BioData Mining

SOFTWARE ARTICLE Open Access

A biologically informed method for detecting rare variant associations

Carrie Colleen Buchanan Moore[1†], Anna Okula Basile[2†], John Robert Wallace[3], Alex Thomas Frase[3] and Marylyn DeRiggi Ritchie[2,3*]

* Correspondence:
marylyn.ritchie@psu.edu
Carrie Colleen Buchanan Moore and Anna Okula Basile are sharing first authorship.
[†]Equal contributors
[2]Department of Biochemistry and Molecular Biology, Center for Systems Genomics, The Pennsylvania State University, University Park, PA 16802, USA
[3]Biomedical and Translational Informatics, Geisinger Health System, Danville, PA 17821, USA
Full list of author information is available at the end of the article

Abstract

Background: BioBin is a bioinformatics software package developed to automate the process of binning rare variants into groups for statistical association analysis using a biological knowledge-driven framework. BioBin collapses variants into biological features such as genes, pathways, evolutionary conserved regions (ECRs), protein families, regulatory regions, and others based on user-designated parameters. BioBin provides the infrastructure to create complex and interesting hypotheses in an automated fashion thereby circumventing the necessity for advanced and time consuming scripting.

Purpose of the study: In this manuscript, we describe the software package for BioBin, along with type I error and power simulations to demonstrate the strengths and various customizable features and analysis options of this variant binning tool.

Results: Simulation testing highlights the utility of BioBin as a fast, comprehensive and expandable tool for the biologically-inspired binning and analysis of low-frequency variants in sequence data.

Conclusions and potential implications: The BioBin software package has the capability to transform and streamline the analysis pipelines for researchers analyzing rare variants. This automated bioinformatics tool minimizes the manual effort of creating genomic regions for binning such that time can be spent on the much more interesting task of statistical analyses. This software package is open source and freely available from http://ritchielab.com/software/biobin-download

Background

Recent advances in sequencing technology and drastic decreases in cost have facilitated the generation of a prolific amount of sequence data. This has presented an opportunity for the investigation of low frequency and rare sequence variants beyond traditional genome-wide association (GWA) based approaches. Rare variants have recently been implicated in multifactorial conditions ranging from neurodegenerative diseases like Alzheimer's and Parkinson's disease, to metabolic disorders, such as obesity, and various cancers, including both prostate and lung cancer [1–6]. Elucidating the influence of rare variants on common diseases may expand our understanding of the heritability of complex traits, and it may eventually provide information that is useful to clinical patient care through the implementation of personalized, preventive practices.

Even with increased data availability, progress toward understanding rare genomic variation and its association to common human disease lags behind technological sequencing advances. Scientists are hindered in exploiting these advances because strategies for analyzing these data are underdeveloped. The growing disparity in rapidly advancing data collection versus slowly developing data analysis methods mandates a more concerted research effort to develop the necessary analytical tools for successful interpretation of genetic and biological data. Tools designed specifically for rare and low-frequency variant analysis require special considerations as these variants are individually uncommon, and often statistically underpowered for detecting phenotypic association [7, 8]. Also, the large sample size requirements may be prohibitive [9]. To increase the composite allele frequency and analyze smaller sample sizes, collapsing or binning methods are commonly utilized. Collapsing methods aggregate variants into a single genetic variable, which can then be used for subsequent statistical analysis, thereby reducing the number of degrees of freedom and also improving power in the analysis.

Many previous strategies developed for rare variants have focused on the statistical analysis of a pre-defined region rather than how to best group variants in an informative manner. Agnostic or un-informed binning approaches can often lead to a decrease in power when there are variants with different directions of effect or too many neutral variants that mitigate the signal. The most successful collapsing method groups variants likely to have an impact on the function of a specific gene or genomic unit and compares the variant distribution or composite genetic score distribution across the trait of interest.

