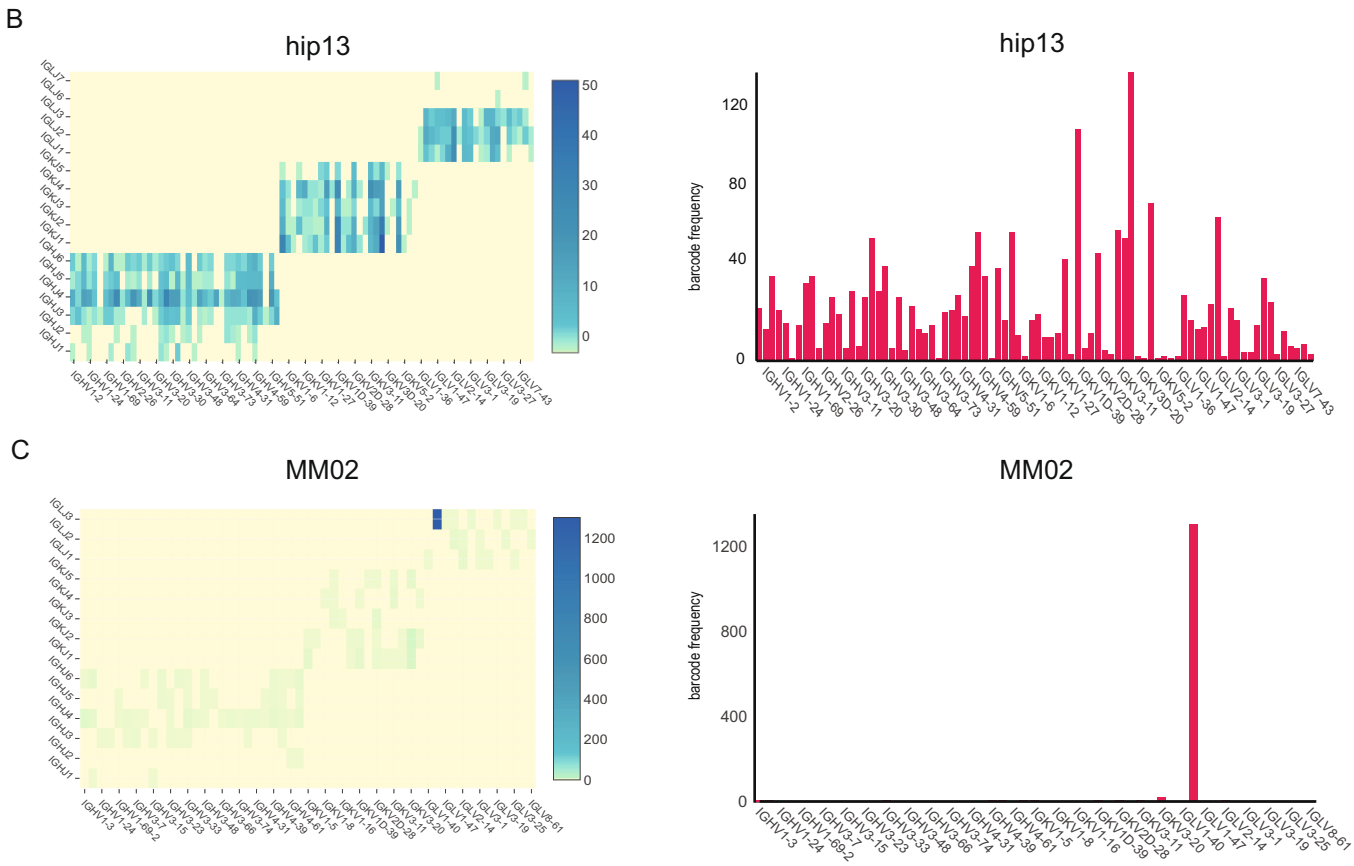

C
hip1
hip2
hip3
hip4
hip5
hip6
hip7
hip8
hip9
MGUS1
MGUS2
MGUS3
MGUS6
MGUS7
MGUS5
MGUS4
SMM1
SMM2
SMM3
SMM5
SMM6
MM1
MM2
MM4
MM6
MM7
MM8
MM9
MM10
MM11
MM12
AL1
AL2
AL3
AL4
0.5
0
-0.5
correlation

Extended Data Fig. 4 | Clustering analysis of single bone marrow plasma cells. a, Heat map showing clustering analysis of 20,586 bone marrow plasma cells sorted from 40 participants (29 newly diagnosed individuals and 11 control donors), featuring normalized single cell expression levels of the 100 most variable genes (Online Methods). **b**, Dendrogram showing the hierarchical clustering of the average transcription profile for all clusters, C1–29 (related to **a**). **c**, Cell-to-cell correlation matrix (Spearman) of 200 randomly selected cells from each subject (with $n > 200$ cells; left); *k*NN adjacency matrix ($k = 100$) showing, for each cell, its 100 nearest neighbors (blue; right).

