CNApp: quantification of genomic copy number alterations in cancer and integrative analysis to unravel clinical implications

Sebastià Franch-Expósito[1#], Laia Bassaganyas[2#], Maria Vila-Casadesús[3], Eva Hernández-Illán[1], Roger Esteban-Fabró[2], Marcos Díaz-Gay[1], Juan José Lozano[3], Antoni Castells[1], Josep M. Llovet[2,4,5], Sergi Castellví-Bel[1*], Jordi Camps[1,6*]

[1] Gastrointestinal and Pancreatic Oncology Team, Institut D'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clínic de Barcelona, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Universitat de Barcelona, Barcelona (08036), Spain

[2] Liver Cancer Translational Research Group, Liver Unit, Institut D'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clínic, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Universitat de Barcelona, Barcelona (08036), Spain

[3] Bioinformatics Unit, CIBEREHD, Barcelona (08036), Spain

[4] Mount Sinai Liver Cancer Program, Division of Liver Diseases, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York (10029), USA

[5] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona (08036), Spain

[6] Unitat de Biologia Cel·lular i Genètica Mèdica, Departament de Biologia Cel·lular, Fisiologia i Immunologia, Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra (08193), Spain

Symbol

[#] Equal contribution.

[*] Corresponding authors.

E-mail: JCAMPS@clinic.cat (Camps J), SBEL@clinic.cat (Castellví-Bel S).

ABSTRACT

Somatic copy number alterations (CNAs) are a hallmark of cancer. Although CNA profiles have been established for most human tumor types, their precise role in tumorigenesis as well as their clinical and therapeutic relevance remain largely unclear. Thus, computational and statistical approaches are required to thoroughly define the interplay between CNAs and tumor phenotypes. Here we developed CNApp, a user-friendly web tool that offers sample- and cohort-level computational analyses, allowing a comprehensive and integrative exploration of CNAs with clinical and molecular variables. By using purity-corrected segmented data from multiple genomic platforms, CNApp generates genome-wide profiles, computes CNA scores for broad, focal and global CNA burdens, and uses machine learning-based predictions to classify samples. We applied CNApp to a pan-cancer dataset of 10,635 genomes from TCGA showing that CNA patterns classify cancer types according to their tissue-of-origin, and that broad and focal CNA scores positively correlate in samples with low amounts of whole-chromosome and chromosomal arm-level imbalances. Moreover, using the hepatocellular carcinoma cohort from the TCGA repository, we demonstrate the reliability of the tool in identifying recurrent CNAs, confirming previous results. Finally, we establish machine learning-based models to predict colon cancer molecular subtypes and microsatellite instability based on broad CNA scores and specific genomic imbalances. In summary, CNApp facilitates data-driven research and provides a unique framework for the first time to comprehensively assess CNAs and perform integrative analyses that enable the identification of relevant clinical implications. CNApp is hosted at http://cnapp.bsc.es.

KEYWORKS: Copy number alterations; Cancer genomics; CNA scores; Shiny app; Pan-cancer; Colorectal cancer; Hepatocellular carcinoma

INTRODUCTION

The presence of somatic copy number alterations (CNAs) is a ubiquitous feature in cancer. In fact, the distribution of CNAs is sufficiently tissue-specific to distinguish tumor entities (Ried et al., 2012; Taylor et al., 2018a), and allows identifying groups of tumors responsive to particular therapies (Cairncross et al., 2013; Davoli, Uno, Wooten, & Elledge, 2017). Moreover, high levels of CNAs, which result from aneuploidy and chromosome instability, are generally associated with high-grade tumors and poor prognosis (Sansregret, Vanhaesebroeck, & Swanton, 2018).

Two main subtypes of CNAs can be discerned: broad CNAs, which are defined as whole-chromosome and chromosomal arm-level alterations, and focal CNAs, which are alterations of limited size ranging from part of a chromosome-arm to few kilobases (Krijgsman, Carvalho, Meijer, Steenbergen, & Ylstra, 2014; Zack et al., 2013). Recently, it has been uncovered that while focal events mainly correlate with cell cycle and proliferation markers, broad aberrations are mainly associated with immune evasion markers, suggesting that tumor immune features might be determined by mechanisms related to overall gene dosage imbalance rather than specific actionable genes (Buccitelli et al., 2017; Davoli et al., 2017; Taylor et al., 2018b). Nevertheless, the precise role of CNAs in tumor initiation and progression, as well as their clinical relevance and therapeutic implications remain still poorly understood.

Interpretation and visualization of CNAs is time-consuming and very often requires complex analyses with clinical and molecular information. Well-established CNA algorithms, such as the gold-standard circular binary segmentation, define the genomic boundaries of copy number gains and losses based on signal intensities or read depth obtained from array comparative genomic hybridization and SNP-array or next-generation sequencing data, respectively (Olshen, Venkatraman, Lucito, & Wigler, 2004). However, the tumor-derived genomic complexity may cause an under- or overestimation of CNAs. This complexity is represented by tumor purity, tumor aneuploidy, and intratumor heterogeneity, which imply high levels of subclonal alterations. Thus, recent segmentation methods improved the accuracy to identify copy number segments in tumor samples either by considering the B allele frequency (BAF), such as ExomeCNV (Sathirapongsasuti et al., 2011), Control-FREEC (Boeva et al., 2012) and SAAS-CNV (Zhang & Hao, 2015), or through adjusting by sample purity and ploidy estimates, such as GAP (Popova et al., 2009), ASCAT (Van Loo et al., 2010) and ABSOLUTE (Carter et al., 2012). However, the state-of-the-art

computational approach for CNA analysis in cancer is GISTIC2.0 (Mermel et al., 2011), which is a gene-centered probabilistic method that enables to define the boundaries of recurrent putative driver CNAs in large cohorts (Beroukhim et al., 2010). Nevertheless, despite ongoing progress on identifying CNAs, to our knowledge none of the existing software packages is readily available for integrative analyses to unveil their biological and clinical implications.

To address this issue, we developed CNApp, the first open-source application to quantify CNAs and integrate genomic profiles with molecular and clinical variables. CNApp is a web-based tool that provides the user with high-quality interactive plots and statistical correlations between CNAs and annotated variables in a fast and easy-to-explore interface. In particular, CNApp uses purity-corrected genomic segmented data from multiple genomic platforms to redefine CNA profiles, to compute CNA scores based on the number, length and amplitude of broad and focal genomic alterations, to assess differentially altered genomic regions, and to perform machine learning-based predictions to classify tumor samples. To exemplify the applicability and performance of CNApp, we used publicly available segmented data from The Cancer Genome Atlas (TCGA) to (i) measure the burden of global, broad, and focal CNAs as well as generate CNA profiles in a pan-cancer dataset spanning 33 cancer types, (ii) identify cohort-based recurrent CNAs in hepatocellular carcinoma and compare them with previously reported data, and (iii) assess predicting models for colon cancer molecular subtypes and microsatellite instability status based on CNA scores and specific genomic imbalances. CNApp is hosted at http://cnapp.bsc.es and the source code is freely available at GitHub (https://github.com/ait5/CNApp).

RESULTS

Implementation

CNApp comprises three main sections: 1- *Re-Seg & Score: re-segmentation, CNA scores computation, variable association* and *survival analysis*, 2- *Region profile: genome-wide CNA profiling, CNA frequencies, correlation profiles* and *descriptive regions*, and 3- *Classifier model: machine learning classification model predictions* (Figure 1). Each of these sections and their key functions are described below. The input file consists of a data frame with copy number segments provided by any segmentation algorithm. Mandatory fields and column headers are sample name (*ID*),

chromosome (*chr*), start (*loc.start*) and end (*loc.end*) genomic positions, and the log2 ratio of the copy number amplitude (*seg.mean*) for each segment. Section 1 incorporates the correction for tumor purity (i.e., fraction of tumor cells in the sample) to measure the actual magnitude of CNAs. Thus, when available, the input file will also include sample purity estimations (*purity*) and BAF values (*BAF*), which correct the accuracy of CNA calls and provide copy number neutral loss-of-heterozygosity (CN-LOH) events. Annotation of variables can be included in the input file (tagged in every segment from each sample) or by uploading an additional file indicating new variables *per* sample.

