BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 14 2013, pages 1793–1800
doi:10.1093/bioinformatics/btt300

Data and text mining Advance Access publication May 28, 2013

A method for finding consensus breakpoints in the cancer genome from copy number data

Laura Toloşi[1,*], Jessica Theißen[2], Konstantin Halachev[1], Barbara Hero[2], Frank Berthold[2] and Thomas Lengauer[1]

[1]Department of Computational Biology and Applied Algorithmics, Max-Planck-Institute for Informatics, Campus E1.4, 66123 Saarbrücken, Germany and [2]Department of Pediatric Oncology and Hematology, University Children's Hospital, Kerpener Straße 62, 50924 Cologne, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Recurrent DNA breakpoints in cancer genomes indicate the presence of critical functional elements for tumor development. Identifying them can help determine new therapeutic targets. High-dimensional DNA microarray experiments like arrayCGH afford the identification of DNA copy number breakpoints with high precision, offering a solid basis for computational estimation of recurrent breakpoint locations.

Results: We introduce a method for identification of recurrent breakpoints (consensus breakpoints) from copy number aberration datasets. The method is based on weighted kernel counting of breakpoints around genomic locations. Counts larger than expected by chance are considered significant. We show that the consensus breakpoints facilitate consensus segmentation of the samples. We apply our method to three arrayCGH datasets and show that by using consensus segmentation we achieve significant dimension reduction, which is useful for the task of prediction of tumor phenotype based on copy number data. We use our approach for classification of neuroblastoma tumors from different age groups and confirm the recent recommendation for the choice of age cut-off for differential treatment of 18 months. We also investigate the (epi)genetic properties at consensus breakpoint locations for seven datasets and show enrichment in overlap with important functional genomic regions.

Availability: Implementation in R of our approach can be found at http://www.mpi-inf.mpg.de/ ~laura/FeatureGrouping.html.

Contact: laura@mpi-inf.mpg.de.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 29, 2012; revised on April 28, 2013; accepted on May 22, 2013

1 INTRODUCTION

DNA breaks occur often during cell replication and are especially dangerous if the DNA repair mechanisms do not succeed in restoring the damaged locus correctly. Erroneous DNA repair events can disrupt essential mechanisms that maintain the integrity of the cell and, on rare occasions, can cause the cell to become cancerous. Insertions, deletions, amplifications, translocations and inversions are types of DNA structural damage frequently observed in cancer genomes that occur as a consequence of unrepaired DNA breaks.

Tumor genomes exhibit hundreds of DNA breakpoints. Some of them promote tumor development by disrupting crucial functional elements, and thus survive the selective pressure that acts on cancer cells. These breakpoints tend to recur in tumors of the same type, suggesting common progression mechanisms. Other breakpoints may have no or less effect on tumor progression, e.g. the contribution of non-recurrent breakpoints that appear at arbitrary locations in the genome is difficult to evaluate. Identifying relevant breakpoint hotspots and characterizing their influence on tumor phenotype can help to identify molecular factors responsible for tumor progression and thus to develop new therapeutic strategies.

Very high-resolution microarray technologies for genome-wide measurement of copy number alterations, such as arrayCGH (Pinkel *et al.*, 1998; Solinas-Toldo *et al.*, 1997) or SNP arrays (Sherry *et al.*, 2001) can reveal DNA breakpoints related to copy-number changes in cancer genomes. The resolution of microarray-based technologies can be very high, with the distance between probes being as little as 2–3 kb, facilitating sufficiently accurate estimation of the location of a DNA breakpoint. The estimation is carried out by performant segmentation algorithms, such as CBS (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007). Presently, large collections of high-resolution arrayCGH data are publicly available at the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/), The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) and other repositories, for many types of cancer.

In this article, we introduce an algorithm for identifying recurrent breakpoints (or consensus breakpoints) in tumor cohorts, based on DNA copy number aberration data. The approach is called *Consensus breakpoints by Kernel Smoothing* (C-KS, pronounce as 'seeks'). C-KS identifies genomic locations around which breakpoints tend to accumulate more frequently than expected by chance and assigns to each a significance *Z*-score. The input of our algorithm consists of the locations of breakpoints of all tumors in the cohort; therefore, the computational running time is independent of the resolution of the experimental assays that produced the data.

In Ritz *et al.* (2011), a competing approach for identifying breakpoint hotspots called Neighborhood Breakpoint Conservation (NBC) is presented. NBC is applied directly to

*To whom correspondence should be addressed.

the raw log-ratios of arrayCGH experiments. The authors devise a special single-array segmentation algorithm, which assigns to each pair of adjacent probes a probability of a breakpoint occurring between them. Then, locations of recurrent breakpoints are analytically estimated and sorted by significance (*P*-value). NBC has two main shortcomings. First, the segmentation models are based on parametric assumptions that are specific to arrayCGH data and, therefore, not applicable to next-generation sequencing. Second, the complexity of the NBC algorithm is high, the segmentation step using a dynamic program that requires quadratic runtime in the number of array probes. Therefore, NBC does not scale well to the most recent microarray platforms, such as Agilent 1M or Affymetrix SNP 6.0.

By identifying consensus breakpoints, our approach can afford *consensus segmentation* of a set of cancer genomes, which is a generalization of the single-sample segmentation procedure. We assume that between two consecutive consensus breakpoints, each tumor has constant copy number, given, for example, by the average log-ratio (in the case of microarrays). We call these regions *consensus regions*.