BioBin [10–12] is a novel bioinformatics tool developed for the multi-level binning of rare variants using a biological knowledge-driven framework. BioBin collapses variants into user-designated biological features such as genes, pathways, evolutionary conserved regions (ECRs), protein families, regulatory regions, and others. Further, BioBin provides the infrastructure to create complex and interesting hypotheses in an automated fashion thereby circumventing the necessity for advanced and time consuming scripting. Simulation testing highlights the utility of BioBin as a fast, comprehensive and expandable tool for the biological binning and analysis of low-frequency variants in sequence data. While multiple biological applications of BioBin have previously been described [10–13], the manuscript herein concentrates on the software features, specifications and various analysis options within the BioBin package. We focus on presenting a comprehensive description of the capabilities of BioBin to provide a resource for users to tailor binning analyses to their specific hypotheses. Additionally, we demonstrate the utility of this software through type I error and power simulations. The BioBin software package has the capability to transform and streamline analysis pipelines for researchers analyzing rare variants in DNA sequencing data. This automated bioinformatics tool minimizes the manual task of curating biologically-relevant regions for binning, such that efforts can instead be spent on subsequent statistical analyses. This software package is open source and freely available from http://ritchielab.com/software/biobin-download.

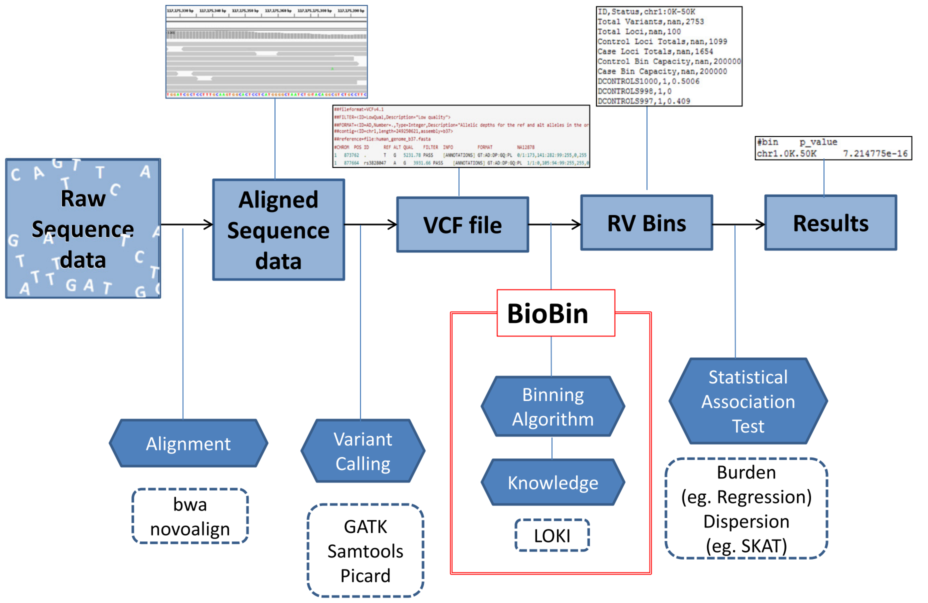
Implementation

BioBin is a unified command line bioinformatics tool for the biologically-inspired binning of rare variants. The novelty of BioBin is the automated multi-level binning process, rather than a focus on a particular statistical test. BioBin frees users from the tedious task of manually curating biologically important regions from multiple sources by using information from publicly available resources. The role of BioBin in a typical rare variant analysis pipeline is illustrated in Fig. 1. BioBin accepts VCF files and utilizes an internal binning algorithm in conjunction with biological information from an internal repository known as LOKI (described in a subsequent section). The algorithm will bin sequence variation in user-selected biologically-defined boundaries. The user then has the freedom to choose a specific statistical test for association.

