Prediction in cancer genomics using topological signatures and machine learning

Conference Paper · January 2019

CITATIONS
3

READS
466

4 authors:

Georgina Gonzalez
University of California, Davis
5 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Arina Ushakova
University of California, Davis
2 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Radmila Sazdanovic
North Carolina State University
71 PUBLICATIONS **401** CITATIONS
SEE PROFILE

Javier Arsuaga
University of California, Davis
74 PUBLICATIONS **2,055** CITATIONS
SEE PROFILE

Prediction in cancer genomics using topological signatures and machine learning

Georgina Gonzalez[1], Arina Ushakova[2], Radmila Sazdanovic[3], and Javier Arsuaga[1,4,*]

[1]Department of Molecular and Cellular Biology, University of California, Davis, CA 95616
[2]Department of Statistics, University of California, Davis, CA 95616
[3]Department of Mathematics, North Carolina State University, NC 27695
[4]Department of Mathematics, University of California, Davis, CA 95616
[*]jarsuaga@ucdavis.edu

Abstract

Copy Number Aberrations, gains and losses of genomic regions, are a hallmark of cancer and can be experimentally detected using microarray comparative genomic hybridization (aCGH). In previous works, we developed a topological based method to analyze aCGH data whose output are regions of the genome where copy number is altered in patients with a predetermined cancer phenotype. We call this method Topological Analysis of array CGH (TAaCGH). Here we combine TAaCGH with machine learning techniques to build classifiers using copy number aberrations. We chose logistic regression on two different binary phenotypes related to breast cancer to illustrate this approach. The first case consists of patients with over-expression of the ERBB2 gene. Over-expression of ERBB2 is commonly regulated by a copy number gain in chromosome arm 17q. TAaCGH found the region 17q11-q22 associated with the phenotype and using logistic regression we reduced this region to 17q12-q21.31 correctly classifying 78% of the ERBB2 positive individuals (sensitivity) in a validation data set. We also analyzed over-expression in Estrogen Receptor (ER), a second phenotype commonly observed in breast cancer patients and found that the region 5p14.3-12 together with six full arms were associated with the phenotype. Our method identified 4p, 6p and 16q as the strongest predictors correctly classifying 76% of ER positives in our validation data set. Although for this set there was a significant increase in the false positive rate (specificity). We suggest that topological and machine learning methods can be combined for prediction of phenotypes using genetic data.

1 Introduction

The cancer genome is characterized by chromosome instability and the formation of chromosome aberrations [40]. Copy number aberrations, that is amplifications and deletions of genomic regions, are particularly relevant in tumor development because they may house proto-oncogenes and tumor suppressor genes. Aberrations containing these genes can be used as prognosis tools [1, 11, 13, 43] but they may be difficult to identify because they are usually accompanied by many passenger aberrations and because they may be hidden by experimental noise. Experimentally the number of copies of the genome can be measured using array comparative genomic hybridization platforms (aCGH) and sequencing (DNAseq) [47]. A number of statistical methods have been proposed to detect copy number changes, these include [6, 12, 18, 23, 29, 30].

In previous works, we proposed a topology based method to identify candidate driver chromosome aberrations, called Topological Analysis of array CGH (TAaCGH). TAaCGH is different from other methods in that it: (1) does not perform a single segmentation of the data but a sequence of segmentations, (2) uses relationships between consecutive probes to determine significance of genomic fragments, (3) identifies copy number changes associated with a specific phenotype, and

(4) allows to detect single [3, 15] and some co-occurring copy number aberrations [2] .

The next step in the development of TAaCGH is determining to what extent the identified genomic regions can be used as patient classifiers. In genetic association studies, machine learning techniques like logistic regression, random forest or support vector machine are often used for classification and feature selection [48, 26, 50, 20, 25]. However, several issues arise that make predictive models challenging for microarray data. For example, data usually consist of a much larger number of co-variates (genotypes) than observations (patients) and copy number data contain numerous highly correlated neighboring probes (co-variates). Additionally, some traits are known to be regulated by many interacting genetic regions located across the genome. Adding a complexity penalty to the loss function (regularization) or using methods such as group Lasso [33] that takes into consideration the correlation among features when assigning the penalties, are some of the approaches used to address these issues; but many of these approaches continue to be affected by correlation bias [48].

In this analysis we introduce a predictive model for binary traits using the output of TAaCGH [3] as a starting set of candidate co-variates. We tested this approach on two data sets consisting of breast cancer patients with different clinical characteristics. Data from [21] was used as a training set and data from [10] as a validation/test set. Here we report our results on two clinical characteristics: over-expression of ERBB2 (denoted by ERBB2+) and of the estrogen receptor gene (ER+). In the ERBB2+ study, TAaCGH found the region 17q11-q22 to be significantly associated with this phenotype, but not with other molecular subtypes like luminals or basals. The region of the genome originally consisted of two sections, when using them as co-variates on data from [21] only, one section was enough for prediction; shrinking the relevant area to 17q12-q21.31 and obtaining a sensitivity of 64% (specificity=96%). When tested in the validation set [10] we obtained sensitivity of 78% (specificity=90%). These results suggest that this section of the genome, which contains the gene ERBB2, discriminates better the true negatives than the true positives. This is most likely due to the fact that over-expression of ERBB2 is not always regulated by a copy number change [9, 51]. In the case of ER+, TAaCGH found section 5p14.3-12 and arms 4p, 5q, 6p, 10q, 16p and 16q in the training set. These regions were validated by either SIRAC [30] or through our validation data set [10]. Our logistic regression study identified 4p, 6p and 16q as the best predictors. Interestingly none of these arms contains the Estrogen Receptor gene (ESR1) suggesting that copy number changes do not regulate the expression of this gene in breast cancer. This model for ER+ had a sensitivity of 79% (specificity=79%). When we validated the model on [10], we obtained a sensitivity of 79% (specificity=52%). Reduction on the specificity might be due to biological differences or differences in the structure between the training and the validation data sets. Based on our results, we suggest that the proposed version of TAaCGH, extended via topological signatures as classifiers, can further provide a framework for other one-class classification methods, and that its expanded capabilities may to be useful for analyzing other phenotypes and genetic interactions.

2 Methods

2.1 Data

Array Comparative Genome Hybridization (aCGH) data measure the difference in the number of DNA copies between a test and a reference sample for regions along the genome. These data are therefore commonly presented as a log-transformed ratio of the two quantities. A log-transform value greater than a threshold >0 indicates an amplification of the genome, while negative numbers signal deletions. Since the physical position along the genome is known for each probe; the $\log_2$ ratio is mapped back to the genome defining what we call the patient's *CGH profile*.

2.1.1 Simulation data

We used simulations to estimate the statistical properties of TAaCGH and of our proposed classification method. We simulated a series of experiments on data sets containing 100 profiles (50 tests and 50 controls); each profile was aimed at recreating a section of the genome with 50 probes. Copy number values for probes in the control group and for probes in the test group that did not belong to a chromosome aberration were drawn from a normal distribution $N(\mu = 0, \sigma_{Ctrl})$. The value of $\sigma \in \{0.2, 0.6\}$ was fixed in any given simulation. Each simulated copy number aberration was determined by three parameters: the mean and standard deviation from a normal distribution $N(\mu, \sigma)$ and the length λ, in probes, of the aberration. For the first, we considered $\mu = 1$ and σ as the test group having aberration's length $\lambda \in \{5, 10, 25\}$. Additionally, and motivated by the fact that the predictor variable is not always present in the test group, we also allowed the number of aberrant profiles within the test group to vary. We called this parameter mix. In our simulations $mix \in \{20\%, 40\%, 60\%, 80\%\}$. In each simulation we tested for specificity and sensitivity of the method for a predetermined combination of parameters $\mu, \sigma, \lambda, mix$.

2.1.2 Horlings data set

As in previous studies, we used the data set published by Horlings and colleagues [21, 22]. BAC Microarrays covered the entire genome with a spacing average of 1 Mb and each BAC clone was spotted in triplicate on every slide (Code Link Activated Slides, Amersham Biosciences). This sample contained a total of 66 patients, 14 of which were ERBB2+ and 38 were ER+. Both phenotypes were determined by clinical diagnosis. The control set consisted of: patient profiles belonging to the remaining cancer patients with ERBB2- (for the ERBB2+ phenotype), and patient profiles for the ER- (for the ER+ phenotype).

2.1.3 Climent data set

This data set [10] was used as a validation set. Arrays were printed on UCSF Hum. Array 2.0, similar to the Horlings data set, had an average coverage of the genome of 1Mb. Preprocessing of the data can be found in [3]. The data set contained 161 patients diagnosed with a stage I/II lymph node-negative breast cancer and with available ER status. Since the ERBB2 status was not reported in the original publication, we classified 9 patients as ERBB2+, those having a copy number change >1 (in log scale) at the clone DMPC-HFF#1-61H8 which contains the ERBB2 gene. The ER+ set consisted of 101 patients and the ER- set consisted of 60 patients.

2.2 Computational topology methods

2.2.1 Foundations of topological data analysis

Our approach is based on the methods developed in persistent homology which we briefly review. A key concept is the mapping of the data into a point cloud, from which simplicial complexes can be derived; And by doing so, obtain structures that capture the shape and geometry of the data.

Let $P \subset \mathbf{R}^d$ denote our point cloud and $\mathbf{d}(p, p')$ the pairwise distance between points p, p' in P. This data structure, consisting of a point cloud P and pairwise distances is used as an input for what is called the Vietoris-Rips (VR) filtration. The VR-filtration of a point cloud is determined by the filtration parameter (commonly denoted by ϵ) that defines the sequence of simplicial complexes that are used for analyzing the data. Therefore, for any value of the filtration parameter $\epsilon \geq 0$, we define $\mathcal{V}R_\epsilon K_P$ to be the simplicial subcomplex of the complete complex K_P that contains only simplices whose vertices are less than ϵ apart. Formally, let $\sigma \subset P$ be a subcollection of points $(p_1, \ldots, p_m)$. Restricting the indices i and j to $\{1, \ldots, m\}$, σ is a simplex in $\mathcal{V}R_\epsilon K_P$ if $\mathbf{d}(p_i, p_j) < \epsilon$ for all i, j,

If τ is a face of the simplex σ, then the set of all pairwise distances between vertices of τ belongs to the set of pairwise distances of σ's vertices, so $\mathcal{V}R_\epsilon$ is a simplicial complex. Practically, in order to construct a filtration we need to ensure that for $\delta > \epsilon$, we have $\mathcal{V}R_\epsilon K_P \hookrightarrow \mathcal{V}R_\delta K_P$ because if all pairwise distances are less than ϵ, they are also less than δ.