Section 1. Re-Seg & Score: re-segmentation, CNA scores computation, variable association and survival analysis

First, CNApp applies a re-segmentation approach to adjust for amplitude divergence due to technical variability and correct for estimated tumor purity (Supplementary Methods). Default re-segmentation settings include *minimum segment length* (100 Kbp), *minimum amplitude (seg.mean) deviation from segment to zero* (0.16), *maximum distance between segments* (1 Mbp), *maximum amplitude (seg.mean) deviation between segments* (0.16), and *maximum BAF deviation between segments* (0.1). These parameters can be customized by the user to better adjust the re-segmentation for each particular dataset. Re-segmented data are then used to calculate the broad (BCS), focal (FCS) and global (GCS) CNA scores, which provide three different quantifications of CNA levels for each sample. To compute these scores, CNApp classifies and weights CNAs based on their length and amplitude. For each sample, BCS is computed by considering broad (chromosome and arm-level) segment weights according to the amplitude value. Likewise, calculation of FCS takes into account weighted focal CNAs corrected by the amplitude and length of the segment. Finally, GCS is computed by considering the sum of normalized BCS and FCS, providing an overall assessment of the CNA burden.

To assess the reliability of CNA scores, we compared each score with the corresponding fraction of altered genome using a TCGA pan-cancer set of 10,635 samples. Both BCS (ranging from 0 to 44) and FCS (values ranging from 5 to 2,466) highly correlated with the fraction of altered genome by broad and focal copy number changes, respectively (Spearman's rank correlation for BCS = 0.957 and for FCS = 0.938) (Supplementary Figure S1A and B). As expected, GCS (values ranged from -1.93 to 12.60) highly correlated with the fraction of altered genome affected by both

broad and focal CNAs (Spearman's rank correlation for GCS = 0.963 (Supplementary Figure S1C). Parametric and non-parametric statistical tests are used to establish associations between CNA scores and annotated variables from the input file. Additionally, Kaplan-Meier survival curves are computed using either CNA scores or additional variables.

Section 2. Region profile: genome-wide CNA profiling

This section transforms segmented data (either re-segmented data from section 1 or original segments uploaded by the user) into genomic region profiles to allow sample-to-sample comparisons. Different genomic windows can be selected to compute the genomic profiles (i.e., chromosome arms, half-arms, cytobands, sub-cytobands or 40-1 Mbp windows). All segments, or either only broad or only focal can be selected for this analysis. Length-relative means are computed for each window by considering amplitude values from those segments included in each specific window. Default cutoffs for low-level copy number gains and losses (i.e., |0.2|) are used to infer CNA frequencies. Genomic profiles are presented in genome-wide heatmaps to visualize general copy number patterns. Up to six annotation tracks can be added and plotted simultaneously allowing visual comparison and correlation between CNA profiles and different variables, including the CNA scores obtained in section 1. CNA frequency summaries by genomic region and by sample are represented as stacked bar plots. Correlation values and hierarchical clusters are optional.

Importantly, assessing differentially altered regions between sample groups might contribute to discover genomic regions associated with annotated variables and thus unveil the biological significance of specific CNAs. To do so, CNApp interrogates descriptive regions associated with any sample-specific annotation variable provided in the input file. Default statistical significance is set to p-value lower than 0.1. However, p-value thresholds can be defined by the user and adjusted p-value is optional. A heatmap plot allows the visualization and interpretation of which genomic regions are differentially altered between sample groups. By selecting a region of interest, box plots and stacked bar plots are generated comparing *seg.mean* values and alteration counts in Student's t-test and Fisher's test tabs, respectively. Additionally, genes comprised in the selected genomic region are indicated.

3. Classifier model: Machine learning classification model predictions

This section allows the user to generate machine learning-based classifier models by choosing a variable to define sample groups and one or multiple classifier variables. To do so, CNApp incorporates the *randomForest* R package (Liaw & Wiener, 2002). The model construction is performed 50-times and bootstrap set is changed in each iteration. By default, annotation variables from the input file are loaded and can be used either by defining sample groups or as a classifier. If *Re-Seg & Score* and/or *Region profile* sections have been previously completed, the user can upload data from these sections (i.e., CNA scores and genomic regions). Predictions for the model performance are generated and the global accuracy is computed along with sensitivity and specificity by group. Classifier models can be useful to point out candidate clinical or molecular variables to classify sample subgroups. A summary of the data distribution and plots for real and model-predicted groups are visualized. A table with prediction rates throughout the 50-times iteration model and real tags by sample is displayed and can be downloaded.

Characterization of cancer types based on CNA scores

First, we evaluated the capacity of CNApp to analyze and classify cancer types according to CNA scores, and assessed whether CNApp was able to reproduce specific CNA patterns across different cancer types. To do so, by using CNApp default parameters we obtained re-segmented data, CNA scores and cancer-specific CNA profiles for 10,635 tumor samples spanning 33 cancer types from the TCGA pan-cancer dataset. The distribution of BCS, FCS and GCS confirmed the existence of distinct CNA burdens across cancer types (Figure 2A). While cancer types such as acute myeloid leukemia (LAML), thyroid carcinoma (THCA) or thymoma (THYM) showed low levels of broad and focal events (GCS median values of -1.67 for LAML, -1.68 for THCA, and -1.52 for THYM), uterine carcinosarcoma (UCS), ovarian cancer (OV) and lung squamous cell carcinoma (LUSC) displayed high levels of both types of genomic imbalances (GCS median values of 2.55, 2.44, and 0.97 for UCS, OV, and LUSC, respectively). Some cancer types displayed a preference for either broad or focal CNAs. For example, kidney chromophobe (KICH) tumors showed the highest levels of broad events (median BCS value of 27), while focal CNAs in this cancer type were very low (median FCS value of 49). In contrast, breast cancer (BRCA) samples displayed high FCS values (median FCS value of 150), while BCS values were only intermediate (median BCS value of 7). Overall correlations between CNA scores were assessed by

computing Spearman's rank test, obtaining values of 0.59 between BCS and FCS, 0.90 between BCS and GCS, and 0.85 between FCS and GCS. In addition, we further assessed the correlation between BCS and FCS for each individual BCS value. While tumors with low BCS displayed a positive correlation between broad and focal alterations, tumors did not maintain such correlation in higher BCS values (Supplementary Figure S2A and B). This correlation between BCS and FCS is maintained across the 33 cancer types (Supplementary Figure S2C).

Subsequent analysis aimed at generating genome-wide patterns for each cancer type based on chromosome-arm genomic windows and the overall corresponding frequencies. In agreement with previous studies (Beroukhim et al., 2010), cancer type-specific patterns of genomic gains and losses determined the tissue-of-origin (Figure 2B). Additionally, we found that chromosome arms altered in more than 25% across all samples were 1q, 7p, 7q, 8q and 20q for copy number gains, and 8p and 17p for copy number losses. Conversely, chromosome arms affected by CNAs in less than 10% of all cancer types included chromosome arms 2q and 19p (Figure 2C). By using a subset of 20 out of the 33 cancer types for which tumor type information was available, we asked CNApp to compute the average arm-region for each cancer type to assess if they clustered according to their CNA profiles (Supplementary Figure S3A). Our analysis showed that correlation values resulting from Pearson's test hierarchically clustered according to the tissue-of-origin from the tumor. Gastrointestinal (colon, rectum, stomach and pancreatic), gynecological (ovarian and uterine) and squamous (cervical, head and neck, and lung) cancers clustered together based on specific CNA profiles for each group (Figure 2D). Intriguingly, correlation profiles using 5 Mb windows and only considering focal alterations showed a very similar degree of clustering based on the tissue of origin (Supplementary Figure S3B and C).

