University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2002 *Paper* 106

Comparative Genomic Hybridization Array Analysis

Annette M. Molinaro* Mark J. van der Laan†
Dan H. Moore‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, annette.molinaro@yale.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley

‡Biomedical Sciences Division, Lawrence Livermore National Laboratory

http://biostats.bepress.com/ucbbiostat/paper106

Comparative Genomic Hybridization Array Analysis

Annette M. Molinaro, Mark J. van der Laan, and Dan H. Moore

Abstract

At the present time, there is increasing evidence that cancer may be regulated by the number of copies of genes in tumor cells. Through microarray technology it is now possible to measure the number of copies of thousands of genes and gene segments in samples of chromosomal DNA. Microarray comparative genomic hybridization (array CGH) provides the opportunity to both measure DNA sequence copy number gains and losses and map these aberrations to the genomic sequence. Gains can signify the over-expression of oncogenes, genes which stimulate cell growth and have become hyperactive, while losses can signify under-expression of tumor suppressor genes, genes whose activity stops the formation of tumors. In order to better understand the progression of cancer and the differences between cancer and non-cancer tissue it is of great importance to fully understand what is happening at the chromosomal level. In the hopes of finding a genetic signature for subtypes of cancer, it is our intention to explore statistical approaches to array CGH data. The Waldman Lab at UCSF-CCC graciously allowed us to access data from their renal cancer study. This project was designed to determine whether microarray information on copy number of genes could be used to discriminate among four subtypes of renal cancer.

1 Introduction

At the present time, there is increasing evidence that cancer may be regulated by the number of copies of genes in tumor cells. Through microarray technology it is now possible to measure the number of copies of thousands of genes and gene segments in samples of chromosomal DNA. Microarray comparative genomic hybridization (array CGH), as described in Section 3.1, provides the opportunity to both measure DNA sequence copy number gains and losses and map these aberrations to the genomic sequence. Gains can signify the over-expression of *oncogenes*, genes which stimulate cell growth and have become hyperactive, while losses can signify under-expression of *tumor suppressor genes*, genes whose activity stops the formation of tumors. In order to better understand the progression of cancer and the differences between cancer and non-cancer tissue it is of great importance to fully understand what is happening at the chromosomal level.

In the hopes of finding a genetic signature for subtypes of cancer, it is our intention to explore statistical approaches to array CGH data. The Waldman Lab at UCSF-CCC graciously allowed us access data to their renal cancer study. This project was designed to determine whether microarray information on copy number of genes could be used to discriminate among four subtypes of renal cancer.

2 Question of Interest

A study utilizing CGH technology results in thousands of covariates (measured by bacterial artificial chromosomes (BACs)) and a relatively small number of observations, i.e., large p and small n. Given this data our goal is to find a way of accurately classifying observations into predefined subtypes of cancer. Thus, there are two questions: "What is the best subset of BACs with which to build a classification rule?" and "Given a subset of BACs, what is the best classification rule one can build?"

The purpose of this manuscript is twofold: First, we examine several approaches for subsetting the given set of BACs and subsequently we explore a clustering technique to classify the subtypes of cancer based on the previously defined subset of BACs. Lastly, we will assess the accuracy of this classification.

Section 3.1 provides a general description of CGH data and technology. In

Section 3.2 the renal data as provided by the Waldman Lab is described. In Section 4 a sampling of subset rules for decreasing the number of covariates, i.e., BACs, are presented. One of the subset rules is Significance Tests, i.e., to choose the BACs which are significant as defined by a parametric or non-parametric test. This choice leads to the question of how to deal with the multiple testing issue. Four possible approaches to multiple testing are addressed in Section 4.2. In Section 5 a clustering algorithm is explained. This algorithm is used to classify the tumors based on subsets of BACs as chosen by the aforementioned subset rules. The classifications are assessed by how many misclassifications are made. A final discussion of the subset and classification rules is in Section 6.

3 Data

3.1 Comparative Genomic Hybridization

Prior to entering mitosis and undergoing division, a cell must exactly replicate its genome by synthesizing a new copy of each chromosome, using the existing DNA as a template (Figure 1). During this process several types of large-scale chromosomal alterations can occur. Regions of DNA can be deleted or fail to be replicated, resulting in a *loss* at that chromosomal locus. Conversely, regions can be duplicated or multiplied, resulting in a *gain* of copy number or amplification. Furthermore, entire segments of chromosomes can be inappropriately fused with other chromosomes in a process called translocation. These mutations are illustrated in Figure 2.

Healthy cells maintain checkpoints to monitor and correct this genomic instability. Mutations can be repaired, or the cell can undergo programmed cell death (apoptosis) to prevent the mutations from being passed to daughter cells. However, if the mutation is not detected, or if a checkpoint has been inactivated by mutation, the cell can survive in its altered state. A particular danger exists if genomic alterations predispose a cell to uncontrolled proliferation. If an *oncogene*, a gene which stimulates cell growth and has become hyperactive, is housed in an area which has been amplified, that oncogene may be over-expressed. On the other hand, the loss of a region that houses a tumor suppressor gene, a gene whose activity opposes uncontrolled cell growth (e.g., $p53$), will result in the inactivation of that gene. Such regional alterations in copy number are a characteristic of solid tumors.

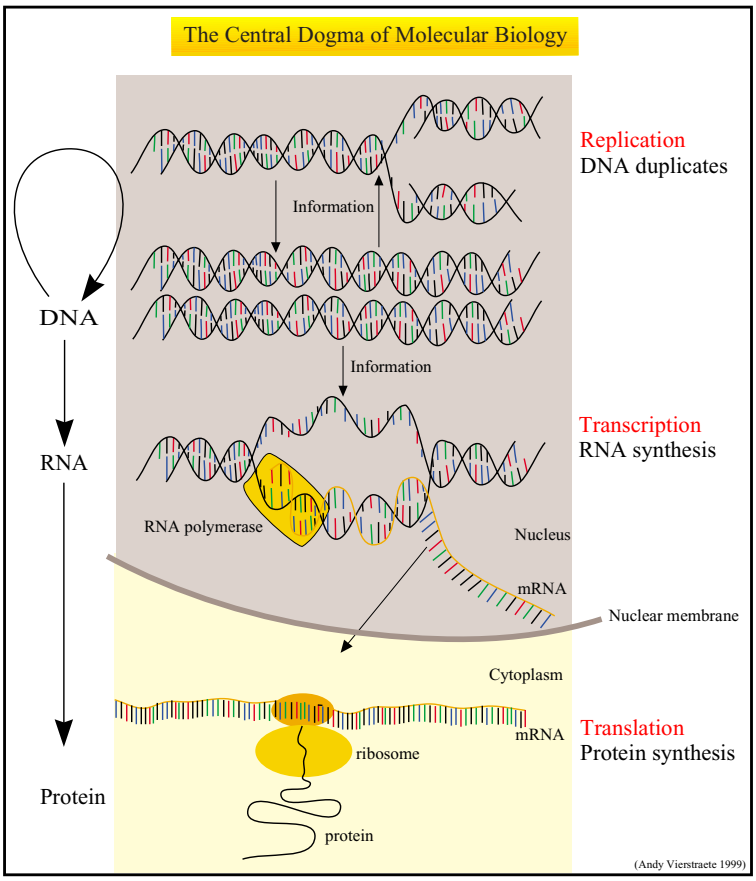

Figure 1: *The Central Dogma of Molecular Biology.* During the cell cycle DNA duplicates (Replication), RNA synthesizes (Transcription), and protein synthesizes (Translation).

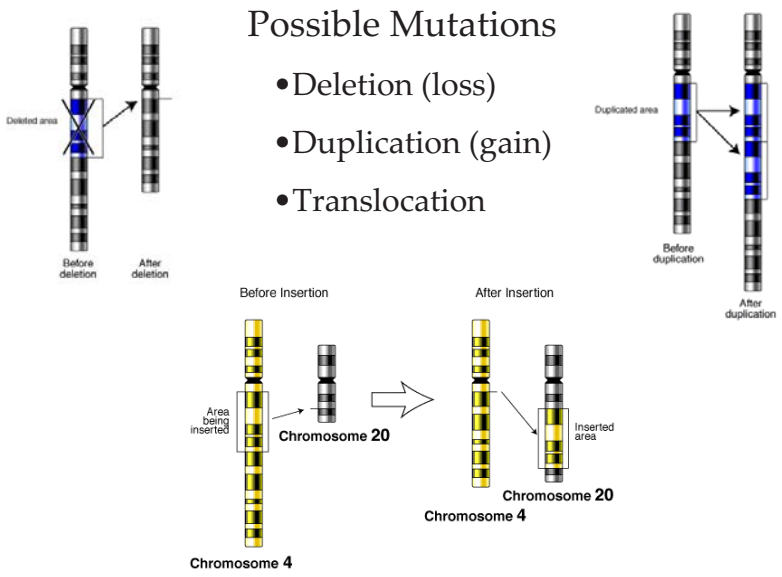

Figure 2: *Possible Mutations that can occur during the cell cycle.* There are three chromosomal abberations of interest: *deletion* resulting in a loss of copy number for a chromosomal region, *duplication* resulting in a gain in copy number, and *translocation* involving both gains and losses for different chromosomal regions.

It is hypothesized that this genomic instability is key to allowing cancerous growth to progress. Thus, there is an inherent need to track and understand these mutations.

