Briefings in Bioinformatics, 21(1), 2020, 272–281

doi: 10.1093/bib/bby096
Advance Access Publication Date: 23 October 2018
Problem solving protocol

Copy number aberrations from Affymetrix SNP 6.0 genotyping data—how accurate are commonly used prediction approaches?

Adriana Pitea, Ivan Kondofersky, Steffen Sass, Fabian J. Theis, Nikola S. Mueller and Kristian Unger

Corresponding author: Nikola S. Mueller, Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, 85764, Germany. Tel: +49-89-3187-1174; Fax: +49-89-3187-3369; Email: nikola.mueller@helmholtz-muenchen.de; Kristian Unger, Research Unit Radiation Cytogenetics, Helmholtz Zentrum München, Neuherberg, 85764, Germany and Clinical Cooperation Group Personalized Radiotherapy in Head and Neck Cancer, Helmholtz Zentrum München, Neuherberg, 85764, Germany. Email: unger@helmholtz-muenchen.de

Abstract

Copy number aberrations (CNAs) are known to strongly affect oncogenes and tumour suppressor genes. Given the critical role CNAs play in cancer research, it is essential to accurately identify CNAs from tumour genomes. One particular challenge in finding CNAs is the effect of confounding variables. To address this issue, we assessed how commonly used CNA identification algorithms perform on SNP 6.0 genotyping data in the presence of confounding variables. We simulated realistic synthetic data with varying levels of three confounding variables—the tumour purity, the length of a copy number region and the CNA burden (the percentage of CNAs present in a profiled genome)—and evaluated the performance of OncoSNP, ASCAT, GenoCNA, GISTIC and CGHcall. Furthermore, we implemented and assessed CGHcall*, an adjusted version of CGHcall accounting for high CNA burden. Our analysis on synthetic data indicates that tumour purity and the CNA burden strongly influence the performance of all the algorithms. No algorithm can correctly find lost and gained genomic regions across all tumour purities. The length of CNA regions influenced the performance of ASCAT, CGHcall and GISTIC. OncoSNP, GenoCNA and CGHcall* showed little sensitivity. Overall, CGHcall* and OncoSNP showed reasonable performance, particularly in samples with high tumour purity. Our analysis on the HapMap data revealed a good overlap between CGHcall, CGHcall* and GenoCNA results and experimentally validated data. Our exploratory analysis on the TCGA HNSCC data revealed plausible results of CGHcall, CGHcall* and GISTIC in consensus HNSCC CNA regions. Code is available at https://github.com/adspit/PASCAL.

Key words: copy number calling algorithm; performance assessment; cancer genomics; copy number aberrations

Adriana Pitea is a PhD student at the Computational Cell Maps, Institute of Computational Biology and at the Integrative Biology Group, Research Unit of Radiation, Helmholtz Zentrum München.
Ivan Kondofersky is a Postdoctoral Fellow at the Computational Cell Maps, Institute of Computational Biology, Helmholtz Zentrum München.
Steffen Sass is former Postdoctoral Fellow at the Computational Cell Maps, Institute of Computational Biology, Helmholtz Zentrum München.
Fabian J. Theis is Head of the Institute of Computational Biology and Group Leader Machine Learning, Helmholtz Zentrum München and associate professor holding the chair of 'Mathematical modeling of biological systems', Department of Mathematics, Technical University of Munich.
Nikola S. Mueller is Group Leader of Computational Cell Maps, Institute of Computational Biology, Helmholtz Zentrum München.
Kristian Unger is Head of the Integrative Biology Group and deputy Head of the Research Unit of Radiation, Helmholtz Zentrum München.
Submitted: 21 May 2018; **Received (in revised form):** 11 August 2018

Introduction

Copy number aberrations (CNAs) are present in all known cancer genomes [1–3]. Unlike copy number variations (CNVs) which occur naturally and originate in germline cells [4–6], CNAs accumulate somatically, emerge after many selection events and have been associated with development and progression of human disease, especially with carcinogenesis: Bardeesy *et al.* showed that the deletion of the tumour supressor gene SMAD4 plays a critical role in progression and tumour biology of pancreatic cancer [7], Witkiewicz *et al.* showed that amplification of the gene MYC is uniquely associated with poor outcome in pancreatic ductal adenocarcinoma [8], Leucci *et al.* showed that the long non-coding RNA (lncRNA) gene SAMMSON is consistently co-gained with MITF in more than 90% of human melanomas [9], while Wells *et al.* showed that deletion of the gene PTGHD1 in the thalamic reticular nucleus only leads to attention deficiency and hyperactivity [10]. Identifying CNAs that are affecting oncogenes or tumour suppressor genes provides knowledge required for the development of new targeted cancer therapies or patient stratification. It is thus of great importance to accurately estimate CNAs from tumour genomes. However, one particular challenge in the accurate estimation of cancer-related CNAs is the presence of confounding variables such as tumour purity and length of CNAs.

The tumour purity represents the ratio between cancerous cells and all the cells present in a tumour sample—comprising both of cancerous and non-cancerous cells. The mixture of cancerous and non-cancerous cells affects the expected allelic fraction between germline and somatic variants and thus influences the accuracy of CNA calling [11]. In simple terms, the higher the non-tumour cell content within the assessed tissue sample, the lower the sensitivity of the copy number calling algorithm gets. Previous studies have shown that the length of a CNA region, i.e. the number of covered base pairs by a genomic region, affects the sensitivity of CNA calling, with longer CNA regions being easier to find [12, 13].

Within this study we focus on algorithms that call CNAs from single-nucleotide polymorphism (SNP) arrays. Nowadays, SNP arrays typically comprise approximately 1.8 million probes and return allele-specific signals at each marker of genetic variation. Affymetrix SNP 6.0 data also come with the great advantage that they can be used for both genotype and copy number analysis. Another advantage of this technology is that it allows us to characterise both copy number changes and allelic imbalances of a sample. To achieve this, the signals resulting from the array genotyping need to be processed and analysed by specific methods. Although numerous methods have been proposed, reliably uncovering cancer-associated CNAs from SNP array data still represents a challenge [3, 14, 15]. One difficulty is that CNA calling algorithms fail to address the effect of known biological confounding variables [16, 17], i.e. the tumour purity of the analysed tissue and the length of underlying CNA regions. GenoCN represents a statistical framework that simultaneously searches for CNAs and CNVs while taking into account the tumour purity but does not account for a chromosomal background that is not diploid [18]. OncoSNP represents a unified Bayesian framework based on a cancer-specific statistical model that classifies SNP array signals into 21 states and accounts for tumour purity, polyploidy and intra-tumour heterogeneity [19]. ASCAT focuses on analysing allele-specific copy numbers in solid tumour initially but requires a threshold-based, model-free segmentation of the SNPs into regions of equal copy number [6]. Another method that is used for finding cancer–related CNAs is CGHcall. CGHcall makes use of breakpoint information from segmentation across all samples and includes information as tumour purity for finding CNAs [20].

