Computational and Structural Biotechnology Journal 20 (2022) 3718–3728

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

# *BoBafit*: A copy number clustering tool designed to refit and recalibrate the baseline region of tumors' profiles

G. Mazzocchetti [a,b,1], A. Poletti [a,b,1], V. Solli [a,b], E. Borsi [a], M. Martello [a], I. Vigliotta [a], S. Armuzzi [a,b], B. Taurisano [a,b], E. Zamagni [a,b], M. Cavo [a,b], C. Terragna [a,*]

[a] *IRCCS Azienda Ospedaliero-Universitaria di Bologna, Istituto di Ematologia "Seràgnoli", Bologna, Italy*
[b] *Department of Specialized, Diagnostic and Experimental Medicine, University of Bologna, Italy*

ARTICLE INFO

*Article history:*
Received 21 April 2022
Received in revised form 28 June 2022
Accepted 28 June 2022
Available online 3 July 2022

*Keywords:*
Copy number alteration
Clustering methods
Multiple myeloma
Breast cancer
Bioinformatic pipeline
Data correction
Baseline region

ABSTRACT

Human cancer arises from a population of cells that have acquired a wide range of genetic alterations, most of which are targets of therapeutic treatments or are used as prognostic factors for patient's risk stratification. Among these, copy number alterations (CNAs) are quite frequent. Currently, several molecular biology technologies, such as microarrays, NGS and single-cell approaches are used to define the genomic profile of tumor samples. Output data need to be analyzed with bioinformatic approaches and particularly by employing computational algorithms.

Molecular biology tools estimate the baseline region by comparing either the mean probe signals, or the number of reads to the reference genome. However, when tumors display complex karyotypes, this type of approach could fail the baseline region estimation and consequently cause errors in the CNAs call. To overcome this issue, we designed an R-package, ***BoBafit***, able to check and, eventually, to adjust the baseline region, according to both the tumor-specific alterations' context and the sample-specific clustered genomic lesions.

Several databases have been chosen to set up and validate the designed package, thus demonstrating the potential of ***BoBafit*** to adjust copy number (CN) data from different tumors and analysis techniques.

Relevantly, the analysis highlighted that up to 25% of samples need a baseline region adjustment and a redefinition of CNAs calls, thus causing a change in the prognostic risk classification of the patients.

We support the implementation of ***BoBafit*** within CN analysis bioinformatics pipelines to ensure a correct patient's stratification in risk categories, regardless of the tumor type.



## 1. Introduction

Copy number alterations are among the most common features of human cancer genome and confer evolutionary advantage to the tumor cells, as the acquisition and/or the loss of specific locus(i) can lead to gene expression changes. In particular, since the activation of oncogenes and the inhibition of tumor suppressor genes can be a consequence of CNAs, these might promote both cell proliferation and resistance to apoptosis and to therapies [1-3].

Overall, tumor clones can have highly different genomic profiles (intra-tumor heterogeneity), with different specific CNAs prevailing in different tumors (inter-tumor heterogeneity): since the presence of specific CNAs or of specific CNAs profiles characterizes tumors' profile, an accurate description of the CNAs landscape provides important information for the disease staging [4-9]. Moreover, both deletions and/or amplifications of either the whole or part of chromosome arm can be used as prognostic factors in the clinical practice [10,11], defining subgroups of patients with different disease outcomes. A proper patients' stratification in risk categories allows the choice of the right and possibly personalized therapy for the patients [12].

*Abbreviations:* CNAs, Copy number alterations; CN, Copy number; NGS, Next Generation Sequencing; SNP, Single-Nucleotide Polymorphism; MM, Multiple Myeloma; CNVs, Copy Number Variations; BAF, B-allele frequency; WES, Whole Exome Sequencing; S-CL, Starting Chromosome List; F-CL, Final Chromosome List; CR, Correction Factor; FISH, Fluorescence In Situ Hybridization; HD, Hyperdiploidy; LOH, Loss of Heterozygosity; HR, High Risk; SR, Standard Risk; R-ISS, Revised International Staging System; WGD, Whole-genome doubling.

* Corresponding author.
*E-mail address:* carolina.terragna@aosp.bo.it (C. Terragna).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2022.06.062

Nowadays, two main molecular technologies allow the detection of genome-wide CNAs, i.e., microarrays and Next Generation Sequencing (NGS). By microarray technology (e.g. SNPs array), the CN changes are evaluated by measuring the mean intensities of the probes that bind to the genome and then by comparing the CN changes to a reference genome, removing all copy number variations (CNVs), common in the healthy population [13]. By NGS, after comparing the reads to the reference genome, the CN profiles are detected either by evaluating the read depth or by measuring the insert size deviation of the paired reads [14]. The output data of both methodologies need to be analyzed by specific computational algorithms, which estimate two parameters: the logR, i.e., the log2 of probes' intensities, and the B-allele frequency (BAF), i.e., the relative frequency of an allele; the two parameters are interdependent, and both define a distinctive pattern of signals describing the baseline regions. However, when these specific computational algorithms use the basic "median-centering" normalization method to estimate the baseline region, assuming that the average value corresponds to the theoretical "2", they might erroneously estimate the regions with diploid CN [14]. This might happen particularly in samples with a complex CN profile, either carrying several and/or large chromosomal aberrations, or presenting very fragmented profile. The incorrect setting of the baseline region leads to errors in the recognition of the sample's amplifications and deletions and, therefore, to a wrong estimation of the overall sample's aberrations profile.

Several bioinformatic tools have been designed to control this bias, mainly by estimating the tumor cells' purity and ploidy, thus leading to an increase of the sensitivity in the identification of CNAs [15,16], or by excluding healthy cells' alterations, as detected in germline samples, in order to get a better resolution of the signal [17]. However, these methods work mainly on raw data, which are not always available and require large amount of computational power and memory, particularly when large cohorts of samples need to be analyzed. To overcome this aspect, a new tool (Mecan4CNA [18]) acts directly on the log2ratio signal, by fixing the baseline region according to the deviation from the normal cell signal, assuming that the tumor cell fraction is at least 50%. Nevertheless, this approach might be limited by cases with very low tumor DNA, such as samples with few tumor cells or circulating free DNA.

Here we present a new R package, aimed at checking the estimated CN value of each chromosome and at recalculating the correct baseline region taking advantage from the patterns of both sample-specific and tumor-specific genomic alterations, thanks to the "clustering" and the "starting chromosome list" strategies, respectively. The package has been named "***BoBafit***" and contains *DRrefit* as main function, which can adjust the wrong baseline regions throughout a clustering method and the "chromosome lists", which include chromosomes that, in the analyzed tumor and sample, are commonly identified with clonal diploid CN (see section 2.3.1). The input data for the package are already computed segmentation files, deriving from both microarray and NGS platforms. Notably, ***BoBafit*** can adjust the baseline region of CN profiles coming from any type of tumors with complex karyotype due to high levels of CNAs; in the present study, it has been applied on 5 different tumors with high genomic complexity to evaluate and confirm the robustness of the method, highlighting the importance of the baseline region adjustment.

## 2. Materials and methods

### 2.1. Datasets

The following datasets have been used for the purposes of the present paper:

- M-M-BO dataset: this dataset includes SNPs array data, as obtained from a cohort of 595 MM patients, analyzed at the "Seràgnoli" Institute of Hematology of Bologna, after informed consent. SNPs array data have been analyzed to obtain the whole genome CNAs profiles [19], which have been used to develop the ***BoBafit*** package.
- CoMMpass dataset: this dataset includes 1044 MM samples, whose CNAs profiles were defined by NGS (Whole Exome Sequencing, WES). The dataset is part of a MMRF (Multiple Myeloma Research Foundation) study (NCT145429) [20], enrolling patients from Canada, Italy, Spain and the United States. CNAs profiles have been used after MMRF authorization and have been used to confirm the performances of ***BoBafit*** package on data produced by a different technology, but in the same clinical context, as compared to that employed to develop the ***BoBafit*** package.
- TCGA datasets: we downloaded from Cancer Genome Atlas Program [21] project 4 CN segment datasets of 4 different tumors with high CNAs load: breast cancer (TCGA-BRCA, 2133 samples, ovarian adenocarcinoma (TCGA-OV, 601 samples), lung adenocarcinoma (TCGA-LUG, 554 samples) and colon adenocarcinoma (TCGA-COAD, 504 samples). All CNAs profiles were defined by SNPs array (Affymetrix Genome-Wide 6.0). CNAs profiles are freely downloadable and have been used to confirm the performances of ***BoBafit*** package on data produced by the same technology, but in a different clinical context as compared to that employed to develop the ***BoBafit*** package.

