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

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

1. Department of Immunology, Weizmann Institute of Science, Rehovot, Israel 2. Department of Internal Medicine "T", 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. 9. Hematology Division, Chaim Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel. 10. Department of Hematology, Assaf Harofeh Medical Center, Zerifin, Israel. 11. Life Sciences Core Facility, Weizmann Institute of Science, Rehovot, Israel. 12. Hematology Institute, Kaplan Medical Center, Rehovot, Israel. 13. Center for Hematological Malignancies, Memorial Sloan Kettering Cancer Center, New York, USA. 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. Clinica Universidad de Navarra, Centro de Investigacion Medica Aplicada; Instituto de Investigación Sanitaria de Navarra, CIBERONC, Pamplona, Spain 18. Integrated Cancer Genomics, Translational Genomics Research Institute, Phoenix, AZ, USA 20. Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA 21. Bench to Bedside Program, Weizmann Institute of Science, Rehovot, Israel 22. Department of Computer Science and Applied Mathematics and Department of Biological Regulation, Weizmann Institute of Science, Rehovot, Israel.

* These authors contributed equally to this work

† To whom correspondence should be addressed: ido.amit@weizmann.ac.il (IA).

Abstract

Multiple myeloma (MM), a plasma cell (PC) 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 scRNA-seq to study the heterogeneity of 40 individuals along MM progression spectrum including 11 healthy controls, demonstrating high inter-patient variability that can be explained by expression of known MM drivers and additional putative factors. We identify extensive sub-clonal structures for 10/29 patients. In asymptomatic patients with early disease and in minimal residual disease post-treatment, we detect rare tumor-PC with similar molecular characteristics of active myeloma, with possible implications for personalized therapies. Single cell analysis of rare circulating-tumor-cells (CTC) allows for accurate liquid biopsy and detection of malignant PC, which reflect the patient BM disease. Our work establishes scRNA-seq for dissecting blood malignancies and devising detailed molecular characterization of tumor cells in symptomatic and asymptomatic patients.

Introduction

Multiple myeloma (MM) is a neoplastic plasma cell (PC) disorder that is characterized by clonal proliferation of malignant PC in the bone marrow (BM). Despite improved survival rates in the last decade, therapy is not curative and almost all patients relapse[2]. The clinical spectrum of the disease includes asymptomatic conditions: Monoclonal gammopathy of undetermined significance (MGUS), a condition with limited suspected malignant PC in the BM, and monoclonal antibodies in the blood (M-protein); or smoldering myeloma (SMM), a more advanced stage with higher proportion of malignant PC and/or M-protein in the blood[3,4]. The rate of progression from MGUS and SMM into active myeloma is approximately 1% and 10% per year, respectively[5]. The genetic landscape underlying myeloma was mapped in several foundational genomic studies[6–10]. Furthermore, the gene expression profiling cohorts of MM patients (e.g. the Multiple Myeloma Research Foundation's CoMMpass Study) have shown to be effective in predicting the risk of disease progression and survival[11–13]. They also highlighted over-expression of several oncogenic drivers and pathways including *CCND1*, *FGFR3*, *NSD2/MMSET*, *MAFB* and others.

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 BM aspiration procedures and accurate 'liquid biopsies'. The large PC 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"[5]. Yet, current strategies for genomic sequencing and transcriptional analysis in cancer were developed for mapping bulk samples from primary tumors and metastases and are therefore lacking the resolution and accuracy for characterizing small tumorigenic sub-populations that are likely driving MGUS, SMM and MM residual disease progression. Single cell genomic technologies are opening the way for the development of such assays[14–19].

Here, we report the first comprehensive single cell RNA profiling of newly diagnosed asymptomatic (7 MGUS and 6 SMM) and symptomatic (12 MM and 4 AL amyloidosis) patients encompassing the different clinical spectra of plasma cell pre-malignant and neoplastic states. We characterized 20,586 single PC from the BM and 3,540 single PC from the blood of 11 control individuals and 29 newly diagnosed patients. We found minimal inter-donor heterogeneity for controls, showing that normal PC from the BM are reproducibly organized into transcriptional states that represent variable activity of genes associated with ER stress and PC physiology[20]. In contrast, MM patients show highly diverse cell states, with every patient defining a unique and individual PC transcriptional program. Importantly, the different patients overexpress common known MM oncogenic drivers: *CCND1* (11/29), *FRZB* (11/29) *NSD2/MMSET* (2/29), and several uncharacterized overexpressed MM markers that can be confirmed in the larger CoMMpass database, including the lysosomal associated membrane protein *LAMP5* (5/29), the endopeptidase inhibitor *WFDC2* (2/29), and a small intronless gene located in the X chromosome, *CDR1 (5/29)*. Intra-patient transcriptional heterogeneity was also observed for 10 patients, in most of the cases with the same immunoglobulin (Ig) clonotype. Using the scRNA-seq transcriptional data we infer copy number alterations (sciCNA) by averaging the relative expression of a large number of genomically adjacent genes and show that in MM transcriptional changes are many times regulated in trans, and may be associated indirectly to genomic aberrations. By profiling single PC from the blood, we further identify efficient molecular markers (e.g. *CD52*) to enrich for circulating tumor cells (CTC), and show that in all cases with paired CTC and BM (15/15), the CTC in the blood reflect the molecular disease observed in the BM. Furthermore, in follow up analysis of 5 of our patients post-therapy, we detect rare malignant cells, and show that the residual malignant PC share most of the 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 MM patients and identifies new pathways and potential targets for diagnosis and therapy in symptomatic and asymptomatic myeloma. Further sampling of a larger cohort of patients pre- and post- therapy can potentially help in prioritizing efficient and personalized treatment for MM patients.

Results

MM patients display unique signatures that converge into defined malignant pathways

In order to better understand the heterogeneity within and across MM patients, we designed a protocol for single cell transcriptomic characterization of the BM PC, as well as circulating PC from MGUS, SMM, MM and primary light chain AL amyloidosis (AL) patients (**Fig. S1**). Our design was focused on maintaining the *in situ* RNA composition of the patients' samples by instant cooling in the operating room of the BM and blood, and immediate sorting of fresh cells for MARS-seq analysis[21]. We calibrated a validated flow cytometry - based method for isolation of PC (CD138+, CD38+) linking the intensity of the markers in each cell with its expression profile using an index-sorting strategy (**Fig. S2**). This allows for retrospective analysis of surface marker combinations for each individual cell[22]. We obtained fresh BM samples from 29 newly diagnosed patients with plasma cell neoplasms (**Table 1** and **Table S1)**. Profiling the normal diversity of plasma cells in an age-matched control cohort is essential to understand the heterogeneity of the normal and malignant PC disease states. To obtain normal BM with similar age to our patient cohort, we selected 11 older-adults and elderly subjects (median age 64 years, range 45-83, 5 males and 6 females) with isolated hip osteoarthritis, and without other medical co-morbidities, or active inflammatory processes, and extracted BM from the proximal femur bony canal during hip replacement surgery[23].

