Published in final edited form as:

Stat Med. 2012 May 20; 31(0): 1203–1220. doi:10.1002/sim.4389.

Local sparse bump hunting reveals molecular heterogeneity of colon tumors‡

Jean-Eudes Dazard^{a,*†}, J. Sunil Rao^b, and Sanford Markowitz^c

^aDivision of Bioinformatics, Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, OH 44106, U.S.A

^bDivision of Biostatistics, Department of Epidemiology and Public Health, The University of Miami, Miami, FL 33136, U.S.A

^cDepartment of Hematology–Oncology, Case Western Reserve University, Cleveland, OH 44106, U.S.A

Abstract

The question of molecular heterogeneity and of tumoral phenotype in cancer remains unresolved. To understand the underlying molecular basis of this phenomenon, we analyzed genome-wide expression data of colon cancer metastasis samples, as these tumors are the most advanced and hence would be anticipated to be the most likely heterogeneous group of tumors, potentially exhibiting the maximum amount of genetic heterogeneity. Casting a statistical net around such a complex problem proves difficult because of the high dimensionality and multi-collinearity of the gene expression space, combined with the fact that genes act in concert with one another and that not all genes surveyed might be involved. We devise a strategy to identify distinct subgroups of samples and determine the genetic/molecular signature that defines them. This involves use of the local sparse bump hunting algorithm, which provides a much more optimal and biologically faithful transformed space within which to search for bumps. In addition, thanks to the variable selection feature of the algorithm, we derived a novel sparse gene expression signature, which appears to divide all colon cancer patients into two populations: a population whose expression pattern can be molecularly encompassed within the bump and an outlier population that cannot be. Although all patients within any given stage of the disease, including the metastatic group, appear clinically homogeneous, our procedure revealed two subgroups in each stage with distinct genetic/molecular profiles. We also discuss implications of such a finding in terms of early detection, diagnosis and prognosis.

Keywords

local sparse bump hunting; mixture density; class discovery; colon cancer patient subtyping; early diagnosis and prognosis

‡Supporting information may be found in the online version of this article.

Copyright © 2011 John Wiley & Sons, Ltd.

*Correspondence to: Jean-Eudes Dazard, Division of Bioinformatics, Center for Proteomics and Bioinformatics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, U.S.A.

†jxd101@case.edu

1. Introduction

The question of how to identify intrinsic structure in high dimensional data is still largely unanswered. To illustrate the problem, consider the following example. A major unresolved question in cancer research is the molecular heterogeneity of the disease and of the tumoral phenotype. The current taxonomy of cancers, mostly based on histopathology, accounts for more than 200 distinct entities arising from diverse cell types. However, because of the large patient tumoral phenotype heterogeneity, more subtypes are suspected to exist. In this regard, molecular profiling by gene expression microarray has proved a useful tool in identifying previously unrecognized subtypes of cancers and/or tumors. For example, molecular profiling has divided human breast cancer into several distinct subclasses [1], human large cell lymphomas into two distinct subclasses [2], and recently, medulloblastoma into four distinct molecular variants [3].

Although classification and class prediction techniques applied to gene expression profiling from microarray experiments have received a lot of attention over the past 10 years, there has been no systematic approach whose objective is to say something about the intrinsic structure of the data, such as in class discovery. For instance, there exist several classifier techniques that have proven powerful predictive tools of cancer patient classification (and patient class prediction) from *previously recognized* tumor types in a variety of cancers [4–8]. The problem, however, is that most of these techniques are only applicable when the tumor subtypes are assumed or known in advance. To our knowledge, only a few exploratory studies have been carried out in cancer research for characterizing *unrecognized subtypes* of tumors, such as in human leukemia [4], breast cancer [1], large cell lymphomas [2], and recently, medulloblastoma [3].

Identifying and describing such molecularly-defined subclasses has important implications as it relates to underlying molecular events in the etiology of the disease, and can relate to subgroups of patients having either different diagnosis or prognosis with respect to disease, and to subgroups having different responses to therapies of the disease. Moreover, the ability to derive prognostic or predictive signatures of disease based on gene expression profiling can obviously be confounded in an apparently single disease actually composed of multiple differing subtypes. Hitherto, human colon cancers have been divided into two molecularly distinct subtypes: microsatellite unstable or MSI colon cancers, which arise due to defects in DNA mismatch repair genes and that account for 15% of colon cancer cases, and microsatellite stable or MSS colon cancers, which account for 85% of colon cancer cases. The application of this study is to address the question of how heterogeneous MSS colon tumors are and provide some molecular basis of the metastatic process in human colon cancer with the help of a new statistical approach (see succeeding text).

Given the high degree of yearly metastasis-related mortality and morbidity in colon cancer, it is of great scientific interest to understand the underlying molecular basis of this phenomenon. Once such level of understanding is at a genomic level where potentially novel gene expression profiles might be identified. We devised a strategy to identify distinct subgroups of samples and determined the gene expression signature that defines them. This involved the use of an optimized bump hunting strategy in a local, rotated, and sparse version of the gene expression space, which we recently developed [9]. Because of the high dimensionality and collinearity of this space, combined with the fact that genes act in concert with one another and that not all genes surveyed might be involved, we made use of a sparse principal components (SPCs) approach, which provides a much more optimal and biologically faithful transformed space within which to search for bumps. A novel bump gene expression signature was derived, which appears to divide terminal-stage (metastatic)

colon cancer patients into two populations: a population whose expression pattern can be molecularly encompassed within the bump and an outlier population that cannot be.

Although the group of metastatic tumors appears clinically homogeneous, our procedure revealed two metastatic subgroups with distinct genetic/molecular profiles, the first of which being close to the most advanced stage of primary tumors, whereas the second represents a more terminal and aggressive metastatic stage. Interestingly, this signature could be tracked all the way back to the earliest stages (primary tumors) of the disease. Because of large metastatic patient survival heterogeneity, these subtypes have been suspected to exist for some time and are potentially of great interest to colon cancer clinicians and biologists. We looked for a stage-by-bump-signature interaction as it relates to cancer-specific survival and showed that the identification of an early gene expression signature in primary tumors could potentially have diagnostic as well as prognostic values.

2. Colon cancer staging and our premise

2.1. Colon cancer staging

We used a large microarray gene expression data set, initially generated from a group of sporadic MSS colon cancer tumor samples. We grouped the tumor samples, and patients from which they were extracted, according to Astler –Coller –Duke’s clinical staging system [10]. These groups included stage ‘B’ *primary* tumors that are tumors that had not penetrated beyond the wall of the bowel and that are 80% curable with surgery. In addition, we selected these tumors from patients who on 3 years of clinical follow-up had not relapsed, and *presumptively*, these tumors were lacking in metastatic ability. These groups also included Duke’s stage ‘C’ tumors that had spread to *regional* lymph nodes and that represent a progressive worsening prognosis. These groups also included Duke’s stage ‘D’ most advanced primary tumors (left over in the colon) that at diagnosis had already spread to distant organ sites (primarily the liver). In addition, we studied a set of ‘METS’ colon cancer samples that were *metastatic* tumors harvested from sites of liver metastasis (as opposed to the associated colon primary tumor).

This staging system is current at the time of this writing, although a slightly more detailed staging nomenclature also exists (see following link). For illustrations and more detailed description on how colon cancer progresses throughout the stages and how cancer cells grow through the layers of the colon wall and spread to lymph nodes and other organs, visit the NIH/NCI VisualsOnline website at <http://visualsonline.cancer.gov/browseaction.cfm?topicid=11> (with permission from the author and right holder ©2011 Terese Winslow LLC, US Government has certain rights).

Note that groups D and METS actually represent *concomitant* groups of tumors from the same underlying population of patients with metastases, so that the four groups of tumors only represent three distinct *stages* in the timeline of disease progression. Duke’s group sample sizes were as follows: B (25), C (21), D (35), and liver METS (23), summing to a total of $n = 104$ samples. We provide complete demographic and clinical information in Supporting_Table 1.