The Cancer Genome Atlas (https://cancergenome.nih.gov) (TCGA) is one of the largest resources providing molecular omics data on multiple levels. TCGA covers various cancer types and aims to improve general knowledge about cancer development and treatment. The commonly used method to estimate copy number states from SNP genotyping data in TCGA studies is GISTIC 2.0 (GISTIC) [21]. GISTIC was designed to primarily estimate significant relative CNAs across a set of patients and not on single patient level. GISTIC eliminates common chromosome arm-level events which are not cancer-specific and focuses on focal events. However, GISTIC does not address the effect of confounding variables on the resulting CNA regions.

Within this study we assessed the performance of the following common-used CNA calling algorithms on Affymetrix SNP 6.0 array data: OncoSNP [19], ASCAT [6], CGHcall [20], genoCNA [18] and GISTIC [21]. All algorithms are commonly used for estimating copy number states in tumour samples and, except for GISTIC, correct for tumour purity, intra-tumour heterogeneity and tumour cell ploidy (ASCAT and OncoSNP). Unlike previous studies that evaluated CNV detection—and not cancer-specific CNAs—for an SNP platform [13, 22] or used a model with 24 parameters for which it is difficult to find a combination that provides realistic data [23, 24], we focused on five different algorithms designed to specifically find CNAs and, moreover, evaluated them on synthetic data derived from Affymetrix SNP 6.0 data. Our contribution consists of

- a pipeline that uses realistic Affymetrix SNP 6.0 array-like synthetic DNA copy number profiles for evaluating the performance of OncoSNP, ASCAT, CGHcall, genoCNA and GISTIC CNA calling algorithms, under the influence of tumour purity, length of CNA and CNA burden (the percentage of CNAs present in the profiled genome, [25])
- the implementation of an adjusted version of the CGHcall algorithm that allows the estimation of CNAs in highly variant genomes.

We applied our pipeline on two real data sets derived from patient samples: a cohort of 522 head and neck squamous cell carcinoma (HNSCC) samples from TCGA [26] and a set of 81 Haplotype Map samples [4]. The pipelines consist of R, Python and shell scripts and can be accessed at https://github.com/adspit/PASCAL. Finally, we provide an appropriate framework to compare CNAs calling algorithms with the aim of finding the algorithm that classifies genomic regions correctly independent of tumour purity, length of a CNA region and CNA burden. Moreover, we developed an improved version of CGHcall that we refer to as CGHcall* and included it in our comparison.

Methods and materials

Preliminaries

The data resulting from Affymetrix SNP 6.0 arrays experiments comprised of fluorescence intensity values of hybridised A and B allele probes for each genetic marker on the array [27]. We obtained and used the following measures from the data:

(i) the log R ratio (LRR) – a log2-transformed value of the total intensity for allele A and allele B for more than 1.8 million markers of genetic variation.

(ii) the B allele frequency (BAF) – the ratio of bases genotyped as variant allele (B allele). BAF ranged from 0 to 1, where 0 represented the *AA*/*A*− genotype, 0.5 represented the heterozygous *AB* genotype and 1 represented the *BB*/*B*− genotype [28].

Realistic synthetic data

We used the jointseg R package [24] to generate realistic Affymetrix SNP 6.0 array-like synthetic tumour data consisting of 400 samples. Each sample comprised of 1.844.399 markers of genetic variation. Jointseg was built to generate realistic synthetic DNA copy number profiles. The framework resamples signals corresponding to genomic regions with manually annotated copy number states from the publicly available lung cancer NCI-H1395 SNP microarray data [24, 29]. We generated 100 samples with each of the following tumour purity levels: 30, 50, 70 and 100%. The tumour purity levels corresponded to the experimental settings of the [29] study. We randomly placed between 1 and 8 breakpoints within each sample. A breakpoint represented a loci where one of the two parental copy number changed. For the resulting regions we sampled the copy number states from a predefined set of copy number states: (0,1), (0,2), (1,1), (1,2), (1,3), (2,2) and (3,2), where (0,1) represented the loss of a single copy, (0,2) and (1,1) represented normal and (1,2), (1,3), (2,2) and (3,2) represented the gain of one, two or three copies.

Haplotype Map data

We started the analysis with 98 Affymetrix 6.0 SNP array profiles of healthy patients from the publicly available Haplotype Map (HapMap) repository: ftp://ftp.ncbi.nlm.nih.gov/hapmap/ [4]. We preprocessed the data with the Aroma Affymetrix Power Tools package [30] and the PennCNV-Affy pipeline [31]. In the preprocessing step, we performed quantile normalisation and generated genotype calls from the Affymetrix spot intensity readout files (CEL format) as output by the Affymetrix microarray scanner files using the Birdseed algorithm [32]. Next, we extracted allele-specific signals, and we calculated the canonical clustering parameters for each marker of genetic variation. We then calculated probe-wise LRR and BAF for each patient sample. Further, we split the signal file into individual files for each patient. We then selected 81 patients that were further experimentally profiled by Redon *et al.* [4].

HNSCC data

We used Level 1 Affymetrix SNP 6.0 array data generated by the TCGA research network (http://cancergenome.nih.gov/) consisting of 522 samples collected from patients suffering from HNSCC [26]. We preprocessed the tumour and normal matched raw HNSCC CEL files with the Affymetrix Power Tools package [30] and the PennCNV-Affy pipeline [31] as described in the previous section.

Genomic copy number calling algorithms

We selected five CNA calling algorithms for comparison: CGHCall (release 3.6), OncoSNP (version 2.1), ASCAT (version 2.4), genoCNA and GISTIC (version 2.0).

OncoSNP

OncoSNP was built upon a statistical model that classifies SNP array signals—both LRR and BAF, from cancer genomes into 21 states covering different combinations of allele loss and amplification. The model includes effects of polyploidy, tumour purity and intra-tumour heterogeneity [19]. We applied OncoSNP on the synthetic data with the arguments specific for Affymetrix SNP array, together with the predefined number of training states and tumour states. We used the intratumour mode and set the tumour purity paramater to 30, 50, 70 and 100% . For the HapMap data, we used the same parameter settings, except for the tumour purity which was set to 0.