Comparative Genomic Hybridization (CGH) was developed as a method for detecting and mapping such alterations in the genome. The basic procedure of CGH begins with purifying genomic DNA from samples of cancerous and normal control tissue (e.g., lymphocytes from a healthy individual). These two DNA preparations are labeled with different fluorochromes which will emit light at easily distinguishable wavelengths–generally red and green. The samples are then mixed and allowed to competitively hybridize to immobilized normal DNA. In regions where there are no amplifications or deletions in the cancer genome, binding of both samples will be equal, and the equal emission of light from both fluorochromes will result in a perceived yellow

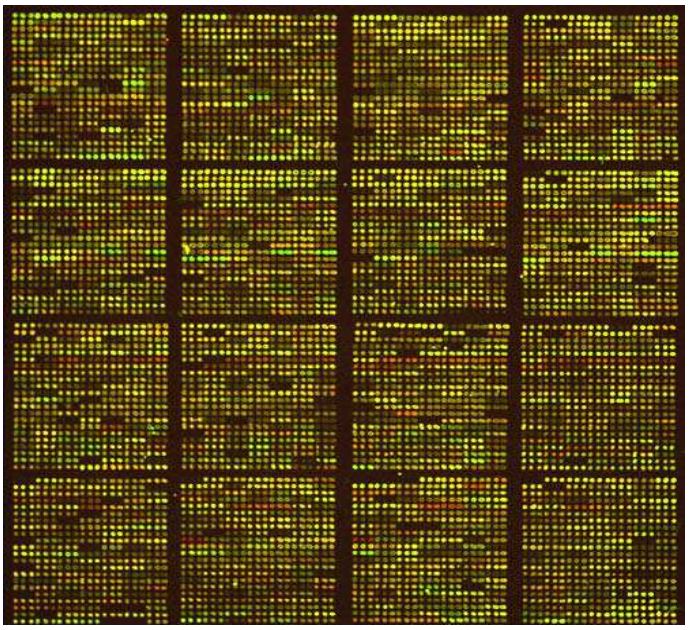

Figure 3: *Microarray-based CGH (array CGH).* In this procedure, probes that map to evenly spaced loci along the entire length of the genome are printed onto glass slides and used as targets for the hybridization of fluorescent DNA.

fluorescence. However, where there are losses in copy number in the cancer DNA, the color with which the normal DNA was labeled (e.g., red) will predominate. Similarly, in regions of DNA copy number gain, the color with which the tumor DNA was labeled (e.g., green) will be apparent. These variations can be seen in Figure 3.

Initially genomic instability was measured with chromosomal-CGH, in which intact chromosomes were used as hybridization targets to map gains and losses of DNA copy number (Figure 4). The resolution of this assay was relatively low, allowing for the detection only of comparatively large gains or losses, and with little information about the locations of the ends of the altered regions. However, recent improvements in the resolution and sensitivity of CGH have been achieved with microarray-based CGH (array CGH). In this procedure, probes that map to evenly spaced loci along the entire length of the genome are printed onto glass slides and used as targets for the hybridization of fluorescent DNA. Array CGH has greater resolution

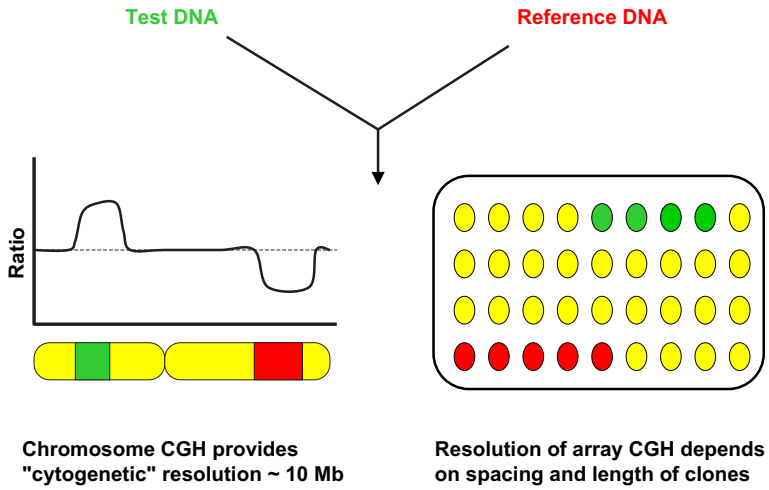

Figure 4: *Hybridization targets for CGH analysis.* Initially measured with chromosomal-CGH now array CGH offers greater resolution and more precise measurements.

than chromosomal CGH, allowing for the detection of smaller amplifications and deletions, and more precise measurements of how these regions are delimited. Additionally, array CGH allows a genome-wide analysis of DNA sequence copy number in a single experiment (Pinkel et al. (1998) Snijders et al. (2001)). Coupled with a physical map of the genome, these data allow more efficient identification of genes that may be involved in cancer progression.

Several types of clones can be used to make the arrays. The UCSF-CCC primarily uses bacterial artificial chromosomes (BACs). While the human haploid genome has 3 billion pairs of bases, each BAC consists of a smaller 100-200 kilobase (kb) region of the DNA. These BAC clones are grown in bacteria, purified, and spotted onto slides by a robot.

In carcinogenesis, researchers compare tumor samples to normal samples in order to examine hypothesized cancer related aberrations of the genome. By comparing test (tumor) to reference (normal) samples, they can identify regions which differ (by loss or gain of copy number) and elucidate the

particular genes housed in those loci.

Method The CGH procedure consists of labeling genomic material from a test sample, i.e. tumor, and a reference sample, i.e. lymphocytes from healthy persons, with different fluorochromes. These fluorochromes are completely distinguishable with no spectral overlap. The two samples are then hybridized to an array containing clones which are designed to cover certain areas of the genome or the entire genome.

Once the area of the genome is decided, these BACs are placed on the arrays to measure the test to reference ratio. The BACs are duplicated thousands of times and then fragmented. Several hundred thousand kb are deposited in each spot of an array, enough DNA to represent the BAC several thousand times. Once the test and reference samples are hybridized to the BAC arrays, it is possible to quantify the signal intensities of the fluorochromes.

1. The background signal intensity is calculated and each fluorochrome is adjusted in each pixel in each spot.

2. The fluorescence ratios on each spot are calculated (e.g. the ratio of means of all pixels in a spot)

3. The ratios on the replicate spots are averaged

4. Ratios of test to reference samples are standardized by dividing by the median of all BACs in the sample

5. The X chromosome can be an internal control

Once these ratios are calculated, missing data is imputed and a log transformation is taken. The data is then examined to find pertinent gains (increases in copy number) and losses (decreases in copy number). For example, Figure 5 displays the results from an array CGH experiment. The x-axis is the chromosomes ordered from 1 to 22 and then X, the sex chromosome. The y-axis is the log2 ratio. From this graph an increase in copy number is apparent for all of chromosomes 2 and 21.

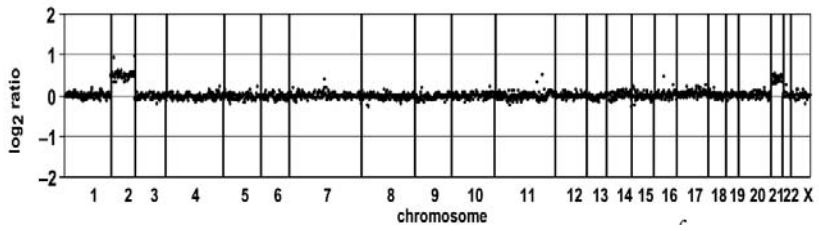

Figure 5: *Plot of array CGH analysis.* The chromosomes are ordered along the x-axis and the $log2$ ratio on the y-axis.

3.2 Renal Cancer Study

The data set provided by the Waldman Lab at UCSF-CCC consists of control tissue (N) and four sub-types of renal neoplasms: Chromophobe (CH), Conventional (CO), Papillary (PA), and Oncocytoma (ON, benign). Each of the respective tumors was measured over 91 BACs representing various areas of each chromosome. There are 5 tumors representing CH, 16 for CO, 13 for PA, 6 for ON, and 2 controls (N) for a total of 42 samples. For this analysis, the oncocytoma sub-type was combined with the control tissue unless otherwise stated.

As described in Veltman et al. (2002), the test sample was made with genomic DNA from 42 renal tumor and normal samples . The 42 samples were derived from frozen tissue or short-term cell cultures and previously histologically characterized. The reference sample was made from normal DNA isolated from lymphocytes of healthy persons.

Methods were similar to those outlined in Pinkel et al. (1998). The array consisted of 93 clones (6 cosmids, 20 BACs and 67 P1 clones). The selected clones represent all 22 autosomal chromosomes and the X-chromosome. The majority of the other targets chosen represent those chromosomes most frequently altered in renal cancer. There was an average coefficient of variation of 3% for the test to reference (T/R) intensity ratio for the quadruplicate spots of each target.

For each set of arrays, six to eight normal vs. normal hybridizations were performed in order to define the normal variation in T/R ratio for each of the target clones. The average T/R ratio of the quadruplicate of each clone was calculated and divided by the median T/R ratio of all targets present on the array in order to normalize the mean T/R value to 1.0. A slight clone to clone variability in the intensity ratios was observed which proved to be reproducible in the normal vs. normal hybridizations. The variability was corrected for by dividing each T/R ratio by the mean T/R ratio of the normal vs. normal hybridizations. The normal range for the T/R ratio for each target was calculated as 2 times the standard deviation (taken from the normal vs. normal hybridizations) from the mean of one. Figure 6 shows the mean gain/loss for each subtype over all of the BACs. In this figure fluorescent green represents CO, dark green represents PA, red represents CH, and blue represents ON.

