Supplementary material

Mutation landscape of multiple myeloma measurable residual disease: identification of targets for precision medicine

Supplementary results

Copy number variation analysis and cellularity

The copy number analysis (CNA) yielded ambiguous results, due to the variation in the sequencing depth ratio between tumor and normal sample (Figure S19). Cellularity results showed median 0,69 (range 0,31 – 0,99). However, cellularity value is influenced by the CNA estimation, thus this measure can be also biased. The only exception in the dataset is the sample M-16-028 where WGA amplified normal sample was used for calculations and resulted in clear pattern of CNAs showing gains on chromosomes 1, 4, 9, 11 and loss of chromosome 13 . Cellularity of the tumor sample was 0,94.

Supplementary methods

Sampling

Samples of bone marrow (BM) and peripheral blood (PB) were taken from patients responding to the therapy and reaching complete remission, very good partial response or partial response after bortezomib based treatment (Table S1). MRD negative samples were not included in the analysis. Patients were followed-up for at least 24 months. The samples for this study were taken after treatment, in most of the cases after transplantation (Figure S1) in multiple myeloma centres in Ostrava, Brno, Pilsen, Hradec Králové, Olomouc (Czech Republic) and Bratislava (Slovakia). 22 samples were used for the final analysis. The study is in accordance with the current version of the Helsinki Declaration and was approved by institutional ethics boards. All the patients were informed about the research activities and signed an informed consent form. Patient's clinical data are available in supplementary table S2. Samples of peripheral blood were frozen and DNA was later isolated by isolator MagCore Automated Nucleic Acid Extractor Magnesia 16 (RBC bioscience) with MagCore 101 Genomic DNA Whole Blood Kit (RBC bioscience) and then used for library preparation as described below. Samples of bone marrow were processed fresh.

Flow cytometry assessment and fluorescence activated cell sorting (FACS)

The level of MRD depth was evaluated in BM by flow cytometry using 2nd generation of EuroFlow panel.[1] Then, samples of BM were processed by centrifugation with Ficoll or with red cell lysing by ammonium chloride to get mononuclear cells or bone marrow deprived of red blood cells. Next, A-PCs were isolated by fluorescence activated cell sorting according to pathological immunophenotype of A-PCs using the following markers: CD38, CD19, CD45, CD56, eventually CD117 (Figure S2) using FACS ARIA^{TM} III machine (BD Biosciences). Median of sorted cells was 2000 (Table S1).

Whole genome amplification, library preparation and sequencing

Sorted A-PCs were subjected to whole genome amplification (WGA) using Repli-g Single cell kit (Qiagen). Absence of contaminating human genomic DNA in negative control was checked by PCR of human fibrinogen gene. Amplified DNA was then purified by QIAquick PCR Purification Kit (Qiagen). One microgram of non-amplified DNA from PB and WGA DNA from A-PCs were used for library preparation with SureSelect V6 kit (Agilent Tech.) and sequenced on Illumina HiSeq4000 platform, 100 cycles in Macrogene Inc. company with target coverage 50x. The only exception was normal sample of the patient M-16-028, where WGA was applied due to the insufficient yield of non-amplified DNA.

Variant calling

The data analysis part was covered by snakemake pipeline[2]. The quality of fastq files was controlled by fastqc[3]. Paired-end reads were mapped to the hg19 reference genome, using BWA MEM[4] with default setting, duplicates were marked using samblaster[5] and reads in the bam file were sorted using sambamba[6] ***(bwa mem -t {threads} -M -R {params.read_group} {params.genome_index} {input.reads} | samblaster -M | sambamba view -S -f bam -l 0 /dev/stdin | sambamba sort -o {output} /dev/stdin).***

Variants were called on both PB and A-PCs using FreeBayes[7] with adjusted setting to capture also low frequency variants ***(freebayes -f data/external/genome/GRCh37/human_g1k_v37.fasta --strict-vcf --pooled-discrete --pooled-continuous --genotype-qualities --report-genotype-likelihood-max --allele-balance-priors-off --no-partial-observations --min-repeat-entropy 1 --min-alternate-fraction 0.05 --min-alternate-count 2 data/processed/MRD-16-028/MRD-16-028_NORMAL.bam data/processed/MRD-16-028/MRD-16-028_TUMOR.bam --region 1:0-100000).***

Somatic variants were then called using a modified version of vcfsamplediff tool. We further filtered the resulting data by bedtools[8] and vcflib[9] to get variants located in baits + 100 bp, in high confidence intervals of the reference sequence and with high somatic score ***(bedtools intersect -header -a data/raw/${SAMPLE_ID}.variants.all.vcf.gz -b data/external/S07604514_Padded_noChr.bed | bedtools intersect -header -v -a - -b data/external/AllRepeats_lt51bp_gt95identity_merged.bed | mmseqtk somatic -t ${SAMPLE_ID}_TUMOR -n ${SAMPLE_ID}_NORMAL | vcffilter -f "MMTKLOD > 3.5" > ${OUTPUT})***; with respect to expected artefacts introduced by WGA, we decided to use a more lenient approach compared to other standard pipelines, and to build a set of high-confidence mutations by applying ad-hoc hard filters. To do that, we first annotated somatic mutations using variant effect predictor[10], then converted this data to mutation annotation format (MAF) using vcf2maf[11] and finally applied the following filters created with pandas library to select variants:

1) not located in homopolymers longer than 5nt
2) not present in PB (0 alternative reads in PB)
3) not present with frequency above 1 % in the databases of human genome variation (gnomAD, ExAC)
4) not present in immunoglobulin genes
5) not identified as tolerated/benign simultaneously in SIFT and PolyPhen databases
6) having Moderate/High impact as predicted by default maftools filter based on http://asia.ensembl.org/Help/Glossary?id=535.
7) Mutations with evidence of expression in our MM cohort (Table S2) were chosen for presentation in the main text.