2.2. Description of the gene expression data set

Let us label the four Duke’s sample groups by $\{G_g\}_{g=1}^G$, so that $G_1 = \text{group B}$, $G_2 = \text{group C}$, $G_3 = \text{group D}$, and $G_4 = \text{group METS}$. Let us further denote a group cardinal by $n_{G_g} = |G_g|$ for $g = 1, \dots, G$, where $\sum_{g=1}^G n_{G_g} = n$. In our data $G = 4$, $n_{G_1} = 25$, $n_{G_2} = 21$, $n_{G_3} = 35$, $n_{G_4} = 23$, and $n = 104$. We represent our complete data by $(\mathbf{X}, \mathbf{y})$, where $\mathbf{y}$ denotes the n -vector

output variable or response on n independent observations (individuals or patients), and $\mathbf{X}=[\mathbf{x}_j]_{j=1}^p$ is the $(n \times p)$ matrix of expression data set. Each input variable $\mathbf{x}_j$ represents the j th gene expression n -vector (assumed to be suitably normalized). We assume each entry y_i of the output vector $\mathbf{y}$ to take on G possible categorical values, corresponding to the aforementioned Duke's group labels $\{G_g\}_{g=1}^G$.

We generated the expression data from custom Affymetrix expression microarrays (PDL Biopharma, Inc. 932 Southwood Blvd. Incline Village, NV 89451, USA.), which represents over 93% of the transcribed human genome. Briefly, the single PDL Hu03 GeneChip contains exactly 59,798 probesets, which represent $\approx 45,000$ mRNAs and expressed sequence tag (EST) clusters along with ≈ 6200 *ab initio* predicted genes from the human genomic sequence not represented in the mRNAs and EST sequences at the time of chip design. The differentially expressing variables (probesets or genes) across the tumor types were selected by a Bayesian ANOVA model (BAM) [11,12]. We thus selected $p = 1500$ such probesets or genes that were further used in our analysis. We provide the corresponding selected data set with gene expressions and complete gene annotations, available at the time of the study, in Supporting_Table 2.

2.3. Our premise

Because of the heterogeneity of the tumoral phenotype, for instance, in cancer-specific survival or in response to treatment, it has long been suspected that multiple subgroups existed, potentially at all stages, beyond the original Duke's classification 2.1. In this study, our goal was to uncover this heterogeneity of tumors in a hypothetical *spectrum* of tumor stages. Our modeling strategy was, first, to examine samples from the most advanced tumors or METS samples, as these tumors should exhibit potentially the maximum amount of genetic heterogeneity and hence would be anticipated to be the most likely heterogeneous group of tumors.

In contrast to traditional approaches, we address this task by carrying an optimized bump hunting procedure within the class of METS samples alone. What hunting for bump(s) in the METS group translates into is to assume that the true response variable covers a *continuum* of classes representing mixed samples from more than one existing class. So, for instance, other non-METS-labeled samples would serve as prototypes to uncover heterogeneity in the METS-labeled samples, which is efficiently carried out through bump hunting.

3. Local sparse bump hunting of colon cancer gene expression data

3.1. Local sparse bump hunting for class discovery in high-dimensional settings

In a previous article, we showed that finding structures in the form of modes or classes in a high-dimensional space is a difficult exercise where many usual supervised and unsupervised methods can fail [9]. For instance, on the one hand, traditional unsupervised procedures (e.g., clustering and density estimation), which would naturally come to mind to address this problem, will have difficulties separating overlapping groups or underlying distributions. On the other hand, traditional *supervised* approaches such as classification procedures (e.g., by binary decision tree or support vector machine), which, one would imagine have an advantage because they use additional information such as the class response, are in fact inherently inappropriate because they are designed to work when the number of classes is fixed or known in advance and therefore *not* for a class discovery task.

Recall that the goal of bump hunting is to identify subdomains of the search space of input variables/predictors, where the target function (continuous or discrete) assumes local extrema or *bumps*. In our approach, we restrict the initial search space of input variables to a

local optimized/transformed space, wherein the sought-after subdomains are box-shaped partitions determined by an algorithm called the patient rule induction method (PRIM) [13,14], and within which the average value of the target function (here, categorical METS class label output variable) is larger than its average over the rest of the search space of input variables.

The PRIM was predicted to not perform well in very high dimension and when the variables (genes) are strongly correlated or collinear with each other [13]. To address these shortcomings in the $p \gg n$ situation, we showed that there are a number of advantages of hunting for bump/modes by PRIM in an optimized space [9]. To deal with the high dimensionality of the input variable space, and to the strong collinearity that exists between those surveyed variables, combined with the fact that much of this correlation is spurious in nature because of the artifact of having to consider so many variables simultaneously with relatively little samples at hand (the so-called $p \gg n$ paradigm; see also [15]), we used a modified version of PCA with a variable selection feature, called sparse PCA (SPCA) [16], which provides a much more optimal and biologically faithful transformed space within which to search for bumps.

In addition, our bump estimation procedure in the context of high dimensionality benefits from a prior estimation of the number of modes/classes within the optimized SPC space and from a recent theoretical justification on how to tune the PRIM box-induction meta-parameters, α and β , which are the box peeling fraction (or patience) and the minimal box support, respectively. The question regarding optimality of the number of clusters and/or classes found in a data by LSBH is carried out by bootstrap hypothesis testing as explained in Section 3.5 and in greater detail in [9]. Regarding the accuracy of the LSBH bumps themselves, these are directly linked to the large-sample properties of PRIM outcomes (i.e., consistency and convergence rates results of encapsulating boxes), which depend on the PRIM tuning meta-parameters involved [14]. In PRIM, and therefore in our LSBH algorithm, we showed how one needs to carefully and jointly tune the box-induction meta-parameters α and β to successfully optimize the definition and accuracy of the bumps [9]. Basically, that the meta-parameter α cannot be chosen (too small nor too large) reflects an inherent bias/variance trade-off.

We called the resulting algorithm LSBH (see Algorithm 1 and [9]). The remaining part of Section 3 provides details on each of the steps of Algorithm 1 as it pertains to the analysis of the colon cancer data set.

Algorithm 1

Local Sparse Bump Hunting.

-
- Partition the input space into $\{P_r\}_{r=1}^R$ partitions, using, for example, CART.
  - While $r \in \{1, \dots, R\}$:
    - If $g_r > 1$
      - ◆ Run a local SPCA: Estimate the local Sparse Principal Components (SPCs), select a first few of them $j = 1, \dots, q$, where $q \ll p$, and choose an optimal amount of shrinkage/sparsity for each of them, resulting in a total number of nonzero loadings s .
      - ◆ Rotate the local space according to SPCs main directions. Denote the corresponding transformation by $\mathcal{T}(\cdot)$ and the transformed partition by $\mathcal{T}(P_r)$.
      - ◆ Test local multimodality $\hat{m}_0$ within transformed partition $\mathcal{T}(P_r)$.
      - ◆ If $\hat{m}_0 > 1$

- Conditioning on $\hat{m}_0$, estimate PRIM meta-parameters α and β , in $\overline{\mathcal{T}}(P_r)$.
- Run a *local* and *tuned* PRIM-based bump hunting within $\overline{\mathcal{T}}(P_r)$ and get descriptive rules of the bumps in the SPC space.
- Rotate the local rules back into the input space to get rules in the form of “sparse linear combinations”.

– $r \rightarrow r + 1$

- Collect the local rules from all partitions to get a global rule giving a full description of the estimated bumps in the entire input space.

3.2. Partitioning the colon cancer gene expression space

The algorithm calls for an initial partition of the input variable (gene) space, over which bumps will be searched for in a recursive and local manner (Algorithm 1). One approach for partitioning the space is to use a binary decision tree algorithm such as the supervised rule induction method of classification and regression tree (CART) [17]. For later use, let us denote by $\{P_r\}_{r=1}^R$ the partition of the input variable space $\mathcal{S}$, generated by a CART partition algorithm, such that $\mathcal{S} = \bigcup_{r=1}^R P_r$ and $\bigcap_{i \neq j} P_i \cap P_j = \emptyset$. Let us denote a tree by T , its size (no. of terminal nodes) by R , and a terminal node by r , which we later termed an individual partition, denoted by P_r . Let us further denote the partition cardinal by $n_{P_r} = |P_r|$ for $r = 1, \dots, R$, such that $\sum_{r=1}^R n_{P_r} = n$. Finally, let $\mathbf{X}^{(r)} = \{\mathbf{x}_i; \mathbf{x}_i \in P_r \subset \mathcal{S}\}$ be the $(n_{P_r} \times p)$ data sub-matrix, corresponding to the r th partition P_r of the entire input space $\mathcal{S}$, and let $\mathbf{y}^{(r)} = \{y_i; \mathbf{x}_i \in P_r \subset \mathcal{S}\}$ be the corresponding response.