Following removal of low-quality cells, unsupervised clustering of 20,568 bone marrow plasma cells sorted from the 29 newly diagnosed patients: 7 MGUS (MGUS01-07); 6 SMM (SMM01-06); 12 MM (MM01-12) and 4 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, Fig. S3, Fig. S4** and **Fig. S5**). Despite a stringent sorting scheme for BM PC, 3,179 contaminating (non-plasma) cells were *in silico* removed, prior to clustering, based on their transcriptional signatures (**Fig. S4**; Methods). Plasma cell sub-populations were based on cluster-specific expression patterns of the 1,500 most variable genes discarding immunoglobulin (Ig) genes (**Fig. S5** and **Table S2**). Clusters C1-2 are associated to the control group of hip replacement individuals, representing normal PC with minor donor specific enrichments in these clusters. Cluster C2 represents long-lived plasma cells, evident by high expression of *CXCR4* and *TXNIP* (**Fig. S5** and **Fig. S6)**[20,24]**.** Cluster C1 shows a similar transcriptional profile, with lower levels of *CXCR4* and *TXNIP* and a gradient of CD81 and CD19 protein surface marker expression (**Fig. S5** and **Fig. S6**). Notably, we detect variable frequency of cells with normal PC phenotype in most patients, especially in the asymptomatic patients (**Fig. 1A-C**). Interestingly, in these patients the normal PC are more similar to short-lived PC expressing low levels of *CXCR4* and *TXNIP* (**Fig. S6**). In contrast, clusters C3-29 represent patient specific transcriptional state(s), with each patient characterized by an almost unique PC transcriptional program (**Fig. 1 A-D**, **Fig. S5** and **Table S3**). Although each patient is unique, we detected common overexpressed pathways shared across sub-groups of patients, such as *CCND1, CCND2* and *NSD2-FGFR3* groups (**Fig. 1E** and **Fig. S6**). *CCND1*-driven malignant PC are observed across 11 patients and can be found in 58.7% (476/811) of the CoMMpass database (**Fig. 1E)**. To validate if these patients harbor the *IGH* translocation, or represent other overexpression mechanisms, we compared the genomic DNA interphase fluorescent *in situ* hybridization (iFISH) data of these

patients (Methods). We find that 10/11 harbor an *IGH* translocation event (**Table S1**). Clusters C20-22, C26, C29 are represented by deregulation of the canonical wingless (Wnt) pathway, including overexpression of *SMAD* genes and the soluble frizzled related protein 3 (*FRZB*, $p<1x10^{-50}$), also found in 68% (553/811) of the CoMMpass database (**Fig. 1E**, **Fig. S6** and **Table S5)**. The tyrosine kinase fibroblast growth factor receptor 3 (*FGFR3*, found in 75/811 patient of the CoMMpass database), a known high-risk oncogene in myeloma, is featured in cluster C19 (patient AL03), and confirmed by t(4:14) iFISH testing (**Fig. 1E** and **Fig. S7**)[25,26]. We also identified putative MM overexpressed genes (all with $p<1x10^{-50}$) not found in the control cohort, including: Lysosome-associated membrane protein like molecule 5 (*LAMP5)*, a protein localized in the ER-Golgi compartment regulated by toll-like receptor signaling[27] in 5/29 patients (over expressed in 52%, 425/811, in the CoMMpass database); Cerebellar degeneration gene 1 (*CDR1*, a protein with a yet unknown function) in 5/29 patients (over expressed in 3.5%, 29/811, in the CoMMpass database); and WAP four-disulfide core domain protein 2 gene (*WFDC2*), a secreted proteinase in 2/29 patients (over expressed in 6.7%, 55/811 in the CoMMpass database (**Fig. 1E** and **Fig. S6**). *LAMP5*, *CDR1* and *WFDC2* were previously implicated in plasmacytoid dendritic cells, paraneoplastic syndromes, and ovarian carcinoma, respectively, but not in MM[28–30]. Since no mutations and/or aberrations are found near the *LAMP5* locus in MM patients, we first checked for overexpression of *LAMP5* in a database of myeloma cell lines (Multiple Myeloma Research Foundation's characterization of myeloma cell lines), and identified a large number of MM cell lines overexpressing *LAMP5* (28/75). We then performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) of histone 3 lysine 4 di methyl regions (H3K4me2), which marks promoter and enhancer regions, in a *LAMP5* MM over-expressing 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 (**Fig. S7**). These may represent regulatory regions specific to *LAMP5* that are activated in MM by trans. Together, we comprehensively profile malignant and normal PC using scRNA-seq and show that even the most careful bulk PC sampling contains significant contaminants from unrelated immune cells and normal PC. Using the single cell resolution data of neoplastic PC, we identify that each patient has its own unique transcriptional signature, and verify previously implicated drivers along with putative overexpressed candidate genes.

The transcriptional state of PC is regulated by the interplay of the genome, epigenome and environmental contexts. To test if DNA mutations and/or copy number alterations (CNA) can contribute to the transcriptional heterogeneity we identified in the different patients, we used a sensitive targeted approach to sequence the DNA regions involved in MM[31]. Profiling 11 patients from our cohort, we find similar DNA aberrations reported in previous MM studies[32,33]. However, most of the transcriptional divergence in our dataset cannot be explained by the DNA mutational status alone, which may suggest that some of the transcriptional changes observed in MM are regulated *in trans*. For example, patient MM08 harboring both t(11:14) translocation with *NRAS* mutated sub-clones, as well as chromosome 13 deletion, is transcriptionally homogeneous (**Fig. 1F, Fig. S7** and **Table S1)**, while patient MM10, with substantial intra-tumor heterogeneity, as we show below, has no sub-clonal mutations (**Fig. 1F, Fig. 2D-E** and **Table S1)**. Together, our data shows that tumor heterogeneity observed in MM cannot be intuitively explained by the DNA mutational status alone. This suggests a possible role for rare intergenic non-coding mutations (that were not captured by the targeted DNA panel), and/or *trans* regulated epigenetic and environmental inputs governing the full extent of transcriptional heterogeneity observed in MM.

Characterizing the intra-tumor heterogeneity of MM patients

Since MM patients display large patient-to-patient heterogeneity, we evaluated if intra-tumor heterogeneity can also be observed in our cohort[9,14]. PC originate from post-germinal center and memory B-cells and thus have a limited capacity to undergo further mutations in the Ig locus[31]. In agreement with this, myeloma cells typically conserve the Ig sequence through the course of tumor evolution[34]. To gain insight into the clonotypic identity of PC within each patient, we used our scRNA-seq data to annotate the variable region of the Ig light chain λ or κ (IGVL/K) and coupled it with the constant region of the Ig heavy chain (IGHC) for every single cell (**Fig. S8**; Methods). PC from control donors were composed of diverse Ig sequences (**Fig. S8**). Across the different controls, the most frequently represented IGVL/K sequences correlated to their prevalence in the general population[35]. Conversely, within MM patients we typically observe one specific Ig clonotype. This was further validated by a microfluidic platform for single cell B-cell receptor (BCR) sequencing[36] (**Fig. S8;** Methods). In order to characterize the intra-tumor transcriptional heterogeneity, we developed a *k*NN-based machine learning classifier that segregates normal from malignant PC (**Fig. 2A**; Methods). The classifier annotates PC based on their similarity to the signatures of normal PC in our data. In the control donors, all but a few cells (≤4/1000 cells) are classified as normal PC, while in MM patients, PC are mostly classified as abnormal (**Fig. 2A**). After removal of normal PC, we then clustered the abnormal PC from every individual patient separately, and analyzed the inter-cluster versus intra-cluster correlations (**Fig. 2B**; Methods). We detected substantial intra-tumor heterogeneity, defined by negative intra-cluster correlation ($r<-0.1$) and positive inter-cluster correlation ($r>0.1$), in 10 patients (**Fig. 2B**; Methods). For example, patient SMM02 displays a single BCR clonotype characterized by two distinct transcriptional states: One dominated by *DEFB1* ($p<1x10^{-50}$), a gene that was not previously implicated in myeloma and reported to be a *CCR6* ligand; and the other by the expression of *FRZB*, implicated in the oncogenic Wnt pathway in MM (**Fig. 2C** and **Fig. S9**)[37]. Patient MM11 exhibits two transcriptional states both expressing high levels of *CCND1* and *FRZB*; one state is characterized by significant ($p<1x10^{-50}$) overexpression of *EDF1* (endothelial differentiation related factor 1), involved in lipid metabolis and PPARγ pathway[38,39], while the second transcriptional clonotype overexpresses *PCBD1*, a transcriptional co-activator of *HNF1*[40]. Conversely, the PC of patient SMM01 (positive serum immunofixation for IgAκ) display two transcriptional clonotypes, a small clonotype with Ig heavy chain class A (IGHA), and a larger one expressing only the κ light chain, each with a distinct transcriptional signature (**Fig. S9**).

