Am J Cancer Res 2018;8(10):2088-2095
www.ajcr.us /ISSN:2156-6976/ajcr0079878

Original Article

The extrema of circulating miR-17 are identified as biomarkers for aggressive prostate cancer

Greg Dyson[1], Batoul Farran[1], Susan Bolton[1], Douglas B Craig[1], Alan Dombkowski[2], Jennifer L Beebe-Dimmer[1], Isaac J Powell[3], Izabela Podgorski[4], Lance K Heilbrun[1], Cathryn H Bock[1]

[1]Karmanos Cancer Institute and Department of Oncology, Wayne State University, Detroit MI, USA; [2]Karmanos Cancer Institute and Department of Pediatrics, Wayne State University, Detroit MI, USA; [3]Karmanos Cancer Institute and Department of Urology, Wayne State University, Detroit MI, USA; [4]Karmanos Cancer Institute and Department of Pharmacology, Wayne State University, Detroit MI, USA

Received May 18, 2018; Accepted June 3, 2018; Epub October 1, 2018; Published October 15, 2018

Abstract: MicroRNAs (miRNAs) constitute short non-coding RNAs that can post-transcriptionally modulate the expression of many oncogenes and tumor suppressor genes engaged in key cellular processes. Deregulated serum miRNA signatures have been detected in various solid cancers including prostate cancer, suggesting that circulating miRNAs could function as non-invasive biomarkers of tumor emergence and progression. To determine whether serum miRNA expression levels are different between patients with aggressive and non-aggressive prostate cancer, we analyzed a panel of miRNAs from the blood of African American (AA) prostate cancer patients using a new recursive partitioning method that allows hypothesis testing of each split. We observed that both extrema of circulating miR-17, i.e. upregulation and downregulation, are associated with aggressive prostate cancer. A similar effect was observed in tumor samples from a separate dataset representing a different population of prostate cancer patients and in AA prostate cancer samples from the TCGA. The dual effect is consistent with the contradictory findings on the role of miR-17 in prostate cancer progression, whereby it controls important oncogenic and tumor-suppressive genes.

Keywords: Prostate cancer, microRNA, miR-17, classification, hypothesis testing, regression trees

Introduction

Prostate cancer (PCa) constitutes the second most frequently diagnosed cancer and fifth leading cause of mortality among men worldwide [1]. Its incidence is higher in the industrialized world, suggesting that diet and lifestyle represent potential risk factors [2]. In fact, nutrition and exercise modulate serum factors that limit growth and elicit apoptosis in androgen-dependent PCa cells, whereas increased body mass index, blood pressure and metabolic factors provoke an increased risk of PCa death [3]. The highest rates of PCa are observed in African American (AA) and Caribbean populations, suggesting that genetic predisposition might contribute to disease development and aggressiveness [1]. Current screening methods for PCa include digital rectal exam, serum level of prostate-specific antigen and transrectal ultrasound guided biopsy [3]. These methods do not have high sensitivity and specificity, often leading to misdiagnosis or over-diagnosis, highlighting the need for improved methods of early detection and management.

Circulating microRNAs (miRNAs) have emerged as promising diagnostic and prognostic biomarkers in various solid cancers including PCa due to their increased stability, bio-availability, and frequent deregulation during tumorigenesis. miRNAs are small non-coding RNA molecules that repress protein expression by cleaving mRNA or inhibiting its translation, which enables them to fine-tune the expression of an intricate network of oncogenes and tumor suppressor genes implicated in key cellular functions, such as cell cycle progression, cell growth and apoptosis. Aberrant expression of miRNAs contributes to tumorigenesis by de-repressing or silencing key regulatory proteins. Extensive research efforts are currently underway to uncover clinically relevant tissue and serum miRNA signatures of diagnostic and prognostic value for the stratification and monitoring of PCa [4-6].

Recursive partitioning (RP) is a statistical tool used to create subgroups of objects (individuals) with differential levels of the phenotype; utilizing binary splits derived from values of the input variables to create the subgroups. A thorough review [7] of RP reveals a rich history of statistical developments since the algorithm was first published [8]. Highlights of this review include a brief account of the incremental advancements in RP theory and a description of modifications to apply to various classes of outcome data. A central part of the research on the RP algorithm is defining a rule for when to stop growing a tree (or how much to prune) to yield an optimal prediction model. Classification and regression trees (CART) [9] were developed to classify objects using binary splits with either a continuous or categorical response based in part on the idea formed from research on RP. The key additions of the CART algorithm were pruning large trees to a parsimonious size using a complexity parameter and utilizing crossvalidation to estimate error in prediction. A crossvalidation heuristic is typically used to select the optimal size of the tree via the complexity parameter; e.g., select the tree that results in the smallest crossvalidation error or select the smallest tree where the crossvalidation error is within one standard deviation of the minimum crossvalidation error [10]. Further enhancements which made an effort to incorporate statistical testing into the CART fitting process focused on methods to optimize the pruning process [11, 12]. Subsequent research led to the development of conditional trees, which embeds RP in a conditional inference framework and includes stopping rules based on multiple testing [13]. Ensemble methods like random forests [14] have also been utilized to increase prediction accuracy while consequently sacrificing the interpretability of the resultant model [15].