Next, let's define the function $g_{\mathcal{V}R} : K_P \to \mathbf{R}$ as follows: $g_{\mathcal{V}R}(\sigma) = \max_{p,q \text{ in } \sigma} \{\mathbf{d}(p,q)\}$ for any simplex σ in K_P; g is monotone since for $\sigma \prec \tau$, we get $g_{\mathcal{V}R}(\sigma) \leq g_{\mathcal{V}R}(\tau)$ simply because the maximum is taken over a larger set. The *sublevelset* of g at the natural number n is defined by $S_n(g) = \{\sigma \in K \mid g(\sigma) \leq n\}$. The *Vietoris-Rips filtration* around $P \subset \mathbf{R}^d$ is the sublevelset filtration of $g_{\mathcal{V}R}$.

Assuming that pairwise distances between points in P are denoted by $0 \leq \epsilon_1 \leq \cdots \leq \epsilon_N$, we get the filtration

$$\mathcal{V}R_{\epsilon_1} K_P \hookrightarrow \mathcal{V}R_{\epsilon_2} K_P \hookrightarrow \cdots \hookrightarrow \mathcal{V}R_{\epsilon_N} K_P = K_P. \tag{1}$$

Since the complete complex K_P contains as many simplices as there are subsets of P, its cardinality is $2^{\#P}$. The Vietoris-Rips filtration is never constructed all the way up to ϵ_N. A description of efficient algorithms for constructing Vietoris-Rips filtrations may be found in [52]. Most persistent homology software packages (Perseus [37, 35], Gudhi [41], Eirene [19], Ripser [5]) are based on these algorithms.

This construction has many advantages since it only requires knowledge of pairwise distances that are easily computable for many data sets. Once an increasing family of *simplicial complexes* around the data points has been built one can record the change of topological features such as connected components, holes, etc., as the filtration parameter is increased. Formally, to each simplicial complex K one can associate a collection of *homology groups*, $\mathsf{H}_d(K)$, where the range of d is determined by the highest dimension of simplices of K.

Homology of a simplicial complex can be computed in the following way. Notice that the vertices of any simplex can be ordered so that the vertices of any given simplex σ can be represented as a d-tuple $(v_0, \ldots, v_d)$ in ascending order of vertices. A d-dimensional *chain* is a $\mathbf{R}$-linear combination of simplices and they form so-called d-dimensional *chain group* $\mathsf{C}_d(K_P)$ of K_P. Next, the *boundary* ∂_d of σ is a $d-1$-chain formed by a collection of $(d-1)$-dimensional proper faces of σ obtained by removing a single vertex. Since ∂_d defines a linear transformation $\mathsf{C}_d(K_P) \to \mathsf{C}_{d-1}(K_P)$ the subspace determined by its kernel is called the d-dimensional *cycle group* $\mathsf{Z}_d(K_P)$, while $(d-1)$-dimensional *boundary group* $\mathsf{B}_{d-1}(K_P)$ is the image. The d-dimensional *homology group* of a simplicial complex K_P is defined as $\mathsf{H}_d(K) = \frac{\mathsf{Z}_d(K)}{\mathsf{B}_d(K)}$, that is elements of homology are cycles but two cycles that differ by a boundary are considered to be the same.

Persistent homology is to filtrations what homology is to simplicial complexes [38]. Homology is functorial; that is, it assigns algebraic objects to simplicial complexes, and algebraic maps to maps of simplicial complexes. In particular, the inclusions between complexes of a filtration induce maps between these homology groups of each level of the filtration. All together, from the filtration (1) we obtain the *persistence module*:

$$\longrightarrow \mathsf{H}_d(\mathcal{V}R_{\epsilon_1} K_P) \xrightarrow{\phi_d^{1\to 2}} \mathsf{H}_d(\mathcal{V}R_{\epsilon_2} K_P) \xrightarrow{\phi_d^{2\to 3}} \cdots \xrightarrow{\phi_d^{(N-1)\to N}} \mathsf{H}_d(\mathcal{V}R_{\epsilon_N} K_P)$$

The p-persistent d-dimensional homology group of the subcomplex $\mathcal{V}R_{\epsilon_m} K_P$ is the quotient of cycles $\mathsf{Z}_d(\mathcal{V}R_{\epsilon_m} K_P)$ in $\mathcal{V}R_{\epsilon_m} K_P$ by the boundaries $\mathsf{B}_d(\mathcal{V}R_{\epsilon_{m+p}} K_P)$ in $\mathcal{V}R_{\epsilon_{m+p}} K_P$:

$$\mathsf{H}_d^p(\mathcal{V}R_{\epsilon_m} K_P) = \frac{\phi^{m\to(m+p)}(\mathsf{Z}_d(\mathcal{V}R_{\epsilon_m} K_P))}{\phi^{m\to(m+p)}(\mathsf{Z}_d(\mathcal{V}R_{\epsilon_m} K_P)) \cap \mathsf{B}_d(\mathcal{V}R_{\epsilon_{m+p}} K_P)}.$$

The intuition behind this construction is simple: any element x in the d-dimensional homology group of $\mathcal{V}R_{\epsilon_m} K_P$ includes into the d-dimensional group of $\mathcal{V}R_{\epsilon_{m+p}} K_P$ by a sequence of maps on homology induced by inclusions. In general, $\mathcal{V}R_{\epsilon_{m+p}} K_P$ contains more simplices than $\mathcal{V}R_{\epsilon_m} K_P$, so the inclusion of x might be filled out by higher dimensional simplices in which case it becomes a boundary, and dies. In this way, we assign to each element x of $\mathsf{H}_d(\mathcal{V}R_{\epsilon_m} K_P)$ a unique interval $[b_x, d_x)$ where the *birth* $b_x \leq m$ denotes the first time x appeared in homology and the *death*

$d_x > m$ the time it became trivial in homology. Collection of *persistence intervals* for all homology generators is called the d-dimensional *persistence diagram* of the filtration (1). The difference $(d_x - b_x)$ quantifies the *persistence* of x across the filtration. The d-dimensional Betti number of $\mathcal{V}R_{\epsilon_m}K_P$ counts d-dimensional persistence intervals which contain the value m.

The collection of homology groups and their ranks, however, are not completely useful by themselves when analyzing data. Hence persistence-based summaries of data are required. Summaries include bar codes or persistence diagrams [38], Betti curves [3], and persistence landscapes [7].

In this paper we combine the 0-dimensional persistence with the sliding window approach. Zeroth Betti number β_0 counts the number of connected components of a topological space. The Betti curve gives us a way to keep track of the number of connected components through the filtration. We will consider β_0s across the filtration as described in Equation (1), independent of their persistence, to obtain so-called *Betti curve* -a function of the point cloud $\beta_0(\epsilon)$- as well as the filtration parameter. Betti curves will play an important role in construction of TAaCGH, described in Section 2.2.2.

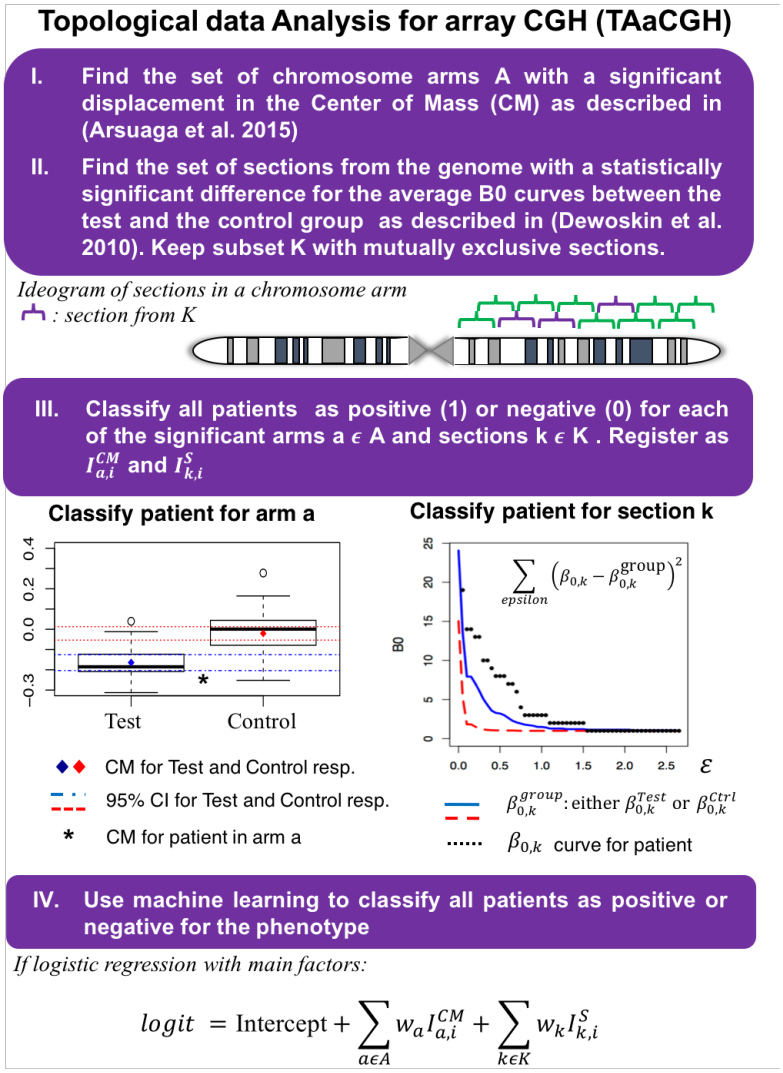

Figure 1: Topological data Analysis for array CGH data (TAaCGH) full methodology to classify a patient for a binary phenotype. TAaCGH finds regions in the genome relevant to discriminating between different phenotypes. Each patient is evaluated for those regions in the genome and the information from all of them is used to derive a classification model using machine learning algorithms.

2.2.2 Topological Analysis of Array CGH (TAaCGH)

TaACGH is designed to identify chromosome aberrations associated with a given phenotype and its key steps (I , II) are illustrated in Figure 1. To achieve this goal the input data needs to include two sets of profiles, one for each phenotype. TAaCGH subdivides chromosomes into overlapping sections that are circularized and analyzed independently of each other. A point cloud is associated with each section of the aCGH profile by means of a sliding window algorithm that maps consecutive copy number measurements along the genome onto a single point. The process is described in Figure 2 [15]. In our previous studies we investigated, through computer simulations, how the size of the window affects our results; we found that a window of size =2 captures the information given by larger window sizes while being computationally more efficient. Furthermore, pairs of consecutive points estimate the norm of the first derivative of the aCGH profile. Next, TaCGH uses the standard filtration algorithm to associate a sequence of Vietoris-Rips (VR) complexes to the point cloud. Traditionally, persistent homology has focused on topological features of the point cloud that persist through the filtration [16]; TAaCGH, on the other hand, uses information of all features that are born during the filtration even if they do not persist. An example is shown in Figure 3 where the data set has two holes at different scales, one of which would traditionally be dismissed.