Identification of recurrent CNAs in hepatocellular carcinoma

Next, we attempted to test the ability of CNApp to identify recurrent broad and focal CNAs in a large cohort, and to assess the impact of the customizable parameters to describe CNA profiles. For that reason, we chose to perform CNA analysis of 370 samples from TCGA corresponding to the Liver Hepatocellular Carcinoma (LIHC) cohort. The pattern of recurrent broad and focal CNAs identified by GISTIC2.0 in the TCGA study (Ally et al., 2017) was similar to earlier reports, confirming the suitability of this cohort and the consistent identification of a CNA profile for hepatocellular

carcinoma (HCC) (Chiang et al., 2008; Guichard et al., 2012; Schulze et al., 2015; Totoki et al., 2014; Wang et al., 2013).

By applying the default parameters of CNApp to the LIHC dataset and selecting chromosome arms as genomic regions to assess broad events, we consistently found copy number gains at 1q (56%) and 8q (46%), and copy number losses at 8p (62%) and 17p (47%) as the most frequent alterations (Figure 3A). These frequencies were slightly lower as compared to those identified by GISTIC2.0 (Supplementary Table S1). Similarly, GISTIC2.0 detected significant gains with rates between 25-40% on eight additional chromosome-arms, including 5p, 5q, 6p, 20p, 20q, 7p, 7q, and 17q, which were identified by CNApp in 20-30% of the samples. Likewise, GISTIC2.0 detected significant broad deletions at a frequency between 20-40% on 18 additional chromosome-arms, of which 4q, 6q, 9p, 13q, 16p, and 16q losses were observed at ≥20% by CNApp, and the rest of them displayed rates between 10-20%. In this case, discrepancies in CNA frequencies were expected considering the lower copy number amplitude thresholds used by GISTIC2.0 in comparison with the CNApp default cutoffs (|0.1| vs |0.2|, corresponding to ~2.14/1.8 copies vs 2.3/1.7 copies, respectively). Indeed, previous reports analyzing CNAs in other HCC cohorts and using greater copy number thresholds, showed frequencies of alterations similar to those estimated by CNApp (Chiang et al., 2008; Guichard et al., 2012; Schulze et al., 2015; Wang et al., 2013).

To assess the impact of customizing the amplitude thresholds of CNApp, we next re-run the software dropping the minimum copy number values to |0.1|. As expected, the overall number of broad alterations increased, reaching frequency values similar or even higher than those reported by GISTIC2.0 (Figure 3B and Supplementary Table S1). Of note, such drop from |0.2| to |0.1| might facilitate the identification of subclonal genomic imbalances, which are very frequent in tumor samples (McGranahan & Swanton, 2017), and it will also be of utility to compensate for low tumor purities, if these are unavailable. Furthermore, we assessed whether the identification of broad events was affected by two additional parameters: (i) the relative length to classify a segment as *arm-level* alteration, and (ii) the re-segmentation provided by CNApp. As expected, increasing the percentage of chromosome arm required to classify a CNA segment as *arm-level* (from ≥ 50% to ≥ 70%) or skipping the re-segmentation step led to an underestimation of some broad events, whereas decreasing the percentage of chromosome arm (from ≥50% to ≥40%) resulted in the opposite (Supplementary Figure S4A-C and Supplementary Table S1).

As far as focal CNAs are concerned, CNApp and GISTIC2.0 use different strategies to quantify their recurrence. Therefore, the comparison between the two methods was evaluated in a more indirect manner. GISTIC2.0 generates minimal common regions (also known as 'peaks') that are likely to be altered at high frequencies in the cohort, which are scored using a Q-value and may present a wide variety of genomic lengths (Mermel et al., 2011). Instead, CNApp allows dividing the genome in windows of different sizes, calculating average copy number amplitudes for all segments included within each window. We reasoned that considering the length of GISTIC2.0 reported 'peaks', CNApp might also be capable of identifying recurrent focal altered regions by dividing the genome in smaller windows. To test our hypothesis, we asked CNApp to calculate the frequency of focal gains and losses by dividing the genome by sub-cytobands. As a result, CNApp consistently localized the most frequently altered sub-cytobands, including gains at 1q21.3 (25%), 8q24.21 (17%, *MYC*), 5p15.33 (13%, *TERT*), 11q13.3 (12%, *CCND1*/*FGF19*) and 6p21.1 (11%, *VEGFA*), and losses at 13q14.2 (20%, *RB1*), 1p36.11 (18%, *ARID1A*), 4q35.1 (17%, *IRF2*) and 9p21.3 (14%, *CDKN2A*), which are in agreement with previous studies in HCC (Figure 3C and Supplementary Table S2) (Chiang et al., 2008; Guichard et al., 2012; Schulze et al., 2015; Wang et al., 2013). Compared to GISTIC2.0, CNApp reported 14 of the 27 significant amplifications and 14 of the 34 significant deletions at rates >10%, and the remaining alterations displaying rates between 4-10% (Supplementary Table S3) (Wang et al., 2013). Most importantly, regions with the highest frequency detected by CNApp showed a good match with lowest GISTIC2.0 Q-residual values, indicating that the most significant 'peaks' identified by GISTIC2.0 were actually included in the most recurrently altered sub-cytobands reported by CNApp.

As previously suggested, recurrent focal alterations often occur at lower frequencies than broad events (Beroukhim et al., 2010). In our analysis, excluding the low-level alterations and evaluating only the moderate and high-amplitude events (≥3 and ≤1 copies), amplifications reached maximum rates of 11%, whereas high-level losses only reached ~2% (Figure 3D and Supplementary Table S2). Top recurrent focal gains involved sub-cytobands 1q21.3 (11%), 8q24.21 (11%, *MYC*), 11q13.3 (7%, *CCND1*/*FGF19*), and 5p15.33 (5%, *TERT*). Recurrent losses estimated at ~2% of the samples included 13q14.2 (*RB1*), 9p21.3 (*CDKN2A*), 4q35.1 (*IRF2*), and 8p23.1. Although slight discrepancies between frequencies might be explained by minimal

variability in the copy number threshold, CNApp results are in high consistence with previous reports (Chiang et al., 2008; Guichard et al., 2012; Schulze et al., 2015).

Classification of colon cancer according to CNA scores and genomic regions

A proposed taxonomy of colorectal cancer (CRC) includes four consensus molecular subtypes (CMS), mainly based on differences in gene expression signatures (Guinney et al., 2015). Briefly, CMS1 includes the majority of hypermutated tumors showing microsatellite instability (MSI), high CpG island methylator phenotype (CIMP), and low levels of CNAs; CMS2 and CMS4 typically comprise microsatellite stable (MSS) tumors with high levels of CNAs; and finally, mixed MSI status and low levels of CNAs and CIMP are associated with CMS3 tumors. A representative cohort of 309 colon cancers from the TCGA Colon Adenocarcinoma (COAD) cohort (Cancer & Atlas, 2012) with known CMS classification (CMS1, N = 64; CMS2 N = 112; CMS3 N = 51; CMS4 N = 82) and MSI status (MSI, N = 72; MSS, N = 225) was analyzed by using CNApp. In agreement with Guinney and colleagues, survival curves generated by CNApp indicated that CMS1 patients after relapse showed the worst survival rates as compared to CMS2 patients (Supplementary Figure S5A) (Guinney et al., 2015). Next, we asked CNApp to perform the re-segmentation step using the default copy number thresholds and excluding segments smaller than 500 Kbp to avoid technical background noise. Then, broad CNAs were considered to generate genomic region profiles using chromosome-arm windows. As expected, the CNA frequency plot displayed the most commonly altered genomic regions in sporadic CRC (Supplementary Figure S5B) (Camps et al., 2008; Cancer & Atlas, 2012; Meijer et al., 1998; Nakao et al., 2004; Ried et al., 1996). Most frequently altered chromosome arms included gains of 7p, 7q, 8q, 13q, 20p, and 20q, and losses of 8p, 17p, 18p, and 18q, occurring in more than 30% of the samples (Figure 4A). On the other hand, focal CNA patterns were obtained by generating genomic profiles by sub-cytobands. Of note, five out of six losses and five out of 18 gains were also identified by GISTIC2.0 in the COAD TCGA cohort (Cancer & Atlas, 2012).