To validate the C-KS algorithm, we use the consensus regions as super-features in classification tasks that predict some binary indicator of tumor phenotype. We applied our method to seven arrayCGH datasets from five cancer types with various phenotypical annotations. We show that in comparison with models trained on probe data, the accuracy of the classification models using the consensus regions is comparable and sometimes higher. Hence, disregarding all breakpoints that are not consensus breakpoints does not adversely impact the predictive power of the model. We also show that consensus segmentation affords substantial dimension reduction of the sample data without significant loss of information.

Moreover, we present an important biological finding that arose from our classification models using C-KS segmentation. Specifically, the tumor genomes of neuroblastoma patients of different age groups are most distinct when the age grouping is defined as '<18 months' against '>18 months'. Our finding has potential impact on differential therapy of neuroblastoma patients, which are currently divided into two prognostic groups: <12 months at diagnosis (good prognosis) and older (poor prognosis). Our result is also supported by other studies (Evans and D'Angio, 2005; London *et al.*, 2005; Schmidt *et al.*, 2005).

For a qualitative validation of the C-KS algorithm, we investigate genomic and epigenomic properties of the genomic locations corresponding to consensus breakpoints and show that they tend to be enriched in functional elements and certain DNA sequence patterns. These associations are interesting and deserve further biological investigation.

2 METHODS

Given are N samples (tumors) and corresponding copy number measurements at p loci, in the form of the logarithm of the ratio between the tumor copy number and the normal copy number (log-ratios). We assume that the samples have been segmented using some procedure that identifies breakpoints and reports segments of constant copy number between them, without calling gains and losses. Examples of such procedures include CBS (Olshen *et al.*, 2004) and GLAD (Hupé *et al.*, 2004). As a consequence, the log-ratio sequence of each sample can be partitioned into a sequence of intervals of constant copy number. A *breakpoint* in this context is defined as the genomic location that marks the boundary between two adjacent intervals of distinct copy number in one particular sample.

Let V be the set of breakpoints observed in the N arrays, represented as triples as $V = \{(v_i, s_i, w_i) \mid 1 \leq i \leq T\}$, where T is the total number of breakpoints, v_i is the genomic location of the i^{th} breakpoint, s_i is the index of the sample on which the i^{th} breakpoint was observed, $s_i \in \{1, \dots, N\}$ and w_i is the difference in copy number at breakpoint v_i (from right to left) and it is called its weight. The weight w_i can be negative or positive, depending on whether the copy number decreases or increases, respectively. The start position of each chromosome is considered a natural breakpoint for each sample, and it is a member of V. Because at the start of the chromosome, there is no proper copy number change, the weight of this breakpoint is considered zero.

In this manuscript, we say that a genomic location B is a *consensus breakpoint* if around location B, more breakpoints occur in the set of samples than expected by chance.

2.1 Algorithm C-KS

We present an approach called C-KS (consensus breakpoints via kernel smoothing) for identifying consensus breakpoints in the set of N samples. C-KS uses a kernel smoothing technique for identifying genomic locations around which unexpectedly large accumulations of breakpoints occur. By sliding a location pointer along the genomic sequence, we observe the breakpoints located within the vicinity of the current location and estimate the likelihood of this observation under a null model (Z-score). The null model assumes that the locations of the breakpoints of each array are uniformly distributed along the genomic sequence and do not depend on the array. The locations at which the estimated Z-score is large enough are reported as candidates for consensus breakpoints.

Formally, let x be a location on the genome and $\mathcal{K}(\cdot;\, \mu, \sigma)$ be a Gaussian kernel with mean μ and standard deviation $\sigma > 0$. We define the score function Γ as $\Gamma(x; \sigma) = \sum_{i=1}^{T} \mid w_i \mid \mathcal{K}(v_i;\, x, \sigma)$. Γ quantifies the abundance of breakpoints around location x. The size of the neighborhood is controlled by the standard deviation of the kernel, σ. The absolute values of the weights w_i increase the contribution of large changes in log-ratio and reduce the contribution of small changes. Consequently, a high $\Gamma(x, \sigma)$ score is attained either by the contribution of many breakpoints of low-weight around location x or by few breakpoints of large weight. Weights also decrease the influence of small log-ratio changes, which are wrongly classified as breakpoints by the segmentation procedure.

Γ assigns a score to each genomic position (in practice, only locations $\{v_i \mid 1 \leq i \leq T\}$ are scored). Local maxima of Γ, which are significantly large, are candidates for consensus breakpoints. The parameter σ (*kernel width*) of Γ specifies how 'tightly' the breakpoints should align to be considered to form a consensus breakpoint. Based on the observation that in real data, breakpoints aggregate in varying degrees of tightness, we apply the scoring function multiply, with different kernel width values. In our experiments, kernel widths can take values from the set $\text{SD} = \{10^3, 10^4, \dots, 10^8\}$.