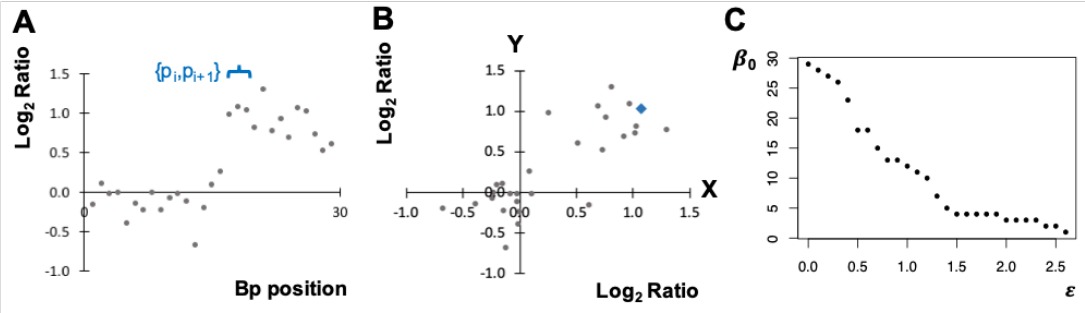

Figure 2: Algorithm to transform an aCGH profile into a zeroth Betti curve. **A)** A simulated aCGH profile with 30 probes consisting of gains in copy number in the second half of the region. Probes are plotted in consecutive order along the genome according to their base pair (bp) position and against their Log_2Ratio from aCGH. A selected pair of consecutive probes has been labeled as $\{p_i, p_{i+1}\}$ in blue with coordinates (18,1.08) and (19,1.04) respectively. **B)** Point cloud associated with the profile from A with a window size equal 2. The set $\{Log_2Ratio_i\}_{i=1}^{n}$ will define a point cloud with n points (here $n = 30$) formed by coordinates $(Log_2Ratio_i, Log_2Ratio_{i+1})$, thus mapping Log_2Ratio information from two consecutive points in A to one point in B. The last and first probes will be considered within one window $(Log_2Ratio_n, Log_2Ratio_1)$. With this point cloud design, two consecutive gains will map to the diagonal in the first quadrant while noise will cluster around the origin. The blue diamond corresponds to the pair of probes $\{p_i, p_{i+1}\}$ in A with coordinates (1.08, 1.04) which correspond to the Log_2Ratio from p_i and p_{i+1} respectively. **C)** Zeroth Betti curve from the point cloud in B applying at each step an incremental value of 0.3 for the filtration parameter ϵ.

In order to retain the information about the birth and death of topological features throughout the filtration TAaCGH uses Betti curves. In the algorithm, the Betti curve for each patient (see C in Figure 2) is calculated using the software jPlex [44], the average of all Betti curves for patients in each group computed, and the average Betti curves compared [15]. Sections of the genome for which statistically significant differences are found are considered aberrant. However, comparing Betti curves does not capture all aberrations. For instance, the only difference between a point cloud associated with a gain or loss of a whole chromosome arm and the control arm is that the first is shifted from the origin. To detect this sort of large scale aberration we included a test that identifies the displacement of the center of masses of the point clouds between the two populations [3].

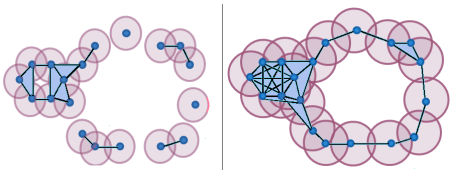

Figure 3: Two stages of the Vietoris-Rips filtration of the (blue) point cloud as the radius of filtration increases are shown in the first two pictures. This illustrates that no single filtration value captures both the smaller and larger loops.

2.2.3 Using machine learning for patient classification

Predicting the phenotype for each patient from the copy number aberration profile is a supervised classification problem, and to address it we followed steps III and IV described in Figure 1. In genomic problems often the number of training examples is small compared to the number of features. Additionally, copy number data contain highly correlated probes. Thus, starting with a subset of aberrant sections that are relevant to the phenotype under study is helpful for building reliable machine learning models. TAaCGH uses as initial co-variates those arms and aberrant sections classified as significant on steps I and II in Figure 1. Next, TAaCGH determines whether or not a given aberration is present in a selected patient. The process is repeated for all patients in a training set and for all the significant sections and chromosome arms with displacement in the center of mass (step III). This creates a set of binary variables as candidate predictors for the classification model. The algorithm to generate the model can be chosen from a variety of machine learning techniques, including logistic regression, random forest, neural networks, and support vector machine. In this paper, we chose to illustrate TAaCGH using logistic regression (step IV). We explain the algorithm in detail next.

During section detection TAaCGH uses an overlapped design (see chromosome ideogram in step II of Figure 1). For any given set of significant overlapping sections, we consider the subset K of non-overlapping ones that covers the exact same regions as the original set. We denote the resulting Betti curves as $\beta_{0,k}^{Test}$, $\beta_{0,k}^{Ctrl}$, with $k \in K$ after averaging the β_0 curves for section k for both patient groups (Test and Control). Next, we classify patients according to the "similarity" between their Betti curve and the Betti curves of the Test and Control groups ($\beta_{0,k}^{Test}$, $\beta_{0,k}^{Ctrl}$) (see blue and red curves in Figure 4) while leaving out the patient i that is being classified. The "similarity" between Betti curves is measured as follows:

$$SS_{k,i}^{G} = \sum_{\epsilon} (\beta_{0,k} - \beta_{0,k}^{G})^2 \tag{2}$$

where $G \in \{Test, Control\}$. Whether a patient i has an aberration or not in section k is encoded by the indicator variable $I_{k,i}^{S}$ which is 1 if assigned to the test group and 0 otherwise. More specifically:

$$I_{k,i}^{S} = \begin{cases} 1, & \text{if } SS_{k,i}^{Test} < SS_{k,i}^{Ctrl} \\ 0, & \text{if } SS_{k,i}^{Test} \geq SS_{k,i}^{Ctrl} \end{cases} \tag{3}$$

An example is shown in Figure 4. Both panels show the averaged Betti zero (β_0) curves for the Test group (blue) and the Control group (red). The panel on the left shows the β_0 curve for a patient that is classified as belonging to the Test group and the panel on the right shows the β_0 curve for a patient that is classified as belonging to the Control group.

The choice of the similarity metric in Equation (2) is directly derived from the test statistic used in TAaCGH to detect aberrant regions [3]. Different metrics were explored in [14]. These include, a metric focusing on relative differences between Betti curves or a weighted metric granting a heavier influence to persistent features (see Table S1 for more details). Computer simulations in [14] revealed that the square of the L_2 norm achieves the best results in terms of detection.

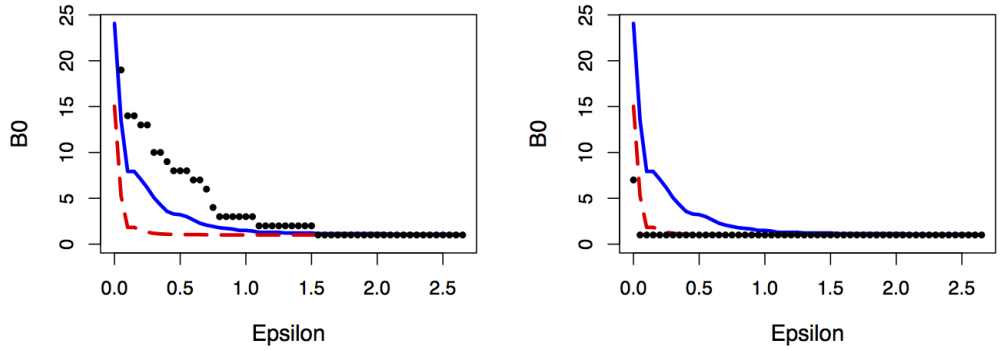

Figure 4: β_0 curve from a patient against averaged β_0 curves for Test and Control groups for the significant section in chromosome 17q for ERBB2 phenotype. Blue solid line for $(\beta_{0,k}^{Test})$ and red dashed line for $(\beta_{0,k}^{Ctrl})$. **Left:** Black pointed line for the $\beta_{0,k}$ curve from patient 8 belonging to ERBB2+ who will be classified for this section of 17q as positive. **Right:** Black pointed line for the $\beta_{0,k}$ curve from patient 37 who does not belong to ERBB2+ and that after this procedure will be classified as negative for this section in arm 17q

One proceeds similarly when using the center of masses (CM) to classify patients (See III in Figure1). If we denote by A the set of all significant arms detected by TAaCGH and $a \in A$. The confidence interval for the CM of the Control group is computed using the mean and standard deviation estimated by TAaCGH. If the value of the center of masses of the patient's point cloud, $\overline{x}_i^a$, falls outside the interval, then the value of the binary variable $I_a^{CM} = 1$ for i; and $= 0$ otherwise. More specifically,

- If the CM for the arm $a \in A$ is a gain

$$I_{a,i}^{CM} = \begin{cases} 1, & \text{if } \overline{x}_i^a > \mu + t_\alpha \sigma/\sqrt{n} \text{ , with } n-1 \text{ d.f.} \\ 0, & \text{Otherwise} \end{cases} \tag{4}$$

- If the CM for the arm $a \in A$ is a deletion

$$I_{a,i}^{CM} = \begin{cases} 1, & \text{if } \overline{x}_i^a < \mu - t_\alpha \sigma/\sqrt{n} \text{ , with } n-1 \text{ d.f.} \\ 0, & \text{Otherwise} \end{cases} \tag{5}$$

where $\overline{x}_i^a = \sum_{probes} x_i^a/n_a$ and n_a is the number of probes in arm a.

We use this information to build a logistic regression model to classify patients for the phenotype given by:

$$logit_i = ln(\frac{\pi_i}{1-\pi_i}) = \text{Intercept} + \sum_{k \in K} w_k I_{k,i}^S + \sum_{a \in A} w_a I_{a,i}^{CM}; i = 1, ..., n \tag{6}$$

were n is the number of patients. Using the predicted value of π_i, we selected a threshold of $\pi_i \geq 0.5$ as our classification criterion for occurrences.

