Unique molecular assay (UMA): a next-generation sequencing targeted panel for efficient and comprehensive genomic profiling and risk stratification of multiple myeloma

Andrea Poletti,[1,2*] Barbara Taurisano,[1,2*] Gaia Mazzocchetti,[3,4] Marta Lionetti,[5] Marina Martello,[1,2] Viola Meixian Vuong,[1,2] Vincenza Solli,[1,2] Giulia Marzocchi,[1,2] Akihiro Maeda,[6] Ilaria Vigliotta,[1] Enrica Borsi,[1] Silvia Armuzzi,[1,2] Ignazia Pistis,[1] Alessio Marella,[5] Sonia Fabris,[6] Paola Tacchetti,[1] Katia Mancuso,[1,2] Ilaria Rizzello,[1,2] Lucia Pantani,[1] Nicoletta Testoni,[1,2] Michele Cavo,[2] Elena Zamagni,[1,2] Niccolò Bolli[5,6] and Carolina Terragna[1]

[1]IRCCS Azienda Ospedaliero-Universitaria di Bologna, Istituto di Ematologia "Seràgnoli", Bologna; [2]Dipartimento di Scienze Mediche e Chirurgiche, Università di Bologna, Bologna; [3]FABIT-Department of Pharmacy and Biotechnology, University of Bologna, Bologna; [4]Computational and Chemical Biology, Italian Institute of Technology (IIT), CMP3VdA, Aosta; [5]Department of Oncology and Hemato-Oncology, University of Milan, Milan and [6]Hematology Unit, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

*AP and BT contributed equally as first authors..

Correspondence: N. Bolli
niccolo.bolli@unimi.it

C. Terragna
carolina.terragna@unibo.it

Received: February 16, 2025.
Accepted: May 5, 2025.
Early view: May 15, 2025.

https://doi.org/10.3324/haematol.2025.287559

Supplementary information

Unique Molecular Assay (UMA): an NGS targeted panel for efficient and comprehensive genomic profiling and risk stratification of Multiple Myeloma.

Summary

1. Sample processing

For each patient, the CD138-positive cells fraction was enriched from BM samples by magnetic bead sorting by AutoMACS® Pro II Separator (Miltenyi Biotec, Bergisch Gladbach, Germany) and was employed both for FISH and molecular testing. The purity of enriched CD138+ cells was performed via flow cytometry using a CD138/38 combination by FACSCanto™ II (BD Biosciences, San Jose, CA, USA), only samples with sufficiently high purity were included in the study. DNA was isolated by Maxwell® DNA extraction kits (Promega Italia Srl, Milan, Italy) and used for SNPs array and NGS panel. In this study all samples achieved a sufficient purity (purity > 50%).

2. NGS library preparation

2.1. UMA panel library preparation

For UMA panel samples, the NGS library preparation requires 100 ng of gDNA input. The Illumina DNA Prep with Enrichment protocol (Illumina Inc, San Diego, CA, USA) was employed, consisting of a first step of DNA tagmentation and a second step of DNA amplification. The amplified libraries were then pooled for hybridization with biotinylated UMA probes (Agilent Technologies, Santa Clara, CA, USA) at 58 degrees for 17 hours and enriched with streptavidin magnetic beads. To check the enriched pooled libraries' quality and quantity, two quality control steps were performed, using Tapestation (Agilent Technologies, Santa Clara, CA, USA) and the Qubit (Thermo Fisher Scientific, Waltham, MA, USA). Enriched libraries were sequenced on MiSeq® System (Illumina Inc, San Diego, CA, USA) platform using either MiSeq® Reagent Kit v2 300 or v3 600, to generate 250 bp paired-end reads.

2.2. Milan panel library preparation

For Milan NGS custom panel samples, NGS libraries were prepared starting from 100 ng of DNA following the KAPA HyperCap Workflow v3.0 protocol (Roche, Basel, CH) by enzymatic fragmentation. For ULP-WGS, to obtain an average genome-wide fold coverage of 0.1X, up to 24 libraries were pooled and sequenced using MiSeq® Reagent Kit v3 and setting 200 cycles for both forward and reverse strands on Miseq® System platform. For targeted NGS, a DNA sequencing of coding regions of 56 MM-driver genes was performed by capture. Pre-capture libraries obtained as described above were combined in an equimolar manner in pools of 16 samples each, which then underwent a "capture" process using biotinylated probes complementary to the genes of interest and synthesized by Roche HyperDesign solution. Target sequences were enriched *via* beads bound to

streptavidin molecules, and then amplified, purified, quantified, and subjected to capillary electrophoresis to assess their size. Finally, each library was sequenced using a MiSeq Reagent Kit v3 and setting 200 cycles for both the forward and reverse strands on a MiSeq® System (Illumina, San Diego, USA).

3. Definition of run parameters, according to the intra-lab validation

We first aimed at defining the best run parameters for UMA panel sequencing runs. We explored a multitude of technical variables, including the number of samples *per* run, the type of flowcell used and the hybridization temperature (hyb-temp), in order to optimize key performance quality metrics such as high on-target coverage, a well-balanced off-target reads percentage (OT%), i.e. between 50% and 70%, and a low Mean Absolute Deviation (MAD) of the CN signal (main noise metric for CN analysis).

Overall, a discernible pattern in sequencing efficiency was highlighted. High on-target coverage is specifically critical for clonal and sub-clonal mutations calls and all technical configurations (number of samples and flowcell type) explored provided both a coverage (>100X) and a number of total reads (>4M) sufficient to detect mutations with VAF ≥ 5%, considering a minimum cut-off of 5 reads to support a mutation call (**Supp. Figure S2a**). Notably, the configurations with 4 samples on a V2 300 (Miseq platform) and that with 10 samples on a V3 600 (Miseq platform) showed lower on-target coverage than others, due to their higher OT% (74% and 72%, respectively) (**Supp. Figure S2a**).

On the other hand, since off-target reads are critical for obtaining good quality broad genome-wide CN profiles, we sought to minimize the MAD of CN profiles derived by off-target reads. Thus, several hyb-temp were tested, highlighting a strong linear relationship between MAD and OT%, whose intercept varied according to the different hyb-temp (**Supp. Figure S2b**). Among all tested temperatures, hyb-temp=58 °C was observed being the optimal one to achieve a low MAD while maintaining a balanced OT%. Finally, we performed a multiple linear regression model to investigate the relationship between MAD and all available sequencing variables (complete list of variables in “Multiple linear regression model on MAD quality” chapter). Following a backward elimination process, four significant ($p<0.05$) variables were selected in the final model (**Supp. Figure S2c**), i.e. increased number of total reads (Mil), OT% and hyb-temp, all associated to a reduction in the MAD value (-0.004, -0.003 and -0.019 MAD decrease *per* variable unit increase). Therefore, run parameters configurations that maximize these values are ideal for the best use of UMA panel. On the contrary, an increase in GC-content % was associated to a MAD increase

(0.004 MAD increase *per* variable unit increase), supporting that sub-optimal hybridization procedure is detrimental for the quality of CN signal.

Taken together, these results suggest that, by using a common bench-top platform (such as Illumina-MiSeq), the optimal configuration would be either 8 samples in a V2 300 flowcell, or 10 samples in a V3 600 flowcell, with 58°C hyb-temp for 17 hours.

2. UMA panel's bioinformatic pipeline development for alteration calls

Three separate bioinformatic calling strategies, illustrated in **Supp. Figure S3, S4, S5**, were developed for each genomic alteration class in MM, i.e. CNAs, t-IgH, and SNVs, and harmonically combined in a single pipeline using Snakemake (v7.20.0) and an interconnected R scripts workflow to ensure scalability and reproducibility of the global analysis. Each calling strategy was tailored to exploit and integrate selected individual tools strengths, aiming to enhance the overall analysis performance, improve alteration calling accuracy and obtain performances comparable to that of traditional GOLD-standard methods (e.g. FISH).

2.1. CNAs calling algorithm

We designed a novel custom algorithm to call CNAs alterations, capable of jointly analyze the CN information contained in all reads generated by targeted sequencing (both on-target and off-target reads). The information retrieved from reads that were either mapped on the panel capture regions (on-target reads) or mapped in other random parts of the genome (off-target reads) were processed in parallel by the algorithm (**Supp. Figure S3**), which uses two different pipelines: 1) "Whole-genome broad CNA pipeline" (steps 1-5 of the CNAs calling algorithm), and 2) "Focal gene-level CNA pipeline" (steps 6-11 of the CNA calling algorithm), represented by blue and red boxes in **Supp. Figure S3**, respectively. These two pipelines contain different and specific analytical strategies and bioinformatic tools appropriate for each read type. Finally, in the last step of the process (step 12), the algorithm entwines both off-target and on-target information to ultimately call CNA events. This computational strategy is of particular interest because the proportion of on-target and off-target reads in experiments is not always stable: even if it typically fluctuates around a ratio range of 0.4-0.6, it's influenced by multiple factors such as hybridization conditions, DNA quality and sequencing platform used[1]. Consequently, when considering cases in which this ratio deviates significantly from the median value, relying on a single read type to call CNAs could lead to substantial loss of information. Hence, we developed a "synergistic dual-read type" approach since the abundance of one read-type corresponds to the scarcity of the other. Consequently, by using this approach one

pipeline's signal precision obtained from a given read-type will be proportionally higher to the noise level of the other, compensating each other performance in precision.

In detail, we employed CNVkit v0.9.9[2] to extract CNAs from off-target reads and CopywriteR v2.6.1[3] to extract whole-genome broad CNAs calls (CN segments) and BOBaFIT v1.0[4] to refit the signal baseline-region. Instead, to extract CNAs from on-target reads we used coverage metrics of all targeted exons generated by HSmetrics GATK v4.3.0 [5] and multiple custom *ad-hoc* R scripts (available at https://github.com/andrea-poletti-unibo/paper_UMA_panel.git). These scripts were used to normalize the CN signal and calculate a log2 ratio based on a reference panel of 13 normal samples, by using a standard "coverage-depth" approach for CN signal computation[6]. Again, we applied the BOBaFIT correction to exclude the baseline-region bias from the CN data (**Supp. Figure S6**).

2.1.1. Dynamic CNAs calls thresholds

Finally, to call CNA events from CN data, we choose to set sample-specific dynamic CN signal cutoffs, based on samples' specific sequencing quality category (defined by the number of off-target and on-target reads obtained per sample By defining these dynamic cutoffs, we enhance UMA panel sensitivity and specificity to accurately distinguish between real CNA events (true positives) and possible artifacts affecting the CN signal (false positives) that may be generated by sequencing noises derived from various sources (either technical and/or biological)[7].

For focal CNAs from on-target reads, we defined the empirical calling thresholds as follows:

1. observing the distribution of **on-target reads** in the entire cohort.
2. assigning a **focal quality** to each sample based on its specific number of on-target reads.
3. assigning a CNA **clonality threshold** for each sample based on its focal quality.

According to the number of reads (n), we identified 3 quality's categories:

- n <= 1 M: *low quality*
- 1 M < n < 2.25 M: *medium quality*
- n >= 2.25 M: *high quality*

In addition, we assigned a specific clonality threshold for CNA assessment to each previously identified category. Specifically: 40% for on-target low quality samples, 30% for medium quality samples, and 20% for high quality samples.

Within the entire cohort, we identified 22/129 (17.1%) low quality samples, 83/120 (64.3%) medium quality samples and 24/129 (18.6%) high quality samples (**Supp. Figure S7a,b,e**).

On-target reads (millions)	Focal quality category	Clonality threshold	n (%)
n <= 1	Low	40%	22 (17,1%)
1 < n < 2.25	Medium	30%	83 (64,3%)
n >= 2.25	High	20%	24 (18,6%)

Focal clonality threshold levels based on sample on-target quality.

Similarly, for broad CNA calls from the off-target reads, the dynamic call thresholds were empirically defined as follows:

1. observing the distribution of **off-target reads** for the whole cohort.
2. assigning a **broad quality** to each sample based on its specific number of off-target reads.
3. assigning a CNA **clonality threshold** for each sample based on its broad quality category.

We identified the following 3 categories depending on the number of off-target reads (n):

- n <= 2 M: *low quality*
- 2 M < n < 3 M: *medium quality*
- n >= 3 M: *high quality*

For each category, a level of clonality was defined accordingly: 40% for low quality samples, 30% for medium quality samples, and 25% for high quality samples.

In total, in our case series: 35/129 (39.5%) are of low quality, 43/129 (33.3%) are of medium quality, 51/129 (27.1%) are of high quality. Consequently, we associated a level of clonality to each category: 40% for low quality samples, 30% for medium quality samples and 25% for high quality samples (**Supp. Figure S7c,d,e**).

Off-target reads (millions)	Broad quality category	Clonality threshold	n (%)
n <= 2	Low	40%	51 (39,5%)
2 < n < 3	Medium	30%	43 (33,3%)
n >= 3	High	25%	35 (27,1%)

Broad clonality threshold levels based on sample off-target quality.