Subsequently, we performed integrative analysis of genomic imbalances, CMS groups, and CNA scores. By using CNApp, we assessed whether CNA scores were able to classify colon cancer samples according to their CMS. While BCS established significant differences between CMS paired comparisons ($P \leq 0.0001$, Student's t-test), FCS poorly discerned CMS1 from 3 and CMS2 from 4 (Figure 4B and Supplementary

Figure S5C). Thus, we reasoned that broad CNAs rather than focal were able to better discriminate between different CMS groups. In fact, the distribution of CMS groups based on BCS resembled the distribution of somatic CNA counts defined by GISTIC2.0 (Guinney et al., 2015).

Next, we integrated the BCS and the CMS groups with the microsatellite status. Our results showed an average BCS of 1.51±2.11 and 10.25±5.92 for MSI (N = 72) and MSS (N = 225) tumors, respectively. By using the *Classifier model*, we assessed the discriminative power of BCS to separate MSI and MSS samples. Our results indicated that a BCS equal to 4 predicted the MSI status with a global accuracy of 82.2%, as seen by the intersection value between BCS distribution MSI- and MSS-predicted samples (Supplementary Figure S5D). Applying this cutoff, 186 out of 225 (83%) of MSS tumors showed BCS values greater than 4 (Figure 4C). In contrast, 39 (17%) MSS tumors showed a BCS of 4 or lower, corresponding to three CMS1, six CMS2, 18 CMS3 and 12 CMS4 tumors, further demonstrating the existence of MSS tumors with a very low CNA burden. On the other hand, seven MSI tumors showed BCS higher than 4. Among them, five samples displayed genomic imbalances typically associated with the CRC canonical pathway, including a focal amplification of *MYC*, unveiling tumors with co-occurrence of MSI and extensive genomic alterations (Trautmann et al., 2006). Our dataset comprised nine out of 51 CMS3 tumors with MSI. Intriguingly, two of them showed focal deletions on chromosome 2 involving *MSH2* and *MSH6*, suggesting the inactivation of these mismatch repair genes through a focal genomic imbalance. In fact, 46% of CMS3 MSS tumors showed BCS below 4, in agreement with the finding that CMS3 tumors display low levels of somatic CNAs.

Moreover, CNApp enabled the identification of possible sample misclassifications by integrating CMS annotation and *BRAF*-mutated sample status. As expected, CMS1 cases were enriched for *BRAF* mutation, although two CMS4 samples also showed mutations in *BRAF*. One of these samples showed a BCS of 11, displaying canonical CNAs. In contrast, the other CMS4 *BRAF*-mutated sample showed MSI and a BCS of 0, similar features as CMS1. Likewise, four *BRAF*-wt samples, classified within the CMS4 group, displayed MSI and a BCS of 0, thus being candidates to be labeled as CMS1 based on the levels of CNAs (Figure 4D). These disparities are of utmost importance since recent studies reported that high copy number alterations correlate with reduced response to immunotherapy (Davoli et al., 2017). Importantly, it has been suggested that MSI status might be predictive of positive immune checkpoint blockade

response in advanced CRC, probably due to the low levels of CNA usually presented by MSI tumors (Le et al., 2015).

We then asked CNApp to compare differentially represented genomic regions between all CMS groups based on a Student's t-test or Fisher's test with adjusted p-value. By applying a Student's t-test, we observed that CMS1 resembled CMS3, except for the gain of chromosome 7 and the loss of 18q, which were regions commonly altered in CMS3 samples with BCS above 4 (adjusted p-value ≤ 0.001, Student's t-test) (Supplementary Figure S5E). Even though only subtle CNA differences between CMS2 and CMS4 were identified, the loss of 14q was significantly more detected in CMS2 (42%) than in CMS4 (17.1%) (adjusted p-value ≤ 0.005, Student's t-test). The gain of 12q was more frequently associated with CMS1 than CMS2 (adjusted p-value ≤ 0.005, Student's t-test), in agreement with previous studies reporting that the gain of chromosome 12 is associated with microsatellite unstable tumors (Supplementary Figure S5E) (Trautmann et al., 2006). Intriguingly, the gain of the chromosome arm 20q alone mimicked the distribution of somatic CNAs defined by GISTIC2.0 across consensus subtype samples (Figure 4E) (Guinney et al., 2015).

Finally, applying machine learning-based prediction models to classify samples by the most discriminative descriptive regions across CMS groups (i.e., 13q, 17p, 18, and 20q), CNApp reached 55% of accuracy to correctly predict CMS. In fact, the occurrence of these genomic alterations was able to differentiate CMS2 from CMS4 with an accuracy of 70%, and CMS1 from CMS3 with a 72.3% accuracy. As expected, this set of genomic alterations distinguished CMS1 from CMS2 samples with an accuracy of 95%.

DISCUSSION

Here we present CNApp, a web-based computational tool that provides a unique framework to comprehensively analyze and integrate CNAs associated with molecular and clinical variables, assisting data-driven research in the biomedical context. Although CNApp has been developed using segmented genomic copy number data obtained from SNP-arrays, the software is also able to accommodate segmented data from any next-generation sequencing platform.

CNApp transforms segmented data into genomic profiles, allowing sample-by-sample comparison and the assessment of differentially altered genomic regions, which can then be selected by the user to assess classifier variables by computing machine

learning-based models. Importantly, besides identifying the impact of specific CNAs, CNApp provides the unique opportunity to establish associations between the burden of genomic alterations and any clinical or molecular variable. To do so, CNApp calculates CNA scores, a quantification of the broad (BCS), focal (FCS) and global (GCS) levels of genomic imbalances for each individual sample. The fact that high levels of aneuploidy may correlate with immune evasion markers in cancer exemplifies the potential association between CNA scores and clinical features (Buccitelli et al., 2017; Davoli et al., 2017; Taylor et al., 2018b). To note, CNA scores are calculated after an optional process of re-segmentation that enables to redefine CNA boundaries and to adjust sample-specific copy number thresholds by correcting for tumor purity estimates.

In agreement with recently reported findings (Beroukhim et al., 2010; Hoadley et al., 2018; Taylor et al., 2018b), CNApp was benchmarked by analyzing 10,635 samples spanning 33 cancer types from the TCGA pan-cancer dataset, and was able to cluster major tumor types according to CNA patterns. Moreover, the software successfully reproduced the well-characterized genomic profile of HCC and CRC, considering both broad and focal events, demonstrating the reliability of CNApp in identifying regions encompassing the most recurrent CNAs (Ally et al., 2017; Cancer & Atlas, 2012).

Finally, applying CNApp to the TCGA colon cancer sample set, for which MSI status and CMS classification was well annotated, we determined that a BCS value of 4 discriminates MSI from MSS tumors with high accuracy, reinforcing the utmost significance of quantifying the CNA burdens. Most importantly, due to the inverse correlation between MSI and aneuploidy in CRC, our results suggest that this BCS value could be established as a cutoff to define the edge between low and high aneuploid tumors. In fact, while high aneuploid tumors show poor response to immunotherapy, it has been suggested that CMS1 microsatellite unstable tumors are likely to show a positive response to immune checkpoints inhibitors (Kalyan, Kircher, Shah, Mulcahy, & Benson, 2018; Le et al., 2015). However, BCS was not associated with overall survival in patients after relapse (data not shown). Moreover, specific genomic regions defined by CNApp contributed to classify the CMS groups, confirming the functional importance of specific genomic imbalances in the pathogenesis of this disease and providing insights into the classification of CRC based on CNA profiles.

In summary, although our results ought to be further validated in independent cohorts, here we show that CNApp enables not only the fundamental analysis of CNA profiles, but also the functional understanding of CNAs in the context of clinical

outcome and their potential use as biomarkers, thus becoming an asset to the cancer genomics community.

MATERIALS AND METHODS

Data set availability

Pan-cancer cohort and clinical annotation

Affymetrix SNP6.0 array copy number segmented data (Level 3) from 10,635 samples spanning 33 cancer types from TCGA pan-cancer dataset were downloaded from Genomic Data Commons (National Cancer Institute, NIH) (Grossman et al., 2016). This dataset included the 370 Liver Cancer-Hepatocellular Carcinoma (LIHC) samples used for the analysis of recurrent CNAs and the subset of 309 samples from Colon Adenocarcinoma (COAD) for which the colorectal cancer consensus molecular subtype (CMS) was known (Guinney et al., 2015).