ASCAT

ASCAT was designed to perform allele-specific CNA analysis in tumour samples. The algorithm corrects for the effects of tumour purity and tumour aneuploidy and infers copy number classes, loss of heterozygosity and homozygous deletions. ASCAT estimates the number of copies for both alleles at all SNP marker positions together with the tumour purity of each sample [6].

We preprocessed the synthetic data and generated the ASCAT-format input tumour LRR and BAF files. Afterwards, we generated corresponding germline genotypes with the ascat.predictGermlineGenotypes R function with the platform parameters set to 'AffySNP6'. Finally, we segmented the data with the ASPCF segmentation algorithm and applied the ASCAT copy number calling function. Next, we applied the same steps to the HapMap data.

GenoCNA

GenoCN was built as a statistical framework that simultaneously searches for CNAs and CNVs while inferring the tumour purity. In this study we used the genoCNA component, which was specifically designed for CNA finding. Applying genoCNA required the following information for each of the genetic markers: name, chromosome, position and population frequency (PFB). We used the genetic marker information as provided by the Affymetrix PFB file corresponding to the human genome assembly hg18. Each input file contained LRR, BAF and PFB values for each genetic marker. We selected the output format 2 which returned the most likely copy number and genotype state of all the genetic markers.

GISTIC

GISTIC was designed to find regions of the genome that are significantly amplified or deleted across a set of samples. The significance measure is based on the amplitude of the CNA, on how frequently the CNA occurs across samples and a user-defined threshold for the discovery rate. GISTIC required as input a segmentation file, a reference genome file and the LRR signals. GISTIC does not use the BAF signals. For all data sets we used the hg18 reference genome and segmentation files obtained by applying circular binary segmentation—further referred to as CBS [33]. For the TCGA HNSCC analysis we used the GISTIC results provided by TCGA as level 3 data.

CGHcall

CGHcall was originally designed for array Comparative Genomic Hybridization (aCGH) data. The algorithm uses breakpoint information from CBS [33] and classifies raw $\log_2$-ratios between reference and tumour DNA into five discrete states: double loss-homozygous (biallelic) deletion, loss-hemizygous deletion (loss of one of the alleles), normal-two copies, gain-three to four copies and amplification–more than four copies [20]. We log-transformed the total copy numbers and we applied the CGHcall

pipeline on resulting signals with adjustment for tumour purity. For the HNSCC TCGA data set, we implemented a Python script to calculate $\log_2$-ratios between tumour and normal matched patient samples. As the HapMap cohort included only healthy patients, we calculated $\log_2$-ratios between each LRR signal and the mean LRR signal of the 81 selected samples.

*CGHcall**

We developed an adjusted version of CGHcall to prevent shifts of the baseline level after global normalisation: CGHcall*. We adjusted the normalisation and post-segmentation normalisation for samples in which the CNA burden exceeded 50% of the sample profile, by considering only the signals included in the [-0.1, 0.1] interval (see Section 3.1). We applied the CGHcall* pipeline on the synthetic data and on the HapMap as described in the previous section for CGHcall. Further, we applied CGHcall* on the $\log_2$-ratios between tumour and normal matched TCGA HNSCC samples. When running CGHcall and CGHcall* on the TCGA HNSCC data, we set the tumour purity parameter to the consensus measurement of TCGA HNSCC estimations derived by Aran *et al*. [34]. For samples with missing derived consensus measurement estimations, we used the immunohistochemistry measurements as tumour purity values.

Performance analysis of genomic copy number calling algorithms

For evaluating the performance of the selected algorithms, we collapsed the resulting calls to three states: loss, normal and gain. For CGHcall, CGHcall* and GISTIC the double loss and loss were collapsed to loss, while the gain and amplification were collapsed to gain. For OncoSNP we collapsed the homozygous and the hemizygous deletion states to loss, and all the states that were defined by more than two copies were considered gain. For ASCAT and genoCNA, the probes with less than two copies were defined as lost, while the probes with more than two copies were defined as gained. We calculated the sample-wise confusion matrix, precision, recall and balanced F-score [35] as follows:

$$\text{precision}_c = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{recall}_c = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$F_c = 2 \cdot \frac{\text{precision}_c \cdot \text{recall}_c}{\text{precision}_c + \text{recall}_c}, \tag{3}$$

where c represented the class: loss, normal or gain. True positives (TP) represent the number of probes that were classified correctly for each class c, while false positives (FP) are the probes classified incorrectly as class c. False negatives (FN) represent the number of probes that belong to class c but were classified as belonging to another class. To test for statistically significant shifts between F-score distributions of the algorithms, we performed non-parametric pairwise comparison Wilcoxon tests [36]. We adjusted the resulting P-values for multiple testing error through Bonferroni correction [37].

Next, we assessed the performance of the CNA calling algorithms on the Affymetrix SNP 6.0 HapMap samples with matched experimentally genomic copy number validated results. Finally, we analysed the results of the CNA calling algorithms on the TCGA HNSCC Affymetrix SNP 6.0 samples in HNSCC consensus regions with focus on the Cyclin D1 (CCND1) and the cyclin dependent kinase inhibitor 2A (CDKN2A) genes.

Results and discussion

Characterising molecular phenotypes in cancer research requires the accurate identification of DNA copy number changes. Although genomics increasingly deploys genome sequencing, there is still a wealth of cost-effective SNP array data available. Thus, making use of these data is important and requires best possible analysis approaches that, among other features, are able to correct for cancer-specific confounding variables such as tumour purity and a wide range of CNA lengths. To benchmark commonly used CNA calling approaches in the presence of such confounding variables, we developed an evaluation pipeline.

To evaluate the CNA algorithms, tumour samples with known true states are required. Since the true copy number states for real cancer data are unknown and experimental validation on genome-wide level is not feasible (the human genome size is about 3.0×10^9 bp and is affected by CNVs), we assessed the performance of the algorithms using synthetic data mimicking Affymetrix SNP 6.0 array experiments (see Methods for details). To make the samples as similar as possible to the real Affymetrix SNP 6.0 array samples, we simulated data for 1.844.399 markers of genetic variation—number of probes comparable to the one present on an Affymetrix SNP 6.0 array. Subsequently, we evaluated the performance of OncoSNP, ASCAT, GenoCNA, CGHcall and GISTIC at SNP level resolution.

