

Topological Analysis of Amplicon Structure in Comparative Genomic Hybridization (CGH) Data: An Application to ERBB2/HER2/NEU Amplified Tumors

Sergio Ardanza-Trevijano¹, Georgina Gonzalez², Tyler Borrman³,
Juan Luis Garcia⁴, and Javier Arsuaga^{2,5}(✉)

¹ Department of Physics and Applied Mathematics, University of Navarra,
31080 Pamplona, Spain

sardanza@unav.es

² Department of Molecular and Cellular Biology, University of California Davis,
One Shields Avenue, Davis, CA 95616, USA

gingonzalez@ucdavis.edu

³ Medical School, University of Massachusetts, 368 Plantation Street,
Worcester, MA 01605, USA

tyler.borrman@umassmed.edu

⁴ Centro de Investigación del Cáncer, Universidad de Salamanca,
37007 Salamanca, Spain

jlgarcia@usal.es

⁵ Department of Mathematics, University of California Davis,
One Shields Avenue, Davis, CA 95616, USA

jarsuaga@ucdavis.edu

Abstract. DNA copy number aberrations (CNAs) play an important role in cancer and can be experimentally detected using microarray comparative genomic hybridization (CGH) techniques. Amplicons, CNAs that extend over large sections of the genome, are difficult to study since they may contain multiple independent and dependent copy number changes. Here, we propose an algorithm to find the CNAs structure within a given amplicon. Our method relies on the observation that co-occurring CNAs can be encoded as 1-dimensional cycles. Applying this method to breast cancer patients known as ERBB2/HER2/NEU amplified we find three regions that can be co-occurring: the first region is in the cytoband 17q12, where the ERBB2 gene is located, the second region expands between 17q21.2 to 17q21.31 and includes the keratin genes, the third one is 17q21.33. We suggest that the first homology group helps uncovering the structure of amplicons.

Keywords: Copy number aberrations · Cancer · Computational homology · First homology group

S. Ardanza-Trevijano and G. Gonzalez contributed equally to this work.

© Springer International Publishing Switzerland 2016

A. Bac and J.-L. Mari (Eds.): CTIC 2016, LNCS 9667, pp. 113–129, 2016.

DOI: 10.1007/978-3-319-39441-1_11

1 Introduction

Cancer is a set of complex genetic diseases whose pathogenesis is not well understood. Initiation and progression of these diseases depend on the misregulation of key genes called *cancer/tumor genes*. Gene misregulation occurs through different mechanisms including the gain and losses of DNA chromosome fragments (e.g. [11, 18, 20, 24]). These events are commonly termed DNA copy number aberrations (CNAs) and are routinely detected in the laboratory through comparative genomic hybridization (CGH) arrays, single nucleotide polymorphism (SNP) arrays and sequencing (e.g. [12–14, 17, 22, 36, 47]). However not all detected CNAs are relevant for tumor initiation and/or progression. It is currently believed that CNAs that contain tumor genes are those that are relevant for tumor progression. These CNAs are called *drivers* while those which appear to have no biological implications are called passengers. Determining which CNAs are driving tumor progression and which ones are just passengers remains an open problem. Certain CNAs expand over large fragments of the genome and are sometimes termed *Amplicons*. These regions are important because contain multiple tumor genes and the presence or absence of certain CNAs within an amplicon has been associated with patient’s prognosis (e.g. [23, 41]). Examples include 9p in breast cancer, colon and glioblastoma tumors and lymphomas [5, 19], 11q in head and neck, breast, oral and liver tumors (reviewed in [46]) and 17q in ERBB2/HER2/NEU (ERBB2+, thereafter) positive breast cancer [4]. The detailed structure of amplicons is complex and difficult to investigate using traditional statistical methods since some amplifications appear to occur simultaneously, hence they are not significant as independent CNAs, and have synergistic effects [1, 28, 43]. In this work we will call *co-occurring CNAs* those that occur simultaneously independently of their functional effects. One potential approach to study the structure of an amplicon and identify potential co-occurring CNAs is to encode combinations of CNAs as a single predictor variable and perform association studies between these new predictor variables and phenotypes of interest.

Here we extend our previously reported supervised approach, termed Topological Analysis of array CGH (TAaCGH), to study the structure of an amplicon. In TAaCGH, we associate a point cloud to each CGH profile (or section of a CGH profile) through a sliding window algorithm [15], build a Vietoris-Rips (VR) simplicial complex [31] and perform an association study between the topological properties of the VR complex and the chosen phenotype. The difference between TAaCGH and other current association studies is that TAaCGH uses the topological properties of the point cloud, instead of the probes, as predictor variables. The advantage of using topological properties as predictors is that they can encode relationships between probes. In previous works we showed that using the rank of the zero homology group (β_0) as a predictor variable in association studies of breast cancer is comparable to other statistical methods [3]. Here we hypothesize that performing association between the rank of the first homology group β_1 and a specific phenotype helps analyze the underlying structure of amplicons. This hypothesis is based on recent analytical and numerical

results that shows that β_1 encodes for periodic patterns [34] and by our own observations that show that neighboring (not-necessarily periodic) regions of amplifications are mirrored by β_1 [10,38].

To test our hypothesis and to illustrate our methodology we analyze the amplicon on 17q in ERBB2/HER2/NEU (ERBB2+, thereafter) positive breast cancer samples. ERBB2+ breast cancer is an aggressive form of the disease that comprises 25% of all breast tumors diagnosed (reviewed in [35]). The ERBB2 gene is located in the region of the genome labeled as cytoband 17q12 (where 17 is the chromosome arm, q denotes the long arm of the chromosome and 12 denotes a specific band that can be detected by chromosome staining). Misregulation of ERBB2 in ERBB2+ tumors commonly occurs through copy number gains of 17q12. In many patients, this amplification is accompanied by gains of other regions in the same chromosome arm. This includes amplifications of 17q21.2 that encompasses the *Top2A* gene [32], chromosome regions 17q21.1, 17q22 [27] and 17q21.33 – q25.1 which is predictive of early recurrence [9] and contains TANC2 (17q23) and PPM1D genes [29,37], two independent co-amplified regions have also been reported in 17q23 [4,39].