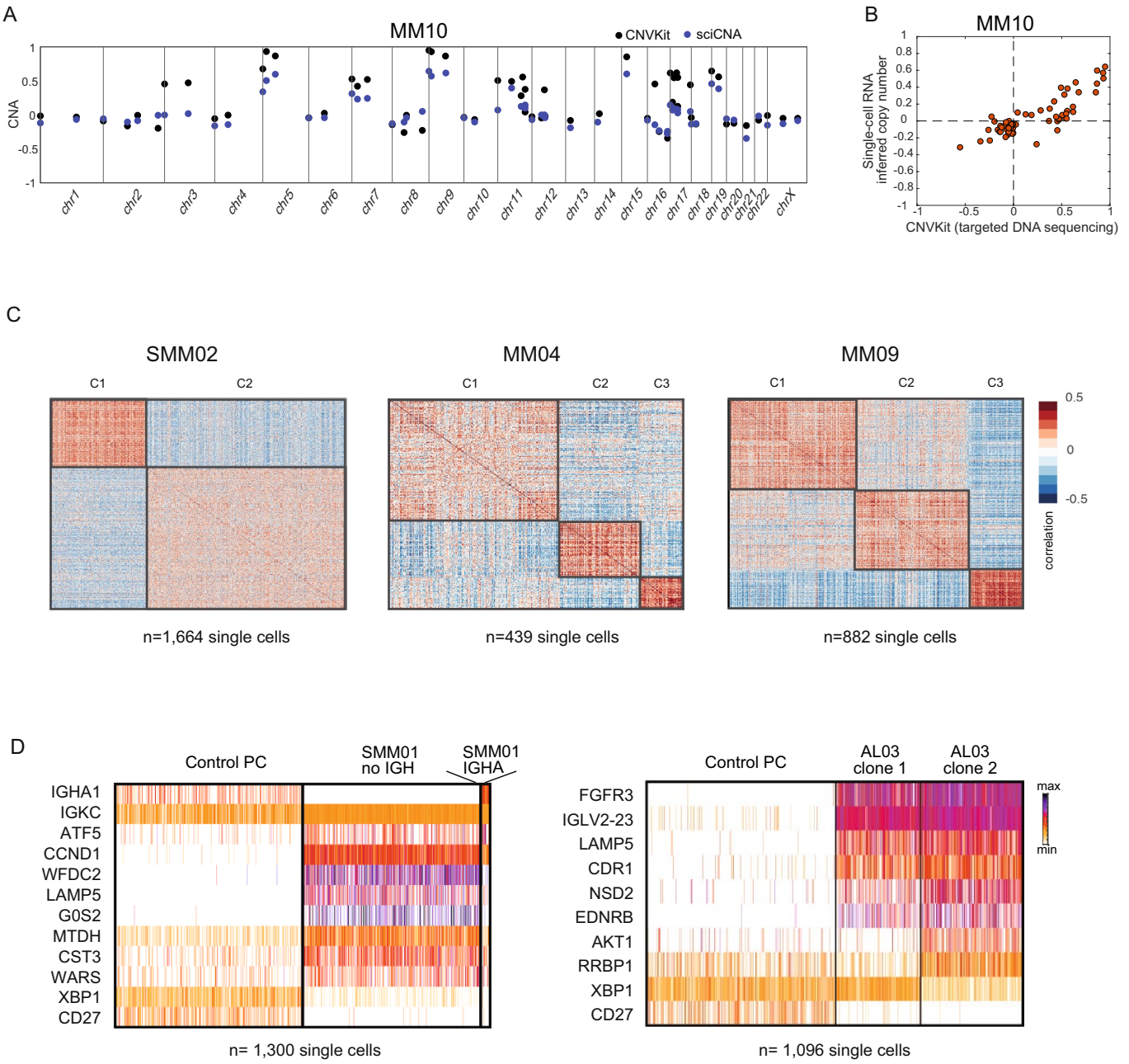


Extended Data Fig. 5 | Genomic FISH examples from selected subjects, and epigenetic analysis of *LAMP5*. a, Genomic iFISH images of bone marrow plasma cells magnetically enriched for CD138+ for subjects AL03 (left), MM08 (middle) and SMM01 (right). Shown are fluorescent probes for the immunoglobulin heavy chain (green: *IGH*) and translocation partners (red: *FGFR3*, *CCND1* and *CKS1B*, from left to right). Each hybridization was performed twice with similar results. **b**, FACS plots showing intracellular staining (gated on live cells) for LAMP5 protein (red) compared with isotype control (blue) for KHM1B and RPMI-8226 cell lines (top and bottom, respectively; see Online Methods). Each experiment was performed twice with similar results. **c**, Genome browser view of normalized H3K4me2 profiles of peaks found in a 1-megabase region in the *LAMP5* locus. Data are from two independent biological replicates (Online Methods).