The basic data visualisation was done by maftools[12], lollipops[13] and lifelines[14].

Copy number variation analysis and cellularity

Copy number variation and cellularity was estimated with Sequenza package[15] by running following sequenza-utils commands:

sequenza-utils gc_wiggle -w 50 –fasta human_g1k_v37.fasta -o human_g1k_v37.gc50Base.wig.gz

sequenza-utils bam2seqz -n NORMAL.bam -t TUMOR.bam --fasta .human_g1k_v37.fasta -gc human_g1k_v37.gc50Base.wig.gz -o out.seqz.gz

sequenza-utils seqz_binning --seqz out.seqz.gz -w 50 -o binned.seqz.gz and Sequenza R package with default settings.

Pathway analysis

The pathway analysis was done with the Fisher's Exact test for the merged list of all mutated genes non-filtered for frequency below 1% in general population and separately for sets of genes from individual MM MRD patients to find out whether some of the pathways were extensively enriched in mutations and to depict potential patterns of mutated genes which were otherwise only rarely shared within the patients cohort. We used 7 gene set collections from MSigDB[16] (gene ontology (GO) biological processes (4436 gene sets), GO cell components (580 gene sets), GO molecular function (901 gene sets), KEGG (186 gene sets), Reactome 1200 gene sets),

Oncogenic signatures (189 gene sets)) and GO slim dataset (135 gene sets) that was generated using the map2slim utility of the OWL tools. Results were corrected for multiple-hypothesis testing using the Benjamini-Hochberg procedure and significance threshold was set to FDR < 0.05 (Table S4).

Survival analysis

Progression free survival (PFS) for individual mutated genes and genesets with at least 1 mutated gene resulting from the previously mentioned pathway analysis were tested by logrank test using lifelines library[14]. Resulting p-values were corrected with benjamini hochberg correction.

Annotation with pharmacological information

OncoKB[17] database and Drug genome interaction database[18] were used to get a list of drugs associated with genes.

Drivers and MM associated genes

Because of small cohort size, we called drivers those genes identified as important in multiple myeloma in previous publications (Bolli et al., 2014; Kortüm et al., 2016; Lohr et al., 2014; Walker et al., 2018 [19–22]), table S3. The frequency of mutations in genes identified in this study was also compared with CoMMpass data available via GDC portal.[23]

Evidence of expression

The mutation data were complemented with the expression value from our unpublished 10 MM patient dataset, not overlapping with patients and samples used in this study (Table S2) and only genes with average baseMean expression > 8 were chosen for presentation in the final results.

Data availability

Data are stored in EGA under number EGAS00001004855.

Supplementary figures

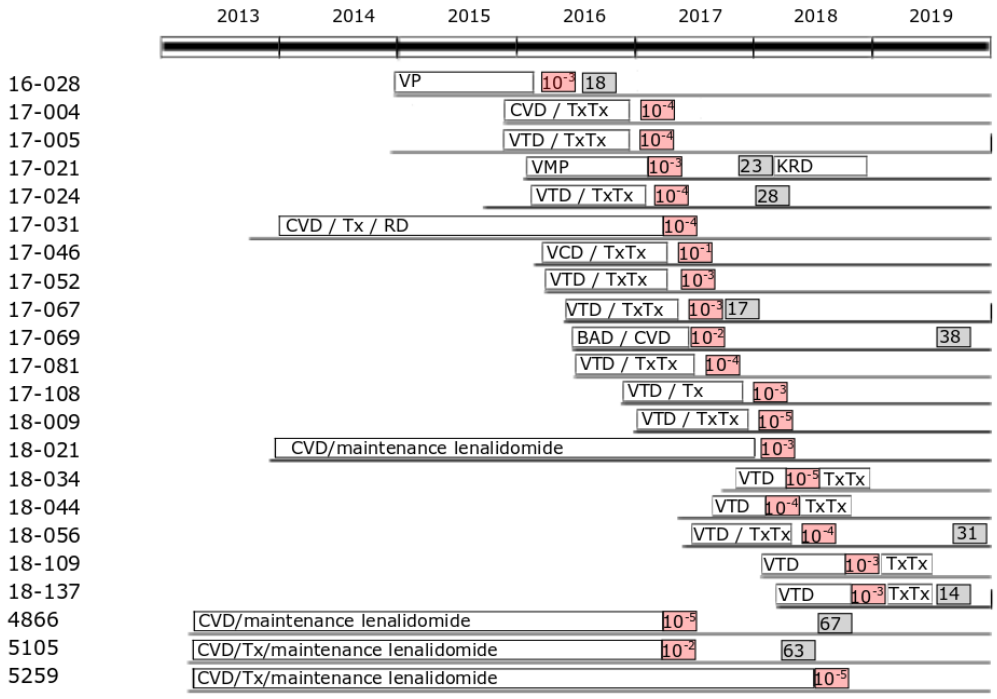


Figure S1: Timeline of patient's treatment. Time of sampling is represented by red ractangle with number showing reached MRD level. Left panel – patien't identification numbers. Treatment regime and its duration is represented by white ractangles (VP – Bortezomib, Prednisone; CVD – Cyclophosphamide, Bortezomib, Dexamethasone; VTD – Bortezomib, Thalidomide, Dexamethasone; VMP – Bortezomib, Melphalan, Prednisone; BAD – Bortezomib, Doxorubicin, Dexamethasone; KRD – Carfilzomib, Lenalidomide, Dexamethasone; Tx – autologous stem cell transplantation; TxTx – tandem autologous stem cell transplantation). Numbers in grey rectangles indicate disease progression and time to progression in months.