To test whether TAaCGH can detect these events, we analyzed two independently published data sets [13,20]. We first confirmed the presence of the amplicon in 17q in both data sets using β_0 , we then identified specific regions within this arm using β_1 analysis. This study revealed two regions of significance delimited by 17q12 and 17q12 – 17q21.33. To further localize the regions of the genome that contributed to the significance of β_1 we calculated the generators of the first homology group and the correspondence between the probes and the generators. Statistical analysis quantifying the over-representations of genomic regions in the generators allowed us to further subdivide the region 17q12 – 17q21.33. A first amplification was detected in between the neighboring regions 17q21.2 – 17q21.31 (extending from base pairs 40,884,763-41,826,877) and the region 17q21.33 (from base-pairs 46,603,678-49,075570). Using the UCSC genome browser we observed that the first region contains the keratin cluster (e.g. [30]) and the second contains, among others the HOXB cluster (see [8] for a review). Both of these clusters have been previously reported in breast cancer studies. Whether their functionality is synergistic in some patients remains to be determined.

2 Data Sets and Methods

2.1 CGH Data

CNAs are defined as gains or losses of genome fragments and can be detected using microarray technologies. Through Comparative Genomic Hybridization (CGH), DNA probes (i.e., fragments of DNA sequences) are spotted on a platform. Tumor DNA, labeled with Cy3, and control DNA, labeled with Cy5, are co-hybridized in a 1:1 ratio. The intensity of the hybridized samples is captured and transformed into a red-green ratio value called the \log_2 ratio. Since the physical position of each probe is known, these \log_2 ratios can be mapped to

the original genome producing a CGH profile (Fig. 1). In traditional statistical approaches each CGH profile is normalized and segmented, and significant copy number aberrations are then identified [6, 33, 45].

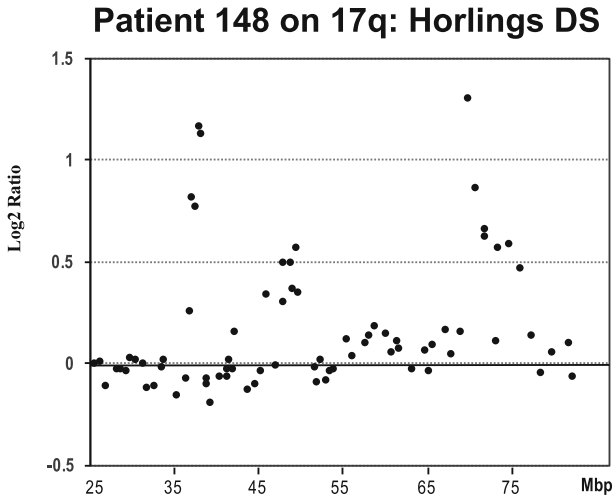


Fig. 1. A CGH profile for chromosome arm 17q. The x-axis indicates the genomic position and the y-axis the \log_2 ratio of the intensity of the tumor and control samples co-hybridized to the same array.

2.2 Simulation Data Set

We simulated single and co-occurring aberrations. A detailed description of the simulation methods for a single aberration can be found in [3, 25, 26]. In brief, each simulation consisted of 200 profiles, 100 in the control set and 100 in the test set. Each simulated profile contained 100 aCGH probes. The value of the copy number along the profile was determined by three parameters: the mean value of the aberration μ , the length of the aberration λ , and the standard deviation associated with noise σ . Probes outside the aberration and in the control set had $\mu = 0$, whereas for those probes inside the aberration was $\mu = 0.6$ or 1. Aberration length λ was equal to 5 and 10 probes. Noise was implemented by drawing samples from a Gaussian distribution of mean 0 and standard deviation σ of values 0.2, 0.6 or 1. The control set for single aberrations was made of profiles without aberrations (i.e. only noise).

Co-occurring aberrations were represented by two aberrations of different lengths. In the first aberration $\mu = 0.6$ or 1 and in the second $\mu = 1$. The control set was made of profiles with no aberrations or with only one aberration.

2.3 Horlings Data Set

This dataset analyzed was published by Horlings and colleagues [20] and was obtained from the supplementary data [21]. Measurements of copy number

variations were performed on microarrays containing 3.5 k BAC, PAC-derived DNA segments covering the entire genome with a spacing average of 1 Mb. Each BAC clone was spotted and triplicated on every slide (Code Link Activated Slides, Amersham Biosciences). Our own preprocessing of the data can be found in [3]. This study contained 14 ERBB2+ patients determined by clinical diagnosis. The control set consisted of the patients belonging to the remaining subtypes.

2.4 Climent Data Set

This data set was used as a validation set. In [13] genome-wide measurements of copy number variations were performed by array CGH (UCSF Hum. Array 2.0) with an average spacing between probes of 1Mb. The study contained 180 patients diagnosed with a stage I/II lymph node-negative breast cancer. The data set was downloaded from the GEO data base with accession number GSE6448. Arrays were preprocessed by averaging/removing probes as follows: 18 clones mapping to chromosome Y or missing genomic location information were removed, 80 probes mapping to identical genomic regions were averaged and represented as single values, 179 probes missing entries for 30% or more patients were removed, and missing values were imputed using the lowest regression method in the aCGH package for R [16]. This resulted in 2,168 unique clones from the original 2,445 printed in the array. We classified as ERBB2+ tumors the subset of 9 patients that showed a copy number change >1 (in log scale) at the clone DMPC-HFF#1-61H8 which contains the ERBB2 gene.

2.5 Multidimensional Analysis of CGH Profiles Using Computational Algebraic Topology

We previously reported a new method to analyze CGH data called topological analysis of array CGH (TAaCGH) [3, 15]. Our method uses a sliding window algorithm that associates a point cloud to a given CGH profile (or section of a CGH profile). The dimension of the point cloud is determined by the size of the sliding window. In this study and based on our previous work [3] we considered windows of size $n = 2$. TAaCGH assigns a β_0 curve to each CGH profile, computes the average $\langle \beta_0 \rangle$ curve for each population of patients (test and control) and performs statistical analysis to determine differences between them (see below). Here we extended TAaCGH by incorporating a similar analysis using $\langle \beta_1 \rangle$ curves. We used the program JavaPlex to perform the calculation of β_1 and its generators [40]. As in the case of β_0 , we generated the function $\beta_1(\epsilon)$ for each patient. In this case ϵ took values between 0 and the value at which $\beta_0 = 1$. Given the $\beta_1(\epsilon)$ for each patient, we computed the average $\langle \beta_1 \rangle$ for the ERBB2 set and the control set (consisting of the remainder of the patients) and test for statistically significant differences between the two $\langle \beta_1 \rangle$ curves.