In order to evaluate if these clonal structures are the product of genomic aberrations we used the single cell genome wide transcriptional data to infer copy number alterations (sciCNA) by averaging the relative expression of a large number of genomically adjacent genes. Using similar strategy to Patel and Tirosh *et al*[18] 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**, Methods). The expression levels were compared to the average signature of the control donor cells. We then compared the sciCNA to the targeted genome sequencing for the same patients (**Fig. 2D** and **Fig.S9**). Our sciCNA analysis shows, as expected, that MGUS patients have less aberrant CNA profile compared to MM patients, that show many 13q deletions, in agreement with the frequency of this aberration in myeloma (**Fig. 2D)**. Importantly, using sciCNA we find that several intra-patient transcriptional clones also harbor genomic aberrations, suggesting that some of the intra-population transcriptional diversity is likely driven by different genome structures (**Fig. 2E**). For example, only one cluster from patient SMM02 showed chromosome 22 deletion, while the other showed a normal sciCNA pattern. For MM06 with 3 transcriptional clones, sciCNA detected 1q amplification together with chromosomes 4 and 14 deletions in all three transcriptional clones, while only in 2 transcriptional clones a chromosome 5 deletion was found (**Fig. 2D-E**). Importantly, in most patients, the differential genes do not necessary reside in the altered chromosomes, suggesting of regulation *in trans*. For example, in patient SMM02,

the *FRZB*, *DEFB1*, *CST3*, *WARS*, are expressed differentially but are not related to chromosome 22 deletion in cluster 1 (**Fig. 2C-E** and **Fig. S9**). Together, these results show that intra-tumor transcriptional and CNA heterogeneity are prevalent in myeloma, and they can be characterized using scRNA-seq profiling. Further, our data suggest that in MM, transcriptional changes are many times regulated *in trans*, and may be associated indirectly to genomic aberrations. It would be important to follow the response to therapy and the risk of disease progression in patients with a single versus multiple transcriptional clones.

Characterizing rare cancer cells in asymptomatic patients and in minimal residual disease

Patients with asymptomatic disease are a highly heterogeneous group with varying risk of developing MM. Currently, limited methods exist for stratifying these patients' molecular signature and risk of progression. We expected these patients to have a lower tumor burden as compared with active MM. We profiled with MARS-seq 7 newly diagnosed patients with MGUS and 6 with SMM (**Table S1**). The disease manifestation of SMM patients, as characterized by MARS-seq, is dramatically different from that of MGUS patients in the number of malignant cells, and closely resembles the profiles of the MM patients. Within the MGUS patients, we detected malignant clusters for 2 patients, MGUS04 and MGUS05, with 69/466 and 482/493 PC displaying a malignant signature, respectively (**Fig. 2B** and **Table S3**). Clustering the plasma cells of MGUS04 shows that the malignant cells (in cluster 4) are characterized by a transcriptional signature overexpressing a known MM driver (*CCND1),* along with *DPEP3* ($p<1x10^{-50}$), a dipeptidase involved in arachidonic acid metabolism, previously associated with triple negative breast cancer, but not with MM (**Table S2**)[41]. These malignant PC 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 asymptomatic patients, 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 patients with residual disease by performing longitudinal scRNA-seq sampling of 5 patients: MM01; MM03; MM07; MM08 and AL01, which were profiled in a dynamic fashion at time of diagnosis and post-treatment. These patients were treated with a bortezomib-based regimen, and all except for AL01 underwent high dose melphalan therapy with autologous stem cell transplantation (**Table S1**). We were able to detect rare (as little as 2%) cancer cells in 5/5 patients with abnormal serum light chain ratios (**Fig. 3 A-C**). Importantly, 2/5 of these patients (MM01 and MM08) were clinically classified as complete responders according to the International Myeloma Working Group criteria (**Table S1**)[32]. Comparing the cancer cells before and after treatment, we find that most of the tumor cells express similar transcriptional programs as compared to the original pre-treatment neoplastic cells of the same patient; specifically, we find that the major MM drivers in these patients such as *CCND1*, *NSD2/MMSET* and *FRZB* are expressed equivalently before and after treatment (**Fig. 3C**). Although most MM overexpressed genes do not change post-treatment, we were able to detect significantly ($p<1x10^{-50}$) differentially expressed genes for a few of the patients, for example the gene *ELK2AP* (member of ETS oncogene family) is overexpressed in MM07 post-treatment compared to the pre-treatment signature; and the gene *LCP1* (lymphocyte cytosolic protein 1), involved in calcium binding, and previously related to several cancers but not to MM, is overexpressed in both AL01 and MM03 post-treatment (**Fig. S10**)[42]. In two additional patients without a baseline sample, we were able to detect as little as 23 tumor PC (1.7% of PC; with a defined MM program): For example, patient MM13 with overexpression of *MAF*, *ITGB7* and *CCND2*, and patient MM14 with *NSD2*/*MMSET*, *PDIA2*, *AZGP1* and *MDK* (**Fig. S11** and **Table S6)**. This demonstrates that scRNA-seq is a powerful tool to dissect PC heterogeneity and identify rare neoplastic states in the setting of low tumor burden and minimal residual disease (MRD). The relative stability of

MM driver genes expression pre- and post-treatment suggests that targeted therapy approaches in MRD settings might be an effective strategy, and warrants further studies.

Circulating tumor cells display similar transcriptional states to the patient BM tumor