BioBin resource requirements

BioBin is a stand-alone command line application written in C++ that relies on a locally built Library of Knowledge Integration (LOKI) database to create knowledge-based bins. Source distributions are available for Mac and Linux operating systems and require minimal prerequisites to compile. The BioBin distribution includes tools that allow the user to create and update the LOKI database by downloading information directly from source websites. BioBin is open-source and publicly available for download on the Ritchie lab website (https://ritchielab.com/software/biobin-download).

To evaluate the computational requirements of BioBin, we randomly selected a number of variants and a number of individuals from the 1000 Genomes Project Phase I low coverage data [14] and applied a BioBin gene binning analysis to the resulting dataset. Because the minor allele frequency dramatically impacts the selection of variants to be binned, we set parameters to include all variants, regardless of rarity, to produce consistent results. Over 10 replicates, Fig. 2 shows that bin generation is highly correlated to the number of loci (or genomic positions) in the study and both the number of loci and bin generation drive the memory requirements. The number of



```
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the or
##contig=<ID=chr1,length=249250621,assembly=b37>
##reference=file:human_genome_b37.fasta
#CHROM  POS ID        REF ALT QUAL     FILTER  INFO          FORMAT          NA12878
1   873762  .         T   G   5231.78 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
1   877664  rs3828047  A   G   3931.66 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
```

```
ID,Status,chr1:0K-50K
Total Variants,nan,2753
Total Loci,nan,100
Control Loci Totals,nan,1099
Case Loci Totals,nan,1654
Control Bin Capacity,nan,200000
Case Bin Capacity,nan,200000
DCONTROLS1000,1,0.5006
DCONTROLS998,1,0
DCONTROLS997,1,0.409
```

```
#bin     p_value
chr1.0K.50K      7.214775e-16
```

Fig. 1 Rare variant analysis pipeline. General rare variant analysis pipeline starting with raw sequence data and ending with association analysis results. BioBin accepts VCF files and is able to bin variation in biologically informed boundaries using an internal biological knowledge biorepository, LOKI. BioBin's output easily facilitates use of various statistical tests

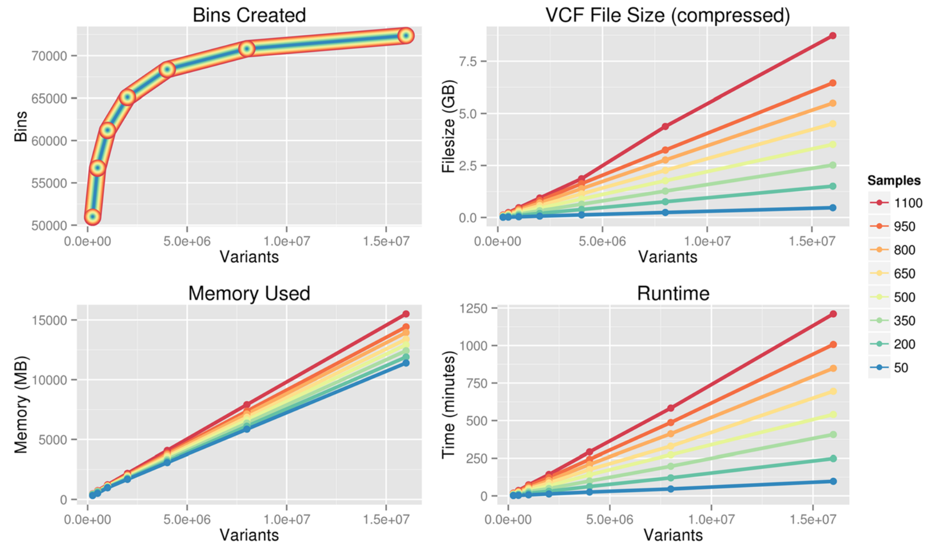


Fig. 2 BioBin resource requirements. BioBin resource requirements for varying numbers of study variants and sample sizes. The number of variants is the primary driver of needed resources, with the number of variants increasing the runtime due to the size of the input file

individuals in a study does not have a large impact on resource requirements, but does increase the size of the input VCF file and thus the time it takes BioBin to read the input VCF file. Even with large datasets, BioBin can be run relatively quickly without access to specialized computer hardware or a computing cluster; however, the number of low frequency variants to be binned is the primary driver of memory usage. Running a gene-based analysis of targeted exome capture of 82 pharmacogenes for 8194 samples [15], BioBin took approximately 10 min and 150 MB of RAM using a single core of an Intel Xeon E5-2670 processor. Using linear extrapolation based on the size of the target, we expect that a gene-based whole exome analysis of a similarly sized population would take approximately 6 h and 6GB of memory.