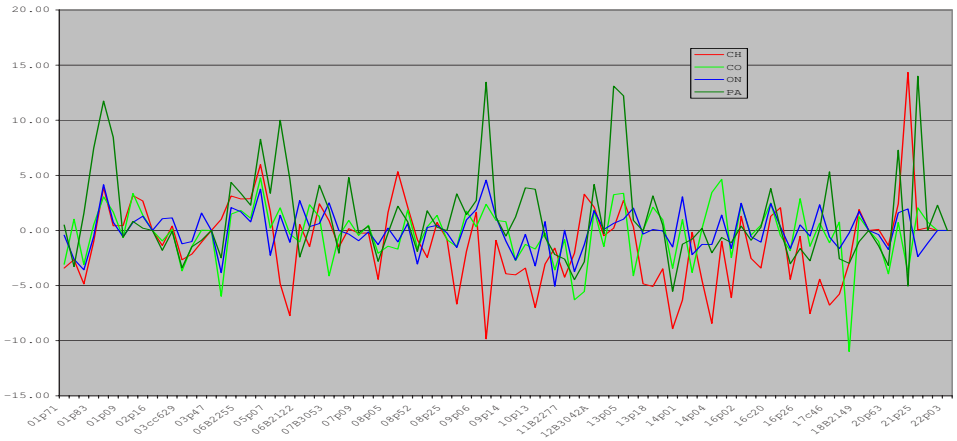

Figure 6: Mean Gain/Loss for each renal cancer subtype.

4 Subsets and Subset Rules

As mentioned in Section 2, our first goal is to identify the best, or *target*, subset of BACs. This target subset is defined as the set of BACs which most accurately predict the cancer classification. In addition to fine-tuning the group of BACs for analysis, this decrease in the size of p (the number of BACs) will help to alleviate computational and multiple testing burdens. If we let X be a p-dimensional vector of T/R ratios, then we observe n i.i.d. tumor samples $X_1, X_2, \ldots, X_n$. That is, for each of the n tumor samples we have p BACs measured. For any missing measurements, we impute the data with the mean of the BAC within each subtype. Since a k-fold gain is the opposite of a k-fold loss, we use the log transformation $Y_j = log(X_j)$, where $j = 1, \ldots, p$. At this point we can denote the expectation, covariance and correlation of Y by μ, Σ, and ρ, respectively. Further, the data is denoted as (μ, Σ), the target subset as S, and the mapping which produces S as a *subset rule* (van der Laan and Bryan, 2001).

Given a data set denoted by (μ, Σ), there are numerous mappings, or *subset rules*, which result in target subsets, S. In the following sections we

shall explore three such subsetting rules: arbitrary cut-offs; significance tests; and variable importance as measured by regression trees. We will compare these three rules to the subset of BACs which itself includes all of the BACs, referred to as *All BACs*.

4.1 Arbitrary Cut-Offs

One approach to selecting subsets of BACs is to decide on a descriptive measure of the data (e.g., mean) and a cut-off value, M, where $M \in [0, 1]$. Then one can choose $M * 100\%$ of the descriptive measure. For this data set, we chose $M = 0.5$ and looked at the variances and the absolute value of the means. After ranking both the variances and the means from largest to smallest, we chose the top 50%. For the means, this results in the 50% of the BACs with the greatest gains and losses and for the variances, the 50% of the BACs with the largest spread.

4.2 Significance Testing

A second approach to finding a subset of BACs is to identify the BACs which are significantly different from each other between subtypes of renal cancer. Significantly different BACs can be found by comparing measures of location (e.g., the mean abberation of CO vs. that of CH) or distributions of the subtypes of renal cancer.

There are numerous significance tests which can be administered for differentiating the significant BACs from the non-significant. These tests can be parametric or non-parametric. As such, we selected a sampling of both. In the following sections the t-test (with equal and non-equal variance assumptions), Wilcoxon Rank Sum Test, Hollander Test of Extreme Reactions, and Kruskal-Wallis are described and implemented.

Given any of these tests and a large number of BACs we are conducting numerous univariate analysis. If a significance level unadjusted for multiple comparisons is used, numerous BACs which do not have copy number aberrations may be identified. These BACs are referred to as "false positives". In order to accommodate for a multiple comparisons issue three different methods are examined: the Bonferroni adjustment, the non-parametric bootstrap, and a permutation test. These three are described in the following paragraphs. For each of the significance test, adjusting for multiple comparisons

via these methods is compared to *no adjustment*, i.e., a significance level $\alpha = 0.05$.

A conventional method for dealing with the multiple testing issue is to use the *Bonferroni adjustment*. This method entails dividing the significance level α by the number of tests k, e.g., if $\alpha = 0.05$ and 90 tests are evaluated the Bonferroni significance level is $\alpha/k = .05/90 = .0005$. This adjustment tends toward conservative if the individual tests are independent, i.e., not correlated. Thus, if the tests are correlated the chance of a Type II error, accepting a false null hypothesis, increases. Given the nature of the CGH data, it seems unwise to make the assumption of independence as several of the BACs are located on the same arm of a chromosome increasing the correlation.

An alternative, the *non-parametric bootstrap*, is to construct a null distribution with means equal to zero and the observed covariance structure. First, the data is centered by subtracting the subtype specific mean and then this centered-data is used to generate a large number of bootstrap samples. Here the tumors are sampled. From this null distribution, a cut-off value is assessed such that no more than $1 - \frac{\alpha}{2}$ of the samples have any BACs with copy number aberrations. This method is less conservative than the Bonferroni adjustment, yet, statistically accommodates for multiple comparisons.

Another alternative, a *permutation test*, is to determine the distribution of a test statistic under the null hypothesis that there is no association between a BAC and tumor subtype. In this method, the subtype identifier, e.g., CO, for each tumor is permuted in order to destroy any relationship between BAC and subtype. For each permutation, the test statistic is repeated. Once a large number of test statistics are generated, the cut-off value is assessed by examining the 95% quantile of the test statistics. The advantage of this method, as well as the non-parametric bootstrap, is that it allows the observed covariance structure to remain intact while making no distribution assumptions.

In order to find potential target subsets, we implemented the t-test (with equal and non-equal variance assumptions), Wilcoxon Rank Sum Test, Hollander Test of Extreme Reactions, and Kruskal-Wallis. Each of these tests was used to compare the six combinations of renal subtypes, i.e., CO vs. N/ON, CH vs. N/ON, P vs. N/ON, CO vs. CH, CO vs. P, and CH vs. P. To address the multiple testing issue we chose from the aforementioned approaches based on the inherent nature or limitation of each test. In addition to using one or two of these approaches, we also looked at simply ignoring

the multiple comparison issue. For each significance test, a BAC qualified for the target subset if and only if it was univariately significant in one of the six combinations.

1. **t-test**

   The t-test is a procedure to test the null hypothesis of an equal population location parameter, μ. Under the following assumptions the t statistic has a known and tabulated distribution, the t distribution:

   - The data consist of a random sample of n observations $x_1, x_2, \ldots, x_n$ from one population and m observations $y_1, y_2, \ldots, y_m$ from another population.
   - The two samples are independent.
   - The measurement scale employed is at least ordinal.

   This implies:

   - The observed variable is a continuous random variable.
   - The measurement scale employed is at least ordinal.
   - The distribution functions of the two populations are normal and differ only with respect to location.

   There are two ways to estimate the denominator of the t-test based on the equality of population variances.

   Equal Variances

   For the assumption of equal variances, the statistical package, `R` (Ihaka and Gentlemen (1996)), uses the following formula to calculate the t-test statistic for two samples $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_m)$:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n-1)\sigma_x + (m-1)\sigma_y}{n+m-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

   where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$, $\sigma_x = \frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}$, and $\sigma_y = \frac{1}{m}\sum_{i=1}^{m} y_i - \bar{y}$.

   This t-test statistic has $n + m - 2$ degrees of freedom.

   In the following, the three methods for approaching the multiple comparisons issue for the t-test with equal variances are compared.

Table 1: *t-test with equal variances: No adjustment for MC.*

Sub-type	Test	Cut-Off	*df*	Sig BACs
1	CO vs. N/ON	2.074	22	11
2	CH vs. N/ON	2.201	11	29
3	P vs. N/ON	2.093	19	31
4	CO vs. CH	2.093	19	32
5	CO vs. P	2.052	27	40
6	CH vs. P	2.119	16	40

Table 2: *t-test with equal variances: Bonferroni adjustment for MC.*

Sub-type	Test	Cut-Off	*df*	Sig BACs
1	CO vs. N/ON	4.077	22	2
2	CH vs. N/ON	4.863	11	9
3	P vs. N/ON	4.187	19	10
4	CO vs. CH	4.187	19	14
5	CO vs. P	3.954	27	19
6	CH vs. P	4.346	16	18

- **No adjustment** For the empirical data, the absolute value of the t-tests with the assumption of equal variances at $\alpha = 0.05$ (equivalent to a two-sided cut-off for the observed t-statistic with $\alpha/2$), the *degrees of freedom*, and the number of significant BACs are shown in Table 1. There are **62** BACs which are significant in one or more of these subtypes.
- **Bonferroni Adjustment** The Bonferroni adjustment with the absolute value of the t-tests, given the assumption of equal variances with $\alpha = 0.05$ for the global, and $\alpha_{bonf} = \alpha/(2 * p) = 0.05/176 = 0.0003$ for the marginal level provides the results shown in Table 2. There are **32** BACs which are significant in one or more of these subtypes.
- **Non-parametric Bootstrap** To form a null distribution, the subtype-specific mean was subtracted from each of the BACs. The tumor samples in this null distribution were bootstrapped 1,000 times. Each time the t-test statistics were recalculated. The cut-

Table 3: *t-test with equal variances: Non-parametric Bootstrap adjustment for MC.*

Sub-type	Test	Sig/1000	Cut-Off	Sig BACs
1	CO vs. N/ON	.049	5.84	1
2	CH vs. N/ON	.05	16.72	0
3	P vs. N/ON	.05	8.15	3
4	CO vs. CH	.05	8.89	1
5	CO vs. P	.05	5.97	8
6	CH vs. P	.047	19.02	0

offs for each group which corresponded to having 950 of the 1,000 samples with no significant t-test statistics are shown in Table 3. There are **8** BACs which are significant in one or more of these subtypes.