A B

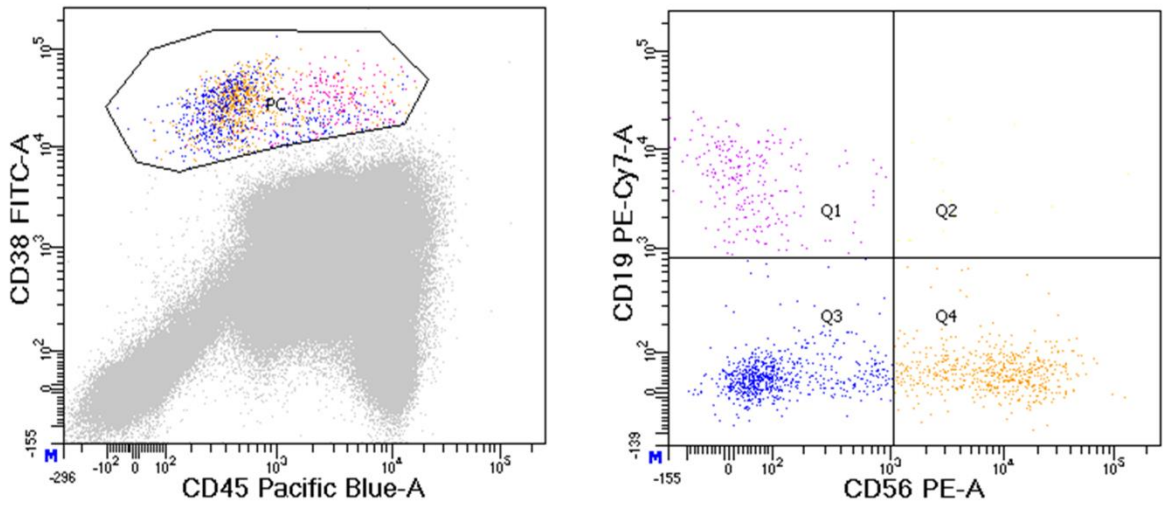


Figure S2: Sorting strategy for A-PCs. (A) PC – identification of PCs in BM; (B) Q1 – normal plasma cells, Q3 and Q4 potential aberrant plasma cells. Q3 or Q4 was sorted after confirmation of aberrant immunophenotype by next generation flow cytometry lab. If needed, CD117 was involved in the analysis as well.

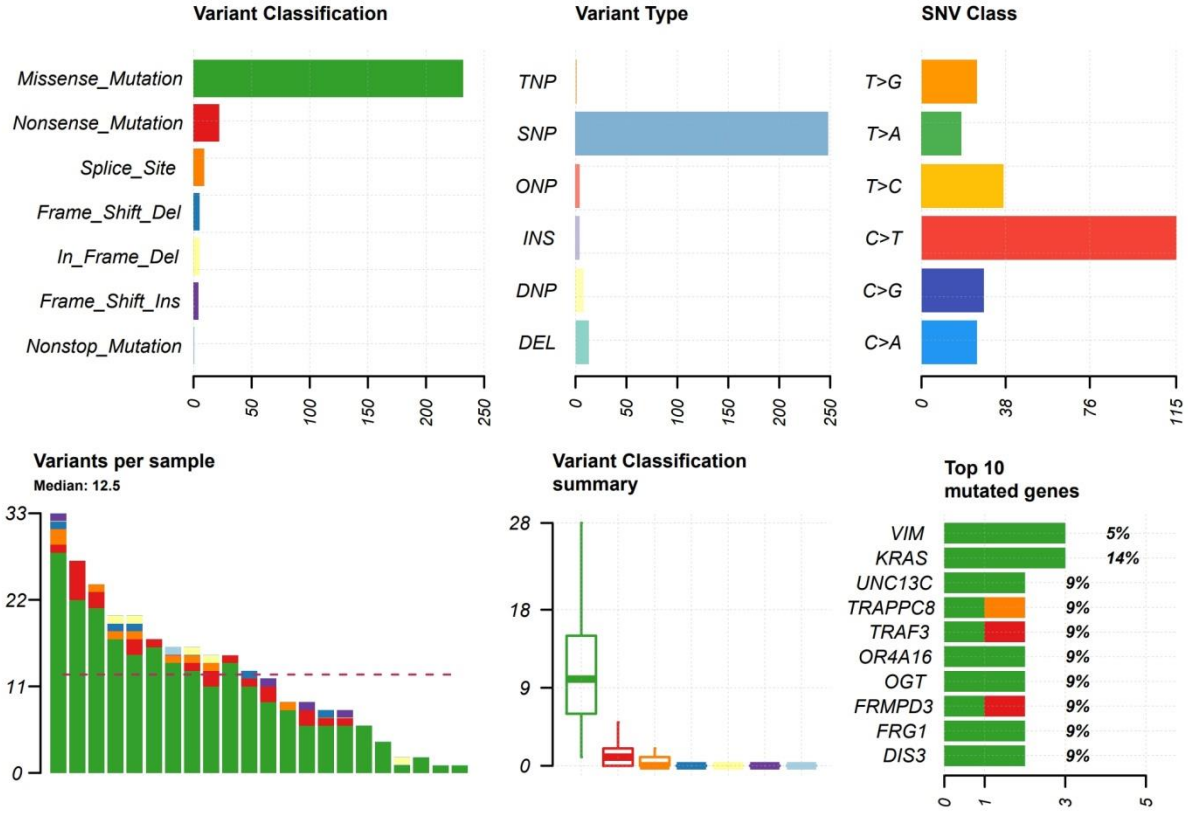


Figure S3: Variant summary. Graphs colored according to Variant Classification. TNP – triple nucleotide polymorphisms, SNP – single nucleotide polymorphism, ONP – Oligo-nucleotide polymorphism, INS – insertion, DNP – Double nucleotide polymorphism, DEL – deletion.

