KBAC with Permutation Testing

Author: James Grover, Golden Helix, Inc.

Overview

The Kernel-Based Adaptive Cluster (KBAC) method by Liu and Leal [Liu and Leal 2010] first catalogs the variant data within each of a number of regions into multi-marker genotypes. Since the variants are rare, only a relatively few different multi-marker genotypes will be found in any given region.

A special case/control test based on these counts is then applied. This test is weighted for each multi-marker genotype according to how often that genotype is expected to occur according to the data and according to the null hypothesis that there is no association between that genotype and the case/control status of the sample. Under this adaptive weighting procedure, the genotypes with high sample risks will be given higher weights which can potentially separate causal from non-causal genotypes. This procedure is meant to attain a good balance between classification accuracy and the number of parameters which are estimated.

Recommended Directory Location

Save the script to the following directory:
*..**\Application Data\Golden Helix SVS\UserScripts\Spreadsheet\Analysis**

Note: The **Application Data** folder is a hidden folder on Windows operating systems and its location varies between XP and Vista. The easiest way to locate this directory on your computer is to open SVS and select the **Tools >Open Folder > UserScripts Folder** menu option. If saved to the proper folder, this script will be accessible from the spreadsheet **Analysis** menu.

Data Requirements

KBAC analysis is to be performed from a marker-mapped spreadsheet containing filtered sequence data and a case/control phenotype as the dependent column (signified by its dependent status).

NOTE:

- You should first filter the markers by rare variants as the script assumes that only rare variant markers are active.
- If you wish to filter based on a user-defined minor-allele frequency threshold and an external reference population provided through a probe annotation track, follow this procedure.
    - From the spreadsheet choose **Quality Assurance > Genotype > Variant Frequency Binning by MAF Track**.

- Enter the minor-allele threshold as the ”bin 0 threshold” to generate a ”binning” spreadsheet.
- Activate only rows in the ”binning” spreadsheet that are in bin 0.
- Go back to the original spreadsheet and choose **Select > Activate or Inactivate Based on Second Spreadsheet**, choosing the “binning” spreadsheet as your second spreadsheet.

- Please see the Variant Frequency Binning by MAF section in the SVS7 manual for more information: http://www.goldenhelix.com/SNP_Variation/Manual/svs7/manualsu77.html#x89-3140004

Input Parameter Description

Kernal Type:

- *Hyper-Geometric*: Normally recommended and gives the most accurate test.
- *Marginal Binomial*: Less compute-intensive and almost as accurate as the hypergeometric kernel.
- *Asymptotic normal*: This is for large sample sizes only, such as more than 400 cases and 400 controls and is not as accurate as the other two kernel methods.

Permutation Parameters for the KBAC Test

- *# Permutations:* The number of permutations to use.
- *Permutation Mode:*
  - *Standard C/C permutation procedure*: This is normally recommended and gives the most accurate test.
  - *KBAC Monte-Carlo approximation:* This is for large sample sizes only (the authors suggest more than 400 cases and 400 controls) and is not as accurate as standard case/control permutation testing.

Outputs from the KBAC Test

- *One-sided statistics:* This tests the one-sided alternative hypothesis of the enrichment of causal variants in cases. This output is recommended and default in the KBAC window.
- *Two-sided statistics*: This tests the two-sided alternative hypothesis of the difference between weighted multisite genotype frequencies between cases and controls.
- **NOTE:** Due to the often asymmetric distribution of the one-sided KBAC test-statistic, two-sided p-values will not always be twice the one-sided p-values and may in fact be exactly the same if the permuted distribution were entirely positive.

Regions for the KBAC Test

- Specify an annotation gene track that will be used to distinguish regions for testing. If possible it is recommended that a local filtered gene track be used to determine the gene list. A local track will allow the spreadsheet to be processed faster than if a network track was used. An example filtered gene track name is: *FilteredRefSeqGenes-UCSC_NCBI_36_Homo_sapiens.idf:1*

Missing Genotype Values for the KBAC Test

- *Impute Wild Type for Genotypic Missing Values:* If there is any missing genotype data in the current region for a given sample, either the wild type (homozygous reference genotype) can be imputed to have been present and substituted for the missing data, or the sample can be skipped and not used. Select **Yes** to impute the wild type or **No** to skip samples containing missing data for the current region.

Output

A marker-mapped spreadsheet will be generated as output containing one row for each region and containing some or all of the following columns, according to what you have requested:

- **Chr**: The chromosome number or label.
- **Start:** The starting genetic position of the region.
- **Stop:** The ending genetic position of the region.
- **Name:** The region or gene name.
- **P-Value (One-Sided):** The p-value resulting from the one-sided KBAC test.
- **-log10 P-Value (One-Sided):** Minus the base-10 log of the above p-value.
- **KBAC (One-Sided):** The one-sided KBAC statistic itself.
- **Bonf. P (One-Sided):** The Bonferroni-adjusted p-value from the one-sided test.
- **FDR (One-Sided):** The false discovery rate for all tests having a one-sided p-value less than or equal to the current test's one-sided p-value.
- **P-Value (Two-Sided):** The p-value resulting from the two-sided KBAC test.
- **-log10 P-Value (Two-Sided):** Minus the base-10 log of the above p-value.
- **KBAC (Two-Sided):** The two-sided KBAC statistic itself.
- **Bonf. P (Two-Sided):** The Bonferroni-adjusted p-value from the two-sided test.
- **FDR (Two-Sided):** The false discovery rate for all tests having a two-sided p-value less than or equal to the current test's two-sided p-value.
- **# Markers:** The total number of markers tested in this region.
- **# Multi-Marker Genotypes:** The total number of distinct multi-marker genotypes discovered in this region.