In myeloma, and especially in its asymptomatic predecessor states, the circulating tumor cells (CTC) load in the peripheral blood is low, complicating a non-invasive and accurate liquid biopsy analysis, due to contaminations of various immune cells and normal circulating PC. Previous studies utilizing whole exome sequencing found that somatic single nucleotide variants are shared between the blood and BM in 84% of patients with active MM[43]. While the latter study utilized a patient-specific sorting strategy in cases with a positive aberrant surface marker, others chose a wider and less-specific approach[44,45]. 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 CTC reflect the BM disease. To test the potential of scRNA-seq applications for accurate CTC characterization, we applied MARS-seq on PC from both BM and blood from 19 different patients and 2 control subjects (hip09 and hip10; **Fig. S3**). In order to develop transcriptional and protein markers for efficient purification of CTC from the patients' blood, we initially clustered the circulating plasma cells from all 21 individuals together. In addition to the 11 patient-specific clusters, we noted a shared cluster ('cPC4') of polyclonal cells with a plasmablast signature common to most patients, including cells from the control donors (**Fig. S12**). Using flow cytometry in a different cohort of relapsed MM patients, we show that CTC with aberrant surface profile have lower protein expression of CD52 compared to non-CTC (**Fig. S13**). In order to compare, for each patient, the CTC with his/her BM tumor cells, we first removed the normal circulating PC by excluding cells with cluster 'cPC4' characteristics (Methods). We also excluded 4 patients with less than 20 CTC from further analysis. Comparing the remaining malignant circulating PC to the malignant BM PC for each patient, we observe that in all cases (15/15), the CTC signatures highly resemble the BM transcriptional state(s), with a few changes likely resulting from the different environments (e.g. expression of *CRIP1* and *KLF6*) (**Fig. 4A-C**). In order to further validate our findings, we compared the BCR clonotype of the patients' BM PC to CTC. The tumor load in the BM and the blood is different by several orders of magnitude, affecting the confidence in our analysis of a few patients with small CTC clones (<20 cells; Methods). Overall, in 11/15 patients we find a good match in the BCR between BM and blood samples (**Fig. S12**). Together, we find that circulating PC in the patients' blood are composed of clonotypic CTC, that reflect the transcriptional status of the BM disease, and additional normal polyclonal plasmablasts. We further devise an efficient sorting strategy for CTC by excluding contaminations of circulating plasmablasts. Our approach can be applied to molecularly characterize a patient's malignant PC in an iterative fashion using liquid biopsies, omitting the need for invasive BM 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 single cell RNA-seq. Data on thousands of PC from 11 control donors is used to characterize plasma cell heterogeneity within normal BM samples, showing polyclonal BCR repertoire and limited inter-individual transcriptional variation. Based on this reference, scRNA-seq provides high sensitivity and confidence to identify and characterize neoplastic PC in low burden disease settings, such as asymptomatic MGUS and SMM, and suggests a direct molecular assay for tracking early MM onset. We find that SMM patients, although asymptomatic, are indistinguishable from active MM patients at the molecular level. 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, providing important potential for guiding and optimizing personalized treatments, and 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 that must be monitored during pre-MM stages or post-treatment.

This study also highlights several remaining challenges. Exploring the immune microenvironment together with the PC from the same patient may highlight potential new targets for immunotherapy, and predict response to specific treatments. MM patients may have a patchy infiltration pattern in the BM, and by sampling a single BM site during a routine clinical diagnostic procedure, we may underestimate the true heterogeneity within the tumor[46]. Our cohort is lacking high risk patients with *TP53* deletion, and the robustness of heterogeneity detection should be tested in that setting as well. We note that we have used a 3' based mRNA sequencing method, and thus are limited to infer coding sequence mutations and splice variants, this can potential be accessed using full length scRNA-seq methods[47]. We show that scRNAs-seq data can accurately infer copy number alterations in MM. Even though aberrant CNA are common in our cohort, our analysis suggests that *trans* acting mechanisms dominate the tumor PC transcriptional state, as we show for *LAMP5*.

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 myeloma patients at various stages of the disease, which may open the way to larger scale studies, and facilitate the design of new and molecularly informed diagnoses and treatment strategies.

Acknowledgments

We thank the patients and their families; the Benoziyo Family Fund, Clalit Health Care Services and Mrs. and Mr. Barry Lang for supporting this study; the clinical coordinators Dikla Yaish, Ido Mamman and Sivan Levy; Dr. Nadia Voskoboinik for iFISH analysis; Mr. Kirill Kogan for help with REDCap installation and to Mr. Amir Giladi for review of the manuscript. I.A. is supported by the Chan Zuckerberg Initiative (CZI), the HHMI International Scholar award, the European Research Council Consolidator Grant (ERC-COG) 724471-HemTree2.0, an 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 European Research Council (ERC) (EVOEPIC). B.P. is supported by European Research Council (ERC) 2015 Starting Grant (MYELOMANEXT). Raw and processed single cell RNA-sequencing data were deposited to NCBI GEO with accession number GSE117156. Raw and processed genomic DNA targeted sequencing data were deposited to NCBI ClinVar with accession number SCVxxx.

Main Figure Legends

Figure 1. MM patients display unique transcriptional signatures that converge into defined malignant pathways. (**A**) tSNE (t-distributed stochastic neighborhood embedding) plot depicting 20,568 single BM PC derived from 29 newly diagnosed patients ('MGUS01-07', 'SMM01-06', 'MM1-12', '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 heatmap in Fig. S5). **(B)** Patients are colored by a severity gradient projected on the tSNE (grey– control; yellow– MGUS; pink– SMM; red– MM and AL, these colors correspond to panel E). (**C**) Index sorting flow cytometry data represented as mean fluorescent intensity (MFI, $\log_{10}$ scale) for specific surface markers, projected onto the tSNE (upper panel CD19; lower panel CD56). (**D**) Bar plot showing distribution of cells from the patients/control donors across the clusters (as in Fig.1 A). Patients are color-coded according to disease severity as in A, names above bars correspond to the patient with the majority of cells in each cluster. (**E**) Boxplots of single cell gene expression for specific genes across the 29 newly diagnosed patients and 11 control donors (left panel). 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. Patients are color-coded according to disease severity. For each gene, corresponding histograms of bulk RNA-seq expression estimates from the CoMMpass study (TPM in log scale) are shown (right panel).

Figure 2. Intra-tumor heterogeneity in myeloma. (A) Shown are P-values to reject the null hypothesis (that a cell belongs to the control PC group) for all 20,586 cells. Dots represent individual PC, classified as either 'normal' (black) or 'abnormal' (red; $p<0.01$). Patients are ordered according to average score value from low to high (Methods). (**B**) Intra-tumor heterogeneity measure for 21 patients with abnormal PC (as classified by the score used in Fig. 2A). This was calculated by plotting the correlations within and between clusters (for each patient separately; Methods). **(C**) Heatmaps showing clustering analysis of BM PC for patients MM04 (upper left, 429 cells), MM09 (upper right, 884 cells), SMM02 (lower left, 1,645 cells) and MM11 (lower right, 437 cells), clustered with the same number of randomly sampled normal BM PC from control individuals. Representative variable genes are shown. (**D)** Heatmap showing copy number alterations inferred from scRNAseq (sciCNA) for each patient, averaged by intra-patient clustering. **(E)** sciCNA profile for SMM02 (top) and MM06 (bottom); each line represents a cluster-averaged CNA profile.

Figure 3. Characterization of rare residual malignant cells post-therapy (A) Histograms of normal to malignant score (Methods) for 5 MRD patients showing the distribution of scores for control cells (from control donors; green, top), the patient's pre-treatment cells (gray, middle) and post-treatment cells (red, bottom). **(B)** Heatmaps showing normalized single cell gene expression of BM PC pre- and post- treatment for patients 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 (Methods) shown in the upper panel, grey background represents the malignant cell fraction. **(C)** Boxplots showing gene expression of representative genes pre- and post-treatment for 5 MRD patients. 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.

Figure 4. Circulating PC are composed of CTC that reflect the BM disease (A) Boxplots showing gene expression of representative genes from 15 patients for whom $N_{CTC}>20$. Each patient is represented by a different color. Shown are pairs of BM PC and circulating plasma cells (cPC) for each patient. 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 BM PC (x-axis; 7,969 cells) and CTC (y-axis; 2,299 cells) across 15 patients. Patients' color codes correspond to Fig. 4A. **(C)** Two-dimensional tSNE views of paired BM PC (upper panels) and CTC in the blood (lower panels) from the same patient. Projection of single cells (black) from a specific patient (SMM02- 1,755 BM PC and 216 CTC; MM07-

274 BM PC and 138 CTC; MM06- 976 BM PC and 128 CTC; MM08- 416 BM PC and 164 CTC; AL03- 589 BM PC and 267 CTC; MM01- 671 BM PC and 141 CTC) is shown on a gray background of all BM PC and CTC (10,268 cells).

Table 1. Clinical characteristics of newly diagnosed patients including risk stratification by disease status.