2.2. t-IgH calling algorithm

We developed a novel strategy to accurately identify canonical translocations in IgH regions, a critical component of MM genomics. This approach is focused on leveraging the individual strengths of two renowned structural variant callers, Manta v1.6.0[8] and DELLY v0.8.5[9], to maximize the detection accuracy. Since both Manta and DELLY are comprehensive tools able to exploit both split-reads and paired-read mapping information to detect structural variants, our approach is designed to capitalize on the unique algorithmic capabilities of each tool, ensuring a comprehensive sweep of potential translocations, where translocation events identified by either Manta or DELLY are considered valid. To maintain high-quality data, we have implemented stringent quality filters, focusing on parameters such as read depth, breakpoint junction clarity, and alignment accuracy (step 3 of **Supp. Figure S4**). These filters are critical in differentiating true translocation events from sequencing artifacts or incidental findings, given the increased likelihood of false positives in a single-tool approach. By merging the output of both tools (step 4 of **Supp. Figure S4**), we aim to cover a broader spectrum of translocation events, mitigating the limitations inherent in relying on a single tool.

2.3. Mutation calling algorithm

A rigorous approach to identify somatic mutations that minimizes false discoveries was implemented (**Supp. Figure S5**). This approach consists in a primary variant calling process (steps 1-3) and a secondary call-validation process (steps 4-7). The dual-process strategy operates on a stringent criterion: a somatic mutation is accepted only if it is independently identified by Mutect2 (GATK v4.3.0)[10] and subsequently corroborated by either Freebayes v1.3.5[11] or VarScan v2.3.9[12]. This intersectional approach leverages on Mutect2's sensitivity and on the confirmatory power of a secondary tool, thereby refining the accuracy of mutation detection. In details, the primary process is centered around Mutect2, a variant calling tool part of the Genome Analysis Toolkit (GATK)[5], which was employed as the primary variant caller due to its demonstrated excellent performance in somatic mutation detection[13–15], especially in tumor-only scenarios[10,16]. Additionally, the rationale for selecting Mutect2 as primary variant calling tool stems from its widespread acceptance in genomic studies, the abundance and customizability of filtering options conferred by its companion tool “FilterMutectCalls” (used for excluding multiple types of variants caused by biological or technical artifacts)[17], continuous updates from the GATK team, and its documented efficacy in peer-reviewed literature, particularly in complex cancer genomes[18–20]. However, as any single-tool approach can potentially lead to the identification of false positives, a secondary validation process using either Freebayes or VarScan2 (tools able to process tumor-only NGS data) is incorporated into the calling strategy. This step is crucial for enhancing the specificity of mutation calls and

avoiding false discoveries, which would be particularly problematic in a clinical setting application. Future directions in this research include the integration of additional bioinformatic tools (e.g. Strelka[21]) to further enhance the sensitivity of the mutation calling process.

2.4. VariantThinker: an R function for mutation pathogenicity assessment

We designed an algorithm named "VariantThinker" to identify and filter out pathogenic somatic variants from variants of uncertain significance (VUS), germline, and benign mutations. This algorithm can exploit the full list of annotations obtained by Annovar[22] to each specific mutation 1) a pathogenic significance (categorized from "A" = pathogenic, to "D" = benign) and a confidence score (quantified from 1= "very confident" to 5 = "low confidence").

We implemented VariantThinker as an R function (available at https://github.com/andrea-poletti-unibo/paper_UMA_panel/blob/main/scripts/4_Mutations_algorithm/2_VariantThinker_annotation.R) that integrates various in-silico prediction tools information (i.e. SIFT, MutationTaster, fathmm-MKL, Polyphen2, LRT, MutationAssessor, MetaLR, MetaSVM, VEST4, PROVEAN, etc. – derived from database dbNSFP v4.1a) along with clinical/population databases information (gnomAD v2.0.1, COSIMC v70, ClinVar version 2021-01-23, InterVar version 2018-01-18) to assess the potential impact of each variant. The function accepts a dataframe containing variant information as input and returns the same dataframe augmented with additional columns that provide insights into variant categorization and prediction confidence information. For each variant in the input dataframe, the function performs a series of evaluations based on the available annotations and predictions. It calculates the proportion of pathogenic predictions from selected tools, assesses the functional impact of the variant (e.g., loss of function, splicing alterations), evaluates clinical significance based on ClinVar annotations (e.g., pathogenic, benign, uncertain significance), and considers allele frequency information from gnomAD to differentiate common polymorphisms from potentially pathogenic variants.

The outcome of these evaluations is summarized in newly added columns to the input dataframe: `VarianThinker_Category`, `VT_notes`, `VT_confidence`, `pred_pathogenic`, `pred_tot_calls`, and `pred_ratio`. These columns provide a categorized assessment of each variant (`VarianThinker_Category`), explanatory notes (`VT_notes`), a confidence level for the categorization (`VT_confidence`), the number of pathogenic predictions (`pred_pathogenic`), the total number of prediction calls made (`pred_tot_calls`), and the ratio of pathogenic predictions to total predictions (`pred_ratio`).

3. Milan panel bioinformatic pipeline

Large-scale CNAs were estimated from ULP-WGS data based on the depth of coverage by using ichorCNA v0.2.0 software with a panel of 34 normal samples. The detection of somatic mutations from targeted NGS data was performed using the bioinformatic pipeline here described. Pre-processing step relied on FastQC for quality control, bbduk for trimming of demultiplexed fastq files, Bwa mem for alignment, GATK/Samtools for postprocess BAM (including mark duplicates, base quality recalibration and sorting) and GATK/PICARD to set the metrics for understanding the depth of coverage, removing PCR duplicates and off-target/unmapped bases.

Mutect2 GATK 4.3.0.0, LoFreq (v. 2.1.5), VarDict (2019.06.04), Freebayes (v. 1.3.6) and Strelka (v. 2.9.2) were used for variant calling, and only variants called by at least 2/5 of these variant callers were further considered. To filter out artifacts and germline polymorphisms, we first excluded variants with at least one of the following characteristics: (1) variant allele frequency (VAF) lower than 0.02 and less than 15 supporting reads; (2) coverage <300x; (3) population allele frequency >0.001 based on the information retrieved from the public database gnomAD (http://gnomad.broadinstitute.org; gnomAD r3.1.2).

4. Multiple linear regression model on MAD quality

In order to explore the effect of multiple sequencing and experimental variables on MAD quality of CN profiles, we first prepared a comprehensive dataset selecting relevant variables from three categories: wetlab metrics, multiQC metrics, and HSmetrics. The selected variables included measures of:

- library and pool concentrations,
- hybridization temperature,
- base pair length,
- cluster density,
- percentage of reads passing filters,
- Q30 percentage,
- average sequence length,
- GC content,
- mean target coverage,
- fold enrichment,

- number of total reads,
- percentage of off-target reads, near-target reads, on-target reads,
- GC dropout,
- AT dropout,
- percentage of unique reads,
- measures of excluded reads due to following reasons: duplicates, low mapping quality, overlap, low base quality, and adapter presence.

Next, a first linear regression model was fitted to the data selecting MAD as dependent variable and all other variables as predictors. Following the initial model fitting, a backward selection process was employed to identify the most significant predictors. The dataset was then further refined to include only variables that were the significant predictors from the initial model: percentage of off-target reads, total reads, hybridization temperature, and GC content. Finally, a second linear regression model was fitted to this reduced dataset and similarly tidied. This model provided a more effective explanation of the data, focusing on the most influential predictors of MAD (**Supp. Figure S2**c).

5. Replicates intra-run and inter-run

In the context of the intra-lab validation process, we performed DNA replicate sequencing both intra-run and inter-run to demonstrate the consistency and reliability of the UMA panel. By sequencing the same DNA samples multiple times within a single run (intra-run) (n=3 replicates in run 3) and across different sequencing runs (inter-run) (n=7 replicates in run 2 and 3), we aimed to assess the reproducibility of the panel's results. The following table describes the replicate samples per run in detail:

DNA	Run 1 samples	Run 2 samples	Run 3 samples
1036	*Run_1_1036*	*Run_2_1036*	
1273	*Run_1_1273*	*Run_2_1273*	
635	*Run_1_635*	*Run_2_635*	
628	*Run_1_628*	*Run_2_628*	*Run_3_628-rep1* *Run_3_628-rep2* *Run_3_628-rep3*

Intra-run and inter-run replicates samples

The subsequent analysis was aimed at evaluating the consistency of detected genomic alterations, including broad and focal CNAs, gene mutations, and t-IgH (**Supp. Figures S8, S9, S10 and S11**). The intra-run replicates were included to confirm the uniformity of results under identical experimental conditions, while the inter-run replicates allowed us to evaluate how changes in sequencing parameters or run conditions might impact the panel's performance. Overall, we observed a great level of reproducibility of genomic alterations calls, in detail:

- **CNAs:** the concordance percentages of broad CNAs were always >95% and focal CNAs Pearson's correlation coefficients (*R*) were always > 0.96 among all replicate pairs (**Supp. Figures S8 and S9**);
- **mutations:** point mutations and indels VAF Pearson's correlation coefficients (*R*) were always > 0.89 among all replicate pairs (**Supp. Figure S10**);
- **t-IgH:** FISH detected t-IgH were also detected in all (100%) replicates. Replicates pairs were fully concordant for t-IgH calls both for Delly and Manta calls (**Supp. Figure S11**).

Overall, the replicate sequencing within the intra-/inter-run validation process affirmed the UMA panel's consistent performance, reinforcing its robustness and reliability for clinical and research applications in MM.

6. BO-MI Validation analysis

6.1. CN analysis

Concordance analysis of replicate profiles consists of several steps, as described below.

6.1.1. Tools concordance analysis

First, we computed a unique CN value between the two CNA calling tools (CNVkit and CopywriteR) enabling a direct segment comparison between BO samples and external MI replicates. An initial concordance analysis was carried out between the two tools, extracting common segments between CNVkit and CopywriteR for each sample. Then, we calculated the difference between the two tools' CN values by segment: if this difference was greater than 0.20 CN units in a given segment, the tools were defined discordant for that segment, otherwise the tools were defined concordant. Next, a tool concordance value per sample was calculated as the ratio between concordant segments and all segments. Since the comparison showed a good concordance between the two methods (mean concordance: 0.88), we choose to employ a "conservative" approach to select a single segment CN value, by picking the CN value closest to the baseline region. This approach guarantees, in case of large discrepancies (difference greater than 0.20 CN units), a reduced risk of falsepositive calls. We also explored another approach that leverages on the two tools' CN average value: we observed that choosing one approach rather than the other does not significantly affect the total concordance of the replicates (data not shown), as the two tools often agree. We ultimately choose to use the conservative approach because, when computing CNA calls, the average approach may occasionally estimate a value very far from the truth due to its strong sensitivity to outliers.

6.1.2. BO-MI concordance analysis

Once we selected a unique CN value per segment, we estimated the concordance between the BO-MI samples pairs using the same conservative approach used in the tools' concordance analysis (see above) and the same difference threshold (0.20 CN units). Finally, a size-weighted average was calculated between the CN values of the segments belonging to each chromosome arm to compute a broad CN call per chromosome arm.

6.1.3. SNP array concordance

The SNP array concordance value was calculated for both BO samples and MI samples. Here we directly compared the broad CN calls of UMA and SNP array. Discordances were defined when the CN difference was greater than 0.40 CN units, to account for platform-specific biases.

6.2. Translocations of IGH region

In this analysis, Manta and DELLY calls were directly compared to FISH data. A positive call was defined when at least one of the two tools detected the alteration, as in the default pipeline.

6.3. Single Nucleotide Variant

For an initial comparison, variants were called by the same criteria as in the default pipeline (VAF ≥ 0.05, exonic, nonsynonymous or unannotated mutation according to Annovar, classified by FilterMutectCalls as “PASS”, “clustered event” or “germline”, called by at least 2 variant callers including Mutect2). Mutation calls were then compared and verified in their own replicate, in particular they had to be present in the same gene, in the same position and with the same nucleotide substitution. We found a list of 145 variants, in which 33/145 were discordant (Only-MI or Only-BO variants). After a manual revision using Integrative Genomics Viewer (IGV) v2.16.2, most discordances (23/33) were resolved as follows: 1) variants on the *PTEN* gene were blacklisted as sequencing artifacts, because the exact same variant was present in the large majority of samples. We inferred it was caused by a systematic sequencing error that went uncaptured by the Panel of Normal, and not a true mutation; 2) mutations classified as “germline”, showing a ~50% VAF, that were not called equally in the replicate (due to the probabilistic nature of the “germline” filter in FilterMutectCall[17]) were classified as real events and flagged as “Mutect filter error”; 3) mutations classified as “clustered events” which are not called so in the replicate (due to a co-occurrence with a real sequencing error within the assembly region[17]) were classified as real events and flagged as “Mutect filter error”. Variants detected in Sample_59 were also removed from the list of variants because its MI replicate did not have any reads (failed sequencing). Only SNV flagged as “subclonal”, “Mutect error filter” and “low coverage” were considered true discordances, interpreted as true somatic mutations that were not found in the replicate.

