Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE BIOMEDICHE E NEUROMOTORIE

Ciclo 35

Settore Concorsuale: 05/H2 - ISTOLOGIA

Settore Scientifico Disciplinare: BIO/17 - ISTOLOGIA

TIMMING: DEVELOPING AN INNOVATIVE SUITE OF BIOINFORMATIC TOOLS TO HARMONIZE AND TRACK THE ORIGIN OF COPY NUMBER ALTERATIONS IN THE EVOLUTIVE HISTORY OF MULTIPLE MYELOMA

Presentata da: Andrea Poletti

Coordinatore Dottorato

Matilde Yung Follo

Supervisore

Matilde Yung Follo

Co-supervisore

Lucia Catani

Esame finale anno 2023

Sapere aude!

Orazio (Epistole I, 2, 40)

Questa tesi è dedicata alla mia famiglia, a chi ha sempre continuato a credere in me.

ABSTRACT

Multiple Myeloma (MM) is a hematologic cancer with a heterogeneous and complex genomic landscape, that includes multiple types of genetic alterations. In particular, Copy Number Alterations (CNAs) play a key role in the pathogenesis and prognostic stratification of the disease. For this reason, it is of particular biological and clinical interest to study the temporal occurrence of early alterations over the developmental history of MM. This, in order to identify specific altered chromosomal regions or genes, which play a disease "driver" function by deregulating key tumor biological pathways. A correct identification of such alterations is especially important for the future development of a “personalized medicine” approach, which is aimed at identifying and targeting the specific biological "driver" alterations that characterize each individual tumor.

In this study, the researcher presents and discuss the development of an innovative suite of five bioinformatics tools (BOBaFit, RemasterCNA, RAPH, ComphyNumber and TestClonality) created for the purpose of harmonizing Copy Number data and tracing the origin of CNAs throughout the evolutionary history of MM. To this aim, the largest available cohorts of newly diagnosed MM (NDMM) and Smoldering-MM (SMM) were aggregated, encompassing in total 1582 MMs and 282 SMMs collected from four different cohorts. This result was made possible by the collaboration of Prof. Irene Ghobrial's laboratory from the Dana-Farber Cancer Institute of Boston.

The suite of tools developed in this study enables the harmonization of Copy Number data as obtained from multiple different genomic analysis platforms (e.g. WGS, WES, SNParray) in such a way that samples from the different cohorts can be merged and consequently a high statistical power of analysis can be obtained. By doing so, the high numerosity of those cohorts was harnessed for both 1) the identification (through the optimized use of the GISTIC tool) of novel of genes characterized as focal "driver" alterations in MM (including *NFKB2*, *NOTCH2*, *MAX* and *EVI5* and *MYC-ME2-enhancer* genes), and 2) the generation of an innovative timing model based on the Bradley-Terry approach, but implement with the introduction of a statistical method to introduce statistical confidence intervals in the analysis of CNAs. This innovative model was developed after a careful review of the existing scientific literature in the field of temporal analysis of cancer, and it is capable of tracing quantitatively, in a confident and precise way, the events considered as "early" or primary in the evolutionary history of cancers.

By applying this model on both NDMM and SMM disease phases, it was possible to identify specific CNAs (amplification 1q(*CKS1B*), deletion 13q(*RB1*), amplification 11q(*CCND1*) and deletion 14q(*MAX*)) and categorize them as "early" and "driver" events with a high precision and confidence. This level of precision was guaranteed by the narrow confidence intervals in the timing estimates obtained. Thus, the identified CNAs were proposed as critical MM alterations, which play a foundational role in the evolutionary history of both SMM and NDMM. Importantly, among the identified events amp *CKS1B* and del *RB1* were previously poorly characterized from an evolutionary point of view and uncertainly classified between primary and secondary events, while del *MAX* represent a completely new discovered MM driver alteration. Finally, a stepwise backward-forward Cox Regression survival model was able to optimally identify all the independent genomic alterations with the greatest effect on patients’ outcomes (Progression Free Survival and Overall Survival), including deletion of *RB1*, amplification of *CKS1B*, amplification of *MYC*, amplification *NOTCH2* and deletion-mutation of *TRAF3*.

In conclusion, the alterations that were identified as both "early-drivers” and correlated with patients’ survival were proposed as new biomarker candidates that, if included in wider multivariate survival models, could provide a better disease stratification and an improved patient prognosis definition.

INDEX

1 INTRODUCTION

1.1 MULTIPLE MYELOMA

Multiple myeloma (MM) is a lymphoproliferative disease that affects plasma cells, a type of white blood cell, in the bone marrow. It is characterized by the production of M-protein (also known as monoclonal immunoglobulin or paraprotein) and can lead to organ dysfunction, including high levels of calcium in the blood, kidney problems, anemia, and destruction of the bones. [1]

Unlike other cancers that spread to the bone, MM does not cause new bone growth in the osteolytic bone lesions it produces. These lesions are the main cause of morbidity in MM and can be detected using imaging techniques, such low-dose whole body computed tomography (WB-CT), fluoro-deoxyglucose (FDG) positron emission tomography/computed tomographic scans (PET/CT), or magnetic resonance imaging (MRI). Other major symptoms of MM include anemia, high levels of calcium in the blood, kidney failure, and an increased risk of infections. Approximately 1-2% of patients have extramedullary disease (disease outside the bone marrow) at the time of diagnosis, while 8% develop EMD during the course of the disease. [2]

1.1.1 Frequency and epidemiology

MM is relatively rare, accounting for only 1% of all neoplastic diseases, but is the second most common hematological cancer in high-income countries, with an annual incidence of 4.5-6 cases per 100,000 people, typically occurring in people around the age of 70. The incidence of MM is higher in western Europe, North America, and Australasia compared to Asia and sub-Saharan Africa, potentially due to differences in diagnosis. From 1990 to 2016, the global incidence of MM increased by 126% due to population growth, an aging population and increased age-specific incidence rates. [1]

1.1.2 Risk factors

Risk factors for MM include obesity, chronic inflammation, and exposure to pesticides, organic solvents, or radiation. Inherited genetic factors may also play a role in the development of MM. Additionally, both inherited and societal influences are identified to contribute to racial and ethnic disparities in the incidence and outcomes of MM and related conditions. [1]

1.1.3 Disease pathogenesis and diagnosis

The causes of MM initiation are complex and associated with intra-clonal tumor heterogeneity. [3,4] This complexity highlights the need for further research on the biology of the disease, particularly on the genetic and the molecular aspects of MM. [5] Most MM cases develop from asymptomatic conditions, known as monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM) (Fig. 1). SMM is a diverse stage of the disease, with some patients having a mild form similar to MGUS (about 30%), some having an intermediate course, and others having an aggressive form known as "early myeloma" (20-30%). [6,7] Individuals with early myeloma SMM were reclassified as having overt MM if they meet the myeloma-defining

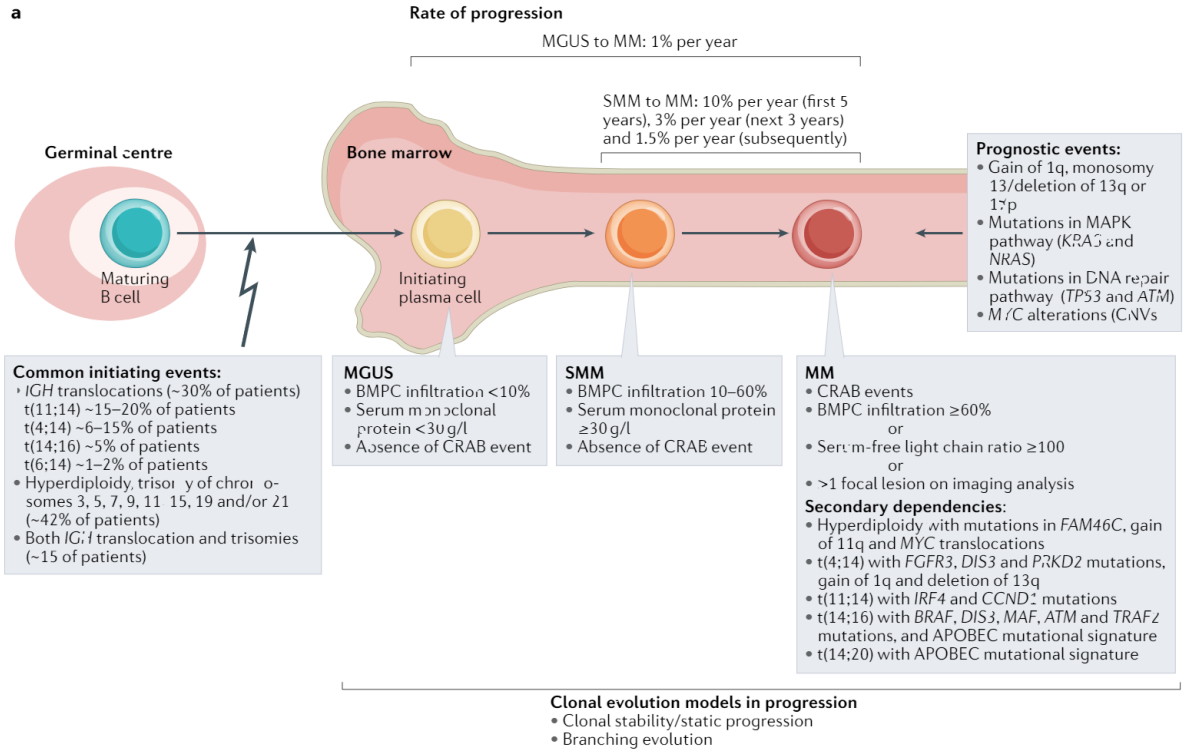


***Figure 1**: Continuum of progression of MM. Multiple myeloma (MM) develops from the precursor stages of monoclonal gammopathy of undetermined significance (MGUS) and smouldering MM (SMM). Although these precursor states are considered asymptomatic, they are currently not precisely defined by clinical parameters alone. Thus, the disease stages of MM can be considered as being on a continuum.*[7]

events that were added to the International Myeloma Working Group clinical criteria for classification. [8] Myeloma-defining events include clonal bone marrow plasma cell infiltration of ≥60% or a serum-free light chain ratio of ≥100 or >1 focal lesion in the skeleton on MRI analysis. These criteria extend the definition of overt MM onset prior to end-organ damage from hypercalcemia, renal insufficiency, anemia and bone lesions (CRAB features) in an effort to capture MMs at an earlier stage for therapy (Fig. 1). [8,9] Despite the extensive knowledge of clinical criteria defining disease stages throughout MM development, the precise genomic changes associated with the progression of MGUS or SMM subgroups to MM remain poorly understood.

1.1.4 Heterogeneity of genetic alterations in MM and their role as risk factors

The onset of conditions that precede MM in a developing B cell clone can be attributed to specific common genetic events, which result in two different types of MM: non-hyperdiploid and hyperdiploid. The non-hyperdiploid group of MM is characterized by specific genetic alterations, including translocations of the immunoglobulin heavy chain (IGH) locus, with the most common being t(11;14), t(4;14), t(14;16), t(6;14) and t(14;20). These alterations occur in 15-20%, 6-15%, 5%, 1-2% and 1% of MM patients, respectively. Hyperdiploidy, on the other hand, refers to the presence of more than 48 chromosomes, and in MM, it is associated with trisomy of chromosomes

3, 5, 7, 9, 11, 15, 19 and/or 21. This benign condition is relatively common, with an estimated prevalence of 3-5.1% among individuals over 50 and 5% among those over 70 (Fig. 1).[7,10]

The use of NGS in MM has greatly expanded our understanding of the genetic diversity, crucial mutations, and evolution of the disease. Several large-scale studies using samples from patients with MM or its precursor stages have shown that MM is characterized by intraclonal heterogeneity and evolutionary changes, with different populations of plasma cells carrying various mutations and the dominance of different clones changing as the disease progresses.

Given the MM genomic alterations heterogeneity it becomes important to distinguish between "passenger" and "driver" genomic alterations:

- **Driver genomic alterations:** genetic changes that are believed to actively contribute to the development and progression of cancer. These changes can include mutations, amplifications, and translocations that occur in oncogenes (genes that promote cell growth and survival) or tumor suppressor genes (genes that normally help to prevent the growth of cancer cells). Examples of driver genomic alterations in cancer include mutations in the KRAS gene in that confers proliferative advantage and deletions of the TP53 gene that confers resistance to apoptosis. [11]
- **Passenger genomic alterations:** the remaining alterations are termed passenger. They show a poorly understood molecular consequences and fitness effects on the tumor growth and emerge simply as victims of genomic instability, occurring as a result of cancer progression. [11]

Driver mutations in genes such as KRAS, NRAS, BRAF, TP53, DIS3, or TENT5C (also known as FAM46C) confer a growth advantage to certain plasma cell clones, leading to their outgrowth and development into MM. The genetic landscape of MM is now well-described, with around 80 driver mutations identified to date.[4,12,13] (Fig. 2).This genetic diversity is a major contributor to the varied outcomes of MM patients and a significant obstacle to finding a single cure.

Studies in patients with high-risk SMM have revealed that specific genomic risk factors can predict progression to MM, such as mutations in MAPK pathway genes (KRAS and NRAS), DNA repair pathway genes (deletion of 17p and TP53 and ATM mutations), and MYC (translocations and copy-number variants).[14]

Whole-genome sequencing (WGS) of MM samples has provided additional insights into the disease biology compared to previous methods such as whole-exome sequencing (WES). For instance, WGS can reveal mutational features such as CNA, structural events, and APOBEC activity that can be used to classify patients into different biological groups.[15]

Furthermore, WGS analyses of NDMM within the CoMMpass study have revealed genomic alterations, including 11 candidate non-coding drivers, IGL light chain locus translocations (t(IGL)) conferring high risk, and complex structural variations, such as chromothripsis and templated insertions, which are key drivers that affect the plasma cell genome (Fig. 2).[15–17]

A comprehensive analysis and understanding of these genomic alterations and their biological implications is crucial for selecting appropriate clinical strategies. Additionally, even rare genetic characteristics of MM can help predict survival and treatment outcomes.

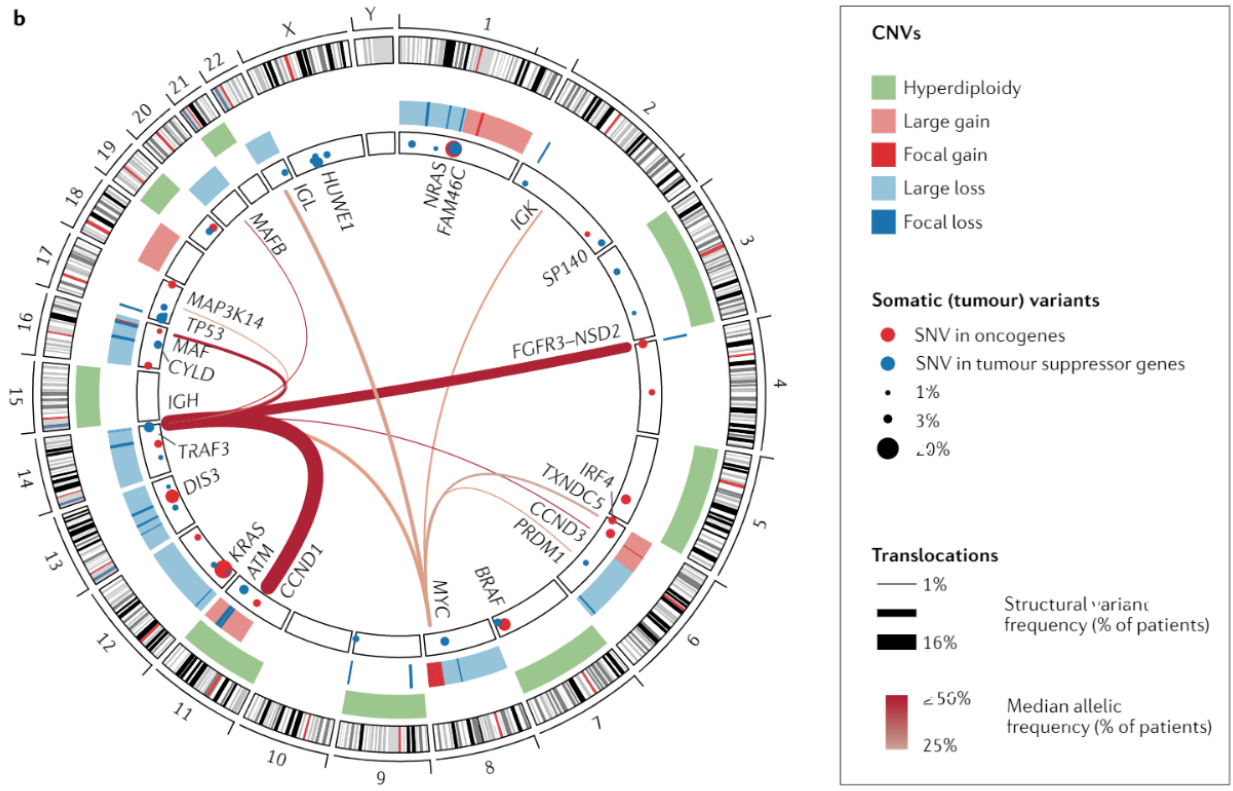


Figure 2: *Circos-plot that recapitulates all the main somatic genomic alterations observed in MM by using WGS technologies.* [18]

1.1.5 Disease evolution trajectories

In patients with MM, PCs are characterized by a high intra-clonal heterogeneity, with multiple sub-clones defined by the presence of different genomic alterations, competing for the access to the limited resources available in the BM. In accordance to the Darwinian evolutionary model, the sub-clonal genomic architecture and spatial heterogeneity both change during the progression of the disease[19,20] throughout at least two evolutionary trajectories[4,21]: 1) linear trajectory, in which the tumor cells population tends to sequentially accumulate genomic alterations [22]; 2) branched trajectory, in which the cells population tends to differentiate into several independent evolutionary lines, each one characterized by different genomic alterations[4,22]. Regardless of the evolutionary trajectory, it is well-accepted that at least 6-7 genomic hits are required for the neoplastic transformation of a healthy cell. These genomic hits can be modeled as stochastic events that can occur over either a short or a long time-span[23].

1.1.6 Therapy, survival, and response assessment

The use of new drugs and/or therapeutic strategies, such as proteasome inhibitors, immunomodulatory drugs, and antibodies targeting cell surface molecules, as well as high-dose therapy and autologous stem cell transplantation (ASCT) in younger patients, has significantly improved the prognosis for patients with MM. The median overall survival for patients eligible for ASCT is about 10 years, compared to 4-5 years for those who are not eligible. Most patients with MM experience multiple relapses of their disease, with each subsequent remission becoming progressively shorter until the disease or treatment-related complications ultimately lead to death.[1]

Until the early 2000s, the therapeutic armamentarium for MM was limited to steroids, alkylating agents (such as melphalan and cyclophosphamide) and 'traditional' chemotherapies. Subsequently, several classes of therapeutic agents have been introduced for the treatment of MM, including immunomodulatory agents, proteasome inhibitors, histone deacetylase inhibitors, and monoclonal antibodies. In addition, several new classes of therapeutics are currently being evaluated in clinical trials[24], including venetoclax (an inhibitor of apoptosis regulator BCL2 and selinexor (an inhibitor of exportin 1). Adoptive cell transfer using chimeric antigen receptor (CAR) T cells targeting B cell maturation protein (BCM) has also shown promise in early-stage clinical trials.[25]

The evaluation of response to treatment in patients with MM has traditionally relied on the quantification of monoclonal protein (a term that refers to the levels of monoclonal antibodies in serum) and on the assessment of residual MM cells within the bone marrow (Minimal Residual Disease, MRD). The effectiveness of treatments has improved over time, and the achievement of deeper responses has become a reality; thus, approaches to disease assessment have also evolved. In the past 2 years, the use of sensitive flow cytometry and/or next- generation sequencing- based approaches to MRD assessment has been reported to enable the detection of up to 1 residual plasma cell in 10^{-6} bone marrow cells.[26]

MRD detection has also been combined with imaging approaches such as PET to provide a more accurate assess-ment of therapeutic effectiveness. The revised IMWG response criteria incorporate MRD assessment.[27]

MRD negativity has been demonstrated to be one of the most important prognostic factors in patients with MM, according to the results of a meta-analysis published in 2017 showing that MRD negativity is associated with improved progression free- (PFS) and overall survival (OS).[26]

According to currently available data [28], MRD negativity can be an excellent surrogate end point for PFS, and potentially OS, and should therefore be incorporated into clinical trials. The clinical utility of MRD negativity as a decision-making tool needs further study, as this parameter seems to be a function of both disease biology and the treatment regimen used. Prospective clinical trials

are currently examining whether changing treatment according to MRD status can affect survival outcomes.[29]

1.2 GENOMIC ALTERATIONS TIMING ANALYSIS IN CANCER

Recently, it has been shown that the integrated use of high-throughput genome analysis technologies (such as whole-exome sequencing (WES), whole-genome sequencing (WGS) or high-resolution SNP arrays) and bioinformatics procedures allows for the tracing of the temporal evolution of single genomic alterations in a tumor. This is accomplished by analyzing the cancer cell fraction (CCF) of the numerous somatic mutations (SNVs and InDels) and CNAs scattered throughout the genome, based on the fact that subclonal (or "late") alterations must occur after clonal (or "early") alterations in any given observed tumor sample (fig. 3). This analysis is often performed in bulk-sequencing samples; however, a more precise approach is to analyze the genomic alterations in cancer at the single-cell level (scWGS). This solution is significantly more expensive and technically challenging but allows for an extremely precise identification of subclonal genomic alterations (fig 3).[30–32]

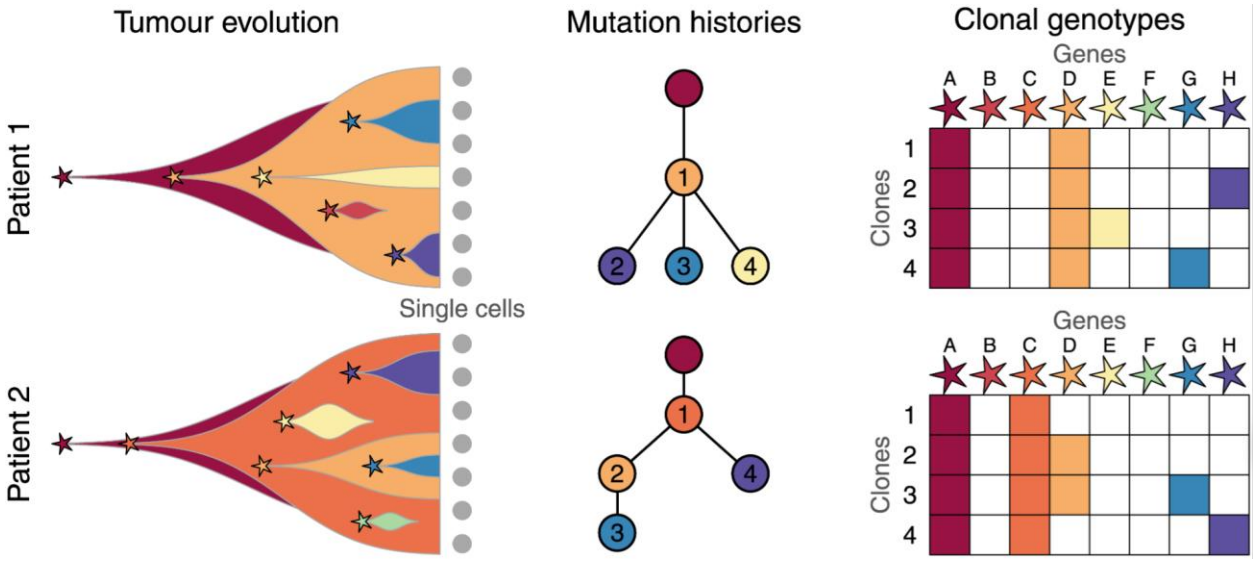


Figure 3: *Cartoon depicting the evolutionary history of two tumors. By using bioinformatic methods it's possible to reconstruct the temporal evolution of the specific genomic alterations in tumors. Clonal alterations which are present in all the tumor cells (e.g. red and orange alterations in patient 1), occurs earlier than sub-clonal alterations that are present only in a fraction of the tumor cells (e.g. yellow, blue and purple alterations in Patient 1).*

To trace the temporal evolution of single genomic alterations, four main bioinformatics strategies can be used:

1. **Duplication timing:** in a single tumor sample, in which it is assumed that mutational processes are active and cause a constant incidence of SNVs over time (or "passenger" mutations), the ratio between duplicated and non-duplicated SNV observed in duplicated genomic regions can inform about the "mutational time" of the given duplicated genomic region[32,33] (Fig. 4);

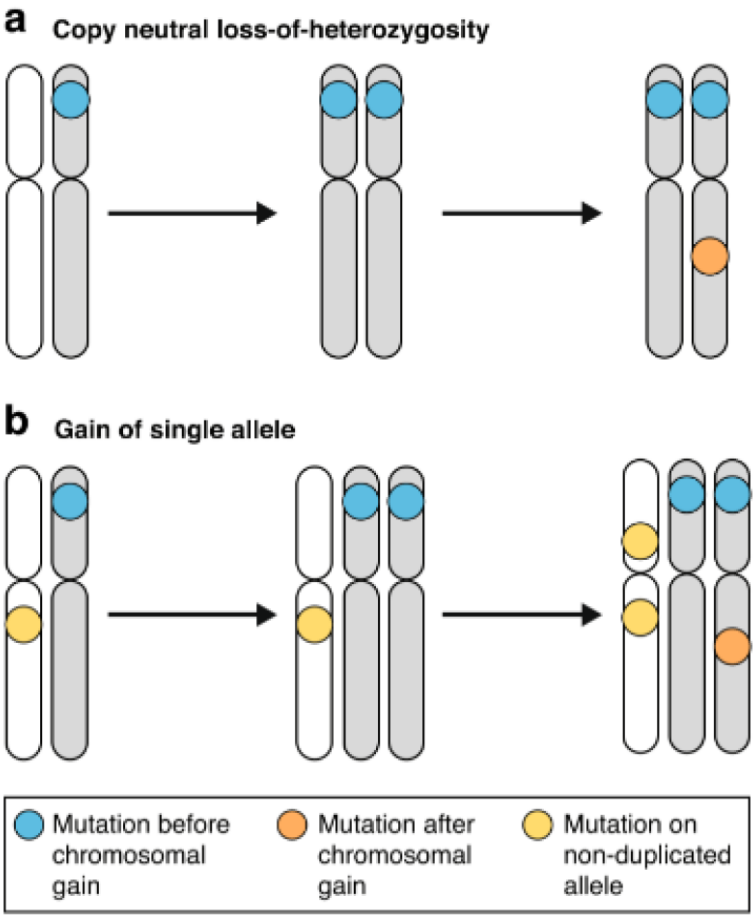


Figure 4: *Timing analysis of gain and LOH events using point mutations (SNV). The relationship between the SNVs and chromosomal events can be used to infer the relative time of the chromosomal alteration.* ***a****. Timing of a LOH event: the mutations highlighted in blue occurred before the LOH event, the mutations in orange occurred after the LOH event. The relationship between the quantity of the two types of mutations allows to derive a relative "mutational time" of the event.* ***b****. Timing of an allele gain event. In this case a further level of complexity is added and it is necessary to normalize the calculation, taking into account that the mutations in single copy can reflect both the mutations that occurred after the gain (orange), and those on the non-duplicated allele (yellow).*

2. **Cohort timing / League model:** in a cohort of tumor samples, the cross-aggregation of coupled temporal estimates of genomic alterations (e.g. clonal vs sub-clonal alterations) can inform on the relative chronology of the tumor's genomic events, throughout the application of specific statistical and probabilistic algorithms (e.g. Bradley-Terry model)[33–35] (Fig. 5);

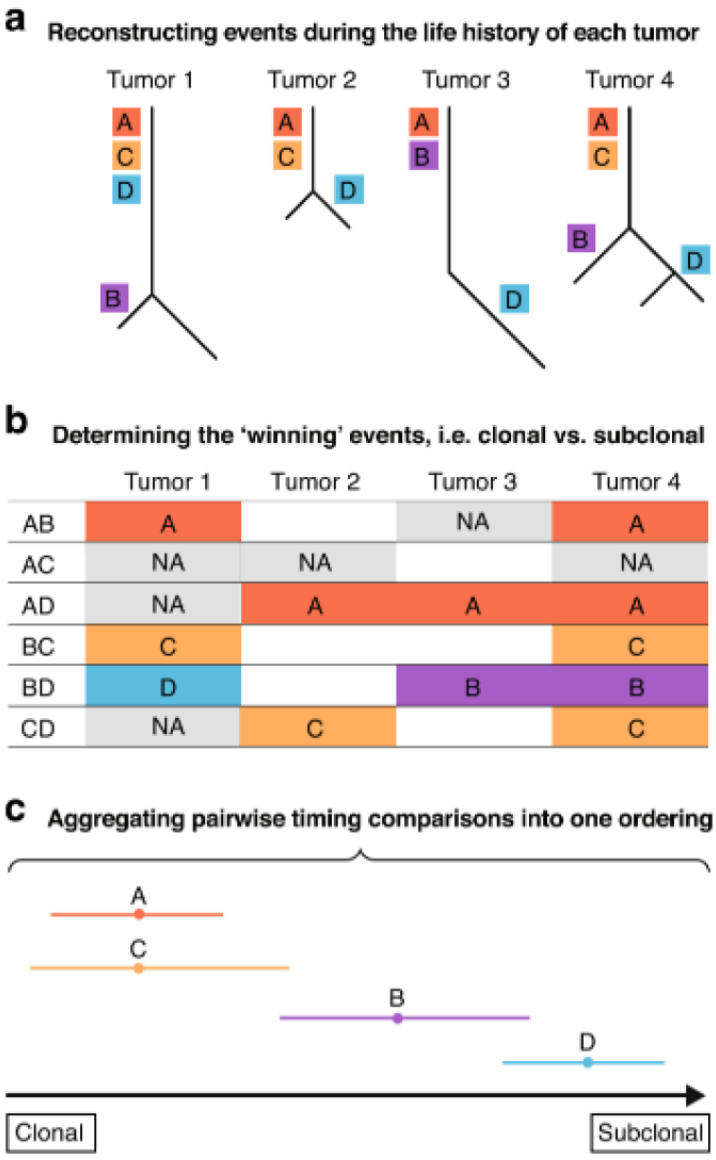


Figure 5: *Aggregation of the relative timing of events between multiple patients of the same tumor type. Once the timing of events in individual patients has been established, individual orders can be cross aggregated over the cohort of tumor samples to determine a probabilistic sequence of events. a. Examples of phylogenetic trees representing the orders of acquisition of events in individual tumor samples. Mutations A-B-C-D are represented near the top or bottom of the trees based on their clonal or sub-clonal status, respectively. b. The results of all the coupled comparisons between the A-B-C-D events within each tumor sample. In these comparisons the most clonal event is considered as the "winner". Comparisons marked as "NA" indicate cases in both events are present, but it is not possible to establish a relative timing between the two as they have the same clonality level. c. Final order: Events A and C are estimated to be the most ancestral ones, as they have often "won" against the other two events B and D (however the confidence interval of A is narrower as it won more comparisons). Next, B event has an intermediate temporal order: it often won against D event, but always lost against C and A events. Finally, D is estimated to be the event that occurs later, because it has never won a competition, except once (in Tumor1).*

3. **Mutational signatures**: the use of stable clock-like mutations characterizing the age-dependent mutational processes allows to transform the relative chronology of events (and the concept of "mutational time") into real years: using linear regression models, it is possible to assess the relationship between the patients age and the mutational burden of age-dependent mutational signatures (SBS1 and SBS5). Thus, it is possible to obtain quantifiable and measurable time coordinates, enabling to enumerate in years the actual ancestrality of genomic alterations. This makes possible to develop time maps of each driver event that can be studied within a specific tumor type[31,34] (Fig. 6).

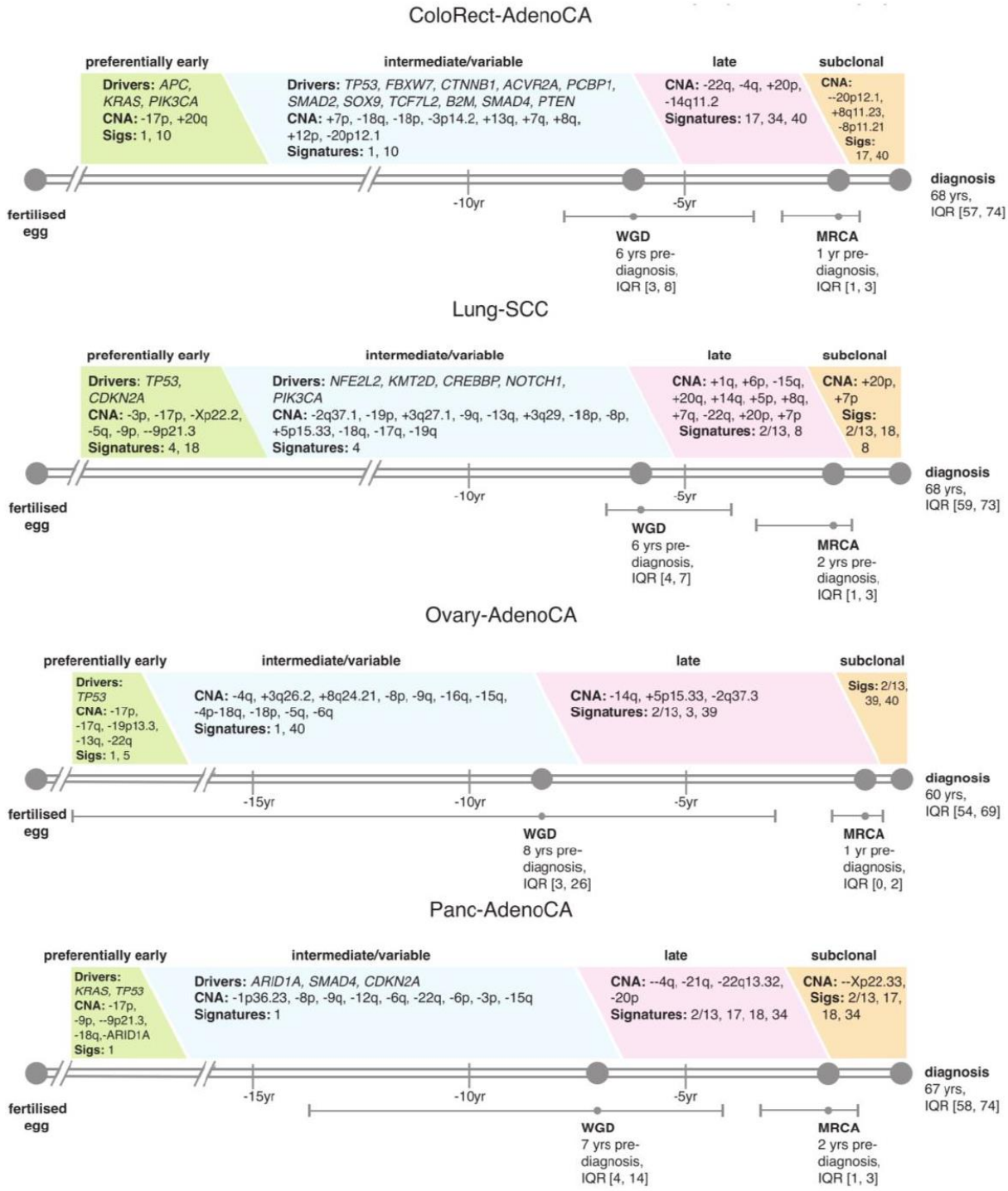


Figure 6: *Chronological maps of different types of cancer studied in the PCAWG pan-cancer project. Each map represents the length of time, in years, between the "fertilized egg" (precursor cell at the origin of tumor evolution) and the median age at diagnosis for each type of cancer. Point estimates for relevant events are used to define different "phases" (early, intermediate, late and subclonal) of tumor clonal evolution, in a chronological real time. Driver alterations are shown associated with each tumor phase according to their own specific timing, defined by the relative chronological order of the alterations in the tumor.*

4. **Multi-sample phylogeny:** by interpreting quantitatively the frequency (CCF) of clonal and subclonal SNVs, it is possible to deduce the existence of linear or branching evolutionary trajectories, tracing specific onco-phylogenetic trees for individual patients by employing algorithms based on the pigeonhole rule (i.e., the sum of CCFs of sibling clones cannot exceed the parent clone's CCF) (Fig. 7).[36,37]

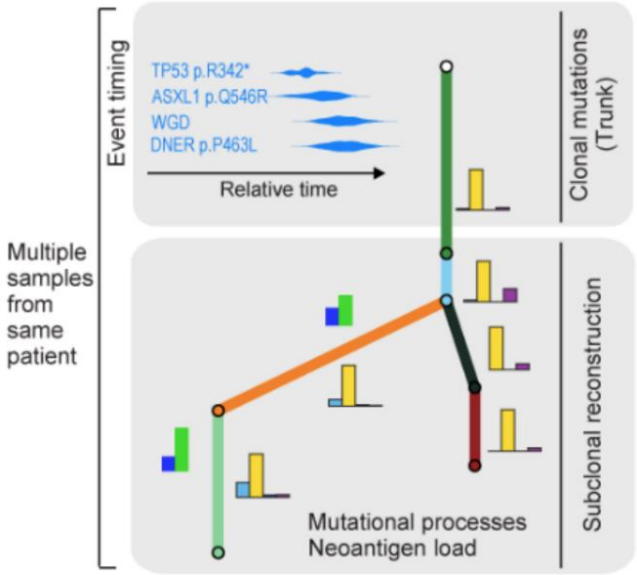


Figure 7: *Schematics of the reconstruction of subclones and phylogeny performed by PhylogicNDT tool [36]. In the upper part, timing of acquisition of alterations (early vs. late) and their phylogenetic relationships in a single patient sample. In the bottom part, generation of most plausible developmental trajectory and its associated phylogenetic tree in a patient with multiple available samples.*

In conclusion, the ancestrality of "driver" events (i.e., events underlying tumor malignancy characteristics) has already been demonstrated in several cancers, suggesting that these events may precede diagnosis by years or even decades. [38,39]. The possibility to analyze the genomic heterogeneity of a tumor on a chronological basis is of great interest in the field of bioinformatic analysis of tumor genomes; in fact, prognosis and response to therapy might depend on both the level of evolution and the degree of tumor heterogeneity, as well as on the presence or absence of certain "driver" events. For this reason, information on the temporal stage of tumor clone(s) might allow to attribute a biological significance to the genomic alterations and to the observed tumor heterogeneity.