The objective of the RP with hypothesis testing (RP-HT) method introduced in this paper is to eliminate the parameters, heuristics, and pruning algorithms utilized by RP methods by utilizing hypothesis testing at each potential split in the tree. In doing so, new splitting objective functions for continuous and categorical responses are developed. The proposed RP-HT method builds upon previous work [16] incorporating hypothesis testing into the Patient Rule-Induction Method [17]. A dataset measuring a panel of circulating microRNAs from African American prostate cancer patients is analyzed using this new technique. The models produced by RP-HT will be compared with those produced by CART and a simulation study is utilized to estimate the overall Type I error.

Statistical methods

The new objective function developed for RP-HT will evaluate the splits that maximize the proportion of each of the observed classes for a categorical response variable, or maximizes the mean rank for a continuous response variable. The rank rather than the observed response variable is used for continuous responses to provide a robust estimate that is not sensitive to outlying responses, and to allow for normal distribution theory of order statistics to be applied to data that may not be normally distributed. RP-HT will continue splitting the data and building the tree until there is no statistically significant split available at each of the existing nodes. Like many other RP-based algorithms, RP-HT is greedy in the sense that it selects the optimal split at a given node without looking forward or going back on earlier decisions [18]. RP-HT is especially greedy as it does not identify the split that results in the smallest *p*-value, but rather the largest result proportion or mean rank.

A support parameter and its selection process are also introduced to regularize the RP-HT fitting process. This parameter is the minimum proportion of the observations under consideration (from the mother node) that are required to be in both of the daughter nodes. A range of support parameters between 0.05 and 0.50 incremented by 0.005 will be considered. The lower bound is set at 5% to ensure that terminal nodes have adequate size. The upper bound is the maximum value that the support parameter could be, given that both daughter nodes have to meet the size requirement. The greediness of the RP-HT algorithm is controlled by this support parameter (which is the same at every node in the tree): the smaller the support parameter, the smaller the created sub-nodes will be. A smaller support parameter will in general lead to a tree with more terminal nodes. For a continuous response, the support parameter that maximizes the Kendall's tau correlation [19] between the observed response and the predicted response is selected (among trees with at least one significant split). For categorical responses, the support parameter that minimizes the sum of one minus the predicted probability for the true class squared is selected (among trees with at least one significant split).

Using a support parameter and the objective function, RP-HT determines the optimal split at a given node. For each potential split, the RP-HT performs exact hypothesis testing for categorical responses and exact or asymptotically exact hypothesis testing for continuous responses to determine if the split was statistically significant, conditional on the number of possible splits from all potential explanatory variables. The derivation of the null distribution for both categorical and continuous responses to construct a *p*-value is detailed in the Supplemental Methods. At each node in the tree-building process, the RP-HT algorithm will determine the most statistically significant split, given the support parameter. If the *p*-value from that split is smaller than the input nominal significance level, then that spilt is incorporated into the definition of the tree. This process will then be conducted on the daughter nodes created from that split. If the *p*-value from a split is not smaller than the nominal significance level, then the algorithm denotes it to be a terminal node and all observations in that node are assigned to that output class. The algorithm continues until there is no split available with a *p*-value less than the nominal significance level at each available node and therefore every observation has been assigned to a terminal node class. For categorical responses, the predicted value for an observation in a particular terminal node is the response category with a plurality of members of that terminal node. For continuous responses, the predicted value for an observation in a particular terminal node is the median of the raw responses of all of the observations that make up the terminal node. The median is used instead of the mean to lessen the impact of outlier observations.

As the implementation of the original CART algorithm is proprietary, we use the version in the rpart package [20] in R programming environment [21] to compare with our proposed algorithm. The rpart version uses the Gini index to select the optimal splits for categorical response variables [22]. Other parameters included in this implementation include "minbucket" (the minimum number of observations required in all daughter nodes), complexity parameter (the minimum decrease in lack of fit required to consider a split), and "minsplit" (the minimum number of observations in a mother node needed to attempt a split) that are used to control the tree-fitting process [22]. For our comparison we used to the default value for theses parameters, with the exception of "minbucket" which we set to 10. We chose not to compare RP-HT with ensemble method like random forests, bagging and boosting to maintain the interpretability of resultant model and ensure a fair comparison.

Clinical methods

miRNA discovery study

A panel of 92 plasma microRNA levels were assayed in a cohort of blood samples from 116 PCa patients from the Karmanos Cancer Institute in Detroit MI, USA using the Exiqon microRNA PCR Cancer Focus panel. Expression levels are normalized using the median polish method [23]. Full details on the cohort, array and pre-processing steps are found in a separate manuscript [24]. As a consequence of the median polish, the median expression within each miRNA will be 0. For this paper we utilize the subset of 93 men with self-identified AA race and the subset of 44 miRNAs with detectable levels on all of them. Of the 93 men, 68 have non-aggressive PCa (defined as a Gleason score of 7 (3 + 4) or lower); while 25 have aggressive PCa (defined as a Gleason score of 7 (4 + 3) or higher). The objective is to identify miRNAs that are associated with disease aggressiveness.

miRNA replication studies

In the absence of publically available circulating miRNA datasets on prostate cancer patients, we chose to use two other studies that examined miRNAs expressed in tumor tissue.

