OXFORD

Briefings in Bioinformatics, 2022, **23(5)**, 1–9
https://doi.org/10.1093/bib/bbac161
Case Study

Refinement of computational identification of somatic copy number alterations using DNA methylation microarrays illustrated in cancers of unknown primary

Pedro Blecua†, Veronica Davalos†, Izar de Villasante, Angelika Merkel, Eva Musulen, Laia Coll-SanMartin and Manel Esteller

Corresponding author: Manel Esteller, Josep Carreras Leukaemia Research Institute (IJC), Badalona, Barcelona, Catalonia 08916, Spain. Tel.: +34-93-557-28-00; Fax: +34-93-282-84-51; E-mail: mesteller@carrerasresearch.org
†Pedro Blecua and Veronica Davalos contributed equally to this work.

Abstract

High-throughput genomic technologies are increasingly used in personalized cancer medicine. However, computational tools to maximize the use of scarce tissues combining distinct molecular layers are needed. Here we present a refined strategy, based on the R-package 'conumee', to better predict somatic copy number alterations (SCNA) from deoxyribonucleic acid (DNA) methylation arrays. Our approach, termed hereafter as 'conumee-K_{CN}', improves SCNA prediction by incorporating tumor purity and dynamic thresholding. We trained our algorithm using paired DNA methylation and SNP Array 6.0 data from The Cancer Genome Atlas samples and confirmed its performance in cancer cell lines. Most importantly, the application of our approach in cancers of unknown primary identified amplified potentially actionable targets that were experimentally validated by Fluorescence *in situ* hybridization and immunostaining, reaching 100% specificity and 93.3% sensitivity.

Keywords: somatic copy number alterations, DNA methylation, gene amplification, actionable target identification, cancers of unknown primary

Introduction

Deoxyribonucleic acid (DNA) methylation alterations, particularly the epigenetic inactivation of tumor suppressor genes, are a hallmark of human tumors and are increasingly used as biomarkers and drug targets [1–3]. For this reason, epigenomic tools that are cost-effective, such as DNA methylation microarrays, are gaining momentum for translational purposes [1–3]. In recent years, as a potential alternative to the use of SNP-based SNP Array 6.0 (SNP6) arrays, several approaches to detect genome-wide Somatic Copy Number Alterations (SCNAs) from Infinium Human Methylation 450K/EPIC arrays have been developed [4–6] and are applied in several fields, including cancer research [i.e. 7–9]. Herein, we present a refined strategy to predict SCNAs using 450K DNA methylation microarrays [10], based on conumee [4], that allow to detect SCNAs quantitatively in cancer. Our approach, hereafter referred as conumee-K_{CN}, refines conumee's calling of SCNAs by estimating a dynamic sample-dependent threshold for different copy number states while accounting for tumor purity-, intra-sample- and copy number state-associated variation. Commonly, SCNAs are detected by a fixed threshold (e.g. >0.3) or by *n* SDs from the sample mean/median and are not further distinguished [4–6]. Previous studies showing associations between high amplifications and high drug response rates have encouraged us to refine the SCNA calling from DNA methylation arrays to further be able to distinguish between gains, moderate amplifications and high amplifications. For instance, the highest responses to crizotinib are observed among non-small cell lung cancer patients harboring high levels

Pedro Blecua, PhD, is an associate researcher at the Josep Carreras Leukaemia Research Institute in Spain.
Veronica Davalos, PhD, is an associate researcher at the Josep Carreras Leukaemia Research Institute in Spain.
Izar de Villasante, MSc, is a bioinformatician at the Josep Carreras Leukaemia Research Institute in Spain.
Angelika Merkel, PhD, is a head of Bioinformatics Unit at the Josep Carreras Leukaemia Research Institute in Spain.
Eva Musulen, MD, PhD, is an associate researcher at the Josep Carreras Leukaemia Research Institute in Spain. She also works as pathologist at the Hospital Universitari General de Catalunya-Grupo Quirónsalud in Spain.
Laia Coll-SanMartin is a PhD student at the Josep Carreras Leukaemia Research Institute in Spain.
Manel Esteller, MD, PhD, is an investigator of the Cancer Epigenetics group at the Josep Carreras Leukaemia Research Institute in Spain, and the Centro de Investigación Biomédica en Red de Cancer (CIBERONC) in Spain. He is also a professor of the Institucio Catalana de Recerca i Estudis Avançats (ICREA) in Spain; and the Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB) in Spain.

Received: October 24, 2021. **Revised:** March 30, 2022. **Accepted:** April 10, 2022

of MET amplification [11]. Similarly, high responses to FGFR inhibitors are detected in high-level clonal FGFR2-amplified gastric cancer patients [12].

The dual detection of DNA methylation and copy number alterations using only the epigenomic platform could be particularly relevant in the clinical setting, considering the common limitations in the amount of available tissue as well as the short timing usually available for oncology patients to receive the best tailored individual treatment. We highlight the application of our algorithm in a cohort of 211 cases of Cancer of Unknown Primary (CUPs), previously profiled by our group using DNA methylation arrays [13]. The intrinsic features of CUPs, including their early dissemination, aggressive clinical course and lack of evident pharmacological targets [14–16], provide an ideal scenario to demonstrate the advantages of identifying potentially druggable somatic amplified targets to broaden the therapeutic alternatives to treat this orphan tumor type and to eventually improve its dismal clinical outcome.

Materials and methods

Affymetrix SNP6 microarray processing

Affymetrix SNP6 array data (.CEL files) from The Cancer Genome Atlas (TCGA) were downloaded from Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/) and were processed in our previous study [17]. In brief, SNP6 array data were processed together, quantile-normalized and median-polished with Affymetrix power tools. Genotyping was performed with the Birdseed algorithm. PennCNV [18] was employed to generate $\log_2$ R ratio and B-allele frequencies. Total- and allele-specific somatic copy number calls, as well as purity estimation, were generated with ASCAT [19].