1.2.1 Genomic timing analysis in MM: state of the art

In MM, the timing analysis of genetic alterations has been carried out in two pioneering studies in recent years. In particular, in the first study[15], a cohort of 30 patients analyzed through WGS, the order and the time window of acquisition of the typical MM odd-numbered chromosomes hyperdiploidy was analyzed through the previously illustrated "Duplication timing" method. In addition, on the same cohort, a Bradley-Terry model was generated to perform a timing analysis of all CNA with the "League Model" method in order to reconstruct the temporal trend of acquisition of different alterations on the entire cohort. Finally, for each patient, a phylogenetic tree was generated using serial samples of the same patient through the "multi-sample phylogeny" timing method, thus allowing to investigate the chronological order of all alterations, including SNV, CNA and structural events such as chromothripsis and translocations [15]. In this study, it has been possible to identify different time windows for multiple acquisitions of hyperdiploid chromosomes, which do not always occur at the same time during tumor evolution. Furthermore, despite the small patient cohort, it has been possible to confirm that hyperdiploidy (in particular the amplification of chromosome 11) is an ancestral event compared to other CNA events, followed by amplification of chromosome 1q and deletion of chromosome 1p and 13q (Fig. 8).[15]

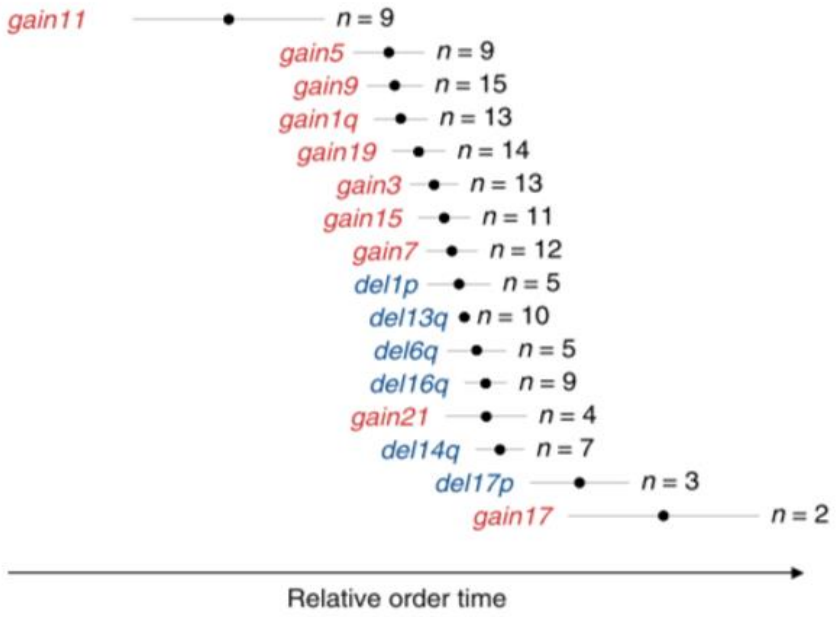


Figure 8 *Chronological order of acquisition of CNA events in the onset of MM, computed using the Bradley-Terry model on a cohort of 30 WGS samples. Earliest event appears on the left while late occurring events appears on the right. 95% Confidence intervals are represented by black lines.*[15]

In the second study[40], the CNA events' acquisition order was investigated employing a novel timing method, similar to the "league model," on a cohort of 336 NDMM analyzed by SNP arrays (Fig. 9). Critically, the timing model used in this study was not the canonical "Bradley-Terry" model, but on the contrary a model that computed the temporal estimation of alterations by carrying out a direct comparison (with 1000 iterations bootstrapping) between the level of clonality of the various observed CNA events. The clonality levels were previously defined in 5 ordered categories ranging from "low subclonal event" to "completely clonal event" according to the Tukey HSD statistical test (a test capable of identifying significant differences between groups of events with different measures)[40]. Despite this different timing model, the results obtained in this study were comparable to those obtained by Maura et al., identifying hyperdiploidy, amplification of chromosome 11 and 1q, and deletion of chromosome 13 as ancestral events in the evolutionary history of MM. (Fig. 9) Furthermore, another fundamental and critical feature of this study was that the timing analyses for the two main categories of MM, namely: hyperdiploid and non-hyperdiploid, were considered apart.

In conclusion, these pioneering studies on the temporal evolution of genomic alterations in the history of MM, provided solid evidence to the role of CNA events' chronological order in the onset of MM, by defining the most ancestral/"early" alterations as MM driver events. However, the low samples size in the first study and the use of a non-canonical bioinformatic method in the second one, both not fully disclose the real and precise chronological order of individual CNA alterations occurring throughout the MM evolutionary history. This is evident by observing the wide 95% confidence intervals of the temporal events estimates, showed in Fig. 8 and Fig. 9, which overlap in most cases. Moreover, the chronological order of either rare or focal events (i.e. involving individual genes and not entire chromosome arms) which were not included in these studies still remains of interest.

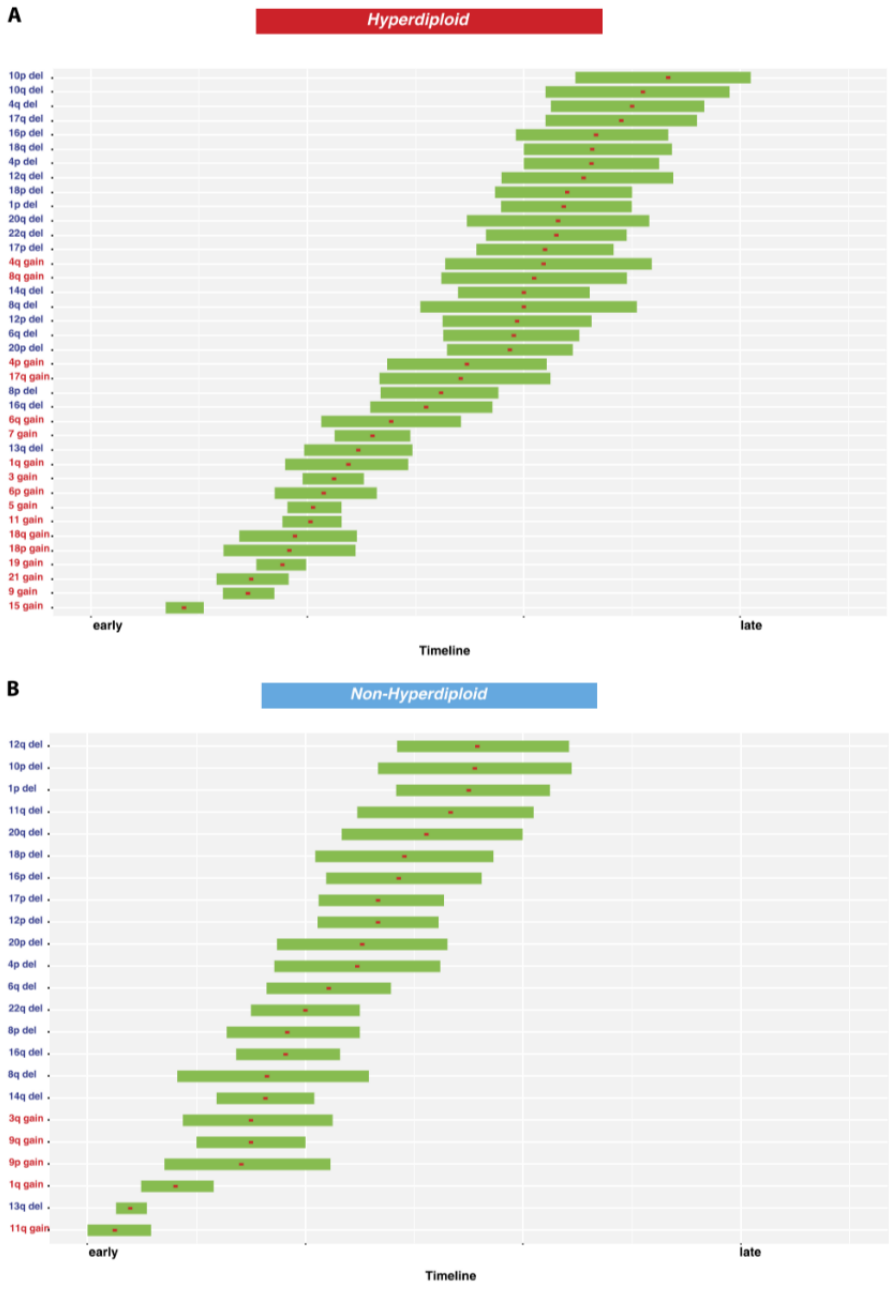


Figure 9: *Chronological order of acquisition of CNA events in the onset of MM, computed using an original statistical timing model on a cohort of 336 SNP array samples. Earliest event appears on the left while late occurring events appears on the right. 95% Confidence intervals are represented by green bars. The patient cohort was divided in two sub-cohorts for this analysis: Hyperdiploid MM and Non-Hyperdiploid MM.*[40]

1.3 COPY NUMBER ALTERATION ANALYSIS

Copy Number Alterations (CNAs) refer to changes in the number of copies of a particular section of DNA within a genome. These changes can range from small losses or gains of a few base pairs, to large deletions or duplications of entire chromosomes. CNAs are a common feature of cancer, as they can disrupt the normal regulation of genes and contribute to the development and progression of the disease. There are several bioinformatic tools and algorithms that can be used for the genomic analysis of CNAs in cancer.[41]

1.3.1 Array based platforms

Array Comparative Genomic Hybridization (aCGH) is one of the most commonly employed method to detect CNAs. This method uses DNA microarrays to compare the CN of genomic regions between a normal sample and a tumor sample. The microarray contains probes that hybridize to specific regions of the genome, and the fluorescence signal generated by these probes is used to infer the CN of the regions. aCGH can detect CNAs at a resolution of several kilobases and is useful for identifying large-scale CNAs, such as whole chromosome gains or losses. However, aCGH is limited in its resolution and cannot detect small CNAs or structural variations. [41]

Single Nucleotide Polymorphism (SNP) array analysis is another popular method, microarray-based as well, that can be used for the detection of CNAs. This method uses SNP markers instead of genomic regions as probes on the microarray. The genotype of each SNP marker is determined by comparing the fluorescence signal generated by the probe to a known reference. By analyzing the genotype of thousands of SNP markers across the genome, SNP arrays can detect CNAs at a resolution of several hundred base pairs. This method also has the advantage of detecting CNAs and genotyping simultaneously. [41,42]

1.3.2 Sequencing based platforms

Next-Generation Sequencing (NGS)-based methods represent another common approach; in particular, whole-genome sequencing (WGS) or targeted sequencing (exome sequencing, panel sequencing) have become increasingly popular for the detection of CNAs. These methods can provide a more comprehensive view of the genome than microarray-based methods and can detect CNAs at a higher resolution. WGS and targeted sequencing can detect CNAs as small as single nucleotide changes and can also detect structural variations such as inversions, translocations, and insertions. However, these methods also require a large amount of computational resources and analytical expertise. The resolution of the analysis is also limited by the depth of the sequencing and, consequently the sequencing costs. Many bioinformatic approaches can be applied to CNAs identification. So far, the NGS based CNAs detection methods can be categorized into four different strategies illustrated in Figure 10 [43,44].

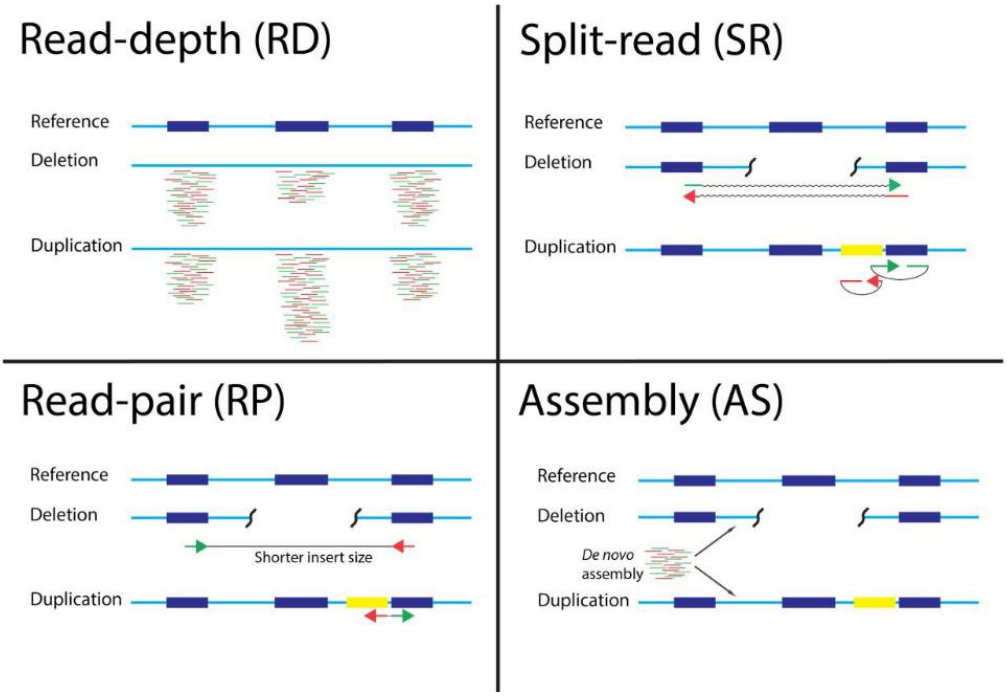


Figure 10: *Four approaches to detect CNAs from NGS short reads. Read depth (RD)-based approach detects CNAs by counting the number of reads mapped to each genomic region. In the figure, reads are mapped to three exome regions. B. Split read (SR)-based methods use incompletely mapped read from each read pair to identify small CNAs. C. Paired-end mapping (RP) strategy detects CNAs through discordantly mapped reads. A discordant mapping is produced if the distance between two ends of a read pair is significantly different from the average insert size. D. Assembly (AS)-based approach detects CNAs by mapping contigs to the reference genome.* [43,44]

1.3.3 Bioinformatics tools used to analyze CNAs

One popular bioinformatic algorithm used to analyze CNAs in cancer is the circular binary segmentation (CBS) algorithm. The algorithm was first described in the paper by Olshen et al in 2004, [45] and it is implemented in the R package DNAcopy [46], which is widely used in tools for the analysis of CN in genomics. For example, the packages *Rawcopy* (analysis of SNP arrays from Affymetrix), *cnvkit* [47] and *CapSeg* [48] (analysis of NGS data) ultimately refer the segmentation of the count data that they generate from sequences to to DNAcopy, and thus to the CBS algorithm. The CBS algorhitm segments the genome into regions of constant CN and identifies breakpoints where CN changes occur. CBS uses a recursive algorithm to divide the genome into smaller segments and then test for changes in CN within each segment. This algorithm is particularly useful for identifying regions of recurrent CNAs across multiple samples. [46,49]

Another commonly used algorithm is GISTIC (Genomic Identification of Significant Targets in Cancer), which uses a significance threshold to identify recurrent CNAs across a large number of samples [50]. GISTIC uses a sliding window approach to identify regions of the genome that are recurrently altered in a significant number of samples. GISTIC2 works by first identifying all the genomic regions that are recurrently altered across a set of samples. These regions are called "peak regions". Next, GISTIC2 identifies regions of the genome that are likely to be "driver" regions, which are regions that are likely to be directly involved in the development of cancer, as opposed

to "passenger" regions, which are regions that are altered in cancer but are not likely to be directly involved in cancer development. This algorithm is particularly useful for identifying CNAs that are likely to be important drivers of cancer [50,51].

In conclusion, CNAs are a common feature of cancer and several bioinformatic tools and algorithms are available, that can be used to analyze them. Array-based methods, like aCGH and SNP arrays, as well as NGS-based methods, can be used for the detection of CNAs. Algorithms like CBS and GISTIC can be used to identify and interpret CNAs in cancer and are particularly useful for identifying recurrent CNAs on driver genes across multiple samples.

1.4 Role of CNAs in Multiple Myeloma

In MM, CNAs can affect the function of tumor suppressor genes, which normally help to prevent the growth of cancer cells, as well as oncogenes, which promote the growth and survival of cancer cells. CNAs are not the only genetic changes that occur in multiple myeloma, in fact other genetic alterations, such as mutations and translocations, can also contribute to the development and progression of the disease. However, CNAs are some of the most clinically relevant genetic changes and show can be used as biomarkers to guide treatment decisions and predict patient outcomes. [25,52]

Relevantly, when analyzed in big cohorts of WGS samples, it was shown that CNAs along with IgH-translocations are the strongest determinants of the structure of the global genomic heterogeneity observable in MM. Thus, they may play a critical role in the genomic classification of different biological sub-types of MM. [15,53]

1.4.1 Relevant MM CNAs

Almost all NDMMs harbor CNAs in their genome; among them, few have been shown to be particularly relevant to define the biology and/or prognosis of the disease:

- **Deletion of TP53**: one of the most relevant CNAs in MM is the deletion of the tumor suppressor gene TP53. This gene plays a key role in regulating cell growth and death and its deletion can lead to the uncontrolled growth of cancer cells. Studies have shown that deletion of TP53 is associated with a higher risk of relapse and shorter overall survival in MM patients.[54] Moreover, the clonality level threshold used to define the TP53 deletion was shown to be critical to reliably describe the impact of this alteration on patients' survival. (Fig. 2) [55]
- **Gain of chromosome 1q:** gain of chromosome 1q (gain1q) is another common CNA in MM. It is one of the most recurrent cytogenetic abnormalities in MM, occurring in approximately 40% of newly diagnosed cases (Fig. 2). Although it is often considered a poor prognostic marker in MM, gain1q has not been uniformly adopted as a high-risk cytogenetic abnormality in guidelines. A major controversy concerns the importance of gain1q copy number, as well as whether gain1q is itself a driver of poor outcomes or merely a common passenger genetic abnormality in a biologically-unstable disease, such as MM. Although the identification of a clear pathogenic mechanism driven by gain1q remains

elusive, many genes included in the 1q21 locus have been proposed to cause early progression and resistance to anti-myeloma therapy. The plethora of potential drivers suggests that gain1q is not only a causative factor or poor outcomes in MM but may be targetable and/or predictive of response to novel therapies. [52]

- **MYC amplification**: gain of MYC oncogene is another important CNAs in MM, commonly caused by structural aberrations involving the MYC locus (e.g. translocations) (Fig. 2). MYC is a transcription factor that regulates the expression of a wide range of genes involved in cell growth and metabolism, and its amplification and/or the amplification of one of the enhancers adjacent to MYC, is considered a poor prognostic factor, associated with aggressive disease and shorter survival.[56]
- **Deletion of chromosome 13q:** chromosome 13q deletion (del13) is present in approximately 50% of patients with newly diagnosed MM. However, despite being the most common copy-number change, its association with prognosis has been debated (Fig. 2). Initially, del13 was associated with a poor outcome, but further study of high-risk abnormalities that co-occur with del13, such as t(4;14), led to the conclusion that it is not associated with poor prognosis. [57] In MM, the main genes of interest on chromosome 13 have been the cell-cycle regulator RB1 and the exonuclease DIS3. RB1 is infrequently mutated but is more frequently bi-allelically deleted (6%), especially in high-risk groups. In contrast, DIS3 is one of the most frequently mutated genes in MM (10%), and biallelic abnormalities are associated with poor outcome. However, due to the high frequency of whole-arm deletion of chromosome 13, and infrequent mutations of the genes contained within, it has been difficult to determine a minimally altered region on this chromosome. [57]

1.4.2 Comparison between MGUS, SMM and NDMM Copy Number landscape

The CNAs landscape of myeloma precursor condition (MGUS and SMM) is not significantly different from that of NDMM; in fact, the same recurrent CNAs regions were found in the different disease stages (Fig. 11) [56,58]. However, some of these alterations were observed at lower prevalence in asymptomatic disease stages as compared to MM, suggesting they may confer higher risk of progression. Although to a lesser degree than the fully malignant counterpart, the genomic CNAs landscape of MGUS and SMM is still complex and heterogeneous, with significant differences in the frequency of CNAs between SMM and MM, providing support with their role as drivers.[14,59]

In studies that included longitudinal samples, focusing at investigating the SMM-MM interface, it has been shown that the majority of CNAs present at progression were already present at the SMM stage. However, copy number events in SMM have a strong driving potential and can eventually become clonal when acquired during progression to MM through clonal evolution processes.[14,59]

Overall, these observations suggest that precursor conditions are in general genetically “mature” entities, whereby most driver genetic alterations have already occurred. Furthermore, specific alterations as univocal driver lesions predictive of the transition to overt MM have not yet been

defined, since the study of SMM genomics is still in its infancy, and advances in this regard are conceivable by continuous research in this field. [58]

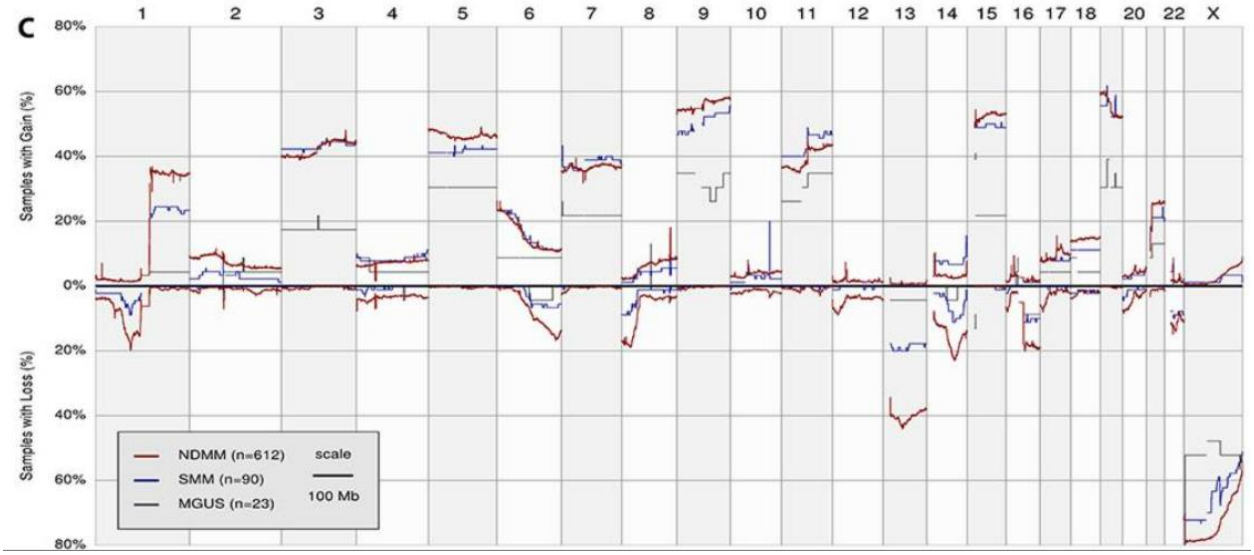


Figure 11*: the genomic Copy Number landscape of NDMM (in red), SMM (in blue) and MGUS (in black). Even if the recurrent regions of alterations are not significantly different, the frequency and clonality of the alterations differs between the different disease phases.*

2 AIMS OF THE STUDY

This study aims to aggregate Copy Number profiles data, obtained by high-throughput genomic analyses performed in the largest NDMM and SMM cohorts of patients available so far, by generating a bioinformatic pipeline designed to harmonize and correct all possible methodological biases observed in the different genomic data collections.

This will support the main objective of the study, that is to generate a timing analysis of MM genomic aberrations, finally generating temporal maps of MM evolutionary history with unprecedented resolution and precision. This will be pursued by using the "League Model" approach.

Secondary objectives of the study will be to implement the procedures currently used in the "League Model" with new developed statistical features, able to

1) improving the definition of "driver" events in MM
2) adding the dimension of "error modeling" to the CNA analysis (calculating 95% confidence intervals for CNA alteration estimates), which in previous studies has been considered very rarely.

At bioinformatic level, the various functions and tools employed to set up the developed bioinformatic pipelines will be designed by R packages published in official repositories (e.g. BioConductor) or in private repositories, accessible on request (e.g. "GitHub"). This will allow to make available for the scientific community both the data harmonization and the "error modeling" pipelines, whose novelty mainly reside in the current absence of informatic tools specifically devoted to Copy Number data analysis in tumors.

A final objective of the study will be to identify either "driver" or ancestral alterations and to correlate the presence of ancestral alterations with survival data (OS and PFS) of patients included in the different cohorts. This will allow to identify the role of the presence of ancestral alterations in the stratification and the prognosis of the disease.

3 PATIENTS AND METHODS

3.1 PATIENTS AND COHORTS

In this study the researcher was able to aggregate four different cohorts of patients, collected at different disease phases (NDMM and SMM), were aggregated, leading to a final dataset including 1867 patients, with genomic profile.

The two cohorts of NDMM samples included the Bologna (BO) cohort, collected at the "L. & A. Seragnoli" Institute of Hematology (IRCCS Azienda Ospedaliero-Universitaria di Bologna, Italy) and the CoMMpass (CoMM) cohort, a widely used public database of MM patients with associated genomic and clinical data, available online by request.

The two cohorts of SMM samples were included in this study thanks to the collaboration with the Irene Ghobrial Laboratory at the "Dana-Farber Cancer Institute" of Boston (BUS cohort and SU2C cohort). (Table 1)

Cohort name	Patients	Phase	Platform	Alteration types available	Origin
CoMM	**832**	NDMM	WGS	CNA + Mutations	**CoMMpass**
BO	**750**	NDMM	SNP arrays	CNA	**Bologna**
BUS	**171**	SMM	Exomes	CNA	**DFCI Ghobrial Lab**
SU2C	**114**	SMM	WGS	CNA + Mutations	**DFCI Ghobrial Lab**

Table 1: *summary of the four cohorts of patients included in this study. Two disease phases (NDMM and SMM) were analyzed, every phase includes two different cohorts, for a total of 1582 NDMM patients and 285 SMM patients. In this study three different genomic platforms were employed to analyze the patients' samples.*

3.1.1 Bologna NDMM cohort

Patients: This cohort included 750 newly diagnosed MM patients, whose CD138+ cell fractions were available at the time of diagnosis. The cohort included 370 and 70 patients previously enrolled in the EMN02/HO9524 and in the GIMEMA MM-BO200525 clinical trials, respectively, as well as 310 patients consecutively treated in our Institution in the context of the daily clinical practice. Median progression free survival (PFS) and overall survival (OS) with respective Inter Quartile Range (IQR) were expressed in months. For the whole cohort of patients, they were respectively 43 (IQR:17-67) and 63 months (IQR: 31-78). Patient baseline clinical characteristics are summarized in Table 2. All patients provided signed consent for the genomic analyses.

BO dataset

Var	*Daily Practice BO n/N (%)*	*Median*	*IQR*	*EMN02 n/N (%)*	*Median*	*IQR*	*BO2005 n/N (%)*	*Median*	*IQR*	*P.Value*
Male	41/73 (56%)	-	-	214/370 (58%)	-	-	47/70 (67%)	-	-	ns
Female	32/73 (44%)	-	-	156/370 (42%)	-	-	23/70 (33%)	-	-	ns
Age, years	72/73 (99%)	66	58.75 - 72.25	370/370 (100%)	58	52-62	70/70 (100%)	59	54 - 62	<0.0001
Age > 65 years	38/72 (53%)	-	-	4/370 (1%)	-	-	0	-	-	<0.0001
Beta 2 microglobulin, mg/L	65/72 (89%)	4.3	3.2 - 6.8	370/370 (100%)	3.5	2.433 - 5.2	70/70 (100%)	3.17	2.4 - 4.6	0.004
Beta 2 microglobulin<3.5 mg/L	25/65 (38%)	-	-	183/370 (49%)	-	-	47/79 (67%)	-	-	0.003
Beta 2 microglobulin >5.5 mg/L	23/65 (35%)	-	-	85/370 (23%)	-	-	9/70 (13%)	-	-	0.008
Albumine, g/dL	64/73 (85%)	3.8	3.375 - 4.1	370/370 (100%)	3.8	3.39 - 4.2	70/70 (100%)	3.94	3.51 - 4.28	ns
Albumine <3.5 g/dL	21/64 (33%)	-	-	110/370 (30%)	-	-	16/70 (23%)	-	-	ns
Creatinine, mg/dL	72/72 (99%)	0.99	0.78 - 1.337	367/370 (99%)	0.9	0.71 - 1.1	70/70 (100%)	1	0.8 - 1.2	0.009
Haemoglobin, g/dL	72/73 (99%)	10.5	9.4 - 11.7	370/370 (100%)	10.95	9.6 - 12.4	70/70 (100%)	11.3	9.9 - 12.2	ns
Haemoglobin < 10.5 g/dL	36/72 (50%)	-	-	155/370 (42%)	-	-	27/70 (39%)	-	-	ns
Platelet count, 10^3/mL	71/73 (97%)	214	176.5 - 275.5	370/370 (100%)	230.5	176 - 278	70/70 (100%)	225	183.2 - 285.8	ns
Platelet count<150 10^3/mL	11/71 (15%)	-	-	51/370 (14%)	-	-	7/70 (10%)	-	-	ns
Lactate dehydrogenase, g/dL	56/73 (77%)	166.5	142 - 207.25	355/370 (96%)	450	240.5 - 480	65/70 (93%)	265	205 - 349	<0.0001
LDH, Upper Limit	7/46 (15%)	-	-	101/260 (39%)	-	-	47/57 (82%)	-	-	<0.0001
Bone marrow plasma cells >60%	26/58 (45%)	-	-	205/358 (57%)	-	-	37/63 (59%)	-	-	ns
IG Isotype IgG	42/70 (60%)	-	-	219/351 (62%)	-	-	45/70 (64%)	-	-	ns
IG Isotype IgA	14/70 (20%)	-	-	73/351 (21%)	-	-	13/70 (19%)	-	-	ns
IG Isotype BJ	13/70 (19%)	-	-	53/351 (15%)	-	-	12/70 (17%)	-	-	ns
Light Chain Kappa	44/68 (65%)	-	-	NaN	-	-	37/62 (60%)	-	-	ns
Light Chain Lambda	24/68 (35%)	-	-	NaN	-	-	25/62 (60%)	-	-	ns
ISS I	22/67 (33%)	-	-	139/370 (38%)	-	-	31/69 (45%)	-	-	ns
ISS II	20/67 (30%)	-	-	143/370 (39%)	-	-	28/69 (41%)	-	-	ns
ISS III	25/67 (37%)	-	-	88/370 (23%)	-	-	10/69 (14%)	-	-	0.008
R-ISS I	8/44 (18%)	-	-	56/348 (16%)	-	-	NaN	-	-	ns
R-ISS II	28/44 (64%)	-	-	241/348 (69%)	-	-	NaN	-	-	ns
R-ISS III	8/44 (18%)	-	-	51/348 (15%)	-	-	NaN	-	-	ns
t(4,14)	6/66 (9%)	-	-	58/356 (16%)	-	-	26/69 (38%)	-	-	<0.0001
t(6,14)	NaN	-	-	3/353 (1%)	-	-	NaN	-	-	ns
t(11,14)	1/40 (3%)	-	-	80/355 (23%)	-	-	NaN	-	-	0.0004
t(14,16)	4/60 (7%)	-	-	18/354 (5%)	-	-	NaN	-	-	ns
t(14,20)	NaN	-	-	7/353 (2%)	-	-	NaN	-	-	ns
FISH Deletion 13	14/50 (28%)	-	-	182/353 (53%)	-	-	26/56 (56%)	-	-	0.007
FISH Deletion 17p	12/67 (18%)	-	-	41/354 (12%)	-	-	3/56 (5%)	-	-	ns
FISH Deletion 1p36	7/59 (12%)	-	-	47/352 (13%)	-	-	NaN	-	-	ns
FISH Amplification 1q	21/62 (34%)	-	-	137/352 (39%)	-	-	10/24 (42%)	-	-	ns
FISH Hyperdiploidy	8/48 (17%)	-	-	170/350 (49%)	-	-	NaN	-	-	<0.0001
Induction (PI)	22/72 (31%)	-	-	370/370 (100%)	-	-	NaN	-	-	<0.0001
Induction (Imid)	9/72 (13%)	-	-	NaN	-	-	25/70 (36%)	-	-	<0.0001
Induction (PI - Imid)	37/72 (51%)	-	-	NaN	-	-	45/70 (64%)	-	-	<0.0001
Induction (Other)	4/72 (6%)	-	-	NaN	-	-	NaN	-	-	<0.0001
ASCT	39/64 (61%)	-	-	211/340 (62%)	-	-	64/68 (94%)	-	-	<0.0001
Single ASCT	20/64 (31%)	-	-	108/340 (32%)	-	-	10/68 (15%)	-	-	0.01
Double ASCT	19/64 (30%)	-	-	103/340 (30%)	-	-	54/68 (79%)	-	-	<0.0001
Maintenance	30/59 (51%)	-	-	271/321 (84%)	-	-	47/67 (70%)	-	-	<0.0001
Consolidation	29/59 (49%)	-	-	118/322 (37%)	-	-	54/67 (81%)	-	-	<0.0001
Progression Free Survival	72/73 (99%)	22.5	13 -43	370/370 (100%)	39.5	16 - 68	70/70 (100%)	48.5	23.2 - 83.8	<0.0001
Overall Survival	72/73 (99%)	27.5	11 - 31.5	370/370 (100%)	64	34 - 78	70/70 (100%)	80.5	551.2 - 128.8	<0.0001

Table 2: *All the clinical and baseline variables available to describe the "BO dataset" were included in this study. The number and percentage of patients' data available for any given variable, along with the median value and inter-quantile range (IQR) are showed here, broken down for each of the three cohorts that composes the dataset ("Daily practice BO", "EMN02" and "BO2005"). Fisher's exact test p-values for frequency comparisons of each variable among the three cohorts are shown in the last column.*

Experiments: 750 bone marrow (BM) aspirates were obtained during standard diagnostic procedures. Mononuclear BM cells were obtained by Ficoll-Hypaque density gradient centrifugation. An immunomagnetic beads-based strategy (MACS system, Miltenyi Biotec, Auburn, CA) was employed to isolateCD138+ plasma cells. The purity of positively selected plasma cells was assessed by flow cytometry using a conventional antibody panel. Total genomic DNA was isolated using Maxwell®16 LEV Blood DNA kit (Promega, Madison, WI) and quality/quantity checked by Nanodrop. SNP array profile experiments were carried out according to the manufacturer's protocols (Cytoscan HD Genome-wide Human Gene Chip, Affymetrix, Santa Clara, CA).