One study (denoted Taylor study) [25] examined miRNAs (on the Agilent-019118 human miRNA microarray) expressed in metastatic and pri-

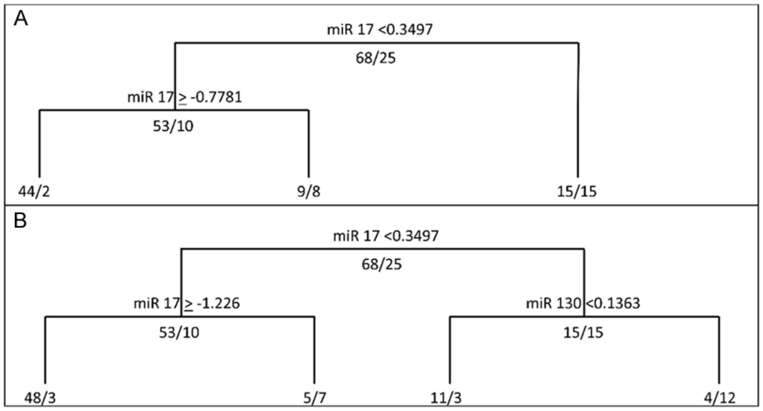

Figure 1. Tree diagrams of the RP-HT (A) and CART (B) analysis of the miRNA dataset. For both trees, the text above a split indicates the Boolean statement that defines the split; taking the left-hand path if the text statement is true and the right-hand path if the statement is false. The two numbers below each split indicate the <number of controls>/<number of cases> at each step in the process.

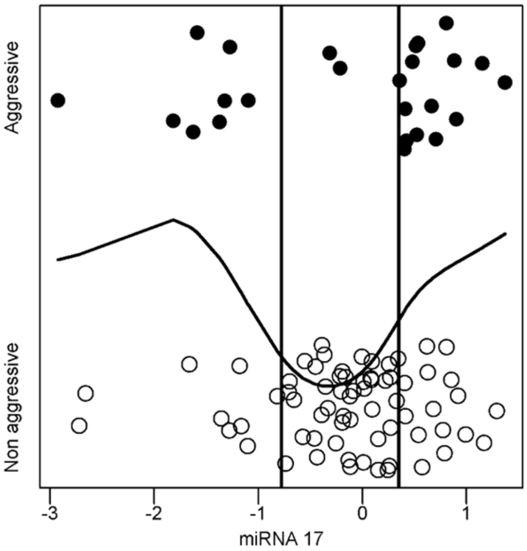

Figure 2. Scatterplot of miR-17 versus jittered aggressiveness status. The vertical lines indicate the cutpoints (-0.778, 0.350) defined by RP-HT. The group of samples at the negative and positive extrema have an aggressiveness incidence of 0.471 (8/17) and 0.500 (15/30), respectively. The middle group of samples has an incidence of 0.043 (2/46). A smoothed spline fit to the non-jittered data emphasizes the U-shaped distribution of aggressive disease incidence.

mary PCa tumors. Combining the data from the supplement to the source paper [25] and the online repository in GEO (GSE21032), we are able to define Gleason's score as we did in our study. As in our study, we normalize all the measured miRNA expression levels using median polish, although we only examine miR-17 as part of this validation exercise. Note that as only 16 of the 110 samples in the Taylor study are AA, results on the whole cohort are presented here. Therefore, although we cannot consider the Taylor study as a true 'replication' dataset, since it measures miRNAs from a different biospecimen type (tumor tissue) in a (mostly) different population using a different assay, we will determine if any significant findings from our miRNA analysis is also detected in this dataset. There are 76 Taylor samples with a Gleason score of 7 (3 + 4) or less and 34 samples with a Gleason score of 7 (4 + 3) or more.

Prostate cancer samples from the Cancer Genome Atlas (TCGA) had miRNA expression quantification measured via miRNA-Seq [26] as part of the data collection process. Linking the phenotype data downloaded from cBioPortal [27], we are able to define Gleason score as we did in our study. Unfortunately (for our validation purposes), only 7 (5 with Gleason score of 7 (3 + 4) or lower and 2 with Gleason score of 7 (4 + 3) or higher) of these samples were AA, while 146 (78 with Gleason score of 7 (3 + 4) or lower and 68 with Gleason score of 7 (4 + 3) or higher) were EA. Like the Taylor study, the TCGA study also cannot be considered a true 'replication' as it is measured in tumor tissue from a diverse population using sequencing (all different than our study). As in our study, we normalize all the measured miRNA expression levels using median polish, although we only examine miR-17 as part of this validation exercise.

Results

The optimal support parameter was identified as 0.26 for this dataset. Three subgroups were created, defined by a single circulating miRNA: miR-17. A CART analysis identified 4 subgroups, defined by 2 miRNAs. A comparison of the RP-HT and CART models is shown in **Figure 1**. **Figure 2** displays a scatterplot of miR-17 versus jittered disease aggressiveness and the identi-

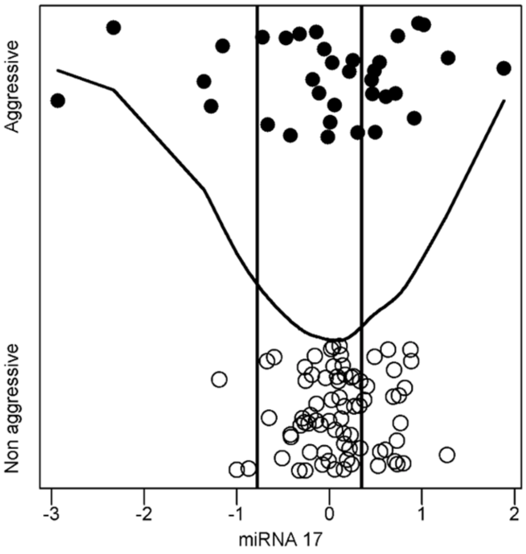

Figure 3. Scatterplot of miR-17 versus jittered aggressiveness status for the Taylor study. A similar U-shaped pattern of aggressiveness incidence is observed in the Taylor dataset. The curve is a smooth spline fit to the non-jittered data, while the vertical lines indicate the thresholds derived from our dataset (-0.778, 0.350). The group of samples at the negative and positive extrema have an aggressiveness incidence of 0.625 (5/8) and 0.406 (13/32), respectively. The middle group of samples has an incidence of 0.229 (16/70).