The output of TAaCGH can be directly used to create the *full model*. However a refinement called, *model selection* is necessary to prevent overfitting. In the work presented here we used two common stepwise methods called *forward addition* and *backward deletion*, both available in R [42]. Forward addition starts with the null model and adds covariates until the best model is found. Backward deletion on the other hand starts with the full model and removes covariates until the best model is found. At each step, stepwise methods uses a specific criterion to measure the change in the goodness of fit by adding or removing a covariate. The most common criteria

are the Akaike Information Criterion (AIC) and the Bayesian Information Criteria (BIC) defined as below. Both criteria consider a penalty associated with the number of covariates included in the model discouraging overfitting.

$$AIC := 2k - 2ln(\hat{L})$$

$$BIC := ln(n)k - 2ln(\hat{L})$$

In these expressions, k is the number of parameters in the model, n is the number of cases in the data set and $\hat{L}$ is the maximum of the likelihood function for the model. Smaller values of AIC or BIC indicate a better fit of the model. BIC uses a heavier penalty in the inclusion of parameters than AIC. Though, the difference between the two criteria lies in their objective. AIC looks for the best model for the sample size at hand, while BIC assumes there is a true model, independent of n, that generated the data. In this case, one must be careful when the sample size is too small ($n/k \leq 40$), since BIC selects the true model only if n is very large and can be quite biased when n is not-large enough [8]. As our data sets are small, we focused on AIC but also visited BIC.

Once the model is selected, we measure its goodness-of-fit using sensitivity=TP/(TP + FN) and Specificity=TN/(TN + FP) where TP, FP, TN and FN are the number of True Positive, False Positive, True Negative and False Negative predictions respectively. These and other common terms used in machine learning are available in Table S2 in a form of a *Confusion Matrix*.

To estimate the bias and confidence intervals for the coefficients in the regression model we used the Jackknife method. Jackknife estimates the coefficients while leaving one (or more) patient(s) out of the sample and recomputes the coefficients of the model. In this work, we used Jackknife delete-one estimation. By repeating this process multiple times, one obtains a set of coefficients from which to estimate the standard error and bias of the coefficients proposed in the model. The Jackknife estimation of the coefficients is then the average of the coefficients across all repetitions. In other words, if $\hat{\theta}_{[i]}$ are the coefficients obtained in the logistic regression after omitting the ith observation, then the Jackknife estimator for the coefficients is:

$$\hat{\theta}_{Jack} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{[i]} \tag{7}$$

the standard error is then estimated by [17]:

$$\hat{SE}(\theta)_{Jack} = (\frac{n}{n-1}\sum_{i=1}^{n}(\hat{\theta}_{[i]} - \hat{\theta}_{Jack}))^{1/2} \tag{8}$$

3 Results

3.1 Computer Simulations

3.1.1 Simulations of TAaCGH

An exhaustive simulation study to estimate the statistical properties of TAaCGH was presented in [3]. Here we extended this study by applying TAaCGH to data sets in which the percentage of Test cases presenting an aberrant chromosome was variable (mix). We performed two studies to estimate the effect of mix on the detection of copy number aberrations. First we tested the detection of copy number aberrations by the Betti curves (Step II in Figure 1) and then tested the performance of TAaCGH to classify each profile for a specific section (Step III in Figure 1). As expected, our results show that mix plays a crucial factor in detection (Figure 5). When the sample included at least 60% aberrant profiles in the test group, the sensitivity was 100%. However when mix decreased to 40% and 20% then the sensitivity also decreased to 83.3% and

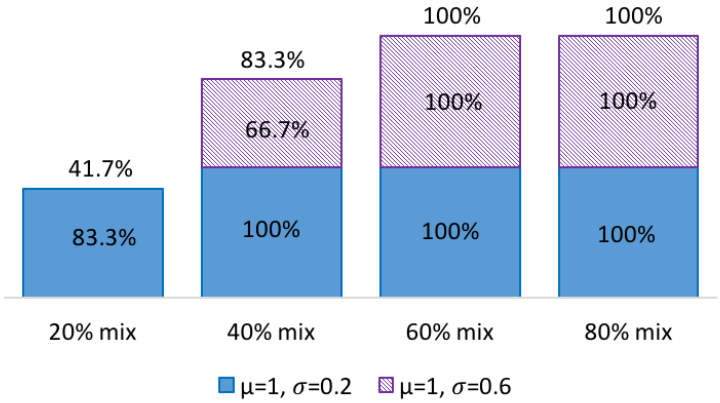

Figure 5: Sensitivity from simulations on detection using β_0 curves for a different *mix* of cases with aberrations in the test group. Solid blue is the percentage of cases when detection was successful with $(\mu = 1, \sigma = 0.2)$, and patterned purple is the percentage of cases from the simulation when detection was successful for $(\mu = 1, \sigma = 0.6)$.

Table 1: Sensitivity (TPR) and specificity (SPC) for patient classification with β_0 curves.

$\mu = 1$	**20%**	*mix*	**40%**	*mix*	**60%**	*mix*	**80%**	*mix*	**100%**	*mix*
$\sigma = 0.2$	50%	70%	56%	80%	65%	88%	78%	94%	98%	97%
$\sigma = 0.6$	49%	55%	54%	64%	60%	71%	73%	75%	76%	79%
Total	**50%**	**63%**	**55%**	**73%**	**63%**	**80%**	**76%**	**78%**	**87%**	**88%**
	TPR	**SPC**	**TPR**	**SPC**	**TPR**	**SPC**	**TPR**	**SPC**	**TPR**	**SPC**

41.7% respectively. When the effects of the ratio between the mean value of the aberration (μ) and the standard deviation (σ) in the data set were explored in [3], sensitivity was found to be close to 100% in a scenario where *mix* was 100% (with a minimum of $\lambda = 5$ aberrant probes). In a mixed environment we expect that the noise will have a big impact in the detection of aberrations. Sensitivity on all experiments $(\mu = 1, \sigma = 0.2)$ was 95.8%. From them, only 1 experiment failed to be detected (with *mix*=20%). The sensitivity value dropped to 66.7% when $(\mu = 1, \sigma = 0.6)$; in this case, all experiments with a *mix*=20% failed to be detected. Yet, all experiments with a *mix* of 60% or 80% were fully detected. The size of the aberrant region λ also played an important role, having a sensitivity of 75 %, 81.3% and 87.5% when the value of λ increased from 5 to 10 and 25 probes (out of 50). As before, experiments with *mix* of 60% or more were fully detected even when only 5 probes were aberrant. Results are shown in Figure 5.

In the second set of simulations we tested the performance of our method at classifying each profile (Step III in Figure 1). In this case, *mix* of aberrant profiles in the Test group had a strong impact in the goodness of fit. For instance, when $\sigma = 0.2$ and all patients were aberrant (*mix* = 100%), the sensitivity (and specificity) were 98% (and 97%) respectively. Even though our model had 100% detection for significant sections when *mix* was 60% or more, sensitivity (and specificity) decreased to 65% (and 88%) for the same parameters. As expected, the difference between the mean (μ) of the aberration and the standard deviation (σ) in the data set also had an impact in the performance. For instance, when all samples in the Test set were aberrant, the sensitivity (and specificity) went from 81% (and 94%) to 70% (and 75%) when σ changed from 0.2 to 0.6. By looking at the difference between sensitivity and specificity, one can tell that the method is better at classifying the negatives than at detecting the positives. This difference is even more dramatic for a smaller standard deviation ($\sigma = 0.2$). Results are summarized in Table 1.

Additionally we explored the performance of a single binary predictor (denoted by I) in two-class classification, where $I = 1$ indicates the presence of the attribute related to the predictor variable and $I = 0$ the absence of it. Whether the case belongs or not to the phenotype of interest is

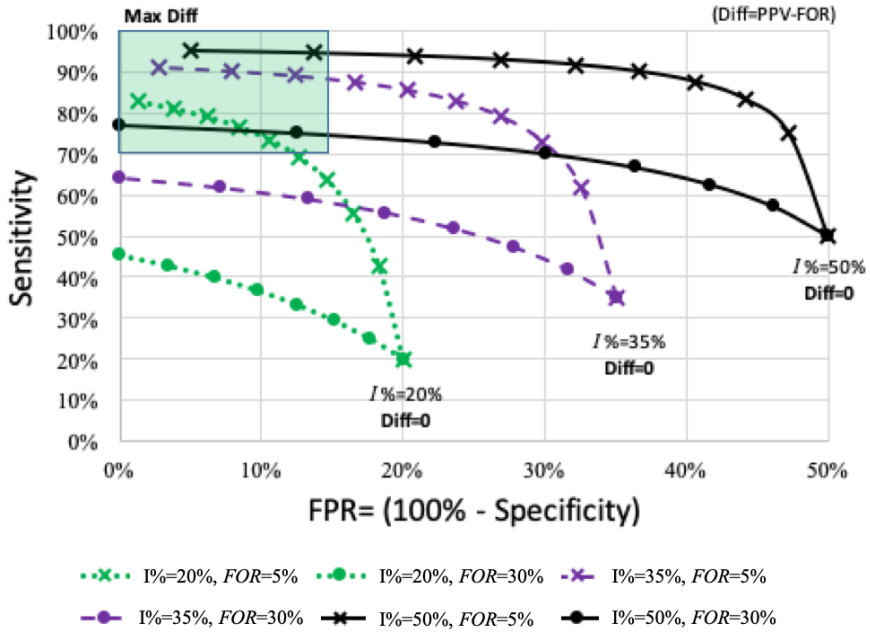

Figure 6: Sensitivity against false positive rate for different binary predictors (I) in an univariate classification model
. Three different scenarios for penetration of the predictor: I%=20% is shown in green dotted lines, 35% in purple dashed lines and 50% in black solid lines. The chart also shows how sensitivity and FPR behaves for 2 different levels of the false omission rate (FOR), 5% and 30%. Each trend was created by decreasing by 10 points Diff=PPV-FOR until Diff=0. A desirable target for combinations of FPR and sensitivity is shown with a green square. Suitable predictors fall in the green square.

denoted by the also binary variable Y. We identified three interlinked factors with a considerable impact in sensitivity and false positive rate $FPR = 1-$specificity:

1) The *penetration of the predictor* (I%), defined here as the percentage of cases for which the predictor variable is equal to 1 (I=1). In Figure 6 we compare sensitivity and FPR for different levels of penetration: I%=20%, I% = 35% and I%=50%.
2) The *False Omission Rate* ($FOR = FN/(FN + TN)$), which is the relative abundance of cases with the phenotype of interest ($Y = 1$) within the group of cases lacking the attribute from the predictor ($I = 0$).
3) The *Difference* in relative abundance of cases with the phenotype of interest ($Y = 1$) between the group of positive ($I = 1$) and negative ($I = 0$) predicted values, defined as Diff=PPV-FOR where $PPV = TP/(TP + TN)$.