3.1.2 CoMMpass NDMM cohort

Download/access: In this study, we used the WGS and WES data obtained from the Multiple Myeloma Research Foundation (MMRF) CoMMpass (Relating Clinical Outcomes in MM to Personal Assessment of Genetic Profile) trial (NCT01454297) database (Interim Analysis 19) (https://themmrf.org/finding-a-cure/our-work/the-mmrf-commpass-study/), which includes whole genome/ exome sequencing data of over 1000 newly diagnosed MM patients with enriched tumor and matched constitutional samples. Somatic CN profiles for the definition of CNAs in MM were generated from 752 NDMM patients from the CoMMpass study, by low coverage long-insert WGS (median 4-8x). SEG files and mutations table were downloaded from the CoMMpass portal, under access request (available at: https://research.themmrf.org/).

Patients: The MMRF CoMMpass study accrued patients from clinical sites in Canada, Italy, Spain and the United States. All patient samples were shipped to one of three biobanking operations: Van Andel Research Institute (VARI) in Grand Rapids, Michigan for all samples collected in Canada or the United States; Salamanca for samples collected in Spain; and Torino for samples collected in Italy (Fig. 12).

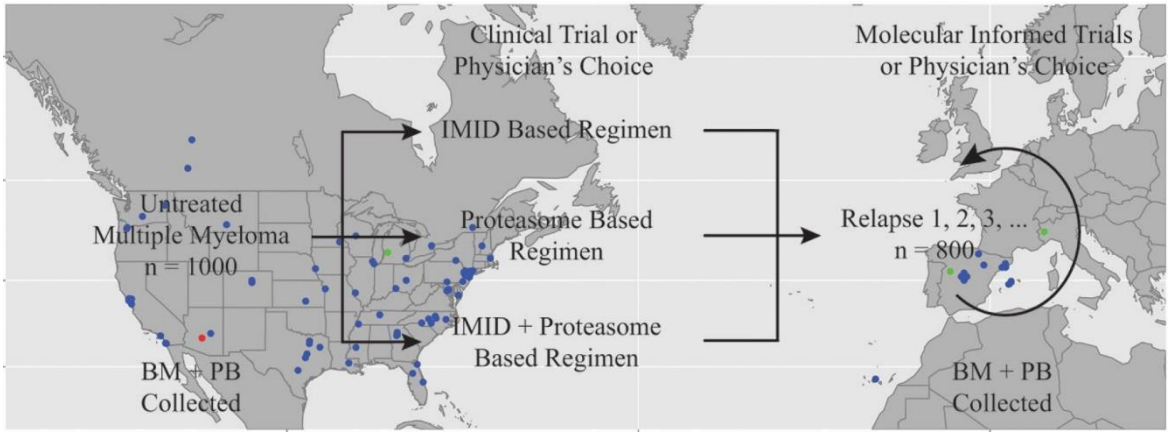


***Figure 12**: Overview of the MMRF CoMMpass Study. The CoMMpass study aimed to enroll 1000 NDMM patients. BM and PB samples were collected and characterized at diagnosis. Patients must have received either an IMiD based, PI based, or combination IMiD and PI based regimen as their first line of therapy. BM and PB samples were also collected and characterized at each progression event, with the aim of collecting data for 800 progression events. Clinical parameters were also collected every 3 months for a minimum 8 year observation period.*

Experiments: The BM samples were gently syringed through a 0.72 mm needle to break up any bony debris and were aliquoted 5 ml at a time into 50 ml conical tubes. Red cell lysis was performed by adding 10-15 volumes of sodium citrate to each tube. The samples and lysis buffer were mixed by inverting the tubes, and lysis was allowed to continue for 15 minutes at room temperature. White blood cells were collected by centrifugation at 1250 rpm for 10 minutes at 25°C after which the supernatant was aspirated off and the pellet suspended in 15 ml of lysis buffer. An aliquot of the cells was removed for n automated cell count on a Coulter AcT diff Analyzer. The remaining cell suspension was filtered through a filter mesh into a 15 ml conical tube and the cells were pelleted at 1250 rpm for 10 minutes at 25°C after which the supernatant was removed by aspiration. The pelleted cells were suspended in80 µl of AutoMACS Running Buffer (Miltenyi Biotec) with 20 µl of anti-CD138 MicroBeads (Miltenyi Biotec) per 2 x 10 7 total cells and were incubated at 4°C for 15 minutes. To wash the cells, 15 ml of AutoMACS Running Buffer was added and, after mixing by inversion, the cells were collected by centrifugation at 1250 rpm for 10 minutes after which the supernatant was removed by aspiration. The cell pellet was suspended in 1 ml of AutoMACS Running Buffer and filtered through a wire mesh before magnetic sorting. Magnetic sorting was performed using an AutoMACS Separator (Miltenyi Biotec) and the positive cell fraction was collected using the appropriate program based on the brightness of CD138 expression and percentage of CD138 + cells, as determined by flow cytometry. The positive fraction count was determined using a Coulter AcT diff Analyzer and aliquots were created for purity analysis along with independent DNA and RNA isolations, as outlined by the North American Biobank. The purity of each enriched CD138 + fraction was determined by flow cytometry using anti-CD138 PE, and only samples with at least 90% positive cells were used. Fractions destined for DNA and RNA extraction were aliquoted into 1.5 ml eppendorf tubes and the cells were collected by centrifugation at 1500 rpm for 5 minutes at 25°C. Supernatants from the DNA fractions were aspirated, and the dry cell pellets were stored at -80°C. Cells designated for DNA extraction were stored as dry pellets at -80°C and extracted with the Qiagen Gentra Puregene Tissue Kit (Qiagen, #158667). DNA pellets were dissolved in Qiagen buffer ATE and stored at -20°C. DNA was quantified by Nanodrop spectrophotometric analysis, as well as by fluorescence using the Qubit 2.0 to determine dsDNA content. Sample quality was determined by agarose gel or Agilent Tapestation Genomic tape. Samples with at least 500 ng of dsDNA were sent to TGen for NGS. The WGS assay used 200 ng of DNA that is fragmented to a target size of 900 bp. LI-WGS libraries were constructed using the Kapa Hyper Prep Kit (Kapa Biosystems, #KK8504). Library molecules were separated on a 1.5% agarose gel. Molecules between 950-1050 bp were either extracted automatically from a Sage Science Pippin Prep 1.5% gel (Sage Science, #CSD1510) or hand punched from a 1.5% agarose gel. One cycle of PCR amplification pre size selection, followed by 6 cycles of amplification post size selection, was performed. Sequencing was performed on Illumina HiSeq2000 or HiSeq2500 instruments at TGen using Illumina HiSeq v3 or v4 chemistry. Diluted library pools with 1% PhiX control libraries were clustered on Illumina cBOT instruments as recommended by the manufacturer. In all cases, sequencing assays utilized a paired-end sequencing format of at least 82x82 nucleotide reads.

3.1.3 Irene Ghobrial's Lab SMM SU2C cohort

Patients: 114 Bone marrow samples were prospectively collected from the Dana-Farber Cancer Institute Observational Precursor Crowd (PCROWD) study (NCT02269592). All patients provided written informed consent for the research use collection of BM samples. Plasma cells were selected with AutoMACS and CD138+ magnetic beads (Miltenyi Biotec).

Experiments: Sorted samples underwent DNA purification (Thermo Fisher PicoPure DNA Isolation Kit) and library preparation using the NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs) with unique dual index adapters (NEBNext Multiplex Oligos) according to the manufacturer's instructions. Final library fragment sizes were assessed using the BioAnalyzer 2100 (Agilent Technologies),with yields quantified by a Qubit 3.0 fluorometer (Thermo Fisher Scientific) and qPCR (KAPA Library Quantification Kit). Final sample libraries were normalized and pooled before WGS was performed on Illumina NovaSeq 6000 flowcells, 300 cycles paired-end reads, at the Genomics Platform of the Broad Institute of MIT and Harvard, targeting 60X genomic coverage for tumor fraction and 30X for matched germline.

3.1.4 Irene Ghobrial's Lab SMM BUS cohort

Patients: A next generation sequencing technology approach was used to study 171 patients with SMM at time of diagnosis. Whole exome sequencing was performed (WES) on 77 matched tumor-normal samples (mean target coverage 109X), and WES on 94 tumor-only samples (with mean coverage 174X). Samples were collected at Dana-Farber Cancer Institute, University College London, Mayo Clinic, and the University of Athens in Greece, in addition to multiple centers in the US and Europe participating in clinical trial NCT02316106. Paired tumor and normal germline DNA were obtained from cases with smoldering multiple myeloma at time of presentation with disease. After approval of the study protocols by the institutional review boards and ethics committees of the participating institutions, samples were obtained after written informed consent. All patients provided signed consent for the genomic analyses.

Experiments: Tumor DNAs were extracted from CD138+ cells from patients' bone marrow. For germline control (normal), DNA was obtained from either buccal mucosa (saliva), or peripheral blood mononuclear cells. Genomic DNA was extracted using QIAamp DNA mini kit (QIAGEN) according to the manufacturer's protocols, and double-stranded DNA concentration was quantified using PicoGreen dsDNA Assay kit (Life Technologies). Libraries were prepared by Agilent SureSelect XT2 Target Enrichment kit. To capture the coding regions, we used the SureSelect XT2 V5+UTR capture probes (Agilent). All sequencing was performed on the Illumina HiSeq 4000 platform at the Broad Institute. For tumor only samples (n= 94), libraries were prepared and hybridized using Agilent SureSelect XT2 V5 capture probes (Agilent).

3.2 METHODS

3.2.1 Genomic data processing and data availability

BO cohort: A raw CEL files was generated for every SNP array experiment. The genomic segments profiles (SEG files) were generated using Rawcopy R package and CBS algorithm, using the reference human genome GRCh37 / hg19. The significance threshold for segmentation (alpha parameter of the CBS algorithm) was set at 10−7.

Data availability: For this study, only SEG files were available to analyze.

CoMMpass cohort: The analysis of all sequencing data is performed at TGen on a high-performance computing system using the automated, "JetStream", analysis system. The workflow supports the analysis of human sequencing samples against the GRCh38 reference genome using ensembl version 98 gene models. The workflow details can be found at: https://github.com/tgen/phoenix .

Data availability: For this study BAM files and SEG files from long insert low-coverage WGS samples were available to analyze. Also, mutational tables from WES samples were downloaded from the CoMMpass web-portal and were available to analyze.

BUS cohort: The output from Illumina software was processed by the Picard data processing pipeline to yield BAM files containing well-calibrated, aligned reads. The Getz Lab CGA WES Characterization pipeline (https://github.com/broadinstitute/CGA_Production_Analysis_Pipeline) developed at the Broad Institute was utilized to call, filter and annotate somatic mutations and copy number variation. The pipeline employs the following tools: MuTect, ContEst, Strelka, Orientation Bias Filter, DeTiN, AllelicCapSeg, MAFPoNFilter, RealignmentFilter, ABSOLUTE, GATK, PicardTools, Variant Effect Predictor, Oncotator.

Data availability: for this study, only SEG files from both tumor-only and matched tumor-normal samples were available to analyze.

SU2C cohort: Short insert paired-end reads/FASTQ files were aligned to the reference human genome (GRCh37) using Burrows–Wheeler Aligner,BWA (v0.5.9). Picard was applied for post-alignment procedures as sorting, indexing, and marking duplicates. The alignments were submitted to base quality score recalibration (BQSR) by using the Genome Analysis Toolkit (GATK) version 4. MuTect and GATK (Haplotype Caller) were used for the single nucleotide variant calling. GATK variants were filtered with the Variant Quality Score Recalibration tool following the best practices on the GATK website. GATK performs the variant calling and filtration in the normal and tumor samples independently, thus the subtraction between the tumor and the normal variants resulted in our first set of candidate somatic variants. To ensure the somatic classification of the SNVs called by GATK, we adapted the Mutect algorithm and applied its LOD_N classifier after the GATK variant calling and filtering. The LOD_N is a Bayesian classifier that compares the likelihood of two models: (1) the mutation does not exist in the normal sample and all non-reference bases are explained by sequencing noise, and (2) the mutation truly exists in the normal sample as a germ-line heterozygous variant. The ratio of these two likelihoods is called LOD (Log Odds) score

and when it exceeds a decision threshold, the mutation can be classified as somatic. For this filtering, we considered only sites that had total read depth greater or equal than 8 in the normal sample and greater or equal than 14 in the tumor sample. Our final candidate list consisted in the union of MuTect and GATK-LOD_N results. The variants were annotated by ANNOVAR, with the Ensembl Gene annotation database for human genome build 38 (http://www.ensembl.org/), and searched for matches in the dbSNP151 and 1000 Genomes data. We selected exonic single nucleotide variants (SNVs) that were non-synonymous, splicing variants or gain or loss of stop codon. Variants present in dbSNP151 and 1000 Genomes with minor allele frequency (MAF) greater than 0.05 were removed. Pathogenic variants were identified by selecting only the variants annotated in clinical databases, such as CLINVAR or COSMIC.

Data availability: for this cohort all the generated FASTQ, BAM files, VCF files and SEG files were available to analyze.

3.2.2 Existing bioinformatic tools

3.2.2.1 GISTIC2

in order to extract focal significant regions of known and novel CNA in the MM genomic landscape, the researcher applied the GISTIC [50] (Genomic Identification of Significant Targets in Cancer) v2 tool on the NDMM SEG files samples. Thanks to the high number of samples in the BO and CoMM cohorts, the researcher was able to maximize the GISTIC analysis resolution and the statistical power (especially for a confident identification of rare MM CNA events) of the analysis. In fact, GISTIC is a computational algorithm that is used to statistically identify and interpret regions of the genome that are recurrently altered in cancer. GISTIC takes as input a matrix of copy number data from a set of samples, where each element in the matrix represents the copy number of a specific genomic interval in a specific sample (SEG files). The algorithm then identifies regions of the genome that are recurrently altered across the set of samples, and assigns a significance score (G-score) to each region based on the frequency and magnitude of the alterations. The algorithm consists of several steps: 1) The input matrix of copy number data is processed to identify regions of the genome that are altered in at least a certain percentage of the samples. 2) The algorithm then assigns a score to each region based on the frequency and magnitude of the alterations, taking into account the underlying copy number variation. 3) The algorithm then identifies recurrently altered regions by applying a threshold on the score, and clusters the regions into "amplification" and "deletion" groups. These regions are called "peak regions". 4) The algorithm then refines the boundaries of the regions by considering the strength and spatial continuity of the alterations. 5) Finally, the algorithm assigns a significance score to each region based on the frequency and magnitude of the alterations, and the background copy number variation. In particular, it assigns a "q-value" to each peak region, which is a measure of the statistical significance of the alteration in that region. The q-value is a measure of false discovery rate (FDR) and ranges between 0 and 1, with lower q-values indicating more significant regions. [50]

GISTIC is used to identify genomic regions that are recurrently altered in cancer, which can help in identifying potential targets for cancer therapy. The algorithm is widely used in cancer genomics studies and has been applied to multiple types of cancer, including breast, lung, prostate, and colon cancer.[50]

Multiple iterations of GISTIC analysis, with different parameters were tested in order to archive an optimal resolution on both the BO and CoMM cohorts analysis. The final GISTIC analysis was performed employing the following input parameters: Join_segments_size = 50, Focal_lenght_cutoff = 0.25, Q-value = 0.01, arm_peel = YES, sample_centering = NO. All the other input parameters were selected as default. In order to correct the false-positives generated by the germline CNVs present in the SNP arrays samples, a list of regions to filter out in the analysis was created by using the DGV (Database of Genomic Variants) version 107 database (available at http://dgv.tcag.ca/dgv/app/home) and selecting for all the region reported in the database in at least 100 samples, across at least 2 different studies which show a reported MAF > 5% in the human population. This filter file was used as input to the GISTIC analysis for the BO cohort analysis.

3.2.2.2 ABSOLUTE

for the task of extracting purity solutions in the cohorts where BAM files were available (SU2C and CoMMpass) the ABSOLUTE [60] tool was employed with default parameters. The purpose of ABSOLUTE is to extract purity and ploidy data from the admixed population of cancer and normal cells in the sample. This process begins by generation of segmented copy number data (SEG files), which is input to the ABSOLUTE algorithm together with pre-computed models of recurrent cancer karyotypes and, optionally, allelic fraction values for somatic point mutations. The output of ABSOLUTE then provides re-extracted information on the absolute cellular copy number of local DNA segments and, for point mutations, the number of mutated alleles.
In this way, the researcher obtained a total of 1172 ABSOLUTE solutions, which were manually inspected in order to select the most appropriate one. This manual review is particularly critical and suggested in samples presenting a complex karyotype, where the ABSOLUTE algorithm outputs low-confident solutions.

3.2.2.3 BradleyTerry2

The Bradley-Terry model [61], also known as the Bradley-Terry-Luce model, is a statistical model used to analyze paired comparison data.[62] It's implemented in R statistical language by the *BradleyTerry2* package.[63] This model assumes that in a "contest" between any two "players", say player i and player j ($i, j \in \{1, \ldots, K\}$), the odds that i beats j are α_i/α_j , where α_i and α_j are positive-valued parameters which might be thought of as representing "ability".

A general introduction can be found in Bradley (1984)[64] . Applications are many, ranging from experimental psychology to the analysis of sports tournaments to genetics (for example, the allelic transmission/disequilibrium test of Sham and Curtis 1995 [65] is based on a Bradley-Terry model in which the "players" are alleles). In typical psychometric applications the "contests" are comparisons, made by different human subjects, between pairs of items. Finally, this model was

also recently applied in the field of genomic alterations timing in cancers.[33] In this context, the "players" are represented by the different genomic alterations, which can be compared in subclonality "contests". Given the fact that clonal alterations happen earlier then sub-clonal ones, in a Bradley-Terry genomic timing model an alteration "wins" if it is found to be more clonal than another one (showing a higher CCF).

The Bradley-Terry can alternatively be expressed in the logit-linear form:

$$\text{logit}[\text{pr}(i \text{ beats } j)] = \lambda_i - \lambda_j$$

where $\lambda_i = \log \alpha_i$ for all i. Thus, assuming independence of all contests, the parameters $\{\lambda_i\}$ can be estimated by maximum likelihood, using standard software for generalized linear models with a suitably specified model matrix. In fact, the Bradley-Terry model is typically fitted to the data using maximum likelihood estimation (MLE), which finds the values of the abilities that maximize the likelihood of the observed data given the model. Once the abilities are estimated, they can be used to make predictions about the outcomes of future pairwise comparisons between the items. The primary purpose of the *BradleyTerry2* package, implemented in the R statistical computing environment, is to facilitate the specification and fitting of such models and some extensions.

Since the Bradley-Terry model employs a 'reference' or 'base' player, on which all the other player abilities are computed. A problem known as the "reference category problem" arises.[66] In order to resolve this problem and to enable the direct comparison between all the players abilities, the *BradleyTerry2* package computes Quasi-Variances (and corresponding quasi-Standard Errors) associated to the Bradley-Terry players abilities, computed by using the *qvcalc* function. [67]

3.2.3 Newly developed bioinformatic tools

RemasterCNA: the RemasterCNA analysis to correct CN profiles for hypersegmentation bias were performed by adapting the parameters to the resolution and mean sequencing quality (MAD or MAPD) of the specific cohorts:

- focalDef (Width in Mbp of regions to consider as small / focal noise) = 1 Mbp for CoMM and BUS cohorts, 0.5 Mbp for BO and SU2C cohorts,
- jump_definition (minimum CN distance required to define a confident breakpoint) = median MAD /MAPD of the cohort (BO = 0.15, CoMM = 0.12, SU2C = 0.08, BUS = 0.15)

BoBaFIT: the BoBaFIT analysis to correct for the baseline region bias were performed using the standard pipeline of the two functions (computeNormalChromosomes and DRrefit) with default parameters as described in the paper.[68]

DRrefit parameters:

- maxCN = 6

- clust_method = "ward.D2"

computeNormalChromosomes parameters:

- tolerance_val = 0.15
- maxCN = 6
- min_threshold = 1.6
- max_threshold = 2.4

RAPH: the RAPH analysis to correct for the purity bias were performed using the standard parameters for all the cohorts, as following:

- clust_type (clustering algorithm of choice) = "DBclust",
- dbs_eps (DBscan minimum search distance) = 0.08,
- dbs_min (DBscan minimum points in a cluster) = 1,
- minimum_region (minimum width of chromosome regions considered as independent observations) = 5*10^6,
- min_purity (minimum purity of the sample) = 0.20

3.2.4 Clinical and statistical analysis

All the analyses were conducted using R language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria) version 4.2. The analysis was performed with a significance level of at least 0.05 and all variables objected of inference were reported together with their 95% confidence intervals (CI). Classifications between samples and patients were explored by comparing characteristics between groups with non-parametric methods such as Kruskal-Wallis's test on the medians (or the parametric t-test on means). For the parametric correlations tests the Pearson’s test was employed, while for non-parametric correlations tests Spearman's rho was employed.

PFS was measured in months, from the start of therapy to the event of first progression of the disease or death. OS considered death as outcome/event and was measured from the same landmark. Univariate survival analysis on both PFS and OS were performed by the Kaplan-Meier method, as for drawing the survival curves. Semi-parametric Cox regression analysis was adopted to estimate hazard ratios (HR) with an 95% CI between predefined possible prognostic groups. Multivariate analysis was performed again by Cox regression analysis to identify the abnormalities independently affecting the prognosis with their HR and 95%.

4 RESULTS

4.1 DEVELOPMENT OF A BIOINFORMATIC PIPELINE FOR MULTI-PLATFORM HARMONIZED CN ANALYSIS AND COHORT-TIMING.

4.1.1 PHASE 1: Data cleaning and harmonization

To perform a reliable timing of CNA alterations in MM and generating comparable timing estimates across the various cohorts included in this study, the integration and the harmonization of copy number data deriving from the 4 cohorts and generated using different genomic platforms was considered a critical requirement.

The harmonization of genomic data is an important and critical issue frequently faced both in bioinformatics and in translational research. In fact, in the era of big-data, the need to integrate data generated by multiple studies and public datasets in order either to create validation cohorts for hypotheses to be proven or to enrich the main study cohort by obtaining greater statistical power, is increasingly present and evident.

At a bioinformatic level, the harmonization should start from the "raw files". In the simplest and most common case, where all data are generated by NGS, raw files are represented by FASTQ (raw sequences) files, or by BAM (aligned sequences) files (Fig. 13). However, in practice, this type of harmonization is rarely feasible, due to the following technical difficulties:

1. high files' dimensions, which, in the case of medium coverage WGS (30x), might reach storage sizes of about 100 GigaBytes per sample (100 TeraBytes for 1000 samples)[69], whose management requires very high economic budget to get storage servers or cloud storage services (about 2000,00$/ month, in the case of raw data storage of the CoMMpass dataset on Amazon AWS S3 Standard bucket). [70]
2. Computational power and times required for preliminary analyses: only high-performance computing (HPC) systems or workstations would be able to process pipelines for raw data analysis, often requiring several hours to complete the analysis of a single WGS sample.[69]
3. Data privacy and security: Integrating raw data from different studies might raise concerns about data privacy and security, as the FASTQ/BAM files contains sensitive information about genetic variants (potentially disease-associated SNPs) of individuals.

Furthermore, when data from different platforms should be harmonized and merged, as in the present study (e.g. NGS platform and SNP array platform), it is not possible to start the analysis from raw files, since the original raw data structures are different.

For these reasons, in the present study, a different strategy was identified, in order to harmonize the Copy Number data, as detailed below:

1. use the first intermediate files that show a common data structure among the analysis pipelines of the different platforms as starting copy number data: these files are the SEG

files, since they have the same structure, and are produced by both the NGS and SNP arrays pipelines (Fig. 13);

2. analyze and correct the data downstream the platform-specific pipelines, using and/or developing a series of *ad-hoc* bioinformatics tools and packages that identify the quality of the starting data and correct the identified methodological and technical biases;
3. merge the harmonized copy number data, enabling the use of a multi-platform dataset for performing further analysis (including Copy Number timing analysis).

In general, starting the copy number analysis from SEG files is a very convenient strategy in order to obtain a data harmonization, since all genomic platforms' pipelines are able to generate this type of file: not only SNP arrays and WGS/WES, but also Ultra-Low-Pass WGS and targeted-Seq (from off-target reads) (Fig. 13).

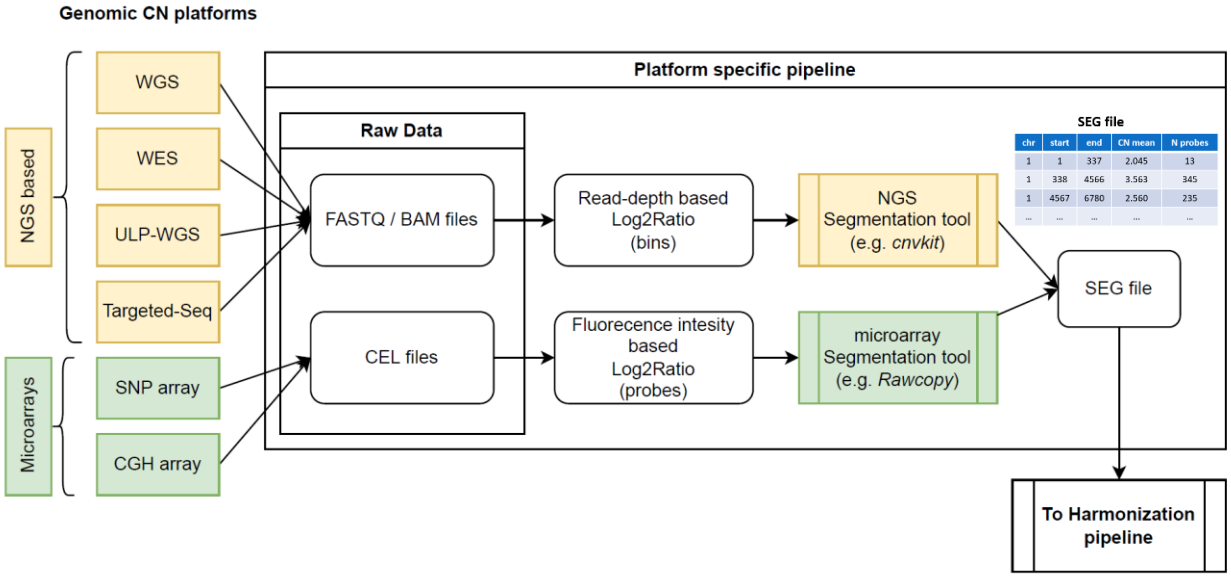


***Figure 13**: the process of generation of the SEG file, starting from all platforms capable to produce genomic Copy Number information. NGS based platforms and tools are colored in yellow, while microarray based platforms and tools are colored in green.*

Additionally, since SEG files, similarly to VCF files, are just simple .tsv or .csv text tables, with one row per chromosome segment (usually no more than a few hundered rows, depending on the complexity of the karyotype) they have the advantage of very small file size (usually about 100 KiloBytes per sample - in comparison to WGS FASTQ/BAM files they are 1 million time smaller), which enable a simple and fast processing and transferring, even in standard desktop computers. They can be easily loaded and visualized using the popular genomic browser IGV (Integrative Genomic Browser - available at https://igv.org/app/) and uploaded on cloud version-control platforms (such as GitHub or GitLab), facilitating the reproducibility of the analysis.

However when harmonizing SEG files from different platforms, is necessary to note that, since they have been generated by different bioinformatic tools, the algorithm performing the actual

segmentation could be different: even if the most popular algorithm is Circular Binary Segmentation (CBS), other algorithms exists, such as the Piecewise Constant Fitting algorithm (used in the ASCAT R package) of the Fused Lasso Regression algorithm (implemented in cghLasso R package).[71] Importantly, since the use of different segmentation algorithms can introduce methodological biases in the data integration process, all the SEG files that need to be harmonized should be segmented by a tool that employs the same algorithm. This check was performed for the data included in the present study, ensuring that the CBS algorithm was employed to generate all the four SEG files of the samples' cohorts to harmonize (Table 3).

COHORT	PLATFORM	SEGMENTATION TOOL	SEGMENTATION ALGORITHM
BO	SNP arrays	Rawcopy [45]	Circular Binary Segmentation
COMM	WGS (low coverage)	Allelic CapSeg [48]	Circular Binary Segmentation
BUS	WES	Cnvkit [47]	Circular Binary Segmentation
SU2C	WGS	Allelic CapSeg [48]	Circular Binary Segmentation

Table 3*: The different platforms and segmentation tools used in each cohort to generate the Copy Number profiles (SEG files) used in this study as input starting data.*

In the first phase (phase 1) of the developed multi-platform CN pipeline (Fig. 14) the input SEG files were subjected to a data-cleaning and harmonization procedure consisting in 4 steps. Each step is associated to a specific identified bias (Table 4) that can potentially affect the CN data quality and, consequently, the downstream CNAs calls and timing analysis. Therefore, in order to correct the data for each bias, four different bioinformatic tools were developed.

IDENTIFIED BIAS	TYPE OF BIAS	DESCRIPTION	EXISTING TOOL	NEW DEVELOPED TOOL
HYPERSEGMENTATION BIAS	Technical	Too many artificial segments in the CN profile, due to poor DNA quality or bad wet-lab processing.	/	RemasterCNA
BASELINE REGION BIAS	Technical	Bad mathematical centering of the baseline region of the CN profile.	/	BOBaFIT
TUMOR PURITY BIAS	Biological	Sub-optimal enrichment of the tumor sample, normal cells contamination.	ABSOLUTE ASCAT SEQUENZA	R.A.P.H.
ERROR ESTIMATION BIAS	Methodological	Not considering the platform resolution, not computing confidence intervals for CN estimates in each segment.	ABSOLUTE	ComphyNumber

Table 4: *four different identified biases in CN data which can possibly affect the data quality in the study. In order to correct for some of these biases (i.e. tumor purity and error estimation) some bioinformatic tools already exist (e.g. ABSOLUTE), while for other biases no tools are currently available to the knowledge of the researcher. In order to perform a full harmonized bias correction process on CN data, four new different bioinformatic tools were developed, each one aimed to correct a specific different CN data bias.*

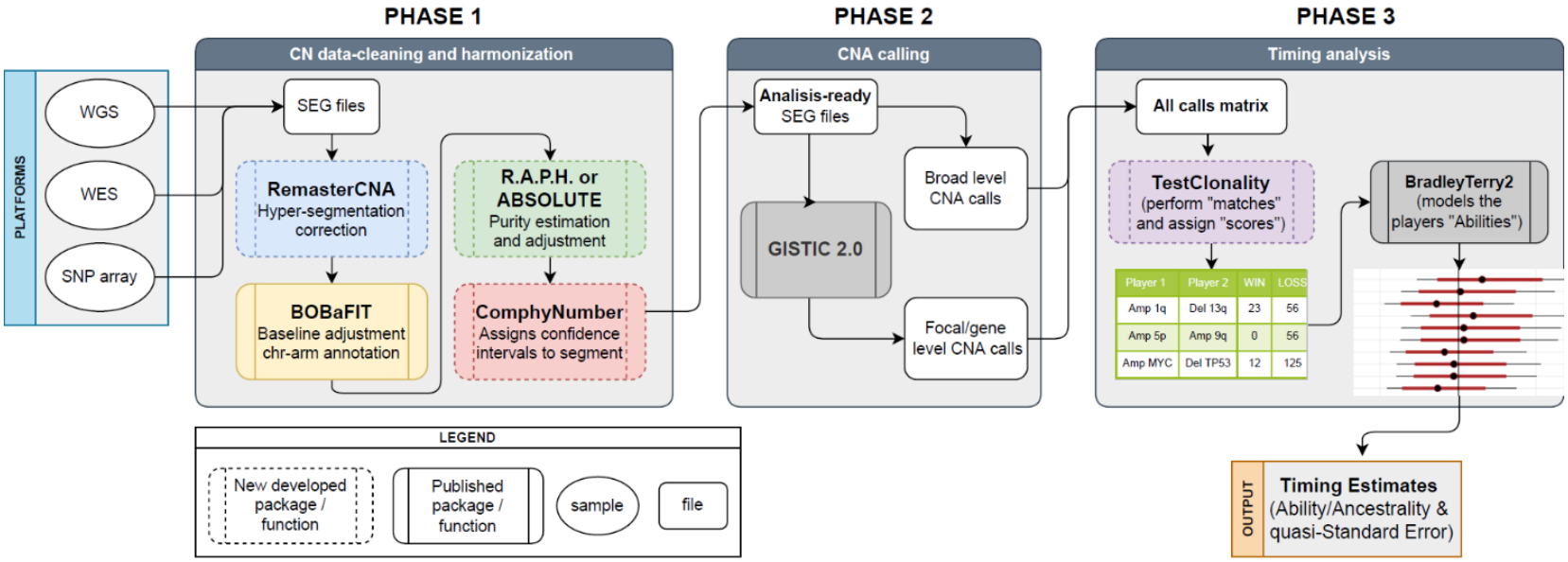


Figure 14: *The complete multi-platform bioinformatic pipeline used in the CNA timing analysis. Data inputs (segments files) generated by multiple platforms were cleaned and pre-processed to ensure homogeneity and harmonization in the Phase 1 of the pipeline ("data pre-processing"). Both arm-level and focal-level CN alterations were called in the Phase 2 of the pipeline ("CNA calling"). Finally, in the phase 3 of the pipeline ("Timing analysis") focal and arm-level calls were merged in a comprehensive calls matrix that was used to compute all the clonality "matches" between calls, thus generating a matches result table which, in turn, was used as the input for the Bradley-Terry league model in order to generate the final Timing Estimates.*

4.1.2 RemasterCNA: a tool to correct the hypersegmentation bias in CN profiles

RemasterCNA is a noise-reduction algorithm created to resolve the "hypersegmentation" bias in CN profiles. The observed «Hypersegmentation phenomenon» is defined as an abnormally high number of segments in a SEG file (usually thousands instead of hundreds), and it is probably generated by a poor starting DNA quality or caused by technical artifacts due to wet-lab sample processing. [72] Additionally it's also possible that a suboptimal parameters choice in the segmentation algorithms might cause a higher number of segments then expected (e.g. the *alpha* parameter used in DNAcopy that regulates the significance levels for the test to accept a breakpoint between two different segments[46]). This highly-skewed number of segments in SEG files have the possible effect of disturbing the constant signal of the CN profiles, thus leading to false CNAs calls. (Fig. 16)

Importantly, this bad-quality effect is hidden and not captured by the standard quality metrics commonly used to assess copy number quality (MAD or MAPD), since it does not affect the log2ratio of probes or bins where those quality metrics are computed, but only the segments subsequently computed from the log2ratios signals. To date, no algorithm or bioinformatic tool is able to solve this type of bias, and commonly bad-segmented copy number profiles are excluded from or not detected in studies involving copy number analysis.

In this study we identified the samples with highly-skewed number of segments by analyzing the distribution of the number of segments in the four cohorts. First, a scree plot was generated for each cohort, then an elbow in each distribution was identified (Fig. 15). In this way it has been possible to detect 3.7%, 0.6%, 0% and 2.7% of samples in the BO, CoMM, Bus, and SU2C respectively, that showed an abnormal number of segments when compared to the base cohort distribution.

