Bioinformatics, 34(3), 2018, 527–529
doi: 10.1093/bioinformatics/btx559
Advance Access Publication Date: 14 September 2017
Applications Note

OXFORD

Genetics and population analysis

Novel features and enhancements in BioBin, a tool for the biologically inspired binning and association analysis of rare variants

Anna O. Basile[1], Marta Byrska-Bishop[2], John Wallace[2], Alexander T. Frase[2] and Marylyn D. Ritchie[2,*]

[1]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA and [2]Biomedical and Translational Informatics Institute, Geisinger, Danville, PA, 17822 USA

*To whom correspondence should be addressed
Associate Editor: Oliver Stegle

Received on April 26, 2017; revised on August 3, 2017; editorial decision on September 3, 2017; accepted on September 13, 2017

Abstract

Motivation: BioBin is an automated bioinformatics tool for the multi-level biological binning of sequence variants. Herein, we present a significant update to BioBin which expands the software to facilitate a comprehensive rare variant analysis and incorporates novel features and analysis enhancements.
Results: In BioBin 2.3, we extend our software tool by implementing statistical association testing, updating the binning algorithm, as well as incorporating novel analysis features providing for a robust, highly customizable, and unified rare variant analysis tool.
Availability and implementation: The BioBin software package is open source and freely available to users at http://www.ritchielab.com/software/biobin-download
Contact: mdritchie@geisinger.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

BioBin (Moore *et al.*, 2016, 2013a) is a bioinformatics tool developed for the biologically informed binning of rare variants in DNA sequence data. Collapsing or binning approaches, which aggregate variants into single genetic units, have proven successful in increasing power in rare variant analyses. BioBin builds on existing methods by utilizing an internal repository, Library of Knowledge Integration (LOKI) (Pendergrass *et al.*, 2013), for the multi-level binning of variants into user-defined biological features, such as genes, pathways, regulatory regions, protein families and others. LOKI unifies and integrates over a dozen public databases, including NCBI Entrez Gene (NCBI Resource Coordinators, 2013), PharmGKB (McDonagh *et al.*, 2011) and KEGG (Kanehisa *et al.*, 2012) into one repository to inform variant binning. The utility of BioBin has been proven in numerous analyses studying rare variant influences on complex phenotypes (Basile *et al.*, 2016; Kim *et al.*, 2015; Moore *et al.*, 2013b). BioBin was originally developed solely as a variant binning tool. Herein, we present BioBin 2.3, a new release of our software which incorporates statistical testing, and implements novel analysis features as well as updates to the binning algorithm thereby providing a comprehensive and unified rare variant analysis tool. The new features of BioBin are highlighted in Supplementary Figure 1.

2 Methods

2.1 Implementation of statistical tests

The framework of BioBin has been expanded to incorporate statistical methods, thereby enabling complete rare variant analysis, from binning to association testing, in one software. Burden and dispersion tests are two categories of statistical tests used in conjunction with binning methods. Burden tests assess the cumulative impact of variants in a genetic region. They are count based, and follow the assumption that all variants influence the trait with comparable

magnitude and direction of effect. Therefore, burden tests suffer a loss of statistical power when variants of mixed trait effects are present. Dispersion methods are robust to these conditions as they test the distribution of variants (Wu *et al.*, 2011). Hence, they maintain power in the presence of effect heterogeneity but can lose power if the majority of variants have a similar impact on the trait (Wu *et al.*, 2011).

To facilitate robust statistical analysis, we have implemented regression and the wilcoxon rank sum test, two standard burden approaches frequently used in rare variant analyses (Lee *et al.*, 2014), as well as a dispersion test, SKAT. SKAT is a widely used method that applies a variance score test within a multiple regression kernel framework to determine the distribution of variants and test for association (Wu *et al.*, 2011). SKAT and regression have been implemented to allow for analysis of binary and continuous phenotypes, as well as covariate adjustment. The addition of statistical tests transforms BioBin into a comprehensive tool which streamlines rare variant analysis, saves time, and also avoids possible time-intensive file conversion issues.

2.2 Updates to binning algorithm

Additional updates in BioBin 2.3 include enhancements to the variant binning algorithm that improve handling of sites containing spanning deletions as well as genotype level filters in a variant call format (VCF) file. Spanning deletions are deletions that overlap a position of interest, such as a single nucleotide polymorphism (SNP). They occur at sites where a SNP in one sample is a part of a deletion in another sample, and are represented with an asterisk ('*') in multi-sample VCFs generated using GATK v3.4-46 and above. If the start of a deletion (e.g. POS = 14, REF = GCCCAC, ALT = G) and the position of a SNP (POS = 18, REF = A, ALT = T, *) are reported separately in a VCF, then the spanning deletion is reported as an '*' at the SNP site. In this case, the SNP site is multi-allelic, at which some samples have a SNP (A→T), while other samples have a deletion that spans that SNP (A→*), reported separately at POS = 14. If a deletion spans multiple SNPs, '*' alleles will be listed for each SNP. Previous BioBin versions did not handle '*' alleles and would count samples with spanning deletions at least twice, when binning a deletion and when binning the overlapping SNP/s. BioBin 2.3 sets '*' alleles to referent before variant binning thereby not biasing the variant count. Alternatively, users can set '*' allele samples to missing in which case they would be excluded from the analysis and counted only when binning a deletion.