Unequal Variances

For the assumption of unequal variances, `R` uses the following formula to calculate the t-test statistic for two samples $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_m)$:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x}{n} + \frac{\sigma_y}{m}}}$$

For this t-test statistic we used degrees of freedom equivalent to the smaller of the two sample sizes minus one (i.e., $min(n, m) - 1$). `R` calculates the degrees of freedom using the Welch modification; however, this changes for each test and is not convenient for the purposes of this exercise. By using the $min(n, m) - 1$, we are allowing a possibly more conservative cut-off than that calculated by `R`.

- **No adjustment** For the empirical data, the absolute value of the t-tests with the assumption of unequal variances at $\alpha = 0.05$ provides the cut-offs shown in Table 4. There are **54** BACs which are significant in one or more of these subtypes.

Table 4: *t-test with unequal variances: No adjustment for MC.*

Sub-type	Test	Test Statistic	*df*	Sig BACs
1	CO vs. N/ON	2.365	7	7
2	CH vs. N/ON	2.776	4	24
3	P vs. N/ON	2.365	7	26
4	CO vs. CH	2.776	4	26
5	CO vs. P	2.1179	12	36
6	CH vs. P	2.776	4	31

Table 5: *t-test with unequal variances: Bonferroni adjustment for MC.*

Sub-type	Test	Test Statistic	*df*	Sig BACs
1	CO vs. N/ON	6.082	7	1
2	CH vs. N/ON	10.307	4	0
3	P vs. N/ON	6.082	7	6
4	CO vs. CH	10.307	4	1
5	CO vs. P	4.716	12	12
6	CH vs. P	10.307	4	4

- **Bonferroni Adjustment** For the empirical data, the Bonferroni adjustment with the absolute value of the t-tests, the assumption of unequal variances at $\alpha = 0.05$ for the global, and $\alpha_{bonf} = 0.05/(2 * 88) = 0.0003$ for the marginal level, provides the cut-offs in Table 5. There are **13** BACs which are significant in one or more of these subtypes.
- **Non-parametric Bootstrap** To form a null distribution, the subtype-specific mean was subtracted from each of the BACs. The tumor samples from this null distribution were bootstrapped 1,000 times. Each time the t-test statistics were recalculated. The cut-offs for each group which corresponded to having 950 of the 1,000 samples with no significant t-test statistics are shown in Table 6. There are **8** BACs which are significant in one or more of these subtypes.

Table 6: *t-test with unequal variances: Non-parametric Bootstrap adjustment for MC.*

Sub-type	Test	Sig/1000	Cut-Off	Sig BACs
1	CO vs. N/ON	.05	6.24	1
2	CH vs. N/ON	.048	19.75	0
3	P vs. N/ON	.049	8.48	3
4	CO vs. CH	.05	10.58	1
5	CO vs. P	.049	6.1	8
6	CH vs. P	.049	22.35	0

Comments on the *t*-test Results In theory, the bootstrap of the null distribution should be more conservative than no adjustment at all and less conservative than the Bonferroni adjustment. However, as seen in the *t*-test results with the assumption of equal variances and unequal variances, this is not the case. In both of these scenarios the number of BACs selected with the non-parametric bootstrap is less that the number selected with the Bonferroni adjustment. Possible reasons are:

(a) *t*-test makes the assumption that the empirical means of the two samples are normally distributed, which is violated in this data.

(b) All of the methods give a common cut-off value. Due to the nature of this data, it might be possible that individual cut-offs are required.

Potential remedies:

(a) Use non-parametric tests with no assumptions on the empirical means, e.g. Median test, Wilcoxon Rank-Sum, Tukey's Quick Test, or Hollander's extreme.

(b) Investigate different methods for getting cut-off values, e.g., permutations tests.

(c) Examine tests which are based on maximum gain and loss which may better fit this data. For instance, instead of testing the difference between means, we could test the difference between the minimums and/or maximums.

2. **Wilcoxon Rank Sum Test** The Wilcoxon Rank Sum Test is a procedure for testing the null hypothesis of equal population location parameters. This test is the non-parametric alternative to the t-test. The advantage it has is not assuming that the empirical means for each sample are normally distributed. However, it does assume that the distribution functions for both samples are the same except for under the null hypothesis. The assumptions made by this procedure are:

   - The data consist of a random sample of n observations $x_1, x_2, \ldots, x_n$ from one population and m observations $y_1, y_2, \ldots, y_m$ from another population.
   - The two samples are independent.
   - The observed variable is a continuous random variable.
   - The measurement scale employed is at least ordinal.
   - The distribution functions of the two populations differ only with respect to location, if they differ at all.

   The hypotheses can either be set up as one or two sided. That is:

   H_0: The populations have identical distributions
   vs.
   H_1: The populations differ with respect to location
   for the two-sided alternative hypothesis
   or
   H_1: The X's tend to be smaller/larger than the Y's.
   for the one-sided alternative hypothesis

   The test statistic is computed by combining the two samples and ranking all sample observations from smallest to largest. Tied observations are given the value of the mean of the rank positions had there been no ties. The sum of the ranks for the X's, is calculated and denoted by R. The Wilcoxon Rank Sum Statistic is:

$$W = R - \frac{n(n-1)}{2}$$

The decision rule is based on the alternative hypothesis. If it is a two-sided test then H_0 is rejected for either a sufficiently small or large value of W at the α level of significance. That is if:

$$w_{1-\frac{\alpha}{2}} < T < w_{\frac{\alpha}{2}}$$

Here $w_{1-\frac{\alpha}{2}} = nm - w_{\frac{\alpha}{2}}$ where $w_{\frac{\alpha}{2}}$ is the critical value of W as given by the Wilcoxon Distribution.

The decision rule for the one-sided H_1 is based on either a sufficiently small or large value of W according to the sign of the alternative hypothesis.

Initially we implemented the two-sided hypothesis. However, as the α level decreases when multiple comparisons are adjusted for multiple test statistics drop to 0 for the left side of the test. This was most evident with the non-parametric bootstrap adjustment. With this adjustment more than 600 samples have significant test statistics up to 0, and then at 0 there are no significant samples. Thus, we replaced the two-sided hypothesis with the one-sided. In this hypothesis the X's denote the cancer sub-types (e.g., X_{CO}, X_{CH}, X_P) and the Y's represent the control-benign observations (Y_N) or the opposing subtype (X_{CH}, X_P) .

The alternative hypotheses are:

(a) H_1: The X_{CO}'s tend to be smaller than the Y_N's

(b) H_1: The X_{CH}'s tend to be larger than the Y_N's

(c) H_1: The X_P's tend to be smaller than the Y_N's

(d) H_1: The X_{CO}'s tend to be smaller than the Y_{CH}'s

(e) H_1: The X_{CO}'s tend to be smaller than the Y_P's

(f) H_1: The X_{CH}'s tend to be larger than the Y_P's

These alternative hypotheses are based on plots of the data where each subtype is compared to the opposing subtype. A difficulty with this approach is that the alternative hypotheses may fit a proportion of the BACs but not all. Both gains *and* losses are expected in comparison to the controls and possibly to the other subtypes.

For the Wilcoxon Rank Sum Test we looked at no adjustment, the Bonferroni adjustment and the non-parametric bootstrap adjustment for multiple comparisons.

Table 7: *Wilcoxon Rank Sum Test: No adjustment for MC.*

Subtype	Test	Cut-Off	Sig BACs
1	CO vs. N/ON	37	13
2	CH vs. N/ON	31	9
3	P vs. N/ON	29	21
4	CO vs. CH	20	13
5	CO vs. P	66	32
6	CH vs. P	49	13

Table 8: *Wilcoxon Rank Sum Test: Bonferroni adjustment for MC.*

Subtype	Test	Cut-Off	Sig BACs
1	CO vs. N/ON	14	2
2	CH vs. N/ON	40	0
3	P vs. N/ON	10	4
4	CO vs. CH	4	1
5	CO vs. P	33	18
6	CH vs. P	62	2

- **No adjustment** For the empirical data, the W statistic at $\alpha = 0.05$ provides the cut-offs shown in table 7. There are **59** BACs which are significant in one or more of these subtypes.
- **Bonferroni adjustment** For the empirical data, the Bonferroni adjustment with the W statistic at $\alpha = 0.05$ for the global and $\alpha_{bonf} = 0.05/88 = 0.0005$ for the marginal level provides the cut-offs shown in table 8. There are **24** BACs which are significant in one or more of these subtypes.
- **Non-parametric Bootstrap adjustment** The results of bootstrapping the null distribution are shown in table 9. There are **7** BACs which are significant in one or more of the subtypes.

3. **Hollander Test of Extreme Reactions** Thus far, we have looked at tests of location which do not show much of a significant difference between the average responses of subtypes and controls. Hollander proposed the Test of Extreme Reactions to detect differences

Table 9: *Wilcoxon Rank Sum Test: Non-parametric Bootstrap adjustment for MC.*

Subtype	Test	Sig/1000	Cut-Off	Sig BACs
1	CO vs. N/ON	.46	6	1
2	CH vs. N/ON	0	40	0
3	P vs. N/ON	0	0	0
4	CO vs. CH	0	0	0
5	CO vs. P	.45	8	7
6	CH vs. P	0	65	0

between control and experimental subjects when some of the latter are expected to react in one way and the others in the opposite way (Hollander, 1963).