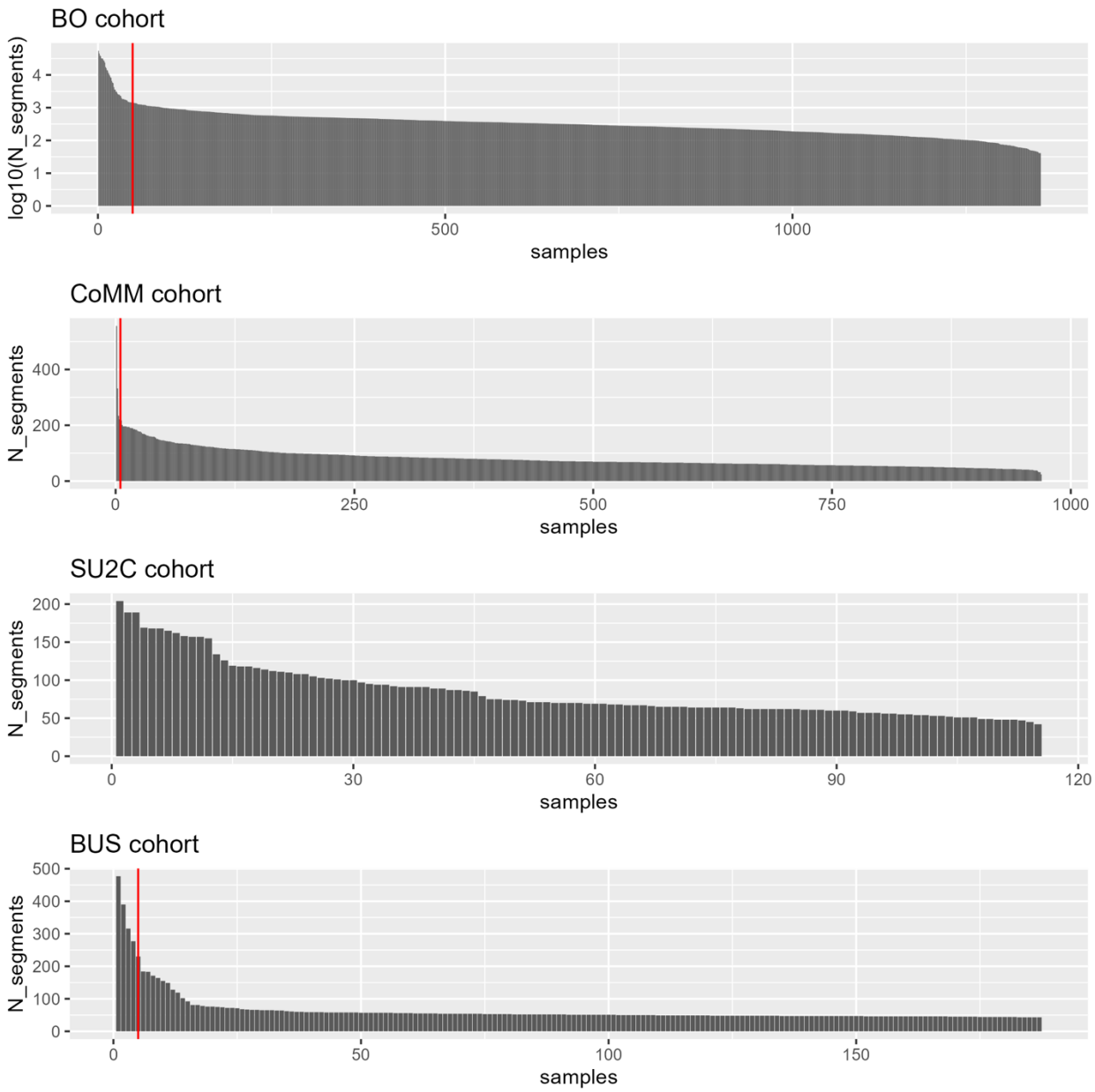


Figure 15*: scree plots that describe the number of segments of the samples included in each of the four cohorts of the study. The red line shows the selected elbow of the distribution.*

Even if number of hypersegmented samples is quite low, instead of discarding those samples, a new tool has been developed, able to correct the SEG files for this bias. In fact, the aim of this study is to gather the biggest cohort of samples as possible, especially given the scarcity and the value of copy number profiles available in the SMM cohorts. The tool was implemented in a R function (that will be published in a R package) and was named *remasterCNA*. This function uses as input the SEG file that the user aims to correct, and it outputs a new hypersegmentation-

corrected SEG file. Its functioning relies on a “dynamic programming approach”, that is, solving a problem by breaking it down in simpler subproblems. In fact, *remasterCNA* consists of 3 sub-algorithms aimed to resolve 3 different subproblems:

- **FocalCleaner**: Cleans the «focal-noise» generated by the presence of extremely small segments. The maximum segment size of such small segments is based on the user defined parameter “width”, which can be easily defined based on the resolution of the platform that generated the input SEG file.
- **BreakPointDetector**: Identifies confident breakpoints, based on the absolute CN distance between every pair of two adjacent segments. If the distance is larger than the user defined parameter “jump” a confident breakpoint is generated. The “jump” parameter can be easily defined based on MAD or MAPD dispersion value (quality metric) of the SEG file.
- **Legolizer:** finally, the algorithm compacts all the cleaned segments between the previously defined confident breakpoints in new “blocks”. The CN value of the new blocks is computed by using a weighted-mean approach (where the weight is the length of the segments in the block, and the value is the CN of the segments).

By using *remasterCNA* on the previously defined bad-quality samples, it has been possible to reduce the number of segments, homogenizing the number of segments according to the mean of the cohorts, while also preserving the Copy Number information required to call CNAs event, as shown in Figure 16.

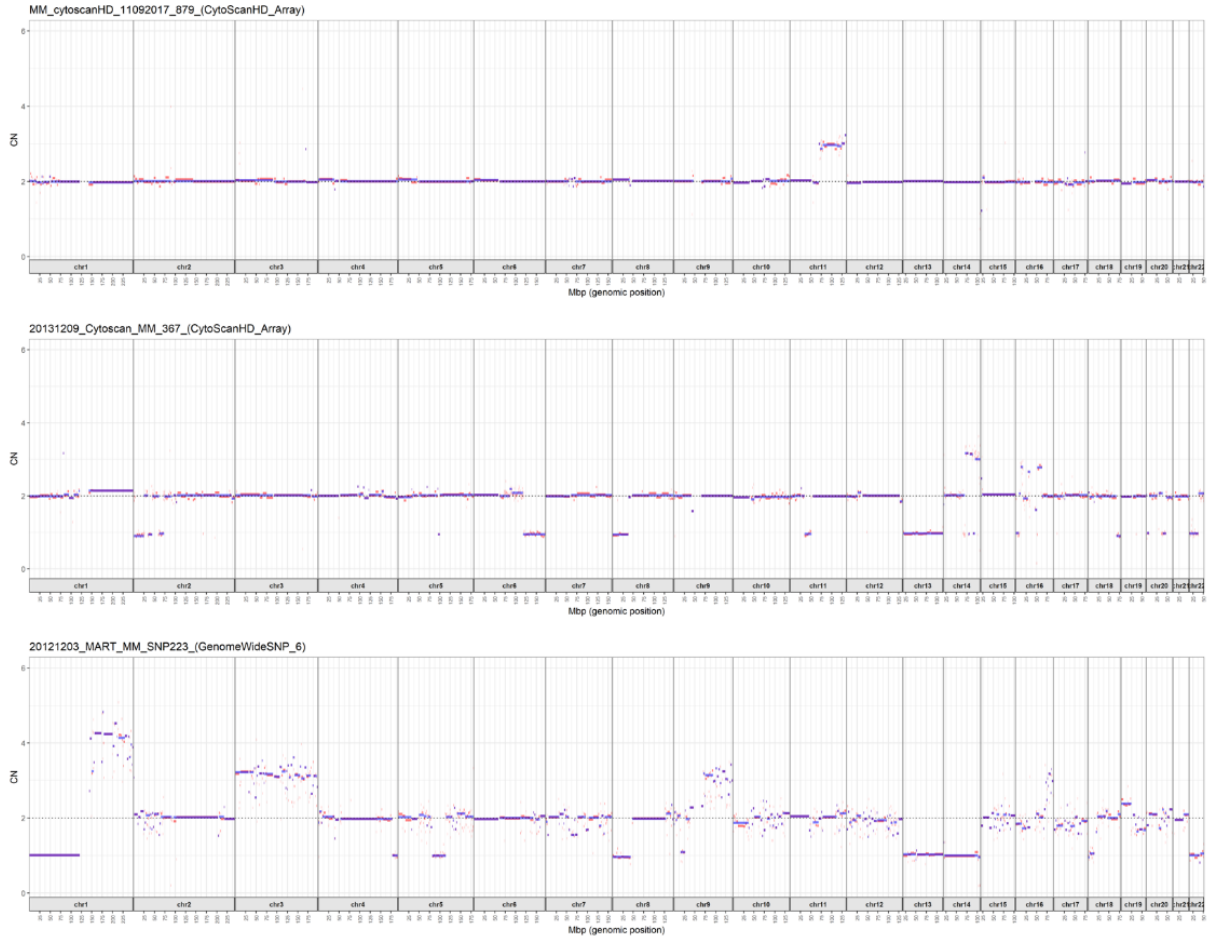


***Figure 16**: remasterCNA anecdotal results on three example samples which show the hypersegmentation phenomenon in their CN profile. Both pre-correction CN profile (orange track) and post-correction CN profile (blue track) are visible and overlapped in the plots. As it's possible to notice, in all three samples the number of segments is evidently heavily reduced, while the global CNAs structure is preserved.*

4.1.3 BOBaFIT: a published R package to refit the baseline region of CN profiles

Another independent bias that was identified by visually inspecting the Copy Number profiles of all the cohorts in this study is the "baseline region bias". This bias is caused by segmentation tools that use the standard "median-centering" normalization method to estimate the baseline region (usually the region with ploidy = 2), assuming that the average value corresponds to the theoretical "2", thus they might erroneously estimate the regions with diploid CN.[73] This might happen particularly in samples with a complex CN profile, either carrying several and/or large chromosomal aberrations. When tumors display complex karyotypes, this "median-centering" approach could fail the baseline region estimation and consequently cause errors in generating the CNAs profile (SEG file). To overcome this issue, we designed and published in the widely known bioinformatic repository "BioConductor" (https://bioconductor.org/packages/devel/bioc/html/BOBaFIT.html) an innovative R-package, named *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. A scientific paper describing this package in detail was also published on the journal "Computational and Structural Biotechnology" and available online on PubMed on 2022 July 3 (Fig. 17). [68]

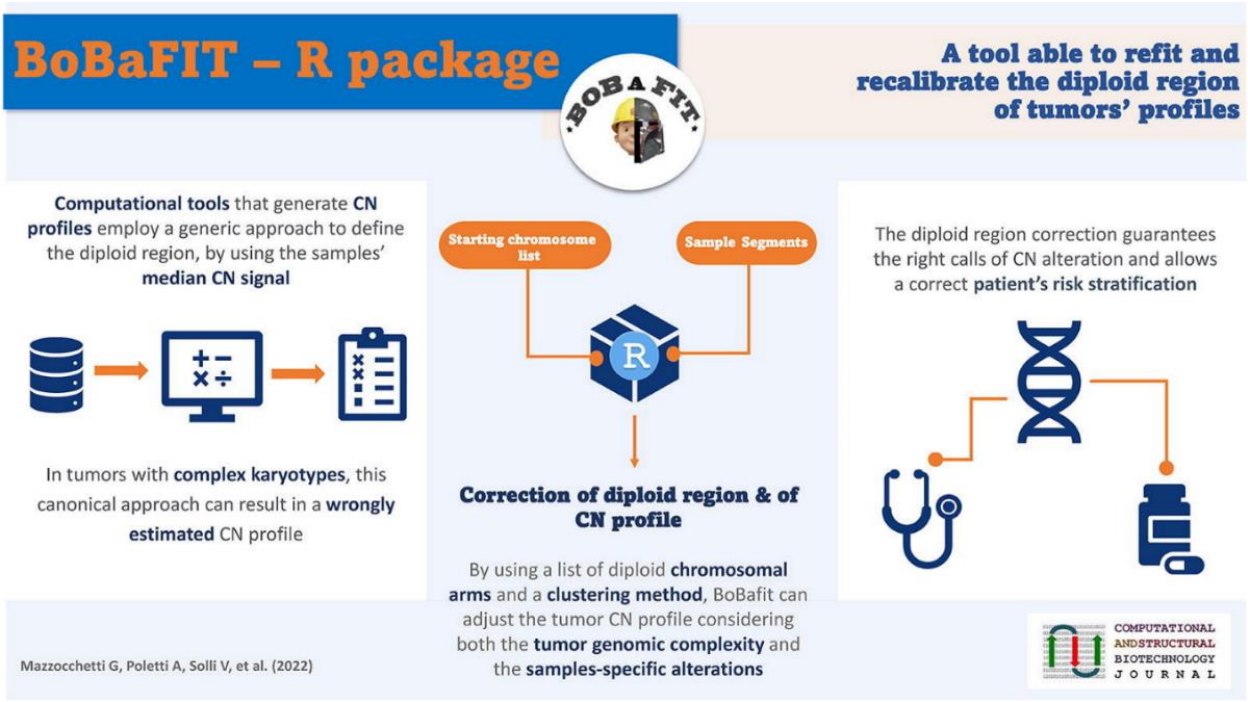


***Figure 17**: Graphical abstract describing the main features and functionalities of BoBaFIT*

The principal function of *BoBaFIT* is named “DRrefit”: throughout a tumor and sample-specific approach it adjusts the CN values of the input SEG file. In addition, BoBafit contains a secondary function “ComputeNormalChromosome” that generates the starting chromosome list (S-CL), important input of DRrefit and cornerstone of the tumor-specific strategy (Fig.18).

To create a tumor and sample-specific method aimed at checking and adjusting the tumor CN profile, DRrefit uses two inputs: (1) the SEG file, 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 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.

The algorithm performs the following steps for each sample (Fig. 18):

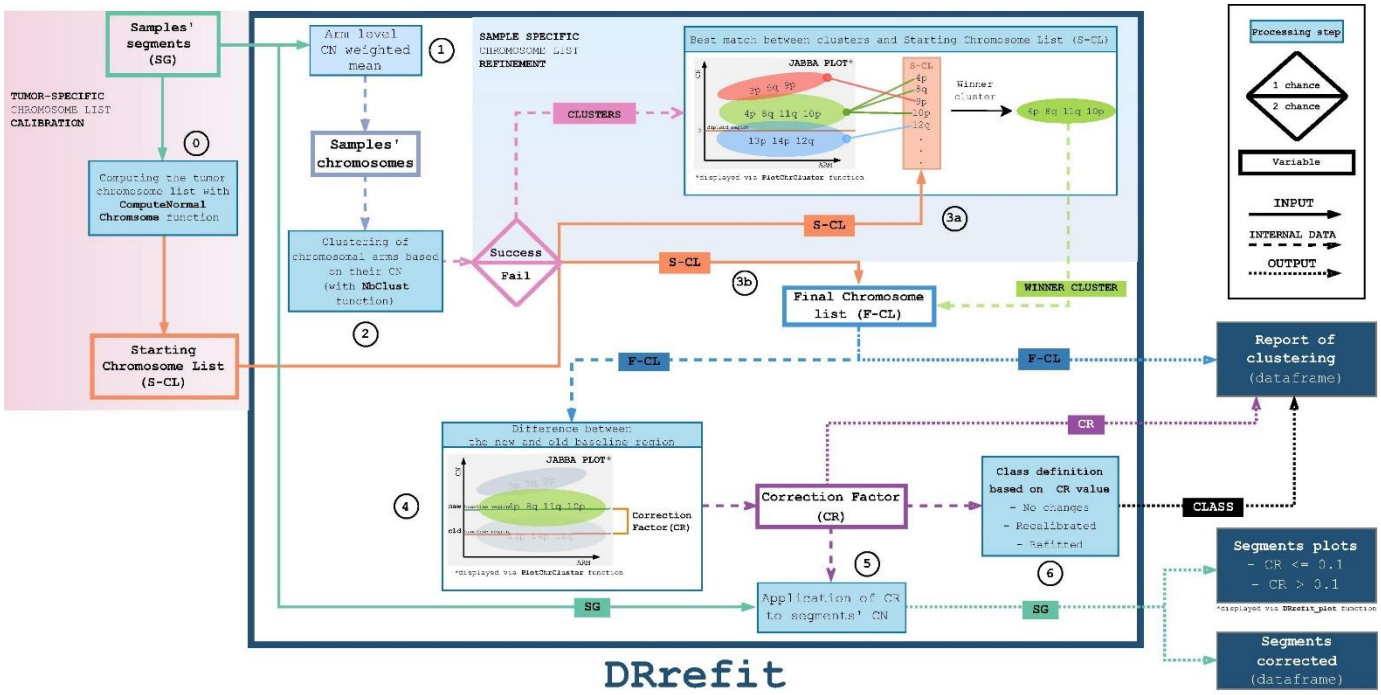


***Figure 18**: BoBafit package workflow. The diagram shows how to organize a BoBafit analysis and DRrefit algorithm steps. 0) First of all, from the SEG file, the tumor specific chromosome list has to be obtained by ComputeNormalChromosome function; 1) Next, the CN mean, weighted on the segments' length, 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. A plot, outputted by a BoBafit 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 (usually CN = 2 or 4) and the median CN value of F-CL (new baseline region). Again, a 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.*

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[74], 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 [75], 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. 18), is chosen as the "winner cluster" and it becomes the "final chromosome list" (F-CL). This step defines the "sample-specific refinement" (Fig. 18) 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 plot 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 [75], 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. 18).

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 *BoBafit* package vignette. 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.19B and 19C); (2) the "CR ≤ 0.1" plot (Fig. 19A), with overlapping segments.

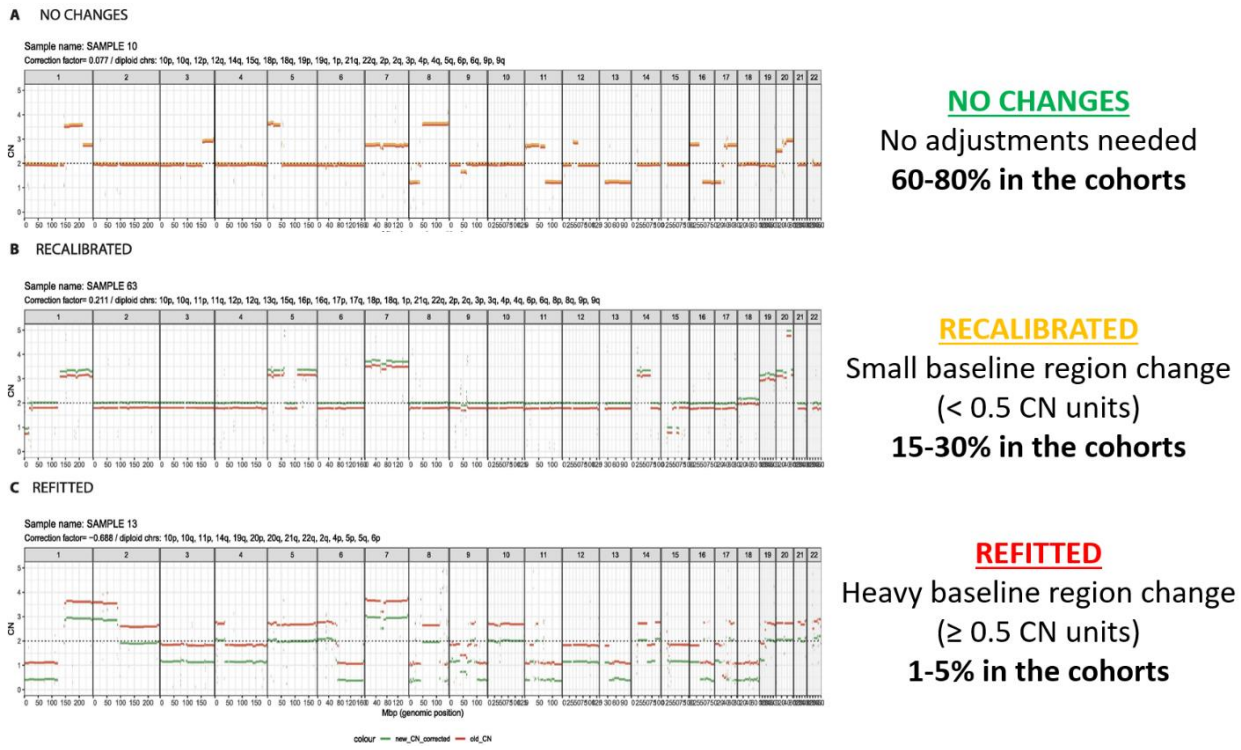


***Figure 19**: DRrefit CN profile plots of three samples, labeled with class identified by the function. In the panel are showed the tree DRrefit classes and how they are plotted. The x-axis reports the chromosomes with their genomic position and the y-axes the copy number value. The plots with CR ≤ 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. On the right, the frequency of the three classes are reported for the samples included in the four cohorts of this study and corrected with BoBafit.*

Consequently, all the SEG files from the four cohorts of this study were corrected using *BoBaFIT* to control for the baseline region bias. Frequency of corrections were approximately the same among all the four cohorts: NO CHANGES = 73.5% on average (range 63.1 – 85.4%), RECALIBRATED = 19.5 % on average (range 14.6 – 34.0%), REFITTED = 3.2% on average (range 0.9 - 6.1%) (Fig. 19).

4.1.4 RAPH: an easy-to-apply and universal purity estimation tool

RAPH (Rapid Assessment of Purity-estimates by Heuristics) is a R function (currently in development to become a R package) and a web-app tool created to resolve the purity bias in SEG files, and consequently to adjust the CN signal relatively to the inferred purity level and the technology platform used for generating the SEG file.

The purity bias is due to non-tumor cell contamination in tumor samples. In fact a correct measurement of the precise level of CNAs subclonality is complicated by the fact that tumors samples often contain multiple populations of both tumor and non-tumor cells (normal cells). In MM, this might be due a not-optimal quality of the plasma cell enrichment procedure, performed prior to sequencing (i.e. enrichment of human tumor cells from primary specimens using cell isolations technologies, for example the Magnetic Activated Cell Sorting (MACS) instrument in this study), or, in solid tumors, to an imperfect biopsy.

The two most popular and widely used tools able to compute the purity level of tumor samples and to correct the CN profile for the purity bias are *ABSOLUTE* [60] and *ASCAT* [76]. Even if both tools are able to analyze both NGS and SNParrays samples, they employ different statistical and computational approaches to assess purity, depending on the sample's platform. In fact, ASCAT is available in two different versions: ASCAT2 (for SNP arrays) where the input BAF track is provided by the user, and ascatNGS (for NGS) where the BAF track is automatically generated from the input BAM file. On the contrary, ABSOLUTE relies on both mutations data (MAF file) and CN data (SEG file) for purity estimation in the NGS version, while relies only on CN data in the SNP array version (Table 5).

In the context of the present study, the harmonization between the purity solutions obtained among all the platforms was intrinsically difficult, since:

a) in an *ASCAT* scenario, the BAM files were not available for all the samples for the NGS version (no BUS cohort BAM), and the BAM files of the CoMMpass cohort show a very low coverage (4-8X) not sufficient to generate a high quality BAF track (not comparable to the superior quality of BAF tracks generated by SNP arrays).
b) in an *ABSOLUTE* scenario, two different computational strategies would be used to generate purity solutions for either NGS and SNP arrays, that is using mutational and CN data for the former, while using only CN data for the latter, introducing in this way a methodological bias and invalidating the aim of this part of this study.

Since neither of the two tools could be homogenously applied to all the samples included in the cohorts, we sought to develop a hybrid strategy to harmonize the purity estimates among all the samples, as follows: 1) Apply the tool of choice (*ABSOLUTE*) on the cohorts were the samples' raw data was available (SU2C and CoMMpass), 2) Develop a new tool, (*RAPH*), capable of computing purity solutions starting from the samples' SEG files, available for the platforms in this study. 3) Validate the *RAPH* computed purity solutions comparing them to the *ABSOLUTE*

computed solutions in the cohorts where both tools can be applied (SU2C and CoMMpass). 4) After a satisfying validation, compute the purity solution with RAPH in the cohorts where ABSOLUTE can't be applied

Purity estimation tool	Input required	Supported platform	unique solutions	Automatic identification of challenging samples	ability to review solutions	Speed of analysis
ABSOLUTE (NGS and SNP array versions)	SEG file (+ MAF file in NGS version)	WGS, WES, SNP array	No	No	Yes	~5-10 min per sample (depending on computing power) + manual review
ASCAT (NGS and SNP array versions)	LogR file + BAF file (in SNP array version) or BAM file (in NGS version)	WGS, WES, SNP array	Yes	No	No	~5-30 min per sample (depending on computing power)
RAPH	SEG file	WGS, WES, SNP array, CGHarray, ULP-WGS, Targeted-Seq	Yes	Yes	Yes (with RAPH-Graph)	1-3 seconds per sample on standard Desktop Computer

Table 5*: comparison between the main features of ABSOLUTE and ASCAT (the two most used tools for purity estimation in tumor samples) and RAPH. The main advantage of RAPH is the ability to be applied on samples generated by every platform that produce a CN profile, since it just requires a SEG file as input. Additionally, since the SEG files are extremely lightweight, the computing process is extremely fast (requiring only a few minutes to analyze thousand samples), facilitating in this way the analysis of big cohorts and the reproducibility of the analysis.*

The *RAPH* tool development started from CN events clustering issue, similarly to what happens in the previously described *BoBaFIT* algorithm. The CN events in *RAPH* are computed at the chromosome-arm level, similarly to *BoBaFIT*. The difference in *RAPH* lies in the CN events measuring here a "deviation from the baseline" metric, measured in CN units (figure 20). For example, a deletion CN value of 1.2 (that is 0.8 units of deviation from the baseline level of CN = 2) is measured like an amplification CN value of 2.8, since it also has 0.8 units of deviation from the baseline level of 2. This choice was motivated by the fact that both deletion and amplification events contribute to generate clusters of CNAs events in the deviation from baseline space. The key observation is that those clusters correspond to the various clonal and/or subclonal states present in the tumor, similarly to what happens when analyzing clonal and subclonal mutations clusters in the CCF space (see *ABSOLUTE* methods and output plots [60]).

Importantly, the identification of the particular "clonal cluster", among all the defined CNAs clusters in the deviation from baseline space, is very important for the purity estimation task, since:

A) if the sample is perfectly pure (100% purity) the deviation from baseline of the clonal cluster of amplifications and deletions would be = 1,
B) on the contrary, if the sample is affected by normal cells contamination, the purity of the tumor sample would be equal to the deviation from the baseline of the clonal cluster of amplifications and deletions. This is because the deviation is less evident the more normal cells contaminate the sample (in the extreme case that a sample would be completely contaminated, the deviation of the clonal cluster would be equal to 0). For example, a sample with a 50% tumor purity and 50% normal cells contamination will show a clonal cluster deviation of 0.5, while a sample with a 80% tumor purity and a 20% normal cell contamination will show a clinal cluster deviation of 0.8.

In conclusion, in order to compute a purity value starting from a SEG file it is necessary to accomplish two tasks: 1) reliably cluster the CNAs events in the deviation from baseline space, 2) reliably identify the "clonal cluster" among all the identified clusters.

Regarding the first task, similarly to the previously described *BoBaFIT* algorithm, *RAPH* uses a NBclust-based clustering approach, in order to identify the various levels of CN events the sample CN profile. The CN events are computed at the chromosome arm level, similarly to *BoBaFIT*. This clustering method is already proven to be capable of reliably identifying the clonal structure, as previously described.[68]

Regarding the second task, we developed a set of three simple logical rules, based on our empirical experience in reviewing purity solutions, which can be implemented in a decision-tree algorithm in order to identify the "clonal cluster". The three rules can be applied on the list of all the clusters present in any given sample, annotated with information on the numerosity and the distance to the integer CN for each cluster in the list (figure 20).

- **RULE A**: The "big-cluster" of alterations is chosen as clonal cluster. Here, a "big" cluster is defined as a cluster showing at least 2 times more CN events then the median of all the clusters (without considering the baseline region cluster), or showing at least 4 events in different chromosomes including both an amplification and a deletion event (figure 20A).

- **RULE B:** If multiple big-clusters are detected, or no big-clusters are detected, the (big-)cluster with the minimum distance to the integer CN is chosen as the clonal cluster (figure 20B).

- **RULE C:** Purity adjusted CN profiles cannot biologically have negative CN values. If the chosen clonal cluster defines a purity that generate negative CN values, it must be wrong. In this case, exclude the chosen cluster from the clusters list and start over with rules A and B.

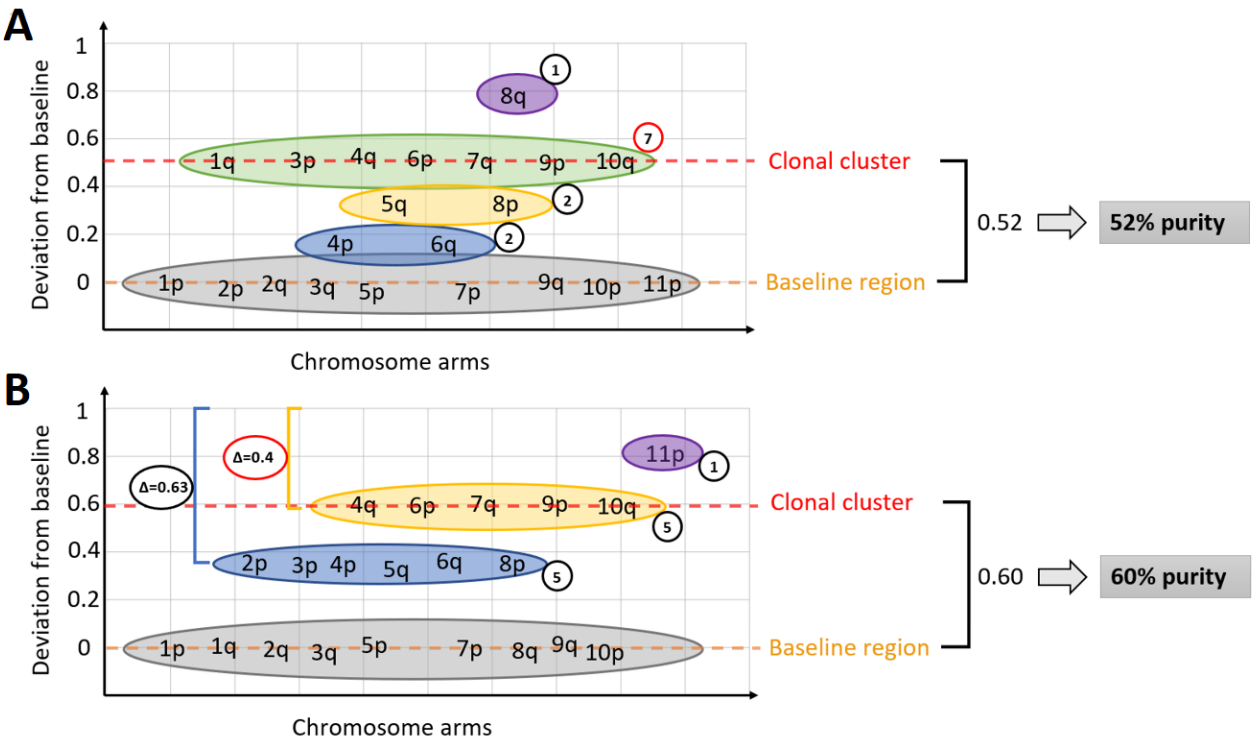


***Figure 20**: the functioning of the RAPH rules for identifying the clonal cluster, and the purity of the sample. CN events are measured in CN units of deviation from baseline, and clustered by using the same approach used in BoBaFIT. A) By using the "rule A" the green cluster is chosen as the clonal cluster, since it shows at least 2 times the median number of CN events among all the clusters (excluding the baseline region cluster, in grey). B) Both yellow and blue clusters are defined as big-clusters by the "rule A", so by using the "rule B" the big-cluster with the minimum distance from the integer CN (the yellow cluster with a distance of 0.4) is chosen as the clonal cluster.*

In addition to the purity solutions and the purity-corrected SEG files, *RAPH* outputs also a table containing the clustering quality statistics of the analyzed samples. This allows to automatically detect complex samples (i.e. samples in which the purity solution is challenging to assign confidently, due to a particularly complex karyotype that reflects to an uncommonly high number of clusters (>6) or to a bad cluster dispersion quality metric (Standard Deviation of CN events in cluster > 0.05)). In order to resolve the purity value of these challenging cases, we developed a Shiny web-application[77] named "RAPH-Graph" (freely and openly available at: https://shirke019.shinyapps.io/RAPH_Graph/) that enables a manual review of any given CN profile. The *RAPH-Graph* allows to upload the user SEG files directly on the web-app. Next, after sample selection, the CN profile of the sample is plotted, and a simple point-and-click interface facilitates the choice of the purity value identified by the user. The operation can be repeated for every sample included in the SEG file. The solutions can be downloaded in the web interface. Figure 21 show a screenshot of RAPH-Graph and illustrates the procedure for selecting the purity for one example CoMMpass sample.

***Figure 21**: a screenshot of the developed web-application "RAPH-Graph". This tool enables a fast and easy visualization of the uploaded SEG file. The user can select of a purity solution (highlighted in the plots with a orange line) by clicking on the CN profiles plots (blue = amplificated segments, red = deleted segments, green = normal segments). The obtained solutions can be both annotated and then downloaded by using a button in the web interface (grey button at the bottom of the web page). The screenshot shows the purity selection for the sample MMRF_1030_1 of the CoMMpass cohort (purity = 0.85, chosen by clicking on the big cluster of hyperdiploid blue chromosomes).*

In order to validate the *RAPH* and purity solutions, a validation cohort was defined (SU2C cohort) in which both RAPH and ABSOLUTE tools could be best applied. The comparison, showed in figure 22A, between the purity solutions generated by both tools indicates a very good overall concordance of the solutions (Pearson's R = 0.87, p<0.0001), indicating a nice quality of *RAPH* solutions when compared with a state-of-the-art widely used tool. In particular, the few discordant samples highlighted with labels in figure 22A,B, are due to the *ABSOLUTE* selection of a purity solution based on the mutation information, which is not available in the SEG file and thus not exploitable by *RAPH* for computing a purity solution. This was demonstrated by comparing ABSOLUTE with manually reviewed purity solutions just by using the CNA information: the same samples were found to be discordant between ABSOLUTE and the CNA-based manually reviewed solutions (figure 22B), confirming that RAPH can correctly identify the right purity solutions when based on the available CN information in the SEG file (figure 22C) (Pearson's R = 0.98, p<0.0001) .

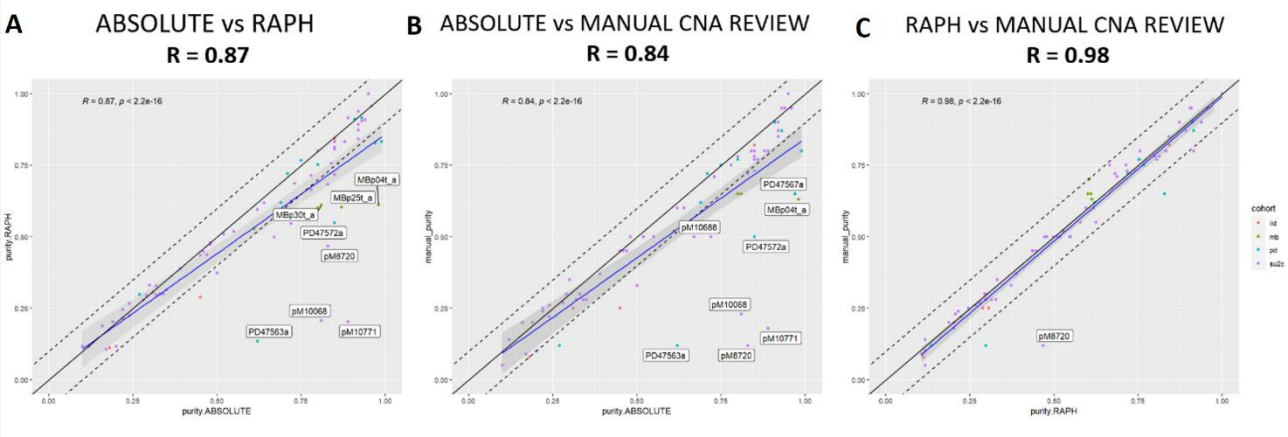


Figure 22: *validation of RAPH purity solutions. Comparison between RAPH and ABSOLUTE solutions on the validation SU2C cohort. A) The direct comparison between the two tools shows a great overall concordance. B) A comparison between ABSOLUTE and the purity solutions generated by reviewing CN profiles shows that the discordant samples were not due a wrong estimation of the CN profiles, but due to the ability of ABSOLUTE to infer purity also from mutations data. C) The comparison between RAPH and the purity solutions generated by reviewing CN profiles shows an almost perfect concordance.*

Furthermore, we sought to demonstrate the ability of RAPH in correctly assessing the purity solution just by using CN information. To this aim we performed a comparison between RAPH solutions and manually reviewed solutions from CN profiles, in a cohort of the first 250 samples from the CoMMpass cohort (figure 23A). This comparison showed an extremely good concordance (Pearson's R = 0.94, p<0.0001), with the only discordant samples being associated to a complex karyotype (figure 23B blue points), as defined by the RAPH's output clustering statistics. This result further demonstrated the potential of RAPH in extracting the correct purity solution when analyzing the CN profile contained in the SEG file.