450K DNA methylation microarrays processing

Illumina Infinium HumanMethylation450 (450K) BeadChip array data from TCGA was downloaded from GDC Data Portal (https://portal.gdc.cancer.gov/) using the R package TCGAbiolinks. From the 7009 TCGA samples with matched genotyping (SNP6 array) and DNA methylation (450K) data, 442 samples across 18 cancer types were used as our training cohort (Supplementary Table S1 available online at https://academic.oup.com/bib), and 151 samples across 11 cancer types were used as our validation cohort (Supplementary Table S2 available online at https://academic.oup.com/bib). The 450K data from 75 normal samples across 18 different tissues were also retrieved from TCGA. The 450K data from 96 whole-blood (WB) samples from healthy individuals were available from a previous study [20]. In addition, 450K data from cancer cell lines (CCLs) were downloaded from the Cancer Cell Line Encyclopedia (CCLE, Broad Institute) using the R-package GEOquery Gene Expression Omnibus (GEO) database (GSE68379). The 450K data from 211 samples of CUPs were available in house from our previous study [13]. For all 450K datasets, we used the R-package minfi [21] for processing and quality control. We applied quantile normalization and a minimum detection threshold with P-value < 0.01 per probe and a maximum of 10% of failed probes per sample. Sex chromosomes and cross-reactive probes were excluded from the analysis, and probes that were located +/− 10 bases away from known SNPs were also filtered out.

Improved quantitative SCNA calling from 450K data based on conumee

We based our SCNA calling strategy on conumee [4], a popular tool for SCNA calling from DNA methylation arrays. Conumee calculates copy number alterations based on the signal intensity (I) ratio between tumor and normal samples ($R = I_{\text{Tumor}}/I_{\text{Normal}}$) using the following steps: (1) normalized intensity values of both the 'methylated' and 'unmethylated' signals are added; (2) a 'best-linear-fit' is performed between the query sample (tumor, in our case) and all normal healthy samples ([20], in our case); (3) the $\log_2$-ratio of probe intensities between the two is calculated; (4) from those median $\log_2$-ratio of all probes intensities for predefined genomic bins are derived ($\log_2[R]$) and (5) copy number alterations are estimated as difference of these ratios from the mean of all bins (mean($\log_2[R]$):

$$cn = \log_2[R] - \mathbf{mean}(\log_2[R]).$$

Note that the intercept, mean($\log_2[R]$), represents the copy number neutral state (baseline), since $I_{\text{Tumor}} = I_{\text{Normal}}$. (6) Final segments are derived from the predefined bins using circular binary segmentation (DNAcopy package [22]). Finally, SCNAs can be detected when the copy number estimate (cn) exceeds a certain threshold, either a fixed value or measured as units of SD.

We refined this approach by (1) estimating dynamic thresholds for different types of SCNAs ('copy number states') and (2) incorporating tumor purity, intra-sample variability and copy-number-state-dependent noise.

To define a threshold for copy number estimates (cn) above which we denote a certain copy number state (CN) ('Amp10', 'Amp', 'Gain', 'HetLoss' or 'HomDel', see Results), we assume that this threshold is array (a) dependent and proportional to the tumor purity (ρ), the intra-sample variation of the array ($sd(\log_2[R_a])$) and a constant (K_{CN}) that reflects the copy number-state-associated biological/technical variation:

$$T_{CN} = \mathbf{mean}(\log_2[R_a]) + K_{CN} * \rho * sd(\log_2[R_a]).$$

Tumor purity (ρ) is calculated using the 'ABSOLUTE' method within the 'RF_Purify' R package [23].

Similarly, we denote that the difference between intensity ratios of segments with copy number alterations ($\log_2[R_{CN}]$) and the overall array mean (mean($\log_2[R_a]$)

varies with the tumor purity (ρ) and the intra-sample variation of the array ($sd(\log_2[R_a])$):

$$\rho * sd(\log_2[R_a]) \sim \log_2[R_{CN}] - \mathbf{mean}(\log_2[R_a]).$$

For each copy number state (CN), we perform a linear regression and retrieve K_{CN}:

$$\rho * sd(\log_2[R_a]) \sim 0 + (\log_2[R_{CN}] - \mathbf{mean}(\log_2[R_a])),$$
$$y = 0 + Bx,$$
$$B = y/x,$$
$$K_{CN} = 1/B \rightarrow x/y,$$
$$*y = \rho * sd(\log_2[R_a]), x = \log_2[R_{CN}] - \mathbf{mean}(\log_2[R_a]),$$
$$K_{CN} = (\log_2[R_{CN}] - \mathbf{mean}(\log_2[R_a]))/\rho * sd(\log_2[R_a])$$

An example of linear regression is shown in Supplementary Figure S1**A** available online at https://academic.oup.com/bib. We have termed our method 'conumee-K_{CN}'.

Software

We run conumee [4] (v. 1.26.0) with default parameters using the option 'exclude' [excluded over 10 000 genomic regions corresponding to common copy number polymorphisms in the otherwise healthy population (arising from both tissue and blood samples), as described by the authors in the conumee online manual]. For the comparative benchmark, we run ChAMP [5] (v.2.22.) and cnAnalysis450k [6] (v.0.99.26) with default parameters, and according to the authors instructions, using a minimum frequency of >0.3 and an effect size of >0.15 to call amplifications, respectively.

Statistical analyses

All statistical analyses were performed under the R package, R version 4.1.2 (2021-11-01)—'Bird Hippie'. The P-values, displayed in Supplementary Table S8 and Supplementary Figure S5**B**, available online at https://academic.oup.com/bib, were calculated with a two-sided Fisher exact test.

Experimental allidation of potential actionable targets

Fluorescence *in situ* hybridization (FISH)

Formalin-fixed, paraffin-embedded (FFPE) tissue sections were analyzed using standard FISH techniques using the following commercial probes: MYC IQFISH Break-Apart Probe (Agilent Technologies, G111623-2, Santa Clara, CA, USA), CCND1 IQFISH Break-Apart Probe (Agilent Technologies, G111622-2), PIK3CA Spectrum Green FISH Probe Kit/CEP3 SpectrumOrange (Vysis, 06N10-001/06J36-003, Abbott Laboratories, Chicago, IL, USA), LSI MET SpectrumRed FISH probe Kit/CEP7 SpectrumGreen (Vysis, 06N05-020/06J37-007), LSI BCL6 (ABR) Dual Color, Break Apart Rearrangement Probe (Vysis, 01N23-020), PathVysion LSI HER-2/neu SpectrumOrange/CEP17 SpectrumGreen (Vysis, 02J01/06J37-017 and CCNE1 BAC-Spectrum Red labeled probe (RP11-104J24). Preparation of slides, hybridization and analysis were performed according to standard procedures [24–27].