fied RP-HT thresholds. The left-hand subgroup consists of 17 individuals, 8 with aggressive disease and 9 without (incidence = 0.47). The right-hand subgroup consists of 30 individuals, half with and half without aggressive PCa (incidence = 0.50). The sample of 46 individuals in the middle of the distribution contains only 2 aggressive cancers (incidence = 0.043). A smoothed spline fit to the non-jittered data (using 0 and 1 as the response variable) emphasizes the U-shaped distribution of aggressive disease incidence.

Taylor dataset

Figure 3 shows a plot of miR-17 versus jittered disease aggressiveness for the Taylor dataset. Utilizing the same thresholds from the analysis of our dataset, we find that the group of samples at the negative and positive extrema have an aggressiveness incidence of 0.625 (5/8) and 0.406 (13/32), respectively. The middle group of samples has an incidence of 0.229 (16/70). The fitted smooth spline (using the non-jittered response) shows a similar, albeit weaker, U-shaped pattern of incidence (even after excluding the right-most point); although the race categorization and biospecimen type were different between the 2 studies. Appling the CART-defined tree to this dataset yielded poor prediction, as miR-130 was not discriminatory (data not shown).

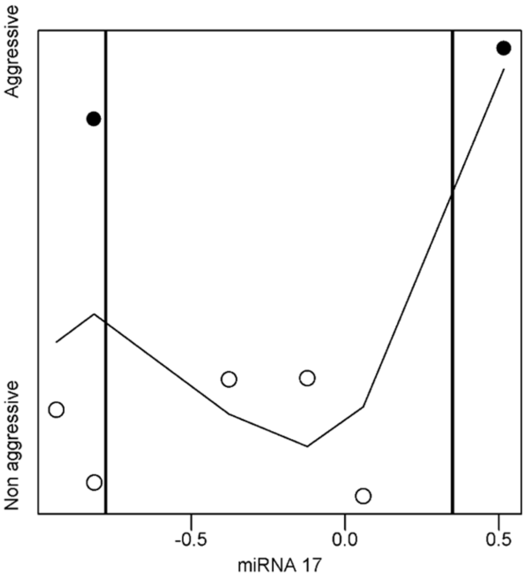

Figure 4. Scatterplot of miR-17 versus jittered aggressiveness status for the AA TCGA study. A similar U-shaped pattern of aggressiveness incidence is observed in the AA TCGA dataset. The curve is a smooth spline fit to the non-jittered data, while the vertical lines indicate the thresholds derived from our dataset (-0.778, 0.350). The group of samples at the negative and positive extrema have an aggressiveness incidence of 0.333 (1/3) and 1.000 (1/1), respectively. The middle group of samples has an incidence of 0.000 (0/3).

Figure 4 shows a plot of miR-17 versus jittered disease aggressiveness for the TCGA AA subset. Utilizing the same thresholds from the analysis of our dataset, we find that the group of samples at the negative and positive extrema have an aggressiveness incidence of 0.333 (1/3) and 1.000 (1/1), respectively. The middle group of samples has an incidence of 0.000 (0/3). We obviously cannot read too much into this result as the sample size is much too small, but the result is at least promising. The U-shaped pattern was not observed in the EA subset (data not shown).

A simulation study was undertaken to estimate an experiment-wide error rate. To that end, we computed the 5th percentile (out of 1000 iterations) of the logistic regression *p*-value for null data sets (the predictor variables are not associated with the binary response) under various conditions. The detailed results from the simulation are in Supplementary Table 1. For our miRNA database, utilizing 1,000 iterations with 68 "0" s, 25 "1" s and 44 null predictor variables, we found that 49 (5%) resulted in a smaller or equal *p*-value than observed in the data analysis of the actual data set. A total of 909 (91%) resulted in an outputted tree; while CART produced a tree for each iteration. So even when there is no association between predictors and response, 90% of the null RP-HT analyses and 100% of the null CART analyses returned a tree, highlighting the need for computation of an experiment-wide error rate, regardless of whether an external validation dataset is available.

Discussion

RP-HT overcomes obstacles inherent in an RP analysis including the lack of hypothesis testing of each potential split, the lack of a consistent mechanism to stop growing the tree and the need for subjective input parameters, and instead a nominal *p*-value is the only requirement. Overfitting will still be an issue (typical of greedy classification models); hence the need for *p*-value calculation, external replication, and validation of any significant finding. The RP-HT technique identified a U-shaped pattern in aggressive disease as a function of miR-17, whereby samples at the positive (upregulation) and negative (downregulation) extrema were associated with aggressive PCa. This duality is consistent with the contradictory findings on miR-17's role in PCa progression, as illustrated by studies reporting both its upregulation and downregulation in cells and tissue [28, 29]. miR-17 belongs to the miR-17-92 cluster, which is implicated in the regulation of many oncogenes and tumor suppressor genes involved in distinct and opposing pathways [30]. This duality encapsulates the complexities of cancer progression and enables miR-17-92 to fine-tune the levels of pro-apoptotic and anti-apoptotic genes respectively.