When conducting a benchmarking study, in addition to realistic synthetic data, we need to use an appropriate measure for the performance of copy number calling algorithms. In general, to show how prediction algorithms perform, receiver operating characteristics (ROC) curves are commonly used [38]. However, when the distribution of the classes is imbalanced, as in our case (Figure S1), ROC curves can present an over-optimistic view on how an algorithm performs, while the recall and the precision have been shown to give a more informative view [39, 40]. Since the F-score represents the balance between the precision and the recall of an algorithm, we selected it as an appropriate criteria and used it to evaluate the performance of the copy number algorithms for each class. The F-score allowed us to determine the algorithm that classified correctly genomic regions independently of the CNA type. This is of great importance, since for a putative future use in personalised medicine, classifying correctly regions overlapping oncogenes or tumour suppressor genes may affect the diagnosis and, thus, the treatment of a patient.

We were interested whether the investigated algorithms can classify precisely the LRR and the BAF signals on probe level into three classes: loss, normal and gain. Therefore, we split the multi-class classification problem into three binary classification problems.

An improved algorithm for copy number calling from Affymetrix SNP 6.0 data: CGHcall*

During manual inspection of the CGHcall pipeline we observed that the normalised signals before and after segmentation in the synthetic samples with more 50% non-normal states covering the sample profiles were incorrectly shifted (either to -1, either to 1). This led to defining an incorrect baseline level in these samples and thus, calling the wrong copy number state. Since cases in which more than half of the genotyped probes are in

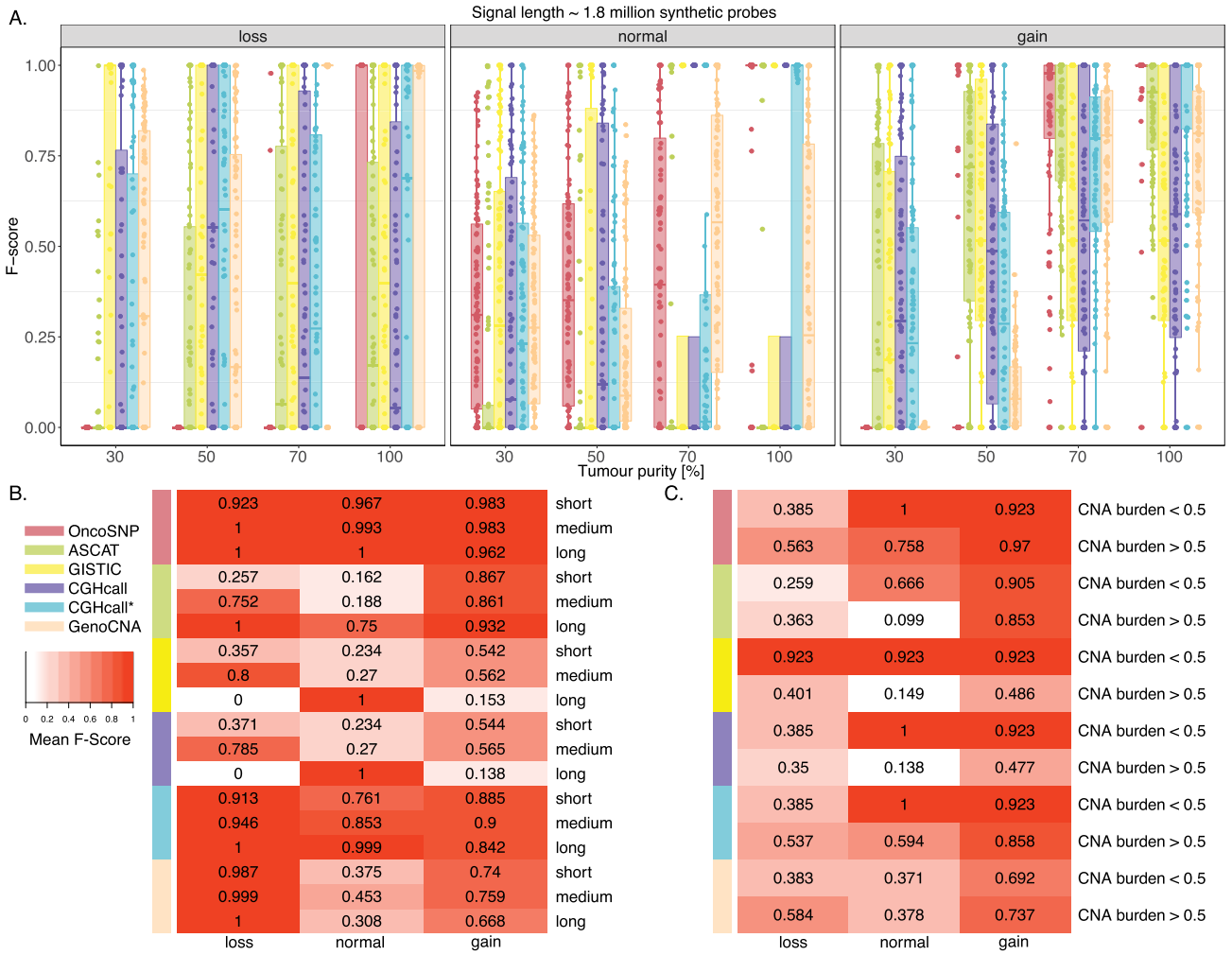

Figure 1. Performance of CNA calling algorithms on synthetic data. We evaluated the performance of six algorithms which are colour-coded as it follows: OncoSNP, coral red; ASCAT, light green; CGHcall, purple; CGHcall*, cyan; GenoCNA, pale pink brown; and GISTIC, yellow. (**A**). The y-axis represents the F-score and x-axis represents the tumour purity level in %. The three facets represent the different classes: loss, normal and gain. Each boxplot consists of F-scores for 100 synthetic samples. The total number of genetic markers covered by the synthetic signal was approximately 1:8 bp. (**B**). Heatmap of mean F-scores for different lengths of copy number regions. (**C**). Heatmap of mean F-scores for samples with CNA burden ratio < 0.5 versus samples with CNA burden ratio > 0.5.

a non-normal state have already been reported in a pan-cancer study on somatic genomic CNAs [14], we set up to correct for the CNA burden effect.

The problem arose from the LRR levels being normalised to the median level over a sample. If more than half of the genome is changed in one direction (loss or gain), CGHcall is unable to correctly estimate the baseline level and assigns the 0 level to what is actually lost or gained. We observed the same behaviour when we applied the post-segmentation normalisation, which assigns the baseline segment to a segment that is either lost or gained. To correct for this effect, we selected three different intervals as constrains for the LRR signals, [-0.1, 0.1], [-0.05, 0.05] and [-0.2, 0.2], and analysed how the performance of the algorithm changes in samples with 100% tumour purity. The resulting F-scores suggested that the LRR signals within the [-0.1, 0.1] interval provided the optimal mean for normalisation and post-segmentation normalisation (Figure S2). As a result, we proposed a solution in which, instead of performing normalisation and post-segmentation normalisation based on all LRR signals, we limit ourselves to LRR signals that fall in the [-0.1, 0.1] interval.