Hybridizations were analyzed using a standard fluorescence microscope (Leica DM5500 B Fluorescence microscopy, Leica Biosystems Newcastle Ltd, Newcastle upon Tyne, UK Nikon Eclipse 50i, Tokyo, Japan, or Zeiss, Oberkochen, Germany) equipped with appropriate filter sets. Acquisition and processing of digital images were performed using CytoVision Imaging System (Leica Biosystems Newcastle Ltd) or the ISIS FISH Imaging System (MetaSystems, Altlussheim, Germany).

Immunohistochemistry (IHC)

FFPE tissue sections were analyzed using standard IHC techniques. The primary antibodies used were: anti-CCND1 (Cyclin D1 P211F11 clone, PA0046), anti-c-MYC antibody (c-MYC Y69 clone, 3PR00355) and anti-BCL6 (BLC-6 LN22 clone, PA0204) from Leica Biosystems (Leica Biosystems Newcastle Ltd); anti-CCNE1 (HE12, sc-247, Santa Cruz Biotechnology, Dallas, TX, USA) and anti-HER2/ERBB2 [PATHWAY HER-2/neu (4B5), TA9145, Ventana, Roche Diagnostics, Basel, Switzerland]. Immunostainings were performed automatically using automatized protocol 'F' on Leica BOND-MAX platform (Leica Biosystems) except for HER2, where the BenchMark ULTRA IHC/ISH system (Roche Diagnostics) was used. Positive staining for anti-CCND1, anti-CCNE1, anti-c-MYC and anti-BCL6 antibody was localized in the nucleus of the neoplastic cells. The evaluation was semiquantitative (weak, moderate or intense), indicating the percentage of positive nuclei. Positive staining for anti-HER2/ERBB2 antibody was localized in the membrane of the neoplastic epithelial cells. The evaluation was performed according to the criteria described in the CAP/ASCO guidelines [25]. For each staining, an external positive control was included.

Results

We developed a refined strategy, based on conumee [4], that allows us to dynamically define a threshold to quantitatively call copy number variations from 450K DNA methylation data in cancer (Figure 1**A**). The algorithm was trained using 442 tumor samples publicly available from TCGA, representing 18 cancer types with sample-matched genotyping (SNP6 array) and DNA methylation (450K array) data (Figure 1**A**).

For our method, we define five copy number states (CN): (1) Homozygous deletions 'HomDel' (complete loss of all the alleles) = 0 copies; (2) Hemizygous deletions 'HetLoss' (loss of one allele) = 1 copy; (3) Gains 'Gain' = 3–4 copies; (4) Moderate amplification 'Amp' = 5–9 copies and (5) High amplification 'Amp10' = ≥10 copies. We used as reference a list of 94 genes that include 83 frequently amplified genes [28] and 11 frequently deleted genes [29, 30] across cancers (Supplementary Table S3 available online at https://academic.oup.com/bib).