Clinical annotation for the 309 COAD samples was retrieved by using *TCGAbiolinks* R package in order to extract survival information for each sample (Colaprico et al., 2016).

GISTIC data from TCGA: LIHC cohort

GISTIC 2.0.22 (Ally et al., 2017) copy number results (Level 4) of the 370 LIHC samples, were downloaded from the Broad Institute GDAC Firehose. Parameters used for the analysis are detailed in the same GDAC repository. Specifically, parameters conditioning the definition of the CNAs and of interest for our comparison were publicly reported with the following values: *amplification* and *deletion thresholds*: 0.1; *broad length cutoff*: 0.7; *joint segment size*: 4.

Software and tool availability

CNApp can be accessed at http://cnapp.bsc.es. It was developed using Shiny R package (version 1.1.0), from R-Studio (Chang, Cheng, Allaire, Xie, & McPherson, 2018). The tool was applied and benchmarked while using R version 3.4.2 (2017-09-28) -- "Short Summer". List of packages, libraries and base coded are freely available at GitHub, and instructions for local installation are also specified.

CNA scores computation

Segments resulting from re-segmentation (or original segments from input file when re-segmentation is skipped) are classified in *chromosomal*, *arm-level* and *focal* events by

considering the relative length of each segment to the whole-chromosome or chromosome arm. Using default parameters, segments are tagged as *chromosomal* when 90% or more of the chromosome is affected; as *arm-level* when 50% or more of the chromosome arm is affected; and as *focal* when affecting less than 50% of the chromosome arm. Percentages for relative lengths are customizable. Broad (chromosomal and arm-level) and focal alterations are then weighted according to their amplitude values (*seg.mean*) and taking into account copy number amplitude ranges defined by CNA calling thresholds (Supplementary Methods).

Broad CNA Score (BCS): for a total *N* of broad events in a sample (*x*), it equals to the summation of segments weights (*A*) in that corresponding sample and being *i* the corresponding segment:

$$BCS(x) = \sum_{i=1}^{N} A_i$$

Focal CNA Score (FCS): same as in BCS, with an additional pondering value *L* included to the summation, which captures the relative size of the chromosome-arm coverage of each focal CNA (according to weights specified in Supplementary_Methods):

$$FCS(x) = \sum_{i=1}^{N} A_i \cdot L_i$$

Global CNA Score (GCS): for a sample *x*, it is calculated as the summation of normalized BCS and FCS values, where *meanBCS* and *meanFCS* stand for mean values of BCS and FCS from total samples, respectively, and *sdBCS* and *sdFCS* stand for standard deviation values of BCS and FCS from total samples, respectively:

$$normBCS(x) = \frac{BCS(x) - meanBCS}{sdBCS} \qquad normFCS(x) = \frac{FCS(x) - meanFCS}{sdFCS}$$

$$GCS(x) = normBCS(x) + normFCS(x)$$

Genomic region profiles computation

Region profiling section allows genome segmentation analysis by user-selected windows (i.e. arms, half-arms, cytobands, sub-cytobands, and 40Mb till 1Mb). In order to do that, windows files were generated for each option and genome build (*hg19* and

hg38). Cytobands file *cytoBand.txt* from UCSC page and for both genome builds was used as mold to compute regions (Casper et al., 2017).

Segmented samples are transformed into genome region profiles using genomic windows selected by user. Segments from each sample are consulted to assess whether or not overlap with the window region. Thus, window-means (*W*) are computed for each genomic window by collecting segments (*t*) overlapping with window-region (*i*). Segments with *loc.start* or *loc.end* position falling within the region are collected, as well as those segments embedding the entire region. At this point, the summation of each segment-mean (*S*) corrected by the relative window-length (*L*) affected by the segment length (*l*) is performed:

$$W(i) = \sum_{t=1}^{n} S_t \cdot \frac{l_t}{L(i)}$$

Descriptive regions assessment

Potential descriptive regions between groups defined by the annotated variables provided in the input file can be studied and *P*-values are presented to evaluate significance in differentially altered regions between those groups. The alterations can be considered as (1) numerical continuous (*seg.mean* values) and (2) categorical variables (gains, losses and non-altered). In the first case, to assess statistical significance between groups Student's T-test is applied, whereas in the second situation the significance is assessed by applying the Fisher's exact test. False discovery rate (FDR) adjustment is performed using the Benjamini-Hochberg (BH) procedure in both cases and corrected *P*-values (*Adj.p-value*) or non-corrected *P*-values (*p-values*) are displayed by user selection.

Machine learning-based classifier models

We used the *randomForest* R package (Liaw & Wiener, 2002) to compute machine learning classifier models. Variables to define sample groups must be selected, as well as at least one classifier variable. Model construction is performed 50-times and training set is changed by iteration. In order to compute model and select training set, multiple steps and conditions have to be accomplished:

i. total *N* samples divided by *G* groups depicted by group-defining variable must be higher than n samples from the smaller group:

$$P = \frac{N}{G} \;\; ; \;\; P > n$$

ii. If condition above is not accomplished, then *P* is set to 75% of n:

$$\text{if } \; P \leq n \;\; \text{then} \;\; P = n \cdot 0.75$$

iii. *P* term must be higher than one, and *N* must be equal or higher than 20:

$$P > 1 \;\; \text{or} \;\; N \geq 20$$

iv. Classifier variables, when categorical, shall not have higher number of tags (*Z*) than groups defined (*G*) by group-defining variable:

$$Z < G$$

v. Training set (*T*) is computed and merged for each group (*g*) from groups (*G*) defined by group variable, extracting *P* samples from *g* as follows:

$$t\,(g) = P \text{ samples from } g \qquad\qquad T = \sum_{i=1}^{g} t_i$$

After model computation, contingency matrix with prediction and reference values by group is created to compute accuracy, specificity and sensitivity by group.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Rodrigo Dienstmann from Vall d'Hebron Institute of Oncology, Barcelona, Spain for providing the CMS and clinical information for the subset of samples included in the COAD cohort from TCGA. The work was carried out at the Esther Koplowitz Centre, Barcelona. We thank the Spanish National Bioinformatics Institute (INB) (PT17/0009/0001) and the Barcelona Supercomputing Center (BSC) for support".

Funding

This work has been supported by the European Commission [PCIG11-GA-2012-321937]; the Instituto de Salud Carlos III and co-funded by the European Regional

Development Fund (ERDF) [CP13/00160, PI14/00783, PI17/01304, PI17/00878]; the CIBEREHD program; the CERCA Program (Generalitat de Catalunya); the Agència de Gestió d'Ajuts Universitaris i de Recerca, Generalitat de Catalunya [2017 SGR 1035, 2017 SGR 21, 2017 SGR 653]; PERIS Generalitat de Catalunya [SLT002/16/00398]; Fundación Científica de la Asociación Española Contra el Cáncer [GCB13131592CAST]; CIBEREHD contract to SF-E; Agència de Gestió d'Ajuts Universitaris i de Recerca -AGAUR- from Generalitat de Catalunya [2016BP00161] to LB and [2018FI B1_00213] to MD-G; Spanish National Health Institute FPI grant [BES-2017-081286] to RE-F. CIBEREHD is funded by the Instituto de Salud Carlos III. This article is based upon work from COST Action [CA17118], supported by COST (European Cooperation in Science and Technology). JML is supported by the European Commission (EC)/Horizon 2020 Program (HEPCAR, Ref. 667273-2), U.S. Department of Defense (CA150272P3), National Cancer Institute (P30-CA196521), Samuel Waxman Cancer Research Foundation, Spanish National Health Institute (SAF2016-76390) and Generalitat de Catalunya/AGAUR (SGR-1162 and SGR-1358).