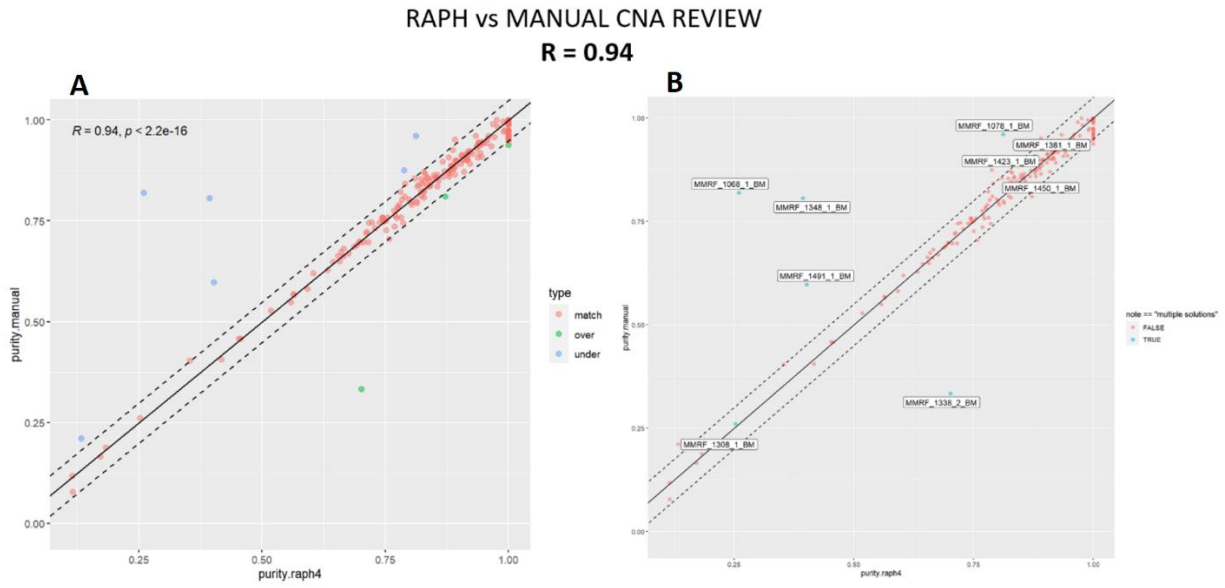


Figure 23: *comparison between RAPH solutions and manually reviewed purity solution in the first 250 samples from the CoMMpass cohort. A) RAPH shows an extremely good concordance with only 5 underestimated samples and 1 overestimated sample. B) The discordant sample are due complex karyotypes in the CN profiles, as defined by the label "multiple solutions" in the RAPH clustering quality output (blue points).*

According to this validation and thanks to the very good concordance between RAPH and ABSOLUTE, the newly developed tool was applied to the cohorts where it was not possible to apply ABSOLUTE, namely BO and BUS. This allowed to get the best harmonized purity solutions and to correct all the samples in the four cohorts from the normal cell contamination / purity bias.

4.1.5 ComphyNumber: a tool to compute confidence intervals to CN estimates

The last step of the harmonization pipeline is to correct the CN profiles for the error estimation bias. This methodological bias arises by the fact that, except for *ABSOLUTE*, all the existing CN computing tools (to our knowledge) do not consider the specific platform resolution, when generating CN segments, and therefore do not compute the confidence intervals (95% confidence intervals referred as "95CI") along with the point estimates of CN or Log2R for each segment of the CN profile. Even ABSOLUTE, even though it assigns a 95CI to every generated segment, does not always compute the correct 95CI, since it employs a default and unchangeable resolution value, instead of dynamically adjusting this parameter to the proper platform resolution.[60]

The segments generated from different platforms or at different seq-coverage can present a (very) different resolution, for example the Affymetrix CytoscanHD SNParray have a resolution of 1 probe every ~1 kB, whereas ULP-WGS have a resolution of 1 probe (bin) every ~2000 kB. In the era of cheap low-resolution platforms for the generation of CN profiles (such as ULP-WGS or targeted-seq) the assignment of a specific 95CI is particularly critical, since the low resolution of the platforms implies large bin sizes (> 1 or 2 MB), hence a low confidence in the assignment of CN values, especially for small segments that contain only few probes. This can cause a high uncertainty when calling the exact subclonality level of CNAs, since the point estimate CN call of segments with a low number of probes is rarely precise and could actually range from a minimum to a maximum, depending on the number of probes (considered as statistical observations) contained in each segment (Fig.24).

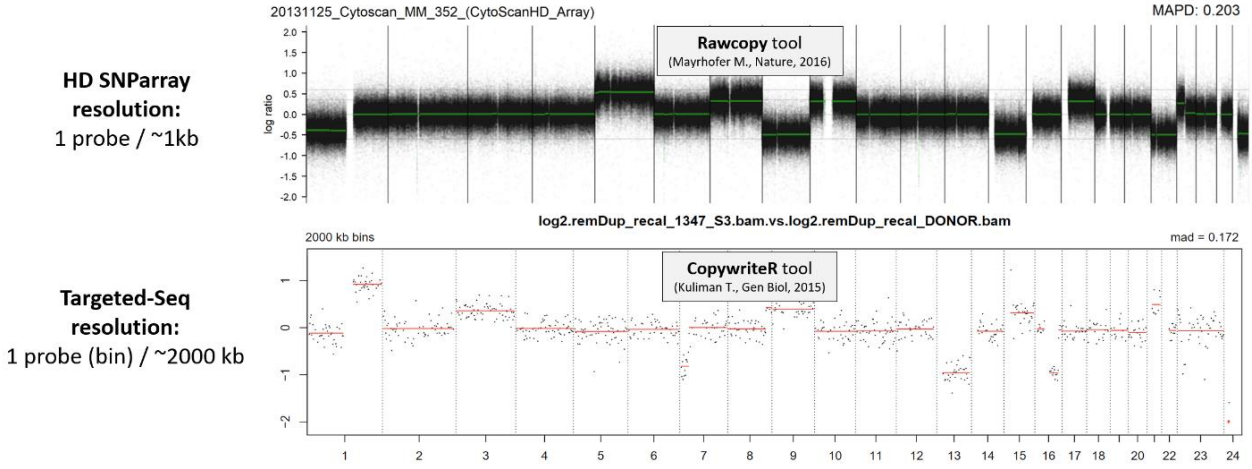


Figure 24: *Two CN profiles generated by two different tools (upper profile: Rawcopy, lower profile CopywriteR). The two profiles show an extremely different resolution, that can be noticed by the evident difference in the number of probes (black dots) that form the log2R track of the profiles. None of the two tools output a confidence interval for the segments they generate (green segments in the upper profile show a elevate number of probes per segment, red segments in the lower profile show a way smaller number of probes per segment), despite being extremely popular and widely used tools.*

Therefore, a new R function was developed, named *ComphyNumber*, in order both to resolve the confidence interval bias and to implement the CNAs calls of this study with a precise error estimation, required later in this study to statistically compare the subclonality level between CNAs with a high precision, thus enhancing in this way the timing results precision.

In order to compute confidence intervals for a SEG file CN point estimates, the following approach was employed. The approach starts from the statistical textbook definition of Confidence Interval (Eq. 1), where $\bar{x}$ is the sample mean, z is the z-score for the chosen confidence interval, s is the sample standard deviation and n is the sample size.

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}} \qquad \textit{Eq. 1}$$

Next, it's possible to observe that in the minimum required information included by every CN segmentation tool that generate a SEG file (i.e. chromosome, start position, end position, CN or log2ratio point estimate and number of probes (or bins) in the segment, Table 6), notably, some of the required information necessary to compute a 95CI for each segment in already available in the SEG file itself:

- The parameter $\bar{x}$ of eq. 1 corresponds to the CN mean, or CN point estimate, which is a mandatory feature of a CN SEG file (fourth column of Table 6).
- The parameter *n* of eq. 1 corresponds to the number of probes included in a segment, which is also a mandatory feature of a SEG file (fifth column of Table 6).

chr	start	end	CN mean	N probes
1	1	337	2.045	13
1	338	4566	3.563	345
1	4567	6780	2.560	235
...	...	...	...	...

Table 6: *example of a standard SEG file with CN information. Five fields are mandatory and found in all SEG file generated by every CN segmentation tool, namely: chromosome, start position, end position, CN or log2ratio point estimate and number of probes (or bins) in the segment*

Then, since the aim is to obtain 95% confidence intervals, the z parameter of eq. 1 is set to a value of 1.959 (Z-score for this confidence level).

Finally, the only missing parameter in eq. 1 required to compute 95CI is *s*, the standard deviation of the sample. Given that this specific statistic is very rarely computed in the sample quality scores of the CN tools, it is possible to derive it from other more common sample statistics often computed by CN segmentations tools. These statistics are the MAPD (mean absolute pairwise deviation, usually computed by SNParray segmentation tools, such as *Rawcopy*, Figure 24) or the MAD

(mean absolute deviation, usually computed by NGS segmentation tools, such as *CopywriteR*, figure 24). Given that the MAD is a consistent estimator the standard deviation of a normal distribution (Eq.2), under minor deviations of the normal distributions its asymptotic variance is smaller than that of the sample standard deviation.[78] Thus, the standard deviation (SD) can be computed from the MAD or the MAPD as in Eq. 2.

$$SD \cong \frac{MAD}{\sqrt{2/\pi}} \qquad \textit{Eq. 2}$$

Finally, both eq. 1 and eq. 2 were implemented in a customized R function, named *ComphyNumber*. The function is able to take as an input a SEG file, and generates an output consisting of a new SEG file which includes a "95CI" variable, corresponding to the 95% confidence intervals of the CN point estimate values for all the segments included in the input SEG file. Thus, by applying this function to all SEG files in the four cohorts of this study, it was possible to correct them for the error estimation bias.

4.1.6 PHASE 2: CNA calling

In the second phase of the pipeline for multi-platform harmonized Copy Number analysis, the goal is to generate homogeneous CNAs calls from the harmonized SEG files outputted from the phase 1 of the pipeline (Figure 14). In order to do so, the researcher judged an appropriate choice to distinguish between "Broad level" or chromosome-arm level CNAs, and "focal level" or gene level CNAs. This reasoning was inspired by the GISTIC algorithm functioning, which also perform a similar discrimination in its computation of CNAs present in the samples.[50] It's also proven that the specific biological mechanisms that generate CNAs can be various and different in nature. Each mechanism generates CNAs of typical sizes and type depending on the specific biological process involved (e.g. mitotic non-disjunction mechanism generate large "broad level" CNAs, while fusion-bridge-amplifications generates small "focal level" CNAs). [79]

4.1.6.1 Broad alteration calls

In order to compute CNA calls for broad level (or chromosome-arm level) alteration events, a specific strategy implemented in a R script was developed. This script executes the broad CNA call procedure, as described in the following pseudo-code:

Pseudo-code for extracting broad CN calls from SEG files

Input: SEG files of samples, containing CN values, CN deviation, chromosome arms and 95CI.

Output: list of broad-level CN calls, including the associated 95CI.

```
For each sample s in cohort do:
        pur ← Extract the purity value of the sample s
        cutoff ← Define a dynamic CN cutoff based on the sample purity: cutoff = 0.1/pur
        for each chromosome-arm arm in s do:
                AltSeg ← filter segments in arm with a CNA event: CN deviation > cutoff
                if number of AltSeg > 0 do:
                        wmCN ← Compute the (size)weighted-mean of the CN values in AltSeg
                        wmCI ← Compute the (size)weighted-mean of the 95CI values in AltSeg
                        Save wmCN and wmCI in the output list
                end if
        end for
end for
```

4.1.6.2

By applying this script to all the SEG files of the samples, it was possible to generate broad-level CN calls with included 95% Confidence Interval. A total of 39 CN calls per sample were computed in this way (figure 25).

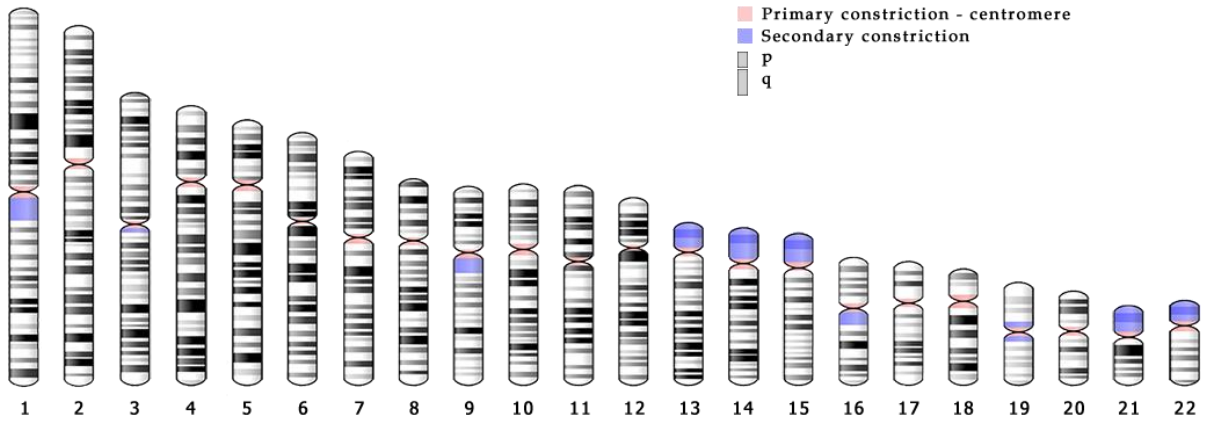


***Figure 25**: Human autosomic chromosomes ideogram (schematic representation of chromosomes). For each autosomic chromosome arm (p and q arms) a broad-level CN call was produced. The 13q, 14q, 15q, 21q, 22q chromosome arms were not included. A total of 39 CN calls per samples were generated in this way.*

4.1.6.3 Focal alteration calls

In order to compute CNA calls for focal level (or gene level) alteration events, a specific strategy implemented in a R script was developed. This script, in addition to the SEG files, uses as an input a pre-computed BED files that describe the specific genomic loci of interest that the user want to extract. The loci should be annotated with the information: Chromosome, start position and end position. Of note, it is critical that the positions must be coherent with the reference genome used for generating the SEG files (usually hg19 or hg38 for the human genome).

This script executes the focal CNA call procedure, as described in the following pseudo-code:

Pseudo-code for extracting focal CN calls from SEG files

Input: SEG files of samples, containing CN values and 95CI.

Input: list of genomic loci defining the focal regions of interest (BED file).

Output: list of focal-level CN calls, including the associated 95CI.

```
For each sample s in cohort do:
    pur ← Extract the purity value of the sample s
    cutoff ← Define a dynamic CN cutoff based on the sample purity: cutoff = 0.1/pur
    for each genomic locus gene in s do:
        geneSeg ← filter segments in s that overlaps with gene locus
        if number of geneSeg > 1 do:
            compute the % of overlap of each segment in geneSeg with gene
            sel_gene_Seg ← select the single segment that has the highest % overlap
            focalCN ← extract the CN value of sel_gene_Seg
            focalCI ← extract the 95CI value of sel_gene_Seg
            Save focalCN and focalCI in the output list
        else if number of geneSeg = 1 do:
            focalCN ← extract the CN value of geneSeg
            focalCI ← extract the 95CI value of geneSeg
            Save focalCN and focalCI in the output list
        end if
    end for
end for
```

By applying this script to all the SEG files of the samples, it was possible to generate focal-level CN calls with included 95% Confidence Interval. A total of 15 focal CN calls per sample were computed in this way, relatively to the GISTIC-defined 10 focal loci for deletions events and 5 focal loci for amplifications events (see results chapter "A GISTIC2 analysis to discover new genes targeted by focal CNA in MM" for more detail about the loci identification).

4.1.7 PHASE 3: Timing Analysis

The third phase of the pipeline for multi-platform harmonized Copy Number Alterations timing, consist in the actual generation of a timing league-model (or "cohort-timing") to estimate the "ancestrality" of the CNAs alterations (defined as a measure of how much a given alteration is likely to be a primary/founder event in the evolutive history of MM) as obtained in the phase 2 of the pipeline (both focal and broad CNAs calls).

To the knowledge of the researcher, two different approaches were already developed and used by other research groups in order to compute league-models in the field of cancer alteration timing: the PhylogicNDT league-model module (PLM) and the Bradley-Terry league model.

- The PLM approach is part of a more general suite of cancer alteration timing tools provided by the PhylogicNDT package. [36,80] In particular, the input required by PLM consists in the output of the previous module of the PhylogicNDT pipeline, the *SinglePatientTiming* module. *SinglePatientTiming*, in turn, requires as input the sequencing BAM files generated by either WGS or WES experiments, and it generates the order and relative timing of the patient's somatic events in a probabilistic manner by using copy number and mutation data together to infer the relative ordering of somatic events (Fig. 26).[36]
  In particular, the PLM module analyze the probabilistic data for single patients generated in this previous module in a very similar fashion to what happens in the Bradley-Terry algorithm (see methods section). In fact, as described in the "results" section of the PhylogicNDT paper, the PLM workflow consists in the following steps:
  1) First, PLM integrates the single patients' information into a pairwise event contingency table, this table represents the probability that a random patient (from the cohort) that harbors a specific pair of events, will have the first event in the pair earlier, later or at a similar / indetermined time as the other event.
  2) This dataset is then sampled in such a way that all individual events play a "sports season" against each other with "matches" played between pairs of events, where the outcome of the "match" is decided by sampling from the pairwise probabilities.
  3) Finally, the method then calculates the odds ratio of events occurring early or late during tumour development (figure 26).

  Unfortunately, no additional information about the specific functioning of PLM could be found by the researcher, due to the fact that, at time, the package documentation is still in development on the GitHub PhylogicNDT public repository.[80] Moreover, the PhylogicNDT paper is still only available on the BioRxiv repository (not published yet in a peer-reviewed journal) and publicly available PDF document is still missing the specific "online method" section that illustrates the more detailed functioning of the PLM module.[36]

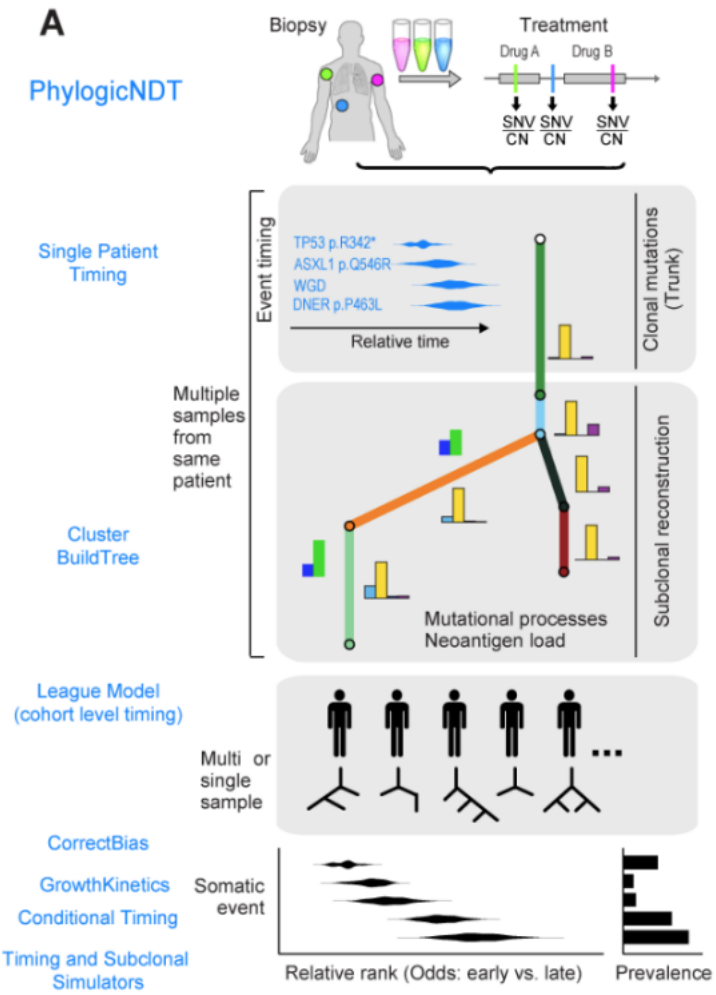


***Figure 26**: infographic of the complete workflow of the PylogicNDT package. The league-model module is computed based on the output of previous modules, which requires CNAs and mutations data, thus NGS generated data.*

- Instead, the Bradley-Terry approach consists in a more general and open workflow: in this approach the user can directly and freely define the "pairwise event contingency table" (also used by PhylogiNDT) with a method of choice. This table is used next as input to the actual Bradley-Terry algorithm, which assumes that individual events (or "players") in the table play a "sports season" against each other with "matches" played between pairs of events (also similarly to what happens in PylogicNDT). Finally, by applying a Maximum Likehood estimation (MLE) model, the Bradley-Terry algorithm computes the specific ability values of all the "players" that played those "matches". Importantly, in the context of cancer alteration timing the matches are played by the alteration events by using a "clonality contest" criterium, in which the more clonal/early events "wins" against the more sub-clonal/late events. Thus, in this scenario the computed ability of the players can be interpreted as "timing estimates".[33]

In addition, in the supplementary material of Gerstung M. et al., 2020 [38] a direct comparison between the performance of the PLM and Bradley-Terry approaches in the timing of cancer alterations was performed on the PCAWG datasets, which includes 27 different types of cancers (Figure 27). This analysis showed a very good correlation and concordance between the two methods, which could be expected due to both methods relying approximately on the same methodological framework and the same conceptual reasoning.

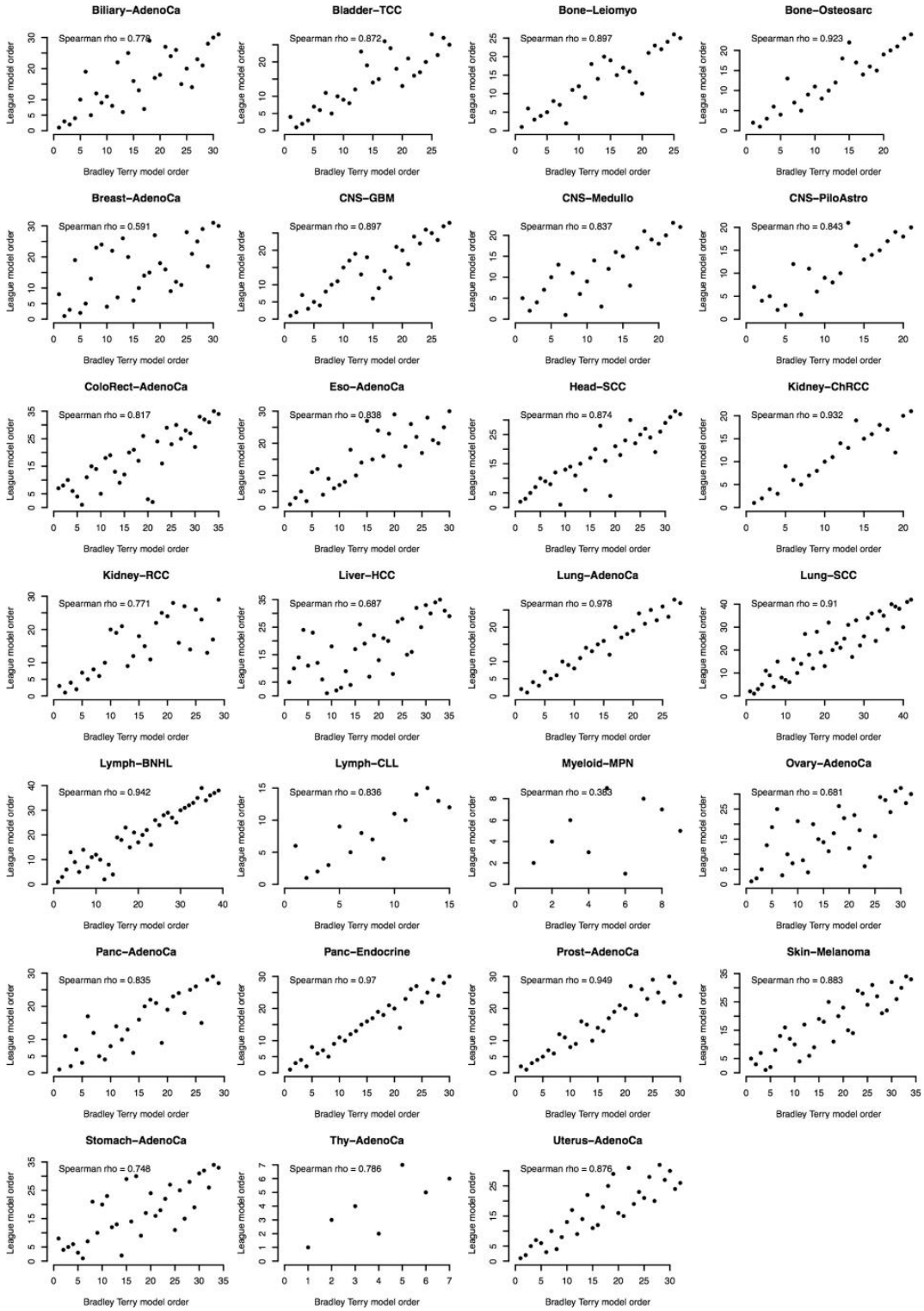


Figure 27: *Bradley-Terry timing analysis versus PylogicNDT League-Model (PLM) timing analysis. The analysis was performed on 27 different cancer types included in the PCAWG dataset. The correlation between the two methods timings are remarkably high, since the Spearman rho is shown to be > 0.70 in 24/27 tumors types, and > 0.80 in 20/27 tumor types.*

After carefully considering all those information, the researcher decided that the application of the Bradley-Terry approach for cohort timing analysis in this study was the most appropriate choice due to the following reasons:

1) The Bradley-Terry approach is the more versatile and controllable approach: this method allows the user to freely define how to build the pairwise event contingency table required for building the model. Additionally the user can fully control how the actual matches between the clonality levels of the alterations are performed (deciding for example how many points will be assigned to the "winner" of each match).
2) The PLM approach, despite being a method included in a solid and compact timing analysis framework (PylogicNDT), can only analyze samples profiled with NGS platforms (WGS or WES), thus it can't be applied to the whole cohort of samples in this study as it's not usable on SNP arrays samples. Additionally, it was not possible to investigate the specific functioning of this method, since the actual bioinformatic tool is still in development and the methods section of the paper that describes it is not yet publicly available.
3) It was demonstrated by Gerstung M. et al. (2017) that the analysis generated by the two methods are highly similar in the task of cohort-timing many different cancer types. Therefore, neither approach is clearly superior, or different, from the other.

4.1.7.1 TestClonality: a function to statistically compare and match the clonality between pairs of alterations

In order to generate the pairwise event contingency table required as input to the Bradley Terry model, it was necessary to process all the harmonized CNAs calls generated by the phase 2 in a specific way. This table is formatted with the objective to describe and summarize the aggregated "matches" results between all the "players" in every "tournament", where the matches are represented by clonality contests, the players are represented by genomic alterations and the tournaments are represented by the tumor samples (Figure 28). The researcher followed the instructions described in the vignette of the *BradleyTerry2* R package, with some modifications, in order to generate such a table.

Player1	Player2	Wins1	Wins2	Matches
Real Madrid	**Juventus**	14	16	30
Chelsea	**Juventus**	12	18	30
Real Madrid	**Chelsea**	21	9	30
…	…	…	…	…

Player1	Player2	Wins1	Wins2	Matches
Amp chr 1q	**Amp chr 4p**	145	21	166
Del TP53	**Amp chr 1q**	74	234	308
Amp chr 4p	**Del TP53**	21	92	113
…	…	…	…	…

Figure 28: *Examples of "pairwise event contingency tables" required as input for the Bradley-Terry model. On the left (in yellow) a table that can be used for computing the different abilities of soccer teams, based on how many times they have won against each other in a sport match. On the right (in green) a table that can be used for computing the different abilities (timing estimates) of genomic alterations, based on how many times they have won against each other in a clonality contest match.*

Since the complex nature of the clonality data, some modifications to the actual procedure for matches computation were need in this context. In particular, it can be noticed that the clonality value of a given alteration consists in both a point estimate, defined by the deviation from the baseline value of the CNA event (e.g. a CN = 2.8 consists in an amplification with clonality = 0.8 and a CN = 1.5 consists in a deletion with clonality = 0.5, with reference to a diploid baseline) and also the associated error estimation, that is the 95% Confidence Interval included in the CNA call.

In this scenario, when performing a clonality contest between a given pair of alterations, the determination of the winner is not a trivial task, since the error in the clonality measurement must be taken into account in the match. This problem, names "matching clonality problem" is visually illustrated in figure 29, and arises when trying to determine if a statistical difference exist between two measurements when only their 95% CI is available.

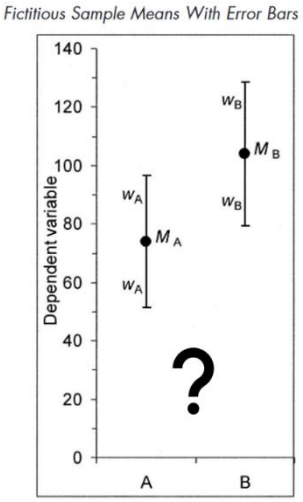


Figure 29*: The "matching clonality problem". A and B measurements (Ma and Mb) are associated with 95%CI error bars (Wa and Wb). In this case determining if Mb is actually different, and superior, to Mb is a statistical challenging task.*

To this aim the *TestClonality* R function was developed. This function can statistically compare and match the clonality between pairs of alterations. To this aim two different solutions are used:

1) The "eye inference" solution: this approach was inspired by the Cumming & Finch (2005) [81] method. This method provides a practical "rule of thumb" useful to assess if two measurements Ma and Mb, with associated 95%CI Wa and Wb are statistically different. Critically, This rule is valid only if Wa and Wb magnitude does not differ more than a factor of 2. The procedure, reported here, was demonstrated by the authors of the paper using simulations and empirical analysis.

    a. First, the *overlap* measure between the two 95%CI is computed as follows in Eq. 3:

$$Overlap = \begin{cases} (Mb + Wb) - (Ma - Wa); & if\ Ma > Mb \\ (Ma + Wa) - (Mb - Wb); & if\ Mb > Ma \end{cases} \qquad Eq.\ 3$$

    b. Second, the *ProportionOverlap* statistic is computed as follows in Eq. 4 (Fig. 30A):

$$ProportionOverlap = \frac{Overlap}{mean(Wa, Wb)} \qquad Eq.\ 4$$

    c. Finally, if the computed *ProportionOverlap* value is < 0.5, also the p-value of the comparison is typically < 0.05, so the difference between the two measurements is significant under the canonical p-value threshold (Fig. 30B).

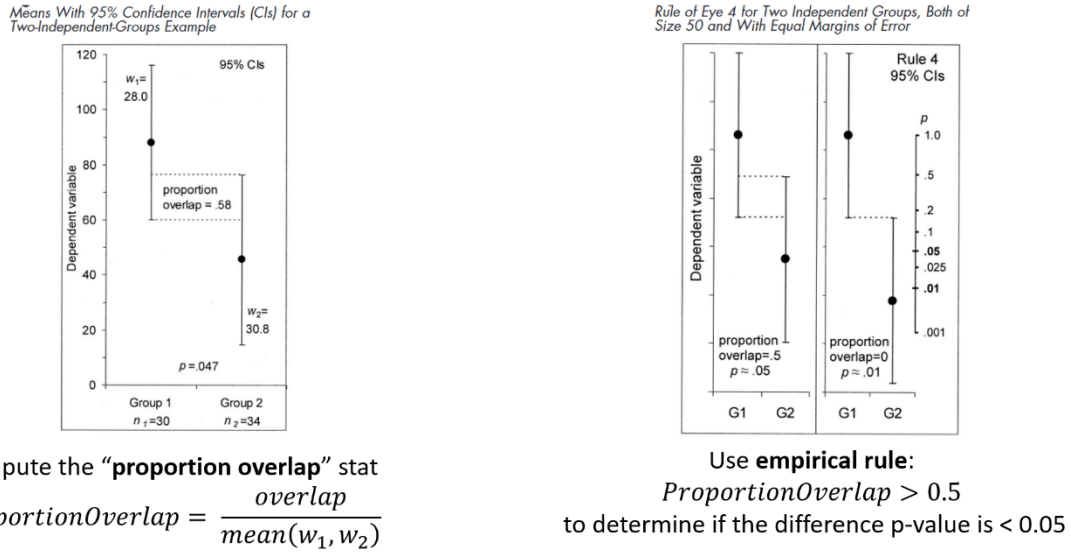


Figure 30: *graphical illustration of the "eye inference" rule-of-thumb approach developed by Cumming & Finch (2005), implemented in the TestClonality R function.*

2) The "Summarized t-test" solution. The researcher observed that a standard statistical 2-sample t-test can be both computed from raw observations, but can also be computed from the summarized statistics of those raw observation, that are the Mean / point estimate μ, and the Standard Deviation σ. [82] This possibility is implemented in R by the *tsum.test* function included the package BSDA (Basic Statistics and Data Analysis).[83] The only limitation of this solution is that Standard Deviation information is missing, but if assuming normality (when the sample size *N* is > 100) it can be estimated indirectly by using the following equation (Eq. 5):

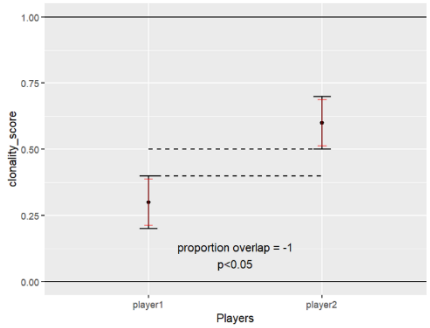
$$\sigma = \sqrt{N} \times \left(\frac{upperCI - lowerCI}{3.92}\right) \qquad Eq.\ 5$$

Where *N* is the sample size. In fact, if *N* is large (bigger than 100 for both events), the normality can be assumed, and the 95% confidence interval is 3.92 standard errors wide ($2 \times \sigma = 3.92$).

In this case the sample size is represented by the number of sequencing reads for mutation data and the number of bins/probes for CN segments data (which is a information present in the SEG file).

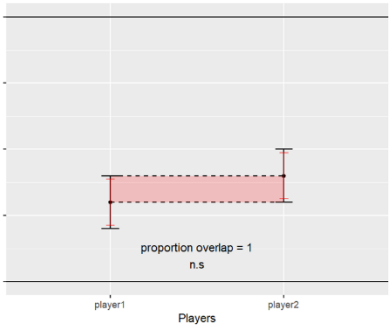
By means of *TestClonality* function, all the clonality matches between alterations could be computed in a formal statistical fashion (Figure 31). In particular, among the two possible *TestClonality* modes the "summarized t-test solution" was preferred, but in the cases when it was not possible to meet the assumption of the underlying t-test (normality, or sample size > 100) the "eye inference" solution was used instead to perform the match. All the matches results computed in this way were used to generate the pairwise event contingency table required as input for the final computation of the Bradley-Terry model.

	CCF	95% CI	N reads
Player 1	0.3	0.1	100
Player 2	0.6	0.1	100



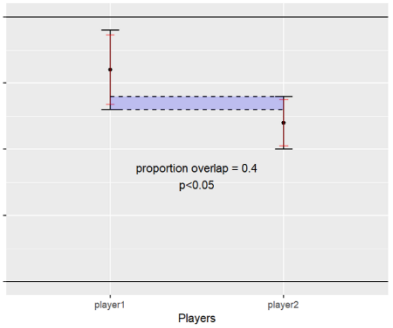
t-test p-value: <0.001

	CCF	95% CI	N reads
Player 1	0.3	0.1	90
Player 2	0.4	0.1	90



t-test p-value: 0.1477

	CCF	95% CI	N reads
Player 1	0.8	0.15	60
Player 2	0.6	0.10	80



t-test p-value: 0.031

***Figure 31**: example of three matches performed by pairs of alterations ("players") using the TestClonality R function. In the first match the t-test assumption is met (N reads / sample size > 100 for both events) so the match was performed using the "summarized t-test" mode. In the second and third matches the "Eye inference" mode was used instead, since the t-test assumption are not met: the second match is not significant while the third match is significant, since the proportion overlap parameter is < 0.5.*

Importantly, the score assigned to the winner of each statistically significant match was not = 1 as in standard Bradley-Terry models, but instead the score was proportional to the difference between the clonality point estimates of the paired alterations. This specific score-assignment strategy was chosen to maximally exploit the timing information from the inferred by the clonality difference between alterations. At the same time, the non-statistically significant matches were considered as ties, regardless of their difference in clonality point estimates.

4.1.7.2 Bradley-Terry models generation and calculation of timing estimates

Finally, after computing the pairwise event contingency tables, the actual Bradley-Terry models could be generated from those tables. The model were computed by using the *BTm* function from the *BradleyTerry2* R package, as detailed in Methods. Quasi-variances and relative Quasi-standard errors (qSE) associated to the computed abilities were extracted by using the *qvcal* function. Critically, in addition to the table, the model also requires the prior definition of a specific player / alteration event to which is assigned an ability equal to zero by default: in this way all the other players / events abilities are computed relatively to the ability of the predefined player. Since this choice is critical for a posterior correct interpretation of the abilities / timing estimates generated by the model, the researcher found appropriate to set the predefined "0 ability" player to the HyperDiploidy event since this event is well-known to represent a real ancestral event in the evolutive history of MM, as demonstrated by multiple previous MM timing studies (see introduction chapter "Genomic timing in MM: state of the art").