Let the kernel width σ be fixed. To identify the local maxima of $\Gamma(\cdot\,; \sigma)$ that are statistically significant, we compare a local-maximum score to a null reference, which is obtained by randomly re-arranging the breakpoints, such that the dependencies between arrays and the dependence on the genomic locations are destroyed. The random re-arrangement is carried out as follows: for each array independently, its breakpoints are re-located to random genomic positions, which are generated by a uniform distribution over the entire chromosome. After all arrays have been processed, a *null instance* of the consensus breakpoint detection problem is generated, of the same size as the initial problem. Let $V^0 = \{(v_i^0, s_i, w_i) \mid 1 \leq i \leq T\}$ be the null instance. The $\Gamma(\cdot\,; \sigma)$ score is computed at the null locations $\{v_i^0 \mid 1 \leq i \leq T\}$. This procedure is repeated P times ($P = 50$ in our experiments), enough for obtaining a large population of null scores covering the chromosome. After appropriate sorting and re-indexing of the P sets of null genomic locations, let $v_1^0, \ldots, v_{PT}^0$ be the genomic locations at which the null scores $\gamma_1^0, \ldots, \gamma_{PT}^0$ have been estimated.

The significance of the observed score at location x for kernel width σ is given by a Z-score as follows:

$$z(x, \sigma) = \frac{\Gamma(x; \sigma) - \text{mean}\,\{\gamma_{i+1}^0, \ldots, \gamma_{i+j}^0 | v_{i+1}^0, \ldots, v_{i+j}^0 \text{ are } j \text{ NN of } x\}}{\text{sd}\,\{\gamma_{i+1}^0, \ldots, \gamma_{i+j}^0 | v_{i+1}^0, \ldots, v_{i+j}^0 \text{ are } j \text{ NN of } x\}} \tag{1}$$

where NN stands for nearest-neighbor. The Z-score indicates how large is the observed score at a certain location x, measured in standard deviations from the mean of the null scores at j NN locations around x. In our experiments, the value of j is 50. Locations with Z-score below some positive threshold $\zeta > 0$ are considered not significant. The C-KS algorithm returns a list of local maxima of the scoring function Γ with positive Z-scores, sorted decreasingly by Z-scores.

Supplementary Figure S1 from the Supplementary Material illustrates the C-KS algorithm.

In this manuscript, we call *consensus regions* the genomic intervals between two consecutive consensus breakpoints.

2.2 Validation of C-KS algorithm

In the absence of a benchmark of annotated consensus breakpoints, we suggest two approaches for validating the C-KS algorithm. One approach is qualitative, investigating key genomic and epigenomic properties of the locations at which we report consensus breakpoints (see Section 4.2). A second approach is quantitative, and it uses the breakpoints $B_1, \ldots, B_m$ as a basis for genome segmentation and dimension reduction by representing the samples $a_i \in \mathcal{R}^p$ in an m-dimensional space as follows. All log-ratios of sample i measured at genomic loci between consecutive consensus breakpoints B_k and B_{k+1} are averaged. The resulting value, denoted by $x_i(k)$, is the k^{th} component of the *reduced representation* x_i of sample i in the space of the consensus regions. With the reduced representations $x_1, \ldots, x_m$ as rows, we form the matrix X. We use X as input data for a supervised prediction of a binary indicator y of tumor phenotype. Examples of phenotype indicators include binary representations of tumor stage, tumor grade, age at diagnosis, survival, response to treatment and so forth. We compare the prediction accuracy of the model to that of a baseline model, using the initial probe data as input and the same phenotype y. If the set of consensus breakpoints is biologically meaningful, then the accuracy of the prediction based on the reduced representation X should not be significantly worse than that based on the baseline. We base this requirement on the assumption that if biologically meaningful consensus breakpoints are missed, then important copy number changes are overlooked and the relation to the phenotype deteriorates. Moreover, if prediction accuracy increases after dimension reduction, then reducing the variance between consecutive consensus breakpoints via averaging is clearly beneficial, and we conclude that the (relatively small) set of consensus breakpoints carries biological significance. This approach is inspired by the existing single-array segmentation procedures, which even out the raw log-ratios between individual breakpoints for more precise estimation of the underlying copy number. The reduced representation X is a multivariate generalization of the single-array segmentation, and we call it *consensus segmentation*. The validation procedure described above will be referred to as *consensus segmentation validation*.

The list of consensus breakpoints identified by C-KS can be sorted in decreasing order by Z-score. We expect that the top-ranking consensus breakpoints are more useful that the ones at the bottom of the list. We propose a supervised selection of the optimal value k, $1 \le k \le m$, such that the consensus segmentation using the top-ranking k consensus breakpoints achieves the largest prediction accuracy. The selection is carried out via classical cross-validation. This supervised selection procedure constitutes a data-driven alternative to choosing a significance cut-off for the Z-score, when phenotypic annotation is available.

Technically, for the prediction of phenotype, we use the Lasso-penalized Logistic Regression method (LLR). LLR models the logarithm of the posterior probabilities of the classes as linear functions of the input features. The parameters w of the model are estimated by maximizing the log-likelihood $L(w; X, y)$ over the observations in the training set. Model sparsity is obtained via the Lasso penalty, weighted by a factor λ, which can be optimized with cross-validation. Feature importance is given by the model weights $w_{\text{LLR}} = \arg\max_w L(w; X, y) - \lambda \sum_{j=1}^{p} | w_j |$.

In our experiments, we used the **R** package *glmnet* (Friedman *et al.*, 2010) for training LLR models. Ten-fold cross-validation was carried out for the simultaneous estimation of optimal parameters λ (penalty) and k (number of consensus breakpoints selected from the list) and prediction accuracy.