Using a systems biology approach, Cloonan, et al. [31] demonstrated this dichotomy by unmasking an extensive yet contradictory network of genes regulated by miR-17-5p, including inhibitors of cellular proliferation such as TSG101, RBL1 and MAPK9 and promoters of cellular proliferation such as MYCN, NCOA3 and NR4A3. Other studies have similarly captured this duality. For instance, miR-17 upregulation was reported to enhance tumor growth by directly repressing PTEN [32] and TIMP3 [33] and to convey chemoresistance to cisplatin in human PCa cell lines [28]. Conversely, others observed a tumor suppressive role for miR-17-5p in PCa cells, revealing that miR17 loss induced progression to anti-androgen resistance by upregulating FGD4, LIMK1, cyclin D1 and SSH1 cell cycle regulatory proteins [29]. Other targets of miR-17, such as E2 F1, have also been reported to function as oncogenes or tumor suppressor genes depending on the nature of the other driving oncogenic mutations present [5]. These contrasting findings highlight the contextual character of miR-17 function and stress the importance of considering cancerous phenotypes as a product of clonally heterogeneous and complex genetic networks subject to multiple layers of dysregulation [31].

Several limitations exist for our study. Firstly, the RP-HT takes longer to run than a standard CART analysis, as it analyzes 91 potential support parameters. The CART analysis of our miRNA dataset was instantaneous using rpart, while HT-RP took approximately 40 seconds to complete. Datasets with a larger number of variables will intensify this issue, while utilizing parallel processing on multiple computing nodes will mitigate it. Secondly, none of the publically available miRNA PCa datasets had a majority of AA samples and had measurements of circulating miRNA. Thus we were compelled to verify our miR-17 result using a dataset of PCa men with mostly European ancestry, measured from tumor samples and a small number of AA TCGA samples. Checking the incidence of aggressive PCa relative to miR-17 in another circulating miRNA dataset of AA men will be necessary before miR-17 is validated as a biomarker for aggressive PCa. Lastly, the sample sizes of both our discovery and replication datasets were small, increasing the chance of a spurious result. Although the procedure we utilized in this paper, including the estimation of an overall experiment-wise error rate and external replication, should lessen that chance.

In conclusion, utilizing a new statistical tool, we have observed that both extrema of circulating miR-17 are associated with aggressive prostate cancer. This effect was observed in tumor samples from separate datasets measured on different assays representing different populations of prostate cancer patients. Our result is consistent with the conflicting findings on the impact that miR-17 has in PCa progression, namely that it controls both oncogenic and tumor-suppressive genes.

Acknowledgements

This work was supported by the National Institutes of Health Cancer Center Support Grant 5P30CA022453, Department of Defense grants W81XWH-09-1-0203 and W81XWH-13-1-0477 & Michigan Fund for Cancer Research.

Disclosure of conflict of interest

None.

Address correspondence to: Greg Dyson, Karmanos Cancer Institute and Department of Oncology, Wayne State University, Detroit MI, USA. E-mail: dysong@karmanos.org

References

[1] Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends--an update. Cancer Epidemiol Biomarkers Prev 2016; 25: 16-27.

[2] Zhou CK, Check DP, Lortet-Tieulent J, Laversanne M, Jemal A, Ferlay J, Bray F, Cook MB, Devesa SS. Prostate cancer incidence in 43 populations worldwide: an analysis of time trends overall and by age group. Int J Cancer 2016; 138: 1388-400.

[3] Vanacore D, Boccellino M, Rossetti S, Cavaliere C, D'Aniello C, Di Franco R, Romano FJ, Montanari M, La Mantia E, Piscitelli R, Nocerino F, Cappuccio F, Grimaldi G, Izzo A, Castaldo L, Pepe MF, Malzone MG, Iovane G, Ametrano G, Stiuso P, Quagliuolo L, Barberio D, Perdona S, Muto P, Montella M, Maiolino P, Veneziani BM, Botti G, Caraglia M, Facchini G. Micrornas in prostate cancer: an overview. Oncotarget 2017; 8: 50240-50251.

[4] Wang BD, Ceniccola K, Yang Q, Andrawis R, Patel V, Ji Y, Rhim J, Olender J, Popratiloff A, Latham P, Lai Y, Patierno SR, Lee NH. Identification and functional validation of reciprocal microRNA-mRNA pairings in african american prostate cancer disparities. Clin Cancer Res 2015; 21: 4970-84.

[5] Ambs S, Prueitt RL, Yi M, Hudson RS, Howe TM, Petrocca F, Wallace TA, Liu CG, Volinia S, Calin GA, Yfantis HG, Stephens RM, Croce CM. Genomic profiling of microRNA and messenger RNA reveals deregulated microRNA expression in prostate cancer. Cancer Res 2008; 68: 6162-70.

[6] Yates C, Long MD, Campbell MJ, Sucheston-Campbell L. miRNAs as drivers of TMPRSS2-ERG negative prostate tumors in African American men. Front Biosci (Landmark Ed) 2017; 22: 212-29.

[7] Loh WY. Fifty years of classification and regression trees. Int Stat Rev 2014; 82: 329-48.