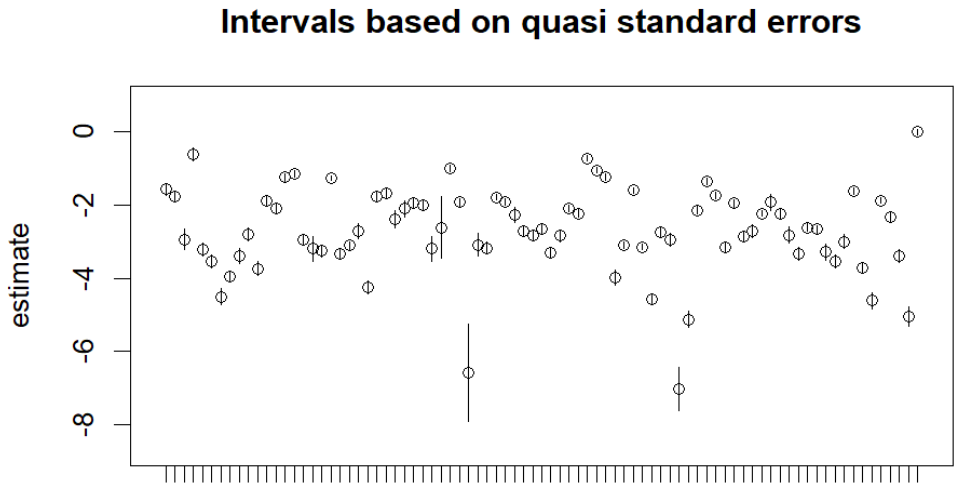


***Figure 32**: example of raw output generated by the Bradley-Terry model. In this bar-plot the ability estimates of the players are shown on the Y axis and the different players are show on the X axis. Bars represent the quasi-standard errors associated with each ability estimate. The abilities are measured relatively to a predefined player, to which is assigned an ability = 0 by default (last point in the plot).*

In this way, it was possible to obtain the raw estimates (and qSE) of the "players" (CNAs alterations) included in the model. An example of the raw output of the BTm and qvcal funcions can be visualized in Figure 32.

It's also important to note that in the case a given alteration didn't play any matches (due to the fact that it was never detected in the dataset) this alteration was excluded by the model output. This was the case for extremely rare alterations (such as the deletions of the typically amplificated chromosomes) in the small cohorts of this study with a relatively small number of available samples (the SU2C and BUS cohorts). Since these alterations were not found to be "early" / ancestral events in the other big cohorts (BO and CoM) included in the study this issue does not represent a problem for the specific aims of this study.

Finally, the ability estimates generated by the model were subject to the "opposite" mathematical transformation: *-f(x)* (i.e. negative values were converted to positive values, and positive values were converted to negative values), thus generating the definitive "Timing Estimates" (TE) parameter that was used as a the main timing metric in all this study. This choice was made in order to facilitate the later interpretation and visualization of the TE, in fact by applying this procedure and considering the previously described definition of the Bradley-Terry abilities, the TE can be interpreted as follows:

Timing Estimate (TE) = *number of units of "molecular time" after the HyperDiploidy reference event, after which a given alteration event occurs.*

4.2 A GISTIC2 ANALYSIS TO DISCOVER NEW GENES TARGETED BY FOCAL CNA IN MM

As described in Methods section, the GISTIC v2 bioinformatic tool was employed in order to define a confident list of focal loci of interest that are statistically targeted by "driver" CNAs, in the MM genomic CNAs landscape. Since the MM patient's genome is extensively affected by CNAs events and given that their CN profile often present a complex karyotype, this type of task is not trivial, due to the high number of "passenger" CNAs that introduce an elevate level of background noise that confounds the correct identification of focal driver CNAs. Despite the availability of a high number of publicly available MM CN profiles (CoMMpass dataset), and despite the popularity of the GISTIC tool, to date there is no published study that specifically applied the GISTIC algorithm for the identification of focal driver CNAs in MM.

Based on the specific GISTIC algorithm functioning (described in the methods section) this is probably due to the following reasons:

A. GISTIC requires a high number of samples in order to significantly identify rare and small focal CNAs: in the case the sample cohort is too small (let's say less than 100-200 samples) it becomes difficult for the algorithm to statistically distinguish true focal alterations from the background noise consisting of random "passenger" CNAs. This happens because the signal of the "drivers" CNAs is too diluted in the "background" noise. Having a big cohort of tumor samples can logically resolves this issue (at least >400-500 samples, depending on the level of genomic instability of the cancer) since with the increase of the observations the signal becomes stronger while the noise becomes weaker. In MM, such big cohorts of CN profiled samples become publicly available only recently from the CoMMpass study.
B. GISTIC requires a high CN profiling platform resolution to correctly identify the small size CN events: if the platform that analyzed the CN profiles of the samples shows a low resolution (let's say 1 or 2 Mb, as in the case of ULP-WGS) focal events with a size smaller than the platform resolution can't be materially detected. In MM a good quality GISTIC analysis of the CoMMpass CN profiles is dataset is challenging due to the low resolution (low-coverage) of the WGS samples of the study. In addition, a non-optimal setting of the segmentation algorithms parameters can also introduce uncertainty in the analysis resolution since it influences the number of CN breakpoints (e.g. the *alpha* parameter of the CBS algorithm, which influence the segmentation p-value). In particular, a too "stringent" segmentation procedure even if it can produce nice and clean CN profiles, can also introduce false-negatives in the called CNAs segments, while on the contrary, a too "aggressive" segmentation procedure can introduce many false-positives in the CNAs segments called.
C. An effective and formal GISTIC analysis requires at least two different datasets of CN profiles of big and equal size, profiled with different genomic platforms: one dataset is used for the main analysis while the second one is used as a validation dataset. This is in order to exclude possible platform-specific bias which could generate false focal CNAs in

the main dataset. In MM, apart from the CoMMpass dataset, no other sufficiently big public datasets are available to this aim.

Starting from this reasoning, the researched proposed that only the BO dataset collected in this study can function as the main dataset needed for a formal GISTIC analysis, since it satisfy both the A) and B) requirements, while the CoMM cohort only satisfy the A) requirement. In this case, the CoMM dataset can instead function as the validation cohort needed for the C) requirement, since even if the low-resolution of the dataset does not allow a "de-novo" confident identification of small CNAs, it could at least enable the validation of their identification from another high-resolution dataset (BO dataset).

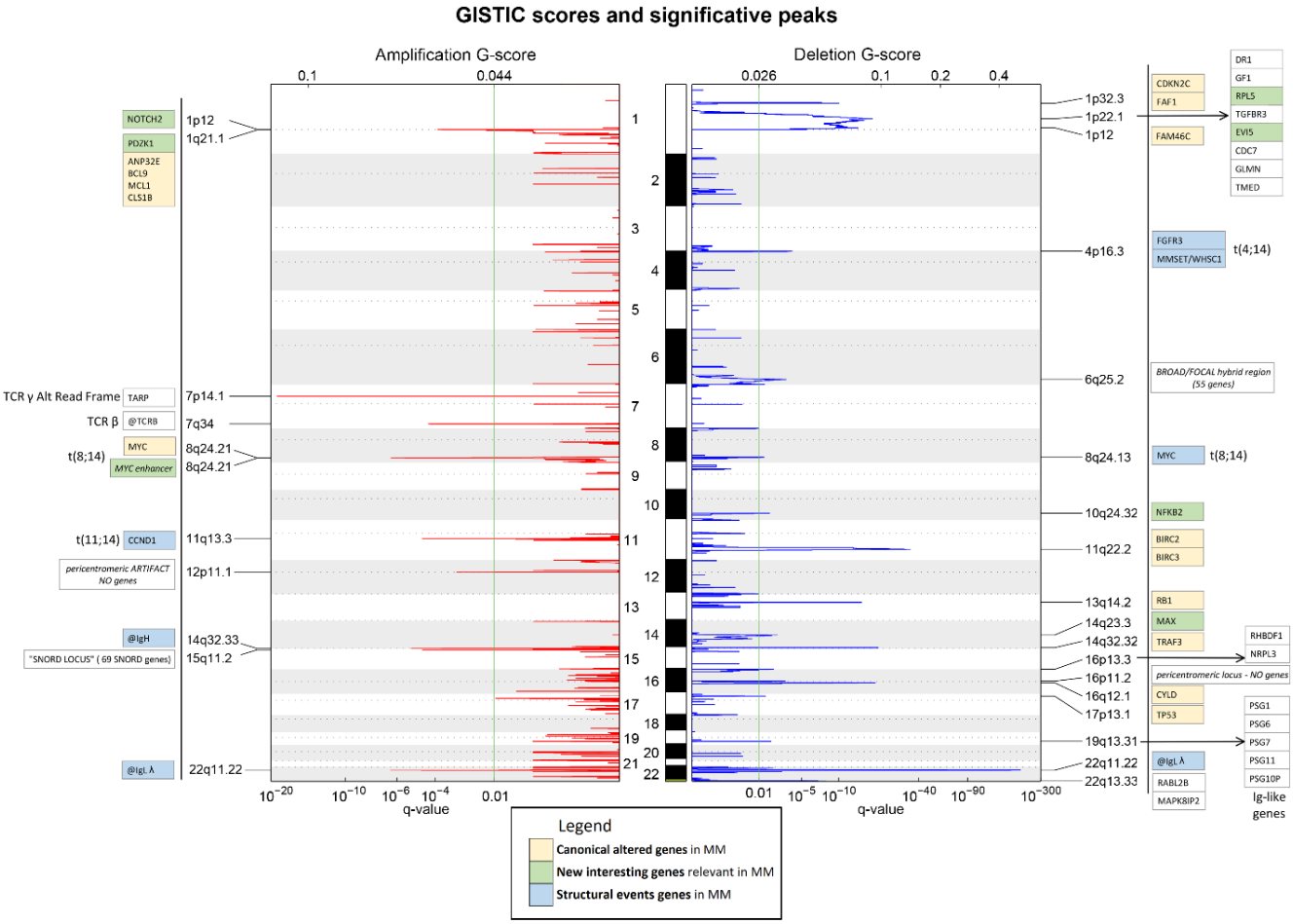


Figure 33: *GISTIC analysis results on the main BO dataset. On the left (red track) and on the right (blue track) are shown the significant deletion and amplifications "peaks", respectively. The peaks are identified by a significative value of the G-score (cutoff q-value < 0.01, as highlighted by the green vertical line). The MM cancer-associated genes included within each peak are annotated near the peaks*

Following this reasoning, the researcher performed two different GISTIC analysis (more details in the Methods section) on the BO (main / high-resolution) dataset and the CoMM (validation / low-resolution) dataset. Results obtained by the analysis of the main BO dataset are presented in Figure 33 and Table 7.

This analysis generated 29 GISTIC-defined peaks in the BO dataset (11 amplification peaks and 18 deletion peaks), which were all subsequently investigated and annotated with the genes included within each of the peaks' boundaries (Fig. 33). Next, after a careful review, the researched was able to pinpoint 15 specific focal CNAs event (5 amplifications and 10 deletions) among all the discovered peaks, which were proposed as "driver" focal CNAs in the MM CNAs landscape (Table 7). The selection of those events was carried out according to the following criteria: a) the peak must include a gene already reported as "driver" in MM, b) the peak must include a new onco-gene or tumor suppressor gene which is involved in a commonly reported deregulated pathway in MM. By using this procedure a total of 5 events were found to be hypothetical novel "focal" CNAs drivers in MM (*NOTCH2*, *MYC ME2-enhancer*, *EVI5*, *MAX*, *NFKB2*). The detection of all these events in the BO dataset were validated by performing another GISTIC analysis in the validation CoMM dataset and ensuring that the peaks identified in BO overlapped with peaks identified in the CoMM dataset analysis (Fig. 35). Of note, the only event that could not be validated in this way was the *MYC ME2-enhancer* amplification, this was probably caused by the high complexity of the MYC region, which is affected by multiple types of translocations in MM. The abundance of structural events in the cohort consequently generates a lot of "noise" in the region, which could in turn confound the GISTIC algorithm, especially when considering the low-resolution of the CoMMpass dataset. The validation of this region was thus performed by a literature review on MYC enhancers ("the MYC enhancer-ome"), which revealed that this specific region is usually amplified in Acute Myeloid Leukemia, another hematological malignancy (Fig. 34). [84]

Peak n.	Focal gene	CNA type	NEW	Chrom	Position hg19	Position hg38
1	**CCND1**	amplification	NO	11	69455855 - 69469242	69641156 - 69654474
2	**CKS1B**	amplification	NO	1	154947129 - 154951725	154974653 - 154979251
3	**MYC**	amplification	NO	8	128747680 - 128753674	127735434 - 127742951
4	**NOTCH2**	amplification	**YES**	1	120454176 - 120612240	119911553 - 120100779
5	**MYC ME2-enhancer**	amplification	**YES**	8	129240986 - 129276648	128228740 - 128264402
6	**BIRC2**	deletion	NO	11	102217942 - 102249401	102347211 - 102378670
7	**CDKN2C**	deletion	NO	1	51426417 - 51440305	50960745 - 50974634
8	**CYLD**	deletion	NO	16	50775961 - 50835846	50742050 - 50801935
9	**EVI5**	deletion	**YES**	1	92974253 - 93257961	92508696 - 92792404
10	**FAM46C**	deletion	NO	1	118148556 - 118170994	117606048 - 117628389
11	**MAX**	deletion	**YES**	14	65472892 - 65569413	65006174 - 65102695
12	**NFKB2**	deletion	**YES**	10	104153867 - 104162281	102394110 - 102402524
13	**RB1**	deletion	NO	13	48877887 - 49056122	48303744 - 48599436
14	**TP53**	deletion	NO	17	7565097 - 7590856	7661779 - 7687538
15	**TRAF3**	deletion	NO	14	103243813 - 103377837	102777449 - 102911500

Table 7: *the focal region "peaks" identified by the BO GISTIC analysis which were selected to be relevant events in MM. The selection identified the genes already reported to be driver in MM (column NEW = NO), or novel oncogenes / tumor suppressor genes involved in pathways reported to be driver in MM (column NEW = YES, in yellow).*

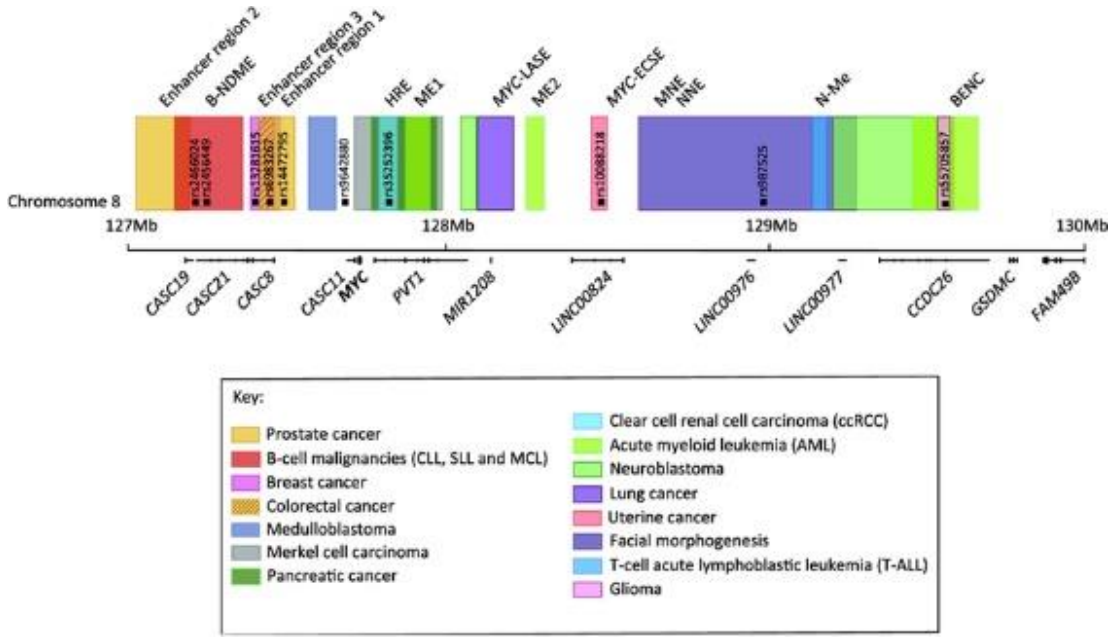


Figure 34: *The "MYC Enhancer-ome" as reported in Lancho O, et al. (2018).*[84]


Chromosome 1
249 mb
Refseq Genes
scores.gistic_CoMM
GISTIC.conf_95_CoMM
scores.gistic_BO
GISTIC.conf_95_BO
Any-DP3
Any-DP1
Any-DP2
Any-AP1
Any-AP2
Chromosome 8
146 mb
Any-AP11
Any-DP11
Any-DP12
Any-DP5
Any-AP5
Chromosome 10
135 mb
Any-DP14
Any-DP7
Chromosome 14
107 mb
Any-DP22
Any-DP21
Any-DP11

.

***Figure 35**: Validation of the novel selected focal "drivers" of MM. Four IGV (Integrative Genomic Viewer) screenshots show the overlapping GISTIC peaks between the BO dataset analysis and CoMM dataset analysis, BO peaks are represented by orange rectangles, CoMM peaks are represented by green rectangles. Black zoom windows are plotted under the peaks of interest to show the specific genes (circled in red). Chromosome 1 shows the validation of EVI5 deletion and NOTCH2 amplification. Chromosome 8 shows MYC and the MYC enhancer peaks. Chromosome 10 shows the validation of NFKB2 deletion, Chromosome 14 shows the validation of MAX deletion*

In conclusion, the peak boundaries of those 15 validated focal MM "drivers" were used to generate the BED file required for the Focal-level call step in the multi-platform harmonized CN analysis pipeline used in this study (see Results, "PHASE 2: CNAs calling" chapter).

4.3 TIMING ANALYSIS OF CNA EVENTS AT NDMM AND AT SMM PHASES

After the complete development and finalization of the multi-platform harmonized CN analysis and timing pipeline, the complete pipeline was executed on the SEG files derived from each of the four cohorts of samples included in the study, namely: 1) the BO cohort of 750 SNP arrays samples from NDMM patients, 2) the CoMM cohort of 832 low-coverage WGS samples from NDMM patients, 3) the BUS cohort of 171 WES samples from SMM patients and 4) the SU2C cohort of 114 WGS samples from SMM patients.

4.3.1 Timing maps of the single cohorts

A total of four full "runs" of the pipeline were carried out in this study. Each run resulted in a different Bradley-Terry Timing model, describing the order of acquisition of broad-level and focal level CNAs in the evolutive history of both NDMM and SMM disease phases. For each model an associated "time map" plot was created, that is a bar-plot-like visual representation of the Bradley-Terry timing model in which all the alteration events with corresponding Timing Estimates (TE) values can be visualized.

One main important result of this study is the fact of having achieved very narrow confidence intervals (quasi-standard errors) in the definition of TE for the various events (Fig. 36, 37, 38, 39), as compared to the confidence intervals previous timing analysis in MM (see Introduction, "Genomic timing analysis in MM: state of the art" chapter). This ensured an high quality and precision in the ranking of CNAs events 79and in the estimation of associated TEs. This observed high resolution of the analysis is due to having provided a lot of information to the Bradley-Terry models, in the form of summarized matches scores (pairwise events contingency tables). This generated a high statistical power for the model computation.

In fact, the matches scores information used by the Bradley-Terry models in this study was particularly abundant thanks to two main features of the study:

1) the high number of samples included in the cohorts (BO and CoMM in particular) which, in turn, logically generated a high number of matches between alterations (important especially for rare CNAs events).

2) The score-assignment strategy (see Results, "TestClonality" chapter) which enabled a maximization of the score (timing information) extracted from each clonality match.

Thanks to this elevate statistical power, the generated time-maps were able to confidently "resolve" the correct timing order of most of the analyzed CNAs, as can be observed by the not-overlapping confidence intervals in Figure 36, 37, 38, 39.

4.3.1.1 Bologna (BO) cohort NDMM timing map

In the BO cohort the top early occurring events (as compared to Hyperdiploidy) were: amp 11q(CCND1), del 13q(RB1), amp 1q(CKS1B), del 1p(EVI5), and del 11q(BIRC2), as shown by the BO time-map in Figure 36. The amp 11q(CCND1) is an event that deregulates the CCND1, a common "driver" of MM that deregulate the cell cycle pathway thus enhancing the tumor plasma cell proliferation. Its deregulation is renown to be an early event in MM since the CCND1 is also the target of the t(11;14) translocation, another well-known early event in MM pathogenesis. Next, del 13q(RB1), amp 1q(CKS1B) are two very frequent alterations that were often, but not always, considered as early alterations in the MM evolutive history. Their role as "drivers" or primary lesions is thus controversial [25,85], but recent studies further validated their role as drivers.[15] They both also deregulate the cell cycle pathway, an early and unifying early event in MM.[86] Del 1p(EVI5) is one of the three deletion peaks identified on chr 1p, EVI5 has been described as being involved in both cell cycle and cell migration regulation: in particular it has a role in the completion of cytokinesis and the safeguarding of genomic integrity during cell division; thus, deletions of EVI5 can results in cell-cycle deregulations. [87] Finally, del 11q(BIRC2) is a well-known CN event, both mutations and deletions of this gene are frequently reported and contribute to carcinogenesis through activation of the noncanonical NF-κB signaling pathway. The NF-kB pathway is reported to play seminal role in the pathogenesis of MM, and its deregulation is widely considered an early-occurring event. [88]

Other interesting observations are the early TE associated with amp of chr 18, which is an event very frequently observed in association with Hyperdiploidy, even though its frequency is not as high as the frequency of the odd-numbered chromosomes typically used to call the hyperdiploidy event in MM.[89] Given the early observed TE of this alteration, in line with the Hyperdiploidy TE, it's possible to speculate that this event could represent an additional amplification that could be included in the Hyperdiploidy definition in MM, but due to its relatively low frequency it was not included to the standard guidelines. Another interesting observation is the different TEs observed for the three focal deletions localized on chromosome 1p. Even if they show a similar alteration frequency (around ~20%), the different observed TEs reveal that they may happen in different time points during the MM evolutive history. In particular the EVI5 deletion is the most ancestral one, followed by TENT5C/FAM46C and CDKN2C. This is particularly interesting since it's possible to imagine that the most relevant "driver" alteration on chr 1p is the EVI5 deletion, and not the commonly studied CDKN2C deletion (used as a proxy for del 1p by FISH studies).

Of note, in figure 36 is possible to notice that the correlation between the TEs and the frequencies of the various alterations is high (R=0.83, p<0.001, data not shown). But while this is true, some outliers exist, such as del 11q(BIRC2), amp 11q(CCND1), amp 18p and 18q. This is particularly important since the frequency of alterations were historically used as a proxy to define their timing of occurrence in the MM evolutive history, following the simple rationale that a high frequent alteration correspond to a early alteration.[25] The timing data obtained in this study show that, while this rationale is generally true, some exceptions emerge when analyzing the timing with a more elaborated method. These exceptions can be of particular interest due to the fact that may represent newly discovered early event in the evolutive history of MM.

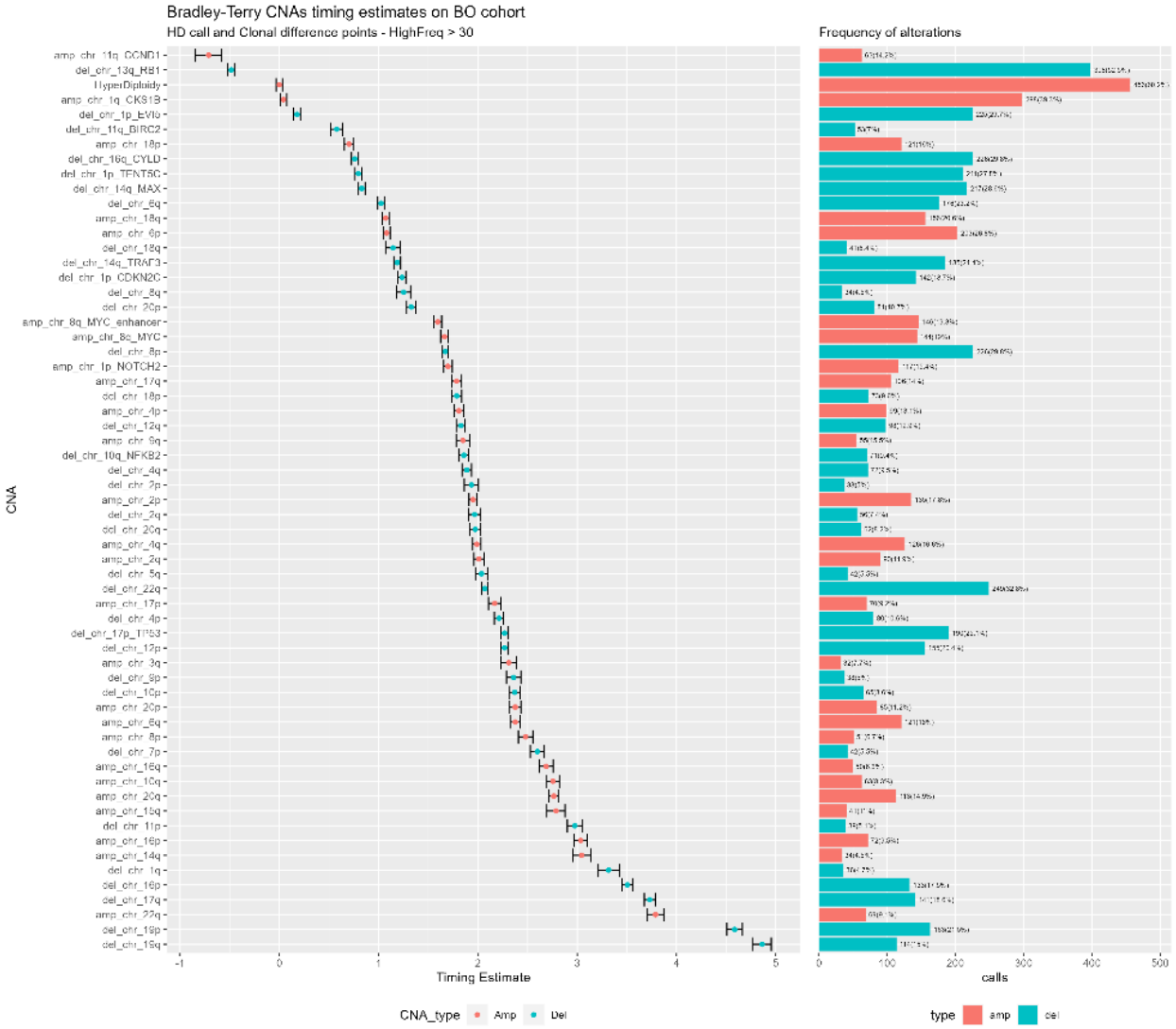


Figure 36: *timing map of the broad and focal CNAs in the NDMM BO cohort. On the left, the Timing Estimates (TE) of both amplifications (in red) and deletions (in blue) are showed with their associated quasi-standard errors (black bars). On the right the frequencies of the various CNAs are shown.*

4.3.1.2 CoMMpass (CoMM) cohort NDMM timing map

In the CoMM cohort the top early occurring events (as compared to Hyperdiploidy) were: amp 11q(CCND1), del 13q(RB1), amp 6p, del 14q(MAX), amp chr 18, del 14q(TRAF3), amp 1q(CKS1B) and del 1p(EVI5), as shown by the CoMM time-map in Figure 37. While most of these events are in common with the early events detected in the BO cohort, thus validating their ancestrality status, some additional events can be identified in this dataset. In fact, differences in TEs can be due the intrinsic diversity of the patients included in the cohorts, which can influence the TEs, especially in the rarer events. In particular the amp 6p was found to be an extremely early occurring event in the CoMM cohort. While, in the BO cohort, even if it was not found to be a top early occurring event, it was ranked among the earliest occurring alterations (13^{th} rank order). This event is interesting since chromosome 6p includes the CCND3 gene, a paralog gene of the Cyclin D gene together with CCND1 and CCND2. This gene is widely known for being targeted by early-occurring translocation t(6;14) in MM, which causes the overexpression of this gene. [25] On the basis of the obtained timing data it's possible to speculate that the early amplification of chr 6p might represent an alternative mechanism that cause the early high expression of the CCND3 gene in MM.

Next, two event localized on chromosome 14q were found to be very ancestral events in the CoMM cohort: del 14q(MAX) and del 14q(TRAF3). Both of them were ranked quite early in the BO cohort as well (10^{th} and 15^{th} rank order, respectively). The del 14q(MAX) event involves MAX (MYC associated factor X), a proposed tumor suppressor "driver" gene in MM, as proposed by recent data.[90] However the recent studies were based only on the inactivating mutations affecting this gene, thus the del 14q(MAX) may represent a novel mechanism by which this important tumor-suppressor gene is inactivated in the early phase of the disease.

Finally, also in the CoMM cohort it's possible to observe that some outliers can be identified when correlating the TEs with the frequencies of events: both amp of chromosome 2(p and q) and amp of chromosome 10 (p and q) were found to be early occurring but rare events. As illustrated in the BO cohort for chromosome 18, they may represent additional rare amplifications associated to the early occurring Hyperdiploidy event, but not detected frequently enough to be included in the guidelines for Hyperdiploidy definition.

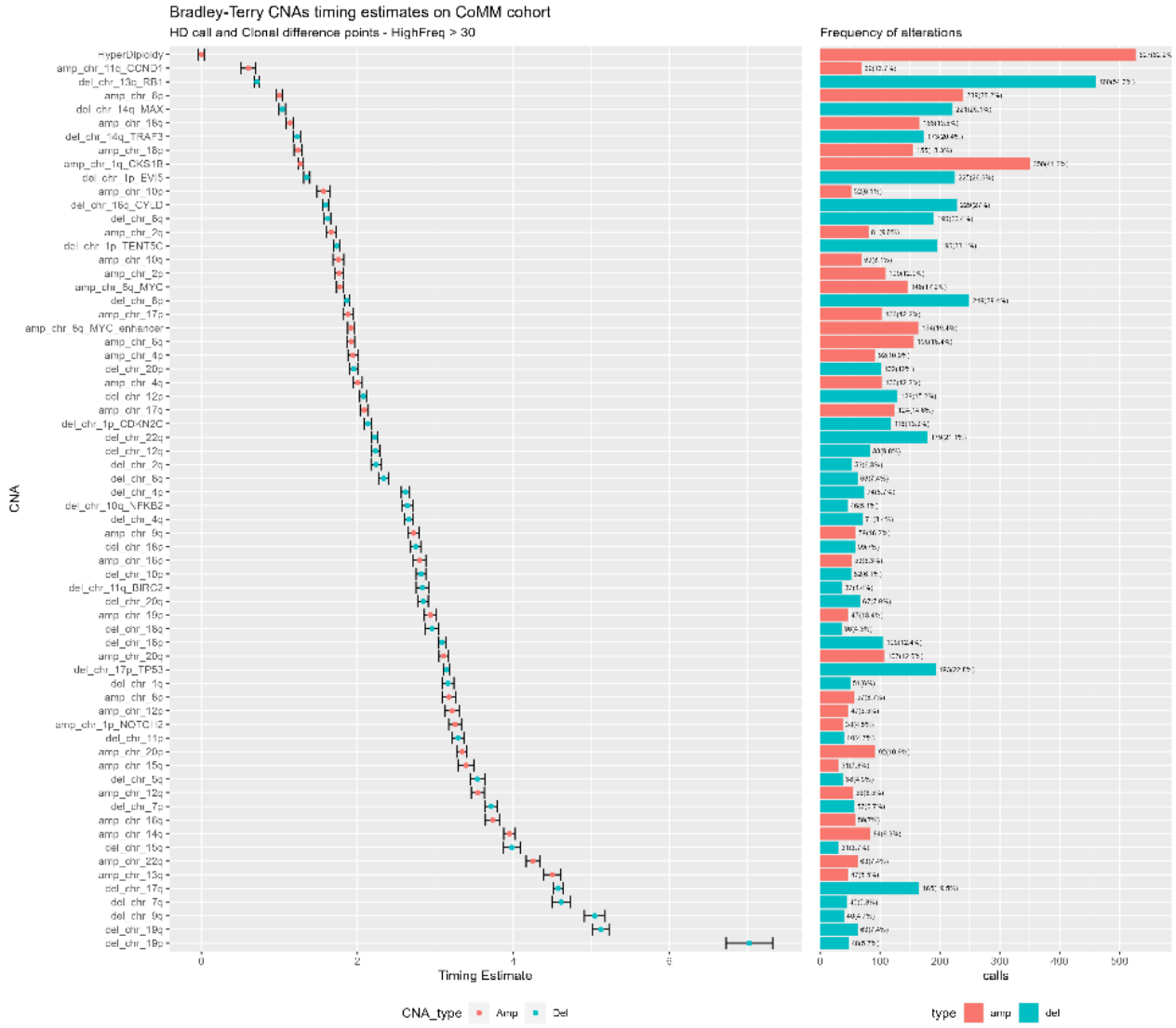


***Figure 37**: timing map of the broad and focal CNAs in the NDMM CoMM cohort. On the left, the Timing Estimates (TE) of both amplifications (in red) and deletions (in blue) are showed with their associated quasi-standard errors (black bars). On the right the frequencies of the various CNAs are shown.*

4.3.1.3 BUS cohort SMM timing map

In the single SMM cohorts (BUS and SU2C) a precise selection of specific top early events is challenging, due to the limited sample sizes of the cohorts (BUS = 171 and SU2C = 114) that could introduce a sampling bias effect when analyzed singularly. Despite this complication, a list of top early occurring events (as compared to Hyperdiploidy) detected in the BUS cohort could be identified, they were: amp 1q(CKS1B), del 13q(RB1), del chr 19, del 14q(MAX), del 11q(BIRC2), amp chr 18 and del 14q(TRAF3), as shown by the BUS time-map in Figure 38. As showed by the data, it was evident that most of the early occurring alterations in the NDMM disease phase can be also defined as early event even when analyzed at the SMM disease phase. This result further supports the notion that SMM is predominantly a genomically "mature" disease at time of diagnosis, since most of the "driver" genomic lesions already happened in the tumor genome at time of SMM diagnosis. [14,58,59]

However, one important difference could be observed when comparing the top early event of this cohort with the ones observed in the NDMM cohorts: that is the absence of chromosome 1p deletions events (EVI5, TENTC5, CDKN2C) in the top early events in SMM. In fact, all three events in this cohort were ranked way later then in the two NDMM cohorts (Figure 46), suggesting an important difference between the timing in the two phases. One possible intriguing explanation is that del 1p could represent an early event in NDMM but, on the contrary, a late event in SMM. This suggests that this alteration could frequently happen in the SMM/MM interface, thus playing a critical role for the progression to active MM.

Another interesting observation is the presence of del 19 in the top early events of this cohort. However, since no currently known driver genes are located on this chromosome, this fact is possibly due to a stochastic sampling bias that may have caused an overrepresentation of clonal alterations of this rare event (only found in N = 15, 8.6% samples) in this relatively small cohort of samples (tot samples = 171). Another possible explanation is that the deletion of this particular chromosome might represent a founder event for the development of the SMM disease clone, but subsequently this clone tends to be lost at time of progression to active MM (since in the NDMM it is found to be a late occurring event). This explanation would be possible if this chromosome deletion implies an indolent disease phenotype clone, and, as a result of clonal competition phenomenon which is known to happen at the SMM/MM interface, this clone gets extinct at time of disease progression.

