BioBin 2.3 Vignette
July 26, 2017

Table of Contents

Overview

BioBin is a software developed for the biologically driven binning and association analysis of rare sequence variants. This vignette is designed to highlight the new features of BioBin 2.3. A summary of these features can be found in **Table 1** below. For use cases of the originally implemented software features please see the BioBin manual under the ***Example Usage*** section as well as the feature-specific ***Example*** sections.
http://www.ritchielab.com/files/RL_software/biobin-manual-2.3.pdf

Example files to accompany this vignette can be found on the software download page www.ritchielab.com/software/biobin-download . These files include:

1. VCF file for Chromosome 22 from the CEU-TSI targeted exome, 1000 Genomes Project “test_vcf.vcf”
2. Phenotype file with 3 fictitious case/control phenotypes “test_multi_phenotypes.phe”
3. Covariate file with 1 fictitious numeric covariate “test_covariates.cov”
4. Custom role file “role_file”
5. Custom region file “region_file”
6. Sample exclusion list “exclude_sample.list”
7. Sample configuration files: “external_regions.config”, “external_roles.config”, “gene_analysis.config”
8. Results files for each example analysis

Analyses can be run by using a configuration file or directly on the command line. The configuration file option requires a plain text file containing all options for BioBin. Each line of this field must either be blank, a comment beginning with (#), or a configuration value pair. Configuration value pairs are specified in the following syntax:

```
config-name=value
```

We have provided example configuration files for a gene analysis, a customized regions analysis, and a customized role analysis. Users are encouraged to consult these configuration files as they list all possible options for a BioBin analysis.

In the sections below, we will provide examples for gene, pathway, interregion, customized role and customized region analyses.

Table 1: This table highlights the major changes that have been implemented in BioBin 2.3 as compared to previous versions of BioBin

Previous BioBin Versions	BioBin 2.3
Single Phenotype analysis	Multiple phenotype analysis (PheWAS)
No parallel computing	Parallel computing for multiple phenotype analyses
No statistical testing	Burden analysis (Regression, wilcoxon rank sum); Dispersion analysis (SKAT)
No on-the-fly sample processing	Sample inclusion/exclusion options
Output =rare variant bins + summary information	Output = rare variant bins, summary information, and p-value of given statistical test OR Output =rare variant bins + summary information (if statistical testing is not desired)

Gene Analysis

A configuration file for a gene analysis using the example data has been provided for users "gene_analysis_2.3.config". To run this equivalent analysis on the command line (without using a configuration file) the following can be used:

```
biobin-2.3 -G 37 -D knowledge.bio -V test_vcf.vcf -p test_multi_phenotypes.phe --covariates test_covariates.cov -F 0.05 --bin-
regions Y --bin-pathways N --bin-interregion N --test SKAT-logistic,logistic –t 2 --weight-loci Y --report-prefix gene_analysis
 --exclude-sources dbSNP,oregano,ucsc_ecr
```

This analysis:

- bins all variants with a MAF < 5% in the case or control populations into gene bins
- adjusts for the given covariate
- tests for association using logistic regression and SKAT
- excludes dbSNP, oregano, and ucsc_ecr from the available LOKI sources used for binning
- weights loci using Madsen and Browning weighting
- uses 2 threads

Four files will be output from this analysis: A locus file, “gene_analysis-locus.csv”, and 3 bins files (1 per phenotype), “gene_analysis-P1-bins.csv”, “gene_analysis-P2-bins.csv”, and “gene_analysis-P3-bins.csv”.

Pathway Analysis

```
biobin-2.3 -G 37 -D knowledge.bio -V test_vcf.vcf -p test_multi_phenotypes.phe --covariates test_covariates.cov -F 0.03 --bin-
regions N --bin-pathways Y --bin-interregion N --test logistic –t 5 --weight-loci Y --report-prefix pathway_analysis
 --exclude-sources biogrid
```

This analysis:

- bins all variants with a MAF < 3% in the case or control populations into pathway bins
- adjusts for the given covariate
- tests for association using logistic regression
- uses 5 threads
- excludes biogrid from the LOKI pathway sources used for binning

Four files will be output from this analysis: A locus file, “pathway_analysis-locus.csv”, and 3 bins files (1 per phenotype), “pathway_analysis-P1-bins.csv”, “pathway_analysis-P2-bins.csv”, and “pathway_analysis-P3-bins.csv”.