COMPETING INTERESTS

Dr. Llovet is receiving research support from Bayer HealthCare Pharmaceuticals, Eisai Inc, Bristol-Myers Squibb and Ipsen, and consulting fees from Eli Lilly, Bayer HealthCare Pharmaceuticals, Bristol-Myers Squibb, EISAI Inc, Celsion Corporation, Exelixis, Merck, Ipsen, Glycotest, Navigant, Leerink Swann LLC, Midatech Ltd, and Nucleix.

Ethics approval and consent to participate

Ethics approval was not required for this study.

AUTHORS' CONTRIBUTIONS

SF-E, LB and JC designed the study and acquired and analyzed the data. SF-E, LB, MV-C and MD-G designed, generated and implemented the package, and analyzed the data. EH-I and RE-F tested the software. JJL supervised the software implementation. AC, and JMLl critically reviewed the software implementation and the data output. SF-E, LB, SC-B and JC wrote the manuscript. All authors read and approved the final manuscript.

REFERENCES

Ally, A., Balasundaram, M., Carlsen, R., Chuah, E., Clarke, A., Dhalla, N., … Laird, P. W. (2017). Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*, *169*(7), 1327-1341.e23. https://doi.org/10.1016/j.cell.2017.05.046

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., … Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, *463*(7283), 899–905. https://doi.org/10.1038/nature08822

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., … Barillot, E. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)*, *28*(3), 423–425. https://doi.org/10.1093/bioinformatics/btr670

Buccitelli, C., Salgueiro, L., Rowald, K., Sotillo, R., Mardin, B. R., & Korbel, J. O. (2017). Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation. *Genome Research*, *27*(4), 501–511. https://doi.org/10.1101/gr.212225.116

Cairncross, G., Wang, M., Shaw, E., Jenkins, R., Brachman, D., Buckner, J., … Mehta, M. (2013). Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: Long-term results of RTOG 9402. *Journal of Clinical Oncology*, *31*(3), 337–343. https://doi.org/10.1200/JCO.2012.43.2674

Camps, J., Grade, M., Nguyen, Q. T., Hormann, P., Becker, S., Hummon, A. B., … Ried, T. (2008). Chromosomal Breakpoints in Primary Colon Cancer Cluster at Sites of Structural Variants in the Genome. *Cancer Research*, *68*(5), 1284–1295. https://doi.org/10.1158/0008-5472.CAN-07-2864

Cancer, T., & Atlas, G. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337. https://doi.org/10.1038/nature11252

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., … Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, *30*(5), 413–421. https://doi.org/10.1038/nbt.2203

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., … Kent, W. J. (2017). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, *46*(D1), D762–D769. https://doi.org/10.1093/nar/gkx1020

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018). *shiny: Web*

Application Framework for R (p. R package version 1.1.0). p. R package version 1.1.0.

Chiang, D. Y., Villanueva, A., Hoshida, Y., Peix, J., Newell, P., Minguez, B., … Llovet, J. M. (2008). Focal Gains of VEGFA and Molecular Classification of Hepatocellular Carcinoma. *Cancer Research*, *68*(16), 6779–6788. https://doi.org/10.1158/0008-5472.CAN-08-0742

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., … Noushmehr, H. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, *44*(8), e71–e71. https://doi.org/10.1093/nar/gkv1507

Davoli, T., Uno, H., Wooten, E. C., & Elledge, S. J. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, *355*(6322), eaaf8399. https://doi.org/10.1126/science.aaf8399

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, *375*(12), 1109–1112. https://doi.org/10.1056/NEJMp1607591

Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I. Ben, … Zucman-Rossi, J. (2012). Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nature Genetics*, *44*(6), 694–698. https://doi.org/10.1038/ng.2256

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., … Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, *21*(October), 1350–1356. https://doi.org/10.1038/nm.3967

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., … Mariamidze, A. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, *173*(2), 291-304.e6. https://doi.org/10.1016/j.cell.2018.03.022

Kalyan, A., Kircher, S., Shah, H., Mulcahy, M., & Benson, A. (2018). Updates on immunotherapy for colorectal cancer. *Journal of Gastrointestinal Oncology*, *9*(1), 160–169. https://doi.org/10.21037/jgo.2018.01.17

Krijgsman, O., Carvalho, B., Meijer, G. A., Steenbergen, R. D. M., & Ylstra, B. (2014). Focal chromosomal copy number aberrations in cancer-Needles in a genome

haystack. *Biochimica et Biophysica Acta - Molecular Cell Research*, *1843*(11), 2698–2704. https://doi.org/10.1016/j.bbamcr.2014.08.001

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., … Diaz, L. A. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine*, *372*(26), 2509–2520. https://doi.org/10.1056/NEJMoa1500596

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

McGranahan, N., & Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, *168*(4), 613–628. https://doi.org/10.1016/j.cell.2017.01.018

Meijer, G. A., Hermsen, M. A., Baak, J. P., van Diest, P. J., Meuwissen, S. G., Belien, J. A., … Walboomers, J. M. (1998). Progression from colorectal adenoma to carcinoma is associated with non- random chromosomal gains as detected by comparative genomic hybridisation. *Journal of Clinical Pathology*, *51*(12), 901–909. https://doi.org/10.1136/jcp.51.12.901

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, *12*(4), R41. https://doi.org/10.1186/gb-2011-12-4-r41

Nakao, K., Mehta, K. R., Fridlyand, J., Moore, D. H., Jain, A. N., Lafuente, A., … Waldman, F. M. (2004). High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*. https://doi.org/10.1093/carcin/bgh134

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, *5*(4), 557–572. https://doi.org/10.1093/biostatistics/kxh008

Popova, T., Manié, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E., & Stern, M. H. (2009). Genome Alteration Print (GAP): A tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biology*, *10*(11), R128. https://doi.org/10.1186/gb-2009-10-11-r128

Ried, T., Hu, Y., Difilippantonio, M. J., Ghadimi, B. M., Grade, M., & Camps, J. (2012). The consequences of chromosomal aneuploidy on the transcriptome of cancer cells. *Biochimica et Biophysica Acta*, *1819*(7), 784–793.

https://doi.org/10.1016/j.bbagrm.2012.02.020

Ried, T., Knutzen, R., Steinbeck, R., Blegen, H., Schröck, E., Heselmeyer, K., … Auer, G. (1996). Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes, Chromosomes and Cancer*, *15*(4), 234–245. https://doi.org/10.1002/(SICI)1098-2264(199604)15:4<234::AID-GCC5>3.0.CO;2-2

Sansregret, L., Vanhaesebroeck, B., & Swanton, C. (2018). Determinants and clinical implications of chromosomal instability in cancer. *Nature Reviews Clinical Oncology*, *15*(3), 139–150. https://doi.org/10.1038/nrclinonc.2017.198

Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S., … Nelson, S. F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, *27*(19), 2648–2654. https://doi.org/10.1093/bioinformatics/btr462

Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., … Zucman-Rossi, J. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nature Genetics*, *47*(5), 505–511. https://doi.org/10.1038/ng.3252

Taylor, A. M., Shih, J., Ha, G., Gao, G. F., Zhang, X., Berger, A. C., … Meyerson, M. (2018a). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell*, *33*(4), 676-689.e3. https://doi.org/10.1016/J.CCELL.2018.03.007

Taylor, A. M., Shih, J., Ha, G., Gao, G. F., Zhang, X., Berger, A. C., … Meyerson, M. (2018b). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell*, *33*(4), 676-689.e3. https://doi.org/10.1016/J.CCELL.2018.03.007

Totoki, Y., Tatsuno, K., Covington, K. R., Ueda, H., Creighton, C. J., Kato, M., … Shibata, T. (2014). Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nature Genetics*, *46*(12), 1267–1273. https://doi.org/10.1038/ng.3126

Trautmann, K., Terdiman, J. P., French, A. J., Roydasgupta, R., Sein, N., Kakar, S., … Waldman, F. M. (2006). Chromosomal instability in microsatellite-unstable and stable colon cancer. *Clinical Cancer Research*, *12*(21), 6379–6385. https://doi.org/10.1158/1078-0432.CCR-06-1248

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., …

Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, *107*(39), 16910–16915. https://doi.org/10.1073/pnas.1009843107

Venkatraman, E. S., & Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, *23*(6), 657–663. https://doi.org/10.1093/bioinformatics/btl646

Wang, K., Lim, H. Y., Shi, S., Lee, J., Deng, S., Xie, T., … Xu, J. (2013). Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. *Hepatology*, *58*(2), 706–717. https://doi.org/10.1002/hep.26402

Zack, T. T. I., Schumacher, S. E. S., Carter, S. L. S., Cherniack, A. D., Saksena, G., Tabak, B., … Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, *45*(10), 1134–1140. https://doi.org/10.1038/ng.2760

Zhang, Z., & Hao, K. (2015). SAAS-CNV: A Joint Segmentation Approach on Aggregated and Allele Specific Signals for the Identification of Somatic Copy Number Alterations with Next-Generation Sequencing Data. *PLoS Computational Biology*, *11*(11), e1004618. https://doi.org/10.1371/journal.pcbi.1004618

FIGURE LEGENDS

Figure 1 CNApp workflow

The diagram depicts the overall processes performed by CNApp and indicates the output for each section.