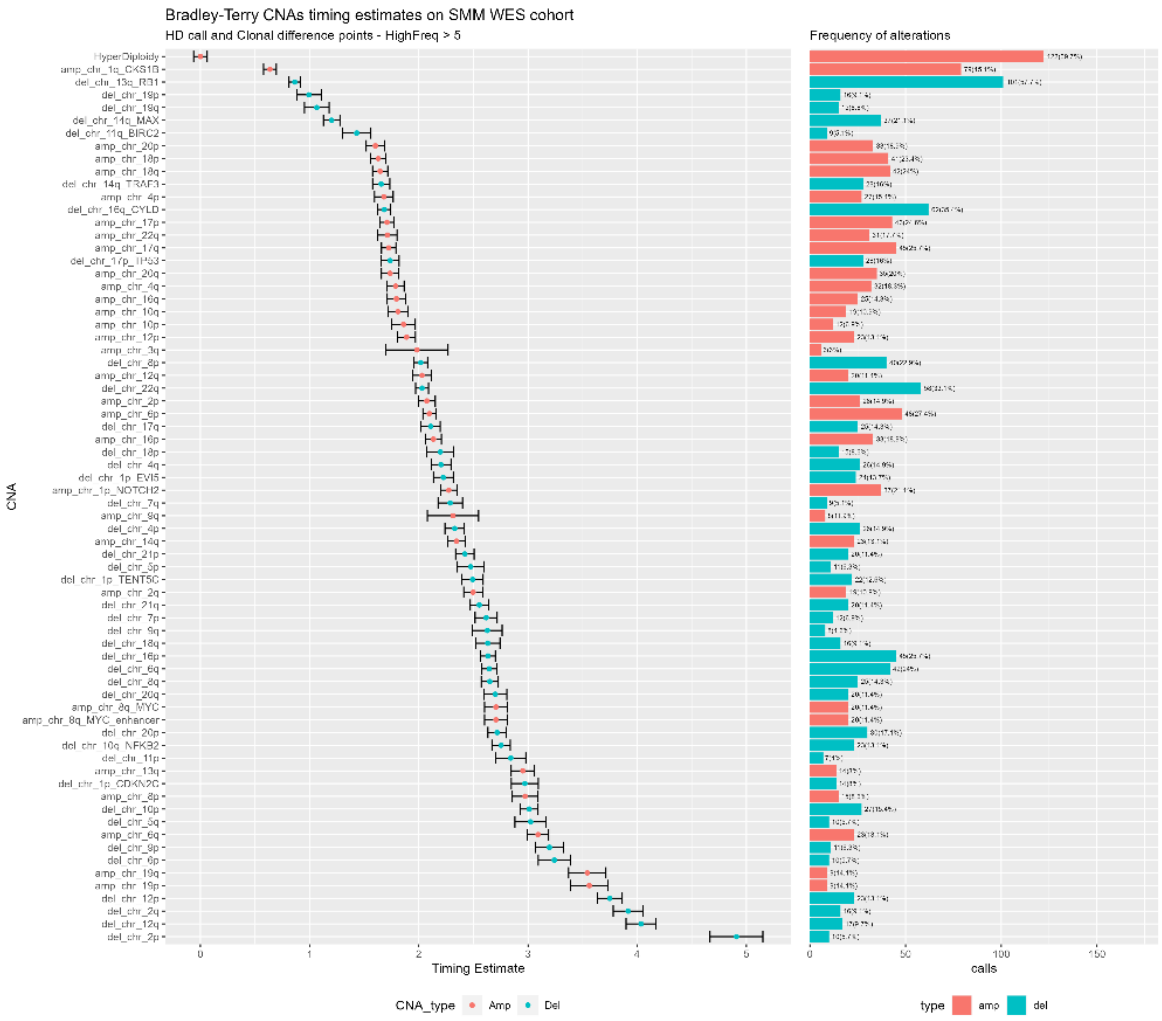


Figure 38*: timing map of the broad and focal CNAs in the SMM BUS cohort. On the left, the Timing Estimates (TE) of both amplifications (in red) and deletions (in blue) are showed with their associated quasi-standard errors (black bars). On the right the frequencies of the various CNAs are shown.*

4.3.1.4 SU2C cohort SMM timing map

Finally, in the smallest cohort out of the four, a precise selection of specific top early events is even more challenging. Nevertheless, the top early occurring events (as compared to Hyperdiploidy) in this cohort were: del 14q(MAX), del 13q(RB1), del 14q(TRAF3), del 1p(EVI5), amp 11q(CCND1) and amp 6p (Fig. 39).

Again, all the early events found in this cohort were found as early events also in the NDMM cohorts, validating their "driver" status once again. Here the main difference consists in the absence of amp 1q(CKS1B) among the top early events, the intermediate timing of the three deletions of chromosome 1p, and the late deletion of 11q(BIRC2). However, given the relatively small sample size as compared to the other cohort, here is not possible to exclude a sampling bias

that introduced either an overrepresentation or underrepresentation of clonal alterations affecting those chromosomes, affecting in turn the rank order of this SMM cohort evolution history.

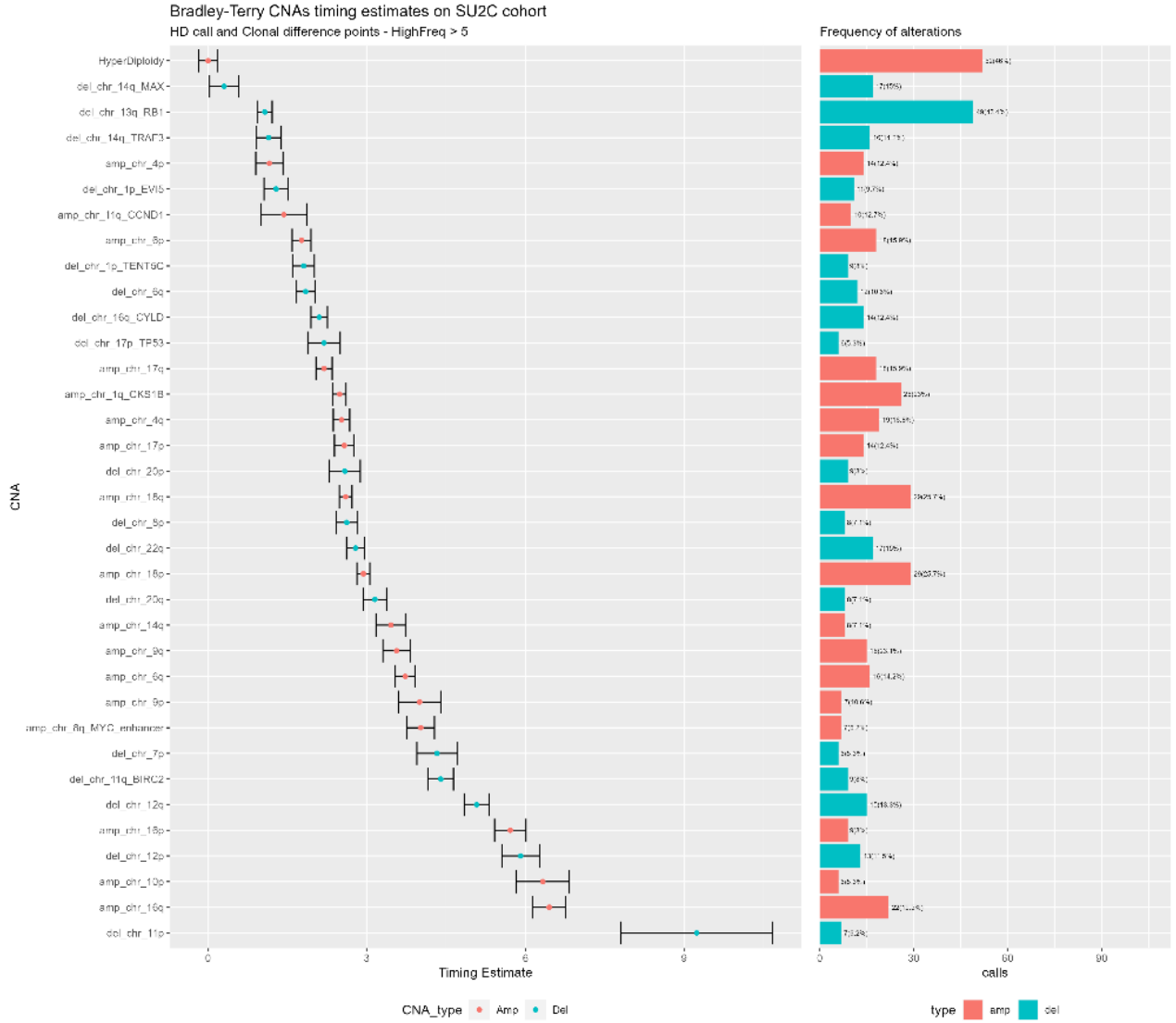


Figure 39: *timing map of the broad and focal CNAs in the SMM SU2C cohort. On the left, the Timing Estimates (TE) of both amplifications (in red) and deletions (in blue) are showed with their associated quasi-standard errors (black bars). On the right the frequencies of the various CNAs are shown.*

4.3.2 Timing Maps of the aggregated cohorts

Next, after the creation of timing models and maps for every specific cohort included in the study, the researcher sought to aggregate the samples belonging to the two different disease phases (SMM and NDMM) together, thus generating two new aggregated-cohorts: the "NDMM-cohort" consisting of BO and CoMM samples, and the "SMM-cohort" consisting of BUS and SU2C samples. These aggregated/cohorts were able to provide even more statistical power for aim of a confident detection of true "driver"/early event. This is particularly important in the case of the SMM disease phase, due to the fact that the single BUS and SU2C SMM cohorts alone showed a low sample size and statistical power, not sufficient for the identification of real early events with the same confidence obtainable in the NDMM cohorts, as illustrated before.

In order to do so, a direct comparison between the TE obtained from the cohorts to merge together was required before merging. This comparison was performed as a quality control step, in order to check the concordance between the TEs obtained from the different cohorts (Figure 40, 42). In fact, a low concordance between the TE estimates would indicate the presence of cohort-specific methodological bias during the process of TEs computation, which is what the researcher would have wanted to exclude by applying the developed CN harmonization pipeline. In addition, possible individual discordant alterations in these TEs comparison could indicate cohort-specific sample selection biases.

4.3.2.1 NDMM-cohort (aggregation of BO and CoMM cohorts)

A correlation analysis between the TEs of the alterations of the BO and CoMM cohorts was performed before merging the two cohorts together, as a quality control check (Figure 40). The correlation revealed a significative concordance between TEs, with a good level of concordance ($R=0.59$, $p<0.0001$). This ensures that no big methodological bias was present in the TEs computation process, validating empirically in this way the data harmonizing capabilities of such pipeline. The only outlier of the comparison, as defined by a residual analysis of the fitted linear model (data not shown) consists in the del 19p event, which even if it is found to be a late event both in BO and CoMM, is significantly more late in the CoMM cohort ($p<0.05$).

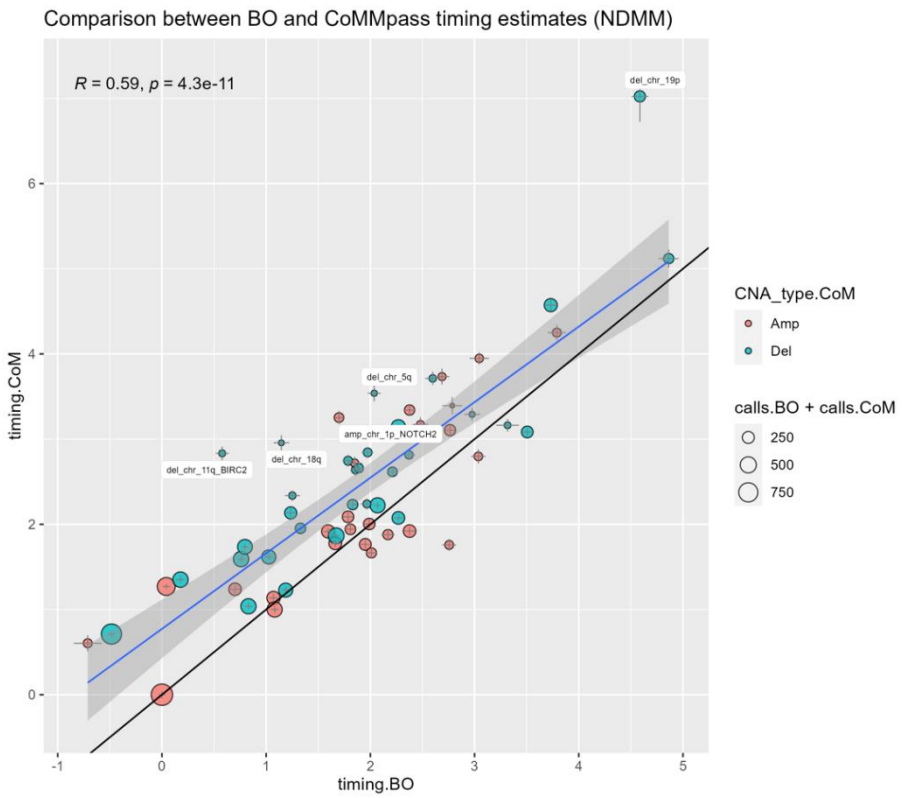


Figure 40: *Correlation analysis between BO and CoMM timing estimates (TE). Deletion events are represented by blue points, amplification events are represented by red points. The size of the points represents the number of total events observed. Grey lines around the points represent the quasi-standard errors associated to the TEs. The black line indicates the perfect correlation. A linear model was fitted to study the correlation (blue line).*

Next, after merging the two cohorts together a new Bradley-Terry model was generated, and a new time map for the complete NDMM cohort was created, as shown in Figure 41. In the NDMM cohort the top early occurring events (as compared to Hyperdiploidy) were: amp 11q(CCND1), del 13q(RB1), amp 1q(CKS1B), del 1p(EVI5), del 14q(MAX), amp chr 18 and amp 6p. This aggregated cohort result further confirms that those events, previously detected and characterized as early events in the single cohort timing result chapter, are indeed early alterations by means of the TE classification.

***Figure 41**: timing map of the broad and focal CNAs in the aggregated NDMM cohort. On the left, the Timing Estimates (TE) of both amplifications (in red) and deletions (in blue) are showed with their associated quasi-standard errors (black bars). On the right the frequencies of the various CNAs is shown, colored by original cohort.*

4.3.2.2 SMM-cohort (aggregation of BUS and SU2C cohorts)

A correlation analysis between the TEs of the alterations of the BUS and SU2C cohorts was performed before merging the two cohorts together, as a quality control check (Figure 42). The correlation revealed a significative concordance between TEs, with a medium level of concordance (R=0.39, p=0.0012). The lower concordance level in comparison to the NDMM analysis is expected, due to the lower number of samples of the SMM cohorts, which logically introduces more noise in the SMM TEs correlation. This second check ensured that no big methodological bias was present in the TEs computation process of either these cohorts. The outliers of the comparison, as defined by a residual analysis of the fitted linear model (data not shown) consists in the following events: del 11p, amp 16q, amp10p, amp16p ($p<0.05$), which are probably caused by a sample selection bias in one of the two cohorts.

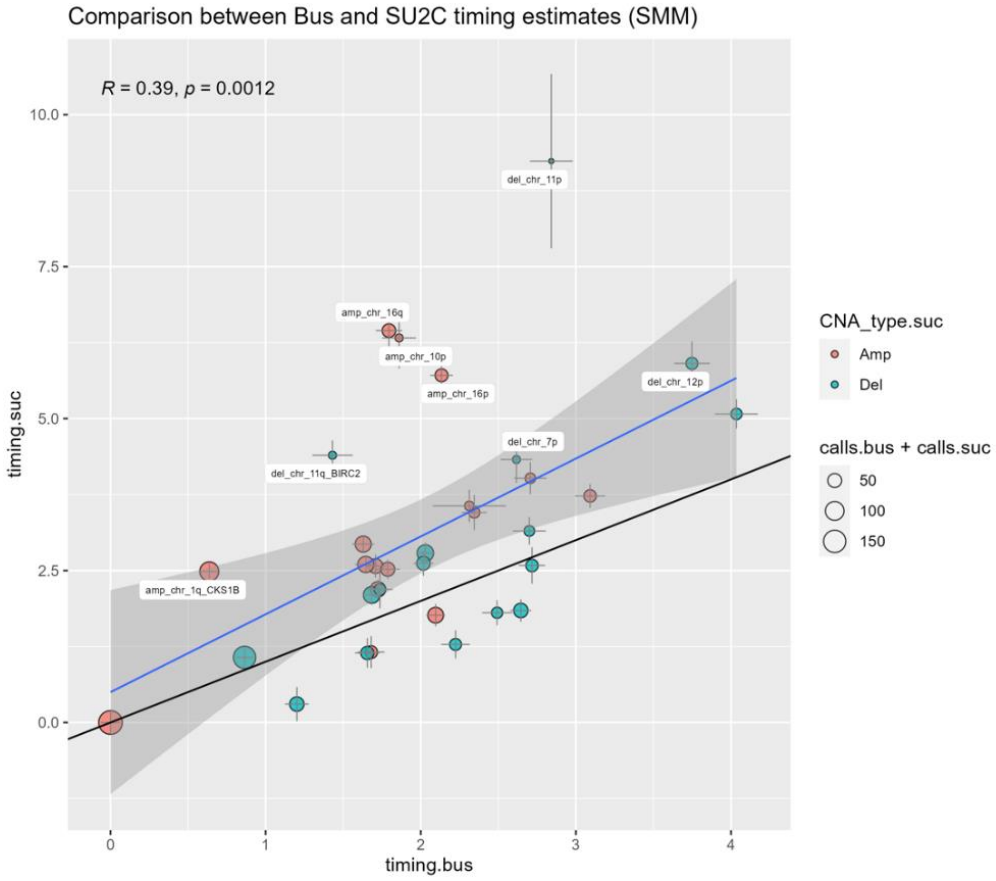


Figure 42: *Correlation analysis between BUS and SU2C timing estimates (TE). Deletion events are represented by blue points, amplification events are represented by red points. The size of the points represents the number of total events observed. Grey lines around the points represent the quasi-standard errors associated to the TEs. The black line indicates the perfect correlation. A linear model was fitted to study the correlation (blue line).*

Then, after merging the two cohorts together a new Bradley-Terry model was generated, and a new time map for the complete SMM cohort was created, as shown in Figure 43. In the SMM cohort the top early occurring events (as compared to Hyperdiploidy) were: amp 11q(CCND1), del 13q(RB1), amp 1q(CKS1B), del 14q(MAX) and del chr 19. This aggregated cohort result confirms that amp 11q(CCND1), del 13q(RB1), amp 1q(CKS1B), del 14q(MAX), that were identified as top early events even in the NDMM cohort, are indeed early alterations even at the SMM disease phase. The identification of those events at both disease phases strongly supports their role as early initiators and "drivers" of the disease pathogenesis.

Additionally, in the aggregated SMM cohort the three deletions of chromosome 1p were confirmed to be late occurring events only at the SMM phase, deepening the hypothesis that these critical alterations (or one of those) could play an important role at the SMM/MM interface.

Intriguingly, the deletion of chr 19, continued to be top early occurring event even in the aggregated SMM cohort, consisting in the event with the biggest difference between the NDMM and the SMM timings, even if it's an extremely rare alteration (N = 19 events, 6.6% in the aggregated SMM cohort).

Figure 43: *timing map of the broad and focal CNAs in the aggregated NDMM cohort. On the left, the Timing Estimates (TE) of both amplifications (in red) and deletions (in blue) are showed with their associated quasi-standard errors (black bars). On the right the frequencies of the various CNAs is shown, colored by original cohort.*

4.4 VALIDATING THE CNAS TIMING ESTIMATES WITH MUTATION DATA

After obtaining the CEs of the different CNAs events among all the cohorts, in order to further validate their timing classification, it was possible to generate two more Bradley-Terry models also including mutations data, for the cohorts where such data was available (CoMM and SU2C). The aim of this analysis was to compare the CNAs TEs with some specific mutations TEs which are known to be early events of MM (e.g. somatic hypermutation mutations, mutations reported with an elevate CCF in other studies) and validate in such a way the early status of the CNAs with a low TE. This "orthogonal" validation was possible since also mutation data, as provided by the MAF file computed by the ABSOLUTE tools, includes 95%CI error estimations and clonality point estimates for each mutation. Consequently, they could be included in the pipeline without

any problem, by adding them in the callset used by the *TestClonality* function in order to perform the clonality contest matches (see Results, "Phase 3: timing analysis" for more details). Specific mutations selected for this analysis consist in the "pathogenic" mutations targeting the top 80 genes more frequently mutated in MM (see methods for more details). After implementing the mutations data in the pipeline, two new Bradley-Terry model and time maps were generated (Figure 44 and 45). The results of this validation analysis show that the IGLL5 mutations were ranked as top early events both in CoMM and SU2C cohorts. IGLL5 is the immunoglobulin light-chain *lambda* gene, which is usually mutated during the somatic hypermutation process, an event that happens very early in the plasma cell development process. Because of this, this mutation can be considered as a proxy of a true early event. The closeness between the IGLL5 mutations and Hyperdiploidy TEs confirmed that the CNAs reference event was correctly timed. Additionally, also DUSP2 mutations, CCND1 mutations and HIST1H1E were classified as top early occurring events, in line with the high CCF level observed by recent molecular characterization studies that imply an early event status.[13]

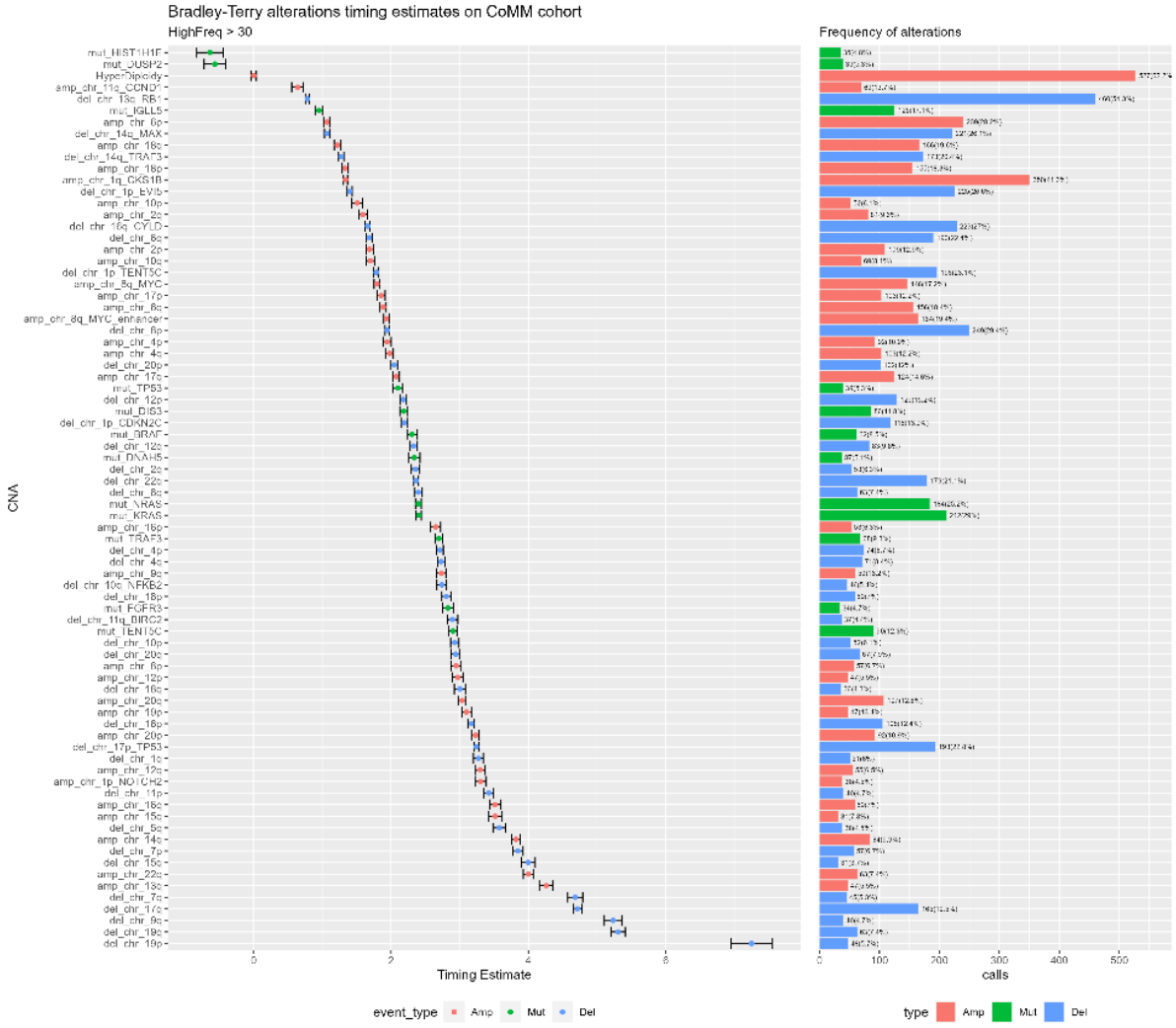


***Figure 44**: timing map of the broad and focal CNAs, with the addition of mutational data, in the CoMM cohort. On the left, the Timing Estimates (TE) of amplifications (in red), deletions (in blue) and mutations (in green) are showed with their associated quasi-standard errors (black bars). On the right the frequencies of the various alterations is shown.*

On the contrary, TP53 mutations, KRAS and NRAS mutations which represent the most frequent mutations in MM, are commonly reported in literature to be subclonal events that happens in a later phase of the evolutive history of MM. [13] This is coherent with the TP53, KRAS and NRAS mutations TEs results. In fact those events are ranked approximately in the middle of the generated timing map, further supporting the validation of the CNAs ranking.

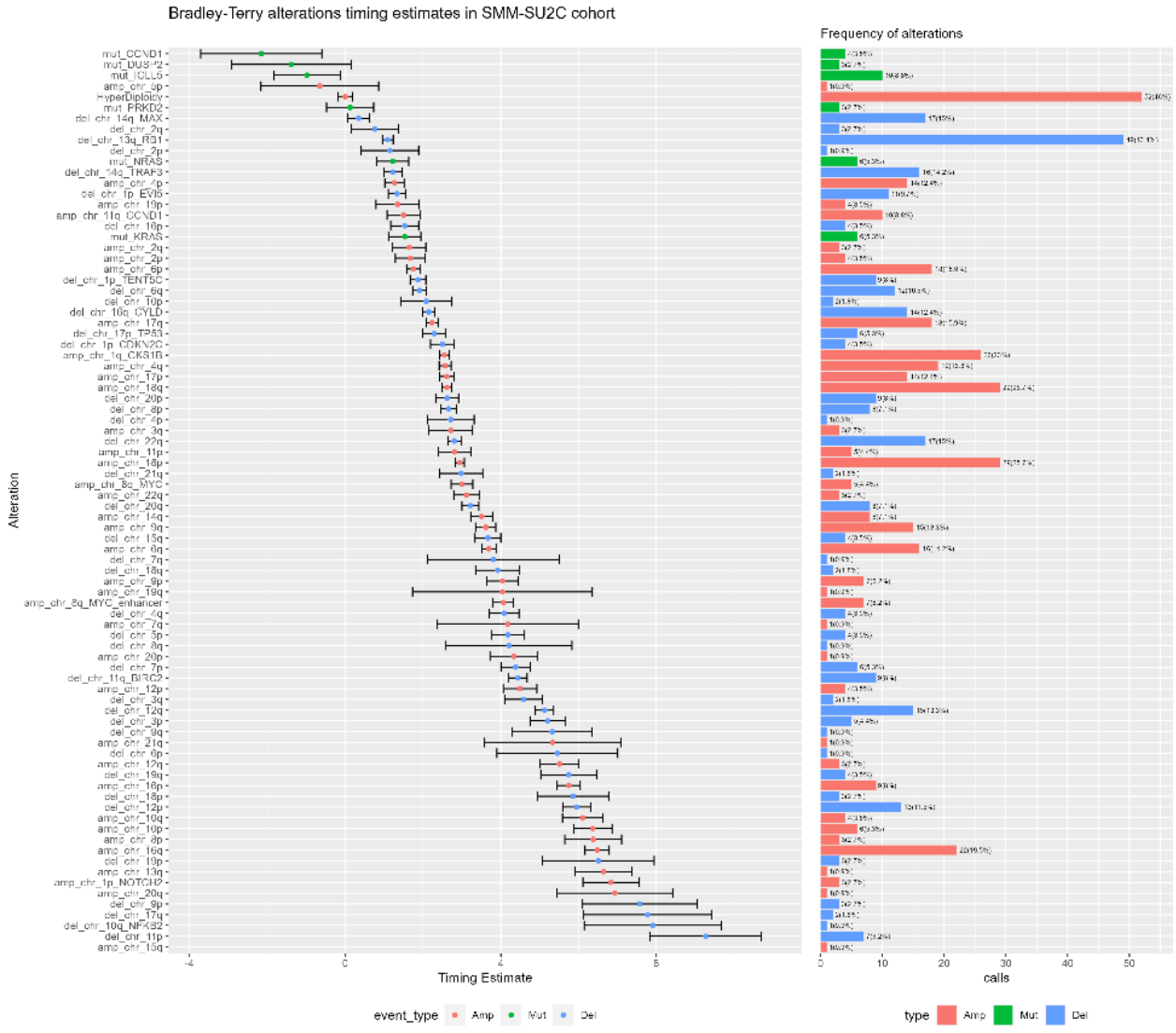


***Figure 45:** timing map of the broad and focal CNAs, with the addition of mutational data, in the SMM SU2C cohort. On the left, the Timing Estimates (TE) of amplifications (in red), deletions (in blue) and mutations (in green) are showed with their associated quasi-standard errors (black bars). On the right the frequencies of the various alterations are shown.*

4.5 COMPARING MM AND SMM TIMING TO STUDY THE DISEASE'S EVOLUTIVE HISTORY

Finally, to investigate the evolutive history of MM in multiple disease phases, the researcher sought to formally compare all the CNAs TEs obtained from the two aggregated cohorts previously generated. This comparison is not a trivial task, given that in order to perform a formal timing comparison between the SMM cohort versus the NDMM cohort, the significant disparity between the number of samples included in the two different cohorts must be taken into account. This means that the numerosity difference, in addition to the noise introduced by the smaller SMM cohort, also causes different scales in the TEs measures (i.e. the Bradley-Terry model computes events abilities (TEs) with more precision in cohorts with many samples, thus potentially generating bigger ability values since more matches are available). For this reason, in order to ensure a correct mathematical comparison between the two aggregated cohorts' timings a scaling normalization transformation of the TEs in both cohorts was performed prior to performing a correlation analysis. This was executed by subtracting the mean and dividing by the Standard Deviation the TEs values (the base *scale* R function was used for this operation). The timing comparison between the scaled TEs values of SMM and NDMM is shown in figure 46.

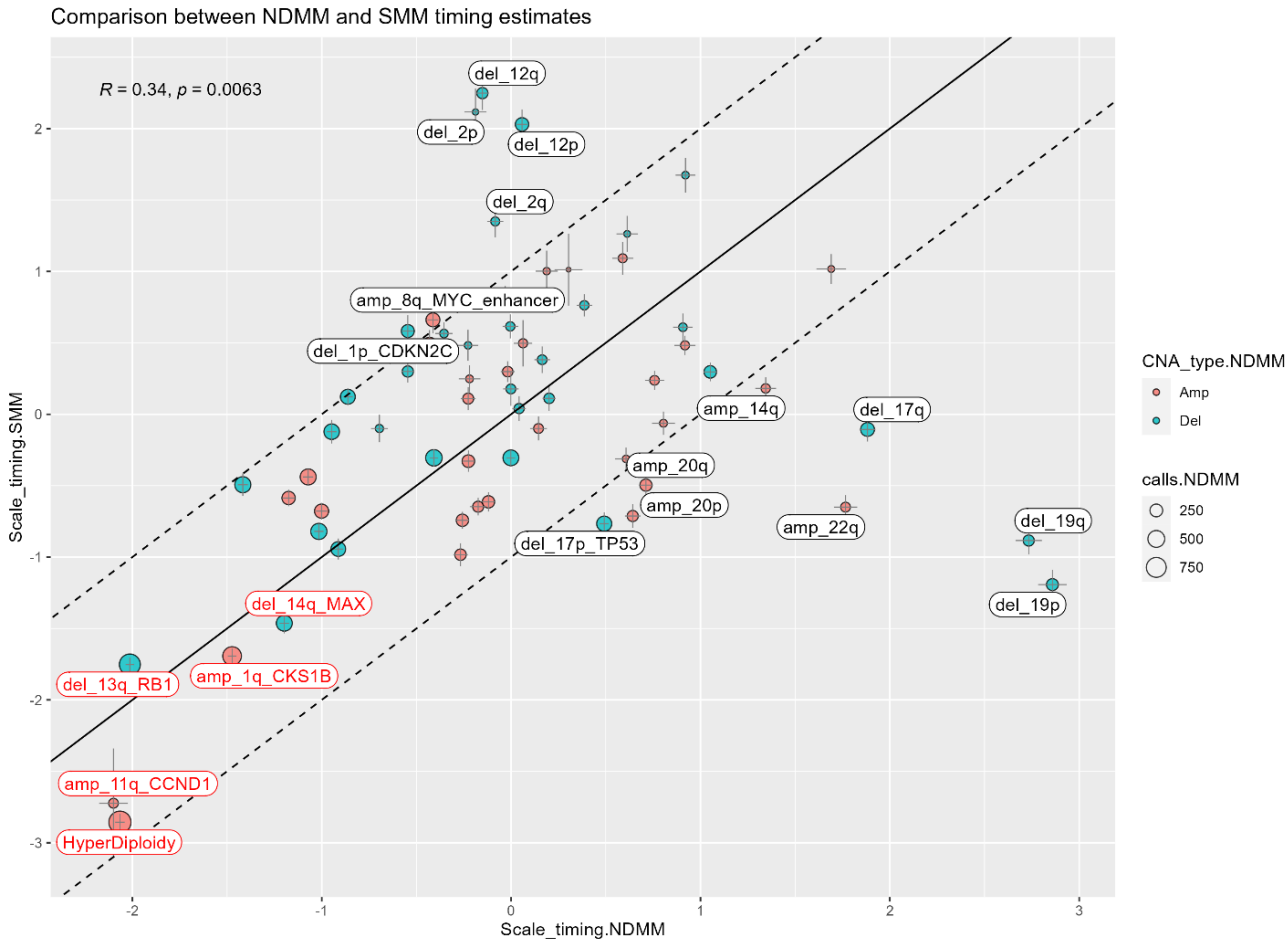


Figure 46: *timing comparison between NDMM and SMM scaled Timing Estimates (sTE). Deletion events are represented by blue points, amplification events are represented by red points. The size of the points represents the number of total events observed. Grey lines around the points represent the quasi-standard errors associated to the TEs. The black line indicates the perfect correlation. Significative outliers in the comparison are highlighted with black labels. The top early occurring events are highlighted with red labels.*

This final analysis showed that the amp 11q(CCND1), amp 1q(CKS1B), del 13q(RB1) and del14(MAX), were confirmed as top early occurring events in both NDMM and SMM disease phases, further validating their role as MM disease driver alterations.

In addition, in order to identify the significative outliers in this comparison, which might represent key events with different timings in the SMM/MM interface, a Z-score was computed for each alteration. In this way events with significative ($p<0.05$) Z scores were identified:

- Early events in NDMM but late event in SMM (Figure 46, upper triangle): including del 1p(CKS1B) and amp8q(MYC enhancer). These events represent genomic alterations which might happen in proximity to the SMM/MM interface, thus potentially contributing to the transition from the asymptomatic SMM phase to the active MM disease phase.
- Late events in NDMM but early event in SMM (Figure 46, lower triangle): including del chromosome 19, amplification of chromosome 20 and del chromosome 17. These events might represent founder events for the development of the SMM disease, however the SMM clones carrying those lesions subsequently tend to be lost at time of progression to active MM.

In any case, it's important to specify once again that these hypothesis on the discordant timing alterations' significance are based on a comparison performed on two very different cohorts, in terms of numerosity. Although a very important effort was made to aggregate the biggest possible cohort of SMM samples, this numerosity difference (NDMM = 1582 samples, SMM = 285 samples) inevitably introduces a substantial intrinsic noise in this final comparison. For this reason, additional studies are required to further integrate the databases of genomic alterations available for SMM. This would be very valuable in order to generate new timing models of equal statistical power, able to elucidate with higher definition the events truly happening during the SMM/MM interface.

4.6 CORRELATIONS OF TIMING RESULTS WITH SURVIVAL DATA

Next, a statistical survival analysis was performed in order to evaluate the clinical significance of the studied genomic events, including the ones which were defined as "drivers" of MM by means of the timing analysis performed in this study. The survival analysis could be performed in all the cohorts where clinical data of patients was available (Overall Survival, OS and Progression Free Survival, PFS), that are the BO and the CoMM cohorts. All the genomic events identified in this study (focal CNAs, broad CNAs and mutations) were tested first in a univariate analysis, including Kaplan-Meier curves and log-rank tests (Figure 47 and 48). Next, all the variables which were found to be significant from the univariate analysis were then included in a multivariate Cox-Proportional Hazard model, automated by a stepwise "backward-forward" variable selection strategy, which was able to optimally identify all the genomic covariates independently associated with outcome (survival) of the patients (Figure 49).

Importantly, among all the identified covariates by this multivariate analysis, the only ones which intersected with the top early events detected in the previous timing analysis were amp 1q(CKS1B) and del13(RB1). On the basis of those results, it is then possible and intriguing to propose that these two alterations represent not only "driver" early alterations in the MM evolutive history, but also important clinical biomarkers, capable of significantly stratify the prognosis of the patients carrying those genomic lesions. Thus, patients with amp 1q(CKS1B) and del13(RB1) might represent a novel biological and clinical entity in MM, which could be integrated and compared with other systems currently used for MM patients' classification.