7. MI-BO validation analysis

7.1. Copy Number – UMA vs ULP-WGS

UMA's broad CN calls were compared with integer CN calls estimated by ichorCNA, a specific tool developed to estimate CN from ULP-WGS profiles. For this comparison, chromosome arms were first classified according to the presence or absence of a CN event. For UMA samples: CN ≥ 2.40 was classified as amplification (AMP), CN ≤ 1.60 as deletion (DEL), 1.60 < CN < 2.40 as normal (NEUT). For ULP-WGS samples: CN > 2 is AMP; CN < 2 is DEL; CN = 2 is NEUT. The concordance value per sample was estimated as the ratio of concordant events to total events.

7.2. Single Nucleotide Variant

As in the BO-MI analysis, the variants were selected by the default criteria: VAF ≥ 0.05, exonic, nonsynonymous or unannotated mutation according to Annovar, classified by FilterMutectCalls as "PASS", "clustered event" or "germline", called by at least 2 variant callers including Mutect2. Initially we found 42 variants, 13 of which were also called in the external panel. All discrepancies were explored and most of them were resolved by directly inspecting BAM files using IGV. Discordant variants due to Mutect2 filter errors (as explained above - chapter 4.3) were recovered. Furthermore, all sequencing errors (e.g. *RPL10* 153629178 C>T), variants in genes not covered by the MI panel, and variants later identified as SNPs were removed from the list. We obtained a new list of 24 SNVs, of which 6/24 were discordant for the following reasons: low coverage, Mutect errors, and indels position mismatch (due to sight differences in variant annotation tools positioning).

8. Dilution test to assess VAF limit of detection

A dilution test was performed to identify the minimum VAF threshold for UMA panel mutation's call. The test was conducted on a peripheral blood sample from a patient with hairy cell leukemia (HCL), which exhibited the *BRAF* V600E mutation and 7q deletion (**Supp. Figure S12a**). One portion of the sample was used in its entirety (in order to achieve a theoretical 100% tumor purity) and another portion was used to perform theoretical 50%, 25% and 5% dilutions of tumour cells with normal cells. For each dilution, a tumor purity assessment was conducted using CD19 and CD103 markers in flow cytometry. Additionally, NGS analysis by using the UMA panel was performed to ascertain the VAF of the *BRAF* V600E mutation in the sample. Each sample was tested in duplicate. Results of this comparison are presented in **Supp. Figure S12c**, where each theoretical dilution is associated with the percentage of purity in HCL found in the cytometric analysis and the corresponding observed VAF for the *BRAF* V600E mutation. The results demonstrate a significant correlation ($R = 0.97$, $p = 7.4e\text{-}0.5$) and the ability of UMA panel to detect mutations even with VAF < 5% (**Supp. Figure S12b,c**).

9. Statistical analysis

A Pearson correlation test was conducted to examine the linear relationship between BO and MI mutations' VAFs. Assumptions of normality were verified with Shapiro-Wilk test. The analysis was performed using R software version 4.2. The correlation coefficient *r* and corresponding p-value were reported, with significance assessed at $\alpha=0.05$.

Supplementary Bibliography

1. Araujo, L. H. *et al.* Impact of Pre-Analytical Variables on Cancer Targeted Gene Sequencing Efficiency. *PLoS One* **10**, e0143092 (2015).

2. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873 (2016).

3. Kuilman, T. *et al.* CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol* **16**, 1–15 (2015).

4. Mazzocchetti, G. *et al.* BoBafit: A copy number clustering tool designed to refit and recalibrate the baseline region of tumors' profiles. *Comput Struct Biotechnol J* **20**, 3718 (2022).

5. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics* **43**, 11.10.1-11.10.33 (2013).

6. Singh, A. K. *et al.* Detecting copy number variation in next generation sequencing data from diagnostic gene panels. *BMC Med Genomics* **14**, 1–12 (2021).

7. Ai, N., Cai, H., Solovan, C. & Baudis, M. CNARA: reliability assessment for genomic copy number profiles. *BMC Genomics* **17**, 799 (2016).

8. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

9. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

10. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *BioRxiv* (2019) doi:10.1101/861054.

11. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint* (2012) doi:10.48550/arXiv.1207.3907.

12. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568–576 (2012).

13. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Medicine 2020 12:1* **12**, 1–13 (2020).

14. Cai, L., Yuan, W., Zhang, Z., He, L. & Chou, K. C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific Reports 2016 6:1* **6**, 1–9 (2016).

15. Bian, X. *et al.* Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics* **19**, 1–11 (2018).

16. He, X. *et al.* Comprehensive fundamental somatic variant calling and quality management strategies for human cancer genomes. *Brief Bioinform* **22**, 1–15 (2021).

17. gatk/docs/mutect/mutect.pdf at master · broadinstitute/gatk. https://github.com/broadinstitute/gatk/blob/master/docs/mutect/mutect.pdf.

18. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462 (2013).

19. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).

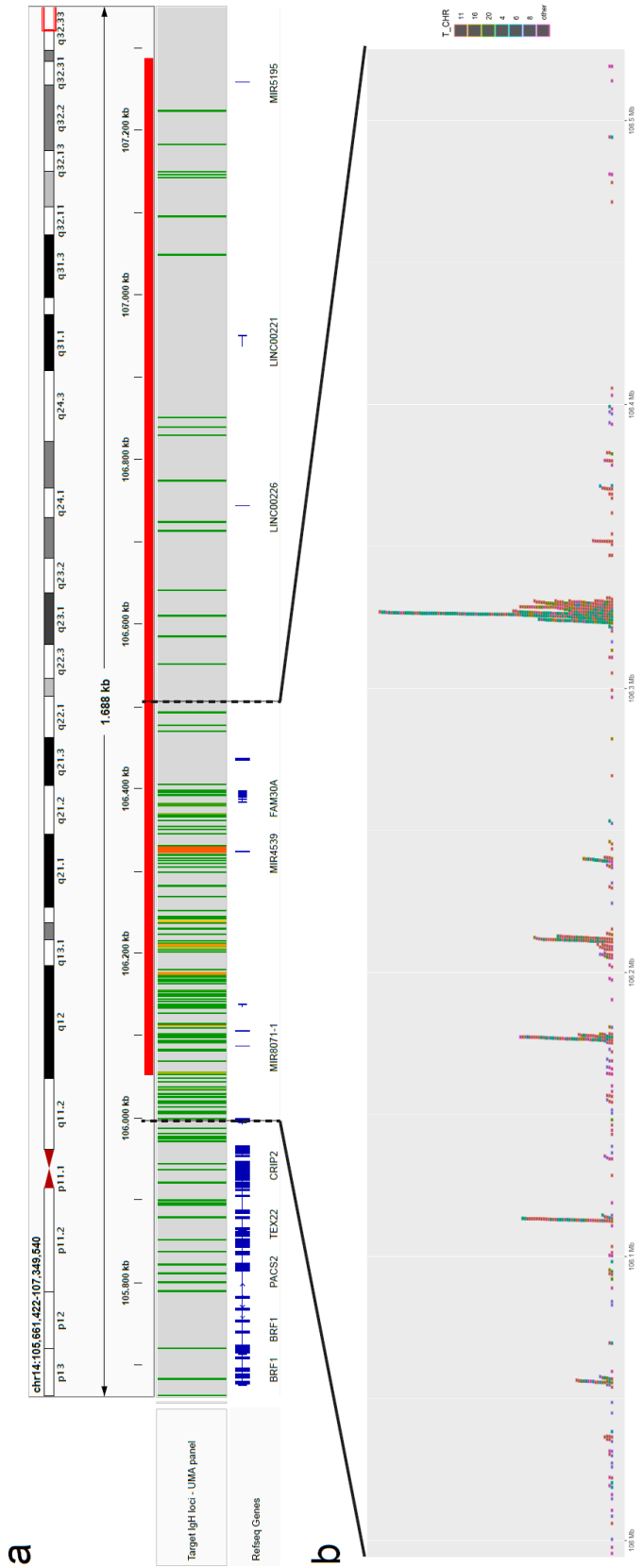
20. Creighton, C. J. *et al.* Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature 2013 499:7456* **499**, 43–49 (2013).

21. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods 2018 15:8* **15**, 591–594 (2018).

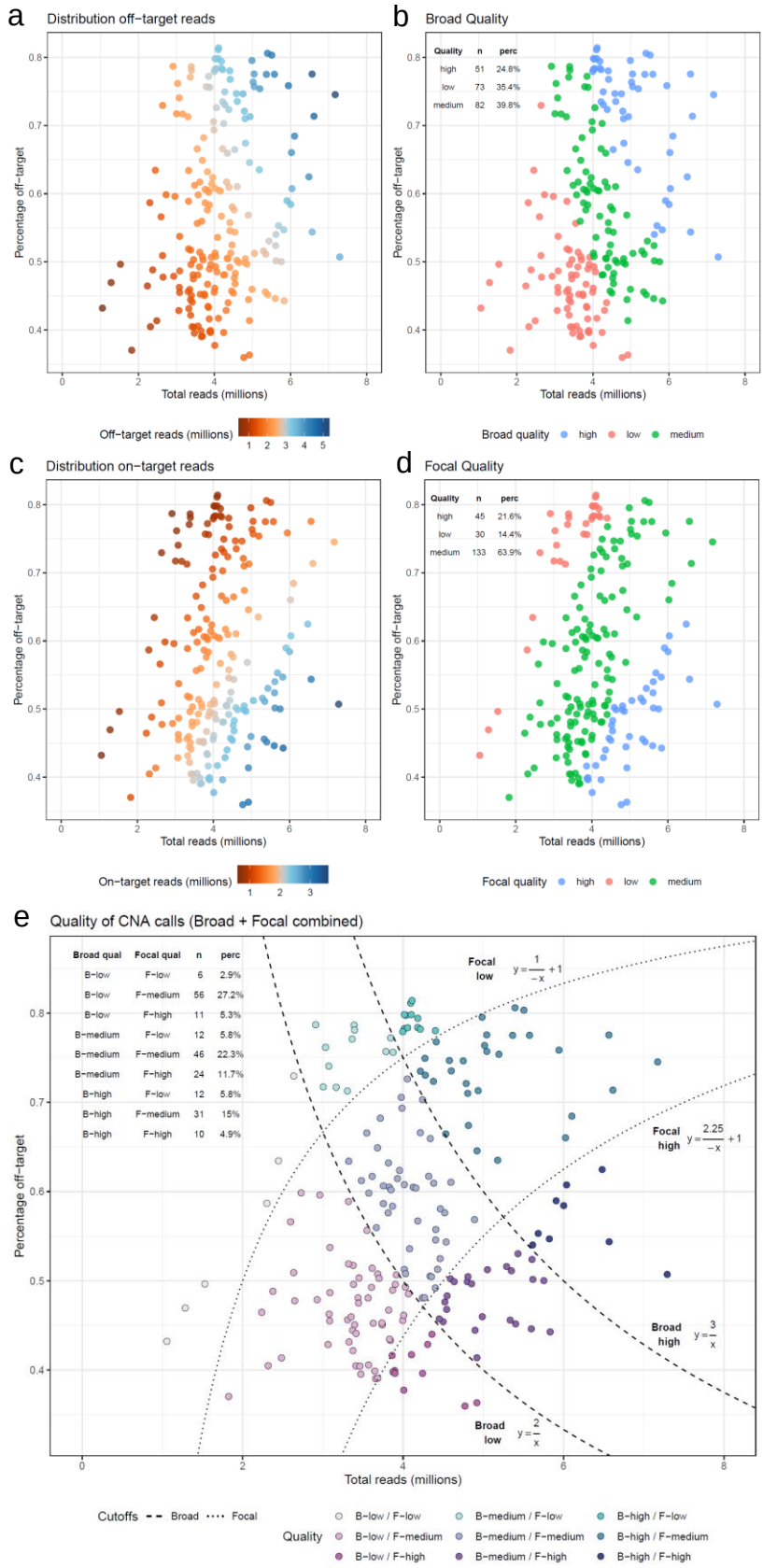
22. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164–e164 (2010).