In our experiments, we report the prediction accuracy of the optimal model as proportion of correctly classified samples. In one specific classification scenario, in which we deal with models trained on unbalanced classes, for a fair comparison, we compute the weighted accuracy, or balanced accuracy:

$$\frac{|\{x \in C_1 | \hat{f}(x) = 1\}|}{2|C_1|} + \frac{|\{x \in C_0 | \hat{f}(x) = 0\}|}{2|C_0|} = \frac{\text{sensitivity} + \text{specificity}}{2},$$

where C_0 and C_1 are the two response classes and $\hat{f}$ is the prediction function. We abbreviate the weighted accuracy as WACC, to distinguish it from the classical accuracy, ACC.

2.3 Summarizing genomic and epigenomic properties of consensus breakpoints with EpiExplorer

We have investigated the genomic and epigenomic properties of the genomic locations at which consensus breakpoints are reported by C-KS. For this purpose, we used EpiExplorer (Halachev *et al.*, 2012), a web-based application for interactive exploration of sets of genomic regions. We explored the overlap with gene promoters, CpG islands, insulators and various histone modifications.

EpiExplorer takes as input a set of genomic regions, which for our study we defined as follows. From each consensus breakpoint, we constructed a genomic region, which is likely to cover the set of individual breakpoints that have a non-negligible contribution to the Γ score of the consensus breakpoint. Specifically, for consensus breakpoint B_k, the region is centered at the location of the consensus breakpoint, and the width is twice the width of the kernel pertaining to consensus breakpoint B_k. For specificity, we restricted to the subsets of consensus breakpoints obtained with small kernel width ($\sigma \in \{1\text{Kbp}, 10\text{Kbp}\}$).

EpiExplorer also allows for direct comparison with control regions, which in our experiments are randomly generated from the genome. For more meaningful results, we used the option of variance estimation offered by EpiExplorer, which means that 10 reference sets are used and corresponding mean and standard deviation are shown.

3 DATA AND CLASSIFICATION SETTINGS

3.1 Neuroblastoma

Neuroblastoma is a tumor that affects the sympathetic nervous system of young children and infants. For the purpose of therapy selection, several factors are currently considered, including classical staging, age at diagnosis and several copy number markers. Five stages are defined for neuroblastoma: 1–4 and 4S. Stages 4 and 4S both correspond to metastasized tumors, but stage 4S neuroblastoma occurs more, most often in infants (aged <1 year), and has a high chance of undergoing spontaneous remission. Stage 1, 2 and some of stage 3 neuroblastomas have good prognosis. For higher stages, prognosis is established taking into account several molecular markers but also, importantly, the age at diagnosis. Until several years ago, patients aged <12 months were included in the good prognosis group, whereas the older ones were considered to have poor prognosis and were administered aggressive treatment. In the recent years, many clinicians have adopted a higher age cut-off, of around 18 months, based on research evidence (London *et al.*, 2005; Evans and D'Angio, 2005; Schmidt *et al.*, 2005). For clinical practice a larger cut-off has a significant impact, as it means that more young children will not receive unnecessary aggressive chemotherapy.

Copy number aberrations have been shown to influence the outcome of therapy for patients with neuroblastoma (Fischer and Berthold, 2010). Hyperploidy is generally indicative of good outcome and is characteristic of tumors occurring in infants. Amplification of the MYCN gene located on chromosome 2, deletion of 11p and 1p and amplification of 17q are the most widely used indicators of poor prognosis (Ambros *et al.*, 2009).

We analyzed 162 neuroblastoma tumors, obtained by two Agilent arrayCGH microarray platforms (44 and 100 k resolution). The tumors are annotated with stage and age at diagnosis (in short, age). We have formulated the following classification tasks: (i) stage 4S (**class 0**) versus stage 4 (**class 1**); (ii) age $\leq t$ (**class 0**) versus age $> t$ (**class 1**). We choose age cut-offs $t \in \{2, 4, 6, \ldots, 32\}$ (months) and for each value of t, we train a classifier. We compute the WACC of each classifier and suggest that the age cut-off resulting in the maximum WACC value corresponds to a maximal difference between the genomes of the tumors from the two groups.

3.2 Breast cancer

From the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/), we obtained public DNA copy number data on three breast cancer cohorts, all using the Agilent Human Genome CGH Microarray 244 A platform (236 000 60-mer oligonucleotide probes).

The first cohort consists of 54 breast tumors (Sircoulomb *et al.*, 2010). In this manuscript, we call this dataset **breast54**. Annotations regarding the estrogen receptor (ER) status and progesterone (PgR) receptor status are available. Also, a binary indicator called metastasis-free survival (MFS) with value 1 if the tumor has not metastasized after 5 years and 0 otherwise is available. The dataset also contains an indicator of tumor subtype, which can be either inflammatory (IBC) or non-inflammatory breast cancer (NIBC). We conducted the following classification experiments: (i) ER negative (**class 0**) versus ER positive tumors (**class 1**); (ii) PgR negative (**class 0**) versus PgR positive tumors (**class 1**); (iii) metastasis-free survival $\geq$ 5 years (**class 0**) versus metastasis-free survival $<$5 years (**class 1**); and (iv) NIBC (**class 0**) versus IBC tumors (**class 1**).