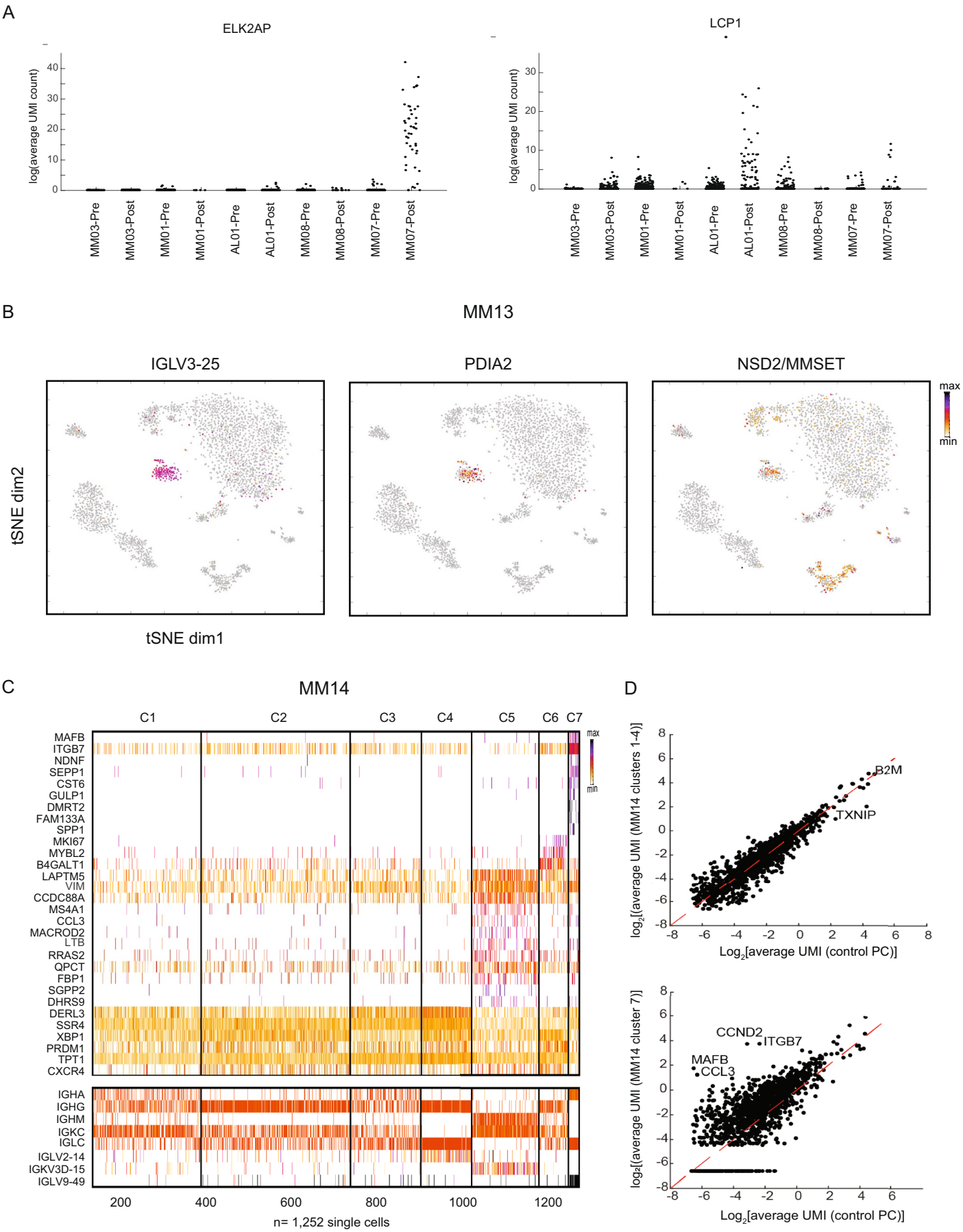

A
BCR sequences
IGLV7-46
IGLV6-57
IGLV4-3
IGLV3-7
IGLV3-25
IGLV3-21
IGLV3-19
IGLV3-10
IGLV3-1
IGLV2-8
IGKV1-13
IGKV1-27
IGKV1-33
IGKV1-39
IGKV1-5
IGKV1-6
IGKV1D-33
IGKV1D-39
IGKV1D-8
IGKV1OR2-108
IGKV2-28
IGKV2D-28
IGKV3-11
IGKV3-15
IGKV3-20
IGKV3D-20
IGKV4-1
IGLV1-36
IGLV1-40
IGLV1-44
IGLV1-47
IGLV1-51
IGLV2-11
IGLV2-14
IGLV2-23
IGLV2-28
IGKC
IGLC
IGHA
IGHG
IGHM
hip01 hip02 hip03 hip04 hip05 hip06 hip07 hip08 hip09 hip10 hip11 MGUS02 MGUS06 MGUS01 MGUS07 MGUS03 MM05 MGUS04 SMM04 SMM03 MM10 AL04 MM07 AL02 MM06 SMM06 SMM05 MM04 MM02 MM11 AL01 MM12 SMM02 SMM01 MM08 AL03 MM01 MM03 MGUS05 MM09


B
hip13
50
40
30
20
10
0
hip13
barcode frequency
120
80
40
0
C
MM02
1200
1000
800
600
400
200
0
MM02
barcode frequency
1200
800
400
0

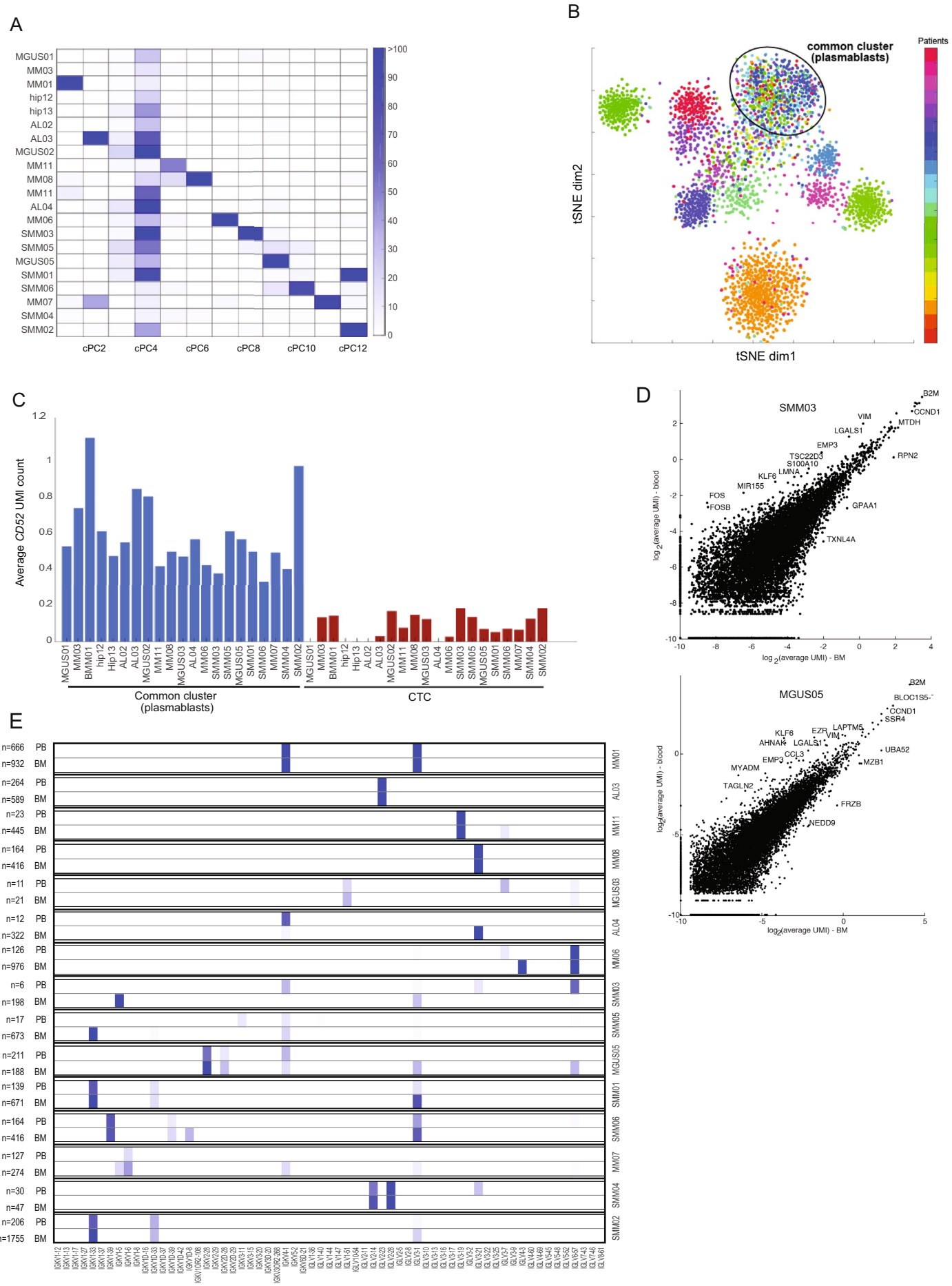
Extended Data Fig. 6 | Clonal assessment by inferring immunoglobulin sequence at single cell resolution. a, Heat map depicting the relative frequency of the immunoglobulin sequences from analyzing 20,586 bone marrow plasma cells of 40 participants (29 newly diagnosed individuals and 11 control donors): top, immunoglobulin light chain variable region (*IGKV* and *IGLV*); middle, immunoglobulin light chain constant region (*IGKC* and *IGLC*); bottom, immunoglobulin heavy chain constant region (*IGHC*). **b**, Chromium 10× single cell BCR clonotype distribution of control donor Hip13, created by Chromium's Loupe V(D)J browser (Online Methods). Left, heat map showing cell frequency for specific *IGHV/IGLV/IGKV* variable (V) region sequences (*x* axis) and *IGHJ/IGKJ/IGLJ* joining (J) region sequences (*y* axis). Right, frequency of different BCR clonotypes (inferred from the heat map). **c**, Plots of single cell BCR data for subject MM02 using the Chromium 10× single cell BCR platform (Online Methods). Left, single *IGLV* for subject MM02. Right, frequency of different BCR clonotypes for subject MM02, generated by Chromium's Loupe V(D)J browser.