Unique Molecular Assay (UMA): an NGS targeted panel for efficient and comprehensive genomic profiling and risk stratification of Multiple Myeloma

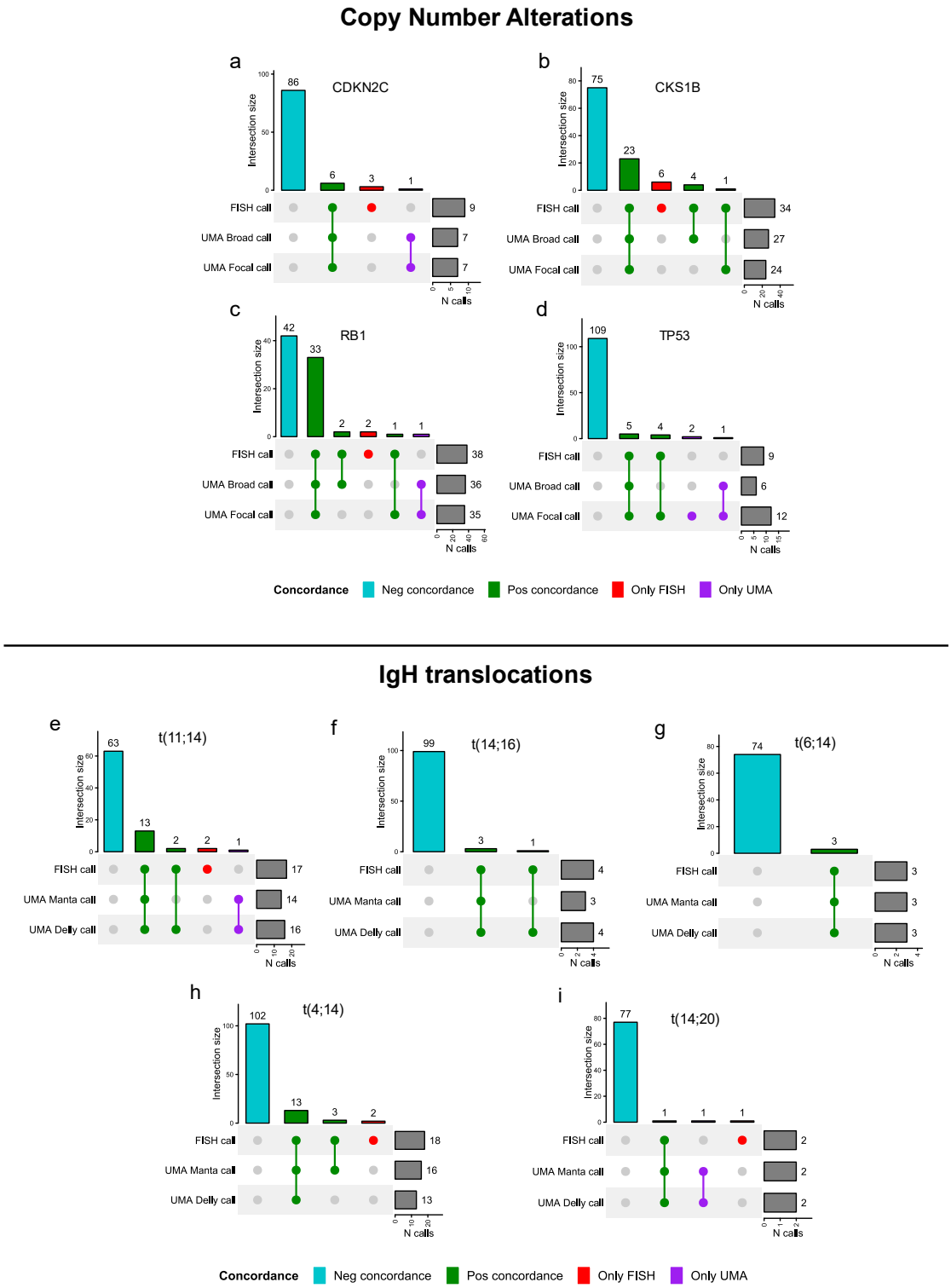
Supplementary figures



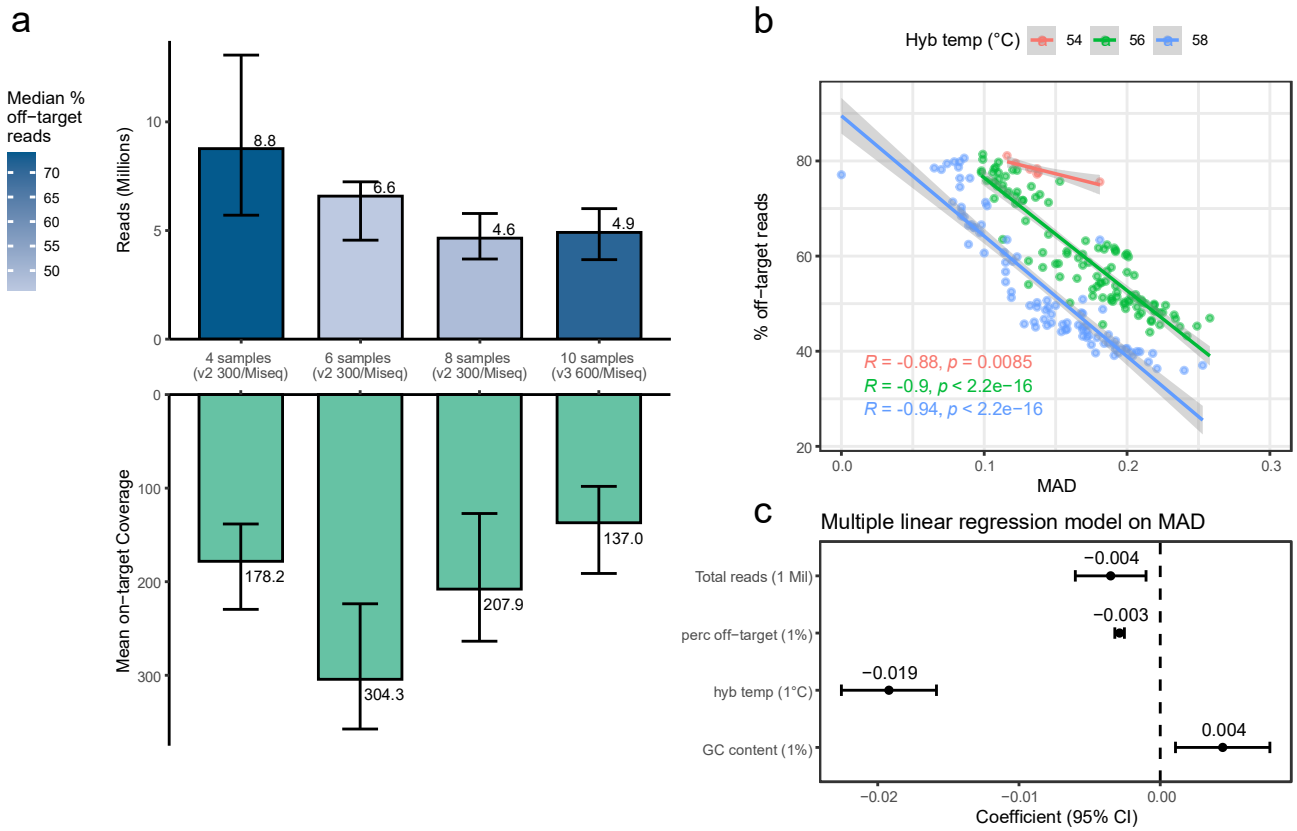
Supplementary figure S1: targeted regions selected for detecting t-IgH events. The distribution of the position of n=757 translocation breakpoints were used to define regions of interest. **a)** IGV screenshot representing the full IGH locus (red bar) and the UMA panel regions selected for capturing t-IgH events (data track). The colors of the data track regions correspond to the number of breakpoints found in those regions: green = 1 event, yellow = 2-50 events, orange = more than 50 events. **b)** Zoom in the IGH locus (chr14:1.60-1.65 Mb) where most breakpoints are concentrated. Stacked squares represent single translocations breakpoints, color coded for type of translocations.



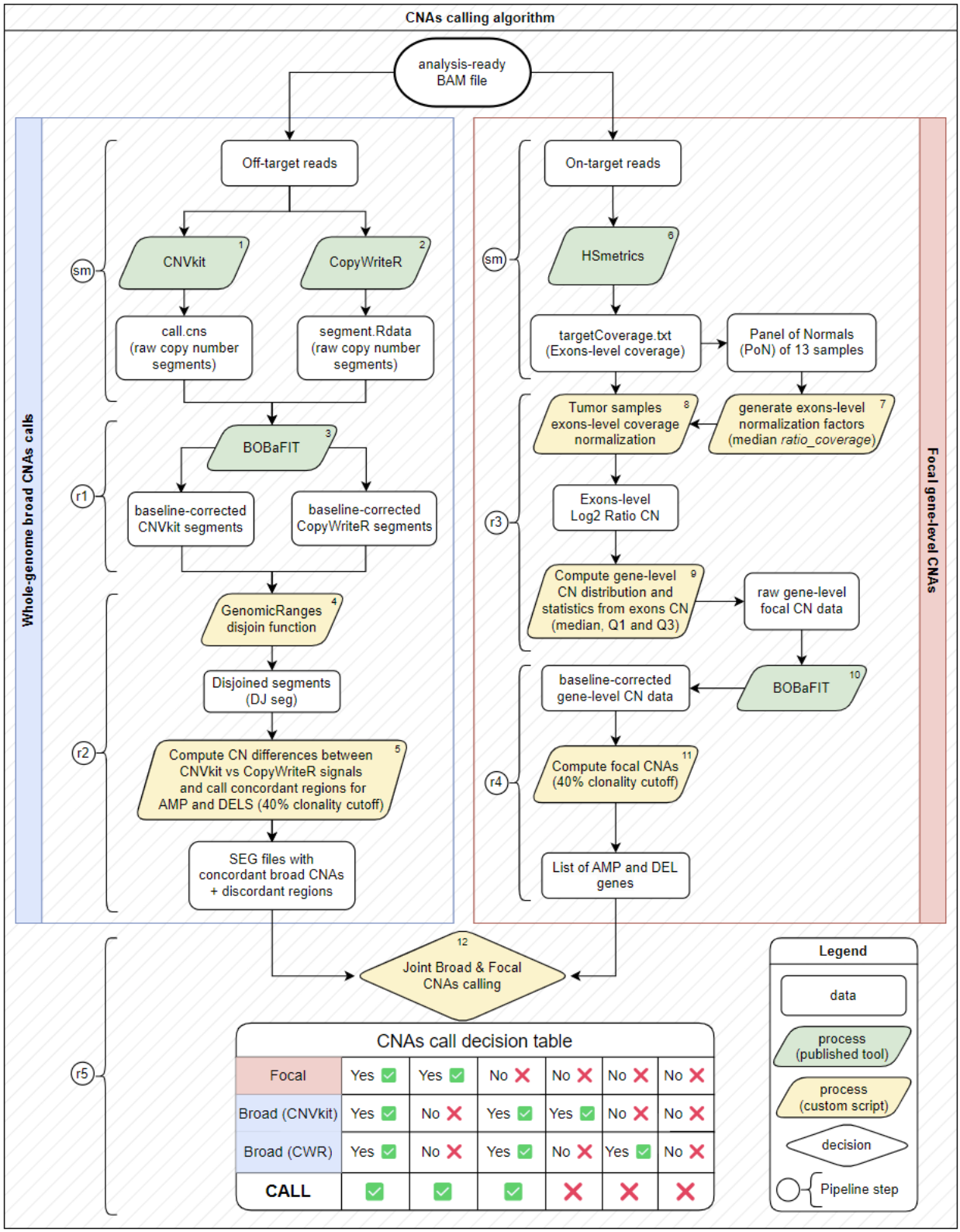
Supplementary figure S2: Scatterplots illustrating the definition of quality categories for broad and focal CN signal, considering the high variability in total reads (x axis) and percentage of off-target (y axis) in our samples. Each dot represent a different sample. **a)** Distribution of the number of off-target reads (color coded) and **b)** broad quality categories assignment based on the number of off-target reads. **c)** Distribution of the number of on-target reads (color coded) and **d)** focal quality categories assignment based on the number of on-target reads. **e)** Summary scatterplot, describing the combination of both focal quality and broad quality categories for each sample. Dashed lines correspond to the thresholds used to define the different categories. Equations can be used to compute the empirical sample quality based on x (total reads) and y (percentage of off-target) sequencing parameters.