2.6 Testing for Statistical Differences

To test for statistically significant differences between $\langle \beta_i \rangle$ curves associated to different patient groups, we assumed the null hypothesis that $\langle \beta_i \rangle$ curves

for a sample of patients was independent of the cancer subtype. We quantified deviations from the null distribution by the statistic S_{exp} , which was defined as the sum of the squares of the differences between the average $\langle\beta_i\rangle$ curves across all radii, i.e.

$$S_{exp,i} = \sum (a_{ij} - b_{ij})^2 \quad \text{for } j = 1, \dots, N$$

where a_{ij} and b_{ij} are the $\langle\beta_i(j)\rangle$ value for each population under study and for the value of the filtration parameter $\epsilon = j$.

2.7 Finding Co-Occurring Aberrations

In order to determine the regions of the genome that contributed to the first homology we found the CGH probes that were mapped to each of the vertices of the generators. First, generators for each patient and value of the filtration coefficient were calculated using JavaPlex [40]. Second, the probes of the CGH profile that mapped to the vertices of the generators were identified. Third, since generators were not necessary minimal and, due to the noise of the data, some generators mapped to different areas of the genome we determined a CNA by measuring the concentration of the probes. Regions with higher concentration of probes than the control set were called CNAs.

2.8 Software for Visualization of Generators

We created an exploratory tool using Shiny app to visualize the generators in the point cloud together with their corresponding probes in the CGH profile. The app highlights the probes and generators as the values of the filtration coefficient changes. The software allows to visualize the dispersion of the probes associated with the probes through the CGH profile. An example is shown in Fig. 5. The software is available from the authors upon request.

3 Results

3.1 Computer Simulations

To better interpret our results we performed computer simulations. Since the analysis of β_0 has been performed elsewhere [3, 15], we focused on simulations concerning the detection of CNAs using β_1 . Figure 2 shows an example of two simulated profiles, one with no aberrations as control (Fig. 2) and a second one with two co-occurring aberrations (Fig. 2B). In both Fig. 2A and B, the x-axis represents the position along the chromosome and the y-axis the \log_2 ratio of the copy number values. The $\langle\beta_1\rangle$ curves (Fig. 2C) obtained from the curves above help understand the growth and disappearance of the first homology. In the case of no amplification (red), the $\langle\beta_1\rangle$ curve starts at $\langle\beta_1\rangle = 0$, since for very small values of ϵ there is no 1-dimensional homology. $\langle\beta_1\rangle$ rapidly increases due to the structure of the noise until it reaches a maximum after which it decays to 0.

The graph for $\langle\beta_1\rangle$ is different when two aberrations are present (blue). For small values of the filtration parameter the graph behaves similarly to the graph without aberrations, however in this case the graph shows more than one local maximum and a lower \log_2 ratio of copy number values at the first maximum.

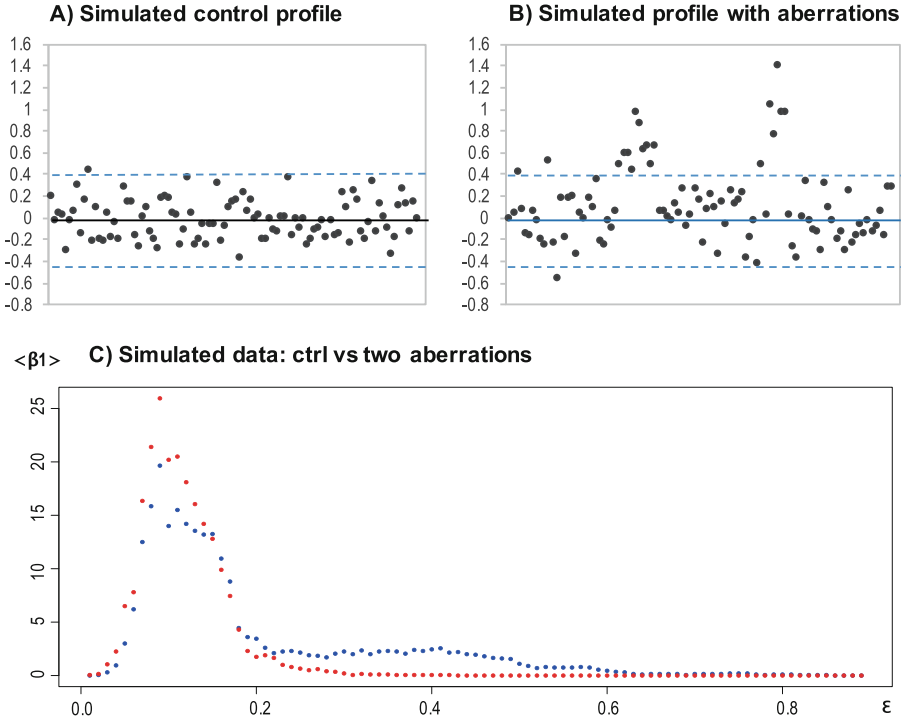


Fig. 2. Examples of simulated aberration profiles and $\langle\beta_1\rangle$ curve. (A) shows a control profile with no aberrations with $\sigma=0.2$. (B) shows a profile with two aberrations with parameters $\lambda = 10$ and 5 , $\mu = 0.6$ and 1 and $\sigma = 0.2$ for both. The blue dashed lines represent two standard deviations. The bottom graph shows in red the $\langle\beta_1\rangle$ for the control group with no aberrations and in blue the $\langle\beta_1\rangle$ curve for a pair of aberrations with $\lambda = 10$, $\mu = 0.6$ and 1 and $\sigma = 0.2$ (Color figure online).

We tested our method by performing a sensitivity and specificity analysis in three different simulation experiments. Each experiment consisted of 200 profiles (100 tests and 100 controls) and all possible combinations of parameters were considered. A successful identification of an aberration was scored when the obtained P-value was less than 0.05 after correcting by FDR. First we considered the case of one single amplification (test set) taking as control set a population with no aberrations. In this case sensitivity was 87.5%. In the second experiment we used profiles with two amplifications as a test set and no amplifications as the control set. In this experiment we got average sensitivity of 95%. In the third

experiment we compared double amplifications with single (as control). Results showed 82.5% in sensitivity. Specificity was measured by comparing two control data sets resulting in 97.5%. Our method has bigger chances to fail when the length of the aberration is small (5 or less) and $\mu = \sigma$.