The pathway bin names will first list an abbreviation of the source followed by the pathway name. For example:

- “go: protein homodimerization activity” = homodimerization activity pathway from Gene Ontology (GO)

- “pfam: DNA topoisomerase” = DNA topoisomerase pathway as annotated by the pfam protein families database

Multiple Biological Feature Analysis

Users may also bin variants by multiple biological features simultaneously in one analysis

Gene and interregion analysis

```
biobin-2.3 -G 37 -D knowledge.bio -V test_vcf.vcf -p test_multi_phenotypes.phe --covariates test_covariates.cov -F 0.04  --bin-
regions Y --bin-pathways N --bin-interregion Y -i 1 -m 2 --test SKAT-logistic -t 5 --exclude-samples exclude_sample.list
 --report-prefix gene_interregion_analysis
```

This analysis:

- bins all variants with a MAF < 4% in the case or control populations into gene bins and interregion bins (These bins consist of loci not contained in any region. They are binned into generic base-pair bounded bins)*
- bins must have a minimum bin size of 2 variants
- interregions will be binned into 1kb bins
- adjusts for the given covariate
- tests for association using SKAT
- excludes 5 samples listed in the “exclude.sample.list”

Four files will be output from this analysis: A locus file, “gene_interregion_analysis-locus.csv”, and 3 bins files (1 per phenotype), “gene_interregion_analysis-P1-bins.csv”, “gene_interregion_analysis-P2-bins.csv”, and “gene_interregion_analysis-P3-bins.csv”.

*Please note that the example VCF is targetted exome, therefore there will not be any interregion bins in the results. This example is simply to provide the user with possible analysis options. If interregion binning is possible, interregion bins will be named in the chr:start-stop format.

Custom analyses

There are multiple ways by which to run customized analyses in BioBin. Two such examples are with a custom region file and a custom role file.

Analysis with custom roles

Variants may be binned by supplying customized role information to BioBin using a role file. This role file allows users to define the roles of variants or regions in the genome. Roles are intended as secondary binning mechanisms following gene binning. Examples of roles include exons, introns, or various functional roles. BioBin currently restricts the number of unique roles to 60.

A configuration file for a custom role analysis using the example data has been provided for users “external_roles_2.3.config”. To run this equivalent analysis on the command line (without using a configuration file) the following can be used:

```
biobin-2.3 -G 37 -D knowledge.bio -V test_vcf.vcf -p test_multi_phenotypes.phe --covariates test_covariates.cov -F 0.05 --bin-
regions Y --bin-pathways N --bin-interregion N --bin-expand-roles Y --bin-expand-size 1 --weight-loci Y  --test logistic –t 2
--role-file role_file --report-prefix role_analysis
```

This analysis:

- bins all variants with a MAF < 5% in the case or control populations into gene-role bins
- adjusts for the given covariate
- tests for association using logistic regression
- bins variants by roles
- bins are expanded into child bins after 1 bin
- uses 2 threads

Four files will be output from this analysis: A locus file, “role_analysis-locus.csv”, and 3 bins files (1 per phenotype), “role_analysis-P1-bins.csv”, “role_analysis-P2-bins.csv”, and “role_analysis-P3-bins.csv”.

Please note this test analysis will issue a warning of the form “WARNING: Unable to lift region at line XX”. This warning is due to genomic build differences when the VCF is in a build prior to build 38. LOKI lifts variants over to build 38, and this warning is specific to variants which were not lifted to this genomic build.

Analysis with custom regions

BioBin provides users with the option of running analyses with biological information outside of the LOKI database using a region file.

A configuration file for a custom regions analysis using the example data has been provided for users “external_regions_2.3.config”. To run this equivalent analysis on the command line (without using a configuration file) the following command can be used:

```
biobin-2.3 -G 37 -D knowledge.bio -V test_vcf.vcf -p test_multi_phenotypes.phe --covariates test_covariates.cov -F 0.05 --bin-regions Y --bin-pathways N --bin-interregion N --weight-loci Y  --test logistic –t 2 –m 2 --region-file region_file --report-prefix region_analysis
```

This analysis:

- bins all variants with a MAF < 5% in the case or control populations into region bins
- adjusts for the given covariate
- tests for association using logistic regression
- uses 2 threads
- minimum bin size of 2

Four files will be output from this analysis: A locus file, “region_analysis-locus.csv”, and 3 bins files (1 per phenotype), “region_analysis-P1-bins.csv”, “region_analysis-P2-bins.csv”, “region_analysis-P3-bins.csv”.