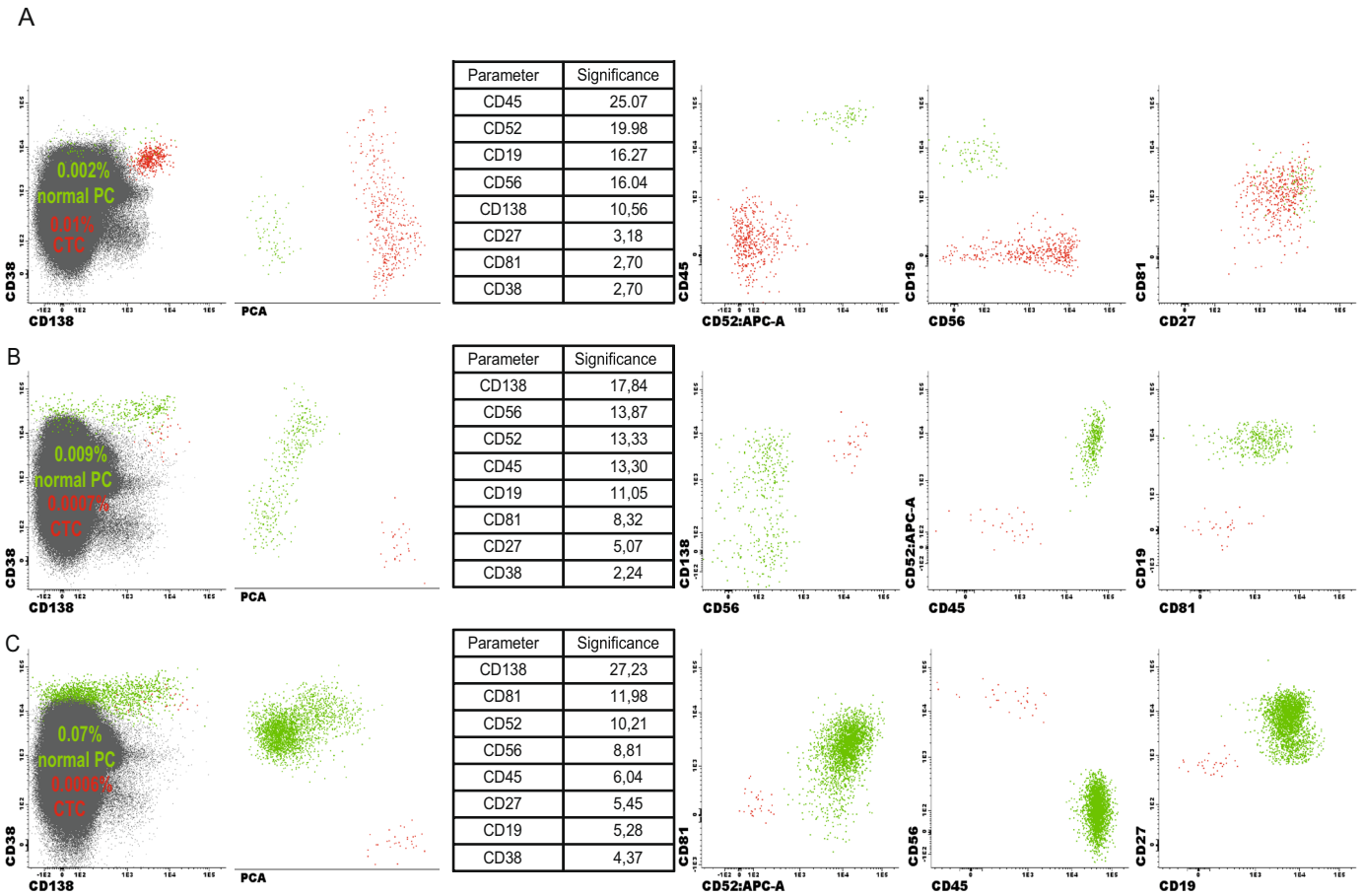
Extended Data Fig. 7 | CNAs deduced from scRNA-seq data. a, Comparison of CNVKit output from genomic DNA-targeted sequencing panel at 500× coverage (black dots) and sciCNA (blue) for subject MM10. **b**, Scatter plot of CNVKit versus sciCNA estimation for each DNA segment for subject MM10. **c**, Cell-to-cell correlation matrix (Pearson correlation on row-normalized log transformed data) for subjects SMM02 (1,664 cells), MM04 (439 cells) and MM09 (882 cells). **d**, Heat maps of normalized single cell gene expression (UMI count, log scale) for bone marrow plasma cells from subjects SMM01 (left, 650 cells) and AL03 (right, 548 cells), clustered with the same number of normal bone marrow plasma cells from control donors. Representative variable genes are shown.