The classification tree generated by the CART partition algorithm should always be pruned/grown to its optimal size (no. of terminal nodes) to avoid under-fitting and over-fitting problems and improve its performance in terms of misclassification error rate. The residual mean deviance (*RMD*) statistic is the usual statistic used for assessing classification tree performance, and the leave-one-out cross-validation is the standard procedure employed for plotting the *RMD* as a function of the tree size. In this plot, one looks for the optimal tree size that minimizes the *RMD* profile [18]. Figure 1 shows the resulting cross-validated number of partitions ($\hat{R} = 3$) found in the entire input (gene) space.

Note that the default full CART classification tree shows a spurious number of $\hat{R} = 9$ terminal nodes, far from the final pruned tree of size $\hat{R} = 3$ terminal nodes (Figure 1). On the one hand, there is biological evidence to support a simpler tree topology with less than $\hat{R} = 9$ terminal nodes, simply because the established Duke’s staging currently supports only four distinct classes, and there is *a priori* little reason to believe that the true number of partitions in the data would be much higher than Duke’s number (Section 2.1). On the other hand, the minimal size of the pruned tree, with only $\hat{R} = 3$ terminal nodes (Figure 1), is consistent with our previous observation that the standard pruning/growing strategy by minimization of the *RMD* statistic does not necessarily estimate accurately the true number of partitions in a data set and results in over-fitting or under-fitting [9]. This inherent uncertainty in the estimation of the true number of structures (here, hidden subclasses) in a data set is the reason for searching them out in a systematical way by local sparse bump hunting (LSBH).

According to the LSBH algorithm, the procedure asks to go recursively over each partition. Because we are first interested in finding whether there is any sample heterogeneity specifically among the METS samples, one wants to focus on those METS-majority class(es) partition(s) only. In the original space of these data, it turns out that such partition

of interest was unique, namely, the terminal node labeled 'M' (Figure 1). We further denote the terminal node of interest by τ_0 and the corresponding partition by P_{τ_0} . In our case, $\tau_0 = 3$.

3.3. Estimating the eligibility for bump(s) in the METS-majority partition (M)

Note that the success of finding a bump(s) in a given partition P_r relies on the initial partition *eligibility*, that is, whether there is initial heterogeneity of class labels of samples in P_r or, formally, whether $g_r > 1$ where g_r is the number of categorical values or class labels

$\{G_{r,g}\}_{g=1}^{g_r} \subseteq \{G_g\}_{g=1}^G$ for the observations in terminal node r . To estimate how likely any given partition of interest is *eligible*, one seeks to estimate the probability $P(r) = \Pr [g_r > 1 \mid r]$ of the r th terminal node, or partition P_r . Note that this probability is not an assessment of the likelihood or potential of finding a bump in the partition of interest, namely, $P'(r) = \Pr [g_r > 1, \hat{m}_0 > 1]$ (Algorithm 1). The probability $P(r)$ can be written as follows:

$$P(r) = 1 - \Pr [g_r = 1 \mid r] \\ = 1 - \Pr \left[G_{g_0} \in \{G_{r,g}\}_{g=1}^{g_r}, \{G_g\}_{g \neq g_0} \notin \{G_{r,g}\}_{g=1}^{g_r} \mid r \right] \quad (1)$$

To get some estimate of the probability $P(r)$, repeat the *whole* CART partition/classification procedure from bootstrapped data sets sampled from the original data $(\mathbf{X}, \mathbf{y})$, a method inspired from bootstrapped confidence levels for phylogenetic trees by Efron *et al.* [19] and bootstrap clustering analysis [20]. We exclude the initial BAM variable selection procedure (Section 2.2) in the simulation, as it would be expected [21], because it was shown that BAM coefficient parameter estimates have much lower mean squared error than traditional least squares estimates [22]. To proceed, generate B bootstrap data sets $(\mathbf{X}^{*1}, \mathbf{y}^{*1}), \dots, (\mathbf{X}^{*B}, \mathbf{y}^{*B})$ and get the corresponding B replicates $\{g_r^{*b}\}_{b=1}^B$, where g_r^{*b} is the cardinal of categorical values for $\mathbf{y}^{(r)*b}$, corresponding to the b th bootstrap sample $(\mathbf{X}^{*b}, \mathbf{y}^{*b})$. Letting $I(\cdot)$ denote the indicator function throughout the paper, we then easily derive a bootstrap estimate of $P(r)$ as the proportion of cases where $g_r^{*b} > 1$; that is,

$$\hat{P}(r) = \frac{1}{B} \sum_{b=1}^B I(g_r^{*b} > 1) \quad (2)$$

Here, with $\tau_0 = 3$ and for $B = 1024$ bootstrap iterations, we found a resulting probability estimate of $\hat{P}(3) = 1 - 47/1024 \approx 0.954$, indicating that we can be fairly confident in the existence of heterogeneous classes in terminal node 3, that is, in the eligibility of partition P_3 .

3.4. High dimensionality and correlation across genes: sparse PCA in the METS-majority partition

After partitioning the space, the LSBH algorithm calls for a complete SPCA within terminal node 3, that is, partition P_3 , where SPCs are to be estimated locally. When a large number of variables ($p = 1500$) are present, even after a BAM variable selection procedure (Section 2.2), we need to select an appropriate number q of leading SPCs, as well as determine the optimal individual amount of shrinkage $\lambda_{1,j}$ for each of them, by using simultaneously a measure of *goodness of fit* and *sparsity*, namely, the cumulative percentage of explained variance (CPEV) and the amount of cumulative percentage of non-zero loadings (CPNZL) (see details in [9]). Figure 2 describes the process for the data in partition P_3 . The general idea is to do for each SPC a stepwise and conditional selection of the amount of shrinkage by optimizing the variance–sparsity trade-off. The amount of CPEV and CPNZL desired dictates the total number q of selected leading SPCs. Here, we selected up to $q = 11$ leading

SPCs, because beyond this point, further SPCs give diminishing returns in terms of explained fit/sparsity trade-off (Figure 2).

Clearly, to ease the biological interpretation, more sparsity can be obtained by allowing larger values of $\lambda_{1,j}$ for each SPC, resulting in a heavier shrinkage, but at the price of losing some fit (explained variance). One can vary sparsity in a biologically sensible manner to ease interpretation of the gene signatures derived. In this example, a medium sparsity has been achieved because a medium shrinkage was imposed (herein called ‘medium’), resulting in a total of $s = 632$ non-zero loadings (genes) with $q = 11$ selected leading SPCs. Here, more or less sparse biological approximations can be found by applying higher or lower arbitrary levels of shrinkage, further termed ‘heavy’ and ‘mild’, respectively.

Similar to previous notations, let us also denote by $\mathbf{Z}^{(g)} = \{\mathbf{z}_i; \mathbf{x}_i \in G_g\}$ the $(n_{G_g} \times q)$ submatrix corresponding to the g th data subset of sample group G_g of the entire SPC space $\mathcal{T}(\mathcal{S})$, where $\mathcal{T}(\cdot)$ denotes the rotation–projection transformation corresponding to the previous SPCA, and by $\mathbf{y}^{(g)} = \{y_i; \mathbf{x}_i \in G_g\}$ the corresponding response. Finally, let us denote by $\mathbf{Z}^{(r)} = \{\mathbf{z}_i; \mathbf{x}_i \in P_r \subset \mathcal{S}\}$ the $(n_{P_r} \times q)$ data submatrix, corresponding to the r th partition P_r of the entire SPC space $\mathcal{T}(\mathcal{S})$.