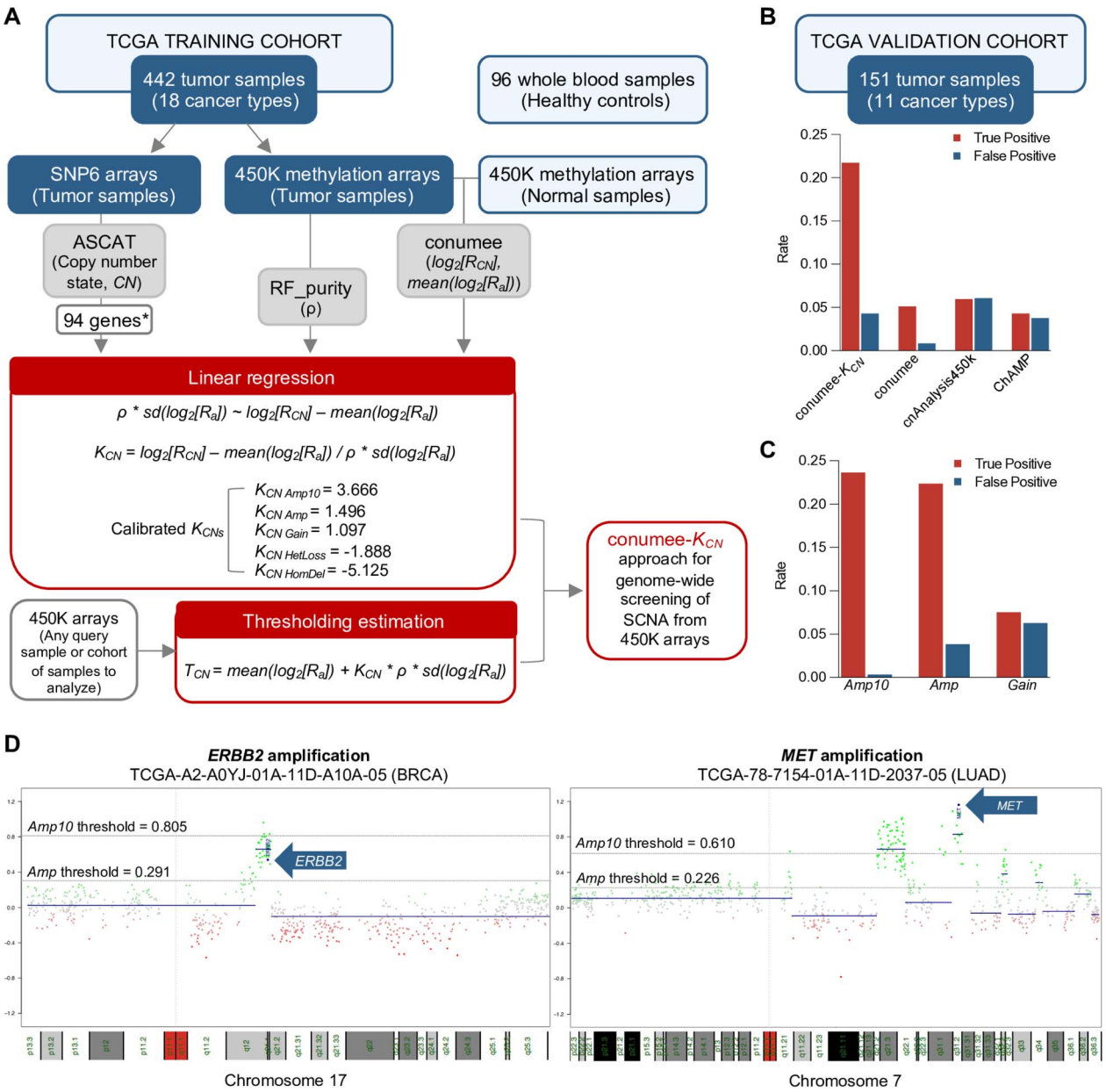

Figure 1. Computational prediction of SCNA from DNA methylation arrays using conumee-K_{CN}. (**A**) Workflow for the developed strategy to refine SCNA detection, 'conumee-K_{CN}', trained over 442 primary tumor samples from TCGA, across 18 cancer types, with matched genotyping (SNP6 array) and DNA methylation array (450K) data. Our thresholding strategy refines conumee outputs and quantitatively calls SCNA from 450K arrays by considering tumor purity ρ (RF_Purity) and rigorous estimation of copy-number-state (*CN*)-dependent constants K_{CN}. *A list of 94 genes frequently amplified or deleted in cancer was used as reference to define the copy number states. Thus, by using calibrated K_{CN}'s and considering tumor purity, intra-sample variability and copy-number-state-dependent noise, thresholds for each *CN* can be estimated for each 450K profiled sample to accurately identify SCNA. (**B**) Benchmarking of our strategy (conumee-K_{CN}) against conumee (fixed threshold of 0.3), cnAnalysis450k and ChAMP in an independent, validation set consisting of 151 TCGA samples, with matched genotyping (SNP6 array) and DNA methylation array (450K) data. True positive (TP) and false-positive (FP) rates of 450K-derived calls versus SNP6-derived calls (ASCAT) for amplifications are depicted, showing the improved performance of our approach. (**C**) TP and FP rates of conumee-K_{CN} versus ASCAT in the TCGA validation set for the three amplification copy number states (Amp10, Amp and Gain). (**D**) Representative examples of gene amplifications in two samples from the TCGA validation cohort. Thresholds estimated by conumee-K_{CN} for Amp and Amp10 are depicted (dotted grey lines). TP = *#True positives/#True positives + #False Negatives*; FP = *#False positives/#False positives + #True Negatives*.

As a first step, for each sample of our TCGA training cohort (Supplementary Table S1 available online at https://academic.oup.com/bib), we called SCNA from genotyping data using ASCAT (see Materials and Methods) and created gene subsets for each copy number state (*CN*) based on the respective calls. Next, we used conumee and the corresponding DNA methylation data to calculate for each of the samples and each of the copy number state: (1) the intra-sample variation of the signal intensity ratio (sd($\log_2[R_a]$), (2) the mean signal intensity ratio (mean($\log_2(R_{CN})$) for all genes in the corresponding gene list and (3) the global array mean signal intensity (mean(($\log_2[R_a]$)). To broaden the applicability of this tool to any tissue type, we used WB samples from healthy

individuals [20] as normal control, as is often done in large scale studies such as TCGA (GISTIC and ASCAT). We verified the suitability of this type of samples as normal counterpart by testing for significant DNA methylation differences between our sample set (96 samples from 48 men and 48 women, all young adults, to avoid biases regarding age and gender) and a set of 75 normal samples spanning 18 tissues. In 97% (1269) of the 1307 CpGs, used to infer the copy number state of the 94 genes frequently amplified or deleted in cancer, we observed no significant differences (Supplementary Table S4 available online at https://academic.oup.com/bib).

In parallel, we estimated the tumor purity (ρ) of each sample using the RF_Purify software [23], results that were strongly correlated with the purity estimates from genotyping data (SNP6 array) derived with ASCAT ($R^2 = 0.573$; $R = 0.76$; Supplementary Figure S2 available online at https://academic.oup.com/bib). Finally, implementing the above estimates, we performed a linear regression and extracted the copy number-state-dependent constants K_{CN} required for our thresholding strategy (see Materials and Methods). This workflow is summarized in Figure 1**A**. We additionally estimated mean K_{CN}'s from 10 000 randomizations with 442 samples each, drawing from the total set of 7009 TCGA samples from 18 cancer types (Supplementary Figure S1**B** available online at https://academic.oup.com/bib). The obtained K_{CN} values highly reflect the values derived from the 442 TCGA samples used as training set.

Next, using a validation cohort consisting of an independent set of 151 TCGA tumor samples across 11 cancer types (Supplementary Table S2 available online at https://academic.oup.com/bib), we benchmark the performance of our thresholding strategy against conumee using a fixed threshold (>0.3) and two other commonly used tools, namely, ChAMP [5] and cnAnalysis450k [6] (Figure 1**B**). We compared the concordance between SCNA calls from SNP6 genotyping arrays (ASCAT) and SCNA calls from 450k methylation arrays using the respective approaches, focusing on amplifications (Gains, Amp and Amp10), the center of our additional analyses. Figure 1**B** shows how, while the call concordance [true-positive (TP) rate] is similar for the three standard approaches (TP = 0.051 for conumee-fixed threshold 0.3, TP = 0.043 for ChAMP and TP = 0.060 for cnAnalysis450k), our strategy conumee-K_{CN} sensibly improves SCNA prediction with a TP = 0.218, demonstrating the superiority of our method in comparison with previously developed tools.