β_0 Significance of 17q

As discussed elsewhere [3, 15], $\langle \beta_0 \rangle$ curves can detect chromosome aberrations. Since we are interested in the entire amplicon in 17q, we applied TAaCGH to full chromosome arms. The chromosome arm 17q was significant in both data sets. In the Horlings data set we found significance on $\langle \beta_0 \rangle$ curves when comparing chromosome arm 1q (P-value = 0.021) and 17q (P-value = 0.004). The graph for chromosome 1q however showed that the control curve was above the test set indicating that the control set (ERBB-) had more CNAs than the test set (ERBB+). Therefore was not relevant in this study. In our validation data set, we found only 17q to be significant with a corresponding P-value after FDR correction of 0.0037. Figure 3 shows examples of $\langle \beta_0 \rangle$ curves for both chromosomes. Since β_0 is the number of connected components of the simplicial complex, $\langle \beta_0 \rangle$ curves start at the value of the number of probes in each chromosome arm for $\epsilon = 0$ and gradually decays with increasing ϵ until a single connected component remains. All blue curves shown in Fig. 3 represent the ERBB2+ population and all red curves represent the ERBB2- population. Results shown in Fig. 3A and B include $\langle \beta_0 \rangle$ curves associated to 17q for the Climent and Horlings data sets respectively; Fig. 3C shows $\langle \beta_0 \rangle$ curves associated to 1q and Fig. 3D $\langle \beta_0 \rangle$ curves associated to the negative control 19q. Chromosome arm 17q showed, as expected, a higher number of chromosome aberrations in the ERBB2+ patients than in the ERBB2- patients.

β_1 Significance of 17q

Next, we analyzed the significance of β_1 in chromosome arm 17q. We considered two approaches. First we tested for β_1 significance of the entire chromosome arm 17q and then for overlapping sections of the chromosome arm. We found important to use both approaches since co-occurring CNAs may be local or spread over the entire arm. Analysis using the whole arm showed 17q to be significant in the Climent data set (with a P-value of 0.040), but not in the Horlings data set (P-value 0.172). Figure 4 shows the corresponding $\langle \beta_1 \rangle$ curves for both studies suggesting that any amplicon structure, if present, would be local.

Following our previous work [3] we subdivided chromosome arm 17q in the Horlings data set into 6 sections, which corresponded to 5 sections in the Climent data set. Each section containing 20 CGH probes with 10 overlapping probes. Results are shown in Table 1. Column 1 shows the section analyzed; columns 2 and 5 the cytogenetic band, columns 3 and 6 the location in base pairs, and columns 4 and 7 the p-values [7]. Both data sets showed some significant sections. In the Horlings data set, Sects. 2 and 3 significant after correction for multiple testing (column 4). In the Climent data set all sections except Sect. 4 were significant (column 7). Based on the reproducibility of these results we concluded

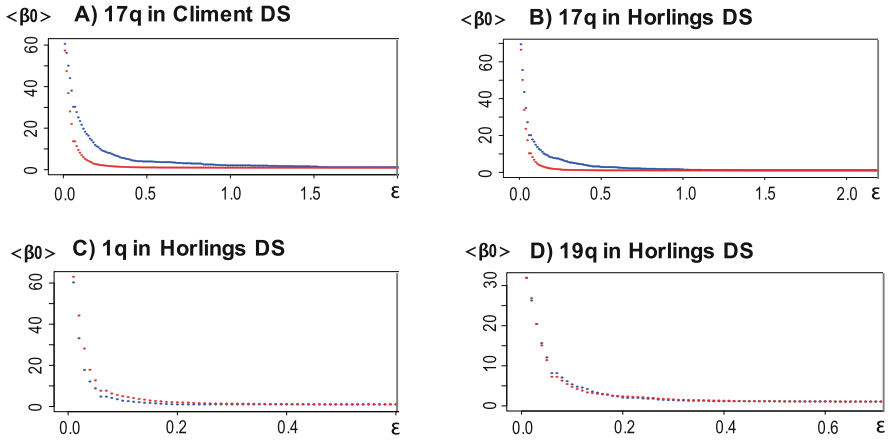


Fig. 3. Examples of $\langle \beta_0 \rangle$ curves in dimension 20. Blue indicates the ERBB2+ population and red the ERBB2-. (A) Arm 17q arm in Climent; (B) Arm 17q in Horlings, (C) Arm 1q in Horlings and (D) Arm 19q in Horlings (Color figure online).

that sections containing cytobands 17q12 to 17q21.33 had co-occurring CNAs and are therefore good candidates for uncovering the underlying structure of the amplicon.

To further identify the regions within 17q12 and 17q21.31 – 17q21.33 we identified the generators of the first homology group for each patient and mapped the probes to the vertices of the corresponding generators. Before we discuss the statistical results we highlight some interesting properties of the generators: (1) probes that made up the generators may be distributed throughout the entire

Table 1. Chromosome Sections. Correspondence between sections, cytobands and base pairs range for each of the sections used to analyze chromosome 17q.

Section	Cytoband (Horlings <i>et al.</i>)	Basepair	(P -value) FDR correction	Cytoband (Climent <i>et al.</i>)	Basepair	(P -value) FDR Correction
17q.s1	q11.1-q12	25440972- 37812853	(0.043) 0.08640	q11.1-q21.2	25530227- 40615955	0.0088
17q.s2	q12-q21.31	32489785- 43339849	(0.0008) 0.00480	q12-q21.33	35669421- 47644854	0.0016
17q.s3	q21.2-q21.33	38428492- 49075570	(0.0116) 0.03480	q21.31-q22	42170022- 55594526	0.0378
17q.s4	q21.31-q22	44084882- 57340119	(0.471) 0.47170	q21.33-q24.3	47968636- 70573094	0.100
17q.s5	q22-q24.2	51080264- 66108804	(0.253) 0.30432	q23.1-q25.3	58025830- 78774742	0.009
17q.s6	q23.1-q25.3	57996713- 80780814	(0.237) 0.30432			