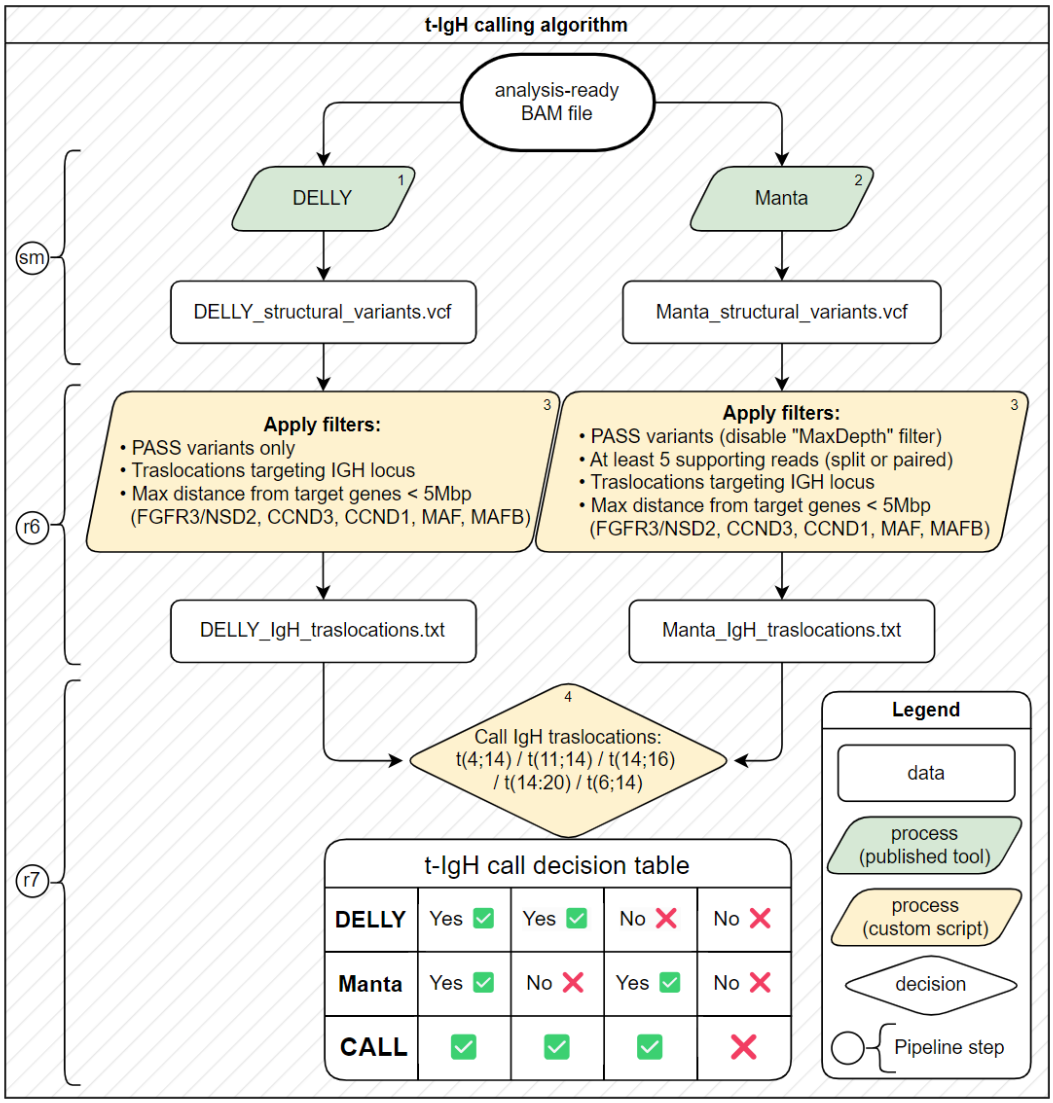
Supplementary Figure S3: Intra-laboratory validation of UMA panel's calls in the BO Cohort (N=129 patients), compared with gold-standard FISH calls. Upper panel shows CNAs comparison, bottom panel shows t-IgH comparison. For each distinct comparison, only patients with both UMA and FISH available data were considered. **a, b, c, d**) UpSet plots depicting all the intersection groups observed among the UMA panel's CNA calling methods (Broad and Focal calls) in comparison with FISH calls, in four target genes: CDKN2C (chromosome 1p), CKS1B (chromosome 1q), RB1 (chromosome 13q), and TP53 (chromosome 17p). Positive and negative concordance intersections groups between UMA and FISH calls are represented in green and light blue, respectively. Discordant intersections between UMA and FISH calls are represented in red (only FISH) and purple (only UMA). **e, f, g, h, i**) UpSet plots depicting all the intersection groups observed among the UMA panel's t-IgH calling methods (Manta and DELLY calls) in comparison with FISH calls, for the five canonical t-IgH. Positive and negative concordance intersections groups between UMA and FISH calls are represented in green and light blue, respectively. Discordant intersections between UMA and FISH calls are represented in red (only FISH) and purple (only UMA).



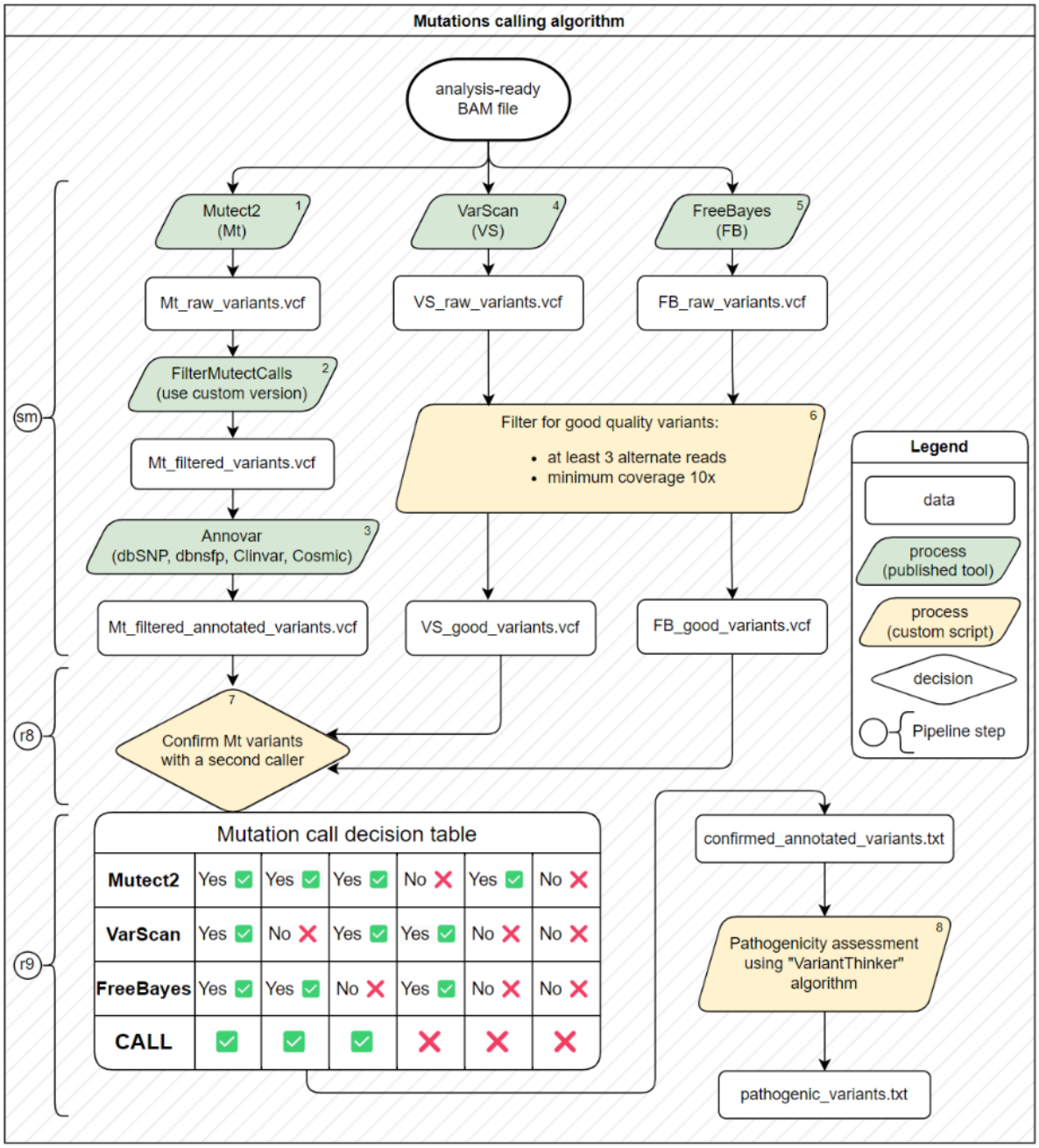
Supplementary Figure S4: Definition of optimal sequencing run parameters. A variety of parameters for UMA panel sequencing runs have been explored (i.e. number of samples, flowcell type, hybridization temperature) to identify the most appropriate trade-off to maximize UMA panel performances in terms of on-target coverage, percentage off-target reads and MAD quality. **a**) Number of samples and flowcell type exploration. The upper bar plot represents the median and standard deviation of total reads (in millions) obtained for each experimental configuration. Darker shades of blue corresponds to higher median percentages of off-target reads observed in each configuration. The bottom bar plot illustrates the median and standard deviation of on-target coverage for each configuration. **b**) Correlation analysis between percentage of off-target reads and MAD quality at various hybridization temperatures. A linear regression line and a Pearson's R correlation test was performed for each hybridization temperature. **c**) Forest plot showing the significant ($p<0.05$) variables in a multiple linear regression model to predict MAD, starting with all sequencing parameters. Error bars correspond to 95% CI.



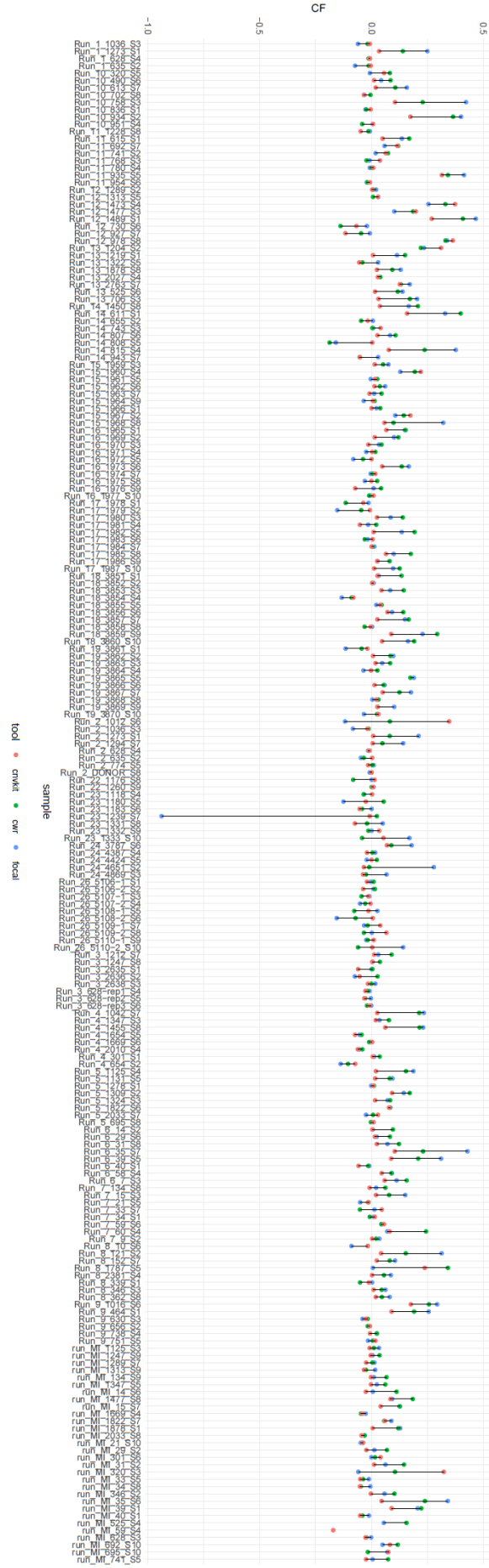
Supplementary figure S5: Flowchart of CNAs calling pipeline. Two CNAs calling approaches (Whole-genome broad CNAs and Focal gene-level CNAs, within blue and red boxes, respectively) were used to generate CNAs calls. The algorithm steps are numbered in order of execution in the top-right corner of processes items. Curly brackets with circles denote different pipeline scripts used to execute processes (sm = snakemake code, rX = R code, where X denotes the number of the script in the GitHub repository). CWR = CopyWriteR.



Supplementary figure S6: Flowchart of t-IgH calling pipeline. Two different tools (DELLY and Manta) were integrated to generate t-IgH calls. The algorithm steps are numbered in order of execution in the top-right corner of processes items. Curly brackets with circles denote different pipeline scripts used to execute processes (sm = snakemake code, rX = R code, where X denotes the number of the script in the GitHub repository).

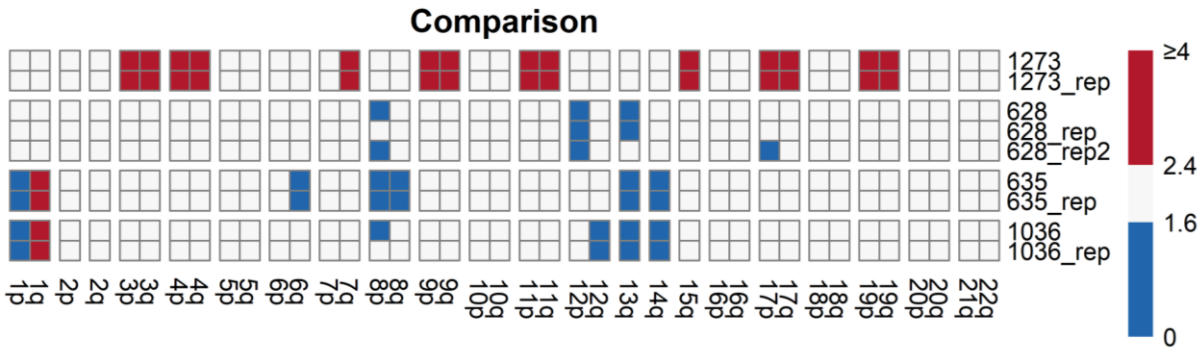


Supplementary figure S7: Flowchart of mutations calling pipeline. Three different tools (Mutect2, VarScan2 and FreeBayes) were integrated to produce high-confidence mutations calls. The algorithm steps are numbered in order of execution in the top-right corner of processes items. Curly brackets with circles denote different pipeline scripts used to execute processes (sm = snakemake code, rX = R code, where X denotes the number of the script in the GitHub repository).