### 2.2. SNPs array experiments

SNPs array experiments have been performed on bone marrow CD138 + enriched cells fractions, as collected from newly diagnosed MM patients; Affymetrix Cytoscan HD or GenomeWide6.0 have been used to obtain the SNPs array profiles, as elsewhere detailed [19]. Output CEL files have been processed by a pipeline including Rawcopy R package [13] and ASCAT [16] algorithms to compute sample's log2 ratio segments corrected for purity. The resulting log2 ratio signals were converted into CN values.

### 2.3. ***BoBafit*** implementation

The R package ***BoBafit*** includes three functions, which overall allow the refit and the recalibration of tumor samples' CN profile. All functions operate on segmentation BED files, derived either by NGS or by microarray data, which need to include the following five basic information related to the samples: sample ID, chromosome arm to which the evaluated segments belong, segment's start, segment's end, and CN value.

The principal refitting function is named *DRrefit*: throughout a tumor and sample-specific approach it adjusts the CN values. In addition, ***BoBafit*** contains two secondary functions, *ComputeNormalChromosome* and *PlotChrCluster*. The first one generates the "starting chromosome list" (S-CL), important input of *DRrefit* and cornerstone of the tumor-specific strategy (Fig. 1). The second one allows the chromosomes clusters' visualization in a plot, in absence of recalibration; *PlotChrCluster* might also potentially highlight the presence of sub-clones in the sample.

#### 2.3.1. The DRrefit function

To create a tumor and sample-specific method aimed at checking and adjusting the tumor CN profile, we developed the function *DRrefit*. It uses two inputs: (1) the BED file, including sample's genomic segments obtained from preceding genomic experiments and (2) the S-CL. This latter is a tumor-specific list of chromosomal arms considered "normal", as being commonly not affected by structural CNAs (e.g. "losses" and "gains" of single chromosomes

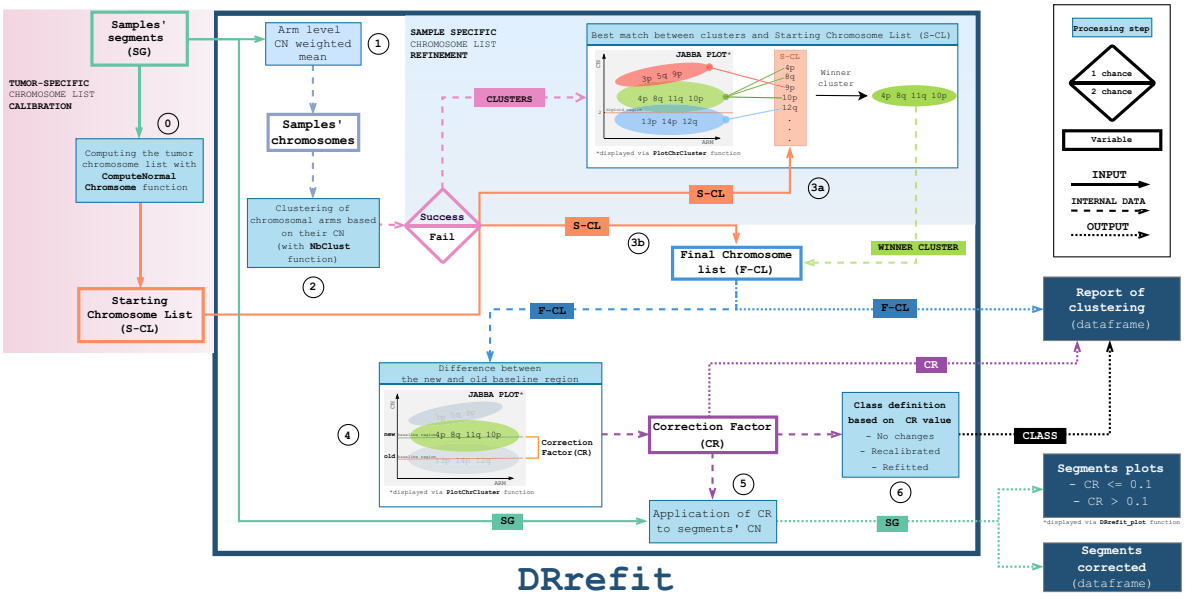


**Fig. 1. *BoBafit* package workflow**. The diagram shows how to organize a ***BoBafit*** analysis and *DRrefit* algorithm steps. 0) First of all, from the segmentation file, the tumor specific chromosome list has to be obtained by *ComputeNormalChromosome* function; 1) Next, the CN mean, weighted on the segments' length, is calculated for each chromosomal arm, thus obtaining the global arm CN. 2) NbClust package perform the clustering procedure based on CN of chromosomal arm. 3a) The clusters, obtained from the previously step, are compared to the S-CL, determining the "winner cluster" and the following F-CL. The JABBA plot, outputted by *PlotChrCluster* function, is used to illustrate how the comparison works. 3b) If NbClust fails the clustering, any cluster is available for the comparison and the S-CL remains the reference list. The S-CL directly becomes the F-CL. 4) At this point, the CR can be estimated as the difference between the old baseline region (CN = 2) and the median CN value of F-CL (new baseline region). Again, a JABBA plot shows the difference between the two baseline regions. 5) The segments CN values are corrected applying the Correction faction (CR), it returns three outputs: the "Report of clustering", where all information about the clustering procedure is reported; "Segments corrected", a data frame with the correct CN values of segments; and a sample plot where is possible to visualize the shift of the baseline region after the correction. 6) The CR value defines three class of profiles: No changes, Recalibrated and Refitted. That information is reported in the Report of clustering data frame.

or chromosome segments) in that specific tumor. The S-CL is used as tumor-specific reference for the possible re-adjustment of the baseline region. Since the S-CL might change according to the tumor type and/or subtypes, *DRrefit* allows accurate and specific control of the CN profiles call, even when obtained from different molecular platforms. To define S-CL, a specific function has been designed, named *ComputeNormalChromosome*, which is included in the ***BoBafit*** package (see below).

The algorithm performs the following steps for each sample (Fig. 1):

1. *Calculation of the CN value for each arm*: the algorithm selects all segments of the same chromosomal arm, then calculates the global arm CN, as the mean of the segments' CN weighted on the segments' length. The weighted mean is calculated for all chromosomal arms, excluding the X and Y chromosomes as they are not always diploid and therefore not helpful to the analysis. We have chosen to perform this simplification step, as it allows to reduce the CN segments profile to a simpler data structure, which results easier and faster to be computationally handled, in particular for the following clustering step. Additionally, providing most CN events happen either on whole chromosomes or on whole chromosomes' arms [1], this weighted mean approach consistently approximates the global chromosomal arm's CN.
2. *Clustering of chromosomal arms*: in order to cluster the chromosomal arms according to their similarities in terms of CN value, *DRrefit* takes advantage of NbClust [22], an R package that defines the best number of clusters resuming the overall chromosome distribution, according to the selected clustering method (e.g., either ward.D2, or complete, or average clustering methods). According to this clustering process, two possible outcomes can be obtained: either a reference list refinement or no reference list change.
3. *Comparison to the S-CL*: (a) The clustering process succeeds, and the clusters are compared to S-CL. The cluster that best matches with S-CL (i.e. the one that has the highest number of chromosomal arms in common with S-CL, Fig. 1), is chosen as the "winner cluster" and it becomes the "final chromosome list" (F-CL). This step defines the "sample-specific refinement" (Fig. 1) of the S-CL, taking into account the intra-tumor heterogeneity phenomenon, as the "winner cluster" includes the baseline and clonal chromosomal arms of the analyzed sample. This is shown by the JABBA plot (obtained by the *PlotChrCluster* function, see section 2.3.3), included in Fig. 1. The plot also shows the correspondence between clusters and the different clonal or sub-clonal CN states. (b) Due to the failure of some statistical indices used by NbClust, see the vignette of the package [22], for a small proportion of samples the chromosome clustering process fails. In this case, the sample will not present clusters and the sample-specific refinement will not be performed. As a consequence, the S-CL directly becomes the F-CL. In these rather infrequent situations (about 6.8% of samples, depending on segmentation quality) the baseline region adjustment is only tumor-specific, and the report of the sample gains a "failed clustering" label.
4. *Definition of a correction factor*: From F-CL, a correction factor (CR) is calculated. The CR highlights the differences between the baseline region assessment before and after *DRrefit* calculation and corresponds to the difference between 2 (the theoretical diploid value) and the median CN value of F-CL (Fig. 1).