3.5. Testing multimodality in the sparse principal component space

A critical step of any bump hunting strategy is to correctly estimate the modality of the data. Following the LSBH algorithm (Algorithm 1), we show in this section how we test multimodality in the transformed partition of interest $\mathcal{T}(P_3)$ of the SPC space. Recall in our case that results are shown for terminal node 3 ($r_0 = 3$), that is, the partition of interest P_3 . We use the hypothesis testing framework for testing multimodality (denoted m_0) by bootstrap hypothesis testing. The idea is borrowed from the problem of testing the number of modes in a univariate density estimation problem, where a kernel density estimator is to be tuned with a single smoothing parameter [23], taking its values on a single interval. Here, one considers a series of sequential one-sided tests for each number of modes $m_0 \in \{1, \dots, M\}$ assumed by the target function $f_{\mathbf{Z}^{(3)}}(\mathbf{z}^{(3)})(\alpha, \beta)$ under the null as a function of the PRIM meta-parameters α and β .

$$\begin{cases} H_0: \text{no. of modes } \left[f_{\mathbf{Z}^{(3)}}(\mathbf{z}^{(3)}) \right] (\alpha, \beta) = m_0 \\ H_1: \text{no. of modes } \left[f_{\mathbf{Z}^{(3)}}(\mathbf{z}^{(3)}) \right] (\alpha, \beta) > m_0 \end{cases} \quad (3)$$

Here, the meta-parameter space denoted by Ω is bidimensional (Figure 3), where $\Omega = \mathcal{A} \times \mathcal{B}$ and where $\mathcal{A} = \{0.01, \dots, 0.50\}$ and $\mathcal{B} = \{0.01, \dots, 0.90\}$ are the individual domains of PRIM meta-parameters α and β (see [13] and [9] for more details). To the null hypothesis H_0 corresponds an estimated significance level or p -value, denoted by $\hat{p}(m_0)$. We previously proposed two alternative stopping rules to the sequential testing procedure: either (i) by taking the value of $\hat{m}_0$ maximizing $\hat{p}(m_0)$ or (ii) by taking the minimum value $\hat{m}_0$ for which we cannot reject the null hypothesis at a certain significance level, say, η (see [9] for details). Formally,

$$\begin{aligned} \text{rule no.1: } \hat{m}_0 &= \arg\max_{m_0} \{\hat{p}(m_0)\} \\ \text{rule no.2: } \hat{m}_0 &= \min_{m_0} \{\hat{p}(m_0) > \eta\} \end{aligned} \quad (4)$$

Using the more conservative stopping rule no. 2, one finds a unimodal response (i.e., $\hat{m}_0 = 1$) in the data $\mathbf{Z}^{(3)}$ of partition $\mathcal{T}(P_3)$ of the SPC space; that is, we would apparently not find any subgroup in the METS observations, and the analysis would stop here (Figure 3). In contrast, using rule no. 1, the estimated modality found is $\hat{m}_0 = 2$. Given the

conservativeness of rule no. 2, and the estimation error attached to the corresponding estimates due to sampling variability, one may actually choose the alternative stopping rule no. 1. For exploratory purposes, we show in the next sections what the results are, should we use rule no. 1, that is, with an estimated modality of $\hat{m}_0 = 2$.

3.6. Fine tuning of the bump hunting meta-parameters in the sparse principal component space

Conditioning on the two modes that we found, and on the PRIM meta-parameters estimates $\hat{\alpha}$ and $\hat{\beta}$ as described in Dazard and Rao's article [9], we carried out a bump hunting step *per se* in the transformed data subset $\mathbf{Z}^{(3)}$, that is, the transformed partition $\mathcal{T}(P_3)$ of the *optimized* SPC space $\mathcal{T}(\mathcal{S})$, (Section 3.4). The partition of interest encompasses $n_{P_3} = 22$ observations, among which the proportion of METS-labeled samples (majority samples) is dominant as compared with the rest of non-METS-labeled samples (i.e., our so-called prototypes: D-labeled samples only in our data). The goal is to identify a subdomain of $\mathcal{T}(P_3) \subseteq \mathcal{T}(\mathcal{S})$ within which the proportion of METS samples would be further enriched in the METS class as compared with the full complement of observations.

To use the bump hunting framework, one first has to recast the problem into a binary classification setup (see details in [9]). Let us therefore consider the dummy output variable $\mathbf{y}_M^{(3)}$ for the data subset $\mathbf{Z}^{(3)}$, such that $\mathbf{y}_M^{(3)} = \mathbf{I}(\text{class}(\mathbf{y}^{(3)}) = \text{'M'})$, where $\mathbf{y}_M^{(3)} = 1$ now represents a majority class of interest, say, with respect to all the other classes combined, where $\mathbf{y}_M^{(3)} = 0$. With the assumption that $\hat{m}_0 = 2$ modes, the LSBH procedure isolated two *bona fide* bumps (i.e., two subdomains of classes) in the colon cancer data set, splitting the observations in data subset $\mathbf{Z}^{(3)}$ into two subdomains corresponding to the bump support and the outside of it (Figure 4). The corresponding 'bump rule' involves here a single SPC $\mathbf{z}_4$:

$$\mathcal{R}_{\mathcal{T}(P_3)}: \mathbf{z}_4 < 2562 \quad (5)$$

So, the bump rule is in that case a hyperplane ($\mathbf{z}_4 = 2562$), but it would take the form of a hyperbox in the general case. Here, it is evident to see how the estimated bump support is described by a single subpartition of $\mathcal{T}(P_3)$ along the $\mathbf{z}_4$ component. Figure 4 shows the estimated bump support, projected into the three-dimensional subspace $[\mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4]$. The scatterplot shows how the $n_{P_3} = 22$ samples in partition P_3 segregate into 15 samples that actually sit together within the bump and 7 samples that do *not* (Figure 4). Interestingly, of these 15 samples, most are METS, except for the one that is a D sample, which, as a prototype, was able to drag onto itself the 14 other METS samples. More investigation and biological insight into the nature of the bump rule and of the segregated samples are given thereafter (Section 4).

Furthermore, there was indication that this bump could not be found by clustering or by a CART binary decision tree (data not shown). CART applied to the data in terminal node 3 alone revealed no significant subpartition, indicating that the bump in the terminal node M is not the result of a trivial tree-based partitioning method (data not shown). In addition, the gap statistic applied to the same data revealed only *one* significant cluster, encompassing all the samples (data not shown), indicating that the bump in the terminal node M does not stem from an underlying cluster structure of samples. The aforementioned results remain true whether in the input variable space or in the SPC space because of rotational invariance.

With similar notations as those previously mentioned, let us denote from now on by $\{B_{b,r_0}\}_{b=1}^B$ the subpartition of P_{r_0} generated by LSBH algorithm, and corresponding to the bumps found in P_{r_0} , such that $P_{r_0} = \bigcup_{b=1}^B B_{b,r_0}$ and $\bigcap_{b_i \neq b_j} B_{b_i,r_0} \cap B_{b_j,r_0} = \emptyset$. Note that $\{B_m\}_{m=1}^B$ is the

PRIM-induced box sequence $\{B_m\}_{m=1}^{\hat{m}_0}$ found in P_{r_0} as described in [9] and in its supplemental materials, that is, B is the modality, or $B = \hat{m}_0$. Let us further denote the cardinal of

subpartition $b = 1, \dots, B$ in P_{r_0} by $n_{B_{b,r_0}} = |B_{b,r_0}|$, such that $\sum_{b=1}^B n_{B_{b,r_0}} = n_{P_{r_0}}$. Finally, denote by $\mathbf{Z}^{(b,r_0)} \subset \mathbf{Z}^{(r_0)} = \{\mathbf{z}_i: \mathbf{x}_i \in B_{b,r_0} \subset P_{r_0}\}$ the $(n_{B_{b,r_0}} \times q)$ data submatrix, corresponding to the r_0 th subpartition B_{b,r_0} of the entire SPC space $\mathcal{T}(S)$, and similarly $\mathbf{y}^{(b,r_0)} = \{y_i: \mathbf{x}_i \in B_{b,r_0} \subset P_{r_0}\}$ for the response. In our case, $r_0 = 3$, $B = \hat{m}_0 = 2$, and $n_{B_{1,3}} = 15$, $n_{B_{2,3}} = 7$, so that the subpartition $\{B_{b,3}\}_{b=1}^2$ of P_3 correspond respectively to the inside and outside bumps (5).

