Carrie Buchanan

BIOS301

Final Project

**Implementation of VAAST Composite Likelihood Ratio Test in R**

**Background:** Sequence analysis is a rapidly evolving field of study in human genetics. Because the memory required to store and manipulate raw sequence data is orders of magnitude larger than genotype data, there is an increasing interest in developing novel techniques to bin or collapse variants at meaningful biological boundaries to perform sequence analysis. One such promising method is called the Variant Annotation and Analysis and Search Tool (Yandell et al. 2011) It is able to annotate and score variant bins, then perform an association test on the data. It has been successful at identifying regions of rare (and some common) variants that are associated with disease. The VAAST method employs a composite likelihood ratio test to test for association. Knowledge about the number of variants in cases and controls is necessary for the CLR. This code applies a VAAST-like CLR to output data from BioBin developed in the Ritchie lab.

BioBin is an algorithm under development in the Ritchie lab to analyze sequence data. It is unique because the user can specify flexible boundaries to bin variants. An example boundary would be the start and stop positions of genes across the genome. In this case, BioBin would collapse variants from sequence data into bins, and the bins would represent all of the genes in the genome. The user can then take BioBin output to test for association. Many researchers are interested in looking for an enrichment of rare variants in disease causing genes. One could take the bins from BioBin and test for an association of increased rare variants in cases over controls. The following example shows the format of the BioBin output and an example application of this program using two gene bins.

**Example:**

**EXAMPLE BIOBIN OUTPUT / CLT INPUT:**

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Status | EGR4 | KCNC3 |
| Totals | 2 | 45 | 7 |
| NA06984.ceu | 2 | 5 | 0 |
| NA06985.ceu | 2 | 5 | 0 |
| NA06986.ceu | 2 | 5 | 0 |
| NA06989.ceu | 2 | 5 | 1 |
| NA06994.ceu | 2 | 5 | 0 |
| NA07000.ceu | 2 | 5 | 1 |
| NA07037.ceu | 2 | 5 | 0 |
| NA07048.ceu | 2 | 5 | 0 |
| NA07051.ceu | 2 | 5 | 0 |
| NA07346.ceu | 2 | 1 | 0 |
| NA07347.ceu | 2 | 1 | 0 |
| NA07357.ceu | 2 | 0 | 0 |
| NA20534.tsi | 1 | 0 | 0 |
| NA20535.tsi | 1 | 0 | 0 |
| NA20537.tsi | 1 | 1 | 0 |
| NA20538.tsi | 1 | 0 | 0 |
| NA20540.tsi | 1 | 0 | 1 |
| NA20542.tsi | 1 | 0 | 0 |
| NA20543.tsi | 1 | 1 | 1 |
| NA20544.tsi | 1 | 0 | 0 |
| NA20585.tsi | 1 | 0 | 0 |
| NA20586.tsi | 1 | 1 | 0 |
| NA20587.tsi | 1 | 0 | 0 |
| NA20589.tsi | 1 | 0 | 0 |
| NA20752.tsi | 1 | 0 | 0 |

The first column is always the individual ID corresponding to all of the samples in your study. In this example, these are individual IDs from 1000 Genomes Project data. The second column indicates each individual’s status. Per convention in the field of genetics, the value 1 indicates control/unaffected status and 2 indicates case/affected status. After the second column, each column represents a bin. As mentioned earlier, this example uses gene boundaries as bins, thus each column header is named for a gene. The first value in the column represents that total number of sites that contain a variant for all of the data. Each value below the total represents the total number of variants that individual contributes to the bin. For example, for EGR4, there are total 45 variant sites in the data set. Individual NA06984 contributes 5 variants (can contribute 0,1,2 minor alleles at each site).

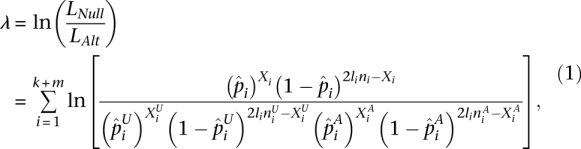
The overall goal of using CLRT is to sum up the variants for cases and controls and see if there is an enrichment of rare variants in cases.

The usage of this program is very simple. Call the function clrt and the only input is the file path for the location of the BioBin output data.

*clrt("~/Desktop/exon\_small.csv")*

As the program runs, it brings up your data to view. Then it will give two tables of output.

The first table will have a row for each row in your dataset. In this example, the first two rows (ID and status) are not interesting. The next two rows are the two bins in this data. The lambda value can be calculated from the following equation (Yandell et al. 2011)



*k* is the number of uncollapsed variant sites which is not relevant in this code (everything is collapsed). *n* represents the number of individuals (ni=total, na=affected, nu =unaffected). *X* corresponds to the total number of minor alleles (can be 0, 1, or 2 at each variant site). *p* is equal to the maximum-likelihood estimates for the frequency of minor alleles in each collapsing category. Assuming that the variant sites are not in LD, -2λ approximately follows a χ2 distribution. Therefore, the non-LD corrected score is what is considered in this program.

The second column of the first table gives the p-values for the each of the bins. The last columns are helpful to see the distribution of variants in each bin between cases and controls. X corresponds to the number of minor allele copies for affecteds, unaffecteds, and total (respectively). AF corresponds to the frequency of rare alleles in affecteds, unaffecteds, and total.

The second table lists the genes significantly enriched in rare variants and their p-value. In this made-up data set, EGR4 has 47 minor alleles in affected individuals and only 3 in unaffected individuals. Therefore, it is associated with the case/control status.

For more information about VAAST or BioBin please refer to the following websites:

<http://www.yandell-lab.org/software/vaast.html>

<http://ritchielab.psu.edu/>