5. *Samples correction*: all segments' CN are corrected for the CR, moving the CN profile to the most likely CN state of that specific sample. The resulting CN profile is shown both in the CN profile plot (Fig. 1) and in the two data frames outputted by *DRrefit*, described in the package vignette [23]. The function returns either one of two possible plots, according to the effective repositioning of the samples' CN profiles and its CR absolute value: (1) the "CR > 0.1" plot, with either green or red colored segments, highlighting segments' distance (Fig. 2 B and C); (2) the "CR $\leq$ 0.1" plot (Fig. 2 A), with overlapping segments.
6. *Class definition*: Based on the CR value, three class of sample are defined:
   - **no changes** (CR $\leq$ 0.1): the new segments overlap the old ones and the CN remains the same, or undergoes a minimal change that doesn't alter significantly the baseline region (Fig. 2 A);
   - **recalibrated** (0.1 < CR $\geq$ 0.5): the new segments positions are different from the old ones, even though the differences between the new and the old CN are not as much significant to impact the overall CN profile (Fig. 2 B);
   - **refitted** (CR > 0.5): the new segments positions are different from the old ones and the CN profile markedly changes, as compared to the original one (Fig. 2 C).

#### 2.3.2. The ComputeNormalChromosome function

*ComputeNormalChromosome* is a secondary function of the ***BoBafit*** package and can be used to define the S-CL. The input BED file consists in a cohort of samples of the same tumor type. As in *DRrefit*, the first step of this function calculates the global arms' CN, as the mean of the segments' CN weighted on the segments' length.

Then the function computes the frequency of alteration of each chromosomal arm among the whole cohort of samples. Only the chromosomal arms which present an alteration frequency below a specific threshold (tolerance value) are selected to create the list of normal chromosomes (S-CL). Since the input consists in samples of the same tumor type, this approach is defined "tumor specific calibration", and takes into consideration the phenomenon of inter-tumor heterogeneity.

*ComputeNormalChromosome* allows to set the tolerance value (expressed as percentage). The tolerance values can range from 5% (stringent analysis) to 20–25% (permissive analysis). The minimum and maximum CN thresholds to define an alteration in each sample can be set according to user requirements.

Finally, the function draws a histogram, showing the alteration rate of chromosomes included in the S-CL (blue bars) (Fig. 3) and stores the output S-CL in a vector, ready to be used as *DRrefit*'s input (Figure 1).

#### 2.3.3. The PlotChrCluster function and the ***JABBA*** plot

*PlotChrCluster* is another secondary function of the ***BoBafit*** package, which can be used to visualize the sample's CN clusters with a specific type of plot (called "JABBA plot"), which allows to have a look to the *DRrefit* process and, additionally, can help the interpretation of clonal and sub-clonal CN states in the tumor sample, in a simple and intuitive way.

This function repeats the two initial steps described in the *DRrefit* (i.e. 1: "Calculating CN value for each arm" and 2: "Clustering of chromosomal arms") (Fig. 1). Afterwards, it creates a visual representation of the clustered chromosomal arms in the JABBA plot. The sample's clusters are pictured as ellipses, where the area corresponds to the cluster confidence interval (Figure 4).

Unlike the main function of the package, *PlotChrCluster* does not modify the segments' CN, but just explore the data and visualize the quality of the clustering procedure.

### 2.4. ***BoBafit*** package settings

The CN thresholds, used as parameters for *ComputeNormalChromosome* function, were set at 2.40 and 1.60 for single copy gain and single copy loss, respectively.

For each dataset that has been analyzed, a specific tumor chromosome list was generated with the *ComputeNormalChromosome* function, starting from the segmentation files. Notably, TCGA-OV, TCGA-LUAD and TCGA-COAD datasets had higher tolerance rates as compared to the MM and TCGA-BRCA ones, since in the CNAs profiles of these tumors, any chromosome had an alteration rate within 15%. Moreover, the TCGA-OV's S-CL was further manually revised by including literature information [26] (Table 1).

The highest number of clusters considered acceptable by *DRrefit* function has been set to 6. "Ward.D2" was used as clustering method by *DRrefit*, since it minimizes the total within-cluster variance.

## 3. Results and discussion

The feasibility of ***BoBafit*** in both refitting and recalibrating CN data has been tested by using six genomic databases including tumor CN profiles, as obtained by different molecular technologies (e.g., SNPs array and NGS) and tumor samples.

### 3.1. Multiple Myeloma datasets

#### 3.1.1. Myeloma-specific S-CL generation and validation

The S-CL generation is the first step of the ***BoBafit*** algorithm set-up. Two MM-specific S-CLs have been generated by applying *ComputeNormalChromosome* with a 15% tolerance rate, chosen in order to balance the S-CL length and the need to include chromosomal arms with the lowest probability of carrying CNAs (Table 1). As shown in Fig. 5 A and B and in Table 1, the S-CLs of MM-BO and CoMMpass datasets were almost superimposable; in addition, the chromosomal regions included in the two S-CLs were in agreement with data of the literature [4–6,9]. The two MM S-CLs pertinence was also validated by manually reviewing the IGV (Integrative Genomics Viewer) CN density plots which identify the chromosomes with the lowest amount of CNAs in the analyzed datasets. Results showed that the same chromosome regions included in the S-CLs (Table 1 and Fig. 5) were observed also in the related IGV CN density plots, thus confirming the power of *ComputeNormalChromosome* to correctly estimate lists of diploid and clonal chromosomes, that can be used as reference in *DRrefit*. Overall, we showed that *ComputeNormalChromosome* was able to generate tumor-specific S-CLs, which, once generated, could be used to analyze any datasets derived from the same tumor, providing the CN data are calculated by median-based algorithms. Importantly, the S-CL should be generated using a dataset large enough to represent the heterogeneity of the disease.

#### 3.1.2. Application and validation of DRrefit on a MM dataset

*DRrefit* was first tested in the MM-BO dataset, including CNAs data derived from SNPs array experiments.

As shown in Fig. 6, overall, 19 out of 595 samples (3.2%) have been refitted and 10 (1.9%) have been recalibrated. To prove the validity of these results, for each sample of the MM-BO cohort, we compared the CN baseline regions, pre- and post-correction, to the corresponding BAF signals: concordant signals, either pre- or post-correction, were considered suggestive of a correct definition of the baseline region.

The comparison showed that, in the pre-correction analysis, 20 samples had BAF and CN discordant values (Supplementary data1), whereas after *DRrefit* correction, just 2 samples remained discor-