PATHWAY ANALYSIS RESULTS – PATHWAYS WITH AT LEAST 1 MUTATED GENE

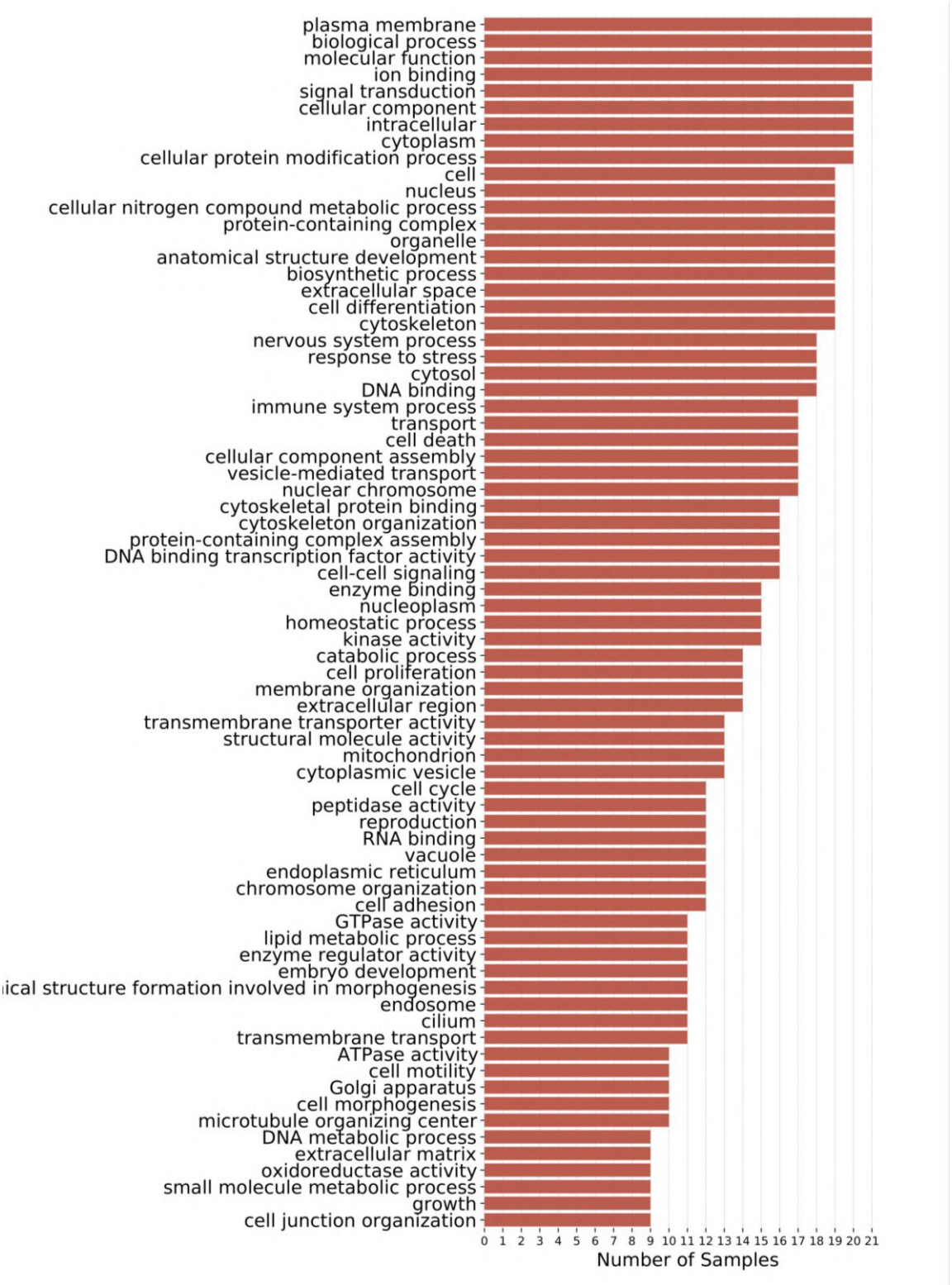


Figure S4: Involvement of pathways form go slim dataset, involved in at least 9 patients (41%).

Figure S5: Pathways involved in at least 7 samples in go slim dataset, exhibited for particular patients.

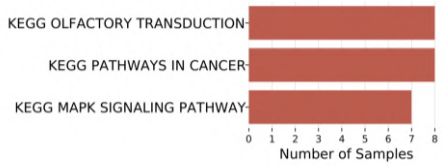


Figure S6: Pathways involved in at least 7 samples in KEGG dataset.

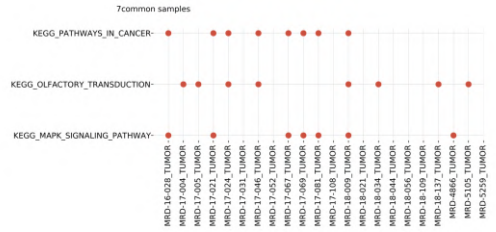


Figure S7: Pathways involved in at least 7 samples in KEGG dataset, exhibited for particular patients.