The second dataset comprises 173 breast tumors. The experimental data were published by Bekhouche *et al.* (2011). We call this dataset **breast173**. ER and PgR status annotations are available for the tumors. We performed the following classification tasks: (i) ER negative (**class 0**) versus ER positive tumors (**class 1**); (ii) PgR negative (**class 0**) versus PgR positive tumors (**class 1**).

The third set of array CGH experiments comprises 167 breast tumors, published by Russnes *et al.* (2010). We will call this dataset **breast167**. The following phenotypic annotations are provided: ER and PgR status, lymph node spread status (yes or no), tumor stage, tumor grade and histological subtype. We performed the following classification tasks: (i) ER negative (**class 0**) versus ER positive tumors (**class 1**); (ii) PgR negative (**class 0**) versus PgR positive tumors (**class 1**); (iii) tumor not spread to lymph nodes (**class 0**) versus tumor spread to lymph nodes (**class 1**); (iv) stage pT1 (**class 0**) versus stage pT2 (**class 1**); (v) grade 2 (**class 0**) versus grade 3 (**class 1**); and (vi) Histological subtype ductal (**class 0**) versus lobular (**class 1**).

3.3 Colon cancer, ovarian cancer and glioblastoma

The following datasets contained no publicly available clinical indicators associated with the tumors. Veeriah *et al.* (2010) investigated a cohort of 98 colon tumors. The experiments based on the Agilent Human Genome CGH Microarray 244 A microarrays have been made public at the GEO. We call this dataset **colon**.

A comprehensive collection of arrayCGH experiments on 290 ovarian cancer samples are publicly available at TCGA. The tumors have been analyzed using the Agilent Human CGH 1 × 1M G4447A arrays, with ~1 million probes covering the genome. We call this dataset **ovarian**.

From the same repository, we have obtained publicly available data on 539 glioblastoma tumors, analyzed using Agilent Human Genome CGH 244 A microarray experiments. This dataset will be called **glioblastoma**.

4 RESULTS

4.1 Validation of the consensus segmentation

We applied the algorithm C-KS to the seven datasets presented in Section 3. Beforehand, we have segmented the data using the CBS (Olshen *et al.*, 2004) method. Table 1 shows the number

Table 1. Number of breakpoints and consensus breakpoints identified in six cancer datasets

Dataset	No. of breakpoints		No. of consensus breakpoints
	Per tumor	Total	Identified by C-KS
Neuroblastoma	54 ± 24	4047	62 ± 7(65)
Colon	168 ± 62	16 985	173 ± 11(98)
Glioblastoma	261 ± 72	64 729	492 ± 35(132)
Breast173	339 ± 115	49 266	320 ± 7(154)
Breast54	394 ± 85	20 093	327 ± 8(62)
Breast167	461 ± 182	37 055	503 ± 11(73)
Ovarian	806 ± 201	125 000	662 ± 8(189)

Note: Second column contains the average number of breakpoints per tumor, with indication of standard deviation. The third column shows the total number of distinct breakpoints in the cohorts. Column four shows the number of consensus breakpoints identified by the C-KS algorithm, with indication of standard deviation computed via 10-fold cross-validation. In brackets, the magnitude of dimension after consensus segmentation reduction is indicated.

of consensus breakpoints with Z-score larger than zero identified in each dataset in comparison with the average number of breakpoints per tumor and the total number of breakpoints in each cohort. We chose a permissive threshold for the Z-score (zero) to ensure that the supervised selection of the optimal threshold is decisive.

Table 1 shows that neuroblastoma stands out from the rest of the cancer types with few breakpoints per tumor (54 on average), a striking difference to other tumors probably related to the fact that neuroblastoma appears early during infancy, and DNA aberrations do not have sufficient time to accumulate. Moreover, segmental gains or losses are not characteristic of neuroblastoma, more frequent events being whole chromosome gain (hyperploidy) or loss (aneuploidy). These are not delimited by breakpoints located strictly within chromosomes, hence, the small counts. Colon cancer, glioblastoma, breast cancer follow, with increasing number of breakpoints. The highest frequency of breakpoints is observed in the ovarian cohort, with 806 breakpoints per tumor.

The total number of breakpoints (excluding duplicates) is the minimal number of features necessary to represent the segmented log-ratio data each tumor in the cohort and the values are evidently large (see column three of Table 1). For the task of phenotype prediction, for example, dealing with the high dimensionality of the data is a challenge (typical $p \gg n$ problem). The fourth column of Table 1 shows the number of consensus breakpoints identified in each cohort, with indication of standard deviation, which is the result of the application of C-KS method independently on 10 sub-folds of the data, each containing 90% of the tumor samples. The number of consensus breakpoints also corresponds to the number of super-features used to represent the samples after consensus segmentation (see Section 2). The dimension reduction is substantial, the fold change being indicated in brackets in Table 1, fourth column.

We tested whether the new models based on the smaller feature set exhibit reduced predictivity of tumor phenotype by comparing the prediction accuracy of LLR models trained on the initial datasets and on the reduced datasets, as explained in Section 2. We performed the classification tasks presented in Section 3. Table 2 shows performance in terms of ACC estimated via 10-fold cross-validation, for LLR and for LLR after C-KS segmentation (LLR + C-KS). Using consensus segmentation results in comparable accuracy with that given by the baseline model. In fact, accuracy improves slightly in 7 of 13 cases. This is so despite disregarding all copy number changes (breakpoints) between consensus breakpoints. The accuracy values that we report are competitive with results presented elsewhere. For example, Chin *et al.* (2007) use a k-NN classifier and report an accuracy of 75.3% for ER status prediction, whereas our method achieves 75.7% on average over the three breast cohorts. The same study reports 74.3% accuracy for prediction of tumor grade, whereas we achieve 75%.