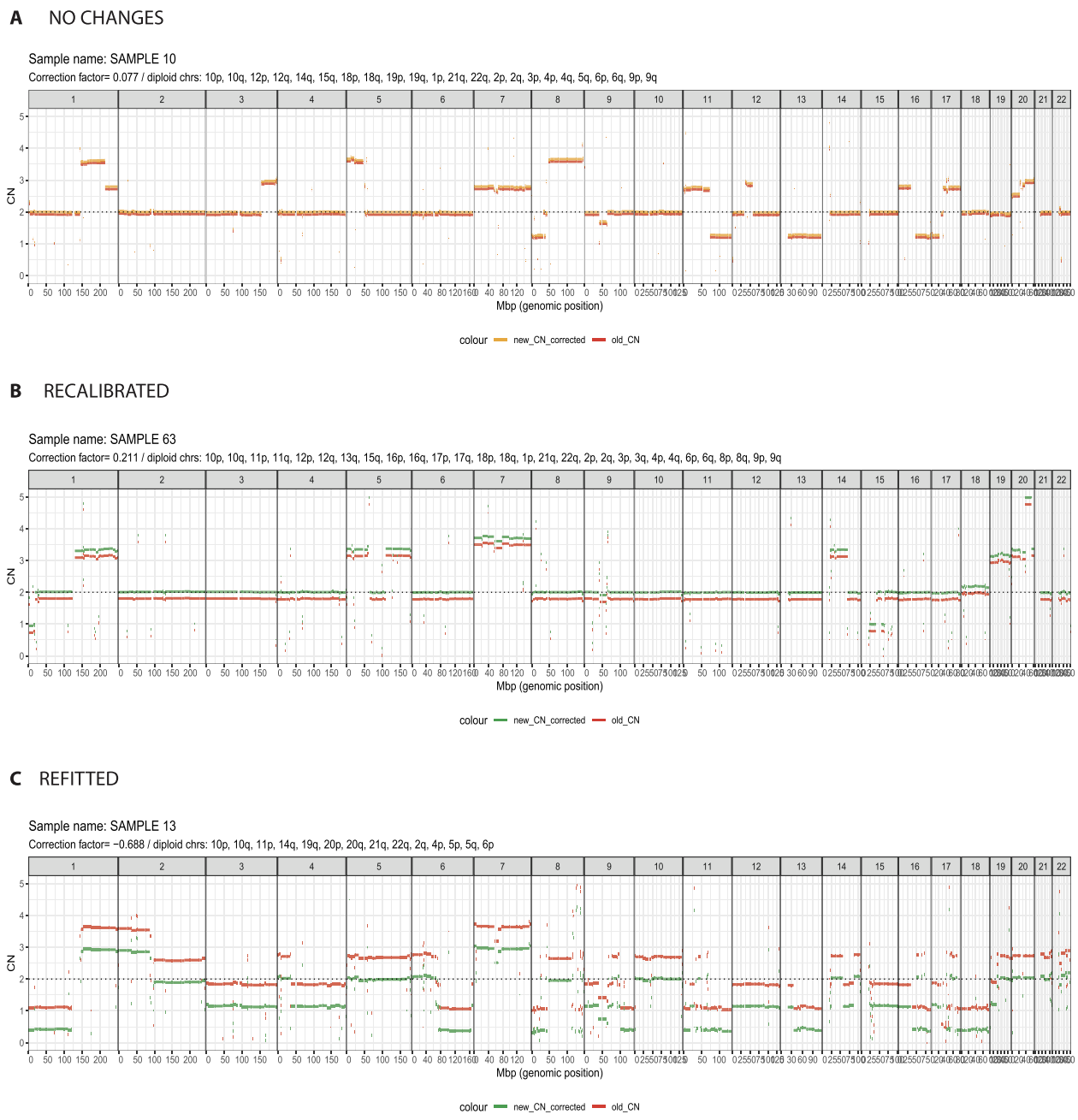


**Fig. 2.** ***DRrefit*** **CN profile plots of** three **samples, labeled with class identified by function**. In the panel are showed the tree *DRrefit* classes and how they are plotted. The x-axis reports the chromosomes with their genomic position and the y-axes the copy number value. The plots with CR $\leq$ 0.1 show that the new segments and the old segments are orange and red colored, respectively; on the contrary, the plots with CR > 0.1 show that the new segments and the old segments are green and red colored, respectively. **a) No Changes** class with CR 0.0077; **b) Recalibrated** class with CR 0.2; **c) Refitted** class with CR −0.688. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

dant. Of these, the first one had very low BAF quality parameters (MM-BO_24), the second one carried a loss of heterozygosity (LOH) event on chromosome 1p, following an amplification event on the same chromosomal arm (MM-BO_25). ***BoBafit*** cannot recognize LOH events because it uses the CN values, whereas BAF is needed to highlight LOH events.

Moreover, a manual review of all samples with *DRrefit*-detected discordances (20 samples) was performed, by comparing raw CN profiles and *DRrefit* plots, confirming the goodness of the refitting process.

Finally, *DRrefit* was applied in a cohort of 102/595 samples with FISH data available. While in 70/102 samples the CNAs FISH and SNP array calls were in agreement both pre and post-correction, in 26/102 the pre-correction discrepancies were resolved by *DRrefit*, whereas in 6/102 the refit led to new small discrepancies between the two types of calls. Notably, in these last 6 cases, the CNAs were subclonal and this pitfall will be further discussed in chapter 3.1.5.

Overall, our results showed that CN profiles, as obtained by SNPs array data might be biased by an incorrect baseline region setting. Therefore, the use of *DRrefit*, which can correctly estimate the baseline region, would help to overcome this issue. This was validated by comparing *DRrefit* refitted data both to BAF values and to FISH results. The limit of the method remains LOH events,

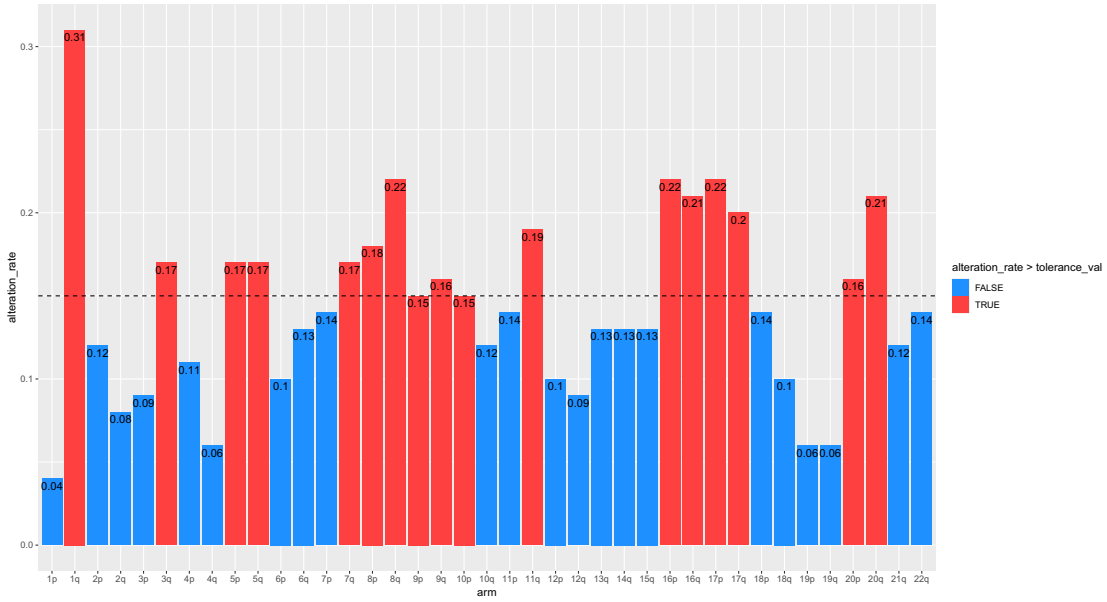


**Fig. 3. Starting Chromosome list (S-CL) plot**. In the x-axes are reported the chromosomal arms and in the y-axes the alteration rate. Each bar reports the alteration rate of each arm. The dotted line indicates the chromosomal alteration rate's tolerance value, below that the chromosomes are selected for the chromosome list as they are little altered. In this plot the chromosomal arms with a tolerance value less than or equal to 15% were selected (it is also the default value of the function).


**Fig. 4. JABBA Plot of a multiple myeloma sample made by *PlotChrCluster*.** The sample's clusters are pictured in as ellipses, where the area corresponds the cluster confidence interval. The chromosomal arm labels are colored based on the cluster which they belong. The x-axes report the CN value and it's possible to see how clusters correspond to different both clonal and sub-clonal CN states.

since, by using just the CN signal, the alleles contribution cannot be discriminated, thus impairing the chromosomes' inclusion among those suitable for the diploid region check.

### 3.1.3. Stability of the DRrefit corrections

*DRrefit* performances depend on the *NbClust* clustering tool, which have been employed to ultimately define the "winner cluster" and then the F-CL (see section 2.3.1). We therefore aimed at evaluating whether different clustering methods (other than ward.D2, which have been used in the present analysis) might impact the results. To this purpose, we analyzed all datasets using two different clustering methods (complete and average) and then compared the samples' correction factors to those obtained with the ward.D2 method. The comparisons showed a complete

**Table 1**
**Starting Chromosome lists (S-CLs) obtained by *ComputeNormalChromosome*.** The function was applied to all database with different tolerance values The tolerance rates have been chosen in order to include a sufficient number of chromosomal arms with the lowest probability of alteration. The TCGA-OV chromosome list was further revised due to the high percentage of alteration per chromosome (minimum 47% - see Fig. 5) in the dataset so it is different from the function output.

| Database | Chromosome list | Tollerance rate |
|---|---|---|
| MM-BO | 1p, 2p, 2q, 4p, 4q, 8q, 10p, 10q, 12p, 12q, 16p, 17p, 17q, 20p, 20q, | 15% |
| CoMMpass | 1p, 2p, 2q, 4p, 4q, 8q, 10p, 10q, 12p, 12q, 16p, 17p, 17q, 18q, 20p, 20q, 22q | 15% |
| TCGA-BRCA | 1p, 2p, 2q, 3p, 3q, 4p, 4q, 9q, 10p, 10q, 11p, 11q, 12q, 14q, 15q, 19p, 19q, 21q | 15% |
| TCGA-OV | 1p, 3p,7p, 7q, 11p,11q, 14q, 21q | 50% |
| TCGA-LUAD | 1p, 32q, 4q, 10q, 11p | 20% |
| TCGA-COAD | 1p, 2p, 2q, 3p, 3q, 6q, 10p, 10q, 11p, 11q, 19p | 20% |

equivalence of the three approaches, as the correction factors were exactly the same, thus demonstrating the method reliability, regardless from the clustering approach employed (Supplementary data2).