Supplementary figure S8: BOBaFIT correction factors used in the CN analysis algorithm for adjusting the baseline regions of both broad Copy Numer segments (from CNVkit and CopywriteR tools - in red and green, resepectively) and focal Copy Number data (from custom analysis, in blue).

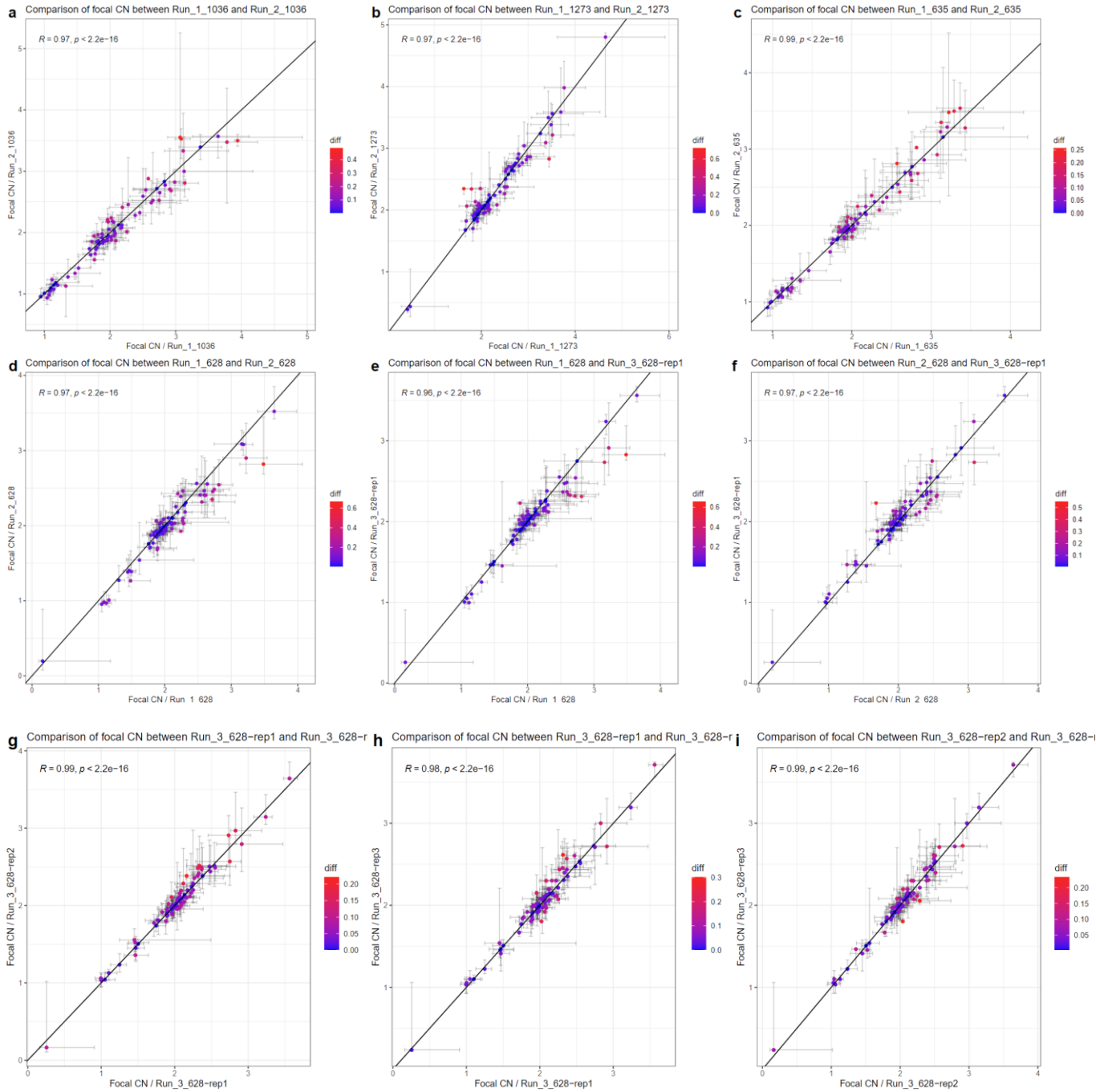
a



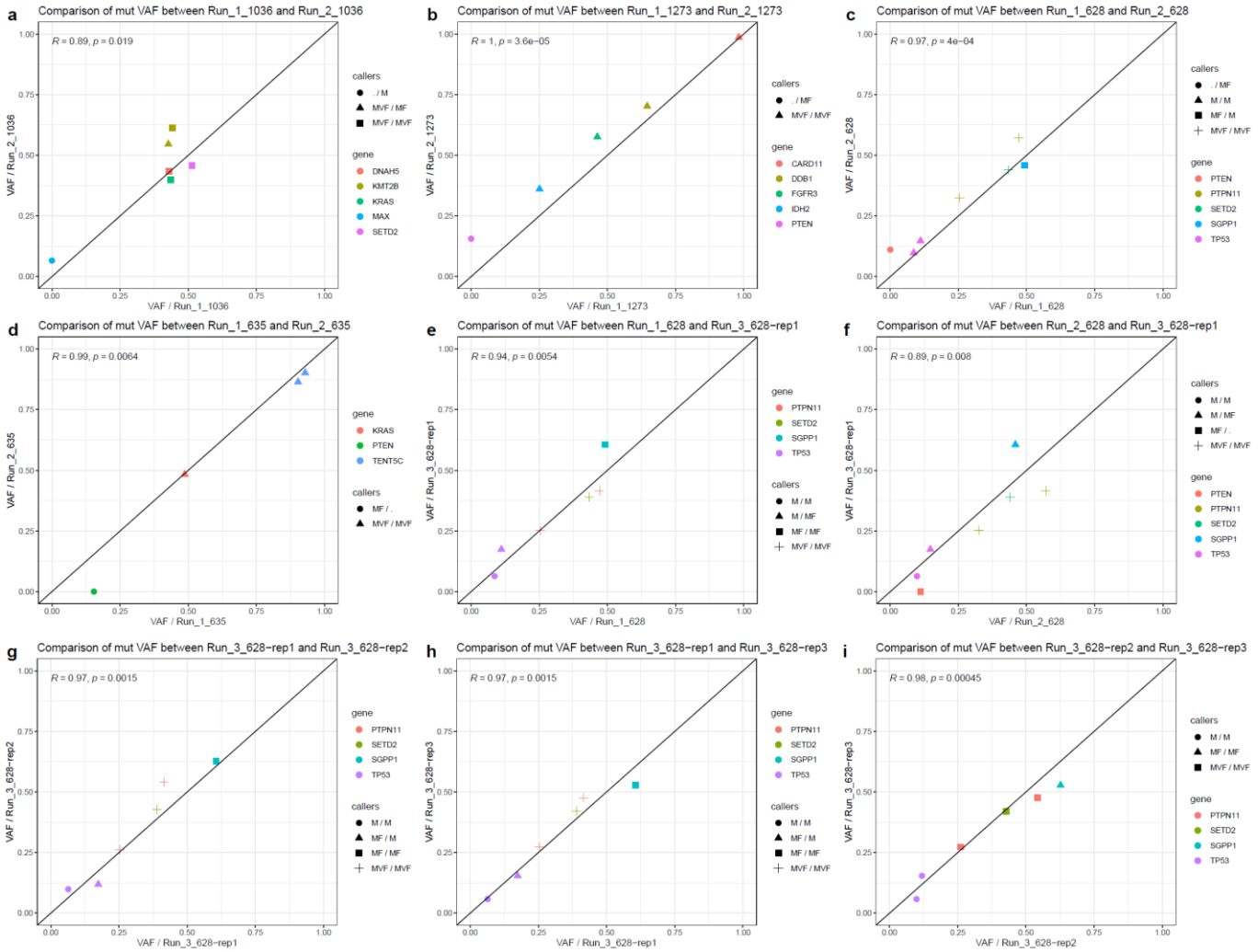
b

Comparison	DNA	Concordance (%)
1273 : Run_1_1273 vs Run_2_1273	1273	100
628 : Run_1_628 vs Run_2_628	628	100
635 : Run_1_635 vs Run_2_635	635	98
628 : Run_1_628 vs Run_3_628-rep1	628	96
1036 : Run_1_1036 vs Run_2_1036	1036	95

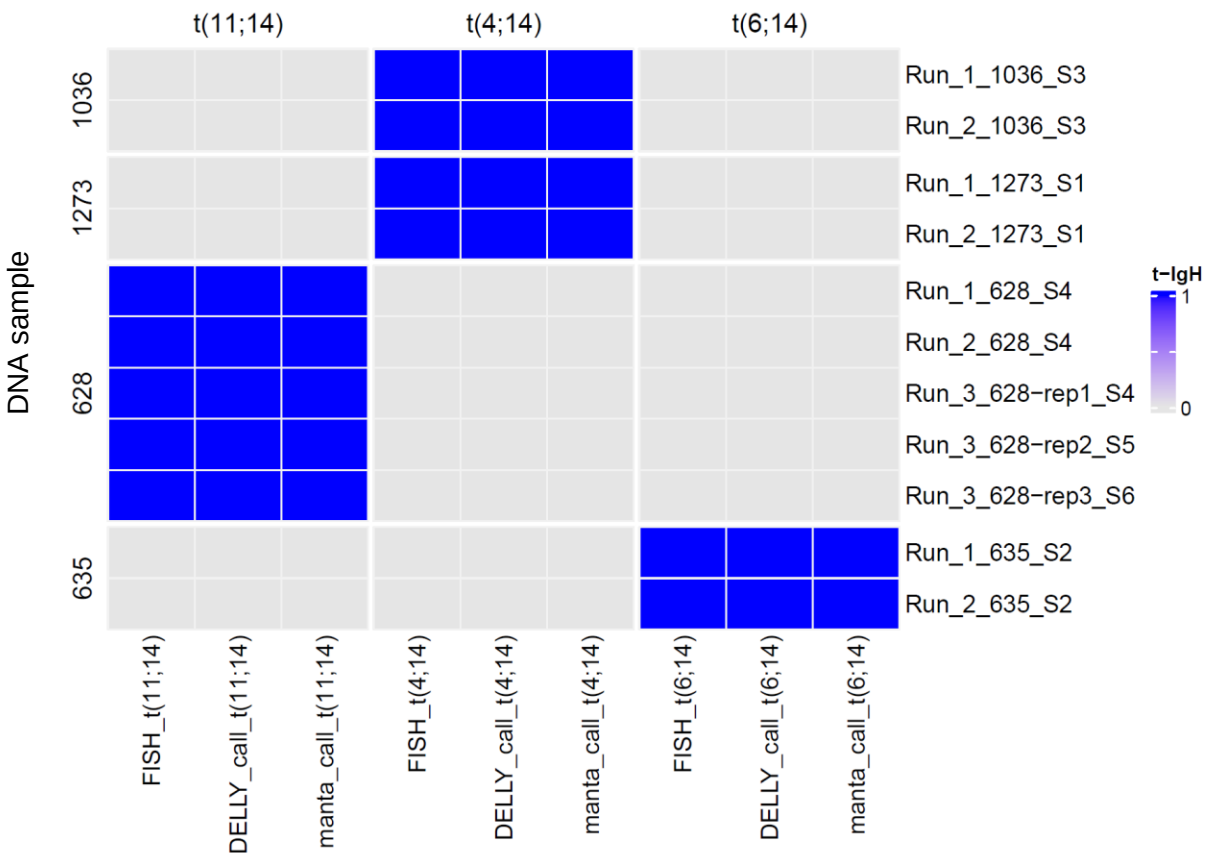
Supplementary figure S9: inter-run and intra-run validation of broad genome-wide CNAs. **a)** Heatmap showing broad CNAs detected in each sample: blue indicates deletion (CN ≤ 1.60) and red amplification (CN ≥ 2.40). **b)** Table showing the concordance of CN profiles between replicates pairs. In detail, the concordance has been calculated as the percentage of concordant CN segments (difference of CN < 0.20) out of the total number of CN segments (see Supplementary Methods).



Supplementary figure S10: inter-run and intra-run validation of focal gene-level CNAs. Scatterplots representing the 82 genes CN values in each replicate pair are shown. Error bars corresponds to inter-quartile ranges (computed from the distribution of exon-level CN values per gene). **a-f)** Inter-run (run 1 vs run 2 vs run 3) comparison analyses on four different DNA samples (i.e. DNA 1036, 1273, 635 and 628. **g-i)** Three intra-run (run 3) comparison analyses on one DNA sample analyzed in triplicate (DNA 628).