3.7. Statistical validation of the sparse bump signature

To convince ourselves that we simply not have over-fitted the data and modeled noise, it is desirable to get a measure of statistical significance of the distributions of the ‘bump-signature samples’ for each group independently. Using the hypothesis framework, we can test the null hypothesis that the observed distribution of the bump-signature samples in the true case for each group (B, C, D, and METS) is equal to the (null) distribution of a random set of samples for that group, further termed ‘random bump-signature samples’, generated from a random bump rule under the following null:

$$\begin{aligned} H_0: & \text{pdf}(\text{bump-signature samples} | \text{samples} \in G_g) \\ & \sim \text{pdf}(\text{random bump-signature samples} | \text{samples} \in G_g) \end{aligned}$$

To simulate the null distribution of random bump-signature samples for each group $\{G_g\}_{g=1}^G$, we use a resampling (permutation) scheme as follows.

- Generate a random gene signature by permuting the s components or loadings of vector $\mathbf{w}_4$, obtained from the matrix of orthogonal eigenvectors $\mathbf{W}$ from the original SPCA procedure in partition P_3 , so that a random s -vector of loadings is formed.
- For each permutation sample, rotate the local SPC space, according to corresponding main (random) SPCs directions, using the pre-determined number of selected SPCs (i.e., $q = 11$), as described in Algorithm 1 of the LSBH article in [9].
- Run a patient bump hunting step in the corresponding random SPC space, as described in Algorithm 1 of the LSBH article in [9], so that a random bump rule is formed.
- Segregate the samples satisfying the random bump rule from those that do not, so that random bump-signature samples are formed for each group (B, C, D, and METS).

We carried out a distributional test against H_0 and calculated the associated statistics and p -values for each group. Specifically, we used a bivariate non-parametric two-sample two-sided Cramer test [24]. We calculated the test statistic for the permutation and its value compared with the observed one. We judge how extreme the observed statistic by counting the proportion of times that more extreme values are obtained from the null distribution. These proportions are our p -values estimates. To compute the corresponding p -values, for each group $\{G_g\}_{g=1}^G$, we repeated the process on $B = 1000$ permutation samples. We show histogram density estimators of the distributions of p -values for each group in Figure 5. Notice immediately for each stage, how few of the p -values are larger than, say, a $\alpha = 0.05$ level. This is a clear confirmation that what we have uncovered is not simply a result of over-fitting the data.

4. Nature of the sparse bump gene expression signature

4.1. Predictive nature of the bump signature of colon tumors heterogeneity

In our case, the bump rule (5) determining the bump support in the terminal node 3 (partition P_3) turns out to be a single univariate inequality. It translates into a single linear combination of s variables (genes or probesets), that is, the s -component vector $\mathbf{w}_4$ of loadings from the matrix of orthogonal eigenvectors, denoted by $\mathbf{W}=[\mathbf{w}_j]_{j=1}^q$ (with $q=11$), which we further term interchangeably ‘bump gene expression signature’ or ‘bump signature’ in short. Although simple and sparse, the biological nature of this bump signature needs to be better understood. To address this question, we analyzed individually in each of the sample group $\{G_g\}_{g=1}^G$ those samples exhibiting the specific bump signature and those not, which we further referred to as ‘bump-signature samples’ and ‘non-bump-signature samples’, respectively, in a given sample group.

Recall that our premise was to look into Duke’s colon cancer primary tumor stage groups B, C, and D to determine if any subgroups once found in the METS samples could be tracked back in stage. We systematically looked for any bump-signature samples among each Duke’s colon cancer tumor group and checked how they distributed comparatively. To that end, we applied the bump rule (5) to each of the data sets $\{\mathbf{Z}^{(g)}\}_{g=1}^g$ corresponding to Duke’s sample groups $\{G_g\}_{g=1}^G$, then segregated the samples accordingly. To our surprise, the bump signature could be detected all the way back to the earliest colon cancer tumor group (B). To show the distributions of samples, a bivariate local polynomial regression model [25] was fit (with fixed bandwidth) by Duke’s stage group for density estimation. We show the resulting plots in Figure 6.

Notice immediately in Figure 6 that the LSBH bump boundary (threshold value $\mathbf{z}_4 = 2562$) nicely separates the two prominent densities of METS samples that were found by the employed density estimator, thereby further supporting the definition of the bump support and out-of-bump support of METS samples. Also, there is *visual* indication that a separate distribution of bump-signature samples exists as well in other Duke’s stage groups. Finally, notice the relative homogeneity of proportions of bump-signature samples throughout the four stage groups (Figure 6 and Table I).

We first used a test of equality or homogeneity of proportions for testing the apparent conservation of proportions across groups. We used the value of Pearson’s chi-squared test statistic for testing equality or homogeneity of proportions for testing the null that the proportions (probabilities of success) in several groups are the same:

$$\begin{cases} H_0: & \text{The four populations from which the patients were drawn} \\ & \text{have the same true proportion of bump–signature samples.} \\ H_1: & \text{This proportion is different in at least one of the populations.} \end{cases}$$

The four-sample test for equality of proportions (without continuity correction) on the bump-signature samples versus the non-bump-signature samples in each sample group gave a p -value of $P \approx 0.0989$. We also performed a chi-squared test for *trend* in proportions across sample groups ($P \approx 0.3442$), as well as between *primary* and *metastatic* tumor groups ($P \approx 0.7945$). All these tests reached the same conclusion: the null hypothesis H_0 could not be rejected at the $\alpha = 0.05$ level; that is, a *common* proportion of bump-signature samples exists in all colon tumors from the earliest stages of the disease. Because this specific bump signature is present in all colon tumors, including the (non-metastatic) primary tumors, this

is indicative that it has a broader biological meaning than just that of a specific metastatic tumor signature. Rather, it points to a sort of a predictive signature of heterogeneity in colon tumors/patients from the earliest stages of the disease.

4.2. Survival correlation to the sparse bump signature

Consider now the dummy outcome variable $\mathbf{y}_{B_1}^{(3)}$ for the data subset $\mathbf{Z}^{(3)}$, such that $\mathbf{y}_{B_1}^{(3)} = \mathbf{I}(\mathbf{X}^{(3)} \in B_{1,3})$, that is, $\mathbf{y}_{B_1}^{(3)}$ is a bump inclusion indicator variable, where $\mathbf{y}_{B_1}^{(3)} = 1$ indicates inclusion within the bump and $\mathbf{y}_{B_1}^{(3)} = 0$ not. We first analyzed the demographic variables available (Duke's colon tumor groups, colon sites of tumor at diagnosis, patient age, and patient gender) to check as to whether a demographic variable would predict the bump status. We fit several models including logistic regression, elastic-net, and tree-based models with the dummy variable $\mathbf{y}_{B_1}^{(3)}$ as the response. We found none of the variables or any combination of them to predict well the bump signature (data not shown). Immediately, this implies a new conclusion: the bump inclusion/exclusion outcome variable is not simply a surrogate, or at least simply explained, by any of these demographic variables, or a combination of them.

We analyzed the censored patient survival time within Duke's colon tumor groups (B, C, D, and METS) to test whether the bump inclusion/exclusion outcome variable could be correlated to patient survival time and, thereby, have some survival predictive value. We ran a two-sample log-rank test for each group on the bump-signature samples versus the non-bump-signature samples. In one group (B), the log-rank test did reach statistical significance at the $\alpha = 0.05$ level (Table II), indicating that a difference in survival time between the bump-signature samples and the non-bump-signature samples is significant in this stage of the disease.

The fact that survival time differences between bump-signature samples and non-bump-signature samples did not quite reach the $\alpha = 0.05$ level of significance in later Duke's stages of the disease may be a reflection of a confounding effect. Indeed, a selection bias is imposed especially on C, D, and METS group samples, which were collected from patients who were selected to undergo therapeutic treatment and/or palliative hepatic resections, in which both the selection process and the intervention favor selecting out individuals who have unusually long survival times.