BioBin software features

Library of Knowledge Integration (LOKI)

BioBin relies on the Library of Knowledge Integration (LOKI), which integrates multiple databases providing a comprehensive biological knowledge platform for variant binning [16]. LOKI is a database that contains biological information from resources including the National Center for Biotechnology (NCBI) dbSNP and gene Entrez [17], Kyoto Encyclopedia of Genes and Genomes (KEGG) [18], Reactome [19], Gene Ontology (GO) [20], Protein families database (Pfam) [21], NetPath- signal transduction pathways [22], and others. Figure 3 provides a complete list of databases within LOKI. LOKI provides a standardized interface and terminology to disparate sources, each containing individual means of representing data [16, 23]. The four main concepts used in LOKI are *position*, *region*, *group*, and *source. Position* refers to the chromosome and base-pair position of single variants, such as single nucleotide variants (SNVs). A *region* represents any genomic segment with a start and stop position including genes, copy number variants (CNVs), insertions or deletions, and evolutionary conserved

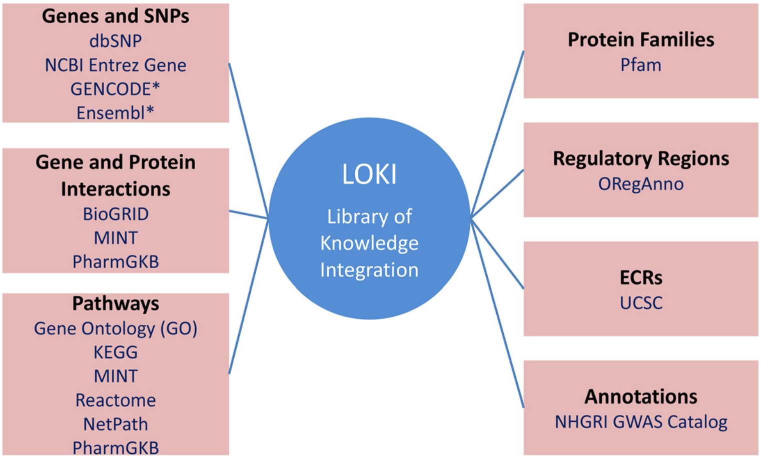


Fig. 3 LOKI. BioBin collapses variants into biological features by consulting LOKI, an internal biorepository that integrates multiple resources from the public domain. All databases within LOKI are listed below their corresponding feature category. Databases annotated with an asterisks (*) represent sources that will be present in the version 3 release of LOKI