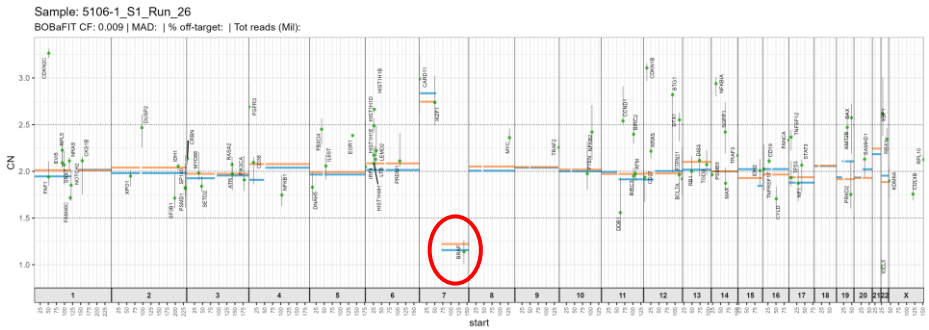


Supplementary figure S11: inter-run and intra-run validation of mutation calls. Scatterplots representing mutations detected in each replicate pair are shown. Point shape corresponds to the list of variant callers which identified the mutation (M = Mutect, F=Freebayes, V = Varscan) in either sample. **a-f)** Inter-run (run 1 vs run 2 vs run 3) comparison analyses on four different DNA samples (i.e. DNA 1036, 1273, 635 and 628. **g-i)** **T**hree intra-run (run 3) comparison analyses on one DNA sample analyzed in triplicate (DNA 628).

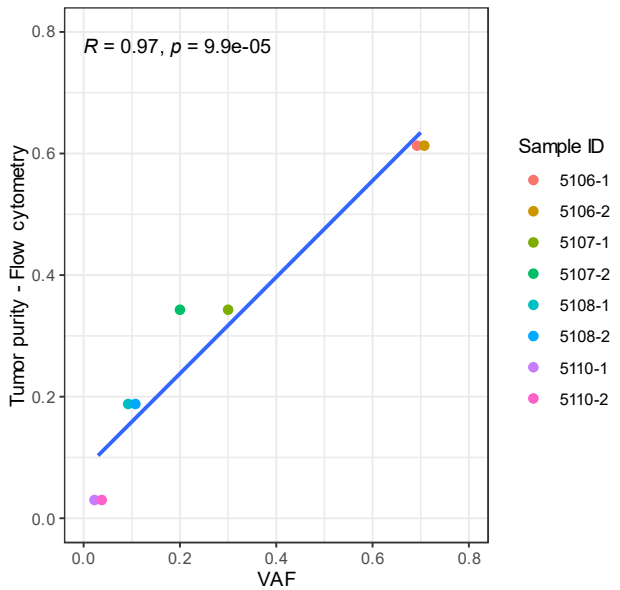


Supplementary figure S12: inter-run and intra-run validation of t-IgH calls. Heatmap representing the t-IgH calls for FISH and UMA panel (Manta and DELLY tools) for all the replicates.

a



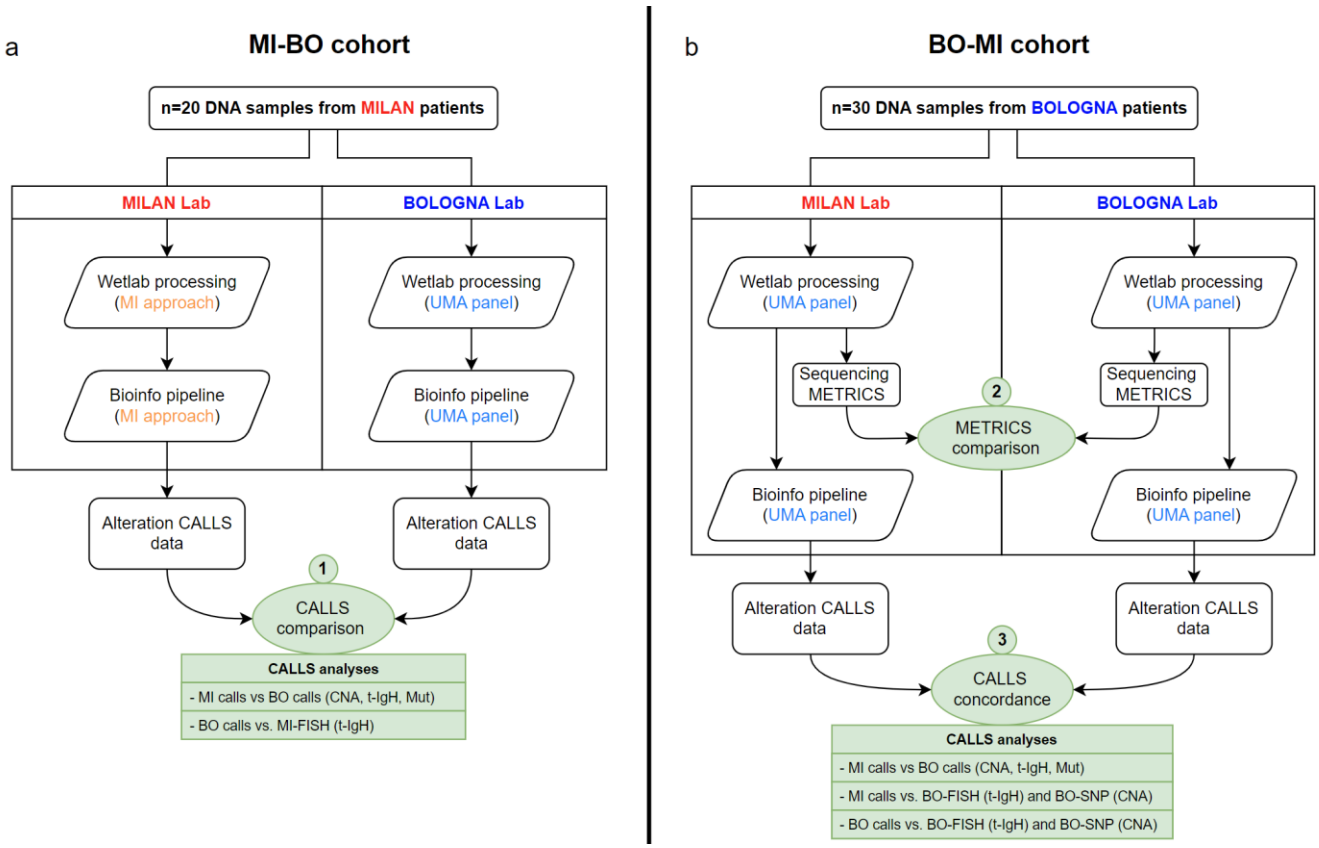
b



c

Sample ID	Expected Purity (%)	Obtained Purity (%)	VAF
5106-1	100	61,3	0,7
5106-2	100	61,3	0,7
5107-1	50	34,3	0,3
5107-2	50	34,3	0,2
5108-1	25	18,8	0,1
5108-2	25	18,8	0,1
5110-1	5	2,9	0,03
5110-2	5	2,9	0,03

Supplementary figure S13: Dilution test to identify the minimum VAF threshold for UMA panel mutation's call. The test was conducted on a peripheral blood sample from a patient with hairy cell leukemia (HCL), which exhibited the BRAF V600E mutation and 7q deletion. a) Copy number profile of the sample, where is evident a deletion on BRAF (del chr 7q, highlighted in a red circle).
b,c) Bulk lysis was performed on the sample to isolate the WBC and to perform cells count. One portion of the sample was used in its entirety (in order to achieve a theoretical 100% purity) and another portion was used to perform theoretical 50%, 25% and 5% dilutions of tumor cells with normal cells. For each dilution, a tumor purity assessment was conducted using CD19 and CD103 markers in flow cytometry. Additionally, NGS analysis using the UMA panel was performed to ascertain the VAF of the BRAF V600E mutation in the sample. Each sample was tested in duplicate. The results of this comparison are presented in the table above, where each theoretical dilution is associated with the percentage of purity in HCL found in cytometric analysis and the corresponding VAF for the BRAF V600E mutation. The results demonstrate a significant correlation (R = 0.97, p = 7.4e-0.5) and the ability of UMA panel to detect mutations even with VAF < 5%.



Supplementary figure S14: schema of the intra-lab analyses. **a)** «MI-BO» comparison of UMA panel alteration calls using external wetlab procedures and analysis pipeline (MI approach). **b)** «BO-MI» validation of UMA panel wetlab procedures and alteration calls, using the same analysis pipeline and wetlab procedures in two different laboratories. Specific analyses steps are highlighted in green boxes.

Supplementary table S1: list and criteria used for the selection of the 82 genes targeted by UMA panel. TSG = Tumor suppressing gene, Amp = amplification, Del = deletion.

Gene info			Four criteria used for gene selection			
Number	Gene name (HUGO)	Gene exons length (bp)	1 - Mutated pathway in MM	2 - Therapy target	3 - GISTIC target (focal CNA) [7]	4 - Role in Cancer (COSMIC cancer gene)
1	BCL7A	6245				fusion
2	CARD11	4366				oncogene
3	IDH1	2790				oncogene
4	IDH2	2694				oncogene
5	IL6ST	9078				oncogene
6	KRAS	5889				oncogene
7	MYD88	2874				oncogene
8	NRAS	4449	MAP-K			oncogene
9	PIK3CA	9093				oncogene
10	PTPN11	6400				oncogene
11	SF3B1	6682				oncogene
12	STAT3	5047				oncogene
13	XPO1	5316				oncogene
14	BRAF	2480	MAP-K			oncogene
15	CCND1	4307	Cell cycle		Amp	oncogene
16	FGFR3	4438			Del	oncogene
17	MYC	2345			Amp Del	oncogene
18	TNFRSF17	994		Yes		oncogene
19	KDM6A	5438				oncogene, TSG
20	NOTCH2	11389			Amp	oncogene, TSG
21	BIRC3	6866	NF-kB non canonical		Del	oncogene, TSG
22	IRF4	5331				oncogene, TSG
23	NFKB2	3317	NF-kB		Del	oncogene, TSG
24	TP53	2712			Del	oncogene, TSG
25	ATM	13238	DNA damage			TSG
26	ATR	8249	DNA damage			TSG
27	BAX	1410				TSG
28	CDKN1B	2657	Cell cycle			TSG
29	CDKN2C	3318	Cell cycle		Del	TSG
30	CYLD	8834			Del	TSG
31	FANCA	6172				TSG
32	MAX	2824			Del	TSG
33	PRDM1	5320				TSG
34	PTEN	9027				TSG
35	RB1	4840			Del	TSG
36	RPL10	2808				TSG
37	RPL5	1043				TSG
38	SETD2	8142				TSG
39	BTG1	2448	Cell Cycle - anti proliferative			TSG
40	IKZF1	6248				TSG
41	NF1	13260				TSG
42	BIRC2	3764	NF-kB non canonical		Del	
43	CD19	1957		Yes		
44	CD27	1338	Cell survival			
45	CD38	5668		Yes		
46	CKS1B	759			Amp	
47	CRBN	2206		Yes		
48	CUL4B	5371	proteasome ubiquitin cell cycle / DNA repair			
49	DDB1	4506	Proteasome ubiquitin cell cycle / DNA repair			
50	DIS3	7439	RNA processing and degradation			
51	DNAH5	15633	Adhesion and motility			
52	DTX1	3455	Notch signaling			
53	DUSP2	1688	MAP-K			
54	EGR1	3138	Differentitation and proliferation			
55	EVI5	7403			Del	
56	FAF1	4367			Del	
57	FAM46C / TENT5C	5751	mRNA stability and gene expression		Del	
58	FBXO4	2459	Cell Cycle Ubiquitination CCND1			
59	HIST1H1B	790	Histone			
60	HIST1H1D	777	Histone			
61	HIST1H1E	785	Histone			
62	HIST1H4H	1034	Histone			
63	IGLL5	1050	Immunoglobluin			
64	KMT2B	10279	Chromatin organization			
65	LEMD2	3185	MAP-K / AKT signaling			
66	LTB	899	Cytokine Signaling in Immune system			
67	NFKB1	4355	NF-kB			
68	NFKBIA	1558	NF-kB			
69	PRKD2	3747	Signaling pathways PI3K/AKT and NF-kB			
70	PSMB5	1464		Yes		
71	PSMD1	3326	Component of Proteasome	Yes		
72	RASA2	5614	MAP-K			
73	RBX1	1188	Proteasome ubiquitination cell cycle			
74	SAMHD1	4784	DNA damage - Cell Cycle			
75	SGPP1	3312	Cell survival			
76	SNX7	1756	Cell survival			
77	SP140	3575	Nuclear body			
78	TGDS	1902	Metabolism			
79	TNFSF12	2823	NF-kB			
80	TRAF2	2409	NF-kB non canonical			
81	TRAF3	7779	NF-kB non canonical		Del	
82	XBP1	1850	Cell survival			

REFERENCES

1 Bolli, N., et al. (2018). Analysis of the genomic landscape of multiple myeloma highlights novel prognostic markers and disease subgroups. Leukemia, 32(12), 2604-2616.
2 Kortüm, K. M.,et al. (2016). Targeted sequencing of refractory myeloma reveals a high incidence of mutations in CRBN and Ras pathway genes. Blood, 128(9), 1226–1233.
3 Maura, F., et al. (2019). Genomic landscape and chronological reconstruction of driver events in multiple myeloma. Nature communications, 10(1), 3835.
4 Barrio, S., et al. (2019). Spectrum and functional validation of PSMB5 mutations in multiple myeloma. Leukemia, 33(2), 447-456.
5 Hoang, P. H., et al. (2018). Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. Leukemia, 32(11), 2459–2470.
6 Hoang, P. H., et al. (2019). Mutational processes contributing to the development of multiple myeloma. Blood cancer journal, 9(8), 60.
7 Terragna, C., et al. (2024). Multi-dimensional scaling techniques unveiled gain1q&loss13q co-occurrence in Multiple Myeloma patients with and adverse clinical features. Nat. Comm., 15(1), 1551.
8 Walker, B. A., et al. (2018). Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma. Blood, The Journal of the American Society of Hemat., 132(6), 587-597.