Moreover, a significant improvement of our strategy in comparison with previous methods [4–6] is that it allows to detect SCNAs quantitatively to distinguish Gains, Amps or Amp10s. As shown in Figure 1**C**, the concordance of Amps (TP = 0.224) and Amp10s (TP = 0.236) further support the added value of our approach. Representative examples of *ERBB2* amplification in a breast cancer (BRCA) sample and *MET* amplification in a lung adenocarcinoma sample, predicted by our approach from the TCGA validation cohort, are shown in Figure 1**D**.

The performance of conumee-K_{CN} was further assessed in CCLs available from the CCLE (Broad Institute). The higher concordance with SNP6-derived calls (TP = 0.247), in comparison with conumee (TP = 0.109), corroborates the superiority of our approach (Supplementary Figure S3 available online at https://academic.oup.com/bib).

Once we confirmed the robustness of our strategy, we applied it to a set of CUP samples aimed to illustrate its usefulness in the clinical setting (Figure 2**A**). First, we applied conumee to the DNA methylation arrays from 211 CUP cases previously profiled by our group [13]. Next, using the K_{CN}'s calibrated from 442 TCGA primary tumors (Figure 1**A**) along with tumor purity estimates from RF_Purity (Supplementary Table S5 available online at https://academic.oup.com/bib), we annotated the different copy number states (*CN*): Amp10, Amp, Gain, HetLoss or HomDel. Aiming to identify gene amplification events with a relevant role in CUP pathogenesis, we dropped gains as it is not uncommon for cancer genomes to undergo whole genome doubling [31] and selected those genes that were amplified (*CN* = Amp or Amp10) in at least 5 CUPs (2.4% of 211 CUPs; Figure 2**A**). Considering the clinical potential of our findings, we focused on gene amplifications given the progress of inhibitory drugs for cancer treatment, and restricted the analysis to protein coding genes (as annotated by Gencode_v34), also taking into account that 450K arrays are enriched in CpGs located in promoters and gene bodies [10]. After this filtering process, a total of 1159 candidate genes amplified in CUPs were selected for further analysis (Figure 2**A**; Supplementary Table S6 available online at https://academic.oup.com/bib).

Going one step further with the CUP cohort, in order to identify actionable gene amplifications with a potential relevance in the clinical setting and thus broaden the therapeutic opportunities for this dismal tumor type, we used publicly available data and online tools to further explore the 1159 amplified genes, including the Cancer Genome Interpreter (https://www.cancergenomeinterpreter.org/) [32], Precision Oncology Knowledge Base (https://www.oncokb.org) [33], Clinical Interpretations of Variants in Cancer (https://civicdb.org/) and Clinical Trials (www.clinicaltrials.gov). The inclusion criteria to identify potential actionable targets were as follows (Figure 2**A**): (1) Is that gene a known or predicted driver in cancer?; (2) Is there a drug (preclinical or clinical) to target that gene? and (3) Is the amplification of that gene a potential biomarker for the drug response?. Preclinical evidences, early and late clinical trials and FDA guidelines were also considered (Figure 2**A**). According to these criteria, we identified 15 potential actionable targets (Figure 2**A**; Supplementary Table S7 available online at https://academic.oup.com/bib), including well-recognized oncogenes such as *MYC*, *CCND1*,