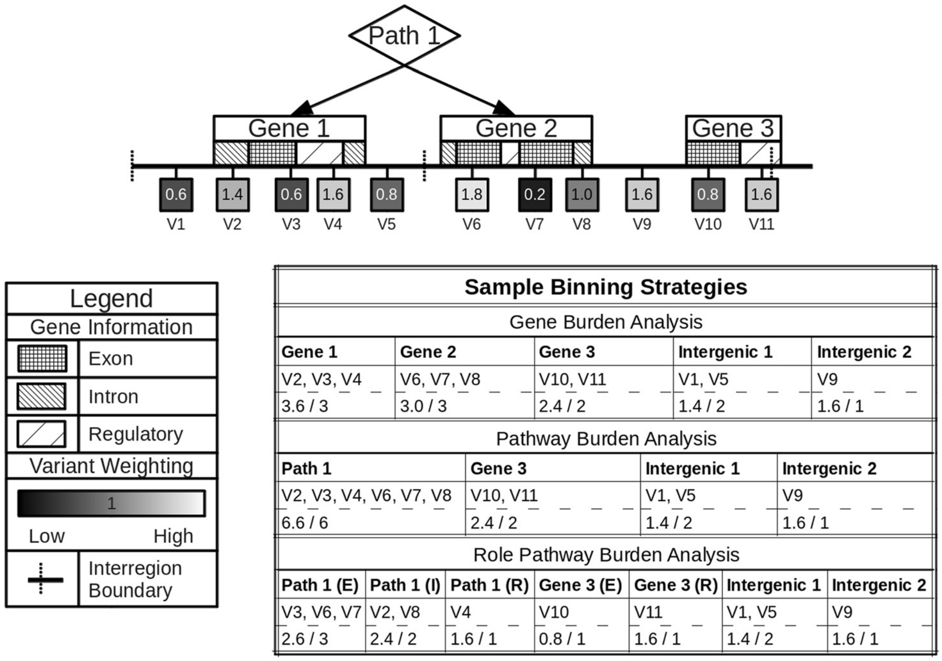
regions (ECRs). *Sources* are the external databases compiled in LOKI that contain *groups* of interconnected information, thus organizing the data in a standardized manner. For example, BioGrid ID:468346 defines a *group* from the BioGrid data *source* which contains the following *regions*: HMGB1P1, CTCFL, and PRMT7.

LOKI is implemented in SQLite, a relational database management system, which does not require a dedicated database server. A system initially building LOKI should have approximately 100GB of disk storage available for the LOKI database file, the LOKI source data, and space for python installer scripts. An updater script will automatically process and combine information from the various sources into a single database file (some of the temporary files are removed during this process). Once the build is complete, the LOKI database file required to run BioBin will be under 25GB. The script to build LOKI is open source, publicly available on the Ritchie lab website, and is included with the BioBin software. Users with knowledge of relational databases can customize their LOKI database by including or excluding sources, providing additional sources, and updating source information as frequently as needed [16].

Multi-level binning and filtering

The novelty of BioBin is its ability to automate bin generation at multiple levels of biological knowledge into one streamlined analysis. Figure 4 provides example binning strategies using biological information in LOKI. Using hierarchical biological relationships and optional functional or role information, BioBin can create many variant combinations to bin. As a standard in the current iteration of LOKI, NCBI dbSNP and NCBI Entrez Gene have been selected as the primary sources of position and regional information due to the data quality, reliability, and clearly defined database schema. These sources also most closely correspond to the region and group IDs provided by other database sources integrated into LOKI.

In addition to binning variants based on knowledge, BioBin also provides an option to bin variants that do not associate with any available knowledge. These are known as inter-region bins, or if generated between gene features, intergenic bins. After feature selection using LOKI and/or external custom files, inter-region bins can be created



Sample Binning Strategies

Gene Burden Analysis

Gene 1	Gene 2	Gene 3	Intergenic 1	Intergenic 2
V2, V3, V4	V6, V7, V8	V10, V11	V1, V5	V9
3.6 / 3	3.0 / 3	2.4 / 2	1.4 / 2	1.6 / 1

Pathway Burden Analysis

Path 1	Gene 3	Intergenic 1	Intergenic 2
V2, V3, V4, V6, V7, V8	V10, V11	V1, V5	V9
6.6 / 6	2.4 / 2	1.4 / 2	1.6 / 1

Role Pathway Burden Analysis

Path 1 (E)	Path 1 (I)	Path 1 (R)	Gene 3 (E)	Gene 3 (R)	Intergenic 1	Intergenic 2
V3, V6, V7	V2, V8	V4	V10	V11	V1, V5	V9
2.6 / 3	2.4 / 2	1.6 / 1	0.8 / 1	1.6 / 1	1.4 / 2	1.6 / 1