Figure 6 shows the trade between sensitivity and false positive rate for different binary predictors when used as the only variable in a classification model. Sensitivity increases as the penetration of the predictor (I%) increases. However, the difference (Diff) in the relative abundance of the response variable with the phenotype of interest between the two groups ($I = 1$ and $I = 0$) needs to be large for the model to be useful. For instance, when the penetration is I%=20%, the difference should be of 40 percentage points to make it to the green square, which is a reasonable target combination of sensitivity and FPR; and when penetration is I%=50% a difference of 80 points is needed.

3.2 A logistic model for ERBB2+ breast cancer

Over-expression of the gene ERBB2 at chromosome 17q, is in many cases a consequence of a copy number gain at the location of the gene. TAaCGH identified four sections in chromosome

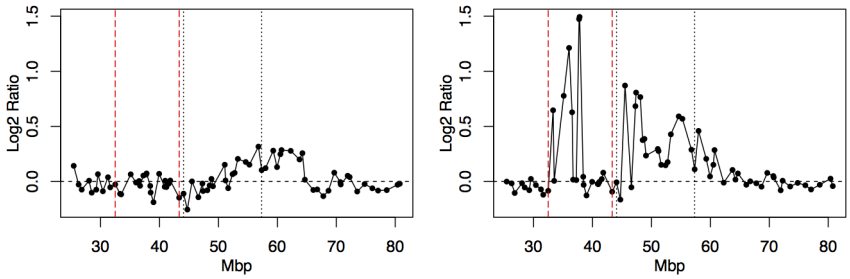

Figure 7: Two profiles from phenotype ERBB2+ with copy numbers for arm 17q with ERBB2 gene located at 38.2Mbp. Section 17$q.s2$ delineated by dashed red lines and section 17$q.s4$ delineated by pointed black lines. **Left:** a clearly non-aberrant profile with ERBB2+ phenotype (patient 153). **Right:** an aberrant profile (patient 308)

17q, ranging from 17q11.1 to 17q22 (25.4 to 57.3 Mbp) [3], that were significant for ERBB2+ patients. As two of the four sections overlapped, thus redundant, we selected only those two that were mutually exclusive and covered the whole region. We denoted them as 17$q.s2$ to refer to section 2 of chromosome 17q ranging from 32.5 to 43.3 Mbp, and 17$q.s4$ to refer to section number 4 ranging from 44.1 to 57.3 Mbp. Each patient was classified as aberrant or non-aberrant for both sections and associated with the indicator variables $I^S_{17q.s2}$ and $I^S_{17q.s4}$ described in step III of Figure 1. Stepwise logistic regression (Section 2.2.3) was used to determine whether both sections contributed to patient classification for the phenotype (Full model available in Table S4). After model selection, only $I^S_{17q.s2}$ was kept. This selection is in agreement with the metrics provided from our simulations and available in Table S3. More importantly, section 17$q.s2$ contains the probe associated with the ERBB2 gene and the region of analysis is reduced to 17q12 to q21.31 (32.49 to 43.3Mbp). Lastly, we used Jackknife delete-one logistic regression to build the following model for the ERBB2+ class phenotype:

$$ln(\frac{\pi_i}{1-\pi_i}) = -2.3 + (3.8)I^S_{17q.s2,i} \tag{9}$$

The bias and 95% confidence intervals for this model are given in Table 2.

Table 2: Horlings Jackknife Coefficients for ERBB2+

	Bias	$\hat{se}$	CI_{lower}	CI_{upper}
Intercept	-1.457611	0.544941	-2.433660	-2.174616
$I^S_{17q.s2}$	1.779060	1.174625	3.533827	4.092198

Predictions with the logistic regression model for ERBB2 produced a sensitivity of 64% (specificity=96%), which was expected considering that only one predictor is being used. The model assigned all individuals being positive in the predictor to one of the classes of the binary response phenotype (Table 3). It is possible that the low detection of positives is related to a high mix of non-aberrant profiles (see Figure 7).

We used the Climent data set [10] as validation set. Using the leave-one-out approach, each patient in the Climent data set was classified for the section (17$q.s2$) used in the model produced with Jackknife with Horlings data set. This resulted in a sensitivity of 78% (specificity=90%). As before, the model is assigning all positive individuals from the predictor as positive for ERBB2. Again, the section in arm 17q discriminates better ERBB2 negatives than positives. In the validation data set the sensitivity increased from 64% to 78%. However, this might be due to the small number of ERBB2 positives in the data set (see Table 3). The complete confusion matrix for both data sets is available in Table S5.

Table 3: Frequencies for ERBB2 against significant sections for TAaCGH with β_0. **Left:** Results for Horlings data set. **Right:** Results for Climent data set.

Horlings	**ERBB2**			**Climent**	**ERBB2**		
	0	**1**			**0**	**1**	
$I^S_{17q.s2}=0$	50	5	FOR=9%	$I^S_{17q.s2}=0$	159	2	FOR=1%
$I^S_{17q.s2}=1$	2	9	PPV=82%	$I^S_{17q.s2}=1$	17	7	PPV=29%
$I^S_{17q.s4}=0$	44	5	FOR=10%	$I^S_{17q.s4}=0$	152	5	FOR=3%
$I^S_{17q.s4}=1$	8	9	PPV=53%	$I^S_{17q.s4}=1$	24	4	PPV=14%

False Omission Rate (FOR=FN/(FN+TN)) and Positive Predictive Value (PPV=TP/(TP+FP) are displayed at the right of each contingency table.

3.3 A logistic regression model for ER+ breast cancer

Estrogen Receptor positive tumors are histochemically characterized by a high level of receptors for the estrogen hormone. The abundance of this receptor is regulated by the gene ESR1, located in chromosome 6q [4]. In the clinical data from the Horlings data set, status for ER was available and we used TAaCGH to find those aberrant regions associated with it. TAaCGH found section $5p14.3-12$ to be significant by Betti curves and arms 2p, 4p, 4q, 5q, 6p, 10q, 14q, 16p and 16q to be significant by the center of masses (see Table 5). We then classified all patients for section $5p14.3-12$ and arms 4p, 5q, 6p, 10q, 16p and 16q, corresponding to regions validated with the Climent data set; and associated them with their indicator variables I^{CM}_{4p}, I^{S}_{5p}, I^{CM}_{5q}, I^{CM}_{6p}, I^{CM}_{10q}, I^{CM}_{16p} and I^{CM}_{16q} (Step III from Figure 1). We finally proceeded to step IV and used the indicator variables as co-variates in modeling. We first created the full model for which results can be found in Table S6. As described in the methods section, the data set used to build the model had a smaller number of ER negatives (28) than positives. Following the widely adopted guidelines of a minimum of 5 to 10 Events Per predictor Variable (EPV) [39, 49], we set up to use no more than 3 predictors (EPV=28/3). Thus, we applied stepwise model selection to reduce the number of predictors resulting in the same model with both AIC and BIC criteria. Results are shown in Table S7. The selected model consists of three covariates: I^{CM}_{4p}, I^{CM}_{6p} and I^{CM}_{16q}. Interestingly, the covariates selected using stepwise regression were in full agreement with the numbers associated with the relevant factors found through our simulations. Table S10 shows that the strongest predictor is I^{CM}_{4p} since it is the one with the highest difference between PPV and FOR and is the second with the smallest penetration of the predictor. We then used Jackknife delete-one to estimate the bias and confidence intervals:

$$ln(\frac{\pi_i}{1-\pi_i}) = 1.9 - (3.3)I^{CM}_{4p,i} - (1.5)I^{CM}_{6p,i} + (2.1)I^{CM}_{16q,i} \tag{10}$$

with the bias and confidence intervals shown in Table 4 for $\alpha = 0.05$.

Table 4: Confidence Intervals for the final model with logistic regression in ER+ for Horlings data set

	Bias	$\hat{se}$	CI_{lower}	CI_{upper}
Intercept	3.689472	1.2639078	1.588530	2.189342
I^{CM}_{4p}	-3.344529	1.3628681	-3.668999	-3.021145
I^{CM}_{6p}	-3.967659	0.9211887	-1.762679	-1.324782
I^{CM}_{16q}	-1.168885	1.1641565	1.788329	2.341724

Predictions with the logistic regression model gave a sensitivity of 79%, (specificity=79%).

Table 5: Results from ER phenotype in Horlings data set and validation with SIRAC, GISTIC[6] and Climent data set after applying TAaCGH for β_0. When known, gains are marked with a plus sign (+) and deletions with (-). GISTIC does not signal amplifications and deletions by phenotype. Instead, for the regions considered aberrant by GISTIC, the percentage of cases for the phenotype is provided if it is higher than 35%.

ER positive

Arm	**TAaCGH**	**SIRAC**	**GISTIC**	**TAaCGH**
5p	5p14.3-12			5p15.33-13.2
16p	16p Arm (+)			16p Arm (+)
16q	16q Arm (-)	16q12.1(-) 16q22.1-23.3 (-)		16q Arm (-)
	Horlings data set	**Horlings data set**	**Horlings data set**	**Climent data set (Validation)**

ER negative

Arm	**TAaCGH**	**SIRAC**	**GISTIC**	**TAaCGH**
2p	2p Arm (+)			
4p	4p Arm (-)	4p15.31 (-)	4p15.2(57%) (-)	4p Arm (-)
4q	4q Arm (-)			
5q	5q Arm (-)	5q33.1 (-)	5q32(50%) (-)	5q Arm (-)
6p	6p Arm (+)	6p33-21.1 (+)		
10p		10p15.1-14 (+)		
10q	10q Arm (-)	10q23.33-24.2 (-)	10q23.32(43%) (-)	
12q		12q13.12-13.2 (-)		
14q	14q Arm(-)			
	Horlings data set	**Horlings data set**	**Horlings data set**	**Climent data set (Validation)**

Next we used our model on the validation data set which, as indicated in the methods section, consists of 101 patients with phenotype ER positive from a total of 161 for which the phenotype was available. It resulted in a sensitivity of 79% (specificity=52%). However, there was a considerable drop in the specificity of the model. It is possible that this is due to the very different frequencies that can be observed in Table S9 and perhaps to a mix with aberrant profiles in the negative Estrogen Receptor group. The complete confusion matrix for both data sets is available in Table S8.