Supplementary Figure Legends

Figure S1

Schematic representation of study design

Figure S2

Representative flow cytometry plots showing sorting strategy (CD38+ CD138+) for PC after doublet exclusion **(A)** BM PC of patient MM08 **(B)** Circulating plasma cells (cPC) of patient MM04 **(C)** BM PC of patient MGUS05 **(D)** BM PC of patient MM13 (active myeloma with minimal residual disease). Plots were generated using FlowJo software (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 (Methods).

Figure S3

BM and peripheral blood single cell QC metrics. **(A)** Shown are number of reads, number of UMIs and % cell analyzed per batch of 384 cells (that were pooled for library construction) for all CD38+ CD138+ BM single cells from 29 newly diagnosed patients and 11 control donors. Cells were sorted into plates, sequenced, underwent QC evaluation, and filtering-out of non-PC contaminants (Methods). **(B)** Shown are number of reads, number of UMIs and % cell analyzed per batch of 384 cells (that were pooled for library construction) for all CD38+ CD138+ peripheral blood single cells from 19 newly diagnosed patients and 2 control donors. Cells were sorted into plates, sequenced, underwent QC evaluation, and filtering-out of non-PC contaminants (Methods). **(C)** Table summarizing the total CD38+ CD138+ cell numbers from BM and blood of newly diagnosed patients and control donors, that were sorted and sequenced, passed QC, and analyzed (after filtering for non-PC contaminants; Methods).

Figure S4

(A) Plot depicting plasma cell *in silico* filtering according to immunoglobulin load and UMI count per cell (Methods) **(B)** Heatmap showing clustering analysis of 3,179 'contaminating' cells that pass QC but do not express Ig gene above the cutoff (100 UMIs per cell). Representative genes (mostly unrelated to PC program) are shown.

Figure S5

(A) Heatmap showing clustering analysis of 20,586 BM PC sorted from 29 newly diagnosed patients and 11 control donors, featuring normalized single cell expression level of the 100 most variable genes (Methods). **(B)** Dendrogram showing the hierarchical clustering of the average transcription profile for all clusters C1-29 (related to Fig. S5). **(C)** Cell-to-cell correlation matrix of 200 randomly selected cells from each patient (with $N>200$ cells) (left panel); kNN ($k=100$) adjacency matrix showing, for each cell, it's 100 nearest neighbors (blue).

Figure S6

(A) Normalized single cell expression (UMI count, log scale) of 16 representative genes, projected onto tSNE map of all 20,586 BM PC from 29 newly diagnosed patients and 11 control donors. **(B)** Distribution of long-lived vs. short lived PC in 40 individuals, donor individuals and MM patients. Short-lived PC bars are colored by a gradient that corresponds to Fig. 1B (grey– control; yellow– MGUS; pink– SMM; red– MM and AL).

Figure S7

(A) Genomic interphase fluorescent in situ hybridization (iFISH) images of BM PC magnetically enriched for CD138+ for patients: 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). **(B)** FACS plots showing intra-cellular staining (gated on live cells) for LAMP5 protein (red) compared with isotype control (blue) for KHM1B and RPMI-8226 cell lines (upper and lower panel, respectively; Methods). **(C)** Genome browser view of normalized H3K4me2 profiles of peaks found in a 1Mb region in the *LAMP5* locus. Data is from two independent biological replicates (Methods).

Figure S8

(A) Heatmap depicting the relative frequency of the Ig sequences from analyzing 20,586 BM PC of 29 newly diagnosed patients and 11 control donors. Upper panel – light chain variable region (IGKV and IGLV); middle panel – light chain constant region (IGKC and IGLC); lower panel – Ig heavy chain (IGHC) constant region. **(B)** Chromium 10x single cell BCR clonotype distribution of control donor 'hip13', created by Chromium's Loupe V(D)J browser (Methods). Left panel: Heatmap showing cell frequency for specific IGH/L/K variable (V) region sequences (x-axis) and IGH/K/L joining (J) region sequences (y-axis). Right panel: Frequency of different BCR clonotypes (inferred from the heatmap) **(C)** Plots of single cell BCR data for patient MM02 using Chromium 10x single cell BCR platform (Methods). Left panel: Single immunoglobulin light chain λ variable region sequence (IGLV) for patient MM02. Right Panel: Frequency of different BCR clonotypes for patient MM02, generated by Chromium's Loupe V(D)J browser.

Figure S9

(A) Comparison of CNVKit output from genomic DNA targeted sequencing panel at 500x coverage (black dots) and single cell RNA-seq inferred copy number alteration (sciCNA; blue) for patient MM10. **(B)** Scatter plot of CNVKit vs. sciCNA estimation for each DNA segment for patient MM10. **(C)** Cell to cell correlation matrix for patients SMM02 (n=1,664 cells), MM04 (n=439 cells) and MM09 (n=882 cells). **(D)** Enriched gene ontology (GO) terms and pathways for up-regulated genes for each of patient's SMM02 clusters (SMM-C1 and SMM-C2) with $\log_2$(fold change) > 1.5 and $\log_{10}P$ > 10 compared with normal PC from control donors (Methods). **(E)** Heatmaps of normalized single cell gene expression (UMI count, log scale) for BM PC from patients SMM01 (left, 650 cells) and AL03 (right, 548 cells), clustered with the same number of normal BM PC from control donors (Methods). Representative variable genes are shown.

Figure S10

(A) Heatmap showing transcriptome differences between pre- and post- treatment PC from patients and normal PC from donors. Clustering was done with the Metacell algorithm (Methods). Differentially

expressed genes in MRD PC are highlighted. **(B)** Scatter plots of single cell expression (average UMI count, log scale) for of *ELKAP2* (left panel) and *LCP1* (right panel) genes in patients pre- and post-treatment.

Figure S11

(A) Chromium 10x 5' single cell expression estimates (Methods) visualized as tSNE plots for patient MM13 BM cells enriched with CD138 magnetic beads (Methods). Shown are expression estimates for immunoglobulin λ variable region V3-25 gene (IGLV3-25; left), PDIA2 gene (middle); NSD2 gene (right). **(B)** Heatmap of normalized single cell gene expression (UMI, log scale) for BM PC from patient MM14 (1,252 cells). Representative variable genes are shown. **(C)** Scatter plots showing average single cell gene expression (log_2 scale) of control donors' BM PC (x-axis) compared with either MM14 cells from cluster 1-4 (related to heatmap in panel B; upper scatter plot) or MM14 cells from cluster 7 (lower scatter plot). Specific gene names are shown.

Figure S12

(A) Heatmap showing the distribution of circulating plasma cells from 21 patients in 12 clusters 'cPC1-12'). **(B)** Two dimensional tSNE view of circulating plasma cells from 21 patients. Each dot represents a single cell. Patients are color coded. **(C)** Bar plots of average UMI count for *CD52* gene (Methods) in common cluster 'cPC4' (left, blue bars) compared to CTC (right, red bars). **(D)** Scatter plots of single cell expression estimates (average UMI count, log scale) for patients' abnormal BM PC (x-axis) to CTC in the blood (y-axis). Upper panel – SMM03, (n=198 BM PC and n=16 CTC); lower panel – MGUS05 (n=188 BM PC and n=211 CTC). **(E)** Ig light chain variable region distribution within each patient's BM and blood tumor cells. Color represents percentage of cells ranging from white to blue (0-0.5). In each box, upper panel represents peripheral blood (PB) and lower panel represents BM. Cell numbers are shown on the left.

Figure S13

(A) Flow cytometry plots of circulating plasma cells from 3 myeloma patients with relapse post-treatment. Plots were generated using Infinicyt software (Methods). CTC (red) are marked by an aberrant surface phenotype, compared to normal PC (green). PCA (principal component analysis) quantifies the significance (contribution to principal component 1) of each surface marker to separate between 'CTC' to 'normal PC'. Each row represents a different patient.