Figure 2 Analysis of the TCGA pan-cancer dataset and clustering by tumor type

CNApp outputs to characterize pan-cancer 10,635 samples including 33 TCGA cancer types. **A**. Broad, Focal and Global CNA scores (BCS, FCS and GCS, respectively) distribution across the 33 cancer types. **B**. Genome-wide chromosome arm CNA profile heatmap for 10,635 samples considering broad and focal events. Annotation tracks for FCS, BCS and GCS are presented. **C**. Arm regions frequencies as percentages relative to the TCGA pan-cancer dataset (red for gains and blue for losses). **D**. Heatmap plot showing 20 out of the 33 TCGA cancer type profile correlations, by Pearson's method, hierarchically clustered by tissue of origin. Gastrointestinal, gynecological and squamous cancers are clustering consistently in their respective groups.

Figure 3 Identification of recurrent broad and focal CNAs

Calculation of broad and focal CNA frequencies using several parameters in CNApp in order to describe the genomic landscape of LIHC. **A**. CNApp frequencies for chromosome arm regions using default cutoffs, corresponding to 2.3/1.7 copies for gains and losses, respectively. **B**. CNApp frequencies for chromosome arm regions relaxing cutoffs to make them equivalent to those of GISTIC2.0. **C**. CNApp frequencies of focal events using default thresholds and sub-cytobands genomic regions. **D**. Frequencies of focal events from moderate- to high-amplitude levels using sub-cytobands genomic regions.

Figure 4 Genomic characterization of colon cancer according to the CMS classification

A. Arm-region frequencies of 309 colon cancer samples using CNApp default thresholds for CNAs. **B**. BCS distribution by CMS sample groups. Wilcox test significance is shown as p-value ≤ 0.001 (***); p-value ≤ 0.01 (**); p-value ≤ 0.05 (*); p-value > 0.05 (ns). **C**. Number of gained and lost chromosome arms for each sample distributed according to the BCS values. Note that a cutoff at 4 is indicated with a black line. Annotation tracks for microsatellite instability (msi), BRAF mutated samples (braf_mut), CMS groups (cms_label), FCS and BCS are displayed. **D**. Genome-wide profiling by chromosome arms distributed according to the CMS group. Annotation tracks for microsatellite instability (msi), BRAF mutated samples (braf_mut), CMS groups (cms_label), FCS and BCS are displayed. Sample-to-sample correlation heatmap plot by Pearson's method is shown below. **E**. Distribution of CNA values affecting 20q according to the CMS groups. Significance is shown as p-value ≤ 0.001 (***); p-value ≤ 0.01 (**); p-value ≤ 0.05 (*); p-value > 0.05 (ns).

Figure 1

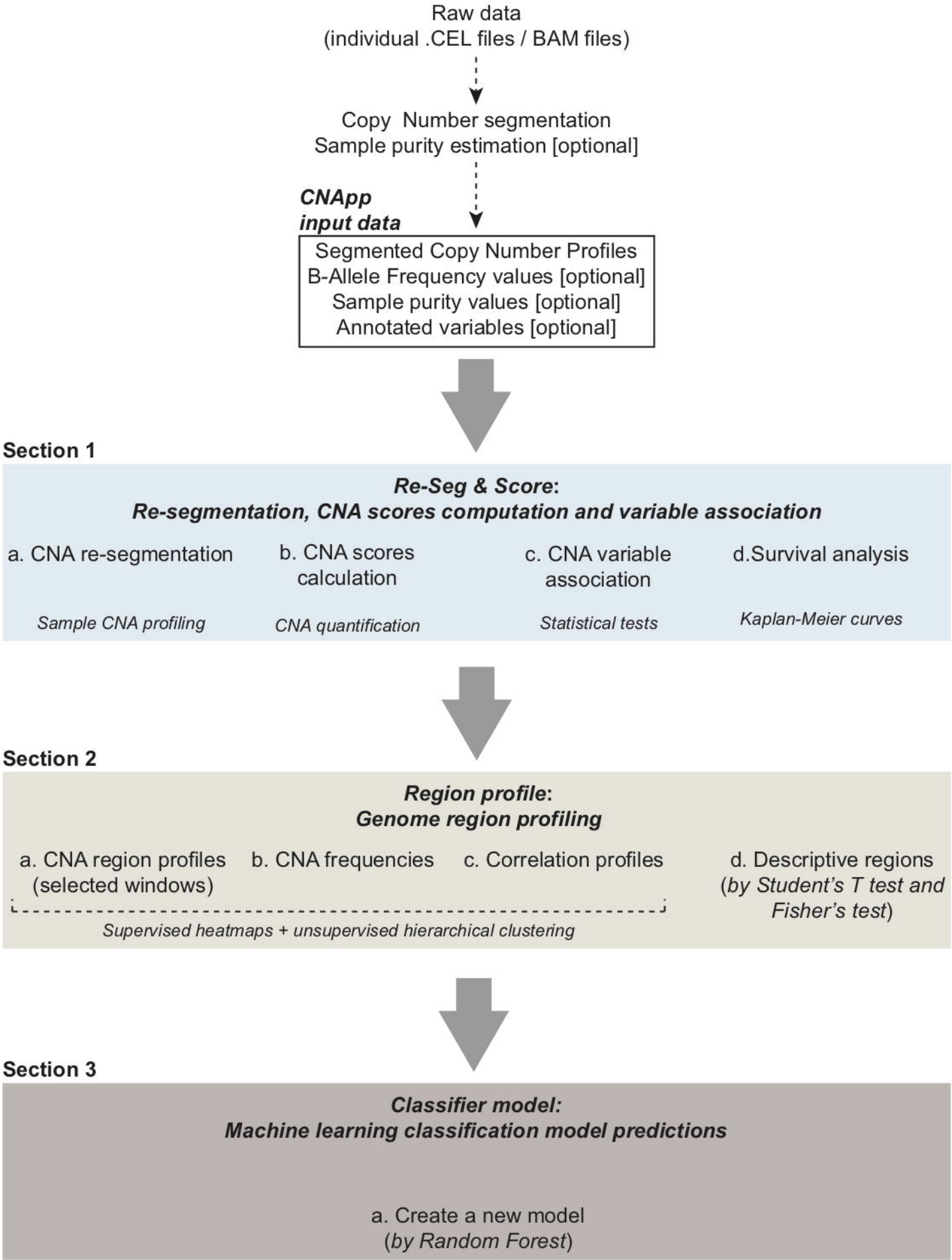
Raw data
(individual .CEL files / BAM files)
Copy Number segmentation
Sample purity estimation [optional]
CNApp
input data
Segmented Copy Number Profiles
B-Allele Frequency values [optional]
Sample purity values [optional]
Annotated variables [optional]
Section 1
Re-Seg & Score:
Re-segmentation, CNA scores computation and variable association
a. CNA re-segmentation
b. CNA scores calculation
c. CNA variable association
d.Survival analysis
Sample CNA profiling
CNA quantification
Statistical tests
Kaplan-Meier curves
Section 2
Region profile:
Genome region profiling
a. CNA region profiles (selected windows)
b. CNA frequencies
c. Correlation profiles
d. Descriptive regions (by Student's T test and Fisher's test)
Supervised heatmaps + unsupervised hierarchical clustering
Section 3
Classifier model:
Machine learning classification model predictions
a. Create a new model
(by Random Forest)