It is noteworthy that we would expect the METS samples to behave similarly to the D samples because by definition they are essentially random samples from the same underlying population of patients (Section 2.1). In line with this, it is of interest to test whether there was any *trend* or difference in the survival/specific mortality of bump-signature samples across groups. To test this hypothesis, we used the Mann–Kendall trend test for monotonic trend in a time series as well as the Cox–Stuart trend test (on the basis of the binomial distribution) that is known to be very robust for trend analysis. In both cases, the null hypothesis was as follows:

H_0 : There is no trend of survival/specific mortality of bump–signature samples across groups/time.

Both tests did not reject the null hypothesis H_0 at the $\alpha = 0.05$ level: the two-sided Mann–Kendall trend test gave a p -value of $P \approx 0.734$, and the two-sided Cox–Stuart trend test gave a p -value of $P \approx 0.750$. This was also confirmed by a time series plot, where a lowess smooth did not suggest any trend, and by the autocorrelation plot in this data that did not show significance either (data not shown). So, even though there is not a *uniformly* strong

evidence of survival time difference across all groups between patients having the bump-signature samples versus those not having it, the uniformity found in the cancer-specific survival time of bump-signature samples across all stages of the disease is indicative that this is what the bump signature tracks across all stages (Section 4.1). Figure 7 illustrates this point specifically for the earliest and latest stages of the disease: B and METS.

We hypothesize that patients exhibiting the bump signature are more likely to have better cancer-specific survival and that this bump signature in the B stage group tracks an improved late-stage condition/prognosis of the disease. This points to a Duke's stage group-by-bump-signature interaction effect. Specifically, we looked at the cancer specific probability of survival difference between non-bump-signature B samples and bump-signature C samples, as well as between non-bump-signature B samples and bump-signature METS samples. The corresponding log-rank test for separation between Kaplan–Meier curves did not reject the null hypothesis of no difference between the probability curves in both comparisons (p -values yielded $P \approx 0.683$ and $P \approx 0.969$, respectively).

We suggested earlier that the specific bump gene expression signature (w_4) at play pointed to a sort of a predictive signature of heterogeneity in colon tumors (Section 4.1). Our working hypothesis is that these bump-signature versus non-bump-signature samples formally represent two distinct phenotypes of tumors, characterized by a specific bump gene expression signature that is not necessarily associated with metastasis but with some survival predictive value. We suggest that non-bump-signature samples may demonstrate a more aggressive course associated with a shorter cancer-specific survival.

5. Functional analysis and biological validation

To further investigate the biological meaning of our finding, we carried out some functional analysis of the sparse bump signature w_4 at the medium level of shrinkage (Section 3.4), resulting in a total of $s = 632$ non-zero loadings (genes). We obtained gene annotations by converting the probe identifiers to NCBI's UniGene cluster identifiers (<http://www.ncbi.nlm.nih.gov/unigene>) or EntrezGene identifiers (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>) as needed. With each identifier come numerous annotations on the transcription locus itself in order to annotate probable or known corresponding gene functions. We interrogated UniGene IDs for known canonical pathways, biological functionality, and network analysis against the Ingenuity Knowledge Database, using the Ingenuity Pathway Analysis (IPA) functional profiling tool (<http://www.ingenuity.com/>). Similarly, we used Entrez-Gene identifiers for gene ontology (GO) categorization using the GOstats database search tool from the Bioconductor project (<http://bioconductor.org/>) against the Gene Ontology Consortium database (<http://www.geneontology.org/>).

In either case, we performed a statistical test to identify a given category of interest (e.g., a GO category or an IPA pathway) with enriched gene numbers. For IPA, we calculated the significance of enrichment from Fisher's exact test, whereas for GO, we used the hypergeometric test. In each test, we calculated p -values of enrichment from the right-tail distribution of the test statistics (one-sided test). To control the FDR, we corrected all p -values for multiplicity by using the Benjamini–Hochberg (BH) method [26]. In either case, the reference set to be used with the statistical test is critical. We used the set of $p = 1500$ pre-filtered probes (see pre-filtering paragraph in Section 2.2) as only those probes (genes) that were pre-filtered and selected as differentially expressed by Bayesian ANOVA between Duke' sample groups (Section 2.1) had a chance of ever being monitored and analyzed for functional analysis, and should therefore be considered. This is to be distinguished from the total set of 59,798 uniquely identified probes present on the chip or from the total number of entries of any knowledge database, both of which would yield incorrect p -values [27, 28].

We assessed the significance of a network by the probability of finding the observed network by random chance. We calculated the p -value from the number of 'network-eligible molecules' contained in the network, the network size, as well as the total number of network-eligible molecules analyzed and the total number of molecules in the Ingenuity Knowledge Database that could potentially be included in networks [29]. Table III lists the significantly enriched biological functions, biochemical pathways, and networks determined by the IPA project database. Complete lists with corresponding bar plots are available as Supporting_Table 3 and Supporting_Figure 1, respectively. We provide the graphical representation of the top-ranking network in Supporting_Figure 2 with complete annotations of molecules within it in Supporting_Table 4 as well.

We carried out a similar statistical analysis as before to identify GO categories with enriched gene numbers. We show in Table IV the most significant GO categories derived from the bump gene expression signature as determined by the corresponding GO tree at the user-selected annotation depth for each root GO (biological process, molecular function, and cellular component). We provide complete lists in Supporting_Table 5 with corresponding bar plots in Supporting_Figure 3).

6. Discussion

In the data, the rule (5) determining the bump support turns out to be a simple one, amounting to a sparse linear combination of s genes, satisfying a single inequality, which we termed bump rule. In general, this may not necessarily be the case because we can potentially expect up to q linear combinations satisfying an inequality (hyperbox). Furthermore, we have found that a specific gene expression signature would manifest itself from the unique majority METS node of samples that was found by CART. In the general case, majority METS nodes may not necessarily be unique. It is also conceivable that some of the METS samples could have been partitioned into non-majority METS tree terminal nodes. This actually happened in our data: terminal node B contained two METS samples. In all these cases, more complex bump rules would result. The description and interpretation of such phenomenon would naturally be more complex.

Previous analysis of gene expression microarrays from human cancers have yielded several important types of insights. Analysis of gene expression profiles of breast cancers and of large cell lymphomas have yielded recognition of new molecularly defined subtypes of these diseases [1,2]. Analysis of breast cancers and of lung cancers selected to be of homogeneous classes, such as stage 1 adenocarcinomas of the lung [30] or node-negative breast cancers [31], have additionally yielded patterns that have defined tumors having better or worse prognosis, with lesser or greater propensity to metastasize. Interestingly, when looking for molecular differences between human primary tumors and metastases of multiple tumor types to unmatched primary adenocarcinomas, researchers have found a gene expression signature of metastasis in primary solid tumors [32]. In our case, the bump signature, which we have defined, divides MSS colon cancers into two distinct subtypes of disease that are defined by differences in patterns of gene expression.

Our underlying premise did not assume gene expression homogeneity in the metastatic group METS. Instead, we assumed potential subgroup(s), from which specific gene signature(s) may be found in other stages. We planned that any subgroups once found in the METS samples would exhibit a specific gene signature that could potentially be tracked back in primary tumor stages and determine when it had developed.

The complexity in the tumoral phenotype could explain why we observe multiple subgroups at all stages of Duke's classification. Our finding relies on the assumption that multiple subgroups exist at all stages, which could not be captured by the initial classification of

Duke, far from being accurate and able to reflect the multiplicity and complexity of the tumoral phenotypes. Our view is that only a combination of multiple measures of gene expression, like our bump gene expression signature, also sometimes defined as a *meta-gene*, might be able to account for the differences in phenotypes and define propensity and predictions.

What is the nature of this bump gene expression signature? We have shown that it is not simply a surrogate of any demographic variables or any combination of them. Thus, the origin of the different expression patterns captured by bump-signature versus non-bump-signature samples is more likely to be biologically based.