Another improvement is the handling of genotype filters in a VCF. An FT flag is a sample-level genotype filter indicating whether a particular sample passed all filters set during quality control. For example, 'FT: PASS' indicates that a sample passed the quality filters. In the event where all samples pass the quality thresholds for a given variant, the FT flag is not listed for any of the samples. In previous versions of BioBin, only variants for which FT flags were not present in the VCF were binned, resulting in the erroneous exclusion of a potentially high fraction of variants. BioBin 2.3 properly bins variants with sample-level FT flags, given that they pass the appropriate quality filters.

2.3 Novel features

To optimize the functionality of BioBin, novel features allowing for multiple phenotype analysis as well as automatic sample processing have been developed. To facilitate simultaneous analysis of multiple traits, a user can now specify the number of parallel threads BioBin should use when generating bins. This expedites computation time when performing a rare variant phenome-wide association analysis (PheWAS). For example, an analysis with 9 K subjects, sequenced for 82 genes, and tested for association with 8 traits using SKAT took 6 min (4.9GB) using 6 threads, while analysis without parallelization took 14 min (1.3GB) to complete. BioBin 2.3 also easily facilitates the inclusion or exclusion of samples from an analysis, thus avoiding tedious pre-processing tasks. Exclusion of samples occurs during reading of the VCF and thus reduces memory overhead in a similar way as if the user provided an already subset VCF. Another related new option removes samples with missing phenotypes from an analysis ensuring that the loci and variants contributing to a bin are solely based on the samples for which phenotype information is available.

3 Conclusions

BioBin was originally developed to perform biologically based binning of rare variants. In its newest 2.3 release, BioBin's framework has been significantly expanded by incorporating statistical tests, upgrading the binning algorithm, and adding novel features to optimize analysis. Statistical test implementation upgrades the status of BioBin from that of a binning method to a tool for complete rare variant analysis, while maintaining the software's customizable nature. For example, a user can run BioBin solely for variant binning, or easily incorporate alternate statistical tests. The addition of novel features, such as on-the-fly sample dropping and multi-phenotype capabilities, helps streamline analysis. Also, keeping up with VCF changes in the representation of spanning deletions, BioBin 2.3 provides the option of counting '*' alleles as referent or missing to prevent variant count inflation. This expands on other tools which do not process '*' alleles nor do they handle multi-allelic sites. Limitations of BioBin include the implementation of fewer statistical tests as compared with other tools, like rvtests and EPACTS, which allow for more testing options (Supplementary Table 1). Future releases will focus on the addition of other statistical tests as well as fine mapping approaches. In conclusion, BioBin 2.3 is an open source, customizable tool that offers automated biological binning and association testing of rare variants. BioBin 2.3 software, a detailed user manual, vignette, and test examples can be accessed freely at http://www.ritchielab.com/software/biobin-download.

Acknowledgments

The authors would like to acknowledge Carrie B. Moore, Sarah A. Pendergrass, Shefali Setia, Anurag Verma, Navya Josyula and Manu Shivakumar for their input.

Funding

This work has been supported by NIH [AI116794 and HG008679]; and Pennsylvania Department of Health [#SAP 4100070267] (partly funded).

Conflict of Interest: none declared.

References

Basile,A.O. *et al.* (2016) Knowledge driven binning and PheWAS analysis in Marshfield Personalized Medicine Research Project using BioBin. *Pac. Symp. Biocomput.*, **21**, 249–260.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Kim,D. *et al*. (2015) Binning somatic mutations based on biological knowledge for predicting survival: an application in renal cell carcinoma. *Pac. Symp. Biocomput*., **2015**, 96–107.

Lee,S. *et al*. (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet*, **95**, 5–23.

McDonagh,E.M. *et al*. (2011) From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark. Med*., **5**, 795–806.

Moore,C.B. *et al*. (2013a) BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. *BMC Med. Genomics*, **6**, S6.

Moore,C.B. *et al*. (2013b) Using BioBin to explore rare variant population stratification. *Pac. Symp. Biocomput*., 332–343.

Moore,C.C.B. *et al*. (2016) A biologically informed method for detecting rare variant associations. *BioData Min*, **9**, 27.

NCBI Resource Coordinators (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **41**, D8–D20.

Pendergrass,S.A. *et al*. (2013) Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData Min*., **6**, 25.

Wu,M.C. *et al*. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet*., **89**, 82–93.