Tumour purity showed strong influence on performance

We first analysed how different tumour purities influenced the performance of the algorithms on synthetic data. We compared the algorithms based on their F-score distributions (Figure 1A). We first showed how the six algorithms (OncoSNP, red; ASCAT, neon green; GISTIC, yellow; CGHcall, purple; CGhcall*, cyan; and GenoCNA, pale orange) identified losses at tumour purity levels (depicted on the x-axis) varying from 30 to 100% (Figure 1A, left panel). OncoSNP was not able to identify losses in samples with tumour purity $< 100\%$ (mean F-score $= 0.03$). ASCAT, GISTIC, CGHcall and CGHcall* showed poor performance when calling losses independent of the tumour purity level (mean ASCAT F-score $= 0.26$, mean GISTIC F-score $= 0.34$, mean CGHcall F-score $= 0.39$, mean CGHcall* F-score $= 0.51$). GenoCNA showed good performance for correctly calling losses in samples with tumour purities $> 50\%$ (mean F-score $= 0.68$). Thus, the performance of CGHcall* and GenoCNA for calling losses increased with the tumour purity.

OncoSNP showed increasing performance for calling normal states as the tumour purity level increased (Figure 1A, middle panel). This may be caused by the log2 ratios being pushed towards the 0 baseline in the presence of normal DNA. Moreover, since the normal state represented the majority class, the improved F-score for OncoSNP when calling normal states suggested that the algorithm may not be able to tackle the imbalance of the classes—represented by the copy number states. ASCAT was unable to classify correctly normal states independent of the tumour puritity. GISTIC, CGHcall and GenoCNA showed poor performance when trying to classify normal states (mean GISTIC F-score = 0.28, mean CGHcall F-score = 0.29, mean GenoCNA F-score = 0.35). CGHcall* showed overall good performance in correctly finding the normal state when compared to the other three algorithms in samples with tumour purity 100% (mean F-score = 0.70, Figure 1A, middle panel).

Next, we compared how the algorithms performed when trying to identify gains (Figure 1A, right panel). OncoSNP showed good performance when the tumour purity was > 50%. This suggests that OncoSNP is not able to correct the effect of tumour contamination > 50% on the signals in gained genomic regions. The performance of all algorithms for calling gains increased as the tumour purity increased. ASCAT was the only algorithm able to correctly call gains in samples with tumour purities > 30% (mean F-score = 0.76). Overall, our adjusted version of CGHcall–CGHcall* showed improved performance with regard to all copy number states and all tumour purities when compared to CGHcall. GISTIC and CGHcall showed comparable results. This can be explained by the fact that both algorithms use CBS segmentation results and do not make use of the BAF. Our analysis suggested that OncoSNP and CGHcall* handled calling CNAs better than the other algorithms in samples with high tumour purities. The main message of this analysis is that tumour purity strongly influences the results of the CNA calling algorithms.This is an important information to be considered in designing a CNA study, since samples with tumour purities markedly below 50% should not be included in the analysis or at least, profiles resulting from such samples should be handled with care.

The effect of copy number region length

Next, we aimed to understand how the length of a copy number region influenced the performance of the calling algorithms. For this purpose, we examined the difference between the mean F-scores of samples with region lengths of $\leq 10^5$probes (short), between 10^5 and 10^6 probes (medium) and region lengths > 10^6 (long) (Figure 1B). In order to eliminate the effect of reduced tumour purity, we selected only samples with 100% tumour purity. The region length was equal to the number of genetic markers with the same copy number state within a chromosomal segment. One chromosomal segment covered from 3 kilo base pairs (kbp) to 1.8 million base pairs (Mbp).

We observed that OncoSNP, GenoCNA and CGHcall* showed little sensitivity to the length of copy number regions. While CGHcall* and OncoSNP performed well for all three states, GenoCNA had difficulty in correctly identifying normal genomic regions. ASCAT performed worse in samples that included short- and medium-length CNA regions than in samples containing long CNA regions. GISTIC was not able to correctly find lost or amplified genomic regions independent of the length. We observed the same behaviour for CGHcall. One reason that may lay at the core of this problem is the fact that both CGHcall and GISTIC use the CBS algorithm. In all, OncoSNP and CGHcall* showed consistency and performed well for all three copy number states across the investigated ranges of copy number region lengths.

The effect of CNA burden

Since we observed that the percentage of aberrated regions in a tumour sample—CNA burden—affected the normalisation of the log2 ratios in the CGHcall pipeline, we investigated whether we observe a similar effect when applying the other copy number calling algorithms.

We therefore grouped the synthetic data into samples with CNA burden > 50% and samples with CNA burden < 50% and calculated the mean F-scores statewise (Figure 1C). We observed that both CGHcall and GISTIC performed poorly for samples with CNA burden > 50%. ASCAT also showed decreased performance for the same scenario, but only for the normal state. The performance of CGHcall* increased in samples with CNA burden > 50% when compared to CGHcall, confirming that we corrected the inaccuracy from CGHcall, especially for predicting normal and gained genomic regions. OncoSNP and CGHcall* were again the best performing algorithms included in this study.

Performance of the copy number calling algorithms on SNP 6.0 array profiles of healthy patients (HapMap)

To assess how OncoSNP, ASCAT, CGHcall, CGHcall*, GenoCNA and GISTIC perform on real data, we would need a gold standard. Due to the size of human genome – 3.0×10^9 bp, we lack a complete Affymetrix SNP 6.0 array gold standard. Since the HapMap project subsequently experimentally validated the CNAs determined from Affymetrix SNP 6.0 data, we defined the copy number profiles annotated by Redon *et al.* [4] as our 'gold standard'. OncoSNP, ASCAT, CGHcall, CGHcall* and genoCNA returned predictions for over 14.500 regions that overlapped the 'gold standard'. When analysing the F-scores of the algorithms corresponding to 81 profiles with matched annotated copy number profiles (Figure 2), we first observed that OncoSNP, ASCAT, CGHcall, CGHcall* and genoCNA performed well for the normal class (mean F-score = 0.91). Unlike the other algorithms, GISTIC returned predictions for only 381 regions overlapping the 'gold standard' and performed poorly for all the classes (mean F-score

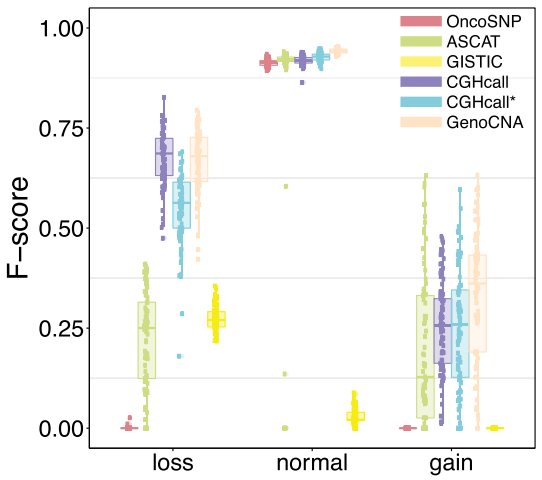

Figure 2. Distribution of F-scores for OncoSNP, ASCAT, CGHcall, CGHcall*, GenoCNA and GISTIC in 81 healthy HapMap subjects.