4 Discussion

In this paper we have used topological signatures to build regression models on a binary response variable. In our proposed approach, we first use TAaCGH to identify regions of the genome that are associated with selected phenotypes. For instance, in previous studies we identified copy number changes associated with specific breast cancer molecular subtypes [3]. In this study, we expand this analysis by first estimating the statistical properties of TAaCGH when the number of aberrant profiles in the test set changes (*i.e. mix* percentage). As expected the sensitivity and specificity of TAaCGH decreases, especially when only 20% of the Test sample has the aberrant region and the copy number value is not very different from the standard deviation ($\mu = 1$ and $\sigma = 0.6$). Second, we developed new algorithms to determine whether a patient has an aberration or not. In

our proposed method, we compare the Betti zero curve of the patient with those of the Control and the Test group after the patient was removed from the corresponding category. We found that when ($\mu = 1$ and $\sigma = 0.2$) and all profiles in the test group are aberrant the sensitivity is 98% but it decreases steadily as the proportion of aberrant profiles decreases. For instance, for *mix* with 60% aberrant profiles the sensitivity was found to be 65%. Detection also decreases as the standard deviation gets closer to the value of the copy number. For example, for values $\mu = 1$ and $\sigma = 0.6$ detection is only 76%. On the other hand, the method has strong specificity. To build the logistic model, we followed standard protocols on statistical genetics and associated a binary variable to each chromosome aberration, with a value =1 if the aberration was detected by TAaCGH in the patient profile and =0 otherwise.

At the introduction of this paper we mentioned some challenges for modeling microarray data such as a large number of co-variates in comparison with the number of samples and highly correlated neighboring probes. TAaCGH reduces dimensionality by creating sections from the genome and by transforming the data within those sections into a point cloud. The structure of the point cloud encodes the correlation between neighboring probes, however, by using topological signatures of the point cloud we strongly believe that are reducing the correlation bias. In this paper, we do not focus on the detection of genetic interactions across the genome associated with a particular phenotype. In [2] we illustrated how the first homology group can be used to detect these interactions.

Models were fine tuned using two standard stepwise protocols from model selection: forward addition and backward deletion. In consideration of the size of our data sets, AIC criterion was used during the process. Bias and confidence intervals for coefficients were estimated using Jackknife. The method was tested on two breast cancer examples: ERBB2+ patients and ER+ patients.

ERBB2+ tumors are characterized by over-expression of the gene ERBB2. Over-expression of the gene ERBB2 is commonly associated with a copy number gain in the region containing the gene in the arm $17q$. We showed that TAaCGH detected this region in ERBB2+ patients [3]. This region originally consisted of four significant overlapping sections, but we kept only the two mutually exclusive ones covering the same region (See chromosome ideogram in Figure 1). After logistic regression we were able to reduce the region to only one (17q12-q21.31) where the probe for the ERBB2 gene is located. This single co-variate classified successfully 78% of the ERBB2+ patients (sensitivity) from our validation data set. It is possible that the method doesn't have a better sensitivity because there is a considerable amount of non-aberrant profiles with ERBB2 over-expresion. An example of this is provided in Figure 7.

We expanded our previous results to include cancer positive for Estrogen Receptor (ER+). These tumors grow faster but may be susceptible for treatment [24]. Using TAaCGH, we identified the full set of co-variates and confirmed with the displacement of the center of mass most regions reported in the initial study by SIRAC [30] : 4p, 5q, 6p, 10q and 16q. Our method did not confirm 10p15.1-p14 nor 12q13.12-q13.2. We also detected and validated with our independent data set two additional regions that have been reported as relevant for breast cancer elsewhere: a section from chromosome 5p (p14.3-p12) previously reported with association with ER [46, 28] detected with β_0 homology, and arm 16p[32, 34] detected by the displacement of the CM which is a common CNA associated with breast cancer[34].

TAaCGH also detected 2p, 4q and 14q; three arms not reported by SIRAC or GISTIC, nor confirmed by the Climent data set, but reported in the literature as connected to breast cancer like 4q25-26, 4q33-34 [45], oncogenes BCL11A[27] at 2p16.1, MYCN [36, 31] at 2p24.3 and RAD51L1 from 14q24.1 Complete results are available in Table 5. Interestingly, Arm 6q containing gene (ESR1) was not significant with TAaCGH nor reported by SIRAC; suggesting that copy number changes do not regulate the expression of this gene in breast cancer.

Some CNAs that are common to ER+ and ER- might not be detected by TAaCGH, as the method focuses on what makes them different. Some of these however can be detected with other methodologies, such as GISTIC [6]. For instance 1q23.3 was detected for an amplification by GISTIC. The aberration was present in 68% of the ER+ patients. However, it was also present in 68% of the ER- patients which might explain why it was not detected by TAaCGH. Complete results for

GISTIC in Horlings data set are available in Table S11.

After determining whether patients had an aberration or not we built the logistic model using only 4p, 6p and 16q after stepwise selection. The model correctly classified 76% of ER+ cases in our validation set (sensitivity). However, there was an unexpected drop of 27 points in the specificity between the validation set and the training data set for which the specificity was originally 79%. From our simulations we learned that having *mix* with less than 60% of aberrant profiles and a high standard deviation could be one of the causes for low detection. In a previous study Tolo̧si and Lengauer [48] achieved 69.6% accuracy with the same data we use for validation [10] by using Lasso Logistic Regression with supervised Feature Clustering to control what they call correlation bias. In their final model they use 195 clusters of probes. Our method resulted in a similar accuracy of 68.9% using 3 sections and a simple model.

In conclusion, by using the topological signature associated with a phenotype, TAaCGH provides a innovative approach to reduce the high dimensionality characteristic in genomics and detect genome fractions that are relevant for differentiation. The classification expansion of TAaCGH to determine patients as positive or negative for specific aberrant regions, allows us to use the signal from the fragments as input for modeling and prediction. More importantly, the new classification capabilities of TAaCGH provide a framework to use in combination with other machine learning tools beyond logistic regression like random forests and support vector Machine, among others. Eventually, the choice of the modeling tool depends on the data at hand. Our previous work [3] illustrate a phenotype with four different cancer molecular subtypes. A natural extension that we are currently exploring is combining the topological signatures in copy number for single aberrations, detected as in here with β_0, with co-occurring aberrations detected with β_1[2], together with other clinical and genotype variables as input for prediction.

5 Software

TAaCGH can be obtained by contacting Javier Arsuaga directly: jarsuaga@ucdavis.edu.

References

[1] Khawla Al-Kuraya, Peter Schraml, Joachim Torhorst, Coya Tapia, Boriana Zaharieva, Hedvika Novotny, Hanspeter Spichtin, Robert Maurer, Martina Mirlacher, Ossi Köchli, et al. Prognostic relevance of gene amplifications and coamplifications in breast cancer. *Cancer research*, 64(23):8534–8540, 2004.

[2] Sergio Ardanza-Trevijano, Georgina Gonzalez, Tyler Borrman, Juan Luis Garcia, and Javier Arsuaga. Topological analysis of amplicon structure in comparative genomic hybridization (cgh) data: an application to erbb2/her2/neu amplified tumors. In *International Workshop on Computational Topology in Image Context*, pages 113–129. Springer, 2016.

[3] Javier Arsuaga, Tyler Borrman, Raymond Cavalcante, Georgina Gonzalez, and Catherine Park. Identification of copy number aberrations in breast cancer subtypes using persistence topology. *Microarrays*, 4(3):339–369, 2015.

[4] Katrina R Bauer, Monica Brown, Rosemary D Cress, Carol A Parise, and Vincent Caggiano. Descriptive analysis of estrogen receptor (er)-negative, progesterone receptor (pr)-negative, and her2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the california cancer registry. *Cancer*, 109(9):1721–1728, 2007.

[5] Ulrich Bauer. Ripser: a lean c++ code for the computation of vietoris–rips persistence barcodes. *Software available at https://github. com/Ripser/ripser*, 2017.

[6] Rameen Beroukhim, Craig H Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S Boehm, Jennifer Dobson, Mitsuyoshi Urashima, et al. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899, 2010.

[7] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.

[8] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[9] S.J.; Hammond M.; Perez E.A.; Burstein H.J.; Allred D.C.; Vogel C.L.; Goldstein L.J.; Somlo G.; Gradishar W.J. Carlson, R.W.; Moench et al. Her2 testing in breast cancer: Nccn task force report and recommendations. *Journal of the National Comprehensive Cancer Network*, 4:S1–S22, 2006.

[10] Joan Climent, Peter Dimitrow, Jane Fridlyand, Jose Palacios, Reiner Siebert, Donna G Albertson, Joe W Gray, Daniel Pinkel, Ana Lluch, and Jose A Martinez-Climent. Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. *Cancer research*, 67(2):818–826, 2007.

[11] Marguerite Cuny, Andrew Kramar, Frank Courjal, Vala Johannsdottir, Barry Iacopetta, Hélène Fontaine, Jean Grenier, Stéphane Culine, and Charles Theillet. Relating genotype and phenotype in breast cancer: an analysis of the prognostic significance of amplification at eight different genes or loci and of p53 mutations. *Cancer research*, 60(4):1077–1083, 2000.

[12] Jorma J de Ronde, Christiaan Klijn, Arno Velds, Henne Holstege, Marcel JT Reinders, Jos Jonkers, and Lodewyk FA Wessels. Kc-smartr: An r package for detection of statistically significant aberrations in multi-experiment acgh data. *BMC research notes*, 3(1):298, 2010.

[13] SL Deming, SJ Nass, RB Dickson, and BJ Trock. C-myc amplification in breast cancer: a meta-analysis of its occurrence and prognostic relevance. *British journal of cancer*, 83(12):1688, 2000.

[14] Daniel DeWoskin. Applications of computational homology to analysis of primary breast tumor cgh profiles. Master's thesis, San Francisco State University, San Francisco, California, USA, 2009.

[15] Daniel DeWoskin, Joan Climent, I Cruz-White, Mariel Vazquez, Catherine Park, and Javier Arsuaga. Applications of computational homology to the analysis of treatment response in breast cancer patients. *Topology and its Applications*, 157(1):157–164, 2010.

[16] Herbert Edelsbrunner and John Harer. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.

[17] Bradley Efron and Gail Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

[18] Jane Fridlyand, Antoine M Snijders, Dan Pinkel, Donna G Albertson, and Ajay N Jain. Hidden markov models approach to the analysis of array cgh data. *Journal of multivariate analysis*, 90(1):132–153, 2004.

[19] Gregory Henselman and Robert Ghrist. Matroid filtrations and computational persistent homology. *arXiv preprint arXiv:1606.00199*, 2016.

[20] Zena M Hira and Duncan F Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015, 2015.