### 3.1.4. DRrefit on a second MM dataset analyzed by NGS

In order to confirm the *DRrefit* power to refit CNAs profiles, a second MM dataset was employed (CoMMpass dataset), including MM genomic data, as obtained by NGS technology.

As shown in Fig. 6, overall 7 out of 1044 samples (0.7%) have been refitted and 76 (7.3%) have been recalibrated. Since the CoMMpass genomic data analysis pipeline includes the tool *tCoNut* [17], which is specifically aimed at correcting the baseline regions, a small, even though appreciable, number of samples remained to be refitted by *DRrefit*. This indirectly confirmed the specificity of *DRrefit*, since any false positive was highlighted within samples already corrected for ploidy. Moreover, since 7.3% of data still needed to be recalibrated, we showed that *DRrefit* might even improve the baseline region estimation, both by using the clustering method and by implementing the S-CL. Of note, *DRrefit* correction can be performed without the need of BAF values or germline samples' profiles (both required by *tCoNut* tool [17]).

In conclusion, we showed that *DRrefit* is able both to recalibrate and refit data, even obtained by NGS technologies, and to produce a baseline regions' output comparable to that produced by an already validated method.

### 3.1.5. Clinical relevance in Multiple Myeloma

CNAs are considered important prognostic factors for many tumor types. Therefore, the right definition of baseline regions is crucial, in order to correctly call CNAs.

To confirm the clinical relevance of correct CN calls, we compared pre- and post-refit CNAs profiles of patients included in the MM-BO dataset. We focused on five alterations (1q amplification, 1p, 13 and 17p deletions, and odd-numbered chromosomes Hyperdiploidy- HD - as defined by the presence of at least two amplificated chromosomes among chr 3, 5, 7, 9, 11, 15, 19 and 21), whose prognostic role has been repeatedly demonstrated [4,6,9,24]. Each CNAs should be present in at least 10% of sample's cells (amplification CN $\geq$ 2.10 and deletion CN $\leq$ 1.90), except for HD, which should be clonal (i.e., present in at least 50% of cells, CN $\geq$ 2.50).

Venn diagrams of pre- and post-correction CN profiles highlighted that the frequency of most alterations and of their co-segregation changed after the ***BoBafit*** correction process (Fig. 7A and B): for example, HD patients were 84 pre- and 112 post-correction process, respectively.

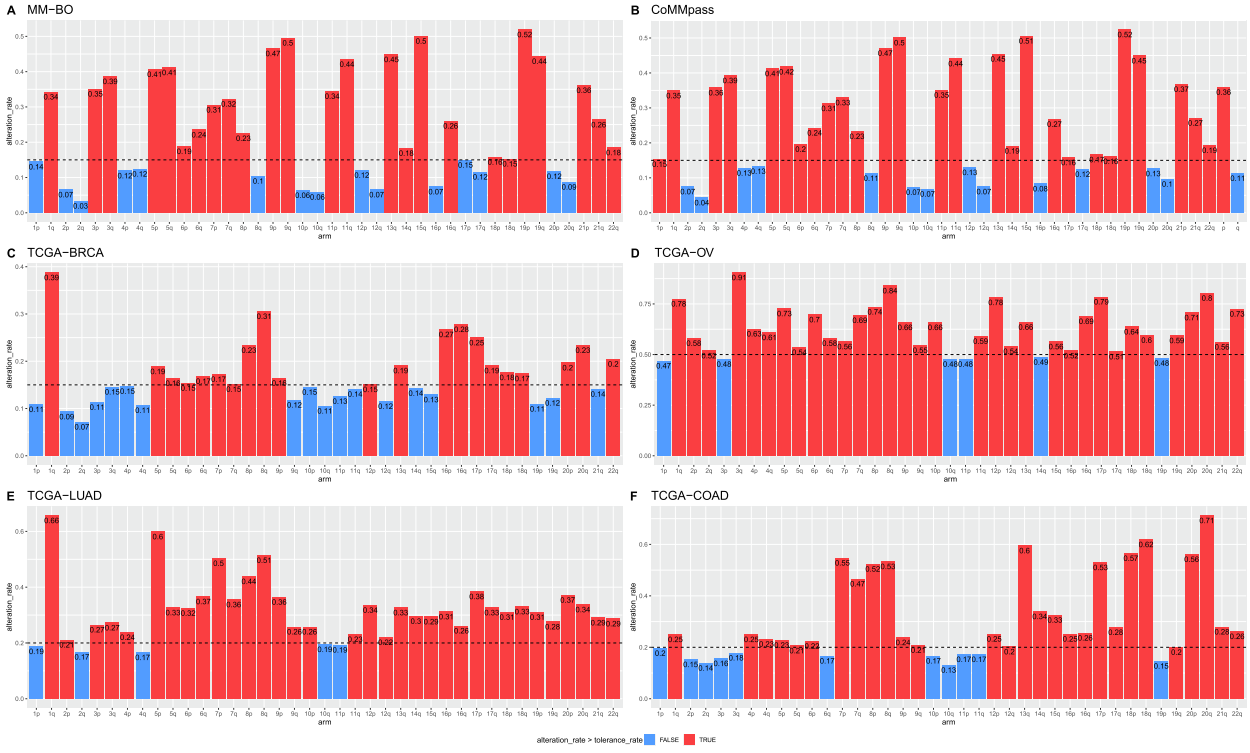


**Fig. 5. Starting chromosome list(S-CL) plots**. In the panel are showed the S-CL histograms of *ComputeNormalChromosome* once performed on each database. In the y-axes is reported the alteration rate of the chromosomal arm and the dotted line highlight the tolerance rate below which the chromosomes are considered "normal". **A) MM Bologna** S-CL, tolerance rate = 0.15**; B) the CoMMpass** S-CL, tolerance rate = 0.15**; C) the TCGA-BRCA** S-CL, tolerance rate = 0.15**; D) the TCGA-OV** S-CL**,** tolerance rate = 0.50; **E) the TCGA-LUAD** S-CL, tolerance rate = 0.20; **F) the TCGA-COAD S-CL**, tolerance rate = 0.20.