= 0.10). OncoSNP could not identify any germline alterations. ASCAT showed a poor performance for identifying gains and losses (mean F-score = 0.20). CGHcall showed a mean F-score of 0.67 for identifying losses, but performed poorly for identifying gains (mean F-score = 0.25). CGHcall* showed a significant improvement only for the normal class compared to the other algorithms. GenoCNA performed best for identifying losses and gains, mean F-score = 0.50. ASCAT, just as OncoSNP and GISTIC, was implemented to find somatic CNAs in cancer samples and was not designed to find germline alterations in the first place. We hypothesise that this might be the reason why OncoSNP, ASCAT and GISTIC perform poorly on healthy patient data.

We are aware that tumour data tailored genomic copy number algorithms are designed to consider CNAs deriving from tumour cell populations. However, HapMap data were generated from blood cells. The genomic copy number changes to be expected from these samples are germline. Therefore, all cells analysed should contain the same alterations. We assume that it would be 'easier' for a tumour data tailored algorithm to pick up copy number changes. The genomic copy number changes present in the HapMap samples were comprehensively experimentally validated. Thereby, HapMap provides added value since the 'gold standard' with regard to genomic copy number is known for these samples and allowed us to calculate the performance of the CNA calling algorithms on real data. Based on the resulting F-scores, genoCNA, CGHcall and CGHcall* were the best performing algorithms.

CNAs in HNSCC

To test the plausability of CNA calling results in tumour samples, we explored the concordance between raw LRR signals from TCGA HNSCC samples and the CNA calls of the six algorithms. Additionally, we compared the results with the HNSCC-specific CNA regions defined in Gollin *et. al* [41]. We focused on two genes: one known to be amplified in HNSCC–CCND1 and one that is known to be lost in HNSCC–CDKN2A (Figures 3 and 4).

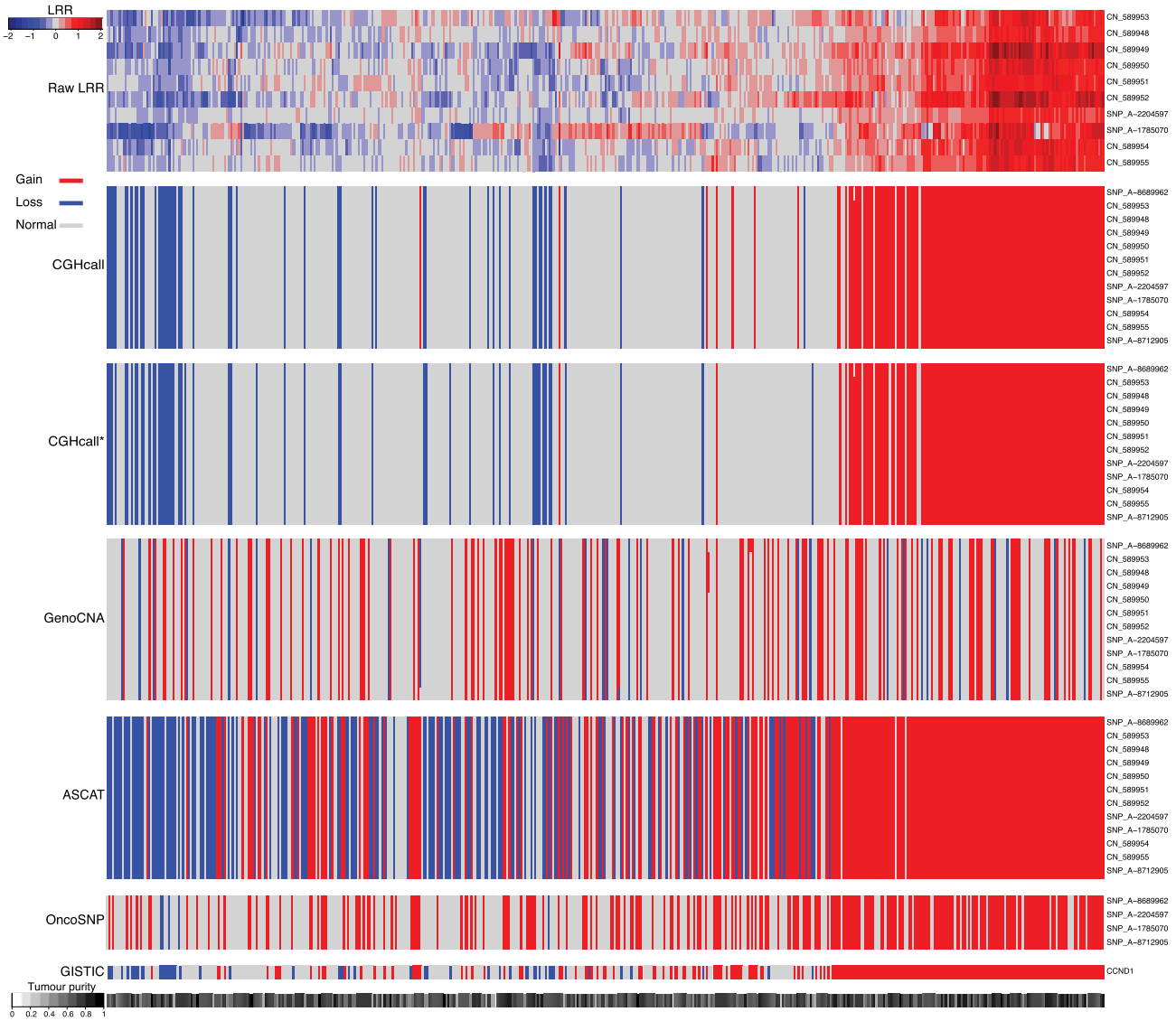

Figure 3. CCND1: Concordance between raw data and algorithm calls in TCGA HNSCC. The heatmap columns represent patients clustered by raw LRR signals. The rows represent the Affymetrix SNP 6.0 probes that overlap the CCND1 region. For CGhcall*, CGHcall, GenoCNA, ASCAT and OncoSNP we also include the neighbouring probe sets of the overlapping region. The lower bar represents the tumour purity of each sample.