Supplementary table S2: list of the 170 IgH locus regions targeted by UMA panel.

chromosome	start	end	width	gene	translocations_overlaps	start_to_next_start_distance
14	104596845	104597245	401	IgH locus	1	380493
14	104977338	104977738	401	IgH locus	1	179450
14	105156788	105157188	401	IgH locus	1	137915
14	105294703	105295103	401	IgH locus	1	62918
14	105357621	105358021	401	IgH locus	1	36151
14	105393772	105394172	401	IgH locus	1	9141
14	105402913	105403313	401	IgH locus	1	6922
14	105409835	105410235	401	IgH locus	1	10240
14	105420075	105420591	517	IgH locus	2	9929
14	105430004	105430404	401	IgH locus	1	1434
14	105431438	105431838	401	IgH locus	1	103554
14	105534992	105535392	401	IgH locus	1	14837
14	105549829	105550229	401	IgH locus	1	37630
14	105587459	105587859	401	IgH locus	1	747
14	105588206	105588606	401	IgH locus	1	48949
14	105637155	105637555	401	IgH locus	1	26081
14	105663236	105663636	401	IgH locus	1	20615
14	105683851	105684251	401	IgH locus	1	36892
14	105720743	105721143	401	IgH locus	1	69272
14	105790015	105790415	401	IgH locus	1	11376
14	105801391	105801791	401	IgH locus	1	857
14	105802248	105802648	401	IgH locus	1	9836
14	105812084	105812484	401	IgH locus	1	11052
14	105823136	105823536	401	IgH locus	1	15448
14	105838584	105838984	401	IgH locus	1	13918
14	105852502	105852902	401	IgH locus	1	28048
14	105880550	105880950	401	IgH locus	1	14093
14	105894643	105895043	401	IgH locus	1	1844
14	105896487	105896887	401	IgH locus	1	4304
14	105900791	105901191	401	IgH locus	1	20818
14	105921609	105922009	401	IgH locus	1	15837
14	105937446	105938083	638	IgH locus	2	7687
14	105945133	105945533	401	IgH locus	1	26854
14	105971987	105972387	401	IgH locus	1	4073
14	105976060	105976460	401	IgH locus	1	2610
14	105978670	105979070	401	IgH locus	1	2655
14	105981325	105981725	401	IgH locus	1	6847
14	105988172	105988572	401	IgH locus	1	12056
14	106000228	106000628	401	IgH locus	1	5645
14	106005873	106006273	401	IgH locus	1	2529
14	106008402	106008802	401	IgH locus	1	6097
14	106014499	106014899	401	IgH locus	1	4087
14	106018586	106018986	401	IgH locus	1	1537
14	106020123	106020523	401	IgH locus	1	1764
14	106021887	106022287	401	IgH locus	1	3849
14	106025736	106026136	401	IgH locus	1	1375
14	106027111	106027511	401	IgH locus	1	2539
14	106029650	106030050	401	IgH locus	1	563
14	106030213	106030613	401	IgH locus	1	1001
14	106031214	106031614	401	IgH locus	1	3487
14	106034701	106035101	401	IgH locus	1	740
14	106035441	106035884	444	IgH locus	2	792
14	106036233	106036899	667	IgH locus	4	1936
14	106038169	106038569	401	IgH locus	1	6530
14	106044699	106045099	401	IgH locus	1	3769

14	106048468	106048868	401	IgH locus	1	5352
14	106053820	106054220	401	IgH locus	1	800
14	106054620	106055050	431	IgH locus	2	792
14	106055412	106056661	1250	IgH locus	14	1290
14	106056702	106057941	1240	IgH locus	6	12594
14	106069296	106069776	481	IgH locus	2	13499
14	106082795	106083195	401	IgH locus	1	1181
14	106083976	106084376	401	IgH locus	1	7920
14	106091896	106092296	401	IgH locus	1	1606
14	106093502	106094200	699	IgH locus	2	1009
14	106094511	106094911	401	IgH locus	1	596
14	106095107	106095507	401	IgH locus	1	2889
14	106097996	106098396	401	IgH locus	1	2133
14	106100129	106100529	401	IgH locus	1	1096
14	106101225	106101625	401	IgH locus	1	2113
14	106103338	106103738	401	IgH locus	1	7288
14	106110626	106111026	401	IgH locus	1	1564
14	106112190	106114211	2022	IgH locus	42	2177
14	106114367	106114819	453	IgH locus	2	14001
14	106128368	106128768	401	IgH locus	1	5828
14	106134196	106134992	797	IgH locus	2	960
14	106135156	106135556	401	IgH locus	1	1138
14	106136294	106136694	401	IgH locus	1	2353
14	106138647	106139047	401	IgH locus	1	4209
14	106142856	106143256	401	IgH locus	1	1441
14	106144297	106144697	401	IgH locus	1	1802
14	106146099	106146499	401	IgH locus	1	1838
14	106147937	106148337	401	IgH locus	1	2367
14	106150304	106150704	401	IgH locus	1	1095
14	106151399	106151799	401	IgH locus	1	2728
14	106154127	106154819	693	IgH locus	2	875
14	106155002	106155402	401	IgH locus	1	1214
14	106156216	106156930	715	IgH locus	3	7968
14	106164184	106164701	518	IgH locus	2	1687
14	106165871	106166271	401	IgH locus	1	678
14	106166549	106166949	401	IgH locus	1	2160
14	106168709	106169110	402	IgH locus	2	640
14	106169349	106169749	401	IgH locus	1	2522
14	106171871	106172271	401	IgH locus	1	879
14	106172750	106173150	401	IgH locus	1	2241
14	106174991	106175561	571	IgH locus	3	835
14	106175826	106178666	2841	IgH locus	56	4662
14	106180488	106181046	559	IgH locus	2	21314
14	106201802	106202202	401	IgH locus	1	658
14	106202460	106202860	401	IgH locus	1	2539
14	106204999	106205573	575	IgH locus	2	625
14	106205624	106206616	993	IgH locus	5	1001
14	106206625	106207025	401	IgH locus	1	1283
14	106207908	106208394	487	IgH locus	4	681
14	106208589	106210142	1554	IgH locus	7	1993
14	106210582	106212995	2414	IgH locus	56	2568
14	106213150	106213957	808	IgH locus	3	1933
14	106215083	106215483	401	IgH locus	1	9110
14	106224193	106224593	401	IgH locus	1	5395
14	106229588	106229988	401	IgH locus	1	407
14	106229995	106230395	401	IgH locus	1	1337
14	106231332	106231732	401	IgH locus	1	5405
14	106236737	106237137	401	IgH locus	1	453
14	106237190	106237590	401	IgH locus	1	594

14	106237784	106238184	401	IgH locus	1	1104
14	106238888	106240760	1873	IgH locus	17	4473
14	106243361	106243761	401	IgH locus	1	1714
14	106245075	106245475	401	IgH locus	1	827
14	106245902	106246302	401	IgH locus	1	6400
14	106252302	106252702	401	IgH locus	1	877
14	106253179	106253579	401	IgH locus	1	15915
14	106269094	106269494	401	IgH locus	1	12983
14	106282077	106282477	401	IgH locus	1	17159
14	106299236	106299636	401	IgH locus	1	6050
14	106305286	106305686	401	IgH locus	1	5499
14	106310785	106311185	401	IgH locus	1	2477
14	106313262	106313662	401	IgH locus	1	3005
14	106316267	106316667	401	IgH locus	1	3765
14	106320032	106320432	401	IgH locus	1	1187
14	106321219	106321995	777	IgH locus	2	1625
14	106322844	106331047	8204	IgH locus	282	8405
14	106331249	106331649	401	IgH locus	1	461
14	106331710	106332112	403	IgH locus	2	14989
14	106346699	106347103	405	IgH locus	2	3843
14	106350542	106350942	401	IgH locus	1	1052
14	106351594	106352238	645	IgH locus	8	2621
14	106354215	106354615	401	IgH locus	1	7416
14	106361631	106362031	401	IgH locus	1	5170
14	106366801	106367201	401	IgH locus	1	1411
14	106368212	106368713	502	IgH locus	2	1878
14	106370090	106370807	718	IgH locus	5	986
14	106371076	106371606	531	IgH locus	2	8910
14	106379986	106380506	521	IgH locus	3	2439
14	106382425	106383232	808	IgH locus	3	10995
14	106393420	106393820	401	IgH locus	1	3107
14	106396527	106396927	401	IgH locus	1	673
14	106397200	106397600	401	IgH locus	1	1884
14	106399084	106399484	401	IgH locus	1	6478
14	106405562	106405962	401	IgH locus	1	65524
14	106471086	106471486	401	IgH locus	1	6873
14	106477959	106478359	401	IgH locus	1	15977
14	106493936	106494410	475	IgH locus	2	58422
14	106552358	106552758	401	IgH locus	1	33665
14	106586023	106586423	401	IgH locus	1	24394
14	106610417	106610817	401	IgH locus	1	31214
14	106641631	106642031	401	IgH locus	1	72539
14	106714170	106714888	719	IgH locus	2	11002
14	106725172	106725920	749	IgH locus	2	49794
14	106774966	106775366	401	IgH locus	1	54676
14	106829642	106830042	401	IgH locus	1	9803
14	106839445	106839845	401	IgH locus	1	12047
14	106851492	106851892	401	IgH locus	1	197340
14	107048832	107049232	401	IgH locus	1	47764
14	107096596	107096996	401	IgH locus	1	46760
14	107143356	107143756	401	IgH locus	1	527
14	107143883	107144283	401	IgH locus	1	3066
14	107146949	107147349	401	IgH locus	1	3193
14	107150142	107150542	401	IgH locus	1	33538
14	107183680	107184372	693	IgH locus	2	40206
14	107223886	107224286	401	IgH locus	1	0

Supplementary table S3: UMA panel's genomic footprint.

	Footprint (bp)
Total genes	367521
Total IgH regions	92922
TOTAL PANEL FOOTPRINT	**460443**

Supplementary table S4: comparison between sequencing metrics obtained by Bologna (BO) and Milan (MI) laboratories, on the BO-MI cohort.
MAD = Median Absolute Deviation; PCT = Percentage (%); TOT = total

Metric	BO, N = 30 Median (IQR)	MI, N = 29 Median (IQR)	p-value Welch Two Sample t-test	SOFTWARE
MAD	0.149 (0.11675, 0.18475)	0.202 (0.18300, 0.21900)	<0.001	COPYWRITER
OFF TARGET	1729002 (1466755, 2364783)	2427054 (2153252, 2937359)	0.4	COPYWRITER
ON TARGET	2001329 (1540910, 2242534)	2404467 (2156828, 2551032)	<0.001	COPYWRITER
TOT READS	3852897 (3506415, 4446469)	4808852 (4354295, 5684380)	0.041	COPYWRITER
PCT OFF TARGET	0.4828 (0.407125, 0.620125)	0.5053 (0.493000, 0.530400)	0.9	COPYWRITER
ON BAIT BASES	169933827 (136476739, 199757114)	247231487 (222064169, 266341738)	<0.001	HS METRICS
NEAR BAIT BASES	53714267 (42496730, 61216257)	100605243 (86851711, 107304761)	<0.001	HS METRICS
OFF BAIT BASES	406376175 (347317561, 479657816)	672679145 (589935099, 790865594)	<0.001	HS METRICS
PCT SELECTED BASES	0.362664 (0.271960, 0.407308)	0.339227 (0.325872, 0.351591)	0.8	HS METRICS
PCT OFF BAIT	0.637336 (0.592692, 0.728039)	0.660773 (0.648409, 0.674128)	0.8	HS METRICS
ON BAIT VS SELECTED	0.763197 (0.753887, 0.788972)	0.716232 (0.710837, 0.721292)	<0.001	HS METRICS
MEAN BAIT COVERAGE	431.683 (346.692, 507.443)	628.043 (564.110, 676.588)	<0.001	HS METRICS