Fig. 4 Binning strategies in BioBin. Alternate binning strategies using biological knowledge (gene information) and functional or role annotations (variant information). Three example binning strategies are shown: gene burden analysis, pathway burden analysis, and functional pathway burden analysis. Note the intergenic bins that collects variants fall outside of the binning strategy

using a configurable width parameter (in kb). These bins catch variants that do not fit into biologically defined feature types (see intergenic bin labels in Fig. 4). For example, if one were testing low frequency burden differences between two groups across genes, all variants in genes would be collapsed into respective gene bins, and variants outside of gene boundaries would be binned based on genomic location in intergenic regions.

Locus selection and models

The framework of a BioBin analysis is to determine biological features upon which data will be binned, such as genes, pathways or intergenic regions, and execute bin generation using LOKI. For locus binning, BioBin follows an allele frequency threshold approach using the non-major allele frequency (NMAF). NMAF is defined as 1 minus the frequency of the most common allele, and at biallelic markers, NMAF and minor allele frequency (MAF) are interchangeable. BioBin allows variants below a user-specified NMAF in the case or the control group to be binned, thereby facilitating the aggregation of both potential risk and protective variants. In order to alleviate increased Type I error, BioBin also gives an option to use the minimum of the NMAF in either case or control group as the value to test against the given NMAF threshold [11].

BioBin provides multiple disease model options for determining individual contribution in a bin. This includes additive, dominant, or recessive encoding allowing the user to test specific hypotheses using these inheritance patterns. The default option utilizes additive encoding, where each allele adds to an individual bin score.

Customization

The power of BioBin becomes apparent in the flexibility provided to the user, which makes the software applicable in a number of low frequency variant analysis pipelines.

In addition to the predefined biologically-informed binning strategies, BioBin allows for customized knowledge, adjustable multi-level feature types, filtering strategies and individual variant weighting.

LOKI contains diverse knowledge from many databases, which together provide variant details, region annotations, and group relationships. To accommodate a wide variety of analyses, the user can choose to include or exclude any source in LOKI. Additionally a user can expand on the predefined knowledge contained within this biorepository as LOKI specification and code are open source allowing the addition of desired database sources. For instance, users may specify additional knowledge through the use of plain text files that can define regions, group or variant weights, and roles. Examples of these input files are provided in the BioBin manual (https://ritchielab.com/software/biobin-download). As part of the customization available, BioBin also accepts custom role files, which contain single variant or region annotations. This file can be used to exclude or specifically include variants based on the results returned from an annotation tool such as Polyphen, SIFT, or SNPEff [24–26].

Variant weighting

To adjust statistical power in a rare variant analysis, BioBin provides the option of weighting loci according to the weighted sum statistic proposed by Madsen and Browning [27], in which the weight of a variant is inversely proportional to its MAF. Multiple weighting schemes are provided which use different populations to calculate these locus weights. For instance, in *control* weighting, weighting is calculated based only on the control population. This weighting represents an exact implementation of Madsen and Browning weighting [27]. Because determining allele rarity solely on the control population has been shown to potentially inflate type I error [28, 29], BioBin implements other weight models allowing the user a means by which to utilize variant weighting while controlling this error. In the *maximum* model, the weight is the maximum calculated for the case and control populations, while the *minimum* model uses the minimum weight in these populations. *Overall* weighting calculates the weight using the entire overall population, regardless of case or control status. The *overall* weighting scheme is nearly equivalent to the Madsen and Browning weighting implementation in SKAT [30, 31]. These methods will be equivalent in the circumstance where there are no cases, or there is completely missing case or control population for a given locus. Finally, BioBin can also incorporate custom weights based on the user's prior knowledge.