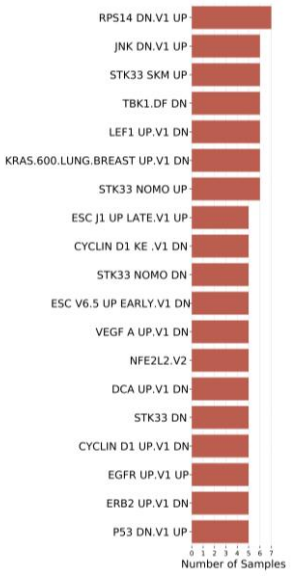


Figure S8: Pathways involved in at least 5 samples in oncogenic signatures dataset.

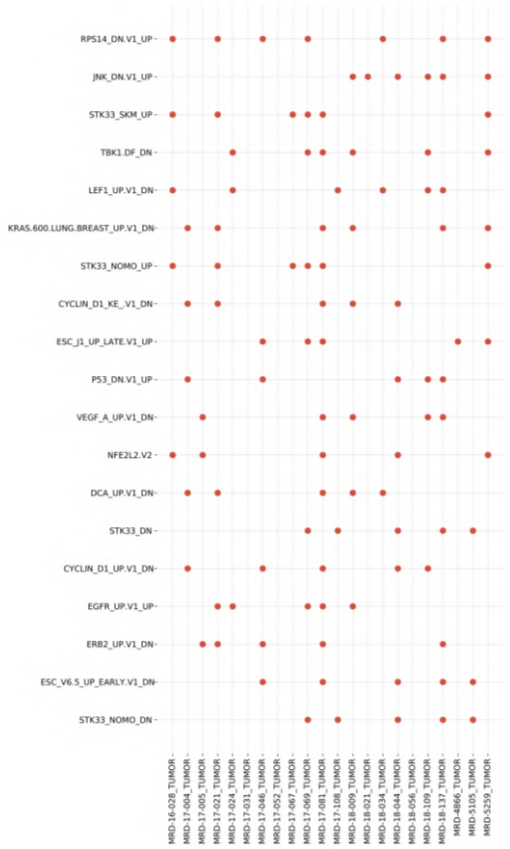


Figure S9: Pathways involved in at least 5 samples in oncogenic signatures dataset, exhibited for particular patients.

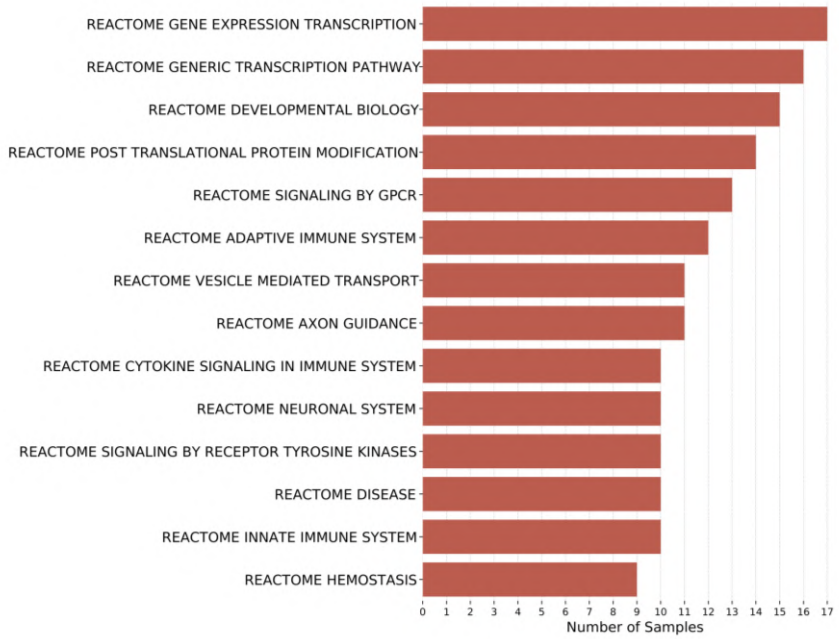


Figure S10: Pathways occurred in at least 9 samples in Reactome dataset.

Figure S11: Pathways involved in at least 9 samples in reactome dataset, exhibited for particular patients.