Supplementary Tables

Table S1

Detailed demographic, clinical, and disease treatment and progression data for all patients.

Table S2

Average UMI per cell for differentially expressed genes, distributed within clusters C1-29. Data is from 20,568 BM PC related to 29 newly diagnosed patients and 11 control donors. Corresponding p-values for Mann-Whitney U a-paramtric test with FDR correction.

Table S3

Patients distribution within and across clusters. Shown are cell numbers that relate to specific patients (rows), and their corresponding clusters (columns). Data is from 20,568 BM PC analyzed from 29 newly diagnosed patients and 11 control donors.

Table S4

Results of genomic DNA targeted sequencing data for myeloma (MyType) for 11 patients, with corresponding clinical iFISH data.

Table S5

Enriched gene ontology (GO) terms and pathways per cluster, compared with cluster C1 of normal PC from control donors.

Table S6

Comparison of differential expression analysis of MRD patient with small malignant clone (MM13) between 10X genomics and MARS-seq. Differential expression was performed relative to normal PC sequenced using the same single cell platform.

Methods

Harvesting of bone marrow plasma cells during hip replacement surgery

Individuals with isolated hip osteoarthritis who are otherwise healthy, were recruited by the orthopedic department in Tel Aviv Medical Center, Tel Aviv, Israel. Procedure was performed in the operating theatre, as described earlier[23], with several modifications. Briefly, after informed consent (in accordance with Helsinki declaration) 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 1000IU. Bony fragments were removed by forcing cells through a metal sieve, diluted 1:1 with ice cold FACS buffer (EDTA pH8.0 2mM, BSA 0.5% in PBS), placed on ice and immediately transported to the lab.

Obtaining patients' plasma cells from iliac crest aspirates and peripheral blood

Patients suspected for plasma cell neoplasm were recruited to the study from hematology departments in 7 medical centers in Israel. After informed consent, bone marrow aspiration and peripheral blood sampling (20ml) were placed in EDTA-containing tubes (Beckton Dickenson). Tubes were mixed, placed on ice, and immediately transported to the lab**.**

Patient's clinical and demographic data

Clinical study data were collected and managed using REDCap electronic data capture tools hosted at Weizmann Institute of Science. REDCap (Research Electronic Data Capture) 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[48].

Single Cell Sorting

Bone marrow cells were diluted 2:1 in ice cold FACS buffer (EDTA pH8.0 2mM, BSA 0.5% in PBS), washed and strained with a 100µm strainer. Peripheral blood cells were diluted 1:1 in ice cold FACS buffer. Mononuclear cell separation was performed by density centrifugation media (Ficol-paque, GE Life Sciences) in a 1:1 ratio with diluted blood or marrow cells. Centrifugation (460g, 25min) 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) for 5min at 4°C and washing, peripheral blood cells were enriched for CD38 with magnetic beads (Miltenyi), washed and stained with antibodies (all from Cytognos or BD Biosciences): CD38, CD138, CD56, CD19, CD117, CD27, CD45, CD81 or CD52.
Bone marrow cells were stained without prior magnetic bead enrichment. Samples were filtered through a 40-µm strainer before commencing sorting. Single cell sorting was performed using either FACS SORP-AriaII or AriaFusion (BD Biosciences, San Jose, CA). After doublets exclusion, isolated cells were single-cell index-sorted into 384-well cell capture plates containing 2µL of lysis solution and barcoded poly(T) reverse-transcription (RT) primers for single-cell RNA-seq. Four empty wells were kept in each 384-well plate as a no-cell control for data analysis. Immediately after sorting, each plate was spun down to ensure cell immersion into 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 FACS Diva 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.

Massively Parallel Single-Cell RNA-seq library preparation (MARS-seq)
Single-cell libraries were prepared as previously described[21,22,49]. Briefly, mRNA from cells sorted into cell capture plates are barcoded and converted into cDNA and pooled using an automated pipeline. The pooled sample is then linearly amplified by T7 *in vitro* transcription, and the resulting RNA is fragmented and converted into a sequencing-ready library by tagging the samples with pool barcodes and Illumina sequences during ligation, RT, and PCR. Each pool of cells was tested for library quality and concentration was assessed as described earlier[21,22,49]. Overall, barcoding was done in 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 single-cell RNA-seq data
MARS-seq libraries, pooled at equimolar concentrations, were sequenced using an Illumina NextSeq 500 sequencer, at a sequencing depth of 50K-100K reads per cell. Reads are condensed into original molecules by counting same unique molecular identifiers (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[49]. Mapping of reads was done using HISAT (version 0.1.6); reads with multiple mapping positions were excluded. Reads were associated with genes if they were mapped to an exon, using the UCSC genome browser 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 less than 500 UMIs were discarded from the analysis. Genes with mean expression smaller than 0.001 UMIs/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 Ig annotated genes) using a cutoff of 100 UMIs per cell. This cutoff was selected based on 2-gaussians mixture model (Fig. S4).

Graph-based clustering analysis

In order to assign cells to homogeneous clusters we used the PhenoGraph clustering algorithm[50]. 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 Uij contains the number of unique molecular identifiers (UMIs) from gene i that were found in cell j. PhenoGraph first builds a *k*-Nearest Neighbors (*k*NN) graph using the Euclidean distance (*k*=30) and then refines this graph with the Jaccard similarity coefficient, where the edge weight between each two nodes is the number of neighbors they share divided by the total number of neighbors they have[50]. 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 FDR correction (Matlab R2016a *ranksum* function).

In order to evaluate the robustness of our clustering analysis, we performed clustering with our more sensitive in-house analysis package Metacell[51]. Briefly, informative genes were identified and used to compute cell-to-cell similarity to build a *k*NN graph to group cells into cohesive groups (or meta-cells). Then, the algorithm uses bootstrapping to derive strongly separated clusters, as previously described[51]. We also compared the results with clustering using Seurat[52]. Our PhenoGraph clustering analysis shows great agreement with Metacell and Surat.

2D projection

Cells are visualized in two dimensions using t-Distributed Stochastic Neighbor Embedding (tSNE, matlab 2017a *tsne* function).

Myeloma cell lines

RPMI-8226 and KHM1B myeloma cell lines were purchased from American Type Culture Collection (Manassas, VA) and the Japan Cell Repository Bank (Osaka, Japan), respectively. Cells were cultured using an aseptic technique in RPMI medium (Gibco) supplemented with 10% heat-inactivated fetal bovine serum, 1mM sodium pyruvate, 2mM L-glutamine, 1% penicillin-streptomycin (ThermoFisher Scientific). Cells were stored in 10-50ml flasks (Corning) in an incubator (ThermoFisher Scientific) with humidified air and 5%CO_2, at 37ºC 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, Beit Ha'emek, Israel). For flow cytometry intra-cellular staining, cells were first stained with Live-Dead Violet (Invitrogen), washed and fixed-permeabilized with Foxp3/Transcription Factor Staining Buffer Set (eBioscience), followed by staining of either control REA(I)-PE or anti-human LAMP5-PE (Miltenyi), and FACS analysis (FlowJo, BD, San Jose, CA).

Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) libraries construction

Chromatin immunoprecipitation and libraries preparation were performed as described earlier, with a few modifications[53]. Briefly, viable cells (negative for Live-Dead Violet, Invitrogen) were sorted into FACS buffer, fixed for 10 minutes with 1% formaldehyde (Sigma) at room temperature, followed by quenching with 0.125M glycine and washing with ice-cold PBS. Cross-linked cells were resuspended in lysis buffer (12 mM TrisCl pH8, 0.1X PBS, 6 mM EDTA) supplemented with protease inhibitor (Roche). Chromatin was sheared using NGS Bioruptor Sonicator (Diagenode). The sonicated cell lysate (Whole Cell Extract) was incubated with 2.5 µg H3K4me2 antibody (Abcam) at 4°C for 5 hr, and for an additional hour with Protein G magnetic beads (Invitrogen). 96 well magnet was used (Invitrogen) in all further steps. Cell lysate was removed and samples were washed 5 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; 200ul per wash), twice with RIPA buffer supplemented with 500 mM NaCl (200ul per wash),

twice with LiCl buffer (10 mM TE, 250mM LiCl, 0.5% NP-40, 0.5% DOC), once with TE (10Mm Tris-HCl pH 8.0, 1mM EDTA), and then eluted in elution buffer (0.5% SDS, 300 mM NaCl, 5 mM EDTA, 10 mM Tris HCl pH 8.0). The elute was treated sequentially with 2ul of RNaseA (Roche) for 30 min and 2.5 ul of Proteinase K (NEB) for two hours, and then reverse crosslinked over night at 65ºC.
DNA was purified by mixing reverse-crosslinked samples with paramagnetic SPRI beads (Agencourt AMPure XP, Beckman Coulter), incubated for 4 minutes. Beads were washed on the magnet with 70% ethanol and then air dried for 4 minutes. 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, A-base addition, adaptor ligation and enrichment) the same SPRI beads cleanup was used. DNA ends are first repaired by T4 polymerase (NEB). Next, T4 polynucleotide kinase (NEB) adds a phosphate group at the 5′ ends. An adenosine base is then added to the blunt-ended fragments, using Klenow enzyme (NEB), and a barcode Illumina compatible adaptor (IDT) was ligated to each fragment using T4 quick ligase (NEB). DNA fragments were amplified by 12 cycles of PCR (Kapa HiFi HotStart PCR ReadyMix, Kapa Biosystems) using specific primers (IDT) 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 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. Normalization of peaks was done by dividing reads inside peaks with the average of all reads**.** For IGV snapshot we used 1 M bp window around *LAMP5* gene.

Genomic interphase fluorescent in situ hybridization (iFISH)
BM 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, Cambridge, UK), per manufacturer instructions. For analysis, 50 nuclei were counted per slide. Karyotype (G-banding) was performed using the non-enriched BM 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, Downington, PA). 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 PC (CD38+, CD138+; **Fig. S2**) into PBS was performed using either FACS SORP-AriaII or AriaFusion (BD Biosciences, San Jose, CA). After centrifugation (300g 10min), supernatant was aspirated, and pellet was snap frozen. DNA extraction was performed using Universal Quick-DNA Miniprep Kit (Zymo Research, Irvine, CA), and quantified using a NanoDrop One spectrophotometer (ThermoFisher Scientific, Waltham, MA).

Preparation of genomic DNA libraries for targeted sequencing

All 11 tumor samples, and other 16 unmatched BM 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, IGH rearrangements as well as arm-level copy number alterations. The target-enrichment design was based on DNA pull-down by cRNA baits (SureSelect, Agilent Technologies, Santa Clara, CA). A total of 11 patient 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-bp paired-end protocol.

Targeted genomic DNA sequencing analysis

Alignment

Short insert paired-end reads were aligned to the GRCh37 reference human genome with 1000 genomes decoy contigs using 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 758.6X per sample

Somatic Mutation Calling

Single base substitutions were called using CaVEMan (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 a third of the alleles containing the mutant must have base quality >=25; If mutant allele coverage >= 10X, there must be a mutant allele of at least base quality 20 in the middle 3rd of a read. If mutant allele coverage is <10X, a mutant allele of at least base quality 20 in the first 2/3 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 than 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, regions of excessive depth (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% mutant allele burden were excluded. Variant annotation was done in Ensembl v74 using VAGrENT.

Small insertions and deletions

Small somatic insertions and deletions (indels) were identified using a modified version of Pindel (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 <200X, mutant variant must be present in at least 8% of total reads; for regions with sequencing depth >=200X, 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: The number of mutant reads mapped by BWA in the tumor sample is greater than 0 or: The number of mutant reads mapped by BWA in the tumor sample is equal to 0 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; Pindel 'SUM-MS' score (sum of the mapping scores of the reads used as anchors) >=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 v74 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 sub-clonal variants at very low frequencies in the tumor samples mutation calling using secondary pipelines were done. Strelka 2 (v2.8.3) [https://doi.org/10.1101/192872] 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 for downstream analysis.

The following filters were applied on calls identified by primary and secondary pipelines: Filter calls with >3% MAF in Exac (Version 0.3) or 1000 Genomes; Filter calls with >0.5% MAF in Exac or 1000 Genomes unless present in COSMIC (v81); Filter calls present in 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 MMRF 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 of a mutation (±9 bp). All calls retained were manually curated.

Structural rearrangements

Given the smaller fragment insert sizes in targeted capture, the 100bp paired-end reads were trimmed to 50bp from the 3' end of the read for better discover of instructural 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 less than 5 reads were excluded. Additionally, translocations in which either of the break-points 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 (Version: 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 break-points 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.

Copy number alterations (CNA)

CNVKit [https://github.com/etal/cnvkit] was used to identify somatic copy number alterations in the data. To negate sample specific biases in CNA analysis, all 16 control samples were combined into a pooled reference. Each tumor sample is then compared with the pooled reference to identify somatic CNA 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 (CBS) algorithm is 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 (blast.ncbi.nlm.nih.gov, version 2.2.26, with e-value bound of 1e-10) to the IMGT reference sequences[35] (updated August 2016). For each cell we choose heavy chain constant region, light chain constant and variable region based on the highest coverage and longest coverage.

Single cell 5' transcriptome and BCR-seq (Chromium's 10x)

For one patient (MM02) and control (hip13), BM mononuclear cell fraction was split, with half proceeding 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 10x Chromium microfluidics system, according to manufacturer's guidelines. Two sets of libraries were prepared from the 10x loaded samples: 5' mRNA library; and single cell BCR-seq library, using custom primers for BCR amplification, according to manufacturer's instructions. 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. Single cell BCR-seq 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 scBCR-seq data, we used Chromium's V(D)J-Loupe for analysis of BCR clonotypes and 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 kNN based classifier. Our method is based on the observation that the transcriptome profiles of malignant cells are very different from normal PCs, as clearly observed in our dataset. We first calculate the similarity between a given PC and all normal PCs using spearman correlation, to prevent potential batch effects, we excluded cells sharing the same source (patient) of a given cell. We then select the top K=100 most similar cells. We chose K=100 to ensure that the selected cells are from the same sub-population, although we only observed two very close sub-populations in our deep analysis on the normal PCs. The distribution of spearman correlation to the K=100 most similar cells is normally distributed. After estimating the normal distribution parameters, we used the normal cumulative distribution function with Bonferroni correction for multiple tests to calculate the p-value of each cell as normal. The lower the p-value is, the cell is more likely to be 'abnormal' (malignant).

Intra-tumor heterogeneity score

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

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

Where i U^g

is the UMI counts of gene *g* in cell *i* and N is the total number of cells. For patients with substantial transcriptional heterogeneity, we would expect a negative inter-cluster correlation and a positive within-clusters correlation. Patients with a uniform transcriptional state, would have near zero inter-cluster correlation. Finally, after devising this score, we have manually inspected each patient's

clustering results which confirmed our analysis and showed that indeed the patients with the most substantial intra-tumor transcriptional heterogeneity show negative inter-cluster correlation.