[8] Morgan JN, Sonquist JA. Problems in analysis of survey data, and a proposal. Journal of the American Statistical Association 1963; 58: 415.

[9] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Belmont, Calif: Wadsworth International Group; 1984.

[10] Venables WN, Ripley BD, Venables WN. Modern applied statistics with S. New York: Springer, 2002.

[11] Cappelli C, Mola F, Siciliano R. A statistical approach to growing a reliable honest tree. Computational Statistics & Data Analysis 2002; 38: 285-99.

[12] Zhang H, Singer B. Recursive partitioning in the health sciences. New York: Springer, 1999.

[13] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. Journal of Computational and Graphical Statistics 2006; 15: 651-74.

[14] Breiman L. Random forests. Machine Learning 2001; 45: 5-32.

[15] Loh WY. Improving the precision of classification trees. Annals of Applied Statistics 2009; 3: 1710-37.

[16] Dyson G, Sing CF. Efficient identification of context dependent subgroups of risk from genome-wide association studies. Stat Appl Genet Mol Biol 2014; 13: 217-26.

[17] Friedman JH, Fisher NI. Bump hunting in high-dimensional data. Statistics and Computing 1999; 9: 123-43.

[18] Curtis SA. The classification of greedy algorithms. Science of Computer Programming 2003; 49: 125-57.

[19] Kendall MG. A new measure of rank correlation. Biometrika 1938; 30: 81-93.

[20] Therneau TM, Atkinson EJ and Ripley B. Rpart: recursive partitioning. R package version 4.1-10. https://CRAN.R-project.org/package=rpart 2015.

[21] R Core Team. R: a language and environment for statistical computing. 2017.

[22] Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routine. Technical Report. Section of Statistics, Mayo Clinic 1997.

[23] Tukey JW. Exploratory data analysis. Addison-Wesley: Reading, Massachussetts, 1977.

[24] Farran B, Dyson G, Craig D, Dombkowski A, Beebe-Dimmer JL, Powell IJ, Podgorski I, Heilbrun L, Bolton S, Bock CH. A study of circulating microRNAs identifies a new potential biomarker panel to distinguish aggressive prostate cancer. Carcinogenesis 2018; 39: 556-61.

[25] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao YH, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y, Mitsiades N, Landers T, Dolgalev I, Major JE, Wilson M, Socci ND, Lash AE, Heguy A, Eastham JA, Scher HI, Reuter VE, Scardino PT, Sander C, Sawyers CL, Gerald WL. Integrative genomic profiling of human prostate cancer. Cancer Cell 2010; 18: 11-22.

[26] Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. Cell 2015; 163: 1011-25.

[27] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012; 2: 401-4.

[28] Zhou P, Ma L, Zhou J, Jiang M, Rao E, Zhao Y, Guo F. miR-17-92 plays an oncogenic role and conveys chemo-resistance to cisplatin in human prostate cancer cells. Int J Oncol 2016; 48: 1737-48.

[29] Ottman R, Levy J, Grizzle WE, Chakrabarti R. The other face of miR-17-92a cluster, exhibiting tumor suppressor effects in prostate cancer. Oncotarget 2016; 7: 73739-53.

[30] Xiang J, Wu J. Feud or friend? The role of the miR-17-92 cluster in tumorigenesis. Curr Genomics 2010; 11: 129-35.

[31] Cloonan N, Brown MK, Steptoe AL, Wani S, Chan WL, Forrest AR, Kolle G, Gabrielli B, Grimmond SM. The miR-17-5p microRNA is a key regulator of the G1/S phase cell cycle transition. Genome Biol 2008; 9: R127.

[32] Dhar S, Kumar A, Rimando AM, Zhang X, Levenson AS. Resveratrol and pterostilbene epigenetically restore PTEN expression by targeting oncomiRs of the miR-17 family in prostate cancer. Oncotarget 2015; 6: 27214-26.

[33] Yang X, Du WW, Li H, Liu F, Khorshidi A, Rutnam ZJ, Yang BB. Both mature miR-17-5p and passenger strand miR-17-3p target TIMP3 and induce prostate tumor growth and invasion. Nucleic Acids Res 2013; 41: 9688-704.

Supplemental methods

Null hypothesis derivation

Categorical responses: The derivation of the null distribution for RP-HT for a categorical response is as follows. Let V indicate the set of valid potential splits (defined by a continuous or categorical explanatoryvariable): subsets that are sized between [$n \times \beta$] and n-[$n \times \beta$] that have a relative frequency ofa response category greater than the input relative frequency of that same response category. Let m_1 denote the frequency of the response category of interest in a sample of size m taken without replacement from the population of size n which has n_1 of the response category of interest. Then the density function of the null distribution [$f(\theta|V)$] as a function of the observed relative frequency values (θ) is defined as:

$$
\begin{aligned}
f(\theta \mid V) &= \sum_{j=[\beta\times n]}^{n\text{-}[\beta\times n]} \{f(\theta, m = j \mid V)\} \\
&= \sum_{j=[\beta\times n]}^{n\text{-}[\beta\times n]} \{f(\theta \mid m = j, V) \times p(m = j \mid V)\} \\
&= \sum_{j=[\beta\times n]}^{n\text{-}[\beta\times n]} \{f\left(\frac{m_1}{m} \mid m = j, V\right) \times p(m = j \mid V)\} \\
&= \sum_{j=[\beta\times n]}^{n\text{-}[\beta\times n]} \{f(m_1 \mid m = j, V) \times p(m = j \mid V)\},
\end{aligned}
$$