Association of shared pathways on survival

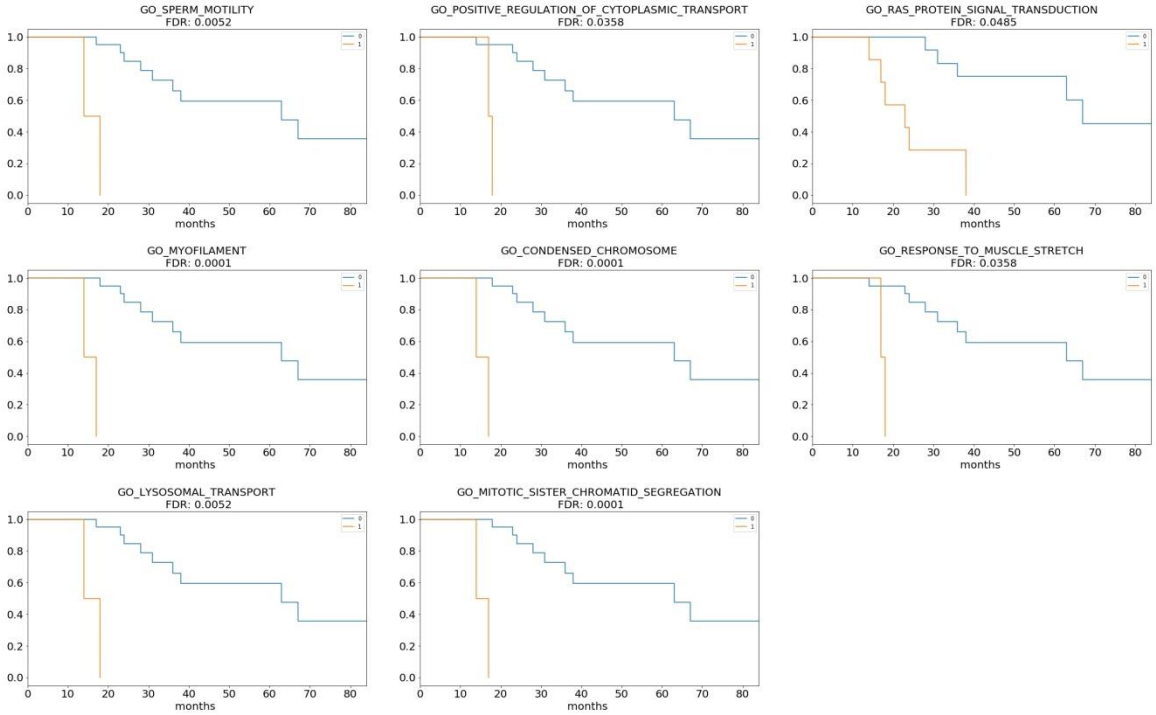


Figure S12: Kaplan-Meier curves showing impact of at least one mutated gene in pathways from GO all dataset on survival. 0 – no gene with mutation, 1 – at least one gene with mutation, name of pathway and resulting false discovery rate (FDR) of log-rank test, adjusted by Benjamini-Hochberg correction.

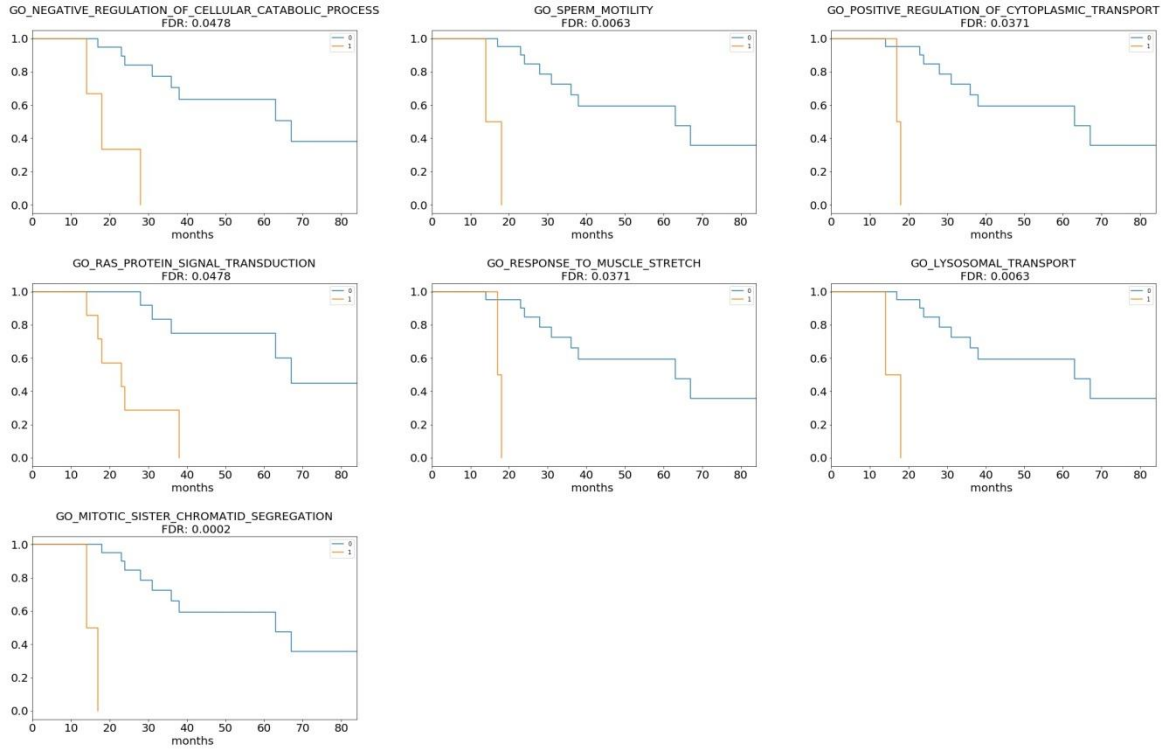


Figure S13: Kaplan-Meier curves showing impact of at least one mutated gene from pathways from GO biological processes dataset at survival. 0 – no gene with mutation, 1 – at least one gene with mutation. FDR – false discovery rate, adjusted by Benjamini-Hochberg correction.

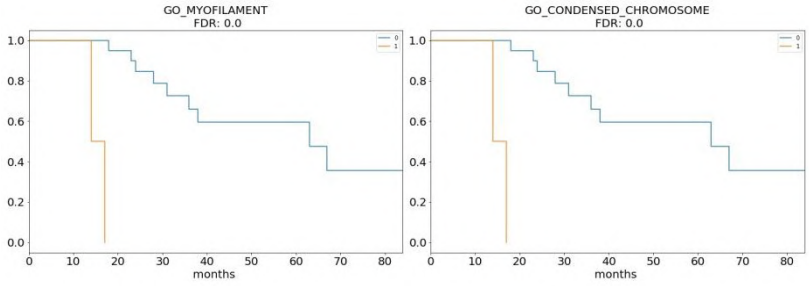


Figure S14: Kaplan-Meier curves showing impact of at least one mutated gene in pathways from GO cellular components dataset at survival. 0 – no gene with mutation, 1 – at least one gene with mutation. FDR – false discovery rate, adjusted by Benjamini-Hochberg correction.