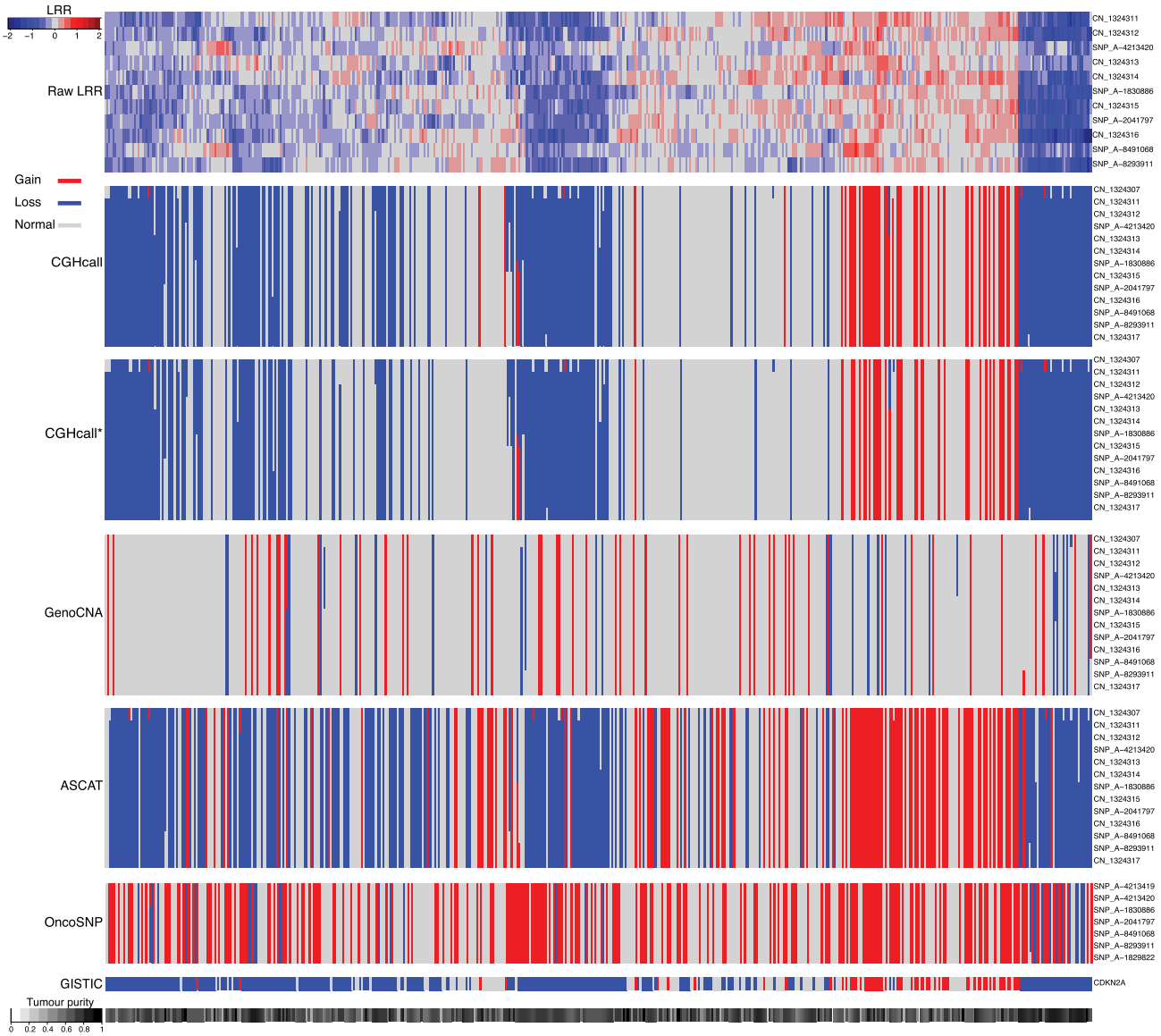

Figure 4. CDKN2A: Concordance between raw data and algorithm calls in TCGA HNSCC. The heatmap columns represent patients clustered by raw LRR signals in the probes overlapping the CDKN2A genomic region. The rows represent the Affymetrix SNP 6.0 probes that overlap the CDKN2A region. For CGhcall, CGHcall, GenoCNA, ASCAT and OncoSNP we also include the neighbouring probe sets of the overlapping region. The lower bar represents the tumour purity of each sample.

The data presented in Figures 3 and 4 show that genomic regions with high LRR values overlapping the CCND1 and CDKN2A genes are called as gained, while genomic regions with low LRR values overlapping the CCND1 and CDKN2A genes are called as lost. The frequencies of CCND1 gains called by CGHcall, CGHcall*, OncoSNP and GISTIC are comparable to the frequencies of CCND1 gains reported from CGH data in Gollin *et. al* [41], 32%; CGHcall, 26.5%; CGHcall*, 24.9%; OncoSNP, 44%; and GISTIC, 43%. CGHcall, CGHcall*, OncoSNP and GISTIC showed a good overlap in frequencies of CDKN2A losses: CGHcall, 39.8%; CGHcall*, 35.4%; and GISTIC, 59%. The tumour purity ranged from 27.9 to 97.7%. Most of the samples present tumour purity > 60%. These results indicate that in a realistic tumour purity range the algorithms that best performed on synthetic data CGHcall* and OncoSNP showed plausible results in the TCGA HNSCC data as well.

Concluding remarks

Within our study we addressed the problem of evaluating the performance of commonly used copy number calling algorithms in the presence of cancer-specific confounding variables. Since we lacked a complete Affymetrix SNP 6.0 array gold standard, we provided a pipeline to evaluate CNA calling algorithms on Affymetrix SNP 6.0 array-like synthetic data. The analysis on the synthetic data revealed that the performance of the CNA calling algorithms is strongly influenced by tumour purity. CGHcall, GISTIC and ASCAT showed high sensitivity to the length of the genomic segments. The CNA burden strongly influenced the performance of ASCAT, GISTIC and CGHcall. We proposed CGHcall*, an adjusted version of CGHcall, in which we correct for the effect of the CNA burden and we showed that indeed the performance of CGHcall* in samples with a CNA burden higher

than 50%. However, the scope of our paper was to benchmark commonly used CNA calling algorithms, and not to develop a new algorithm.