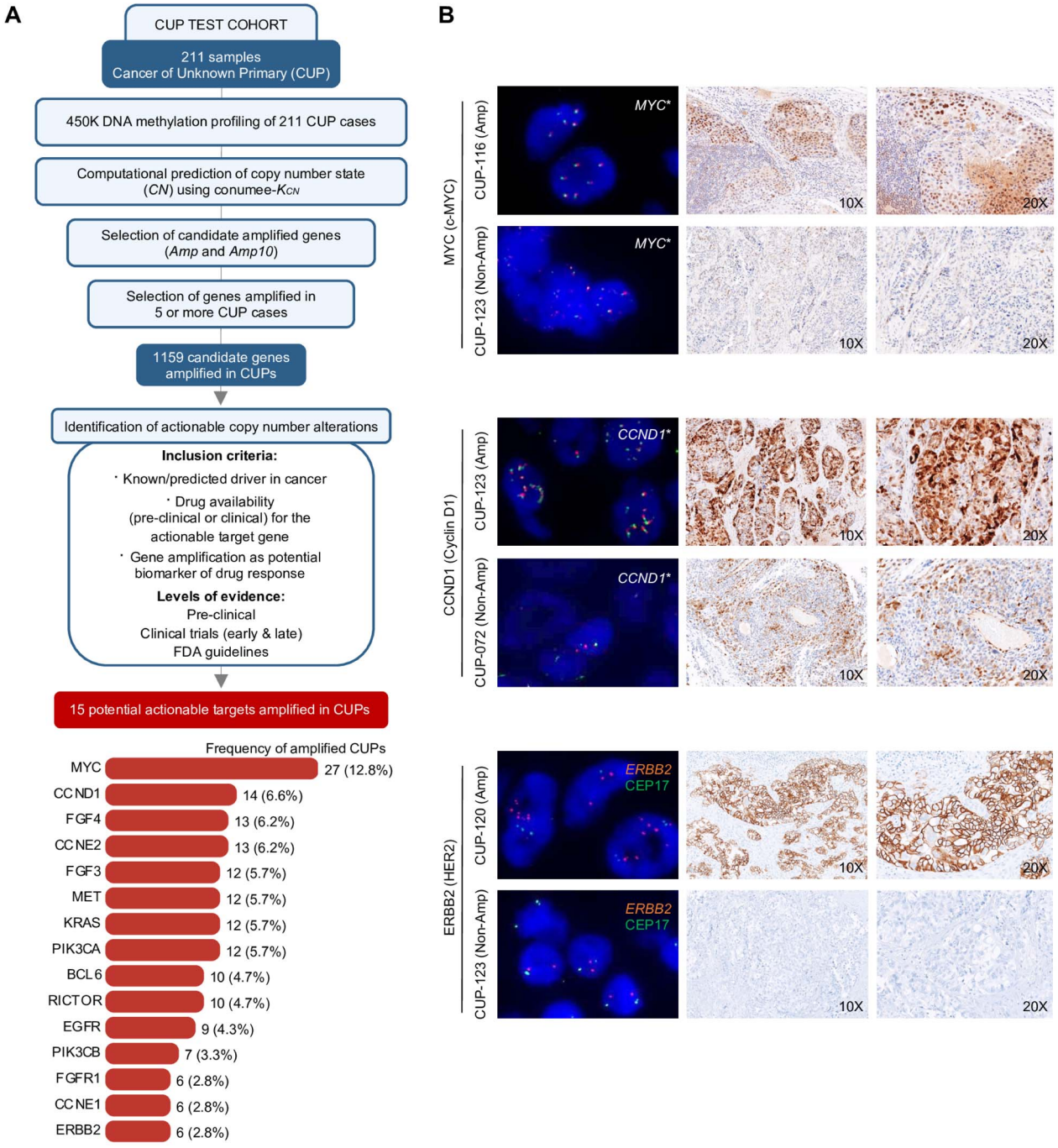

Figure 2. Epigenomic-based computational prediction of SCNA in CUP using conumee-K_{CN}. (**A**) Workflow for the detection of SCNA in CUPs using conumee-K_{CN} and further identification of gene amplifications with clinical relevance. (**B**) Experimental validation of conumee-K_{CN} predicted SCNA in CUP patient samples. Representative images of copy number state validation by FISH and protein expression by IHC are shown for three well-recognized oncogenes: *MYC* (c-MYC), *CCND1* (Cyclin D1) and *ERBB2* (HER2). *Orange/green break-apart FISH probes were used.

ERBB2 (Supplementary Figure S4 available online at https://academic.oup.com/bib), *BCL6*, *PIK3CA*, *MET* and *CCNE1*. Furthermore, we found some of these genes to be significantly co-amplified, not only those within the same amplicon (e.g. role of amplicon 8q24 in cancer [34], Supplementary Figure S4 available online at https://academic.oup.com/bib) but also genes located on different chromosomes [i.e. *FGF3/FGF4* (11q13.3) co-amplified with *PIK3CA/PIK3CB* (3q26.32/3q22.3)] (Supplementary Figure S5, Supplementary Table S8 available online at https://academic.oup.com/bib).

Finally, to demonstrate the robustness of our strategy in a *bona fide* manner, we sought to experimentally validate our predictions using the gold standard approach to assess copy number: FISH. The results of the experimental validation of 21 cases are summarized in Supplementary Figure S6 available online at https://academic.oup.com/bib. FISH assays confirmed

the predicted SCNA calls in 95.2% of cases (14/14 amplification events and 6/7 non-amplifications). In terms of the confusion matrix, we obtained a sensitivity of 93.3% and a specificity of 100%. Representative examples of amplified and non-amplified CUP cases for three well-recognized oncogenes, *ERBB2* (HER2), *MYC* (c-MYC) and *CCND1* (Cyclin D1), are shown in Figure 2**B**. Importantly, in comparison to non-amplified cases, these gene amplification events were clearly associated with protein upregulation detected by IHC (Figure 2**B**). Additional validations of BCL6 and MET are shown in Supplementary Figure S7 available online at https://academic.oup.com/bib. Thus, our experimental results confirmed the SCNA predictions provided by our DNA methylation microarray-based strategy conumee-K_{CN} herein presented.

Conclusion and Discussion

The occurrence of shared molecular alterations in different tumor types has supported the emergence of tissue-agnostic approaches, which have led to a revolutionary paradigm shift in drug development and cancer treatment. Novel precision oncology trial designs, such as basket trials, which eligibility is based on the presence of specific genetic alterations, irrespective of histology, are opening bright opportunities for patients with tumors harboring actionable alterations. The tissue-agnostic arsenal of FDA-approved drugs is increasing, and considering this scenario, optimization of computational strategies to maximize the identification of actionable alterations that guide the targeted treatment choice is paramount. The more actionable alterations identified, the more tools for precision medicine.

Here, we described conumee-K_{CN}, an improved strategy based on conumee [4], to detect SCNA in cancer using DNA methylation microarrays as a surrogate for genotyping arrays. Our approach shows significant improvement over similar approaches, namely, conumee (standard) [4], ChAMP [5] and cnAnalysis450k [6], that previously have been reported to have low reliability [35]. Moreover, the refinement of our approach estimates a dynamic threshold for each copy number state, allowing differentiation between gains, moderate amplifications and high amplifications, a functionality that does not exist in the currently available tools. Thus, with conumee-K_{CN}, the user can estimate sample-dependent thresholds for Gains, Amps or Amp10s by using calibrated K_{CN}'s and considering tumor purity and intra-sample variability to accurately identify SCNA from any 450K profiled sample. Considering previous studies showing highest drug responses among patients harboring high levels of amplifications of the target gene [i.e. 11, 12], this feature of conumee-K_{CN} is an added value that could better guide treatment decision-making in healthcare of cancer patients. Furthermore, whereas genotyping-based SCNA callers generally take tumor purity into account, among current DNA methylation-based SCNA calling tools, this is unique for conumee-K_{CN}. Incorporation of tumor purity in the thresholding function could improve SCNA prediction, considering that cancer-cell intrinsic features could be hidden in biological samples with a low proportion of tumor cells, which is not uncommon in clinical setting.

The usefulness of conumee-K_{CN} was illustrated in CUP, a heterogeneous group of metastatic tumors that lack an identifiable primary tumor despite a standardized diagnostic work-up. Most CUP patients (80–85%) have an unfavorable prognosis with a dismal survival of 3–6 months despite empirical chemotherapy treatments [16, 36, 37]. Thus, CUP management is an unmet medical need. Using our epigenomic-based computational predictions of SCNA in CUPs, we identified many potential drug-actionable targets, including well-recognized oncogenes, such as *MYC*, *CCND1*, *ERBB2*, *BCL6*, *PIK3CA*, *MET* and *CCNE1*. Noticeably, some of these genes had been previously identified in CUPs with very similar amplification frequencies [38–43], confirming the power of our approach. More relevant, successful examples of CUP cases harboring *MET* [43, 44] or *ERBB2* [45] amplifications with favorable response to the corresponding target therapies (crizotinib or trastuzumab, respectively) reinforce the importance of intensifying the research in this field. Currently, an ongoing clinical trial is comparing the efficacy of molecularly guided therapies versus platinum-based standard chemotherapy in CUP patients [38].