A
ELK2AP
LCP1
log(average UMI count)
MM03-Pre
MM03-Post
MM01-Pre
MM01-Post
AL01-Pre
AL01-Post
MM08-Pre
MM08-Post
MM07-Pre
MM07-Post
B
MM13
IGLV3-25
PDIA2
NSD2/MMSET
max
min
tSNE dim2
tSNE dim1
C
MM14
C1
C2
C3
C4
C5
C6
C7
MAFB
ITGB7
NDNF
SEPP1
CST6
GULP1
DMRT2
FAM133A
SPP1
MKI67
MYBL2
B4GALT1
LAPTM5
VIM
CCDC88A
MS4A1
CCL3
MACROD2
LTB
RRAS2
QPCT
FBP1
SGPP2
DHRS9
DERL3
SSR4
XBP1
PRDM1
TPT1
CXCR4
IGHA
IGHG
IGHM
IGKC
IGLC
IGLV2-14
IGKV3D-15
IGLV9-49
200
400
600
800
1000
1200
n= 1,252 single cells
D
log2[(average UMI (MM14 clusters 1-4)]
Log2[average UMI (control PC)]
B2M
TXNIP
log2[(average UMI (MM14 cluster 7)]
CCND2
ITGB7
MAFB
CCL3

Extended Data Fig. 8 | scRNA-seq of rare malignant plasma cells in individuals with minimal residual disease. a, Scatter plots of single cell expression (average UMI count, log scale) for of *ELKAP2* (left) and *LCP1* (right) genes in subjects pre- and post-treatment. **b**, Chromium 10× 5′ single cell expression estimates (Online Methods) visualized as *t*SNE plots for subject MM13 ($n=1$) bone marrow cells enriched with CD138 magnetic beads (Online Methods). Shown are expression estimates for *IGLV3–IGLV25* (left), *PDIA2* (middle) and *NSD2* (right). **c**, Heat map of normalized single cell gene expression (UMI, log scale) for subject MM14 ($n=1$) bone marrow plasma cells (1,252 cells). Representative variable genes are shown. **d**, Scatter plots showing average single cell gene expression ($\log_2$ scale) of control donors' bone marrow plasma cells (*x* axis) compared with either MM14 cells from C1–4 (top scatter plot; related to heat map in **b**) or MM14 cells from C7 (bottom scatter plot). Selected gene names are shown.