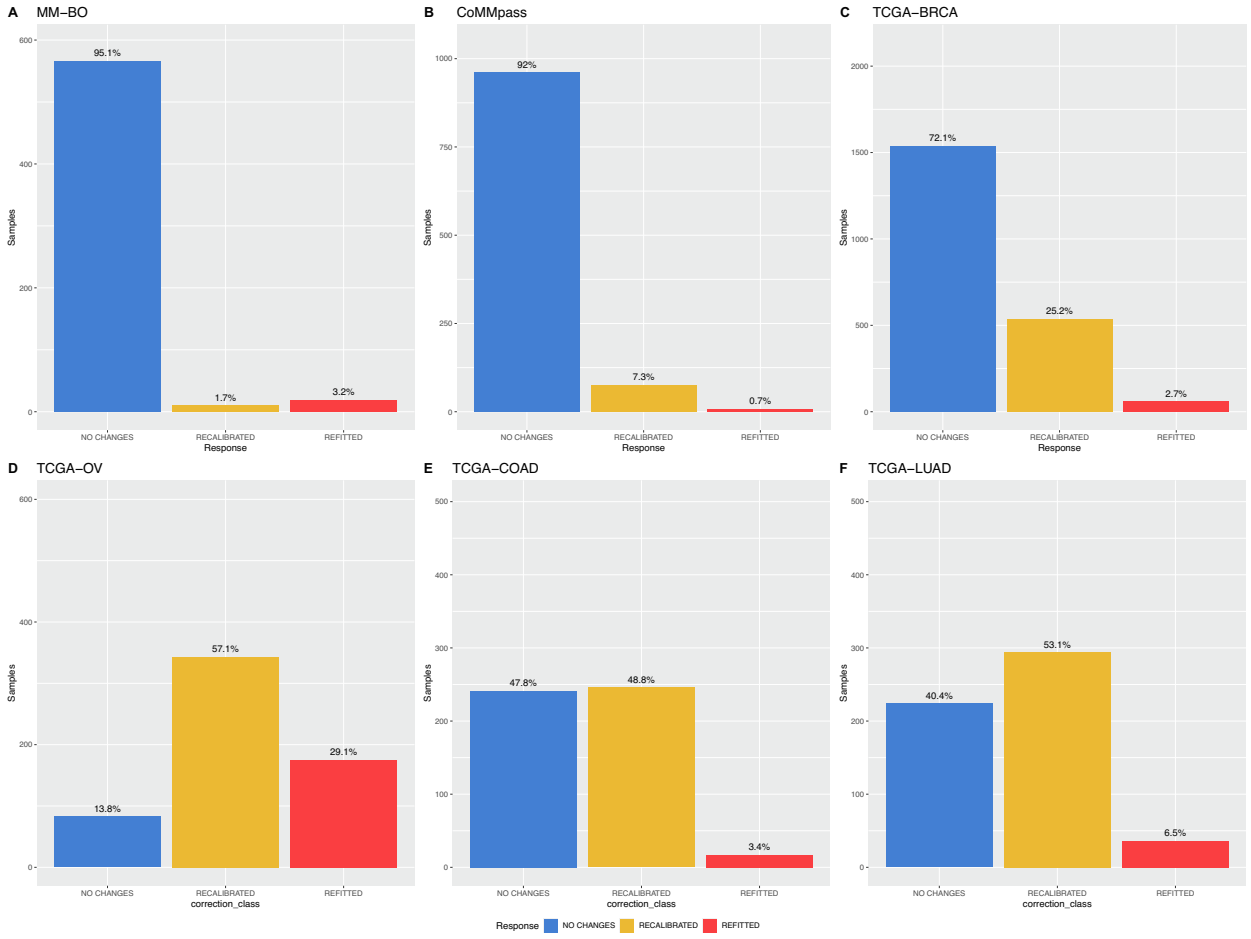


**Fig. 6. Summary of *BoBafit* results performed on MM Bologna, CoMMpass and all TCGA databases**. The histograms report, for each database analyzed in the present study, the percentages and the number of samples (x-axis) belonging to the three ***BoBafit***'s classes (Refitted, Recalibrated and No changes) **A) MM- BO samples; B) CoMMpass samples; C) TCGA-BRCA samples; D) TCGA-OV samples; E) TCGA-LUAD samples; F) TCGA-COAD samples.**

In order to describe each sample's refitting trajectory, all pre- and post-correction groups generated by the Venn diagrams were used to define the trajectory's starting and ending points (Fig. 7C). As expected, the refitting trajectory remained stable for most samples (438/595, 73.6%), whose baseline region was not adjusted by *DRrefit* correction. On the contrary, in a remarkable number of patients (157/595, 26.4%), a shift from one group to another was observed. In particular, both HD and 1q amplification were the most frequently misclassified CNAs (Fig. 7C), thus affecting their overall frequency, either as single or as co-occurring alterations. These results showed that in more than a quarter of patients the definition of clinically relevant prognostic factors was not accurate.

Since 17p deletion and, more recently also 1q amplification, have been included in the most frequently employed MM risk scoring systems [10,25], their correct CN call is crucial for patients' prognostic stratification. Therefore, we checked how the ***BoBafit*** correction process may affect the assignment of patients to the different risk categories, as defined by the Revised International Staging System (R-ISS) [10] and/or the mSMART guidelines [25]: in particular, according to the presence of 17p deletions and/or 1q amplification, patients were stratified in High (HR) or Standard Risk (SR), (Table 2).

In 29/595 (4.87%), the 17p deletion was corrected (Fig. 7 A and B), thus causing a transition from R-ISS HR to SR and from SR to HR in 27 and 2 samples, respectively, (Table 2 and Fig. 8). The results of 17p deletion calls' corrections were compared to FISH data (available in the context of the daily clinical practice for 102/595 included in the study), in order to confirm the corrections' accuracy. In 23/29 samples, the corrections allowed to get the same results provided by FISH analysis. On the contrary, in the remaining 6 cases (also mentioned in chapter 3.1.2), results were discordant, despite the CN calls correction. In these cases, 17p deletion was sub-clonal, with frequencies very close to either FISH or SNPs array CN detection cut-offs, thus impairing the correct CN call by SNPs array, not rectifiable even after the ***BoBafit*** correction. Notably, most cases have a "no changes" profile and very small CR (<0.1), leading to a slight CN value shift around the pre-defined SNPs array cut-off (Supplementary data 3).

Concerning the mSMART stratification, minor changes in the SR and HR samples were highlighted post ***BoBafit*** correction (Table 2 and Supplementary Fig. S2).

These results support the importance to accurately call these critical alterations, in order to correctly stratify patients according to their prognostic risk. Mostly, it will become important in case MM patients, enrolled in risk-oriented clinical trials, would be treated differently according to their prognostic risk score.

### *3.2. DRrefit application on the CN profiles of different tumors*

*DRrefit* was finally tested on datasets derived from solid tumors, in order to check its performances on highly fragmented and complex genomic profiles. Four solid tumors' datasets were down-