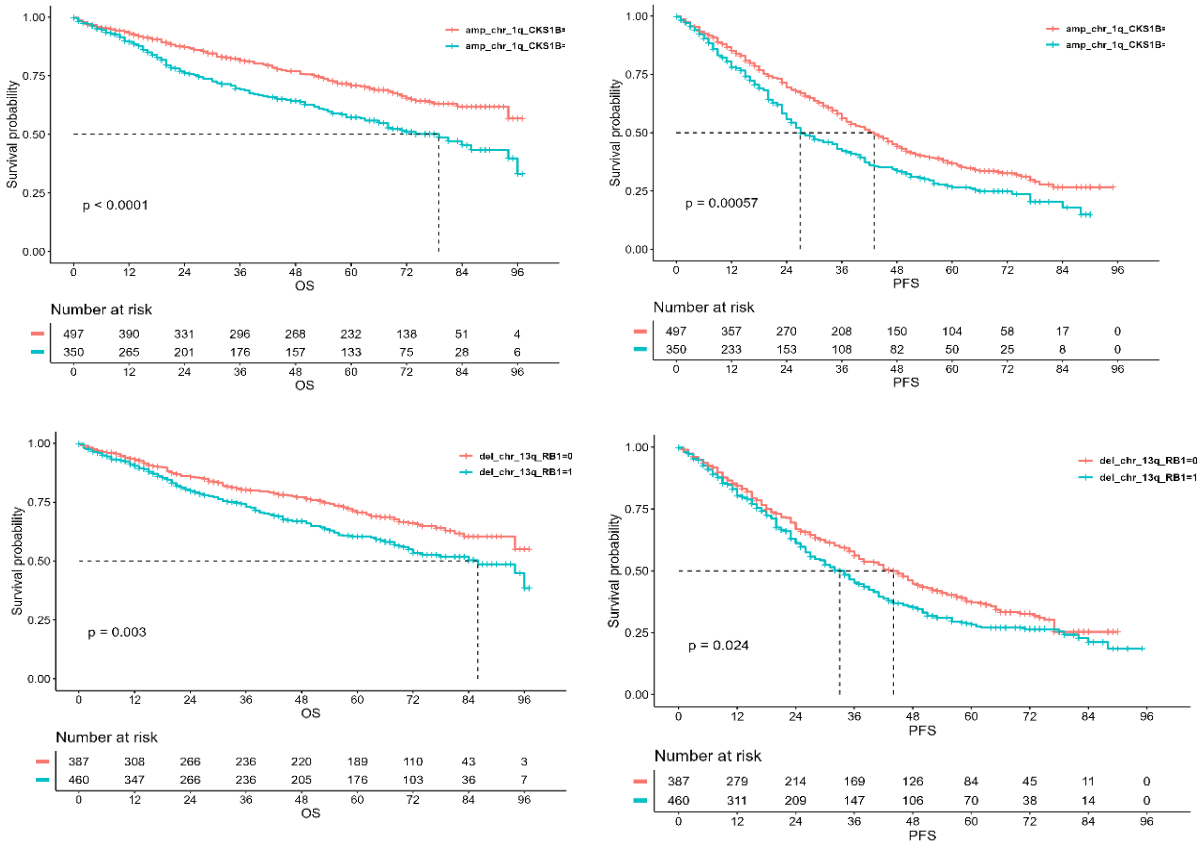


***Figure 47**: Kaplan-Meier curves in the CoMM cohort for the amp 1q(CKS1B) and del 13q(RB1) genomic events. OS curves on the right column and PFS curves on the left column. Every curve shows the associated log-rank test p-value.*

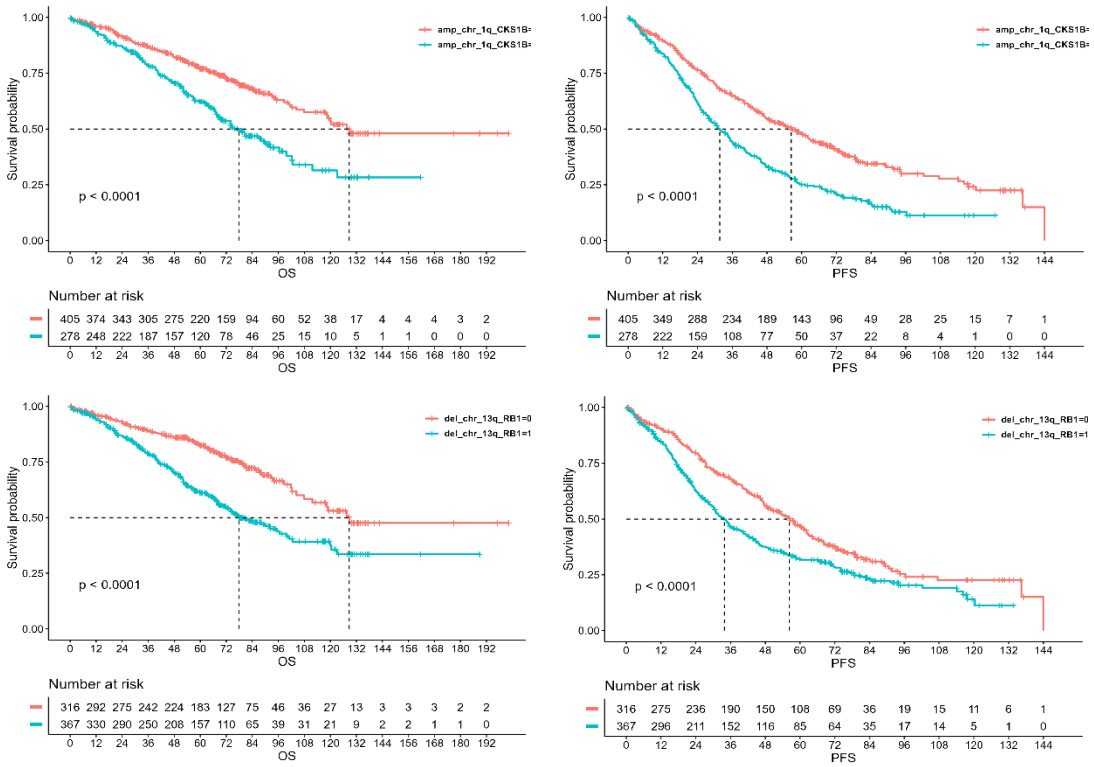


Figure 48: *Kaplan-Meier curves in the BO cohort for the amp 1q(CKS1B) and del 13q(RB1) genomic events. OS curves on the right column and PFS curves on the left column. Every curve shows the associated log-rank test p-value.*

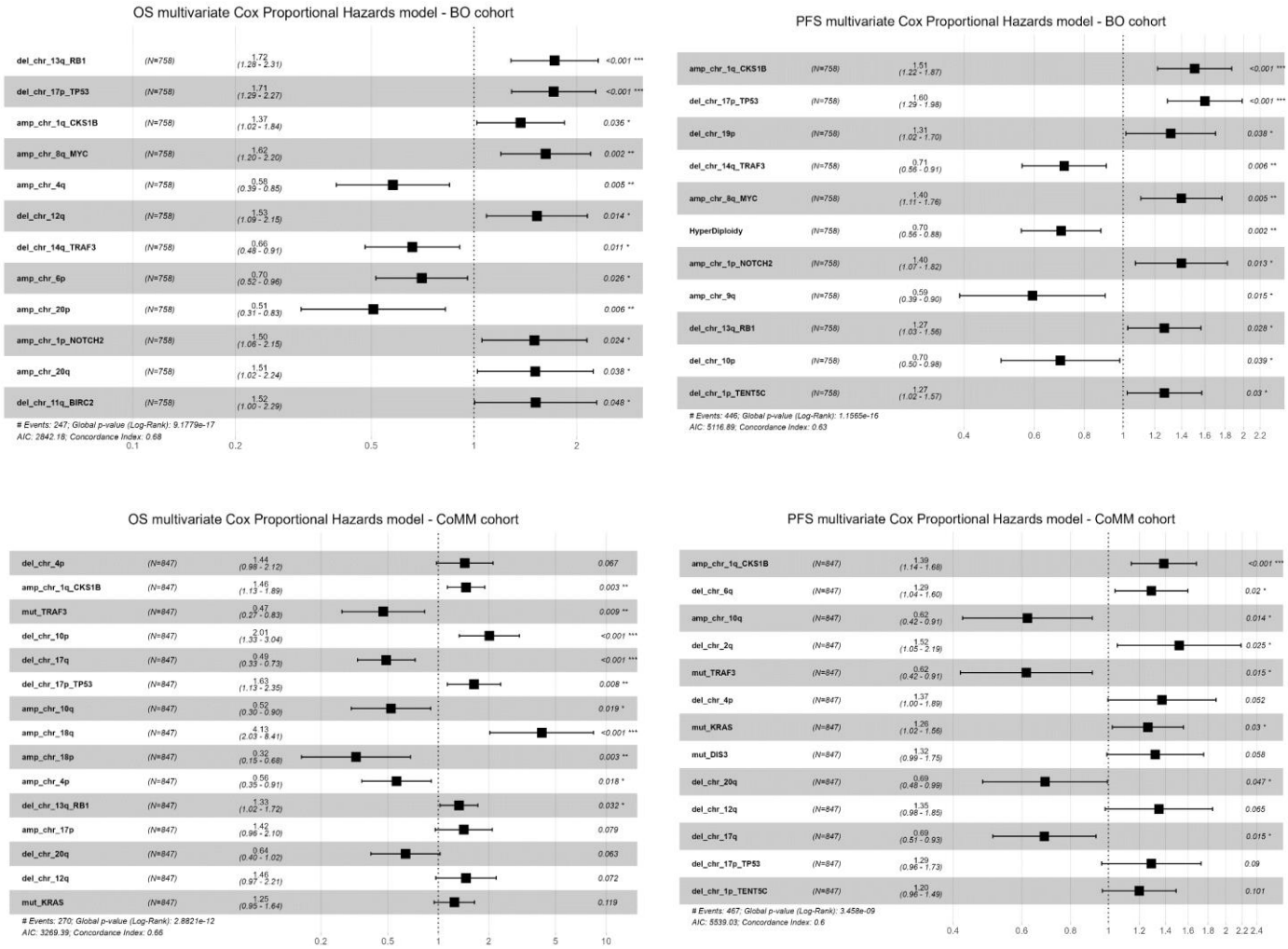


***Figure 49**: Forests plots representing the Multivariate Cox Proportional Hazard models results obtained for both the CoMM and BO cohorts. OS models in the left column, PFS models in the right column.*

5 DISCUSSION

5.1.1.1 Multi-platform CN harmonizing pipeline

In this study, an innovative pipeline for harmonizing copy number data generated by different genomic platforms has been designed and developed. This pipeline is particularly useful for harmonizing data produced by different genomic platforms, in particular NGS and SNP arrays, which represent the two most used platforms for the generation of CN profiles. The development of this pipeline is particularly important in the era of "big data" in bioinformatics, where it is increasingly crucial to effectively integrate data deriving from new prospective studies with those coming from retrospective ones or from other public studies. This pipeline consists of a suite of bioinformatic tools that are interconnected to identify and correct specific biases present in the CN data, including the "baseline-level" bias resolved by the tool *BoBaFIT*, the "Hypersegmentation" bias resolved by the tool *RemasterCNA* and the "purity" bias resolved by the tool *RAPH*. Notably, unlike conventional data harmonization approaches, this pipeline does not act on raw data but on secondary standard files (SEG files, or segmentation files), generated by all the existing platforms while producing CN data. This unconventional approach has the advantage of saving both a huge amount of storage space (reducing the data size of a factor of about 1 million) and a lot of time and computational power in the process of harmonizing CN data. Additionally, the download of raw files is not always feasible, due to their huge size and to the storage costs, whereas, on the contrary, all the main public repositories of tumor genomic data (e.g., GDC Data Portal, EGA archieve) make SEG files readily available and easily accessible. Therefore, the SEG files can be considered as alternative convenient starting points for harmonizing CN data. One important limit of this approach is that the SEG files must be generated by the same segmentation algorithm, in the case of this study the CBS algorithm. This is often not a problem due to the extreme diffusion of the use of CBS algorithm in the tools that generate SEG files.

The pipeline designed in the present study has been applied to harmonize the SEG files generated from three different genomic platforms in four different sample cohorts. Whenever possible, all the obtained results have been validated by performing a comparison with other orthogonal methods, such as with FISH data in the case of results obtained by BOBaFit, or with the popular tool ABSOLUTE in the case of results obtained by RAPH. Furthermore, the quality of harmonization was empirically demonstrated downstream the analysis, by cross-comparing the timing results obtained from different cohorts and from different platforms. This comparison demonstrated good correlation between the Timing Estimates obtained from different platforms and cohorts, thus confirming the quality of harmonization. This is particularly relevant because Timing Estimates rely on the quality of CNAs calls from which they are generated, both in terms of events' presence or absence (qualitative information) and in terms of event's clonality (quantitative information).

The value represented by the suite of tools proposed in this study is double; in fact its modular and versatile structure allows to apply either one or multiple tools of choice, in order to correct

particular biases detected in CN data, regardless of the data's original platform. In addition, the entire suite of tools can be applied as a whole to the SEG files generated by different platforms, in order to obtain fully harmonized CN data across those platforms. To facilitate the use and the accessibility of the suite of tools, all the packages and functions are publicly available on the main bioinformatic repositories (GitHub and soon BioConductor) and freely downloadable upon request.

5.1.1.2 League-model improvement using ComphyNumber and implementation using a GISTIC analysis

Another important milestone achieved in this study was the implementation and improvement of the "league-model" approach, commonly used for calculating the timing of tumor alterations in the evolutionary history of tumors. [31,34] This development has resulted in a better identification of specific "driver" alterations present in the genomic landscape of MM, and an increase in the precision of the timing of CNA alterations in MM compared to previously developed models present in literature to date.

Specifically, the improvement of the timing model involved the introduction of confidence intervals in the calculation of CNA events (by using the *ComphyNumber* tool developed in this study), which were then subsequently considered and utilized when generating "clonality contests", or matches, within the Bradley-Terry model. These matches, in fact, have the particularity of implementing a statistical ad-hoc analysis to establish the winning event of every match, based on the confidence intervals of the events that have to be compared in the comparison. In this way, it has been possible to formally ensure that the data used by the Bradley-Terry model for timing analysis were truthful and of high quality.

Secondly, the implementation of the timing model involved the introduction of focal CNAs events, that have never been considered in previous MM timing models [15,40], but that have been shown to be present and crucial in various types of cancer[79]. In this regard, a formal analysis was accomplished by using the GISTIC tool for the precise identification of "driver" focal events within the genomic landscape of MM. The analysis was conducted on the highest resolution dataset available (BO cohort) and the results obtained were subsequently confirmed on a secondary validation dataset (CoMM cohort). This analysis allowed the identification of 15 focal regions of CNAs (10 deletions and 5 amplifications) categorized as "driver" through an in-depth biological interpretation of the genes contained within the regions, including: CKS1B, MAX, BIRC2, TRAF3, NFKB2, CDKN2C, TP53, RB1, CCND1, MYC, MYC ME2-enhancer, CYLD, EVI5, TENT5C and NOTCH2.

In conclusion, these developments in the timing model have allowed a comprehensive analysis of CNA alterations, including both broad and focal alterations. The improved model has also been able to generate extremely precise timing estimates. This precision was demonstrated by the observation that the obtained confidence intervals of the temporal estimates almost never overlapped, and the intervals were extremely smaller, as compared to the confidence intervals

obtained in the previous timing studies of MM. [15,40] The potential of this improved and implemented Bradley-Terry timing model opens the way for the analysis of timing of other types of tumors using the same approach, in order to significantly improve the resolution of analysis. The GISTIC analysis of "driver" focal alterations is also of great relevance for the study and interpretation of the biological mechanisms underlying the development of MM, as this analysis was not only able to correctly identify all the main known targets of "driver" CNAs alterations (e.g. TP53, CKS1B, MYC, RB1) but also to identify new potential targets, including genes involved in pathways of great importance in MM, such as: EVI5 on chromosome 1p (involved in the cell-cycle pathway), NFKB2 on chromosome 10q (involved in the NF-kB pathway), MAX on chromosome 14q (involved in the Myc-signaling pathway), and NOTCH2 (involved in the NOTCH-signaling pathway).

5.1.1.3 Timing analysis

Finally, regarding the ultimate timing aim of the study, by using the harmonization pipeline and the Bradley-Terry model developed, four different timing analyses were generated for each of the cohorts studied. In each cohort, the top early-occurring events were identified, including amplifications of 11q (CCND1), 13q (RB1), 1q (CKS1B), 14q (MAX) and deletion of 1p (EVI5) in the NDMM (BO and CoMM) cohorts. In particular, CCND1 is a known “driver” of MM that deregulate the cell cycle pathway thus enhancing the tumor plasma cell proliferation. Its deregulation is renown to be an early event in MM since the CCND1 is also the target of the t(11;14) translocation, another well-known early event in MM pathogenesis. Next, del 13q(RB1), amp 1q(CKS1B) are two very frequent alterations that were often, but not always, considered as early alterations in the MM evolutive history. Their role as “drivers” or primary lesions is thus controversial [25,85], but recent studies further validated their role as drivers.[15] They both also deregulate the cell cycle pathway, an early and unifying early event in MM.[86] Del 1p(EVI5) is one of the three deletion peaks identified on chr 1p, in particular EVI5 was described as being involved in both cell cycle and cell migration regulation: in particular it has a role in the completion of cytokinesis and the safeguarding of genomic integrity during cell division; thus, deletions of EVI5 can results in cell-cycle deregulation as well. [87] Finally, del 14q(MAX) involves MAX (MYC associated factor X), a proposed tumor suppressor “driver” gene in MM, as proposed by a recent study.[90] However the study data were based only on the inactivating mutations affecting this gene, thus the del 14q(MAX) may represent a novel mechanism through which this important tumor-suppressor gene is inactivated during the early phase of the disease development. Instead, in the SMM cohorts (SU2C and BUS), the top early-occurring events identified were: amp 11q(CCND1), del 13q(RB1), amp 1q(CKS1B), del 14q(MAX) and del chr 19. The same timing analyses were also performed on the "aggregate cohorts", that is the NDMM cohort composed of BO and CoMM, and the SMM cohort composed of BUS and SU2C. The aggregate analyses allowed on the one hand to increase the statistical power of the analyses by increasing the sample size in the Bradley-Terry models, and on the other hand to demonstrate the quality of the previously developed data harmonization pipeline, since the timing of the alterations in the cohorts to be aggregated showed a good and significant statistical correlation. Ultimately, the aggregate analyses also confirmed with greater confidence the identification of the same previously described top early-occurring

events for both stages of the disease. Of note, the effective timing estimates precision and accuracy in the correct measurement of early alterations was validated by generating two additional Bradley-Terry models in which true “early” early mutations (e.g. IGLL5 mutations generated during the early somatic hypermutation process) events were included as a control. This validation analysis confirmed that the timing estimates were coherent with the expected mutation timings, further supporting the validity of the developed timing models used in this study.

Next, the NDMM timing was compared to the SMM timing, this final analysis showed once again that amp 11q(CCND1), amp 1q(CKS1B), del 13q(RB1) and del14(MAX), were confirmed as top early occurring events in both NDMM and SMM disease phases, further validating their role as MM disease driver alterations. However, the analysis also highlighted events with a significant difference in timing estimates between the two disease phases, including del 1p(CKS1B) which was found as an early event in NDMM but as a later event in SMM. Consequently, it might be speculated that this event might represent a genomic alteration occurring close to the SMM/MM interface, thus potentially contributing to the transition from the asymptomatic SMM to the active MM disease phase. Other events with a significant different timing included deletion of chromosome 19, amplification of chromosome 20 and deletion chromosome 17. Which were found as late events in NDMM but early event in SMM. A possible explanation for this might be that these events might represent founder early events for the development of the SMM disease, however the SMM clones carrying those lesions subsequently tend to get extinct at time of progression to active MM, appearing as subclonal/late after the disease progression. In any case, it is important to note that the conclusions drawn about the significance of these alterations showing a different timing are based on a comparison between two very different groups of patients, in terms of sample size. Despite efforts to increase the sample size of the SMM cohort, the difference (1582 samples for NDMM vs 285 samples for SMM) introduced significant noise in the comparison. Therefore, more research is needed to expand the available genomic data for SMM and thus creating more precise comparisons between timing models from different disease phases.

Finally, a statistical survival analysis was performed in order to evaluate the clinical significance of the identified top early-occurring events. The survival analysis found that both amp 1q(CKS1B) and del13(RB1) which were identified as the earliest "drivers" of MM in the timing analysis also had a significant and independent impact on patient survival, both on OS and PFS.

Based on the results obtained, it is plausible to suggest that amp 1q(CKS1B) and del13(RB1) not only play a significant role in the evolution of MM, but also represent significant biomarkers capable of significantly impact the prognosis of the patients carrying those genomic lesions. Thus, the presence of those genomic alterations could potentially define a new biological and clinical subgroup within MM which would be interesting to integrate and/or compare with other classification systems currently used for MM patients’ stratification.

6 CONCLUSION

This study aimed to achieve three main goals: 1) to develop a multi-platform data harmonization pipeline for CN data in order to harmonize the genomic profiles belonging to four different patient cohorts in two MM disease phases (two NDMM cohorts and two SMM cohorts), 2) to improve and develop existing MM genomics timing models, and 3) to finally generate a new timing model, capable of confidently and accurately identifying "driver"/early events in the evolutive history of MM. The identification of such events is crucial since they are able to deeply define the biology and genetics of the disease biological mechanisms. However, their identification has always been a challenging and complicated task due to the presence of a multitude of "passenger" events that introduce noise in the genomic landscape of MM. To this aim, timing analysis represent an innovative approach, as it is capable to reliably identify “driver” events among the genomic noise.

In this study, five new bioinformatics tools were developed for CN profile harmonization, each with the specific goal of identifying and correcting a different potential bias in CN data. These five tools can be used either individually or integrated together in a suite, which can be downloaded publicly and freely from main bioinformatics repositories. The harmonized data produced by the suite of tools was used to develop an innovative Bradley-Terry timing model, improved with a novel statistical approach which increases the confidence and the quality of results, and implemented by the introduction of focal CNAs events identified through a formal GISTIC analysis. Indeed, the GISTIC analysis performed in this study revealed unexpected relevant results, such as the identification of novel MM “driver” genes alterations previously not discovered, but strongly implicated in MM tumor biology. Some of these alterations consists in deletion of NFKB2 on chromosome 10q, amplification of NOTCH2 on chromosome 1p and deletion of MAX on chromosome 14q. Finally, this innovative timing model was used to generate and validate temporal maps in both the NDMM and SMM phases. This analysis allowed the identification of specific "top early-occurring events" among both disease phases, including amp 1q(CKS1B), del 13q(RB1), amp 11q(CCND1) and del 14q(MAX). These events were validated as primary and ancestral events in the evolutionary history of MM by means of a comparison with early-mutation events used as timing controls. Among these four identified top-early events, it is important to note that the classification of del14(MAX) as an ancestral event of MM is an innovative discovery, as this event has never been detected nor considered as an MM driver before this study. Importantly the survival analysis to characterize the clinical impact of these driver events revealed that amp 1q(CKS1B) and del 13q(RB1) can also play a role as important MM biomarkers, since they are capable of strongly predict the patients’ survival outcomes.

This study significantly contributes to the advancement of the discovery of the true the biological alterations underlying the development and progression of cancer, by defining a method to precisely characterize the precise timing at which CNAs events occur. Additionally, this study contributes to the development of the bioinformatics CN analysis field, by introducing a completely novel suite of tools that can easily be applied in other projects where CN harmonization or data bias-cleaning is required. Finally, the importance of this study relies in the identification of new, previously unknown, MM-driving genetic alterations and the validation of them as early

events in the disease chronology, also demonstrating the clinical utility of some of them as strong biomarkers.

This study remains open to further implementations, including the possibility of introducing in the timing model different event types, such as structural events, translocations, and complex structural events (e.g. chromothripsis and chromoplexity). Even though this would be of extreme interest, the implementation is not an easy task due to the technical difficulties in establishing precisely the exact clonality of this type of alterations. Additionally, the study leaves open the possibility of applying the same timing pipeline to CNAs and mutations in other types of cancer. Finally, the future developments of this study will involve the search and aggregation of larger SMM sample cohorts, in order to further improve the obtainable timing precision in this crucial stage of the disease, improving in such a way the timing estimates and making them even more comparable to those calculated in NDMM phase. Another possible future development will involve the search and aggregation of a sufficiently powerful cohort of relapsed MM samples to introduce in the timing study of a third temporal point, thus improving even more the potential of analysis of the timing model that can be generated.

7 BIBLIOGRAPHY

1. van de Donk, N. W. C. J., Pawlyn, C. & Yong, K. L. Multiple myeloma. *The Lancet* **397**, 410–427 (2021).

2. Rajkumar, S. V. Multiple myeloma: 2020 update on diagnosis, risk-stratification and management. *Am J Hematol* **95**, 548–567 (2020).

3. Walker, B. A. *et al.* Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. *Leukemia* **28**, 384–390 (2014).

4. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**, (2014).

5. Lohr, J. G. *et al.* Widespread genetic heterogeneity in multiple myeloma: Implications for targeted therapy. *Cancer Cell* **25**, 91–101 (2014).

6. Kyle, R. A. *et al.* Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma. *N Engl J Med* **356**, 2582–2590 (2007).

7. Dutta, A. K. *et al.* Single-cell profiling of tumour evolution in multiple myeloma — opportunities for precision medicine. *Nature Reviews Clinical Oncology* vol. 19 223–236 Preprint at https://doi.org/10.1038/s41571-021-00593-y (2022).

8. Rajkumar, S. V. *et al.* International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol* **15**, e538–e548 (2014).

9. Kyle, R. A., Buadi, F. & Vincent Rajkumar, S. Management of Monoclonal Gammopathy of Undetermined Significance (MGUS) and Smoldering Multiple Myeloma (SMM). *Oncology (Williston Park)* **25**, 578 (2011).

10. Manier, S. *et al.* Genomic complexity of multiple myeloma and its clinical implications. *Nat Rev Clin Oncol* **14**, 100–113 (2017).

11. Ji, X. *et al.* Distinguishing between cancer driver and passenger gene alteration candidates via cross-species comparison: A pilot study. *BMC Cancer* **10**, 1–13 (2010).

12. Walker, B. A. *et al.* Mutational spectrum, copy number changes, and outcome: Results of a sequencing study of patients with newly diagnosed myeloma. *Journal of Clinical Oncology* **33**, 3911–3920 (2015).

13. Walker, B. A. *et al.* Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma. *Blood* **132**, 587–597 (2018).

14. Bustoros, M. *et al.* Genomic Profiling of Smoldering Multiple Myeloma Identifies Patients at a High Risk of Disease Progression. *J Clin Oncol* **38**, 2380–2389 (2020).

15. Maura, F. *et al.* Genomic landscape and chronological reconstruction of driver events in multiple myeloma. (2019).

16. Hoang, P. H. *et al.* Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. *Leukemia* **32**, 2459–2470 (2018).

17. Barwick, B. G. *et al.* Multiple myeloma immunoglobulin lambda translocations portend poor prognosis. *Nature Communications 2019 10:1* **10**, 1–13 (2019).

18. Dutta, A. K. *et al.* Single-cell profiling of tumour evolution in multiple myeloma — opportunities for precision medicine. *Nature Reviews Clinical Oncology* Preprint at https://doi.org/10.1038/s41571-021-00593-y (2022).

19. Pawlyn, C. & Morgan, G. J. Evolutionary biology of high-risk multiple myeloma. *Nat Rev Cancer* **17**, 543–556 (2017).

20. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).

21. Weinhold, N. *et al.* Clonal selection and double hit events involving tumor suppressor genes underlie relapse from chemotherapy: myeloma as a model. *Blood* **128**, blood-2016-06-723007 (2016).

22. Corre, J. *et al.* Multiple myeloma clonal evolution in homogeneously treated patients. *Leukemia* (2018) doi:10.1038/s41375-018-0153-6.

23. Beerenwinkel, N., Schwarz, R. F., Gerstung, M. & Markowetz, F. Cancer evolution: Mathematical models and computational inference. *Syst Biol* **64**, e1–e25 (2015).

24. Kumar, S. Emerging options in multiple myeloma: targeted, immune, and epigenetic therapies. *Hematology: the American Society of Hematology Education Program* **2017**, 518 (2017).

25. Kumar, S. K. & Rajkumar, S. V. The multiple myelomas — current concepts in cytogenetic classification and therapy. *Nature Reviews Clinical Oncology 2018 15:7* **15**, 409–421 (2018).

26. Munshi, N. C. *et al.* Association of Minimal Residual Disease With Superior Survival Outcomes in Patients With Multiple Myeloma: A Meta-analysis. *JAMA Oncol* **3**, 28–35 (2017).

27. Kumar, S. *et al.* International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma. *The Lancet Oncology* vol. 17 e328–e346 Preprint at https://doi.org/10.1016/S1470-2045(16)30206-6 (2016).

28. Majithia, N. *et al.* Early relapse following initial therapy for multiple myeloma predicts poor outcomes in the era of novel agents. *Leukemia* **30**, 2208 (2016).

29. Anderson, K. C. *et al.* The role of minimal residual disease testing in myeloma treatment selection and drug development: Current value and future applications. *Clinical Cancer Research* **23**, 3980–3993 (2017).

30. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

31. Jolly, C. & Loo, P. van. Timing somatic events in the evolution of cancer. 1–9 (2018).

32. Durinck, S. *et al.* Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov* **1**, 137–143 (2011).

33. Jolly, C. & Loo, P. van. Timing somatic events in the evolution of cancer. 1–9 (2018).

34. Gerstung, M., Jolly, C., Leshchiner, I. & Dentro, S. C. The evolutionary history of 2,658 cancers. (2017) doi:10.1101/161562.

35. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209–2221 (2016).

36. Leshchiner, I. *et al.* Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. *bioRxiv* 508127 (2019) doi:10.1101/508127.

37. Demeulemeester, J. *et al.* Tracing the origin of disseminated tumor cells in breast cancer using single-cell sequencing. *Genome Biol* **17**, 1–15 (2016).

38. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature 2020 578:7793* **578**, 122–128 (2020).

39. Nik-zainal, S. *et al.* The Life History of 21 Breast Cancers. (2011) doi:10.1016/j.cell.2012.04.023.

40. Aktas Samur, A. *et al.* Deciphering the chronology of copy number alterations in Multiple Myeloma. *Blood Cancer J* **9**, (2019).

41. Spence, T. & Dubuc, A. M. Copy Number Analysis in Cancer Diagnostic Testing. *Clin Lab Med* **42**, 451–468 (2022).

42. Kendall, J. & Krasnitz, A. Computational methods for DNA copy-number analysis of tumors. *Methods Mol Biol* **1176**, 243 (2014).

43. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics* **14**, 1–16 (2013).

44. Gabrielaite, M. *et al.* A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *bioRxiv* 2021.04.30.442110 (2021) doi:10.1101/2021.04.30.442110.

45. Mayrhofer, M., Viklund, B. & Isaksson, A. Rawcopy: Improved copy number analysis with Affymetrix arrays. *Sci Rep* **6**, 36158 (2016).

46. Bioconductor - DNAcopy. https://bioconductor.org/packages/release/bioc/html/DNAcopy.html.

47. 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).

48. aaronmck/CapSeg: CapSeg - Copy Number from Exome Sequencing. https://github.com/aaronmck/CapSeg.

49. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).

50. Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L. & Beroukhim, R. GISTIC2 . 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).

51. Dobson, J. GISTIC2 Documentation. 1–9.

52. Schmidt, T. M., Fonseca, R. & Usmani, S. Z. Chromosome 1q21 abnormalities in multiple myeloma. *Blood Cancer Journal 2021 11:4* **11**, 1–11 (2021).

53. Bolli, N. *et al.* Analysis of the genomic landscape of multiple myeloma highlights novel prognostic markers and disease subgroups. *Leukemia* (2017) doi:10.1038/leu.2017.344.

54. Jovanović, K. K. *et al.* Deregulation and targeting of TP53 pathway in multiple myeloma. *Front Oncol* **9**, 665 (2019).

55. Martello, M. *et al.* Clonal and subclonal TP53 molecular impairment is associated with prognosis and progression in multiple myeloma. *Blood Cancer J* **12**, (2022).

56. Misund, K. *et al.* MYC dysregulation in the progression of multiple myeloma. *Leukemia* **34**, 322 (2020).

57. Walker, B. A. The Chromosome 13 Conundrum in Multiple Myeloma. *Blood Cancer Discov* **1**, 16–17 (2020).

58. Lionetti, M. *et al.* Genomics of Smoldering Multiple Myeloma: Time for Clinical Translation of Findings? *Cancers 2021, Vol. 13, Page 3319* **13**, 3319 (2021).

59. Boyle, E. M. *et al.* The molecular make up of smoldering myeloma highlights the evolutionary pathways leading to multiple myeloma. *Nat Commun* **12**, (2021).

60. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413–421 (2012).

61. Bradley, R. A. & Terry, M. E. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* **39**, 324 (1952).

62. Augustin, T. Bradley-Terry-Luce models to incorporate within-pair order effects: representation and uniqueness theorems. *Br J Math Stat Psychol* **57**, 281–294 (2004).

63. Bradley-Terry Models [R package BradleyTerry2 version 1.1-2]. (2020).

64. Bradley, R. A. 14 Paired comparisons: Some basic procedures and examples. *Handbook of Statistics* **4**, 299–326 (1984).

65. SHAM, P. C. & CURTIS, D. An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* **59**, 323–336 (1995).

66. Overcoming the Reference Category Problem in the Presentation of Statistical Models on JSTOR. https://www.jstor.org/stable/1519851#metadata_info_tab_contents.

67. Quasi-Variances on JSTOR. https://www.jstor.org/stable/20441079#metadata_info_tab_contents.

68. 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).

69. Storage and Computation Requirements | Strand NGS. https://www.strand-ngs.com/support/ngs-data-storage-requirements.

70. Amazon S3 Simple Storage Service Pricing - Amazon Web Services. https://aws.amazon.com/s3/pricing/?nc1=h_ls.

71. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 1–16 (2012).

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

73. Rasmussen, M. *et al.* Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol* **12**, R108 (2011).

74. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. **463**, 899–905 (2010).

75. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J Stat Softw* **61**, 1–36 (2014).

76. van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910–16915 (2010).

77. Fawcett, L. Using Interactive Shiny Applications to Facilitate Research-Informed Learning and Teaching. *Journal of Statistics Education* **26**, 2–16 (2018).

78. W. J. Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics* 448–485 (1960).

79. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature 2020 578:7793* **578**, 112–121 (2020).

80. broadinstitute/PhylogicNDT. https://github.com/broadinstitute/PhylogicNDT.

81. Cumming, G. & Finch, S. Inference by eye confidence intervals and how to read pictures of data. *American Psychologist* **60**, 170–180 (2005).

82. Higgins JPT *et al. Cochrane Handbook for Systematic Reviews of Interventions (updated February 2022)*. (Cochrane, 2022).

83. Basic Statistics and Data Analysis [R package BSDA version 1.2.1]. (2021).

84. Lancho, O. & Herranz, D. The MYC enhancer-ome: Long-range transcriptional regulation of MYC. *Trends Cancer* **4**, 810 (2018).

85. Chiecchio, L. *et al.* Timing of acquisition of deletion 13 in plasma cell dyscrasias is dependent on genetic context. *Haematologica* **94**, 1708–1713 (2009).

86. Bergsagel, P. L. *et al.* Cyclin D dysregulation: an early and unifying pathogenic event in multiple myeloma. *Blood* **106**, 296 (2005).

87. Terragna, C., Remondini, D., Martello, M. & Zamagni, E. The genetic and genomic background of multiple myeloma patients achieving complete response after induction therapy with bortezomib , thalidomide and dexamethasone (VTD) The genetic and genomic background of multiple myeloma patients achieving complete . (2016) doi:10.18632/oncotarget.5718.

88. Roy, P., Sarkar, U. A. & Basak, S. The NF-κB Activating Pathways in Multiple Myeloma. *Biomedicines* **6**, (2018).

89. Skerget, S. *et al.* Genomic Basis of Multiple Myeloma Subtypes from the MMRF CoMMpass Study. *medRxiv* 2021.08.02.21261211 (2021) doi:10.1101/2021.08.02.21261211.

90. Barrio Garcia, S. *et al.* Role of MAX As a Tumor Suppressor Driver Gene in Multiple Myeloma. *Blood* **130**, 4347–4347 (2017).