Table 2. Accuracy of the various classification models

Dataset	Phenotype	LLR	LLR + C-KS
Breast54	ER	0.68 ± 0.03	0.72 ± 0.03
	PgR	0.64 ± 0.03	0.70 ± 0.03
	MFS	0.56 ± 0.04	0.69 ± 0.03
	Type	0.59 ± 0.03	0.57 ± 0.03
Breast167	ER	0.70 ± 0.02	0.72 ± 0.02
	PgR	0.62 ± 0.02	0.64 ± 0.02
	Lymph	0.54 ± 0.02	0.64 ± 0.02
	Stage	0.60 ± 0.02	0.62 ± 0.02
	Grade	0.76 ± 0.01	0.75 ± 0.02
	Hist	0.76 ± 0.01	0.76 ± 0.02
Breast173	ER	0.84 ± 0.01	0.83 ± 0.01
	PgR	0.77 ± 0.01	0.76 ± 0.01
Neuroblastoma	4vs4S	0.80 ± 0.02	0.79 ± 0.02

Note: Columns three and four show the mean and standard deviation of the prediction accuracy.

For the case of prediction of age at diagnosis in neuroblastoma, in order to make the models comparable in the face of imbalances in class cardinality, we computed the WACC performance measure for each classifier. Figure 1a shows the WACC values for each age cut-off. A maximum WACC is obtained at a cut-off of 18 months, meaning that the age groups are most different for this dichotomization. Not significantly lower performance is achieved also by the cut-off of 16 months. Our result is consistent with the recent recommendations for neuroblastoma treatment, which suggest 18 months cut-off for differential therapy. To our knowledge, we are the first to present arguments in support of this choice based solely on the copy number profiles of the tumors.

The insights into neuroblastoma that our prediction models provide reach further. We have analyzed the weights of the features in the models, in order to assess the predictive value of each consensus region. To make the weights comparable, we performed the following simple normalization: for each model, we computed the absolute value of the weights and divided by the maximum absolute weight. Hence, the most relevant feature of each model has a weight of 1, and the features that do not enter the model have a weight equal to 0. All other feature weights fall between 0 and 1. Figure 1b shows a heatmap representation of the normalized feature weights. The most predictive features are consensus regions situated at chromosomes 1q, 2q, 3p, 4p, 6q, 11, 14, 17 18 and 20. The importance of these regions changes smoothly with the increase of the cut-off from 2 to 32 months. Some features are more useful for discriminating between young patients and the rest, such as the region at 2q, which is the strongest predictor for a cut-off of 8–10 months. This region is consistently gained in the younger patients, as a consequence of the whole chromosome 2 gain. The 17p also seems predictive if the cut-off value is of 8–10 months, for a similar reason: whole chromosome 17 gain is characteristic for younger patients. When the age split is increased to 12 months, 2q is not dominant anymore and instead a combination of 2q, 3p, 4p, 6q, 11, 17 and 18 contributes to the classification model. All these regions are predictive because they indicate full chromosome gains or losses (polyploidy), which are characteristic to many neuroblastoma tumors. Models corresponding to cut-offs from 14 to 18 months involve roughly the same regions, with slight variations. For cut-offs larger than 18 months, region 2q is no longer predictive, but chromosomes 14 and 20 are added to the model.