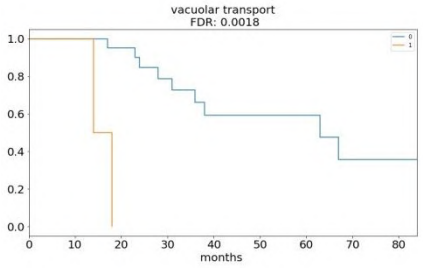


Figure S15: Kaplan-Meier curves showing impact of at least one mutated gene in pathways from GO slim dataset at survival. 0 – no gene with mutation, 1 – at least one gene with mutation. FDR – false discovery rate, adjusted by Benjamini-Hochberg correction.

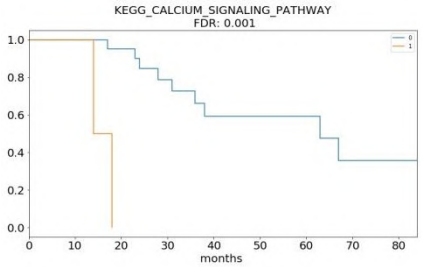


Figure S16: Kaplan-Meier curves showing impact of at least one mutated gene in pathways from KEGG dataset at survival. 0 – no gene with mutation, 1 – at least one gene with mutation. FDR – false discovery rate, adjusted by Benjamini-Hochberg correction.

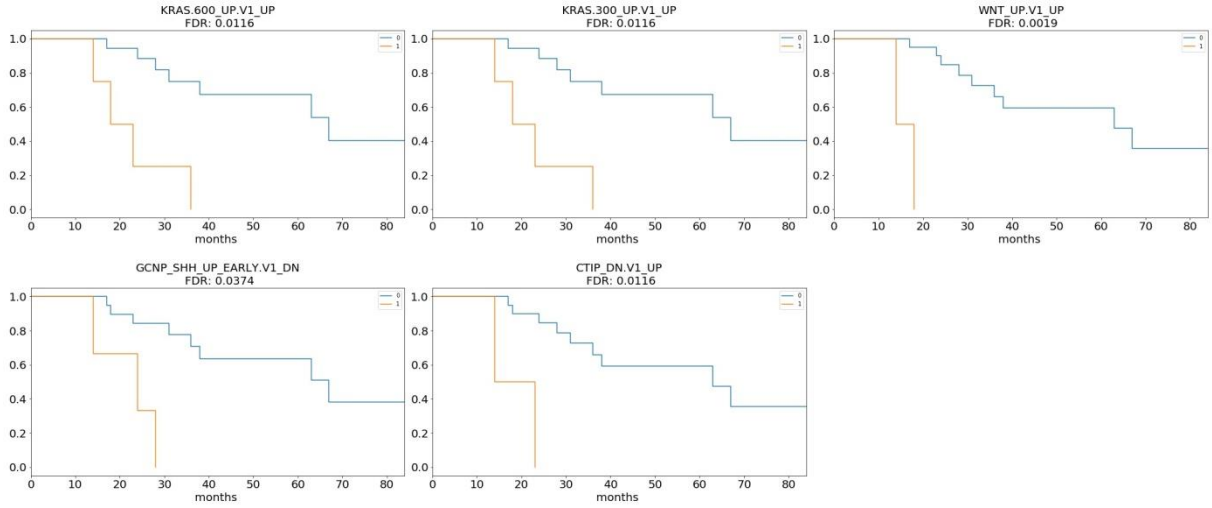


Figure S17: Kaplan-Meier curves showing impact of at least one mutated gene in pathways from oncogenic signalization dataset at survival. 0 – no gene with mutation, 1 – at least one gene with mutation. FDR – false discovery rate, adjusted by Benjamini-Hochberg correction.

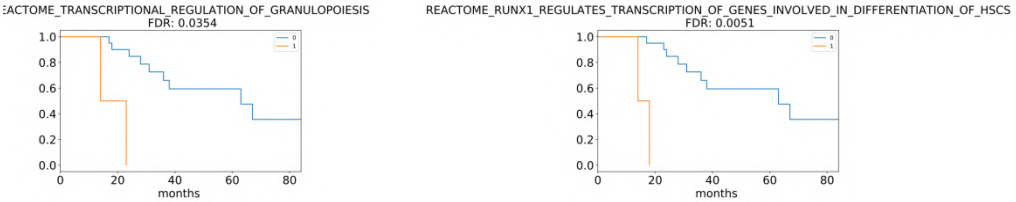
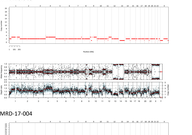
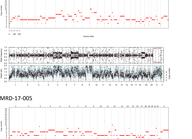


Figure S18: Kaplan-Meier curves showing impact of at least one mutated gene in pathways from reactome dataset at survival. 0 – no gene with mutation, 1 – at least one gene with mutation. FDR – false discovery rate, adjusted by Benjamini-Hochberg correction.