The assumptions for this test are as follows:

- The data consist of two independent random samples $X_1, X_2, \ldots, X_n$ and
  $Y_1, Y_2, \ldots, Y_m$ of control and experimental subjects, respectively.
- The variable of interest is continuous.
- The strength of the measurement is at least ordinal.

The hypotheses:

H_0: The two samples may be considered as having been drawn from the same population.

vs.

H_1: One population consists of observations resulting from extreme reactions in both directions.

The test statistic, G, is computed by combining the observations from both samples and ordering them from smallest to largest, while keeping track of which are X's and which are Y's. Then:

$$G = \sum_{i=1}^{n} (r_i - \bar{r})^2,$$

Table 10: *Hollander Test of Extremes: No adjustment for MC.*

Subtype	Test	Cut-Off	Sig BACs
1	CO vs. N/ON	129.9	6
2	CH vs. N/ON	63.88	27
3	P vs. N/ON	129.9	31
4	CO vs. CH	44.8	21
5	CO vs. P	267.7	13
6	CH vs. P	224.9	31

where r_i is the rank of the ith largest X value and $\bar{r}$ is the mean of the ranks assigned to the n X values (i.e., $\bar{r} = \frac{\sum_{i=1}^{n} r_i}{n}$).

If the reactions of the experimental subjects are extreme, the responses of the control subjects tend to be compressed with respect to their ranks, and G is relatively small.

The null hypothesis is rejected if the computed value of G is less than or equal to C_α given by the tabled values for this test. Because the published tabled values are only for $\alpha = 0.05$ and $\alpha = 0.01$ we cannot estimate the Bonferroni adjustment. The cut-off values based on these two α (No adjustment and Limited Adjustment) as well as the permutation test are reported.

- **No Adjustment** For the empirical data, the G statistic at $\alpha = 0.05$ provides the cut-offs shown in table 10. There are **61** BACs which are significant in one or more of these subtypes.
- **Limited Adjustment at** $\alpha = 0.01$ For the empirical data, the G statistic at $\alpha = 0.01$ provides cut-offs as shown in table 11. There are **41** BACs which are significant in one or more of these subtypes.
- **Permutation Test** The permutation test of the G statistic provides the cut-offs shown in table 12. There are **56** BACs which are significant in one or more of these subtypes.

4. Kruskal-Wallis one-way analysis of variance by ranks

An alternative to six individual tests is to test the null hypothesis that

Table 11: *Hollander Test of Extremes: Limited Adjustment at* $\alpha = 0.01$ *for MC.*

Subtype	Test	Cut-Off	Sig BACs
1	CO vs. N/ON	91.88	3
2	CH vs. N/ON	49.88	16
3	P vs. N/ON	91.88	17
4	CO vs. CH	23.2	12
5	CO vs. P	222.9	10
6	CH vs. P	190.9	14

Table 12: *Hollander Test of Extremes: Permutation Test Adjustment for MC.*

Subtype	Test	Cut-Off	Sig BACs
1	CO vs. N/ON	102.9	3
2	CH vs. N/ON	25.88	0
3	P vs. N/ON	76.4	15
4	CO vs. CH	47.2	22
5	CO vs. P	552.9	37
6	CH vs. P	242.9	35

several samples have been drawn from the same or identical samples. This test is the *Kruskal-Wallis one-way analysis of variance by ranks.*

The assumptions for this test are as follows:

- The data consist of k samples of sizes $n_1, n_2, \ldots, n_k$.
- The observations are independent both within and among samples.
- The variable of interest is continuous.
- The strength of the measurement is at least ordinal.
- The populations are identical except for a possible difference in location for at least one population.

The hypotheses:

H_0: The k population distribution functions are identical.

vs.

H_1: The k populations do not all have the same median.

The test statistic, H, is computed by combining the observations from all k samples and ordering them from smallest to largest, while keeping track from which sample the observations originate. Then:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} \Big[R_i - \frac{n_i(N+1)}{2} \Big]^2$$

where R_i is the sum of the ranks assigned to observations in the ith sample, and $n_i(N+1)/2$ is the expected sum of ranks. If the null hypothesis is true, we expect the k sum of ranks to be about equal when adjusted for unequal sample sizes. This test statistic is a weighted sum of squares of deviations of sums of ranks from the expected sum of ranks, using reciprocals of sample sizes as the weights.

Kruskal (1952) showed that for large n_i ($n_i > 5$)and k, H is distributed approximately as chi-square with $k-1$ degrees of freedom.

Given the nature of the test a permutation test approach to multiple comparisons is compared to no adjustment.

Table 13: *Kruskal-Wallis one-way ANOVA by ranks: Permutation Test adjustment and No adjustment for MC.*

Test	α	Cut-Off	Sig BACs
No Adjustment	0.05	7.814	53
Permutation Test	NA	14.85	26

Table 14: *Number of BACs selected by each significance test and MC procedure.*

Test	T-test Equal Var	T-test Non-equal Var	Wilcoxon Rank-Sum	Hollander Test	Kruskal Wallis
No Adj	62	54	59	61	52
Bonferroni Adj	32	13	24	41	NA
Bootstrap/Perm	8	8	7	56	26

- **Permutation Test and No Adjustment** For the empirical data, the H statistic at $\alpha = 0.05$ with the chi-square distribution and the H statistic with the permutation test adjustment and no adjustment provide the cut-offs shown in table 13.

4.2.1 Results of Significance Testing

In Table 14 the number of significant BACs selected by each test and multiple comparisons procedure is displayed. This table illustrates the fluctuations in number of BACs selected relative to which adjustment is made for multiple comparisons. Interestingly, Hollander's Test of Extreme Reactions appears to be the least conservative.

4.3 Variable Importance as Measured by Regression Trees

A third approach to finding subsets is to evaluate variable, i.e., BAC, importance as measured by a regression tree algorithm. Possible algorithms include Classification and Regression Trees (CART) (Breiman et al., 1984)

and *Random Forests* (Breiman, 2001). For purposes of this project we chose to look at *Random Forests*.

Random Forest constructs trees by first obtaining a root node which is a bootstrap sample of the original data. The user provides an integer K, which is the number of variables (BACs) which will be randomly selected at each node. These K variables are used to find the best binary split, defined as a split such that the two resulting nodes are as homogeneous as possible in regard to subtype. This random selection of variables and best split continues until the largest tree possible is grown and not pruned. Eventually, N trees are grown, all with different bootstrap samples of the original data for root nodes. Once the "forest" is complete, the observations left out of each tree are classified by running those observations down their respective trees and getting a classification. The forest then chooses the classification having the most out of N votes. An internal error rate is computed by comparing the tree's classification to that which is known.

In order to estimate variable importance, in the left out observations for the wth tree, the values of this mth variable are randomly permuted. The "new" covariate values are put down the tree and classifications are made. Another internal error rate is computed and the mth variable is rated by how much the new error rate exceeds the original.

Implementing the Random Forest's algorithm with $5,000$ trees provided a list of **23** BACs deemed most important. The internal error rate for this "forest" was 14.83.

5 Clustering for Classification

Once potential target subsets have been identified, the next goal is to cluster those subsets for purposes of classification. Here we want to identify groups of BACs which can be used to classify the subtypes of cancer. A target subset of BACs can be defined by arbitrary cut-offs, significance tests, or variable importance as measured by tree regression. In this section we will explain the algorithm(s) implemented for clustering and subsequently detail the results of this clustering based on target subsets of BACs previously found in Section 4.

Algorithm The primary algorithm we used is partitioning around the medoids (PAM). This algorithm, described in Kaufman and Rousseeuw (1990),

takes as input a dissimilarity matrix and user-defined number of clusters, k. It first finds the k data points which are most centered in relation to the rest of the data and defines those as the medoids. Next, each data point is measured in relation to those medoids and put into a cluster with the medoid with the minimum distance from the data point. To compare different clusterings we are interested in the average silhouette width, a type of goodness of fit. The average silhouette width is calculated by comparing the distance of the data point to the elements in its neighbor cluster in relation to the elements in its own cluster. The silhouette width is evaluated as follows:

$$s_j = \frac{a_j - b_j}{max(a_j, b_j)},$$

where $s_j \in [-1, 1]$, a_j is the average of dissimilarity of j to all other objects in its cluster and b_j is the minimum of the average of dissimilarity of j with objects in neighboring clusters. When s_j is close to 1 it implies that the within dissimilarity a_j is much smaller than b_j the smallest of the between dissimilarities; thus, j is well classified. On the other hand, when s_j is close to 0 then a_j and b_j are approximately equal and hence it is not clear at all whether the observation should be in the same cluster or a neighbor. The worst case scenario is when s_j is close to 1. Then a_j is much bigger than b_j so the observation lies on the average much closer to a neighbor cluster than to cluster it is in; thus, j is misclassified.

In addition to using PAM on its own, we worked with hierarchical clustering with PAM, as described in van der Laan and Pollard (2001). The function we used from this hierarchical clustering was RunDownConverge, which takes the initial clustering from PAM and then tries to split and collapse the clusters with the goal of maximizing the average silhouette width. Initially *RunDownCollapse* was run up to the second level of clustering.