The functional analysis of the bump gene expression signature points to pro-inflammatory mediators and immune response genes, all found in either biological functions, pathways, or gene ontologies (Tables III and IV). For instance, the only significant pathway is the acute-phase response, that is, a rapid inflammatory response that can be triggered by a variety of causes (tissue injury, trauma or surgery, or immunological disorders) including *neoplastic growth*. Typically, it consists of an increase in pro-inflammatory cytokines and a change in concentration of several plasma proteins due largely to an *altered hepatic metabolism* (reviewed in [33]). This is totally consistent with the known genetic makeup of all three stages of tumor development: initiation, progression, and metastasis (reviewed in [34–36]). In addition, inflammation is a well-known risk factor for colorectal tumors. Note that these observations are also totally in line with the executive summary of the inflammation and cancer by the NIH and its specific recommendations for the NCI (<http://www.cancer.gov/think-tanks-cancer-biology/page8>).

It does not appear that these specific subtypes differ in intrinsic propensity for metastasis, as similar proportions of bump-signature samples were observed, for instance, in Duke's B cases that did not generate metastasis, as in Duke's D cases that did generate metastasis (60% and 80%, respectively). However, non-bump-signature samples may have an intrinsic biological difference that translates into more aggressive growth and shorter patient survival time once tumors reach the stage of generating tumor metastases (Duke's D and METS cases). Therefore, the distribution of the specific non-bump-signature samples indicates that it is not a tumorigenic or metastatic progression signature but a signature associated with malignancy conversion potential. We will need analysis of further samples to validate this hypothesis. Moreover, we bear in mind that a more rational approach for investigating a correlation to any response variable (such as a clinical outcome of interest like survival) would warrant a *complete supervision* of the bump hunting algorithm by the response variable.

Our observations further serve to highlight the molecular complexity of the colon cancer phenotype. This complexity, as reflected in different clinical stages and now as reflected in emergence of the difference between bump-signature versus non-bump-signature patients, may well explain challenge in thus far deriving expression-based predictive or prognostic rules for colon cancer. We certainly favor future attempts to derive such expression-based predictors of tumor behavior by using larger samples sets of colon tumors in which samples are first segregated into bump-signature versus non-bump-signature categories.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

J.-E. Dazard is an assistant professor in the Bioinformatics Division of the Center for Proteomics and Bioinformatics at Case Western Reserve University (CWRU). This research was conducted in part while he was a postdoctoral fellow in the Division of Biostatistics of the Department of Epidemiology and Biostatistics at CWRU, mentored by J. Sunil Rao, under NIH grant R25-CA04186. J. Sunil Rao is a professor in the Division of Biostatistics, Department of Epidemiology and Public Health at The University of Miami. He was partially supported by NSF grant DMS-0405072 and by NIH grant K25-CA89867. Sanford Markowitz is an investigator of the Howard Hughes Medical Institute and is the Ingalls Professor of Cancer Genetics in Hematology–Oncology and Molecular Biology with co-appointment in the Department of Molecular Biology at Case Western Reserve University (email: sxm10@po.cwr.edu). He was partially supported by NIH grant RO1-CA120237. Additional support came from grants of the Case Comprehensive Cancer Center (NIH-National Cancer Institute P30-CA043703) and the Clinical and Translational Science Award (NIH-National Center for Research Resources UL1-RR024989).

References

1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. *Nature*. 2000; 406(6797):747–752. [PubMed: 10963602]
2. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JJ, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000; 403(6769):503–511. [PubMed: 10676951]
3. Northcott PA, Korshunov A, Witt H, Hielscher T, Eberhart CG, Mack S, Bouffet E, Clifford SC, Hawkins CE, French P, Rutka JT, Pfister S, Taylor MD. Medulloblastoma comprises four distinct molecular variants. *Journal of Clinical Oncology*. 2011; 29:1408–1414. [PubMed: 20823417]
4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999; 286(5439):531–537. [PubMed: 10521349]
5. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*. 2000; 406(6795):536–540. [PubMed: 10952317]
6. Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D, Van de Rijn M. Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet*. 2002; 359(9314):1301–1307. [PubMed: 11965276]
7. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Science of the United States of America*. 2002; 99(10):6567–6572.
8. Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E. A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society*. 2002; 64(Series B):717–736.
9. Dazard J-E, Rao JS. Local sparse bump hunting. *Journal of Computational and Graphical Statistics*. 2010; 19(4):900–929. [PubMed: 22399839]
10. Cohen, A.; Minsky, B.; Schilsky, R. Cancer of the colon. In: DeVita, VTJ.; Hellman, S.; Rosenberg, S., editors. *Cancer: Principles and Practice of Oncology*. 5. Lippincott-Raven; Philadelphia, PA: 1997.
11. Ishwaran H, Rao JS. Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association*. 2003; 98(462):438–455.
12. Ishwaran H, Rao JS. Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association*. 2005; 100(471):764–780.

13. Friedman JH, Fisher NI. Bump hunting in high-dimensional data. *Statistics and Computing*. 1999; 9:123–143.
14. Polonik W, Wang Z. Prim analysis. *Journal of Multivariate Analysis*. 2010; 101(3):525–540.
15. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society*. 2008; 70(5 Series B):849–911. [PubMed: 19603084]
16. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*. 2006; 15(2):265–286.
17. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *The Wadsworth Statistics/Probability Series*. Chapman and Hall/CRC; Boca Raton, FL: 1984. *Classification and Regression Trees*.
18. Ripley, BD. *Pattern recognition and neural networks*. Cambridge University Press; Cambridge: 1996.
19. Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Science of the United States of America*. 1996; 93(14):7085–7090.
20. Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Science of the United States of America*. 2001; 98(16):8961–8965.
21. Ambrose C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99(10):6562–6566. [PubMed: 11983868]
22. Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*. 2005; 33(2):730–773.
23. Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*. Chapman & Hall/CRC; London: 1993.
24. Baringhaus L, Franz C. On a new multivariate two-sample test. *Journal of Multivariate Analysis*. 2004; 88:190–206.
25. Loader C. Local likelihood density estimation. *The Annals of Statistics*. 1996; 24:1602–1618.
26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. 1995; 57(Series B):289–300.
27. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*. 2005; 21(18):3587–3595. [PubMed: 15994189]
28. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*. 2008; 9(7):509–515.
29. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, Miller-Graziano C, Moldawer LL, Mindrinos MN, Davis RW, Tompkins RG, Lowry SF. A network-based analysis of systemic inflammation in humans. *Nature*. 2005; 437(7061):1032–1037. [PubMed: 16136080]
30. Beer DG, Kardias SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*. 2002; 8(8):816–824.
31. Van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart y, Voskuil y, Schreiber y, Peterse y, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*. 2002; 347(25):1999–2009. [PubMed: 12490681]
32. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. *Nature Genetics*. 2003; 33(1):49–54. [PubMed: 12469122]
33. Ray LB. Inflammation and tumor progression. *Science STKE*. 2007; 394:246.
34. Balkwill F, Mantovani A. Inflammation and cancer: back to virchow? *Lancet*. 2001; 357(9255):539–545. [PubMed: 11229684]
35. Coussens LM, Werb Z. Inflammation and cancer. *Nature*. 2002; 420(6917):860–867. [PubMed: 12490959]
36. Karin M. Inflammation and cancer: the long reach of Ras. *Nature Medicine*. 2005; 11(1):20–21.

37. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data mining, Inference, and Prediction. Springer Science; New York: 2009.

Figure 1.

Partitioning of the original input space. Left: full CART classification tree. Middle: leave-one-out cross-validated residual mean deviance (RMD) statistic profile as a function of the CART tree size R (no. of terminal nodes). The RMD profile of the CART tree reaches its minimum for a tree size of $\hat{R} = 3$ terminal nodes and its maximum for the full tree with $\hat{R} = 9$ terminal nodes. Based on the aforementioned minimizer found for the RMD , the estimated optimal tree size is $\hat{R} = 3$ terminal nodes (red arrow indicates the result of the stopping rule). In case a minimum cannot be found over a reasonable range for R (as here with $R \in \{1, \dots, 9\}$), a useful objective stopping rule to estimate the tree size R can be applied [37]. Right: final pruned CART classification tree with $\hat{R} = 3$ terminal nodes. Terminal node class labels are determined by the usual majority class label rule. Here, after growing/pruning the initial tree, there was no majority-group D terminal node.

Figure 2.