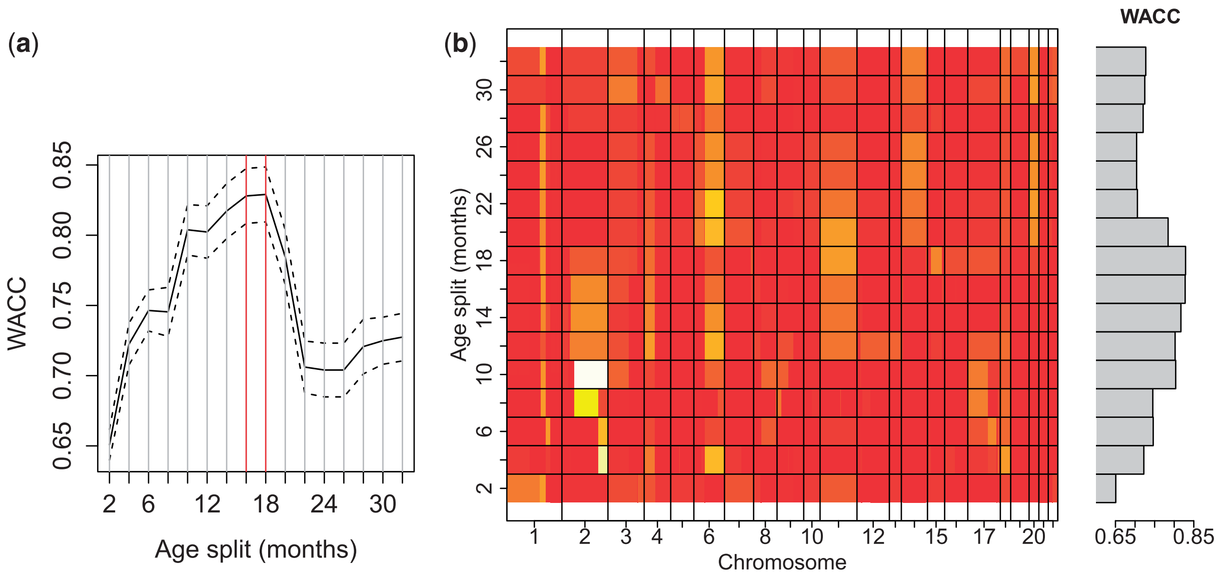

Fig. 1. Prediction of age group-based copy number data from the neuroblastoma dataset. The input features were the consensus regions resulting from segmentation with C-KS. (**a**) Class-weighted accuracy as a function of the age cut-off. (**b**) Feature weights (in absolute value) for all models. Dark red corresponds to weights close to zero (irrelevant features) and shades of yellow to white mark increasingly relevant features for prediction. The features are ordered according to the genome position. On the right side, the weighted class accuracy is shown

4.2 Genomic and epigenomic properties of consensus breakpoints

Using EpiExplorer, we investigated the (epi)genomic properties of tight consensus breakpoints with significant *Z*-score from all seven datasets. Figure 2 summarizes our findings. Comparing with a random reference, we observe an enrichment in overlap with gene promoters. We also notice an enrichment in overlap with CpG islands, which has been reported previously (Abeysinghe *et al.*, 2003). Because the CpG islands are often found in gene promoters, we excluded the regions overlapping with gene promoters and re-evaluated the overlap with CpG islands (result not showed here). Interestingly, the enrichment in CpG islands persisted.

We investigated also the overlap with insulators and histone modifications. As these are tissue specific and there are no public annotations of sequence data from cancer tissues, we used data from human embryonic stem cells. Our results show slight enrichment of insulator regions overlapping with consensus breakpoints. Insulators are proteins that bind to DNA in specific locations to establish transcription boundaries, and insulator regions are the genomic regions insulators bind to. An example is the CTCF insulator protein, the function of which is often disrupted in cancer, e.g. by hypermethylation of its binding site (Feinberg and Tycko, 2004). We observe the enrichment of histone marks that are associated with dynamic regulation of gene transcription (H3K4me3, H3K27me3 and H3K36me3).

To guarantee that the particular selection of the probes on the microarray chip does not bias the enrichment analysis, for example, by being located mostly within gene-rich regions, we ran an additional control experiment. We collected all individual breakpoints from each dataset and repeated the analysis using EpiExplorer (white bars in Fig. 2). The overlap with the (epi)genomic properties was in general lower, notably in the case of CpG islands. This suggests that the selection of breakpoint locations performed by the C-KS algorithm is biologically meaningful.

Although many studies show that the location of the DNA breakpoint in fusion transcripts is critical, little is known about the biological relevance of the breakpoints associated with DNA gain or loss. It is commonly accepted that the oncogenes or tumor suppressors located within the gained or lost segments are responsible for tumor development, and the location of the breakpoint is not essential. This hypothesis is probably true in most of the cases, as real data show that recurrent aberrations may vary greatly in their extent. However, many tumors display focal aberrations with tightly aligned breakpoints, which suggests that the local structure of the chromatin ‘forces’ the breaks to occur within certain regions by hindering DNA repair (Soria *et al.*, 2012) (see also Supplementary Figs S2 and S3 for examples from our datasets and TightConsensusBreakpoints.xls available as Supplementary Material).

The enrichment analysis that we performed supports the hypothesis that the DNA locations of tightly aligned consensus breakpoints have interesting biological properties that deserve to be further investigated.

5 DISCUSSION

We have introduced the algorithm C-KS for identifying consensus breakpoints from DNA copy number alteration data. The algorithm takes as input the list of breakpoints occurring in the tumor cohort. C-KS uses a weighted kernel counting for evaluating the abundance of breakpoints around genomic locations.