As mentioned above, PAM takes as arguments k and a dissimilarity matrix. This matrix is defined by a distance measure (Kaufman and Rousseeuw, 1990). Potential distance metrics are:

Euclidean: $$d_{i,j} = \sqrt{(x_{i1}, x_{j1})^2 + (x_{i2}, x_{j2})^2 + \ldots + (x_{ip}, x_{jp})^2}$$

Manhattan: $$d_{i,j} = \mid x_{i1} - x_{j1} \mid + \mid x_{i2} - x_{j2} \mid + \ldots + \mid x_{ip} - x_{jp} \mid$$

Table 15: *Average Silhouette widths and Cluster Size by Distance Metric with All BACs.*

Rule	Euc	Clus	Abs Euc	Clus	Eis	Clus	Abs Eis	Clus
All BACs	0.180	3	0.177	3	0.160	5	0.159	5

$$\text{Cosines of the Angle:} \qquad d_{i,j} = 1 - \frac{\frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj}}{\sqrt{\frac{1}{n}\sum_{k=1}^{n} x_{ki}^2 \times \frac{1}{n}\sum_{k=1}^{n} x_{kj}^2}}$$

To evaluate the contribution of different measures, we ran PAM for $k = 2, ..., 6$ with each of the following distance metrics: Euclidean, absolute Euclidean, cosine of the angle, and absolute cosine of the angle. After the initial cluster, *RunDownConverge* was employed for hierarchical clustering (HC). For each of these attempts the average silhouette width and number of clusters have been reported.

Each clustering attempt was done with all 42 tumors which includes the cancer subtypes and control tissue. In order to visualize the clustering results and gain an initial assessment of classification accuracy, a plot is drawn with the axis representing tumor labels and lines showing the separation of clusters. Within each cluster the tumor to the farthest left is the chosen medoid and the tumors to its right are ordered closest to farthest to that cluster's medoid. The gradation in color represents the "distance" of a tumor to the remaining tumors. Bright red signifies tumors in closest proximity, while bright green represents tumors farthest away.

In the following sections, clustering was performed using each distance metric for each target subset.

5.1 All BACs

A subset of all 88 BACs is in fact the entire set of 88 BACs. As a means of comparison for how well smaller target subsets do to **not** subsetting, we included the clustering of all BACs. The average silhouette widths for each of the distance measures as well as number of clusters chosen (from $k = 2, \ldots, 6$) for the respective distance measure is shown in the Table 15.

Table 16: *Average Silhouette widths and Cluster Size by Distance Metric with Arbitrary Cut-Offs.*

Rule	Euc	Clus	Abs Euc	Clus	Eis	Clus	Abs Eis	Clus
Larg Abs Mean	0.226	3	0.224	3	0.148	3	0.147	3
Larg Var	0.249	4	0.242	4	0.216	5	0.215	5

Although we did use both *RunDownCollapse* and *RunDownConverge* functions, neither improved on the clustering results and, thus, we have not included those results. For the initial level of PAM, the Euclidean distance measure provides the highest average silhouette width with three clusters. The cosine of the angle and absolute cosine of the angle distance allowed for more than three clusters for a total of five. Figures 7 and 8 show the plots for these distances at the initial level of PAM.

In Figure 7, the chromophobes are almost all separated out from the rest, while the majority of papillaries are apart from the conventionals. Due to the intermingling of the normals and benigns, this clustering is not sufficient for our purposes. In Figure 8, we can see that all five chromophobes group together into one cluster and the papillaries are together in a cluster with a few other subtypes. However, the differentiation of the conventionals, benigns, and normals is not sufficient. Thus, the subset of all BACs is not acceptable for classifying the subtypes.

5.2 Arbitrary Cut-off

The arbitrary cut-offs that we chose were the largest absolute BAC means and variances. The results of each distance metric at the initial level of PAM are shown in Table 16.

For both of these cut-offs, the *RunDownCollapse* function did not make a substantial improvement, if an improvement at all. In comparing the silhouette widths it appears as though the Euclidean measure did the best for both cut-offs. Figures 9 and 10 show the clustering results for the Euclidean and cosine of the angle metrics for the absolute means. In the first, it is apparent that the only advantage this cut-off and metric offer are to *almost* separate out the papillaries. In the second, we can see a fairly strong clustering of the papillaries and a few chromophobes. However, neither of these two clustering

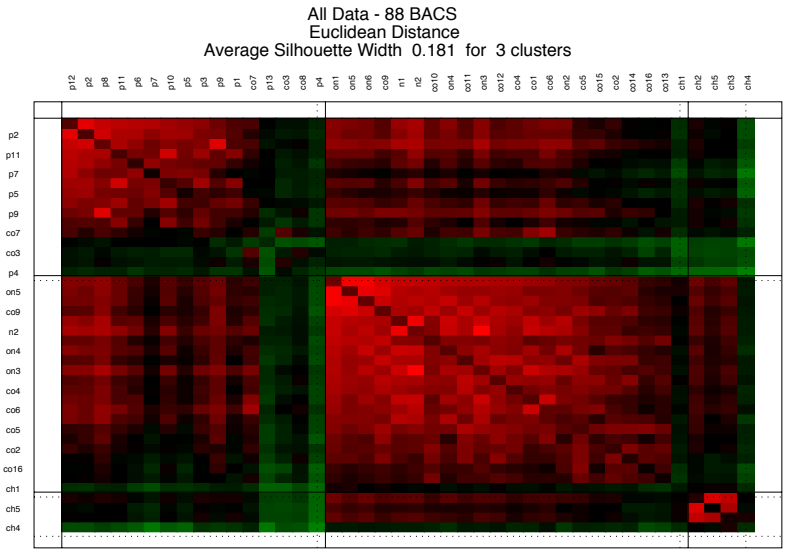

Figure 7: All BACs with Euclidean measure.

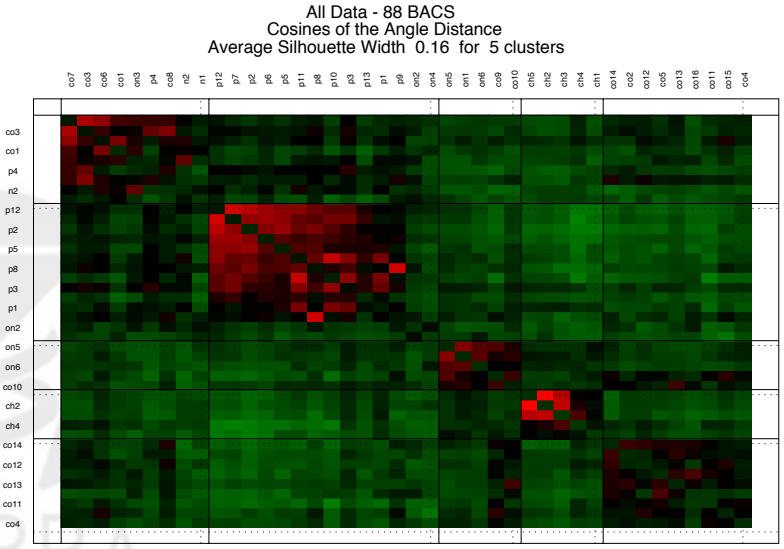

Figure 8: All BACs with Cosines of the Angle Distance.

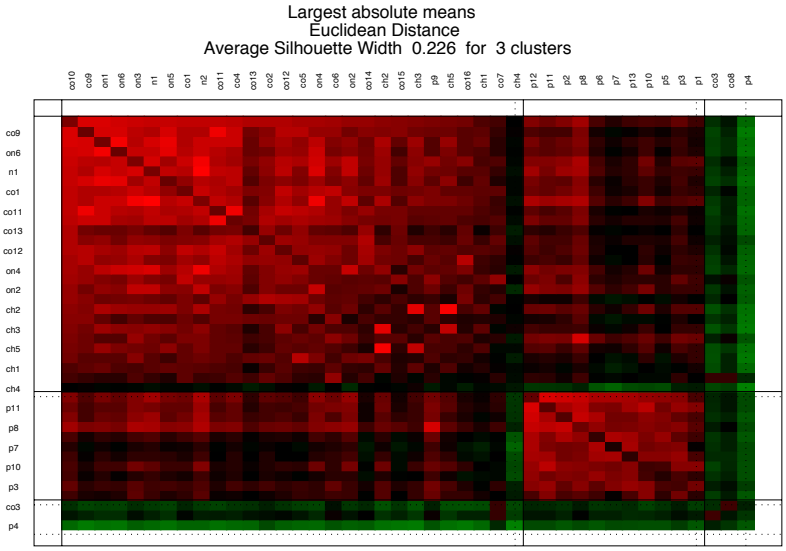

Figure 9: Largest absolute mean with Euclidean measure.

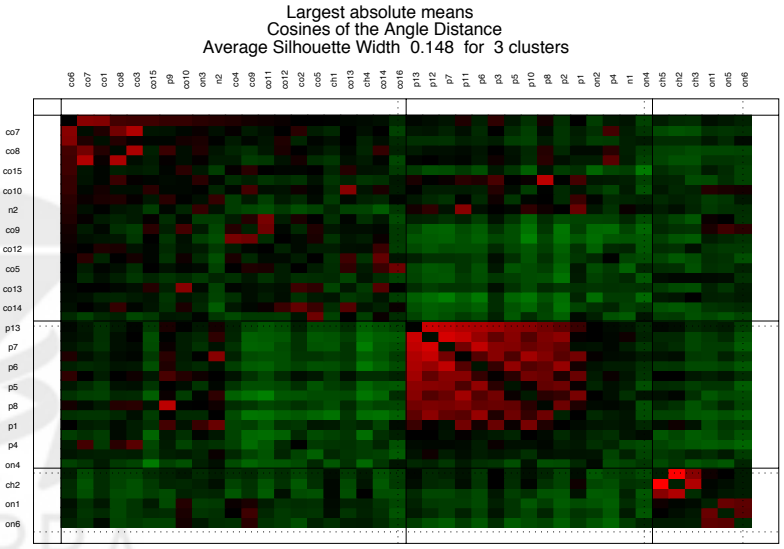

Figure 10: Largest absolute means with cosines of the angle distance.