where m is assumed the be marginally discrete uniform and m_1 follows a hypergeometric distribution. Additionally, define N as the number of potential valid subsets for a response category and $F(\theta|V)$ as the cumulative distribution function of θ values that satisfy V for a response category. The algorithm then proceeds to enumerate all possible combinations of m_1 and m that satisfy V to derive the null distribution. As the number of observable combinations could be very large, we used 10,000 values evenly spaced between the input relative frequency and maximal observable relative frequency to estimate the density functions for a response category. The computed density is rounded up to the nearest of these 10,000 categories for storage. Then the null distribution of the maximum relative frequency is given by $N \times f(\theta|V) \times F(\theta|V)^{N-1}$ for a response category.

From the null distribution for each of the response categories, a p-value is computed given the observed maximal relative frequency obtained from each response category. If multiple splits within a response category result in the same maximal response, the split with the larger number of observations is selected. If there are multiple splits with the same maximal response and number of observations, one of the splits is chosen at random to be selected. The split/response category combination that results in the smallest p-value is then outputted. If multiple splits/response category combinations result in the same minimum p-value, a similar algorithm examining the relative frequency and number of observations is used to select the outputted split/response category, randomly outputting a split if all elements are tied. As the response categories will likely have different input relative frequencies, a single null distribution is not possible. However, to allay concerns about multiple testing, the outputted p-value is multiplied by the number of response classes as a Bonferroni correction. The simulation study in section 5 explores this issue of multiple testing in further detail.

Continuous responses: We rank the response breaking ties randomly, since tied ranks are not consistent with the distribution theory requirements used in the hypothesis testing component of the RP-HT algorithm. As a consequence, continuous response variables with tied values may result in different resultant RP-HT objects, even with the same support parameter. The responses are re-ranked at each node in the tree. Therefore, the rank for the same observation will change at each node.

Mathematically, the objective function at each node is to select from the input set of integers from 1 to *n*, a subset of size *m* with a statistically significantly higher mean than in the input population. Potential valid subsets (denoted as *V*) must be sized between [$n * \beta$] and n-[$n * \beta$] and have a mean response greater than the input mean (n + 1)/2. The null distribution for this algorithm will be either a weighted enumeration of all valid subsets or a normal approximation if *n* gets too large. The enumeration of valid subsets is weighted to ensure that the marginal distribution of *m* is discrete uniform. Other formulations of the marginal distribution of *m* can be entertained, including an unweighted enumeration which is equivalent to treating all possible combinations of subsets across all possible *m* values as equally likely. Once the approximation gets close enough (in terms of Kullbeck-Leibler [KL] divergence) to the exact distribution, then RP-HT will switch over to the approximation to save computational time and resources. The exact calculation involves enumerating the mean responses of all possible valid subsets given *n* and β to construct the null distribution. As with the categorical responses, the enumerated mean values are rounded up to the nearest of the 10,000 values for storage. Finally, the distributions for all possible *m* values are combined to produce a single density function.

We use normal distributional properties of order statistics to construct the approximation. A randomly selected subset of size *m* taken from a set of integers from 1 to *n* without replacement has a normal distribution with mean (n + 1)/2 and variance (n-m) × (n + 1)/(12 × m) in the limit. The observable mean responses (θ) given n and m that satisfy *V* will be the midpoints of the nonoverlapping intervals of width 1/*m* used to estimate the density. The probability that the above normal distribution falls within each of these intervals corresponding to each observable θ is then computed. This value is then scaled by the relative frequency of the marginal distribution of *m* for valid subsets to ensure that *m* is marginally discrete uniform. As with the categorical response, 10,000 values evenly spaced between the mean of the input response and maximal observable response across the range of *m* values are used to estimate the density function. Computed density values from the proper normal distribution are rounded up to the nearest of the 10,000 values for storage. Finally, the densities for all possible *m* values are combined to produce a single density function. When m is less than 4 or greater than *n*-4, the exact calculation is used since the approximation poorly matches the exact distribution near the extrema of m. When *n* is greater than or equal to 22, the RP-HT will use the approximation rather than the exact computation as the KL difference between the two distributions is less than 0.001 for the range of possible support parameters (data not shown). Since the objective function selects the subset with the largest θ from all valid subsets, the null probability distribution function of the largest θ is given by the probability density function of the maximum order statistic of the null distribution, namely $N \times f(\theta|V) \times F(\theta|V)^{N-1}$, where *N* is the number of potential valid subsets for one of the response categories and $f(\theta|V)$ and $F(\theta|V)$ are the distribution and cumulative distribution functions of observable θ values that satisfy *V*.