Simulations

Simulation testing was performed to evaluate type I error and power with the various weighting schemes within BioBin (control only, maximum, minimum, overall, and no weighting) using two standard statistical tests: logistic regression and the Wilcoxon rank sum test. SeqSIMLA2 [32] a tool commonly used to simulate unrelated case control sequence data for various genetic analyses, was used to generate all sequence files, which then served as input for BioBin in the present analyses. To generate a reference sequence for SeqSIMLA2, the 1000 Genomes Project Phase I VCF file was parsed to obtain allele frequencies for all gene regions in the autosomes specific to individuals of European descent. A customized python script was then used to randomly select sites from this allele frequency file and generate a reference sequence in the format required by SeqSIMLA2, a binary zip file with each row being a sequence and each column a site. This custom script can be found in Additional file 1. For each SeqSIMLA2 dataset simulation, a sequence file of 10,000 reference samples was created with the number of generated markers varying in relation to the biologically-based bin size being tested (specified under *Type I Error Analysis and Power Analysis*). SeqSIMLA2 was then run with simulation parameters specific to the analysis being performed (see Table 1 for parameters). The output plink files were converted to VCF, and a BioBin variant binning analysis was performed with a MAF cutoff of 5 %. This was followed by statistical analysis using logistic regression and the Wilcoxon rank sum test. Type I error and power were evaluated for each weighting method.

Type I error analysis

Parameters for the type I error simulation analysis are listed in the left pane of Table 1. Type I error was assessed by performing three different tests, each varying in the size of the biological bin, as we attempted to simulate datasets that roughly correspond to gene-level and pathway level analyses. The choice of size for gene-based simulations is largely debated, and we decided to test three different bin sizes to accommodate various binning analyses, and to explore the relationship between bin size and type I error. These tests include a 25 kb gene-sized bin (referred to as average gene) composed of 50 variants (standard deviation = 5), a large 100 kb gene-sized bin (referred to as XL gene throughout this work) composed of 200 variants (standard deviation = 5), and a pathway bin composed of 2–50 gene-sized bins, or 100–2500 variants (standard deviation = 5). We chose 50 variants to represent an average sized gene bin by consulting the autosomal variant site statistics reported by 1000 Genomes Project [14, 33] and calculating a rough estimate for the number of possible variants expected in 25 kb, an approximation for median gene size [34]. For each simulation, the specific number of variants was randomly determined. For example, each pathway dataset simulation could

Table 1 Simulation parameters. Parameters for the type I error analysis and the power analysis simulations performed using SeqSIMLA2

Testing parameter	Type I error analysis	Power analysis
Bin size assessed	*Gene-sized bin:* 25 kb (50 ± 10 variants) *XL_Gene sized bin:* 100 kb (200 ± 10 variants)	*Gene-sized bin:* 25 kb (50 ± 10 variants)
	Pathway sized bin: 2–50 gene-sized bins (100–2500 ± 10 variants)	
Number of simulations	1000	1000
Sample size	500 cases, 500 controls	500 cases, 500 controls
Disease prevalence	5 %	5 %
Number of causal variants	N/A	10
Odds ratio (OR)	1	1.25, 1.5,1.75, 2, 2.5, 3, 4, 5
Variant weighting	No weighting	No weighting
	Control only weighting	Control only weighting
	Minimum weighting	Minimum weighting
	Maximum weighting	Maximum weighting
	Overall weighting	Overall weighting
Statistical test	Logistic regression	Logistic regression
	Wilcoxon	Wilcoxon

contain anywhere from 100 to 2500 variants. Type I error was estimated with 1000 null dataset simulations for each bin size using an odds ratio (OR) of 1, and assessing significance with an α of 0.05 for both regression and Wilcoxon.

Power analysis

To assess the statistical power of each weighting method, power analyses were performed with 1000 simulations of an average sized 25 kb gene bin, containing 50 variants (standard deviation = 5), as described in the right pane of Table 1. For each simulation, 10 causal variants or disease sites were randomly selected in the binned locus. Eight independent simulation tests were performed for each weighting scheme in which the OR of the causal variants was varied as 1.25, 1.5, 1.75, 2, 2.5, 3, 4, and 5. Power was assessed for each of these OR analyses with logistic regression and Wilcoxon using a significance criteria of 0.05.