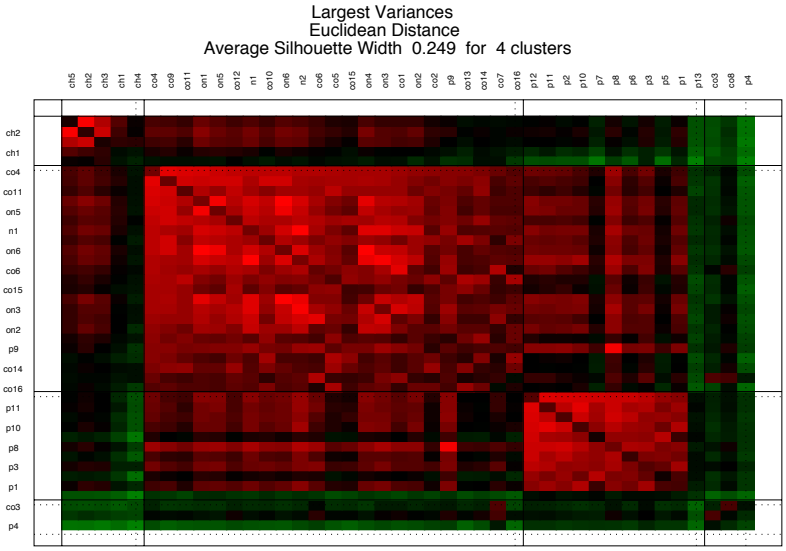

Figure 11: Largest variances with Euclidean measure.

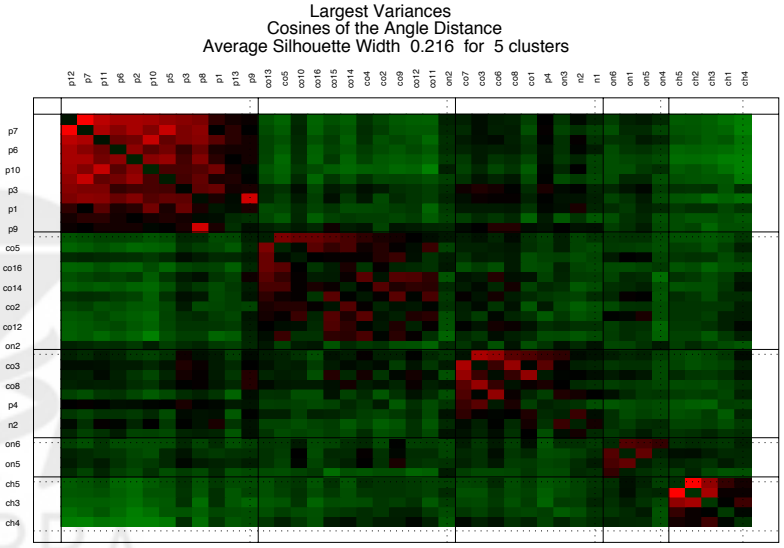

Figure 12: Largest variances with cosines of the angle distance.

results is satisfactory.

Figures 11 and 12 show the clustering for the largest variances with the Euclidean and cosine of angle metrics, respectively. The first demonstrates a nice separation of chromophobes and all but two papillaries with the conventionals mixed together with the normals and benigns. This offers a definite improvement over the absolute means with the same metric. The second plot is by far the best thus far. The chromophobes are clustered together, all but one of the papillaries are together in one cluster, and four of the benigns are separated out, while the conventionals are mixed with the remaining normals and benigns.

5.3 Significance Test

As described in Section 4.2, we looked at several significance test including the t-test (for equal and unequal variance assumptions), Wilcoxon Rank Sum test, Hollander's test of extremes, and Kruskal-Wallis test. The results of clustering each with the initial level of PAM, all four distance metrics, and three adjustments for the multiple comparisons problem are in Table 5.3.

For the t-test with equal variances, the Euclidean metric with the non-parametric bootstrap adjustment has an average silhouette width of 0.421 with four clusters. This plot shows the best differentiation of this group's plots. It only misclassified **three** of the conventionals and puts them with the benign/normal cluster. Figure 13 shows the plot of this clustering.

Similarly, for the t-test for which unequal variances are assumed, the Euclidean metric is employed, and the non-parametric bootstrap adjustment, we again get the best result of the group's clusterings. Figure 14 shows the results of this clustering, where all but **four** conventionals are properly clustered.

For the Wilcoxon Rank Sum test, the Bonferroni adjustment with the cosine angle metric has an average silhouette width of 0.256 with six clusters. Figure 15 shows the results of this clustering where only three tumors are misclassified. Interestingly, the conventional are split into two clusters as well as the normal/benign. For comparisons, we have included the plot for the Wilcoxon Rank Sum test where no adjustment is made and the cosine angle metric is employed. This plot is shown in Figure 16. Here, six tumors are misclassified.

When reviewing the results from the Hollander's Test of Extreme Reactions, the best plot for classification is that with $\alpha = 0.01$ and absolute

Table 17: *Average Silhouette widths and Cluster Size by Distance Metric with Significance Tests*

Rule	Adjustment	Euc	Clus	Abs Euc	Clus	Eis	Clus	Abs Eis	Clus
t-test	No adj.	.218	3	.234	3	.185	4	.185	4
Equal	Bonferroni	.326	3	.316	3	.257	3	.243	5
Var	NP Bootstrap	.421	4	.419	2	.466	3	.397	2
t-test	No adj.	.255	3	.249	3	.203	4	.201	4
Unequal	Bonferroni	.372	2	.360	2	.365	3	.307	3
Var	NP Bootstrap	.421	4	.419	2	.466	3	.397	2
	No adj.	.231	2	.246	4	.194	5	.194	5
Wilcoxon	Bonferroni	.328	2	.294	2	.256	6	.243	6
Rank Sum	NP Bootstrap	.231	2	.246	4	.194	5	.194	5
	No adj.	.210	2	.247	4	.181	4	.180	4
Hollander	$\alpha = 0.01$	.307	3	.300	3	.25	6	.25	6
	Perm. Test	.204	2	.251	4	.207	5	.207	5
Kruskal-	No adj.	.26	3	.256	3	.226	6	.226	6
Wallis	Perm. Test	.348	3	.334	3	.275	6	.259	6

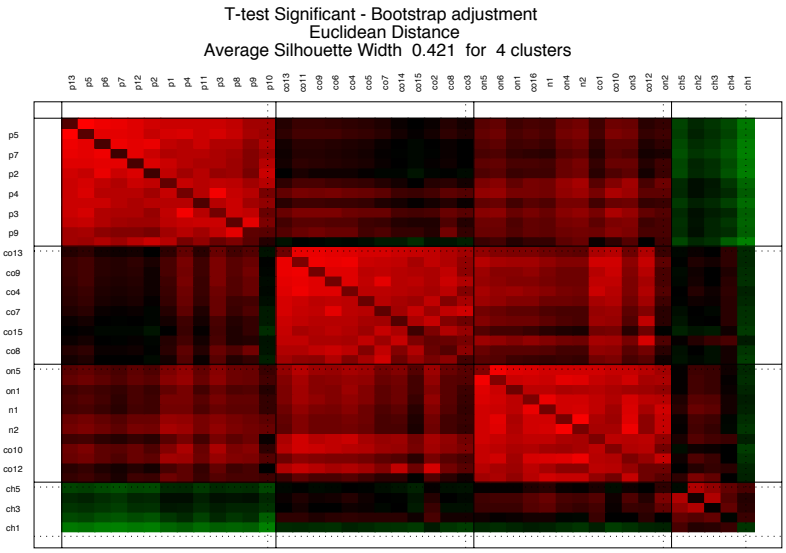

Figure 13: *t*-test equal variance, Euclidean measure, and bootstrap adjustment.

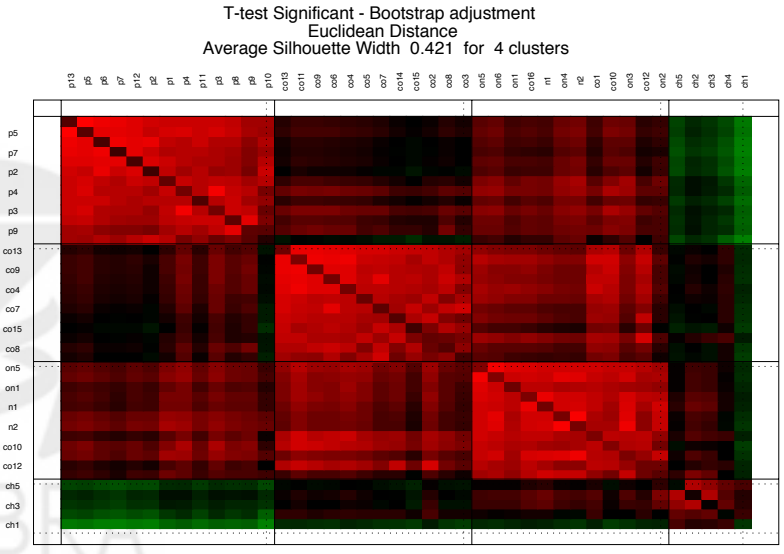

Figure 14: *t*-test unequal variance, Euclidean measure, and bootstrap adjustment.