Figure 2

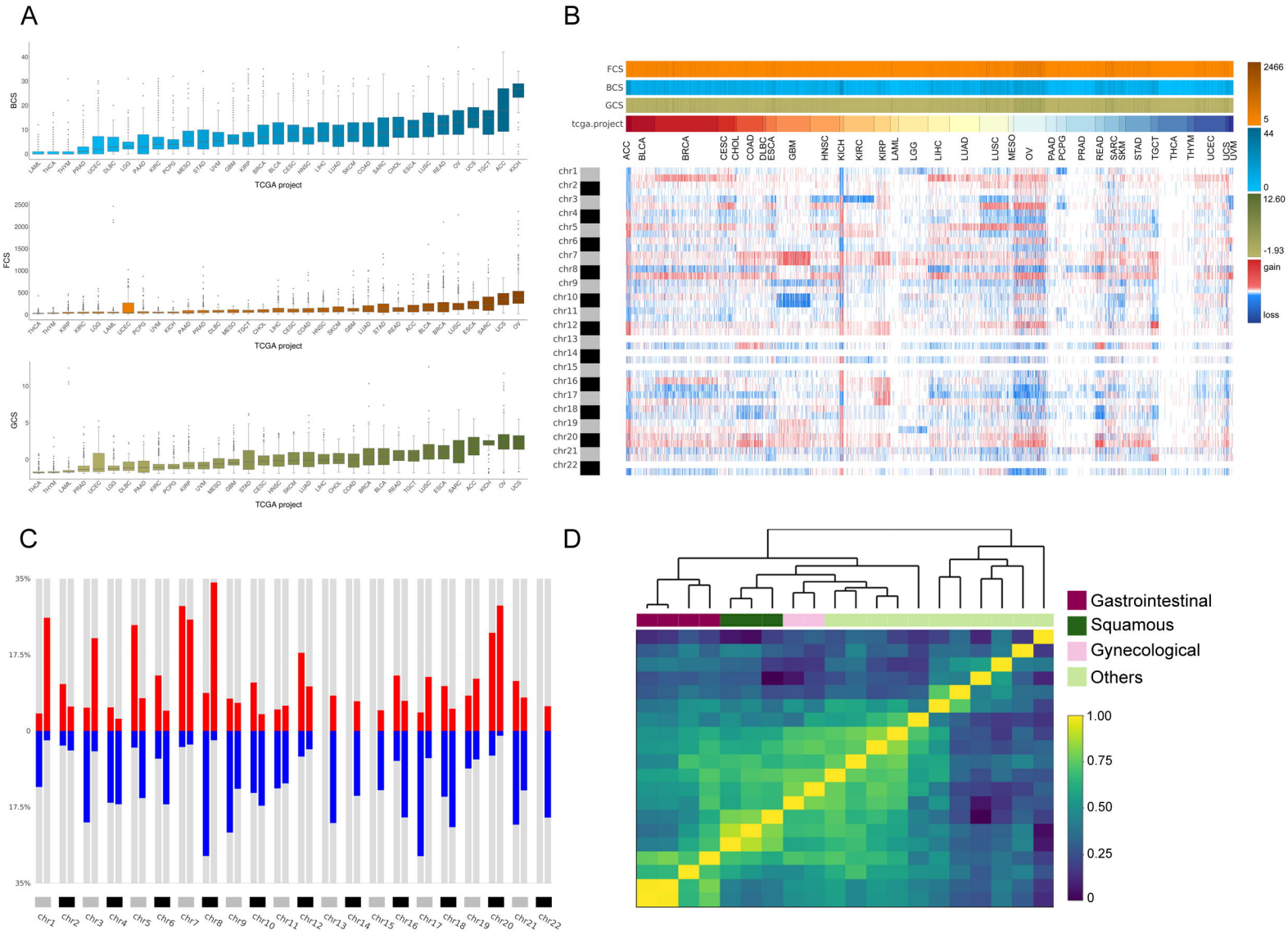

Figure 3

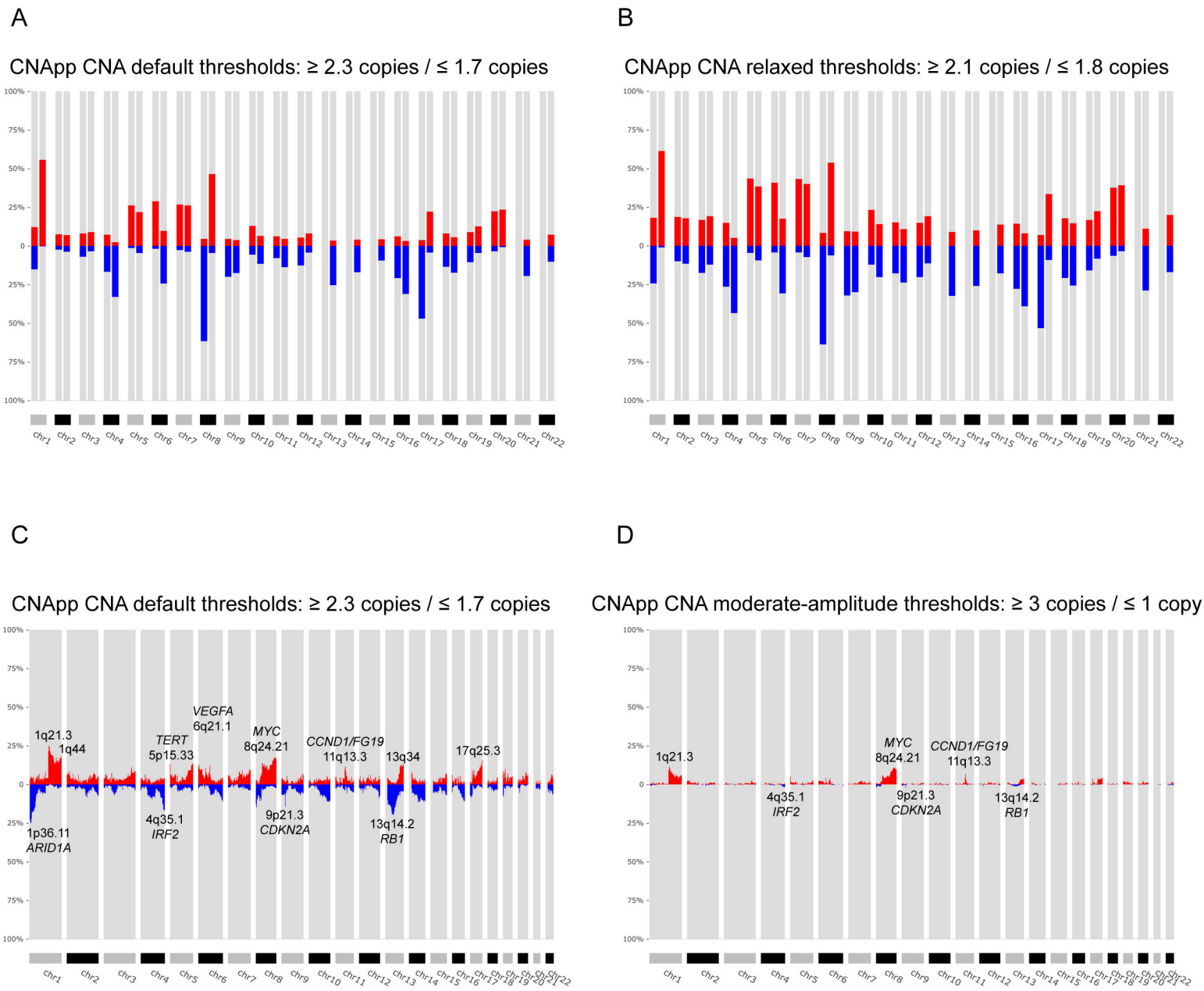

Figure 4