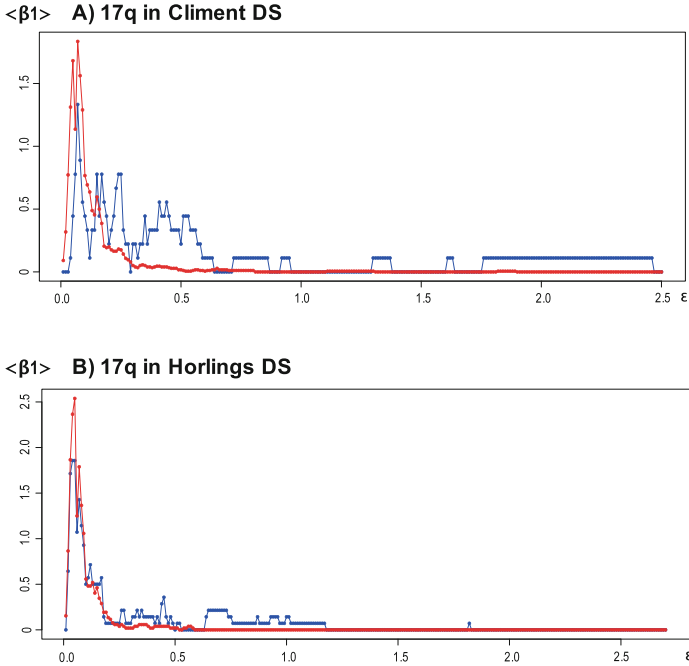


Fig. 4. $\langle \beta_1 \rangle$ Significance of $17q$ in the climent and the horlings data sets. (A) The figure shows the $\langle \beta_1 \rangle$ curves for ERBB2+ (blue) and ERBB2- (red) in the Climent data set (significant). (B) Here we show the $\langle \beta_1 \rangle$ curves for both categories for the Horlings data set (non-significant) (Color figure online).

arm or localized in a specific region (2) unlike β_0 generators do not necessarily detect the global maximum in the profile but different regions that contribute to several local maxima (3) neighboring maxima or even sections of the same maximum are detected at different values of the filtration parameter. Figure 5 shows the profile of a patient for $17q$ and the point cloud. Probes in blue are those that were mapped to the generators at two different filtration coefficient values. The corresponding 2D point cloud (with edges included) and with the vertices in each cycle highlighted in blue are also shown.

These inherent variability of the generators and the noise of the data motivated us to use a statistical approach. As detailed in the methods sections for each patient and value of the filtration parameter we computed the cycles and the probes that defined those cycles. The frequency at which a probe was mapped to a particular region of the genome is represented by a histogram (see Fig. 6). The top graphs show the histograms for the Horlings data set and the bottom ones the histograms for the Climent data set. The histograms on the left are the control and the ones on the right correspond to the ERBB2+. The most remarkable feature is the difference between the control and the ERBB2 data sets. While the control show no significant concentration of the probes that

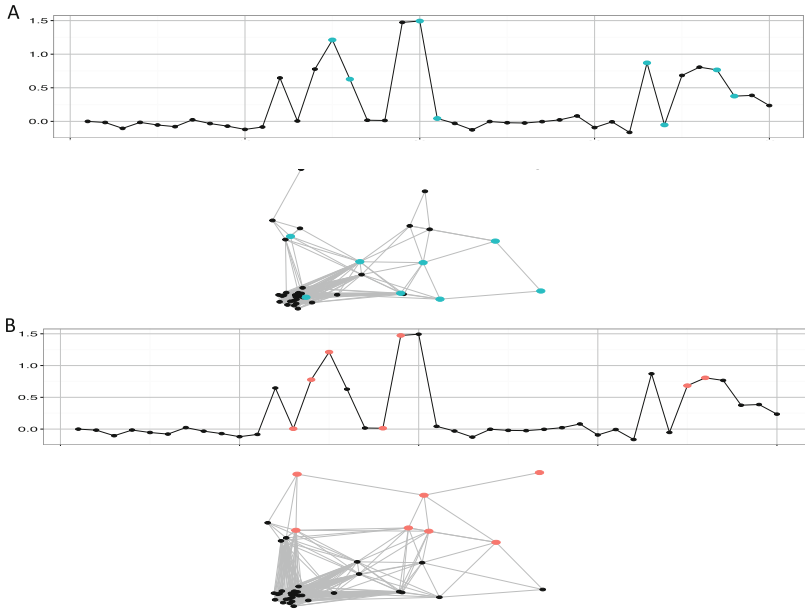


Fig. 5. Correspondence between CGH probes and generators. Different values of the filtration parameter detects different generators which corresponds to different probes in the genome. Panel A shows the profile of one patient and its associated point cloud. The probes highlighted in blue correspond to the vertices of the single generator, also in blue. The filtration coefficient was $\epsilon = 0.78$. Panel B shows the same patient and point cloud for a different value of the filtration coefficient $\epsilon = 0.83$

belong to cycles the ERBB2+ clearly show three regions of interest. 17q12 has a significant concentration of cycle elements and corresponds to the position of the gene *ERBB2*. Two regions extend beyond the position of *ERBB2*. The first one is in the boundary between 17q21.2 and 17q21.31. The Horlings data set suggests that the region of interest is more localized in 17q21.31 while the Climent data set suggest a region contained in 17q21.2. The last region is located at 17q21.33 and is common to both studies.

Since our simulations show that the first homology group can also identify single amplifications one may argue that the found amplifications correspond to single independent events. To address this problem we analyzed the distribution of the cycles-forming-probes. Figure 7 show some examples of the distribution of cycles in the genome for specific patients. Each plate corresponds to one patient, the x-axis is the position along the genome and the y-axis the “life” of the cycle. Each color represents a different cycle. If the amplifications were independent events one would expect to see single colors concentrated at specific regions. However we see cycles dispersed over the entire profile indicating the presence of co-occurring CNAs.