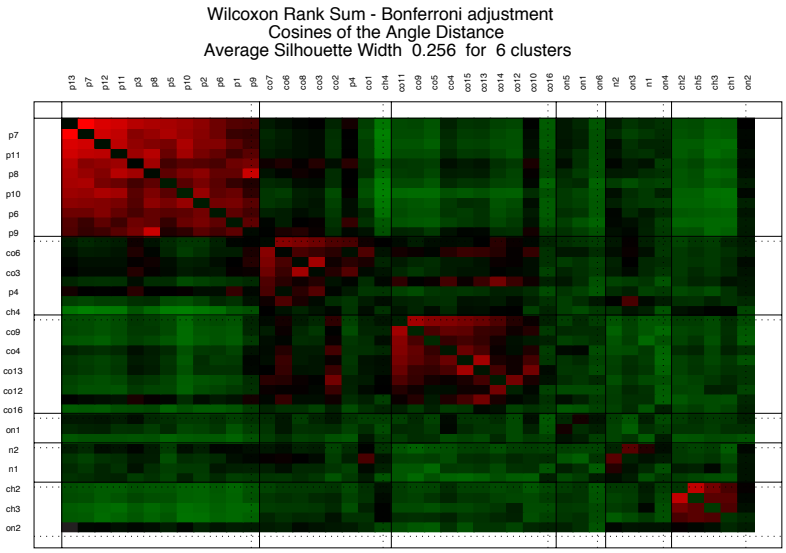

Figure 15: Wilcoxon Rank Sum with cosine angle measure and Bonferroni adjustment.

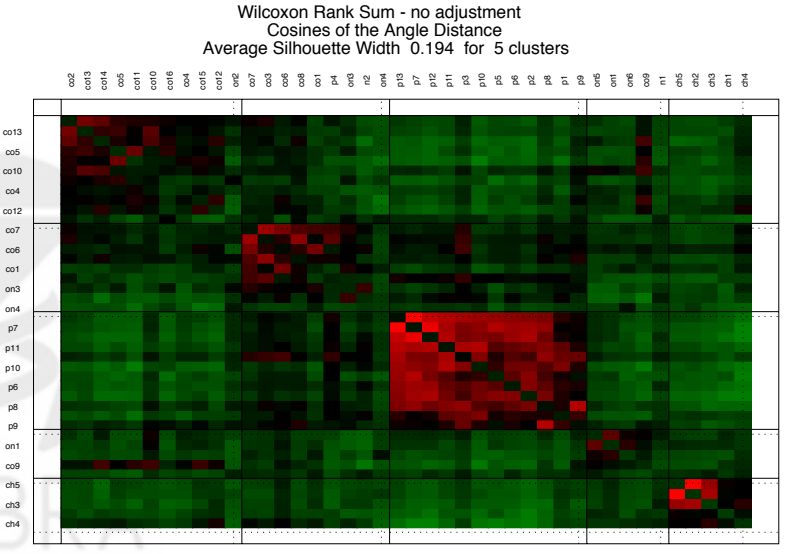

Figure 16: Wilcoxon Rank Sum with cosine angle measure and no adjustment.

cosine angle distance metric. This plot is shown in Figure 17. Although the conventional and normal/benign subtypes are split into two groups each, there is only a total of one misclassification in the entire clustering.

The Kruskal Wallis test with the permutation test adjustment and cosine angle metric also supersedes the previous results with only **one** misclassification. As one can see in Figure 18, this clustering also splits the conventional and normal/benign groups into two each.

5.4 Variable Importance as Measured by Regression Trees

The result from clustering the 23 BACs chosen by the tree regression measure for variable importance in 5,000 trees is shown in Figure 19. This clustering employed the Euclidean metric and besides splitting the conventional and chromophobe subtypes into two each, it has perfect classification.

6 Conclusions

The purpose of this exercise was to find a subset of BACs to classify chromosomal aberrations into tumor subtype with a CGH data set. We approached this by applying standard statistical methods for subsetting and then using hierarchical clustering for classification.

One of the most typical approaches to this type of data is to subset using univariate tests between subtypes. However, as we have seen from the results, the most favorable clusterings came from tests which took as much data into account as possible (i.e., Kruskal Wallis and Random Forests). Although, it should be noted that the Hollander Test of Extremes fared well at the $\alpha = 0.01$ level. Our inclination is, however, for the other two tests, as the results from the Kruskal Wallis are based on a permutation test which accommodates for the multiple comparisons problem and, by design, Random Forests also does.

[0]This research has been supported by a grant from the Institute for Scientific Computing Research at Lawrence Livermore National Laboratory.

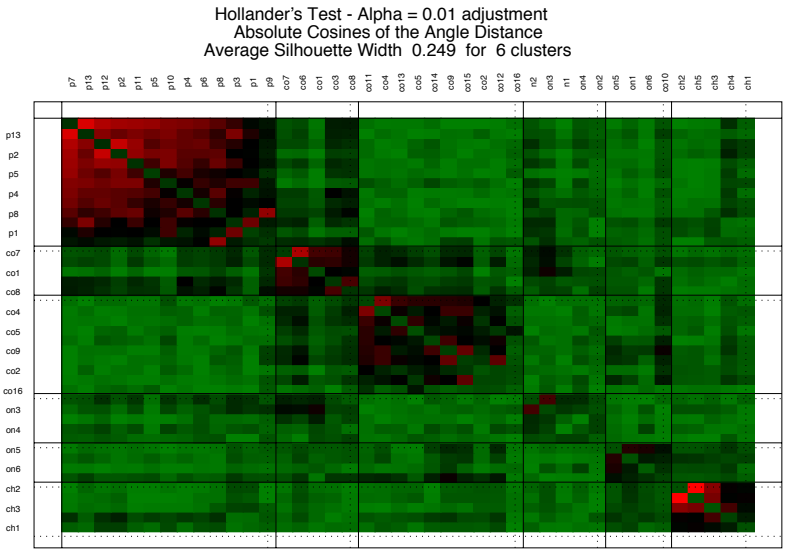

Figure 17: Hollander's test with absolute cosine angle measure and $\alpha = 0.01$ adjustment.

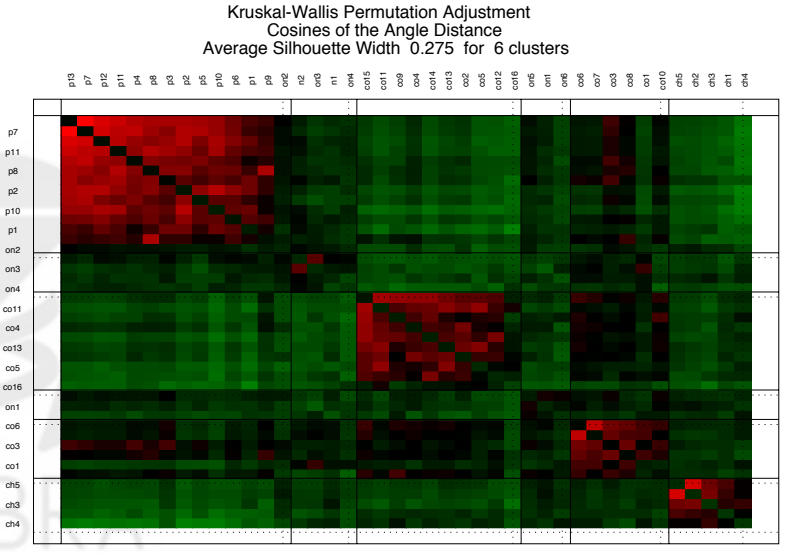

Figure 18: Kruskal Wallis with cosine angle measure and permutation test adjustment.

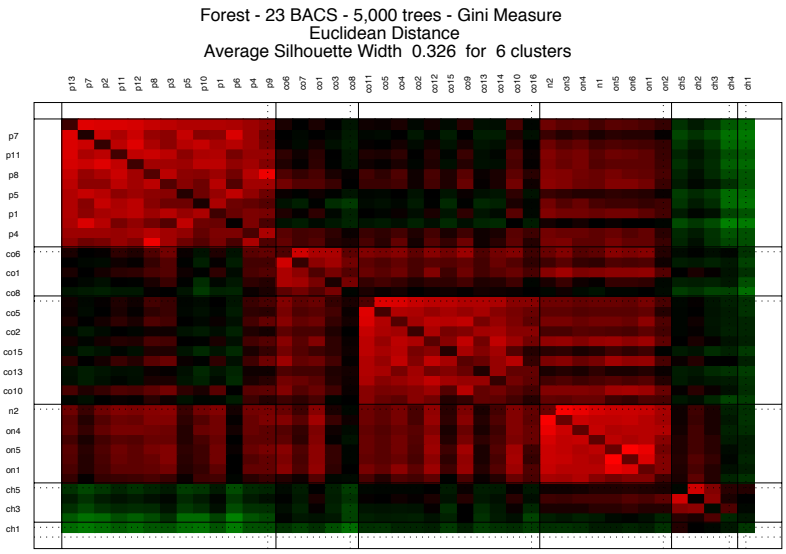

Figure 19: Random Forest, gini measure with the Euclidean metric.

References

L. Breiman. Random forests. Technical report, Statistics Department, University of California, Berkeley, 2001.

L. Breiman, J. H. Friedman, R.A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey, CA, 1984.

M. Hollander. A nonparametric test for the two-sample problem. *Psychometrika*, 28: 395–403, 1963.

R. Ihaka and R. Gentlemen. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, NY, 1990.

W. H. Kruskal. A nonparametric test for the several sample problem. *Ann. Math. Statist.*, 23:525–540, 1952.

D. Pinkel, R. Segraves, S. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Z. Zhai, S. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20:207–211, 1998.

A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel, and D. G. Albertson. Assembly of microarrays for genome-wide measurement of dna copy number. *Nature Genetics*, 29:263–264, 2001.

M. J. van der Laan and J. F. Bryan. Gene expression analysis with the parametric bootstrap. *Biostatistics*, 2:1–17, 2001.

M. J. van der Laan and K. Pollard. Hybrid clustering of gene expression data with visualization and the bootstrap. Technical Report 93, Biostatistics Department, University of California, Berkeley, 2001. URL `www.bepress.com/ucbbiostat/paper93/`.

W. M. Veltman, A. B. Olshen, A. N. Jain, D. H. Moore, J. C. Presti, G. Kovacs, and F. M. Waldman. Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Research*, 62(4):957–960, 2002.