Determining an optimal degree of sparsity in high dimension. Upper row: shrinkage plot sequence of local sparse bump hunting in the microarray data set. The superimposed sequences of cumulative percentage of explained variance (CPEV) and cumulative percentage of non-zero loadings (CPNZL) scree plots are shown for the data in terminal node 3 (partition P_3 , $n_{P_3} = 22$). The chosen amount of shrinkage is shown for each sparse principal component (SPC). A total of 12 leading SPCs were entertained (only the first two and the last one are shown for illustration purposes). Bottom row: final scree plots of CPEV and CPNZL surfaces as a function of the number of selected SPCs and the amount of shrinkage for the data in terminal node 3. The CPEV and CPNZL surfaces are shown at step 11 of the selection process, conditioned on the choices of λ_1 made for all previous SPCs. At this step, an optimal $\lambda_{1,j} = 4.9 \times 10^5$ and $q = 11$ leading SPCs were selected, resulting in an overall ‘medium’ level of shrinkage ($s = 632$).

Figure 3.

Testing multimodality in the partition $\mathcal{T}(P_3)$ of interest. Left: contours of the number of modes in the meta-parameter space $\Omega = \mathcal{A} \times \mathcal{B}$, where $\mathcal{A} = \{0.01, \dots, 0.50\}$ and $\mathcal{B} = \{0.01, \dots, 0.90\}$ for the target function in partition $\mathcal{T}(P_3)$ of the sparse principal component space. Right: step function of p -values for each mode tested under H_0 . The red solid line represents the threshold of significance at the $\eta = 0.05$ level.

Figure 4.

Graphical illustration of local sparse bump hunting in the microarray data set. Scatterplot of observations in data set $Z^{(3)}$ from partition $\mathcal{T}(P_3)$, projected into the three-dimensional sparse principal component subspace (z_2, z_3, z_4) . The estimated bump support is shown (hashed red) with bump membership. Red and green labels denote the inside and outside bump, respectively. Bump boundary: black dotted line.

Figure 5.

Histogram density estimates of the estimated distributions of p -values per group at a medium level of shrinkage ($s = 632$).

Figure 6.

Scatterplots with superimposed contours of smoothers of the distributions of samples by Duke's stage group. All observations are from data set $\mathbf{Z}^{(3)}$, that is, from partition $\mathcal{T}(P_3)$, projected into the bidimensional sparse principal component subspaces $(\mathbf{z}_3, \mathbf{z}_4)$ or $(\mathbf{z}_2, \mathbf{z}_4)$. Group cardinals: group B, $n_{G1} = 25$; group C, $n_{G2} = 21$; group D, $n_{G3} = 35$; and group METS, $n_{G4} = 23$. In each plot, the black dashed line represents the bump boundary of the bump rule (threshold value $\mathbf{z}_4 = 2562$ at the medium level of shrinkage $s = 632$), showing the separation of bump-signature samples versus non-bump-signature samples. Counts of bump-signature samples and proportions with respect to groups are indicated.

Figure 7.

Left and middle: Kaplan–Meier curves (solid lines) with 95% confidence intervals (dotted lines) for the earliest and latest stages of the disease: B and METS. The specific censoring indicator is the death event: alive = 0, death with disease (DWD) = 1. Curves are marked at each censoring time. Survival time is in months. Right: scatterplot of survival time (in months) by bump signature in the METS group of samples, showing the separation of bump-signature samples versus non-bump-signature samples by the bump rule ($z_4 = 2562$), where the vertical black dashed line represents the bump boundary or threshold value $z_4 = 2562$ of the bump rule (at the medium level of shrinkage $s = 632$).

Table I

Sample proportion estimates of the bump-signature samples versus the non-bump-signature samples in each Duke's stage sample group.

Duke's stage	B	C	D	METS
Sample proportion estimates	0.600	0.905	0.800	0.739

Table II

Two-sample log-rank test on the bump-signature samples versus the non-bump-signature samples in each Duke's stage sample group.

Duke's stage	B	C	D	METS
Log-rank p -value	0.0359	0.1741	0.1257	0.1354

Table III

Top-enriched biological functions, canonical pathways, and corresponding networks derived from the bump gene expression signature at the ‘medium’ level of shrinkage ($s = 632$).

	n_o	$\frac{n_o}{n_e}$	P
<i>Biological functions</i>			
Cancer	190		3.95×10^{-4}
Gastrointestinal disease	90		5.73×10^{-3}
Inflammatory response	78		3.93×10^{-2}
Cellular movement	104		3.93×10^{-2}
Inflammatory disease	144		3.94×10^{-2}
Hematol. syst. dev. and function	68		3.94×10^{-2}
Immune cell trafficking	63		3.94×10^{-2}
<i>Canonical pathways</i>			
Acute-phase response signaling	26	$\frac{26}{178} \approx 0.146$	7.06×10^{-4}
<i>Networks (top functions)</i>			
Cancer, tissue development, cell–cell signal. and interaction	35	$\frac{26}{35} \approx 0.743$	$\approx 10^{-39}$
Molec. transport, metab. disease, and cell death	35	$\frac{26}{35} \approx 0.743$	$\approx 10^{-39}$
Neurol. disease, organismal injury and abnormalities, cell–cell signal. and interaction	35	$\frac{23}{35} \approx 0.657$	$\approx 10^{-35}$

For each entry, the gene number, the ratio of gene enrichment, and its corresponding BH-corrected p -value are shown. Only those entries with at least two genes and with $P < 0.05$ are shown, but for conciseness, only the top-ranking networks (3) are shown. If we let s and p be the total number of genes in the bump gene signature set and reference set respectively, j be the total gene number in the category of interest for the reference set, and n_o , n_e be the observed and expected gene numbers respectively in the category of interest for the gene signature set, then $n_e = j \frac{s}{p}$, and the ratio $\frac{n_o}{n_e}$ gives the gene enrichment ratio for the category of interest. When applicable, the hypergeometric test p -value is given (when $\frac{n_o}{n_e} > 1$ by

$P = \sum_{i=n_o}^s \Pr(I=i|s, p, j)$, where I is the random variable of observed number of genes in the category of interest for the gene signature set,

$$\Pr(I=i|s, p, j) = \frac{\binom{p-j}{s-i} \binom{j}{i}}{\binom{p}{s}}$$

and

Table IV

Top-ranking GO categories derived from the 'bump gene expression signature' at the 'medium' level of shrinkage ($s = 632$) and at the user-defined annotation depth (three for biological process and two for molecular function and cellular component).

	Odds ratio	Exp. count	Gene count	Category size	P
<i>Biological process</i>					
Immune response	2.816	18.478	46	697	9.91×10^{-9}
Immune system process	2.380	26.033	55	982	9.47×10^{-8}
Triglyceride metabolic process	9.357	1.060	8	40	8.29×10^{-6}
Response to external stimulus	2.220	20.201	41	762	1.26×10^{-5}
Acute-phase response	10.894	0.822	7	31	1.32×10^{-5}
<i>Molecular function</i>					
Antigen binding	9.487	1.442	11	56	1.51×10^{-7}
Structural constituent of ribosome	4.723	4.068	17	158	6.78×10^{-7}
Protein binding	1.603	196.603	241	7636	3.33×10^{-6}
Cell surface binding	11.512	0.669	6	26	4.17×10^{-5}
Vitamin transporter activity	21.834	0.283	4	11	1.23×10^{-4}
<i>Cellular component</i>					
Extracellular region	2.164	51.144	95	1942	1.02×10^{-9}
Extracellular region part	2.569	23.149	53	879	1.39×10^{-8}
Extracellular space	2.833	16.117	41	612	4.13×10^{-8}
Platelet alpha granule lumen	14.145	0.869	9	33	1.23×10^{-7}
Platelet alpha granule	10.795	1.185	10	45	2.02×10^{-7}

For conciseness, only the top-ranking gene ontology (GO) categories (5) are shown. 'Odds ratio' is the odds ratio for each category term tested. 'Exp. count' is the expected number of genes in the selected gene list to be found by random chance at each tested category term. 'Gene count' for each category term tested is the number of observed genes from the gene set that are actually annotated at the term. 'Category size' for each category term tested is the number of genes from the gene universe that are annotated at the term. P is the raw p -values for each category term tested.