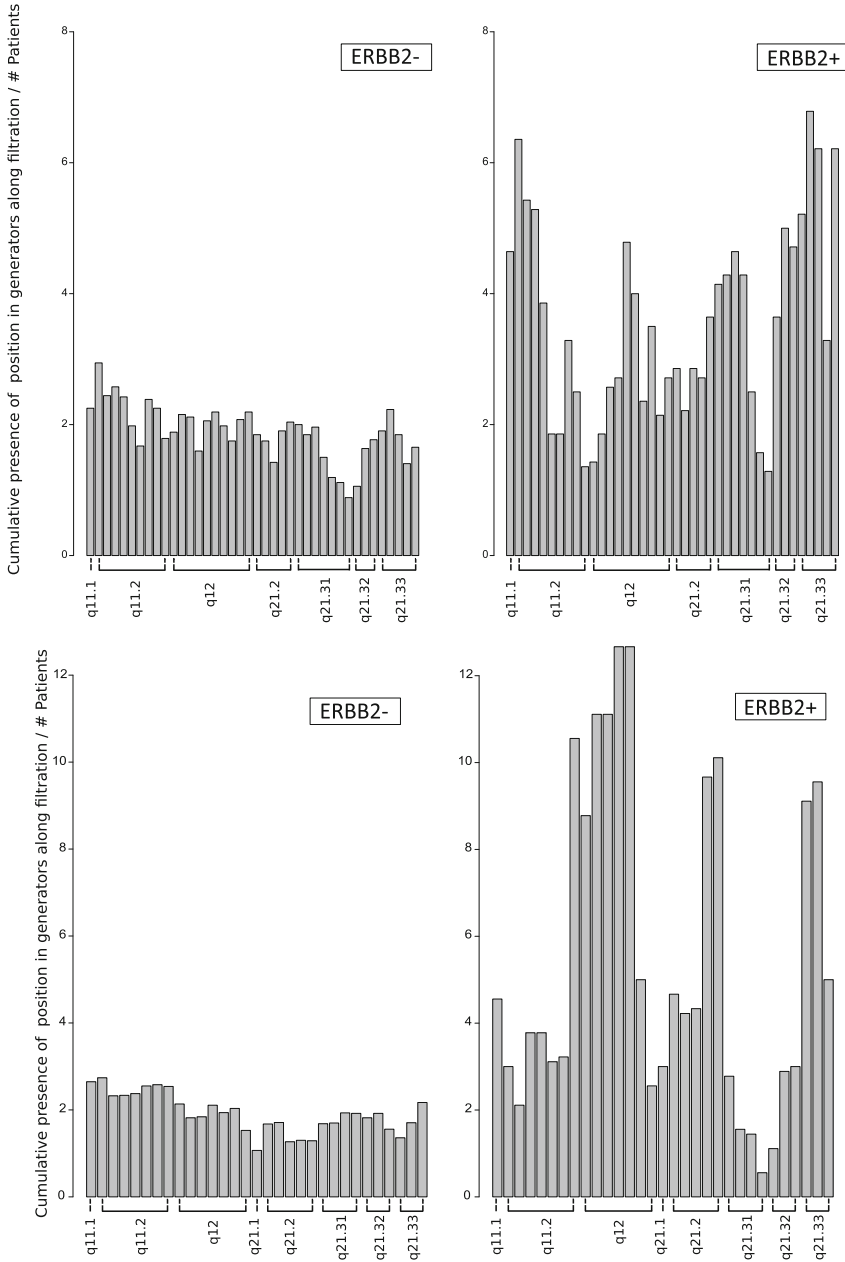


Fig. 6. Comparison of ERBB2- (left) and ERBB2+ (right) patients at the generator level. The top histograms correspond to the Horlings data set and the bottom to the Climent data set. Each bar in the histogram represents a probe. Its height represents the cumulative presence of that probe on the generators of the first homology group divided by the number of patients. The cumulative presence is calculated by counting the number of cycles in which the probe is part of the generator for each value of the filtration parameter (multiplied by the number of generators if they were more than one).

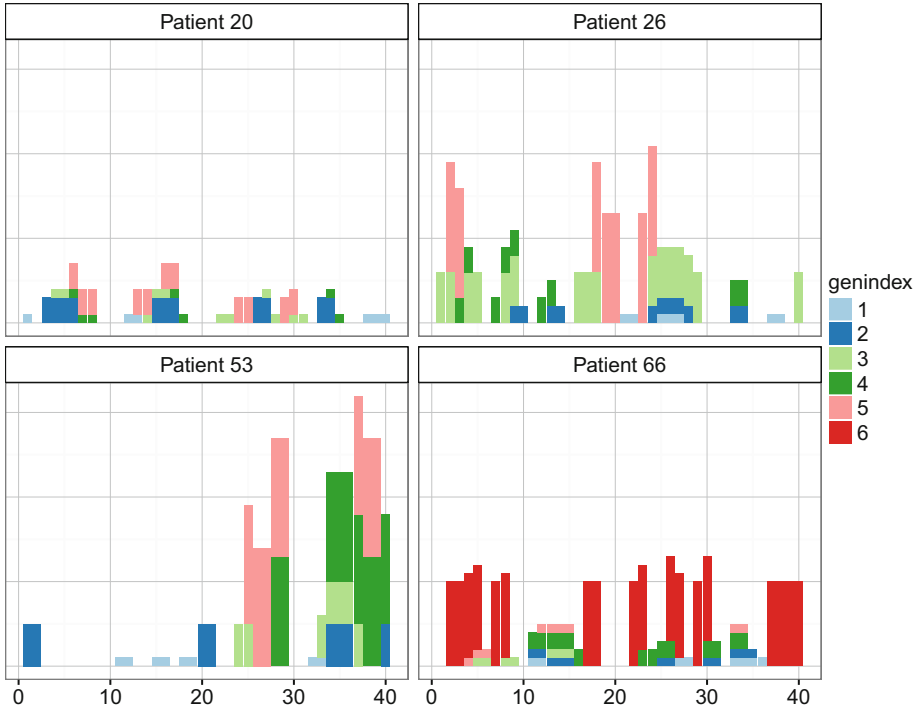


Fig. 7. Distribution of cycles in CGH profiles. Each plate corresponds to the CGH profile of a patient and how the vertices of the cycles are mapped back to the profile. Different colors indicate different cycles and do not represent the same cycle in each plate. The height of the bars represent the life of the cycle (Color figure online).