We further evaluated how the algorithms performed on a real data set comprising of 81 healthy patients HapMap samples that were subsequently experimentally validated. CGHcall and CGHcall* were able to detect germline alterations, unlike OncoSNP and ASCAT. Finally, we examined how comparable were the results of the CNA calling algorithms with the annotated CNAs in CCND1 and CDKN2A, when evaluated on the TCGA HNSCC data set. The results indicated that CGHcall, CGHcall* and GISTIC return comparable calls to what has been reported so far.

In conclusion, we provided a benchmarking pipeline for CNA calling algorithms from Affymetrix SNP 6.0 array tumour profiles together with CGHcall*—an adjusted version of CGHcall for finding CNAs in highly variant genomes.

Key Points

- CNAs are tumour-specific DNA changes that play an important role in cancer research.
- The accurate identification of CNAs is affected by biological confounding variables like tumour purity, the length of a chromosomal segment and the percentage of CNAs present in a genome.
- Within this benchmarking study we provide a pipeline through which we evaluated the performance of six CNA calling algorithms (OncoSNP, ASCAT, CGHcall, CGHcall*, GenoCNA and GISTIC) in the presence of biological confounding variables.
- We provide an adjusted version of CGHcall–CGHcall* that accounts for a high CNA burden.
- We identify tumour purity and CNA burden to significantly influence the performance of all the CNA calling algorithms.

Supplementary Data

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

Acknowledgments

The authors would like to thank Gökcen Eraslan, Richa Batra, Linda Krause, Michael Strasser and Michael Laimighofer for helpful discussions and feedback.

Funding

German Federal Ministry of Education and Research (BMBF) (02NUK045A).

References

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* April 2009;**458**(7239):719–24.
2. Pleasance ED, Cheetham RK, Stephens PJ, *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010;**463**(7278):191–6.
3. Beroukhim R, Mermel CH, Porter D, *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* 2010;**463**(7283):899–905.
4. Redon R, Ishikawa S, Fitch KR, *et al.* Global variation in copy number in the human genome. *Nature* 2006; **444**(7118):444–54.
5. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet* 2015; **16**(3):172–83.
6. Van Loo P, Nordgard SH, Lingaerde OC, *et al.* Allele-specific copy number analysis of tumors. *PNAS* 2010;**107**(39):16910–5.
7. Bardeesy N, Cheng K-H, Berger JH, *et al.* Smad4 is dispensable for normal pancreas development yet critical in progression and tumor biology of pancreas cancer. *Genes Dev* 2006; **20**(22):3130–46.
8. Witkiewicz AK, McMillan EA, Balaji U, *et al.* Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat Commun* 2015;**6**:6744.
9. Leucci E, Vendramin R, Spinazzi M, *et al.* Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* 2016; **531**(7595):518–22.
10. Wells MF, Wimmer RD, Schmitt LI, *et al.* Thalamic reticular impairment underlies attention deficit in Ptchd1(y/-)mice. *Nature* 2016;**532**(7597):58–63.
11. Cibulskis K, Lawrence MS, Carter SL, *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotech* 2013;**31**(3):213–9.
12. Y, Zhao L, Wang Y, *et al.* SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data. *BMC Bioinformatics* 2017;**18**:147.
13. Zhang X, Du R, Li S, *et al.* Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinformatics* 2014; **15**:50.
14. Zack TI, Schumacher SE, Carter SL, *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 2013; **45**(10):1134–40.
15. Zhou W, Zhao Z, Wang R, *et al.* Identification of driver copy number alterations in diverse cancer types and application in drug repositioning. *Mol Oncol* 2017;**11**(10):1459–74.
16. Carter SL, Cibulskis K, Helman E, *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotech* 2012;**30**(5):413–21.
17. Cai TT, Jeng XJ, Li H. Robust detection and identification of sparse segments in ultrahigh dimensional data analysis. *J Roy Stat Soc Ser B Stat Methodol* 2012;**74**(5):773–97.
18. Sun W, Wright FA, Tang Z, *et al.* Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res* 2009;**37**(16):5365–77.
19. Yau C, Mouradov D, Jorissen RN, *et al.* A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol* 2010;**11**:R92.
20. van de Wiel MA, Kim KI, Vosse SJ, *et al.* CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics* 2007;**23**(7):892–4.
21. Mermel CH, Schumacher SE, Hill B, *et al.* GISTIC 2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;**12**(4):R41.
22. Metzger J, Philipp U, Lopes MS, *et al.* Analysis of copy number variants by three detection algorithms and their association with body size in horses. *BMC Genomics* 2013; **14**:487.

23. Mosén-Ansorena D, Aransay A, Rodríguez-Ezpeleta N. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC Bioinformatics* 2012;**13**:192.
24. Pierre-Jean M, Rigaill G, Neuvial P. Performance evaluation of DNA copy number segmentation methods. *Brief Bioinform* 2015;**16**(4):600–15.
25. Hieronymus H, Schultz N, Gopalan A, *et al*. Copy number alteration burden predicts prostate cancer relapse. *PNAS* 2014;**111**(30):11139–44.
26. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 2015;**517**(7536):576–82
27. Lin C-F, Naj AC, Wang L-S. Analyzing copy number variation using SNP array data: protocols for calling CNV and association test. *Curr Protoc Hum Genet* 2013;**79**:Unit–1.27.
28. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;**12**(5):363–76.
29. Rasmussen M, Sundström M, Kultima HG, *et al*. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol* 2011;**12**:R108.
30. Lockstone HE. Exon array data analysis using Affymetrix power tools and R statistical software. *Brief Bioinform* 2011;**12**(6):634–44.
31. 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.
32. Korn JM, Kuruvilla FG, McCarroll SA, *et al*. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 2008;**40**(10):1253–60.
33. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostat* 2004;**5**(4):557–72.
34. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun* 2015;**6**:8971.
35. Van Rijsbergen CJ. *Information Retrieval*, 2nd edn. Newton, MA, USA: Butterworth-Heinemann, 1979.
36. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull* 1945;**1**(6):80–3.
37. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni R Istituto Superiore Scienze Economiche Commerciali Firenze* 1936;**8**:3–62.
38. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett* 2006;**27**(8):861–74.
39. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*, *ICML '06*. pp. 233–40. ACM, New York, NY, USA, 2006.
40. Lever J, Krzywinski M, Altman N. Points of significance: classification evaluation. *Nat Meth* 2016;**13**(8):603–4.
41. Gollin SM. Cytogenetic alterations and their molecular genetic correlates in head and neck squamous cell carcinoma: a next generation window to the biology of disease. *Genes Chromosomes Cancer* 2014;**53**(12):972–90.