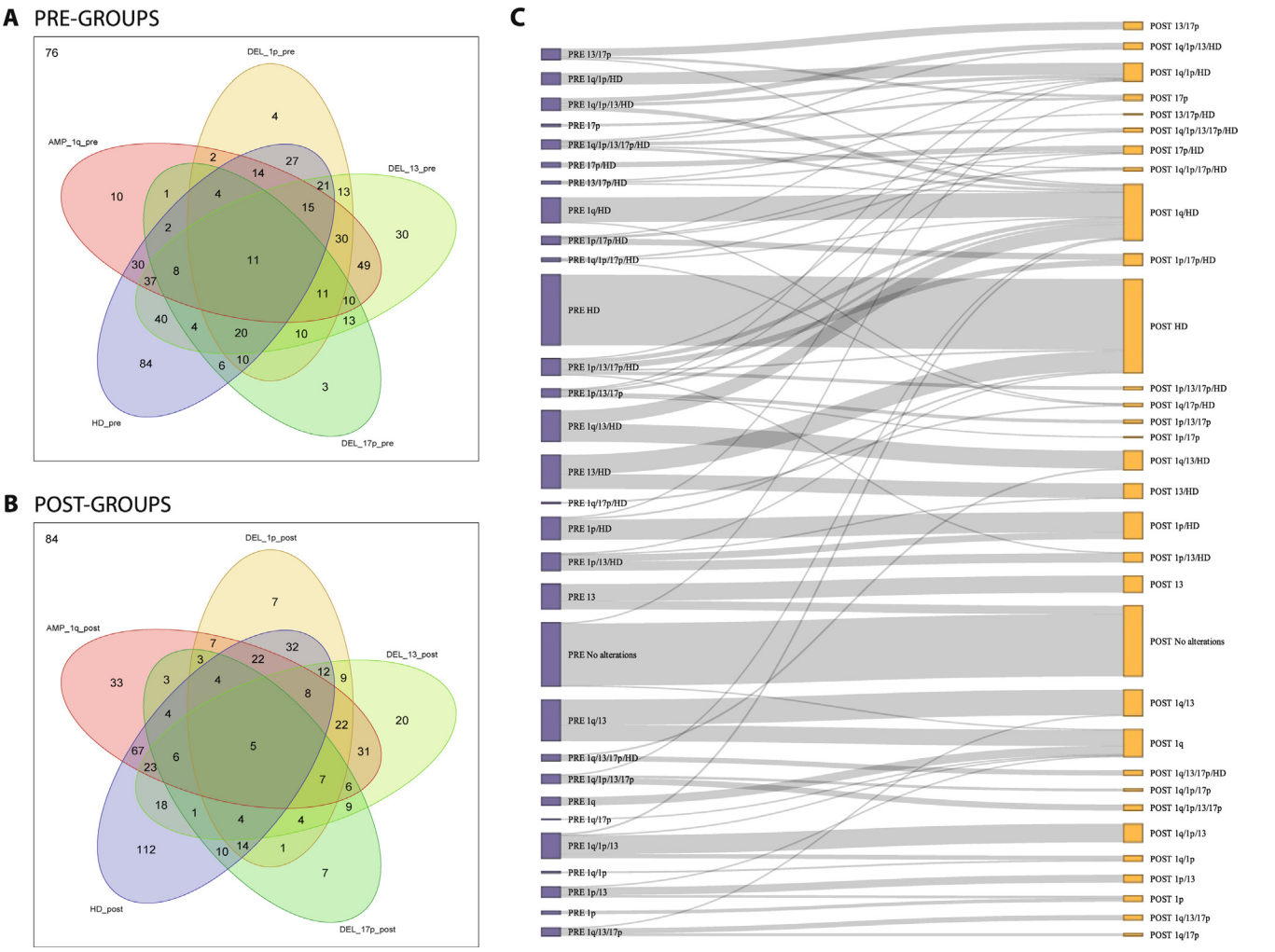


**Fig. 7. Clinically relevant alterations pre and post *BoBafit* correction in MM-BO samples. a) Five-way Venn diagram of pre-correction alterations and b) Five-way Venn diagram of post-correction alterations,** which allow to appreciate the number of samples that belongs to each alteration group. In the top left are indicated the number of samples without alterations**; c) Sankey network diagram**, on the left are represented the starting alteration groups (pre-correction, purple) and on the right the final alteration groups (post-correction, yellow). The gray bands indicate the flow of the samples, some remain in the starting group while others acquire / lose alterations and change their alteration group. The thickness of the line changes according to the number of samples in the trajectory. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

loaded from the TCGA project, namely TCGA-BRCA, TCGA-OV, TCG-COAD and TCGA- LUAD and, for each dataset, a tumor-specific S-CL was generated by *ComputeNormalChromosome* function (Table 1).

**Table 2**
MM-BO samples stratified according to R-ISS and mSMART cytogenetic guidelines pre and post ***BoBafit*** correction.

| **R-ISS** | | |
|---|---|---|
| **PRE** | **POST** | **SAMPLES** |
| pre HR | post HR | 86 |
| pre HR | post SR | 27 |
| pre SR | post HR | 2 |
| pre SR | post SR | 480 |
| High risk (HR) = presence 17p deletion; Standard risk (SR) = no deletion | | |
| mSMART | | |
| PRE | POST | SAMPLES |
| pre HR | post HR | 297 |
| pre HR | post SR | 3 |
| pre SR | post HR | 4 |
| pre SR | post SR | 291 |
| High risk (HR) = presence 17p deletion and/or 1p amplification; Strandard risk (SR) = none of the two. | | |

Due to their higher genomic complexity, the tolerance rates employed for the generation of the S-CL of TCGA-OV, TCG-COAD and TCGA- LUAD were higher, as compared to those of MM and TCGA-BRCA datasets: in fact, for these tumors the lowest tolerance value between chromosomal arms was over 15% (Fig. 5). For the TCGA-OV dataset the tolerance rate was increased to 50%, and the output data were revised and corrected according to disease-related published data [26] (Table 1), since the use of a very high tolerance rate might increase the probability to include false references in the S-CL, thus limiting the applicability of *DRrefit* in cancers with very high frequencies of alterations *per* chromosome arm. In these cases, either the S-CL list should be manually generated, or the function output should be revised, with the support of disease-related published data.

Once the S-CLs were generated and reviewed, *DRrefit* was applied to the segmentation files of the cancers' datasets: all TCGA datasets had higher percentage of both refitted and recalibrated alterations, as compared to MM datasets (Fig. 6). This was mostly due to the high number of CNAs that characterizes all the analyzed tumors, causing an overall baseline region distortion; in addition, we observed that TCGA CN profiles were overall more fragmented, as compared to those obtained from MM (Supplementary data 4),

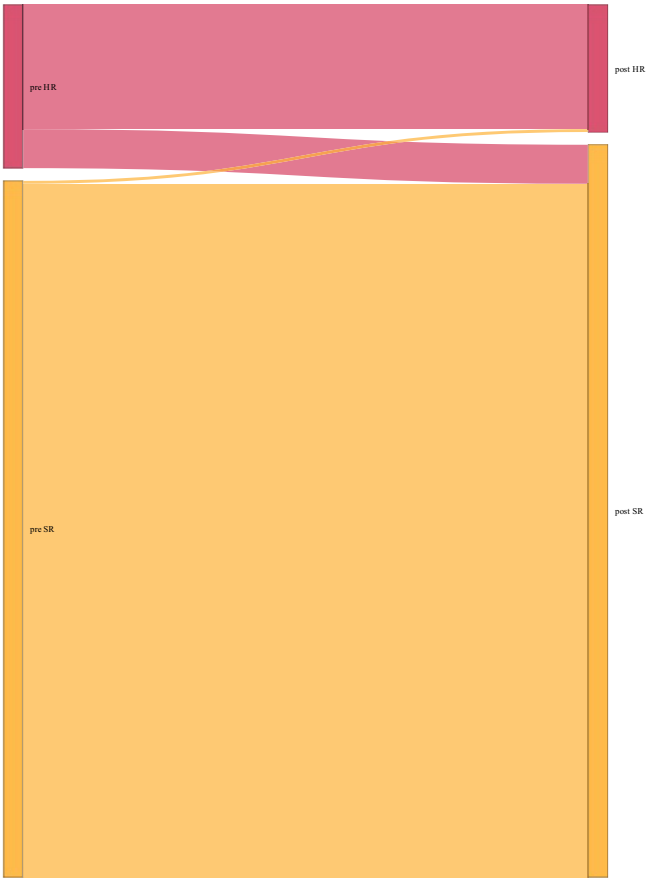


**Fig. 8. Sankey Network diagram of MM-BO samples**. The diagram shows how samples change risk class from the start profile (pre) and end profile (post) according to R-ISS cytogenetic guidelines.

thus significantly impacting the correct baseline region definition (p value < 2e-16, Supplementary data 5).

These results show that ***BoBafit*** can be employed to correctly estimate the baseline regions of CN profiles derived from a wide range of genomically heterogeneous diseases, even those with very fragmented and low-quality CNAs profiles.

## 4. Conclusions

By analyzing genomic datasets including data related to five types of tumors with complex karyotypes and obtained by using two different molecular technologies (SNPs array and NGS), we have been able to show the benefit of ***BoBafit*** as implemented in the analysis workflow of tumor samples' CNAs profiles and the reliability of *DRrefit* in correcting the baseline region assignment. In particular, the implementation of a S-CL within ***BoBafit*** allowed the use of this analysis pipeline regardless from the tumor type, thus highlighting the universality of this bio-informatic approach. Moreover, the use of ***BoBafit*** pipeline has been shown to be simple and reproducible, as it requires just CN data as input, without the need of either germline samples, or BAF values, as against other similar tools (e.g., *tCoNut* [16]). By employing chromosomes arms instead of raw segments as basic information unit, ***BoBafit*** has also the advantage to be computationally manageable and applicable to data already processed. Particularly, the generation of S-CL from clinical knowledges and the visualization of both clonal and subclonal CN clusters for any tumor sample (JABBA plot), both make ***BoBafit*** very user-friendly.

We observed that, overall, CN profiles from samples of tumors with high numbers of alterations (e.g. OV cancer) and with low quality parameters (e.g. highly fragmented) need baseline region adjustment. In these cases, a wrong baseline region estimate might lead to a bias in patients' prognostic risk assignment based on the presence of chromosomal aberrations, thus supporting the need of a bio-informatic adjustment of CNAs output data. By refitting and recalibrating CNAs results, as derived from high throughput molecular approaches' data, ***BoBafit*** significantly reduces the incorrect estimate of whole-chromosome, arm-level and focal CNAs, thus leading to correct CNAs calls.

Of note, we are aware that both Whole Genome Doublings (WGD) and Loss of Heterozygosity (LOH) events, by altering the tumor ploidy, remain bias that cannot be corrected by ***BoBafit***,. In fact, in bulk analyses, when WGD occurs, the CN state appears equivalent to itself with the CN state doubled [27], thus impairing WGD events to be considered; on the contrary, to detect LOH events, allele frequency data are needed. The total CN signal is the only data required for the analysis by ***BoBafit***, which therefore possess this intrinsic ambiguity in its results. However, our approach does not claim to solve these issues, which in turn are well-managed by other tools, specifically designed to these aims (ASCAT [16], ABSOLUTE [14], FACETS [28]). On the contrary, we support the application of ***BoBafit*** to raw segments (fractional CN values) generation either by ASCAT [16] or by any other segmentation tools.