4 Discussion

Copy number measurements provide an unparalleled opportunity to identify the underlying mechanisms of cancer. Previous efforts in analyzing copy number data have mainly focused on the identification of single, independent chromosome copy number aberrations. These approaches however are known to be deficient in the identification of co-occurring copy number changes since there is a large number of combinations of probes that one needs to interrogate. In this study, we have presented a methodology that helps circumvent the search for simultaneously occurring CNAs by encoding copy number data as topological objects. In particular we have used the rank of the first homology group to perform this association. To test this hypothesis, we searched for co-occurring aberrations in ERBB2+ breast cancer patients. Our results show β_1 significance in chromosome cytobands that extend from 17q12 to 17q21.33. By identifying the probes that form the generators and measuring their concentration along the CGH profiles we were able to further narrow this significant region to three amplifications. The first is 17q12 which contains the ERBB2 gene. The second and the third

have also been reported in ERBB2+ patients. The second amplification is in the boundary between 17q21.2 and 17q21.31 and according to our estimation is delimited by the Top2A and BRCA1 genes (base pairs 40, 884763 – 41, 826, 877). This region encompasses the type I keratin gene cluster. Finally we identified 17q21.33 (base pairs 47, 400, 368 – 49, 075570) a large region that contain multiple tumor associated genes including the HOXB cluster [42], Prohibitin [44] and amplification of this region has been associated with poor prognosis [41]. Unfortunately at this point, due to the small sample size, we cannot determine how common these co-occurring CNAs are in the general population of ERBB2+ patients or whether they form subtypes within the ERBB2+ subtype. Nevertheless the fact that these regions are significant in two independent data sets is encouraging. It is therefore our immediate plan to scale up this study on larger data sets.

Our work presents also new tools for the topological analysis of time series. We and others [34] independently introduced the concept of using the sliding window algorithm to analyze time series. In our previous work we noted that: (1) the overall shape of the point cloud already provides information of the data [2, 3, 15], (2) The point cloud can be seen as the reconstruction set of the dynamical system induced by the sliding window algorithm [2], (3) the zero homology group identifies large step increments between consecutive measurements [15]. Our contributions in this work is the development of algorithms that (1) detect the single and co-occurring maxima in the data in non-necessarily periodic signals using the first homology group (2) Identify local maxima by computing the concentration of the pre-images (by the sliding window algorithm) of the vertices that form the cycles. It is our belief that the use of topological methods for the analysis of signals using simple construction techniques, such as the commonly used sliding window algorithm, can provide new insights in the analysis of time series.

Acknowledgments. We would like to thank H. Bengtsson and T. Speed for very helpful comments during the development of this methodology. T.B and J.A. were partially supported by NSF grant 1217324 and by NIH-RIMI (Research Infrastructure in Minority Institutions) grant 2P20MD000544-06. SA was partially supported by the *Ministerio de Economía y competitividad* grant MTM2013-42486-P.

References

1. Arriola, E., Marchio, C., Tan, D.S., et al.: Genomic analysis of the HER2/TOP2A amplicon in breast cancer and breast cancer cell lines. *Lab Invest.* 88(5), 491–503
2. Arsuaga, J., Baas, N.A., DeWoskin, D., et al.: Topological analysis of gene expression arrays identifies high risk molecular subtypes in breast cancer. *Appl. Algebra Eng. Commun. Comput.* **23**(1), 3–15 (2012)
3. Arsuaga, J., Borrman, T., Cavalcante, R., Gonzalez, G., Park, C.: Identification of copy number aberrations in breast cancer subtypes using persistence topology. *Microarrays* **4**(3), 339–369 (2015)

4. Barlund, M., Tirkkonen, M., Forozan, F., Tanner, M.M., Kallioniemi, O., Kallioniemi, A.: Increased copy number at 17q22-q24 by CGH in breast cancer is due to high-level amplification of two separate regions. *Genes Chromosom. Cancer.* **20**(4), 372–376 (1997)
5. Barrett, M.T., Anderson, K.S., Lenkiewicz, E., et al.: Genomic amplification of 9p24.1 targeting JAK2, PD-L1, and PD-L2 is enriched in high-risk triple negative breast cancer. *Oncotarget* **6**(28), 26483–26493 (2015)
6. Bengtsson, H., Ray, A., Spellman, P., Speed, T.P.: A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics* **25**(7), 861–867 (2009)
7. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**(1), 289–300 (1995)
8. Bhatlekar, S., Fields, J.Z., Boman, B.M.: HOX genes and their role in the development of human cancers. *J. Mol. Med. (Berl)* **92**(8), 811–823 (2014)
9. Bilal, E., Vassallo, K., Toppmeyer, D., et al.: Amplified loci on chromosomes 8 and 17 predict early relapse in ER-positive breast cancers. *PLoS One* **7**(6), e38575 (2012)
10. Cavalcante, R.: Using Homology and networks to locate copy number aberrations associated to recurrence in breast cancer. MA Thesis, San Francisco State University (2012)
11. Chin, K., DeVries, S., Fridlyand, J., Spellman, P.T., Roydasgupta, R., et al.: Genomic and transcriptional aberrations linked to breast cancer pathophysiology. *Cancer Cell* **10**, 529–541 (2006)
12. Ching, H.C., Naidu, R., Seong, M.K., Har, Y.C., Taib, N.A.: Integrated analysis of copy number and loss of heterozygosity in primary breast carcinomas using high-density SNP array. *Int. J. Oncol.* **39**(3), 621–633 (2011)
13. Climent, J., Garcia, J.L., Mao, J.H., Arsuaga, J., Perez-Losada, J.: Characterization of breast cancer by array comparative genomic hybridization. *Biochem Cell Biol.* **85**(4), 497–508 (2007)
14. Desmedt, C., Voet, T., Sotiriou, C., Campbell, P.J.: Next-generation sequencing in breast cancer: first take home messages. *Curr Opin. Oncol.* **24**(6), 597–604 (2012)
15. DeWoskin, D., Climent, J., Cruz-White, I., Vazquez, M., Park, C., et al.: Applications of computational homology to prediction of treatment response in breast cancer patients. *Topology Appl.* **157**, 157–164 (2010)
16. Fridlyand, J., Dimitrov, P.: aCGH: Classes and functions for Array Comparative GenomicHybridization data. R package version 1.34.0
17. Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G., Jain, A.N.: Hidden Markov models approach to the analysis of array CGH data. *J. Multivar. Anal.* **90**, 132–153 (2004)
18. Fridlyand, J., Snijders, A.M., Ylstra, B., Li, H., Olshen, A., et al.: Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer* **6**, 96 (2006)
19. Green, M.R., Monti, S., Rodig, S.J., et al.: Integrative analysis reveals selective 9p24.1 amplification, increased PD-1 ligand expression, and further induction via JAK2 in nodular sclerosing Hodgkin lymphoma and primary mediastinal large B-cell lymphoma. *Blood* **116**(17), 3268–3277
20. Horlings, H.M., Lai, C., Nuyten, D.S.A., et al.: Integration of DNA copy number alterations and prognostic gene expression signatures in breast cancer patients. *Clin Cancer Res.* **16**(2), 651–663 (2010)