Gene Ontology (GO) analysis

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

Inferring copy number alterations from scRNAseq data

Copy number alterations were inferred from single cell RNA-seq as previously described[18] with modifications to Metacells (clusters of cells). 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+UMI)) and absolute values of fold change were bound by 3. For this analysis we only used 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 MRD 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 equivalent size group of normal PC (sampled from the healthy cohort). For each patient i we define the group G^i to contain N pre-treatment malignant cells and N normal control PCs (total $2N$ cells). For each cell c from patient i post-treatment we calculate the correlation to all cells in group G^i. Next, we sort the vector of correlations and save 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:

M M M M H M H H H H

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 by

$$UU = RR - \frac{NN(NN+1)}{2}$$

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:

$RR = 1 + 2 + 3 + 4 + 6 = 16$ and $UU = 16 - \frac{5\times6}{2} = 1$

U is approximately normally distributed, we standardize the values of U and gives a score between -1 to 1 where 1 represents all malignant cells proceeds the healthy cells in our rank vector (MMMMMHHHHH) and -1 the opposite.

Flow cytometry analysis of CTC from relapsed myeloma patients

BM and blood from 3 relapsed myeloma patients were analyzed, as previously described[54]. Aberrant PC were identified either by antigen under-expression (CD19, CD27, CD38, CD45, CD52, CD81) or antigen over-expression (CD56, CD138). Data acquisition was performed in a FACSCantoII flow cytometer (BD, San Jose, CA) using the FACSDiva 6.1 software (BD Biosciences). Data analysis was performed using the Infinicyt software (Cytognos SL, Salamanca, Spain). PCA (principal component

analysis) quantifies the significance (contribution to principal component 1) of each surface marker to separate between 'CTC' to 'normal PC'. Each row represents a different patient.

References

1. Palumbo, A. & Anderson, K. Multiple myeloma. *N. Engl. J. Med.* **364,** 1046–60 (2011).
2. Rajkumar, S. V. Multiple myeloma: 2016 update on diagnosis, risk-stratification, and management. *Am. J. Hematol.* **91,** 719–34 (2016).
3. 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,** 1121–7 (2010).
4. Morgan, G. J., Walker, B. A. & Davies, F. E. The genetic architecture of multiple myeloma. *Nat. Rev. Cancer* **12,** 335–48 (2012).
5. Dhodapkar, M. V. MGUS to myeloma: a mysterious gammopathy of underexplored significance. *Blood* **128,** 2599–2606 (2016).
6. Bolli *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).
7. Chapman, M. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471,** 467–472 (2011).
8. 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).
9. Lohr, J. G. *et al.* Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell* **25,** 91–101 (2014).
10. 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).
11. Laganà *et al.* Integrative network analysis identifies novel drivers of pathogenesis and progression in newly diagnosed multiple myeloma. *Leukemia* (2017). doi:10.1038/leu.2017.197
12. Shah *et al.* Prediction of outcome in newly diagnosed myeloma: a meta-analysis of the molecular profiles of 1905 trial patients. *Leukemia* (2017). doi:10.1038/leu.2017.179
13. 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,** 2276–84 (2007).
14. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5,** 2997 (2014).
15. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* (2016). doi:10.1038/nrg.2015.16
16. Giustacchini, A. *et al.* Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat Med* **23,** nm.4336 (2017).
17. Melchor, L. *et al.* Single-cell genetic analysis reveals the composition of initiating clones and phylogenetic patterns of branching and parallel evolution in myeloma. *Leukemia* **28,** 1705–15 (2014).
18. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–401 (2014).

19. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–96 (2016).
20. Paiva *et al.* Differentiation stage of myeloma plasma cells: biological and clinical significance. *Leukemia* (2016). doi:10.1038/leu.2016.211
21. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–9 (2014).
22. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163,** 1663–77 (2015).
23. 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).
24. Halliley, J. *et al.* Long-Lived Plasma Cells Are Contained within the CD19−CD38hiCD138+ Subset in Human Bone Marrow. *Immunity* **43,** 132–145 (2015).
25. 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,** 260–4 (1997).
26. Pawlyn, C. & Morgan, G. Evolutionary biology of high-risk multiple myeloma. *Nat Rev Cancer* **17,** 543–556 (2017).
27. Combes, A. *et al.* BAD-LAMP controls TLR9 trafficking and signalling in human plasmacytoid dendritic cells. *Nat Commun* **8,** 913 (2017).
28. Defays, A. *et al.* BAD-LAMP is a novel biomarker of nonactivated human plasmacytoid dendritic cells. *Blood* **118,** 609–17 (2011).
29. 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. U.S.A.* **88,** 3451–4 (1991).
30. Hellström, I. *et al.* The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma. *Cancer Res.* **63,** 3695–700 (2003).
31. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nat. Rev. Immunol.* **15,** 160–71 (2015).
32. Kumar, S. K. & Rajkumar, S. V. The multiple myelomas - current concepts in cytogenetic classification and therapy. *Nat Rev Clin Oncol* (2018). doi:10.1038/s41571-018-0018-y
33. Rajan, A. M. & Rajkumar, S. V. Interpretation of cytogenetic results in multiple myeloma for clinical practice. *Blood Cancer J* **5,** e365 (2015).
34. Puig *et al.* The predominant myeloma clone at diagnosis, CDR3 defined, is constantly detectable across all stages of disease evolution. *Leukemia* **29,** 1435–1437 (2015).
35. Lefranc, M.-P. *et al.* IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res* **43,** D413–D422 (2015).
36. Zheng, G. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8,** 14049 (2017).
37. 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,** 2483–94 (2003).
38. Zhao, X.-Y. Y. *et al.* Long noncoding RNA licensing of obesity-linked hepatic lipogenesis and NAFLD pathogenesis. *Nat Commun* **9,** 2986 (2018).
39. Leidi, M., Mariotti, M. & Maier, J. Transcriptional coactivator EDF-1 is required for PPARγ-stimulated adipogenesis. *Cell Mol Life Sci* **66,** 2733–2742 (2009).
40. Simaite, D. *et al.* Recessive mutations in PCBD1 cause a new type of early-onset diabetes. *Diabetes* **63,** 3557–64 (2014).

41. Chen, X. *et al.* Prognostic value of diametrically polarized tumor-associated macrophages in multiple myeloma. *Oncotarget* **8,** 112685–112696 (2017).
42. 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).
43. Mishima, Y. *et al.* The Mutational Landscape of Circulating Tumor Cells in Multiple Myeloma. *Cell Rep* **19,** 218–224 (2017).
44. Lohr, J. *et al.* Genetic interrogation of circulating multiple myeloma cells at single-cell resolution. *Science Translational Medicine* **8,** 363ra147–363ra147 (2016).
45. Manier *et al.* Whole-exome sequencing of cell-free DNA and circulating tumor cells in multiple myeloma. *Nat Commun* **9,** 1691 (2018).
46. Rasche *et al.* Spatial genomic heterogeneity in multiple myeloma revealed by multi-region sequencing. *Nat Commun* **8,** 268 (2017).
47. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–8 (2013).
48. 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,** 377–81 (2009).
49. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169,** 1276–1290.e17 (2017).
50. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 184–97 (2015).
51. Giladi, A. *et al.* Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat Cell Biol* **20,** 836–846 (2018).
52. 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).
53. 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,** 539–54 (2013).
54. Flores-Montero *et al.* Next Generation Flow for highly sensitive and standardized detection of minimal residual disease in multiple myeloma. *Leukemia* **31,** 2094 (2017).