A
MGUS01
MM03
MM01
hip12
hip13
AL02
AL03
MGUS02
MM11
MM08
MM11
AL04
MM06
SMM03
SMM05
MGUS05
SMM01
SMM06
MM07
SMM04
SMM02
cPC2
cPC4
cPC6
cPC8
cPC10
cPC12
>100
90
80
70
60
50
40
30
20
10
0
B
common cluster
(plasmablasts)
Patients
tSNE dim2
tSNE dim1
C
Average CD52 UMI count
1.2
1
0.8
0.6
0.4
0.2
0
MGUS01
MM03
BMM01
hip12
Hip13
AL02
AL03
MGUS02
MM11
MM08
MGUS03
AL04
MM06
SMM03
SMM05
MGUS05
SMM01
SMM06
MM07
SMM04
SMM02
Common cluster
(plasmablasts)
CTC
D
SMM03
B2M
CCND1
VIM
MTDH
LGALS1
EMP3
TSC22D3
S100A10
RPN2
LMNA
KLF6
MIR155
FOS
FOSB
GPAA1
TXNL4A
log2(average UMI) - blood
log2(average UMI) - BM
MGUS05
B2M
BLOC1S5-
CCND1
SSR4
KLF6
AHNAK
EZR
LAPTM5
VIM
LGALS1
UBA52
CCL3
EMP3
MZB1
MYADM
TAGLN2
FRZB
NEDD9
log2(average UMI) - blood
log2(average UMI) - BM
E
n=666 PB
n=932 BM
n=264 PB
n=589 BM
n=23 PB
n=445 BM
n=164 PB
n=416 BM
n=11 PB
n=21 BM
n=12 PB
n=322 BM
n=126 PB
n=976 BM
n=6 PB
n=198 BM
n=17 PB
n=673 BM
n=211 PB
n=188 BM
n=139 PB
n=671 BM
n=164 PB
n=416 BM
n=127 PB
n=274 BM
n=30 PB
n=47 BM
n=206 PB
n=1755 BM
MM01
AL03
MM11
MM08
MGUS03
AL04
MM06
SMM03
SMM05
MGUS05
SMM01
SMM06
MM07
SMM04
SMM02

Extended Data Fig. 9 | Circulating plasma cells reflect the bone marrow disease. a, Heat map showing the distribution of circulating plasma cells from 21 individuals with multiple myeloma as well as controls in 12 clusters (cPC1–cPC12). **b**, Two-dimensional *t*SNE view of circulating plasma cells from 21 subjects and controls. Each dot represents a single cell. Patients are color coded. **c**, Bar plots of average UMI count for the *CD52* gene (Online Methods) in common cluster cPC4 (left; blue bars) compared with circulating tumor cells (right; red bars). **d**, Scatter plots of single cell expression estimates (average UMI count, log scale) for subects' abnormal bone marrow plasma cells (*x* axis) to circulating tumor cells in the blood (*y* axis). Top, subject SMM03, (198 bone marrow plasma cells and 16 circulating tumor cells); bottom, subject MGUS05 (188 bone marrow plasma cells and 211 circulating tumor cells). **e**, Immunoglobulin light chain variable region distribution within each subject's bone marrow and blood tumor cells. Shown are 15 individuals for whom we were able to reconstruct BCR data from rare circulating tumor cells. Color represents percentage of cells, ranging from white (0) to blue (0.5). In each box, the top panel represents PB and the bottom panel represents bone marrow. Cell numbers are shown on the left.



Parameter	Significance
CD45	25.07
CD52	19.98
CD19	16.27
CD56	16.04
CD138	10,56
CD27	3,18
CD81	2,70
CD38	2,70

Parameter	Significance
CD138	17,84
CD56	13,87
CD52	13,33
CD45	13,30
CD19	11,05
CD81	8,32
CD27	5,07
CD38	2,24

Parameter	Significance
CD138	27,23
CD81	11,98
CD52	10,21
CD56	8,81
CD45	6,04
CD27	5,45
CD19	5,28
CD38	4,37

Extended Data Fig. 10 | CD52 segregates polyclonal circulating plasma cells from malignant ones. a, Flow cytometry plots of circulating plasma cells from individuals with relapsed myeloma ($n=3$) post-treatment. Plots were generated using Infinicyt software (Online Methods). Circulating tumor cells (red) are marked by an aberrant surface phenotype compared with normal plasma cells (green). Principal component analysis quantifies the significance (contribution to principal component 1) of each surface marker to separate between 'circulating tumor cells' and 'normal plasma cells'. Each row represents a different subject.

nature research

Corresponding author(s): Ido Amit

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	CoMMpass IA9 database (mmrf.org), UCSC, IMGT, BLAST, 1000Genomes, COSMIC
Data analysis	PehnoGraph, matlab custom code (deposited in bitbucket.org), MetaCell, Suerat, FlowJo, Diva6.1, Infinicyt, CaveMan, Pindel, Strelka, BRASS, Delly, CNVkit, Cell Ranger, V(D)J-Loupe, GO (metascape), Picard, MetaMorph

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Raw and processed single cell RNA-sequencing data and ChIP-seq data were deposited to NCBI's GEO with accession number GSE117156.
Raw and processed genomic DNA targeted sequencing data were deposited to NCBI's SRA with accession number SRP165705.