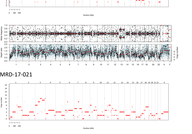
MRD-16-028



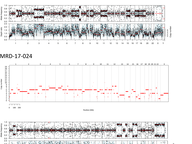
MRD-17-004



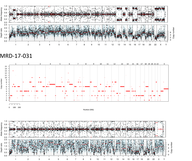
MRD-17-005



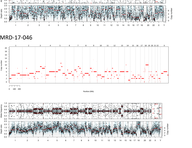
MRD-17-021



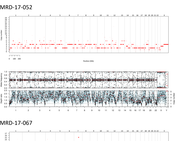
MRD-17-024



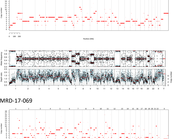
MRD-17-031



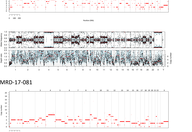
MRD-17-046



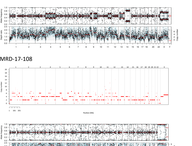
MRD-17-052



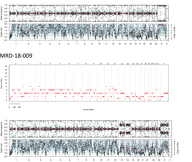
MRD-17-067



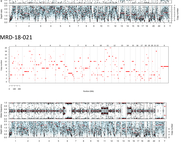
MRD-17-069



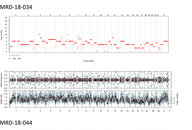
MRD-17-081



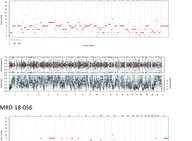
MRD-17-108



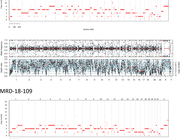
MRD-18-009



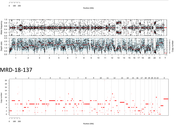
MRD-18-021



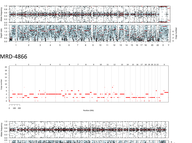
MRD-18-034



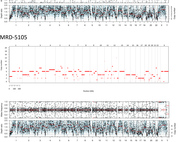
MRD-18-044



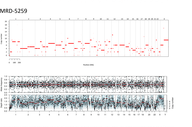
MRD-18-056



MRD-18-109



MRD-18-137



MRD-4866

MRD-5105

MRD-5259

Figure S15: Copy number analysis of all samples in the dataset. Red bars indicate chromosome gains/losses, blue mark allele frequency distribution and depth ratio distribution between normal and tumor sample.

Supplementary tables in excel file

S1: Clinical data and genes. This table shows patients' clinical data and presence of mutations in genes in particular patients.

S2: All variants. The MAF file showing all variants after filtering mentioned in supplementary methods.

S3: MM associated genes. Overlap of our dataset with genes previously identified as relevant for MM.

S4: All pathways. All pathways with at least 1 mutated gene, all 9 datasets merged.

S5: Significant pathways. All significantly mutated pathways. Tested by Fisher test, corrected by Benjamini-Hochberg procedure as mentioned in supplementary methods.

S6: MM associated pathways. Details for mutated positions in genes involved in pathways previously associated with multiple myeloma.

S7: Genes survival. Association of mutated genes and survival according to log-rank test. Both normal and Benjamini-Hochberg adjusted p-values shown.

S8: Pathways survival. Association of shared pathways with at least 1 mutated gene with progression free survival according to log-rank test. P-value and Benjamini-Hochberg adjusted p-value shown.

S9: DGIdb. Intersection of our results with The Drug-Gene Interaction Database.

S10: Drugability. Three tables showing intersection of our results with OncoKB database, TARGET v3 database and Intersection of our genes marked MM associated with records in DGIdb database.

S11: Drugable genes: Genes with known drug associations and expression evidence selected for further preclinical investigation.

References:

1. Flores-Montero, J. *et al.* Next Generation Flow for highly sensitive and standardized detection of minimal residual disease in multiple myeloma. *Leukemia* (2017) doi:10.1038/leu.2017.29.
2. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
3. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2015), 'FastQC,' https://qubeshub.org/resources/fastqc.
4. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
5. Faust, G. G. & Hall, I. M. SAMBLASTER: Fast duplicate marking and structural variant read extraction. in *Bioinformatics* vol. 30 2503–2505 (Oxford University Press, 2014).
6. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
7. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).
8. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
9. Garrison, E. Vcflib, a simple C++ library for parsing and manipulating VCF files. 2016. https://github.com/vcflib/vcflib.
10. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).
11. Kandoth, C. mskcc/vcf2maf: vcf2maf v1.6.18. [online] https://github.com/mskcc/vcf2maf doi:10.5281/zenodo.593251. (2020).
12. Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Research* **28**, 1747–1756 (2018).

13. Jay, J. J. & Brouwer, C. Lollipops in the Clinic: Information Dense Mutation Plots for Precision Medicine. *PLOS ONE* **11**, e0160519 (2016).
14. Davidson-Pilon, C. *et al.* CamDavidsonPilon/lifelines: v0.25.4 [online] https://github.com/CamDavidsonPilon/lifelines. (2020) doi:10.5281/ZENODO.4002777.
15. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology* **26**, 64–70 (2015).
16. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550 (2005).
17. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology* 1–16 (2017) doi:10.1200/po.17.00011.
18. Cotto, K. C. *et al.* DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Research* **46**, D1068–D1073 (2018).
19. Walker, B. A. *et al.* Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma. *Blood* **132**, 587–597 (2018).
20. Kortüm, K. M. *et al.* Targeted sequencing of refractory myeloma reveals a high incidence of mutations in CRBN and Ras pathway genes. *Blood* **128**, 1226–1234 (2016).
21. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature Communications* **5**, 2187–2198 (2014).
22. Lohr, J. G. *et al.* Widespread genetic heterogeneity in multiple myeloma: Implications for targeted therapy. *Cancer Cell* **25**, 91–101 (2014).
23. Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine* **375**, 1109–1112 (2016).