In conclusion, we propose the implementation of ***BoBafit*** as crucial step of the standard CN analysis pipelines for data derived from all type of molecular platforms, particularly in the daily clinical routine analysis, in order to guarantee an unbiased patient's stratification, based on the CNAs prevalence in the examined population. Including ***BoBafit*** can only become an advantage, as the computational time for the diploid correction is very minimal and does not affect costs, as the R packages is currently available for download from the Bioconductor open-source repository [23].

## Availability of data and materials

- MM Bologna
  The data underlying this article cannot be shared publicly due to for the privacy of individuals that participated in the study.
- CoMMpass
  The data underlying this article were provided by Multiple Myeloma Research Foundation (MMRF) by permission and downloaded at https://research.themmrf.org. The dataset version of analyzed data is IA13a.
- TCGA-BRCA
  The data underlying this article are available in National Cancer Institute (NIH) GDC Data portal, at https://portal.gdc.cancer.gov/projects/TCGA-BRCA.

## Author contributions

Conception and study design: G.M., A.P.; Experimental analyses: G.M., A.P and V.S.; Manuscript preparation: G.M., A.P., V.S and C.T.; Statistical analysis G.M., A.P., and V.S. M.M, E.B., I.V., S. A. and B.T. provided MM genomic data. E.Z. and M.C. provided MM clinical data. G.M., A.P and V.S. and C.T. discussed and interpreted data. All authors implicated in the present study have read and agreed to the last version of the submitted manuscript.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We gratefully thank the Multiple Myeloma clinical group of Institute of Hematology "L. and A. Seràgnoli": Lucia Pantani, Paola Tacchetti, Serena Rocchi and Katia Mancuso, who provided the medical support and the clinical risk classification information for the patients included in this study. We also thank the cytogenetic group of Institute of Hematology "L. and A. Seràgnoli": Nicoletta Testoni and Giulia Marzocchi who performed the FISH experiments and provided the FISH results data used in this study.

## Funding

Associazione Italiana per la Ricerca sul Cancro (AIRC, Grant Awards: IG 15839, IG 22059), Ministero della Salute (Grant awards: RF-2016-02362532, RC- 2773350), Associazione Italiana Contro le Leucemie - Linfomi e Mieloma (AIL BOLOGNA ODV).

Lucia Pantani, MD at the Institute of Hematology "L. and A. Seràgnoli", in Bologna, Italy.

Paola Tacchetti, MD at the Institute of Hematology "L. and A. Seràgnoli", in Bologna, Italy.

Serena Rocchi, MD at the Institute of Hematology "L. and A. Seràgnoli", in Bologna, Italy.

Katia Mancuso, MD at the Institute of Hematology "L. and A. Seràgnoli", in Bologna, Italy.

Nicoletta Testoni, Associate Professor at the University of Bologna, Italy.

Giulia Marzocchi, PhD, a senior fellow at the University of Bologna, Italy.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.062.

## References

[1] Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers Published online. Nature 2010. https://doi.org/10.1038/nature08822.
[2] Stratton MR, Campbell PJ, Futreal PA. The cancer genome Published online. Nature 2009. https://doi.org/10.1038/nature07943.
[3] Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet 2013;45(10):1134–40. https://doi.org/10.1038/ng.2760.
[4] Cardona-Benavides IJ, de Ramón C, Gutiérrez NC. Genetic Abnormalities in Multiple Myeloma: Prognostic and Therapeutic Implications. Cells 2021;10 (2):336. https://doi.org/10.3390/cells10020336.
[5] Barwick BG, Gupta VA, Vertino PM, Boise LH. Cell of origin and genetic alterations in the pathogenesis of multiple myeloma Published Online. Front Immunol 2019. https://doi.org/10.3389/fimmu.2019.01121.
[6] Morgan GJ, Walker BA, Davies FE. The genetic architecture of multiple myeloma Published online. Nat Rev Cancer 2012. https://doi.org/10.1038/nrc3257.
[7] Bergamaschi A, Kim YH, Wang P, et al. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer Published online. Genes Chromosom Cancer 2006. https://doi.org/10.1002/gcc.20366.
[8] Tsang JYS, Tse GM. Molecular Classification of Breast Cancer Published online. Adv Anat Pathol 2020. https://doi.org/10.1097/PAP.0000000000000232.
[9] Pawlyn C, Morgan GJ. Evolutionary biology of high-risk multiple myeloma. Nat Rev Cancer 2017;17(9):543–56. https://doi.org/10.1038/nrc.2017.63.
[10] Palumbo A, Avet-Loiseau H, Oliva S, et al. Revised international staging system for multiple myeloma: A report from international myeloma working group Published online. J Clin Oncol 2015. https://doi.org/10.1200/JCO.2015.61.2267.
[11] Hieronymus H, Murali R, Tin A, et al. Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. eLife 2018;7:1–18. https://doi.org/10.7554/eLife.37294.
[12] Smith JC, Sheltzer JM. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. eLife 2018;7:1–26. https://doi.org/10.7554/eLife.39217.
[13] Mayrhofer M, Viklund B, Isaksson A. Rawcopy: Improved copy number analysis with Affymetrix arrays Published online. Sci Rep 2016. https://doi.org/10.1038/srep36158.
[14] Rasmussen M, Sundström M, Göransson Kultima H, et al. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity Published online. Genome Biol 2011. https://doi.org/10.1186/gb-2011-12-10-r108.
[15] Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 2012;30(5):413–21. https://doi.org/10.1038/nbt.2203.
[16] Van Loo P, Nordgard SH, Lingjærde OC, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci USA 2010;107(39):16910–5. https://doi.org/10.1073/pnas.1009843107.
[17] Translational Genomics Research Institute (TGen). tCoNut. Published 2016. https://github.com/tgen/tCoNuT.wiki.git.
[18] Gao B, Baudis M. Minimum error calibration and normalization for genomic copy number analysis. Genomics 2020;112(5):3331–41. https://doi.org/10.1016/j.ygeno.2020.05.008.
[19] Martello M, Poletti A, Borsi E, et al. Clonal and subclonal TP53 molecular impairment is associated with prognosis and progression in Multiple Mieloma. *Blood Cancer J*. Published online 2022. doi:In press.
[20] Keats JJ, Speyer G, Christofferson A, et al Published online. In: Molecular Predictors of Outcome and Drug Response in Multiple Myeloma: An Interim Analysis of the Mmrf CoMMpass Study. https://doi.org/10.1182/blood.V128.22.194.194.
[21] The Cancer Genome Atlas Program - National Cancer Institute. Accessed September 22, 2021. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.
[22] Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust : An R Package for Determining the. *J Stat Softw*. Published online 2014. Accessed January 14, 2022. https://www.jstatsoft.org/article/view/v061i06.
[23] Poletti A, Mazzocchetti G, Solli V. BOBaFIT: Refitting diploid region profiles using a clustering procedure. Published 2021. Accessed January 11, 2022. https://bioconductor.org/packages/devel/bioc/html/BOBaFIT.html.
[24] Haverty PM, Hon LS, Kaminker JS, Chant J, Zhang Z. High-resolution analysis of copy number alterations and associated expression changes in ovarian tumors. BMC Med Genomics 2009;2(21). https://doi.org/10.1186/1755-8794-2-21. Published 2009 May 6.
[25] Avet-Loiseau H, Li C, Magrangeas F, et al. Prognostic significance of copy-number alterations in multiple myeloma Published online. J Clin Oncol 2009. https://doi.org/10.1200/JCO.2008.20.6136.
[26] Mikhael JR, Dingli D, Roy V, et al. Management of newly diagnosed symptomatic multiple myeloma: Updated mayo stratification of myeloma and risk-adapted therapy (msmart) consensus guidelines 2013. Mayo Clin Proc 2013;88(4):360–76. https://doi.org/10.1016/J.MAYOCP.2013.01.019/ATTACHMENT/5F645F0E-5199-4053-9DFD-9176BD2353C0/MMC1.MP4.
[27] Tarabichi M, Salcedo A, Deshwar AG, et al. A practical guide to cancer subclonal reconstruction from DNA sequencing. Nat Methods 2021;18(2):144–55. https://doi.org/10.1038/s41592-020-01013-2.
[28] Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res 2016;44(16):e131–e. https://doi.org/10.1093/nar/gkw520.