[21] Hugo M Horlings, Carmen Lai, Dimitry SA Nuyten, Hans Halfwerk, Petra Kristel, Erik van Beers, Simon A Joosse, Christiaan Klijn, Petra M Nederlof, Marcel JT Reinders, et al. Integration of dna copy number alterations and prognostic gene expression signatures in breast

cancer patients. *Clinical Cancer Research*, 16(2):651–663, 2010.

[22] Hugo M Horlings, Carmen Lai, Dimitry SA Nuyten, Hans Halfwerk, Petra Kristel, Erik van Beers, Simon A Joosse, Christiaan Klijn, Petra M Nederlof, Marcel JT Reinders, et al. Integration of dna copy number alterations and prognostic gene expression signatures in breast cancer patients. supplementary material. *Clinical Cancer Research*, 16(2):651–663, 2010. http://clincancerres.aacrjournals.org/content/16/2/ 651/suppl/DC1/.

[23] Philippe Hupé, Nicolas Stransky, Jean-Paul Thiery, François Radvanyi, and Emmanuel Barillot. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–3422, 2004.

[24] National Cancer Institute, September 2018. https://www.cancer.gov/types/breast/patient/breast-treatment-pdq.

[25] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. Big data analytics in bioinformatics: A machine learning perspective. *arXiv preprint arXiv:1506.05101*, 2015.

[26] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.

[27] Walid T Khaled, Song Choon Lee, John Stingl, Xiongfeng Chen, H Raza Ali, Oscar M Rueda, Fazal Hadi, Juexuan Wang, Yong Yu, Suet-Feung Chin, et al. Bcl11a is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nature communications*, 6:ncomms6987, 2015.

[28] Hyung-cheol Kim, Ji-Young Lee, Hyuna Sung, Ji-Yeob Choi, Sue K Park, Kyoung-Mu Lee, Young Jin Kim, Min Jin Go, Lian Li, Yoon Shin Cho, et al. A genome-wide association study identifies a breast cancer risk variant in erbb4 at 2q34: results from the seoul breast cancer study. *Breast Cancer Research*, 14(2):R56, 2012.

[29] Christiaan Klijn, Henne Holstege, Jeroen de Ridder, Xiaoling Liu, Marcel Reinders, Jos Jonkers, and Lodewyk Wessels. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array cgh data. *Nucleic acids research*, 36(2):e13–e13, 2008.

[30] Carmen Lai, Hugo M Horlings, Marc J van de Vijver, Eric H van Beers, Petra M Nederlof, Lodewyk FA Wessels, and Marcel JT Reinders. Sirac: Supervised identification of regions of aberration in acgh datasets. *BMC bioinformatics*, 8(1):422, 2007.

[31] Florence Lerebours, Ivan Bieche, and Rosette Lidereau. Update on inflammatory breast cancer. *Breast Cancer Research*, 7(2):52, 2005.

[32] Jirong Long, Qiuyin Cai, Xiao-Ou Shu, Shimian Qu, Chun Li, Ying Zheng, Kai Gu, Wenjing Wang, Yong-Bing Xiang, Jiarong Cheng, et al. Identification of a functional genetic variant at 16q12. 1 for breast cancer risk: results from the asia breast cancer consortium. *PLoS genetics*, 6(6):e1001002, 2010.

[33] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[34] John Mendelsohn, Peter M Howley, Mark A Israel, Joe W Gray, and Craig B Thompson. *The Molecular Basis of Cancer E-Book*. Elsevier Health Sciences, 2014.

[35] Konstantin Mischaikow and Vidit Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete & Computational Geometry*, 50(2):330–353, 2013.

[36] Y Mizukami, A Nonomura, T Takizawa, M Noguchi, T Michigishi, S Nakamura, and T Ishizaki. N-myc protein expression in human breast carcinoma: prognostic implications.

Anticancer research, 15(6B):2899–2905, 1995.

[37] Vidit Nanda. Perseus, the persistent homology software. http://www.sas.upenn.edu/ vnanda/perseus, Accessed: 04/08/2019.

[38] Vidit Nanda and Radmila Sazdanović. Simplicial models and topological inference in biological systems. In *Discrete and topological models in molecular biology*, pages 109–141. Springer, 2014.

[39] Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R Holford, and Alvan R Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12):1373–1379, 1996.

[40] Daniel Pinkel and Donna G Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature genetics*, 37(6s):S11, 2005.

[41] The GUDHI Project. Gudhi: User and reference manual. 2015.

[42] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[43] Jorge S Reis-Filho, Kay Savage, Maryou BK Lambros, Michelle James, Dawn Steele, Robin L Jones, and Mitch Dowsett. Cyclin d1 protein overexpression and ccnd1 amplification in breast carcinomas: an immunohistochemical and chromogenic in situ hybridisation analysis. *Modern pathology*, 19(7):999, 2006.

[44] Harlan Sexton and Mikael Vejdemo-Johansson. jPlex, December 2008. http://comptop.stanford.edu/programs/jplex/.

[45] Narayan Shivapurkar, Sanjay Sood, Ignacio I Wistuba, Arvind K Virmani, Anirban Maitra, Sara Milchgrub, John D Minna, and Adi F Gazdar. Multiple regions of chromosome 4 demonstrating allelic losses in breast carcinomas. *Cancer research*, 59(15):3576–3580, 1999.

[46] Simon N Stacey, Andrei Manolescu, Patrick Sulem, Steinunn Thorlacius, Sigurjon A Gudjonsson, Gudbjörn F Jonsson, Margret Jakobsdottir, Jon T Bergthorsson, Julius Gudmundsson, Katja K Aben, et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor–positive breast cancer. *Nature genetics*, 40(6):703, 2008.

[47] Zhifu Sun, Yan W Asmann, Krishna R Kalari, Brian Bot, Jeanette E Eckel-Passow, Tiffany R Baker, Jennifer M Carr, Irina Khrebtukova, Shujun Luo, Lu Zhang, et al. Deep sequence analysis of the relationship between gene expression, cpg island methylation, and gene copy number in breast cancer cells, 2011.

[48] Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.

[49] Eric Vittinghoff and Charles E McCulloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American journal of epidemiology*, 165(6):710–718, 2007.

[50] John W Whitaker, Zhao Chen, and Wei Wang. Predicting the human epigenome from dna motifs. *Nature methods*, 12(3):265, 2015.

[51] Antonio C Wolff, M Elizabeth H Hammond, David G Hicks, Mitch Dowsett, Lisa M McShane, Kimberly H Allison, Donald C Allred, John MS Bartlett, Michael Bilous, Patrick Fitzgibbons, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline update. *Archives of Pathology and Laboratory Medicine*, 138(2):241–256, 2013.

[52] Afra Zomorodian. Fast construction of the Vietoris-Rips complex. *Computers & Graphics*, 34(3):263–271, 2010.

6 Supplementary Materials

Table S1: Four different metrics comparing the average Betti curves from the patients in the control group ($\beta_{0,\epsilon}^{Ctrl}$) against the average for the test group ($\beta_{0,\epsilon}^{Test}$) using filtration parameter ϵ. After simulations for these metrics, DeWoskin [14] finds that the best results are provided by SS_1

.

$$SS_1 = \sum_{\epsilon} (\beta_{0,\epsilon_i}^{Test} - \beta_{0,\epsilon i}^{Ctrl})^2 = \left\| \beta_0^{Test}, \beta_0^{Ctrl} \right\|_{L_2}^2$$

$$\begin{aligned} SS_2 = & \frac{\Delta\epsilon}{2} \left| \beta_{0,\epsilon_1}^{Test} + 2\beta_{0,\epsilon_2}^{Test} + ... + 2\beta_{0,\epsilon_{n-1}}^{Test} + \beta_{0,\epsilon_n}^{Test} \right| \\ & - \left| \beta_{0,\epsilon_1}^{Ctrl} + 2\beta_{0,\epsilon_2}^{Ctrl} + ... + 2\beta_{0,\epsilon_{n-1}}^{Ctrl} + \beta_{0,\epsilon_n}^{Ctrl} \right| \end{aligned}$$

$$SS_3 = \frac{\Delta\epsilon}{2} \left(\frac{\beta_{0,\epsilon_1}^{Test}}{\beta_{0,\epsilon_1}^{Ctrl}} + 2\frac{\beta_{0,\epsilon_2}^{Test}}{\beta_{0,\epsilon_2}^{Ctrl}} + ... + 2\frac{\beta_{0,\epsilon_{n-1}}^{Test}}{\beta_{0,\epsilon_{n-1}}^{Ctrl}} + \frac{\beta_{0,\epsilon_n}^{Test}}{\beta_{0,\epsilon_n}^{Ctrl}} \right)$$

$$\begin{aligned} SS_4 = & \frac{\Delta\epsilon}{2} \left| (\epsilon_1)\beta_{0,\epsilon_1}^{Test} + 2(\epsilon_2)\beta_{0,\epsilon_2}^{Test} + ... + 2(\epsilon_{n-1})\beta_{0,\epsilon_{n-1}}^{Test} + (\epsilon_n)\beta_{0,\epsilon_n}^{Test} \right| \\ & - \left| (\epsilon_1)\beta_{0,\epsilon_1}^{Ctrl} + 2(\epsilon_2)\beta_{0,\epsilon_2}^{Ctrl} + ... + 2(\epsilon_{n-1})\beta_{0,\epsilon_{n-1}}^{Ctrl} + (\epsilon_n)\beta_{0,\epsilon_n}^{Ctrl} \right| \end{aligned}$$

SS_1 corresponds to the square of the L_2 norm, SS_2 measures the difference between the areas of both curves, SS_3 uses relative differences before finding the area under the curve and SS_4 is similar to SS_2 but assigns heavier weights as the filtration parameter increases.

Table S2: Confusion Matrix

Predicted	**True Condition** Negative=0	**True Condition** Positive=1	
Negative=0	TN True Negative	FN False Negative	**FOR** $= \frac{\text{FN}}{\text{Predicted Negative}} = \frac{\text{FN}}{\text{TN+FN}}$
Positive=1	FP False Positive	TP True Positive	**PPV** $= \frac{\text{TP}}{\text{Predicted Positive}} = \frac{\text{TP}}{\text{FP+TP}}$
	SPC $= \frac{\text{TN}}{\text{Condition Negative}}$	**TPR** $= \frac{\text{TP}}{\text{Condition Positive}}$	**Diff** =PPV-FOR

True negatives (TN), false negatives (FN), false positives (FP) true positives (TP), specificity (SPC), sensitivity (TPR), false omission rate (FOR) and positive predictive value (PPV).

Table S3: Three factors with heavy impact in the relevance of the variables to become good predictors. The indicator variables listed corresponds to the validated (with SIRAC and Climent data set) significant sections and arms after applying TAaCGH to Horlings data set **(left)** for ERBB2+ phenotype. The same metrics are provided for the variables from Climent data set **(right)**.