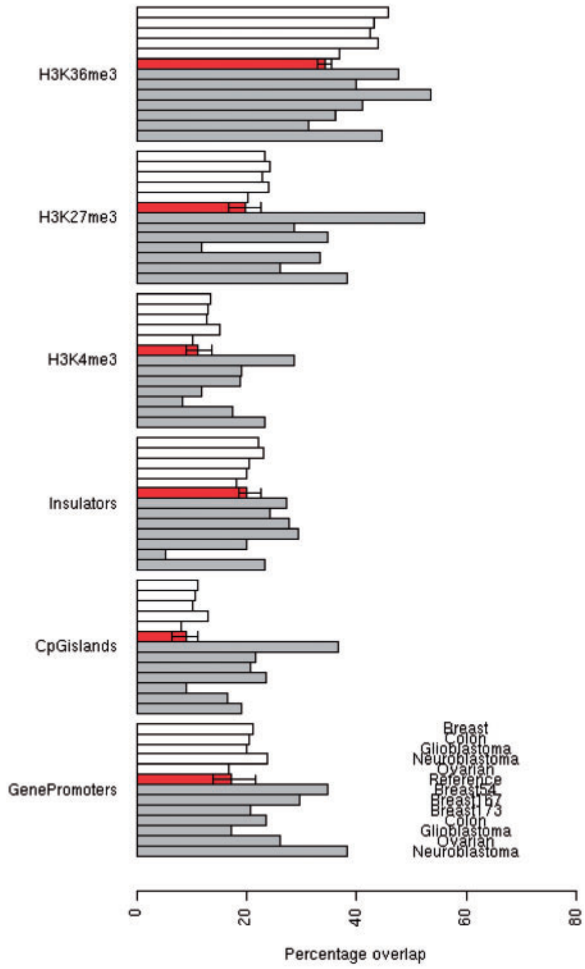

Fig. 2. Properties of consensus breakpoint regions given by EpiExplorer. On the *y*-axis, various genomic and epigenomic properties are listed. On the *x*-axis, we show the percentage of consensus breakpoint regions overlapping with each of the properties. In gray, the consensus breakpoints from each dataset are represented. In red, the random reference is shown, with standard deviation bars. In white, summary on the overlap of individual breakpoints with (epi)genomic properties

A significance *Z*-score is associated to each consensus breakpoint, resulting from a comparison with null scores obtained via random permutation.

We also introduced the concept of consensus segmentation, which is a generalization of the single-sample segmentation procedure. We used the consensus segmentation for representing the data into a space of lower dimensionality and carried out classification tasks with tumor phenotype as outcome variable.

We applied the C-KS algorithm to seven arrayCGH datasets. We showed that despite the significant dimension reduction, the consensus segmentation driven by the consensus breakpoints results in comparable accuracy of prediction of the tumor phenotype. Often, the prediction accuracy increases after segmentation, especially in cases in which the number of samples is small.

Based on our classification models, the copy number aberration profiles of neuroblastoma tumors are most distinct between age groups defined using a cut-off of 18 months. Our finding confirms recent recommendations from the research community, based on arguments not incorporating copy number alteration profiles. We believe that our result is important, as it provides with a biologically founded explanation to the impact of age at diagnosis on outcome.

We also showed that the genomic regions identified as tightly aligned consensus breakpoints are enriched in overlap with functional elements, such as gene promoters, CpG islands, several histone modifications and insulators, which constitute interesting biological findings that deserve further investigation.

To conclude, we describe an algorithm for identifying consensus breakpoints and a method for consensus segmentation of copy number aberration data. The substantial dimension reduction achieved by consensus segmentation with no loss of essential predictive information can constitute the basis of efficient downstream analysis and data integration.

ACKNOWLEDGEMENT

The authors thank Jörg Rahnenführer for valuable comments on this work. We highly appreciate the collective efforts of the members of the TCGA portal (http://cancergenome.nih.gov/), who continue to publish high-quality data on numerous cancer types, some of which we used in this study.

Funding: German National Genome Research Network (NGFNplus) (01GS08100 to L.T.).

Conflict of Interest: none declared.

REFERENCES

Abeysinghe,S.S. *et al.* (2003) Translocation and gross deletion breakpoints in human inherited disease and cancer I: nucleotide composition and recombination-associated motifs. *Hum. Mutat.*, **22**, 229–244.

Ambros,P.F. *et al.* (2009) International consensus for neuroblastoma molecular diagnostics: report from the international neuroblastoma risk group (INRG) biology committee. *Br. J. Cancer*, **100**, 1471–1482.

Bekhouche,I. *et al.* (2011) High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes. *PLoS One*, **6**, e16950.

Chin,S. *et al.* (2007) Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. *Oncogene*, **26**, 1959–1970.

Evans,A.E. and D'Angio,G.J. (2005) Age at diagnosis and prognosis in children with neuroblastoma. *J. Clin. Oncol.*, **23**, 6443–6444.

Feinberg,A.P. and Tycko,B. (2004) Timeline: the history of cancer epigenetics. *Nat. Rev. Cancer*, **4**, 1–11.

Fischer,M. and Berthold,F. (2010) The role of complex genomic alterations in neuroblastoma risk estimation. *Genome Med.*, **2**, 31.

Friedman,J.H. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Halachev,K. *et al.* (2012) EpiExplorer: live exploration and global analysis of large epigenomic datasets. *Genome Biol.*, **13**, R96.

Hupé,P. *et al.* (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

London,W.B. *et al.* (2005) Evidence for an age cutoff greater than 365 days for neuroblastoma risk group stratification in the children's oncology group. *J. Clin. Oncol.*, **23**, 6459–6465.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pinkel,D. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Ritz,A. *et al.* (2011) Detection of recurrent rearrangement breakpoints from copy number data. *BMC Bioinformatics*, **12**, 114.

Russnes,H.G. *et al.* (2010) Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci. Transl. Med.*, **2**, 38ra47.

Schmidt,M.L. *et al.* (2005) Favorable prognosis for patients 12 to 18 months of age with stage 4 nonamplified MYCN neuroblastoma: a children's cancer group study. *J. Clin. Oncol.*, **23**, 6474–6480.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Sircoulomb,F. *et al.* (2010) Genome profiling of ERBB2-amplified breast cancers. *BMC Cancer*, **10**, 539.

Solinas-Toldo,S. *et al.* (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Soria,G. *et al.* (2012) Prime, repair, restore: the active role of chromatin in the DNA damage response. *Mol. Cell*, **46**, 722–734.

Veeriah,S. *et al.* (2010) Somatic mutations of the Parkinson's disease-associated gene PARK2 in glioblastoma and other human malignancies. *Nat. Genet.*, **42**, 77–82.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.