The dual use of DNA methylation arrays to identify both epigenomic and genomic SCNA alterations is a significant advantage particularly evident in the clinical practice, considering the need to maximize the use of scarce tissues and the limited time to obtain the most comprehensive molecular information to efficiently guide therapeutic decision-making. In addition, a genome-wide approach, as the herein described, boosts the possibility of identifying co-amplification events, which might open new venues for combination of targeted therapies in CUPs in a tissue agnostic manner. Intriguingly, preclinical studies have shown that the combined use of FGFR and PI3K inhibitors resulted in an enhanced efficacy in comparison with single treatments [46, 47]. Furthermore, although we focused the CUP analysis on the most promising actionable targets considering the current evidences, further studies of the remaining candidates identified by our conumee-K_{CN} approach could reveal genes or pathways with key roles in CUP pathogenesis that might open novel therapeutics opportunities for this dismal cancer type.

The robustness of conumee-K_{CN} is reflected by the high sensitivity (93.3%) and specificity (100%) obtained through the experimental validation by FISH. In this regard, selective tissue-agnostic treatment of CUPs focusing on the gene targets identified in this study might improve the very scarce landscape of effective therapies against this tumor type. Importantly, our

approach is in principle flexible and generalizable to other DNA methylation microarrays (e.g. EPIC Methylation Arrays) as well as potentially useful for any cancer type. Thus, the application of conumee-K_{CN} in further biocomputational studies is warranted.

Data availability

Data and code to reproduce key components of Figure 1 are publicly available at https://github.com/izarvillasante/SCNA.git. SNP6 array data to determine total Copy Number, as well as 450K DNA methylation microarrays, were obtained from TCGA at https://portal.gdc.cancer.gov/. CUP and WB 450K array data were obtained from [13, 20], respectively.

Competing interests

M.E. is a consultant to Ferrer International and Quimatryx. The remaining authors declare that they have no competing interests.

Key Points

- In the era of precision medicine, maximizing the use of commonly scarce tumor tissue to offer the most comprehensive molecular landscape from one biological sample is a critical need.
- We improved SCNA prediction from DNA methylation arrays with our conumee-K_{CN} strategy to further distinguish gains, moderate amplifications and high amplifications, a tool potentially useful for any cancer type.
- The added value of our strategy is significant, particularly since none of the existing approaches using DNA methylation arrays stratify gene amplifications, a feature especially relevant in clinical setting.
- The usefulness of conumee-K_{CN} strategy was illustrated in samples from CUP patients, an aggressive tumor type with dismal prognosis.
- The identification of 15 candidate actionable targets amplified in CUPs highlights the potential of our approach in the clinical setting.
- The conumee-K_{CN} SCNA predictions in CUPs were validated by experimental procedures (FISH, IHC), demonstrating a very high sensitivity (93.3%) and specificity (100%).

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Authors' contributions

P.B., I.d.V. and A.M. took care of algorithm design, implementation and computational analysis; V.D. was in charge of target identification and curation and sample preparation; E.M. and L.C.-S. were in charge of sample preparation and handling and experimental results curation. P.B., V.D. and M.E. wrote the manuscript. All authors read, edited and approved the final manuscript.

Acknowledgements

We thank Jonathan Garcia Barrasa for the technical support.

Funding

Sarah Jennifer Knott Foundation Research Award (to M.E.) Health Department PERIS project no. SLT/002/16/00374 and AGAUR project no. 2017SGR1080 of the Catalan Government (Generalitat de Catalunya); Ministerio de Ciencia e Innovación; Agencia Estatal de Investigación; European Regional Development Fund project no. RTI2018-094049-B-I00; Cellex Foundation; 'la Caixa' Banking Foundation (LCF/PR/GN18/51140001).

References

1. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. *Nat Rev Genet* 2019;**20**(2):109–27.
2. Mancarella D, Plass C. Epigenetic signatures in cancer: proper controls, current challenges and the potential for clinical translation. *Genome Med* 2021;**13**(1):23.
3. Locke WJ, Guanzon D, Ma C, *et al.* DNA methylation cancer biomarkers: translation to the clinic. *Front Genet* 2019;**10**:1150.
4. Hovestadt V, Zapatka M. Conumee: enhanced copy-number variation analysis using Illumina DNA methylation arrays. R package version 190. http://bioconductor.org/packages/conumee/.
5. Feber A, Guilhamon P, Lechner M, *et al.* Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biol* 2014;**15**(2):R30.
6. Knoll M, Debus J, Abdollahi A. cnAnalysis450k: an R package for comparative analysis of 450k/EPIC Illumina methylation array derived copy number data. *Bioinformatics* 2017;**33**(15):2266–72.
7. Gao Y, Widschwendter M, Teschendorff AE. DNA methylation patterns in normal tissue correlate more strongly with breast cancer status than copy-number variants. *EBioMedicine* 2018;**31**:243–52.
8. Haider Z, Landfors M, Golovleva I, *et al.* DNA methylation and copy number variation profiling of T-cell lymphoblastic leukemia and lymphoma. *Blood Cancer J* 2020;**10**(4):45.
9. Wang X, Grasso CS, Jordahl KM, *et al.* Copy number alterations are associated with metastatic-lethal progression in prostate cancer. *Prostate Cancer Prostatic Dis* 2020;**23**(3):494–506.
10. Sandoval J, Heyn H, Moran S, *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 2011;**6**(6):692–702.
11. Camidge DR, Otterson GA, Clark JW, *et al.* Crizotinib in patients with MET-amplified NSCLC. *J Thorac Oncol* 2021;**16**(6):1017–29.
12. Pearson A, Smyth E, Babina IS, *et al.* High-level clonal FGFR amplification and response to FGFR inhibition in a translational clinical trial. *Cancer Discov* 2016;**6**(8):838–51.
13. Moran S, Martinez-Cardus A, Sayols S, *et al.* Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol* 2016;**17**(10):1386–95.
14. Kato S, Alsafar A, Walavalkar V, *et al.* Cancer of unknown primary in the molecular era. *Trends Cancer* 2021;**7**(5):465–77.
15. Laprovitera N, Riefolo M, Ambrosini E, *et al.* Cancer of unknown primary: challenges and progress in clinical management. *Cancers (Basel)* 2021;**13**(3):451.
16. Rassy E, Pavlidis N. Progress in refining the clinical management of cancer of unknown primary in the molecular era. *Nat Rev Clin Oncol* 2020;**17**(9):541–54.