21. Horlings, H.M., Lai, C., Nuyten, D.S.A., et al.: Supplementary Data. *Clin. Cancer Res.* **16**(2), 651–663 (2010b). <http://clincancerres.aacrjournals.org/content/16/2/651/suppl/DC1>
22. Hupe, P., Stransky, N., Thiery, J.P., Radvanyi, F., Barillot, E.: Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**(18), 3413–3422 (2004)
23. Jacot, W., Fiche, M., Zaman, K., Wolfer, A., Lamy, P.J.: (2013) The HER2 amplicon in breast cancer: Topoisomerase IIA and beyond. *Biochim. Biophys. Acta.* **1**, 146–157 (1836)
24. Jonsson, G., Staaf, J., Vallon-Christersson, J., Ringner, M., Holm, K., et al.: Genomic subtypes of breast cancer identified by array comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res.* **12**(3), R42 (2010)
25. Lai, W.R., Johnson, M.D., Kucherlapati, R., Park, P.J.: Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* (2005). doi:10.1093/bioinformatics/bti611
26. Lai, C., Horlings, H., van de Vijver, M.J., et al.: SIRAC: supervised identification of regions of aberration in aCGH datasets. *BMC Bioinform.* **8**, 422 (2007)
27. Latham, C., Zhang, A., Nalbanti, A., et al.: Frequent co-amplification of two different regions on 17q in aneuploid breast carcinomas. *Cancer Genet. Cytogenet.* **127**(1), 16–23 (2001)
28. Leiserson, M.D., Vandin, F., H-T, Wu, et al.: Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015)
29. Mahmood, S.F., Gruel, N., Chapeaublanc, E., et al.: A siRNA screen identifies RAD21, EIF3H, CHRAC1 and TANC2 as driver genes within the 8q23, 8q24.3 and 17q23 amplicons in breast cancer with effects on cell growth, survival and transformation. *Carcinogenesis* **35**(3), 670–682 (2014)
30. Martin-Castillo, B., Lopez-Bonet, E., Bux, M., et al.: Cytokeratin 5/6 fingerprinting in HER2-positive tumors identifies a poor prognosis and trastuzumab-resistant basal-HER2 subtype of breast cancer. *Oncotarget* **6**(9), 7104–22 (2015)
31. Niyogi, P., Smale, S., Weinberger, S.: Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39**, 419–441 (2008)
32. Nielsen, K.V., Muller, S., Miller, S., Schonau, A., Balslev, E., Knoop, A.S., Ejlertsen, B.: Aberrations of ERBB2 and TOP2A genes in breast cancer. *Mol. Oncol.* **4**(2), 161–168 (2010)
33. Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M.: Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**(4), 557–572 (2004)
34. Perea, J., Harer, J.: Sliding windows and persistence: An application of topological methods to signal analysis. *Found. Computat. Math.* **15**(3), 799–838
35. Perou, C., Borresen-Dale, A.L.: Systems biology and genomics of breast cancer. *Cold Spring Harbor Perspect. Biol.* **3**, a003293 (2011)
36. Pinkel, D., Albertson, D.G.: Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.* **37**(Suppl), S11–S17 (2005)
37. Rauta, J., Alarmo, E.L., Kauraniemi, P., et al.: The serine-threonine protein phosphatase PPM1D is frequently activated through amplification in aggressive primary breast tumours. *Breast Cancer Res. Treat.* **95**(3), 257–263 (2006)
38. Rebouh: Exploring topological methods to study topological imbalance in breast cancer. San Francisco State University MA thesis (2012)

39. Sinclair, C.S., Rowley, M., Naderi, A., Couch, F.J.: The 17q23 amplicon and breast cancer. *Breast Cancer Res. Treat.* **78**(3), 313–322 (2003)
40. Tausz, A., Vejdemo-Johansson, M., Adams, H.: JavaPlex: A research software package for persistent (co)homology. In: Hong, H., Yap, C. (eds.) *Mathematical Software – ICMS 2014*. LNCS, vol. 8592, pp. 129–136. Springer, Heidelberg (2014)
41. Thompson, P.A., Brewster, A.M., Kim-Anh, D.: Selective genomic copy number imbalances and probability of recurrence in early-stage breast cancer. *PLoS One* **6**(8), e23543 (2010)
42. Torresan, C., Oliveira, M.M., Pereira, S.R., et al.: Increased copy number of the DLX4 homeobox gene in breast axillary lymph node metastasis. *Cancer Genet.* **207**(5), 177–187 (2014)
43. Ulz, P., Heitzer, E., Speicher, M.: Co-occurrence of MYC amplification and TP53 mutations in human cancer. *Nat. Genet.* **48**(2), 104–106 (2016)
44. Webster, L.R., Provan, P.J., Graham, D.J., et al.: Prohibitin expression is associated with high grade breast cancer but is not a driver of amplification at 17q21.33. *Pathology* **45**(7), 629–636 (2013). doi:[10.1097/PAT.0000000000000004](https://doi.org/10.1097/PAT.0000000000000004)
45. Willenbrock, H., Fridlyand, J.: A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**(22), 4084–4091 (2005)
46. Wilkerson, P.M., Reis-Filho, J.S.: The 11q13-q14 amplicon: clinicopathological correlations and potential drivers. *Genes Chromosom. Cancer* **52**(4), 333–355 (2013)
47. Zhou, X., Rao, N.P., Cole, S.W., Mok, S.C., Chen, Z., Wong, D.T.: Progress in concurrent analysis of loss of heterozygosity and comparative genomic hybridization utilizing high density single nucleotide polymorphism arrays. *Cancer Genet. Cytogenet* **159**(1), 53–57 (2005)