Supplementary Table 1. Estimated experiment-wide error rate utilizing a variety of number of observations, percentage of cases, covariates, alpha level and splits. The first row in the table indicates the result for the dataset analyzed in the manuscript

Number of observations	Percentage of 'cases'	Number of covariates	Alpha	Number of splits	5% experiment wide error threshold
93	0.27	44	0.05	2	1.745E-06
50	0.25	20	0.01	1	1.321E-04
100	0.25	20	0.01	1	1.272E-04
200	0.25	20	0.01	1	8.903E-05
50	0.5	20	0.01	1	1.425E-04
100	0.5	20	0.01	1	7.889E-05
200	0.5	20	0.01	1	6.714E-05
50	0.25	50	0.01	1	5.676E-05
100	0.25	50	0.01	1	3.944E-05
200	0.25	50	0.01	1	3.354E-05
50	0.5	50	0.01	1	5.157E-05

100	0.5	50	0.01	1	3.653E-05
200	0.5	50	0.01	1	3.025E-05
50	0.25	100	0.01	1	5.064E-05
100	0.25	100	0.01	1	1.973E-05
200	0.25	100	0.01	1	1.333E-05
50	0.5	100	0.01	1	1.797E-05
100	0.5	100	0.01	1	1.584E-05
200	0.5	100	0.01	1	1.095E-05
50	0.25	20	0.05	1	1.838E-04
100	0.25	20	0.05	1	1.272E-04
200	0.25	20	0.05	1	8.060E-05
50	0.5	20	0.05	1	1.425E-04
100	0.5	20	0.05	1	1.089E-04
200	0.5	20	0.05	1	6.714E-05
50	0.25	50	0.05	1	5.676E-05
100	0.25	50	0.05	1	6.007E-05
200	0.25	50	0.05	1	2.915E-05
50	0.5	50	0.05	1	5.157E-05
100	0.5	50	0.05	1	4.585E-05
200	0.5	50	0.05	1	2.476E-05
50	0.25	100	0.05	1	3.119E-05
100	0.25	100	0.05	1	2.403E-05
200	0.25	100	0.05	1	2.735E-05
50	0.5	100	0.05	1	2.710E-05
100	0.5	100	0.05	1	1.910E-05
200	0.5	100	0.05	1	1.581E-05
50	0.25	20	0.01	2	7.108E-05
100	0.25	20	0.01	2	3.981E-05
200	0.25	20	0.01	2	9.983E-06
50	0.5	20	0.01	2	1.259E-04
100	0.5	20	0.01	2	1.357E-05
200	0.5	20	0.01	2	1.282E-05
50	0.25	50	0.01	2	5.676E-05
100	0.25	50	0.01	2	5.723E-06
200	0.25	50	0.01	2	1.284E-06
50	0.5	50	0.01	2	5.157E-05
100	0.5	50	0.01	2	3.233E-06
200	0.5	50	0.01	2	1.383E-06
50	0.25	100	0.01	2	3.119E-05
100	0.25	100	0.01	2	1.676E-06
200	0.25	100	0.01	2	4.753E-07
50	0.5	100	0.01	2	1.797E-05
100	0.5	100	0.01	2	4.851E-07
200	0.5	100	0.01	2	3.181E-07
50	0.25	20	0.05	2	1.374E-05
100	0.25	20	0.05	2	4.405E-06
200	0.25	20	0.05	2	3.422E-06
50	0.5	20	0.05	2	8.345E-06
100	0.5	20	0.05	2	3.139E-06

200	0.5	20	0.05	2	2.262E-06
50	0.25	50	0.05	2	3.019E-06
100	0.25	50	0.05	2	1.104E-06
200	0.25	50	0.05	2	7.479E-07
50	0.5	50	0.05	2	1.337E-06
100	0.5	50	0.05	2	3.871E-07
200	0.5	50	0.05	2	2.384E-07
50	0.25	100	0.05	2	7.422E-07
100	0.25	100	0.05	2	1.796E-07
200	0.25	100	0.05	2	1.467E-07
50	0.5	100	0.05	2	4.355E-07
100	0.5	100	0.05	2	1.034E-07
200	0.5	100	0.05	2	1.270E-07
50	0.25	20	0.01	3	1.229E-04
100	0.25	20	0.01	3	1.099E-05
200	0.25	20	0.01	3	3.298E-06
50	0.5	20	0.01	3	1.259E-04
100	0.5	20	0.01	3	9.955E-06
200	0.5	20	0.01	3	3.070E-06
50	0.25	50	0.01	3	5.676E-05
100	0.25	50	0.01	3	3.338E-06
200	0.25	50	0.01	3	6.830E-07
50	0.5	50	0.01	3	4.828E-05
100	0.5	50	0.01	3	1.939E-06
200	0.5	50	0.01	3	4.971E-07
50	0.25	100	0.01	3	1.506E-05
100	0.25	100	0.01	3	7.059E-07
200	0.25	100	0.01	3	1.307E-07
50	0.5	100	0.01	3	1.504E-05
100	0.5	100	0.01	3	1.489E-07
200	0.5	100	0.01	3	9.257E-08
50	0.25	20	0.05	3	3.805E-06
100	0.25	20	0.05	3	1.698E-06
200	0.25	20	0.05	3	5.017E-07
50	0.5	20	0.05	3	2.186E-06
100	0.5	20	0.05	3	2.087E-07
200	0.5	20	0.05	3	1.404E-07
50	0.25	50	0.05	3	6.965E-07
100	0.25	50	0.05	3	9.309E-08
200	0.25	50	0.05	3	2.502E-08
50	0.5	50	0.05	3	2.210E-07
100	0.5	50	0.05	3	3.725E-08
200	0.5	50	0.05	3	1.457E-08
50	0.25	100	0.05	3	8.472E-08
100	0.25	100	0.05	3	2.410E-08
200	0.25	100	0.05	3	3.982E-09
50	0.5	100	0.05	3	9.685E-08
100	0.5	100	0.05	3	3.474E-09
200	0.5	100	0.05	3	2.419E-09