17. Riaz N, Blecua P, Lim RS, *et al.* Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nat Commun* 2017;**8**(1):857.
18. Wang K, Li M, Hadley D, *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007;**17**(11):1665–74.
19. Van Loo P, Nordgard SH, Lingjaerde OC, *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* 2010;**107**(39): 16910–5.
20. Voisin S, Almen MS, Zheleznyakova GY, *et al.* Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers. *Genome Med* 2015;**7**:103.
21. Aryee MJ, Jaffe AE, Corrada-Bravo H, *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014;**30**(10):1363–9.
22. Olshen AB, Venkatraman ES, Lucito R, *et al.* Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 2004;**5**(4):557–72.
23. Johann PD, Jager N, Pfister SM, *et al.* RF_Purify: a novel tool for comprehensive analysis of tumor-purity in methylation array data based on random forest regression. *BMC Bioinform* 2019;**20**(1):428.
24. Ventura RA, Martin-Subero JI, Jones M, *et al.* FISH analysis for the detection of lymphoma-associated chromosomal abnormalities in routine paraffin-embedded tissue. *J Mol Diagn* 2006;**8**(2): 141–51.
25. Wolff AC, Hammond MEH, Allison KH, *et al.* Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline focused update. *J Clin Oncol* 2018;**36**(20): 2105–22.
26. Noonan SA, Berry L, Lu X, *et al.* Identifying the appropriate FISH criteria for defining MET copy number-driven lung adenocarcinoma through oncogene overlap analysis. *J Thorac Oncol* 2016;**11**(8):1293–304.
27. Simons A, Shaffer LG, Hastings RJ. Cytogenetic nomenclature: changes in the ISCN 2013 compared to the 2009 edition. *Cytogenet Genome Res* 2013;**141**(1):1–6.
28. Santarius T, Shipley J, Brewer D, *et al.* A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer* 2010;**10**(1): 59–64.
29. Pertesi M, Ekdahl L, Palm A, *et al.* Essential genes shape cancer genomes through linear limitation of homozygous deletions. *Commun Biol* 2019;**2**:262.
30. Cheng J, Demeulemeester J, Wedge DC, *et al.* Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors. *Nat Commun* 2017;**8**(1):1221. https://doi.org/10.1038/s41467-017-01355-0.
31. Quinton RJ, DiDomizio A, Vittoria MA, *et al.* Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* 2021;**590**(7846):492–7.
32. Tamborero D, Rubio-Perez C, Deu-Pons J, *et al.* Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* 2018;**10**(1):25.
33. Chakravarty D, Gao J, Phillips SM, *et al.* OncoKB: precision oncology knowledge base. *JCO Precis Oncol* 2017;**1**.
34. Raeder MB, Birkeland E, Trovik J, *et al.* Integrated genomic analysis of the 8q24 amplification in endometrial cancers identifies ATAD2 as essential to MYC-dependent cancers. *PLoS One* 2013;**8**(2):e54873.
35. Kilaru V, Knight AK, Katrinli S, *et al.* Critical evaluation of copy number variant calling methods using DNA methylation. *Genet Epidemiol* 2020;**44**(2):148–58.
36. Fizazi K, Greco FA, Pavlidis N, *et al.* Cancers of unknown primary site: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2015;**26**(Suppl 5): v133–8.
37. Conway AM, Mitchell C, Kilgour E, *et al.* Molecular characterisation and liquid biomarkers in carcinoma of unknown primary (CUP): taking the 'U' out of 'CUP'. *Br J Cancer* 2019;**120**(2): 141–53.
38. Ross JS, Sokol ES, Moch H, *et al.* Comprehensive genomic profiling of carcinoma of unknown primary origin: retrospective molecular classification considering the CUPISCO study design. *Oncologist* 2021;**26**(3):e394–402.
39. Gatalica Z, Xiu J, Swensen J, *et al.* Comprehensive analysis of cancers of unknown primary for the biomarkers of response to immune checkpoint blockade therapy. *Eur J Cancer* 2018;**94**: 179–86.
40. Clynick B, Dessauvagie B, Sterrett G, *et al.* Genetic characterisation of molecular targets in carcinoma of unknown primary. *J Transl Med* 2018;**16**(1):185.
41. Varghese AM, Arora A, Capanu M, *et al.* Clinical and molecular characterization of patients with cancer of unknown primary in the modern era. *Ann Oncol* 2017;**28**(12):3015–21.
42. Loffler H, Pfarr N, Kriegsmann M, *et al.* Molecular driver alterations and their clinical relevance in cancer of unknown primary site. *Oncotarget* 2016;**7**(28):44322–9.
43. Ross JS, Wang K, Gay L, *et al.* Comprehensive genomic profiling of carcinoma of unknown primary site: new routes to targeted therapies. *JAMA Oncol* 2015;**1**(1):40–9.
44. Palma NA, Ali SM, O'Connor J, *et al.* Durable response to Crizotinib in a MET-amplified, KRAS-mutated carcinoma of unknown primary. *Case Rep Oncol* 2014;**7**(2):503–8.
45. Lee KK, Kim M, Kim KM, *et al.* Next-generation sequencing for better treatment strategy of cancer of unknown primary (CUP). *Ann Oncol* 2019;**30**:v766.
46. Holzhauser S, Lukoseviciute M, Andonova T, *et al.* Targeting fibroblast growth factor receptor (FGFR) and phosphoinositide 3-kinase (PI3K) Signaling pathways in medulloblastoma cell lines. *Anticancer Res* 2020;**40**(1):53–66.
47. Holzhauser S, Lukoseviciute M, Papachristofi C, *et al.* Effects of PI3K and FGFR inhibitors alone and in combination, and with/without cytostatics in childhood neuroblastoma cell lines. *Int J Oncol* 2021;**58**(2):211–25.