Horlings	I%	**FOR**	**Diff**	**Climent**	I%	**FOR**	**Diff**
$I^S_{17q.s2}$	17%	9%	73	$I^S_{17q.s2}$	13%	1	28%
$I^S_{17q.s4}$	26%	10%	43	$I^S_{17q.s4}$	15%	3	14%

I%: The penetration of the predictor; that is, the percentage of cases equal to 1. False Omission Rate (FOR=FN/(FN+TN)) tells the abundance of positive cases from the response variable (phenotype) within the set of cases where the characteristic from I is absent ($I = 0$). The Difference (Diff) between the Positive Predictive Value (PPV=TP/(TP+TN))) and FOR: Diff=PPV-FOR, represents the difference in the abundance of the positive response variable (Y=1) between the two groups formed by the values of the predictor ($\%Y = 1$ when $I = 0$ vs $\%Y = 1$ when $I = 1$).

Table S4: Full logistic regression model for ERBB2 phenotype: sensitivity=64%, specificity=96%.

Coefficients	**Estimate**	**Std. Error**	**Z-value**	**Pr(>|z|)**
Intercept	-2.4830	0.5279	-4.704	2.56e-06
$I^S_{17q.S2}$	3.2795	1.0224	3.208	0.00134
$I^S_{17q.S4}$	0.9136	0.9269	0.986	0.32428

Table S5: Accuracy and confusion matrix for the final logistic regression model for ERBB2+ phenotype created with Horlings as training data set and tested with Climent data set.

	Horlings		**Climent**	
Predicted	**ERBB2-**	**ERBB2+**	**ERBB2-**	**ERBB2+**
Negative	TN=50	FN=5	TN=159	FN=2
Positive	FP=2	TP=9	FP=17	TP=7
	SPC=96.2%	**TPR=64.3%**	**SPC=90.3%**	**TPR=77.8%**
	ACC=89.4%		**ACC=89.7%**	

True negatives (TN), false negatives (FN), false positives (FP), true positives (TP), specificity (SPC), sensitivity (TPR) and accuracy (ACC=(TP+TN)/total).

Table S6: Full logistic regression model for ER+ phenotype: sensitivity=84%, specificity=75%

Coefficients	**Estimate**	**Std. Error**	**Z-value**	**Pr(>|z|)**
Intercept	3.7001	2.2017	1.681	0.09284
I_{4p}^{CM}	-3.3548	1.0639	-3.153	0.00162
I_{5p}^{S}	0.1725	0.7881	0.219	0.82670
I_{5q}^{CM}	-0.6384	0.9797	-0.652	0.51466
I_{6p}^{CM}	-1.5036	0.7977	-1.885	0.05944
I_{10q}^{CM}	-1.3610	1.0661	-1.277	0.20174
I_{16p}^{CM}	-0.3021	0.9322	-0.324	0.74589
I_{16q}^{CM}	1.8620	0.9510	1.958	0.05025

Table S7: Logistic regression model for ER+ after forward and backward stepwise selection.

Coefficients	**Estimate**	**Std. Error**	**Z-value**	**Pr(>|z|)**
Intercept	1.8842	1.0225	1.843	0.065359
I_{4p}^{CM}	-3.3343	0.9377	-3.556	0.000377
I_{6p}^{CM}	-1.5420	0.7234	-2.132	0.033038
I_{16q}^{CM}	2.0582	0.9258	2.223	0.026214

Table S8: Accuracy and confusion matrix for the final logistic regression model for ER+ phenotype created with Horlings as training data set and tested with Climent data set.

	Horlings		**Climent**	
Predicted	**ER-**	**ER+**	**ER-**	**ER+**
Negative	TN=22	FN=8	TN=31	FN=21
Positive	FP=6	TP=30	FP=29	TP=80
	SPC=78.6%	**TPR=78.9%**	**SPC=51.7%**	**TPR=79.2%**
	ACC=78.8%		**ACC=68.9%**	

True negatives (TN), false negatives (FN), false positives (FP), true positives (TP), specificity (SPC), sensitivity (TPR) and accuracy (ACC=(TP+TN)/total).

Table S9: Frequencies for ER against significant CM and sections for TAaCGH with β_0. **Left:** Horlings data set. **Right:** Climent data set.

	ER				**ER**		
Horlings	**0**	**1**		**Climent**	**0**	**1**	
$I_{2p}^{CM}=0$	7	22	FOR= 76%	$I_{2p}^{CM}=0$	34	68	FOR=67%
$I_{2p}^{CM}=1$	21	16	PPV=43%	$I_{2p}^{CM}=1$	26	33	PPV=56%
$I_{4p}^{CM}=0$	2	26	FOR= 93%	$I_{4p}^{CM}=0$	16	57	FOR=78%
$I_{4p}^{CM}=1$	26	12	PPV=32%	$I_{4p}^{CM}=1$	44	44	PPV=50%
$I_{4q}^{CM}=0$	3	14	FOR= 82%	$I_{4q}^{CM}=0$	18	40	FOR=69%
$I_{4q}^{CM}=1$	25	24	PPV=49%	$I_{4q}^{CM}=1$	42	61	PPV=59%
$I_{5p}^{S}=0$	18	13	FOR= 42%	$I_{5p}^{S}=0$	32	33	FOR=51%
$I_{5p}^{s}=1$	10	25	PPV=71%	$I_{5p}^{s}=1$	28	68	PPV=71%
$I_{5q}^{CM}=0$	4	15	FOR= 79%	$I_{5q}^{CM}=0$	16	47	FOR=75%
$I_{5q}^{CM}=1$	24	23	PPV=49%	$I_{5q}^{CM}=1$	44	54	PPV=55%
$I_{6p}^{CM}=0$	7	25	FOR= 78%	$I_{6p}^{CM}=0$	35	75	FOR=68%
$I_{6p}^{CM}=1$	21	13	PPV=38%	$I_{6p}^{CM}=1$	25	26	PPV=51%
$I_{10q}^{CM}=0$	3	8	FOR= 73%	$I_{10q}^{CM}=0$	24	46	FOR=66%
$I_{10q}^{CM}=1$	25	30	PPV=55%	$I_{10q}^{CM}=1$	36	55	PPV=60%
$I_{14q}^{CM}=0$	4	12	FOR= 75%	$I_{14q}^{CM}=0$	16	43	FOR=73%
$I_{14q}^{CM}=1$	24	26	PPV=52%	$I_{14q}^{CM}=1$	44	58	PPV=57%
$I_{16p}^{CM}=0$	20	17	FOR= 46%	$I_{16p}^{CM}=0$	5	32	FOR=87%
$I_{16p}^{CM}=1$	8	21	PPV=72%	$I_{16p}^{CM}=1$	55	69	PPV=56%
$I_{16q}^{CM}=0$	11	4	FOR= 27%	$I_{16q}^{CM}=0$	22	21	FOR=49%
$I_{16q}^{CM}=1$	17	34	PPV=67%	$I_{16q}^{CM}=1$	38	80	PPV=68%

False Omission Rate (FOR=FN/(FN+TN)) and Positive Predictive Value (PPV=TP/(TP+FP) are displayed at the right of each contingency table.

Table S10: Three factors with heavy impact in the relevance of the variables to become good predictors. The indicator variables listed corresponds to the validated (with SIRAC and Climent data set) significant sections and arms after applying TAaCGH to Horlings data set **(left)** for the positive Estrogen Receptor (ER+) phenotype. The same metrics are provided for the variables from Climent data set **(right)**.

Horlings	$I\%$	**FOR**	**Diff**	**Climent**	$I\%$	**FOR**	**Diff**
I_{2p}^{CM}	56%	76%	-33	I_{2p}^{CM}	37%	67%	-11%
I_{4p}^{CM}	58%	93%	-61	I_{4p}^{CM}	55%	78%	-28
I_{4q}^{CM}	74%	82%	-33	I_{4q}^{CM}	64%	69%	-10
I_{5p}^{S}	53%	42%	29	I_{5p}^{S}	60%	51%	20
I_{5q}^{CM}	71%	79%	-30	I_{5q}^{CM}	61%	75%	-20
I_{6p}^{CM}	52%	78%	-40	I_{6p}^{CM}	32%	68%	-17
I_{10q}^{CM}	83%	73%	-18	I_{10q}^{CM}	57%	66%	-5
I_{14q}^{CM}	76%	75%	-23	I_{14q}^{CM}	63%	73%	-16
I_{16p}^{CM}	44%	46%	26	I_{16p}^{CM}	77%	87%	-31
I_{16q}^{CM}	77%	27%	40	I_{16q}^{CM}	73%	49%	19

$I\%$: The penetration of the predictor; that is, the percentage of cases equal to 1. False Omission Rate (FOR=FN/(FN+TN)) tells the abundance of positive cases from the response variable (phenotype) within the set of cases where the characteristic from I is absent ($I = 0$). The Difference (Diff) between the Positive Predictive Value (PPV=TP/(TP+TN))) and FOR: Diff=PPV-FOR, represents the difference in the abundance of the positive response variable (Y=1) between the two groups formed by the values of the predictor ($\%Y = 1$ when $I = 0$ vs $\%Y = 1$ when $I = 1$).

Table S11: Regions detected as amplified or deleted by GISTIC [6]; a standard methodology used to detect anomalies in copy number. GISTIC does not compare between two phenotypes, therefore it doesn't provide what differentiate them. However, it is informative to look at those not detected with our method because they could be common ground for different cancer phenotypes. In an effort to associate the aberrations from GISTIC to a phenotype, we computed the proportion of aberrant profiles within ER+ and within ER- for each of the regions detected by GISTIC. Here we report only aberrant regions when present in at least 35% of the cases.

Arm	**Aberration Type**	**ER negative**	**ER positive**
1q	Amplification	1q23.3 (68%)	1q23.3 (68%)
		1q41 (54%)	1q41 (66%)
3p	Deletion	3p14.3 (57%)	
3q	Deletion	3q27.2 (46%)	
4p	Deletion	4p15.2 (57%)	
5q	Deletion	5q32 (50%)	
7q	Amplification	7q34 (46%)	
8p	Deletion	8p23.2 (57%)	8p23.2 (47%)
8q	Amplification	8q24.11 (64%)	8q24.11 (68%)
10q	Deletion	10q23.32 (43%)	
12p	Amplification	12p13.33 (36%)	
13q	Deletion	13q14.11 (64%)	13q14.11 (61%)
17q	Amplification		17q23.1 (45%)
17q	Amplification	17q24.3 (36%)	17q24.3 (39%)
18q	Deletion	18q12.2 (36%)	