Source code used for scRNAseq analysis can be found at https://bitbucket.org/amitlab/multiple-myeloma-2018/.
The patients' clinical information is available in Supplementary Table 1.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	This was an exploratory study of high throughput scRNAseq in myeloma. For intra-patient heterogeneity at the RNA level in MM, we based our sample size after sampling a pilot cohort of n=4, and found heterogeneity in 2/4 of our cohort.
Data exclusions	As some patients with lymphoma or an autoimmune disease have gammopathy, they were excluded. These criteria were pre-established.
Replication	Reproducibility of scRNAseq data per batch for each patient was performed, and confirmed.
Randomization	Allocation was not random. Covariates were controlled by processing the samples at the same place by the same people
Blinding	Blinding was not possible, as the control donors and the patients samples were samples from the operating room and a bed-side aspirate, respectively.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

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

Antibodies

Antibodies used	anti-human CD19-PE-Cy7, Beckman Coulter, clone J3.119, Cat#IM3628U, dilution 1:100 anti human CD27-BV510, Biolegend, clone O323, Cat#302836, dilution 1:100 anti-human CD38-FITC, Cytognos, clone ME, Cat# CYT38-F2, dilution 1:100 anti human CD56-PE, Cytognos, clone: C5.9, Cat#CYT-56PE, dilution 1:100 anti-human CD45-PerCP-Cy5.5, Biolegend, clone EO1, Cat#368504, dilution 1:100 anti-human CD81-APC-H7, Cytognos, clone M38, Cat#CYT-81AC750, dilution 1:100 anti-human CD117-APC, BD, clone 104D2, Cat#341096, dilution 1:100 anti-human CD138-BV421, BD, clone MI15, Cat#562935, dilution 1:100 anti-human CD52-APC-H7, BD, clone 4C8, Cat#563611, dilution 1:100 anti-human LAMP5-PE, Miltenyi Biotech, clone REA590, Cat#130-109-203, dilution 1:50 Anti-Histone H3 (di methyl K4) antibody [EPR17707] - ChIP Grade (ab176878)
Validation	This panel of antibodies was extensively validated by the Euroflow group of investigators, please see reference 51 (Flores-Montero et al)

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	ATCC (RPMI-8226), JCRB (KHM1B)

Authentication	Bulk RNA-seq
Mycoplasma contamination	These cell lines were tested negative for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	None of these cell line appear in the misidentified cell line list

Human research participants

Policy information about studies involving human research participants

Population characteristics	The age, gender, diagnosis and other clinical data (disease related BM and serum studies) are detailed in Supplementary Table 1
Recruitment	Subjects were recruited by the orthopedics hip disease clinic (control donors with isolated osteoarthritis) and 5 hematology clinics in Israel (patients). All subjects were screened according to inclusion and exclusion criteria by the physicians, and were asked to participate and sign consent

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	ChIP-seq data were deposited to NCBI's GEO with accession number GSE117156
Files in database submission	N/A
Genome browser session (e.g. UCSC)	For IGV snapshot we used 1 M bp window around LAMP5 gene

Methodology

Replicates	2 biological replicates for each cell line
Sequencing depth	50M reads per sample
Antibodies	H3K4me2 antibody (Abcam), cat#ab176878
Peak calling parameters	Reads were aligned to the human reference genome (hg38) using Bowtie2 aligner version 2.3.4.1 with default parameters. The Picard tool 'MarkDuplicates' from the Broad Institute (http://broadinstitute.github.io/picard/) was used to remove PCR duplicates. To identify regions of enrichment (peaks) from H3K4me2 reads, we used the Homer package (http://homer.ucsd.edu/homer/) 'makeTagDirectory' followed by the 'findPeaks' command with the histone parameter using appropriate whole cell extract control. Peaks from all samples were merged using 'mergePeaks' from Homer package. Reads from all samples counted using 'annotatePeaks' from Homer with the default homer genome data hg38, merged peaks area file and the parameter -raw in order not to normalize by the default read count.
Data quality	N/A
Software	Picard, Homer

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	BM cells were processed using ice-cold FACS buffer and density gradient centrifugation, followed by red blood cell lysis, and antibody staining.
Instrument	FACS AriaII-SORP and FACS AriaFusion (BD)

Software	Data was collected using FACSDiva 8.0, and analyzed using FlowJo
Cell population abundance	Cells were single cell sorted
Gating strategy	Bone marrow mononuclear cells were gated using FSC/SSC plot, doublets exclusion was performed using FSC-A/FSC-W and SSC-A/SSC-W gating. Plasma cells were gated as CD38+CD138+.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.