Results and discussion

BioBin is an innovative variant collapsing method that provides a flexible infrastructure for biologically informed variant binning adaptive to individual user needs. In this work, we evaluated four weighting schemes provided within BioBin: control, minimum, maximum and overall weighting, in addition to the no locus weighting option. These weighting methods were examined using two standard burden tests: regression and the Wilcoxon rank sum. While multiple studies have performed exhaustive comparisons of statistical tests for rare variant analyses [35–37], the focus of BioBin is to build versatile and biologically relevant bins rather than to implement a particular statistical analysis. BioBin can provide the necessary files for a user to implement his or her statistical test of choice; this provides the user with freedom to choose the statistical test that is most appropriate for his/her hypothesis. We chose to specifically focus on regression and the Wilcoxon rank sum test as these are very commonly used methods in rare variant analyses [27, 38–41].

Type I error analysis

Results of the type I error analysis using logistic regression and Wilcoxon for bins of all biological-based sizes are presented as quantile-quantile (QQ) plots in Figs. 5 and 6, respectively. Both figures are comprehensive plots combining simulation p-value results from the average gene, XL gene and pathway analyses. The simulation results indicate that weighting using only the control population (*CTRL_ONLY_weight*) drastically inflates the type I error in both tests examined. Similar to the observations made by Lemire [28] and Pearson [29] in which allele rarity based solely on the control population introduces a bias, we also observe that weight calculations using only this phenotypic class increase type I error. A variant selection bias is created since there is an upper limit for the frequency of variants in the controls, but there is no bound for variant frequencies in the case population. This error becomes even more inflated when the size of the bin is increased from that of an average gene to a pathway, as evident in the additional material, thus introducing a spurious correlation that can confound results. BioBin implements other weight models where frequency thresholds are established using cases and controls, thereby imposing an upper frequency bound in both phenotypic classes, providing the user a means by which to utilize variant weighting

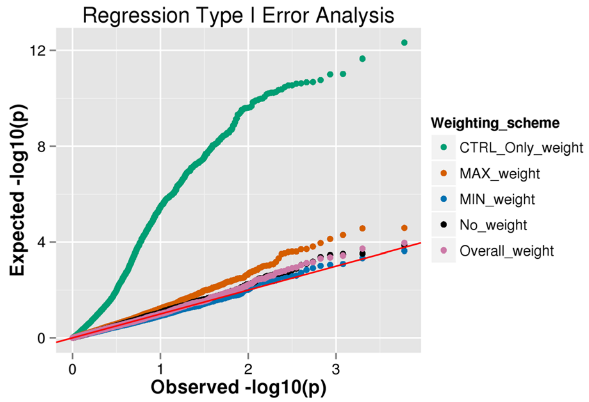


Fig. 5 Logistic regression type I error analysis. A comprehensive quantile-quantile plot for the type I error logistic regression analysis showing the combined p-value distribution from the average gene, XL gene, and pathway simulations. The different colors are representative of the various weighting schemes in BioBin that were analyzed

while controlling type I error. Of the weighting models tested, minimum weighting (*MIN_weight*) for the case and control populations presented the lowest type I error rate. The weights calculated from the overall allele frequency (*Overall_weight*), which is a common implementation of the Madsen and Browning test in current online tools, are mostly well controlled. Maximum weighting (*MAX_weight*) had a greater type I error rate than overall, minimum, and no weighting, but still a lower false positive rate when compared with control only estimates, especially when the bin size was increased.

No correlation between significance and bin size (except with control weighting)

Table 2 displays the type I error results for the different weighting schemes and different bin sizes. The majority of the weighting methods have a type I error controlled around 5 %, with the exception of maximum weighting which is closer to 8–9 % and control weighting which is dramatically higher. As seen in Figs. 5 and 6, control

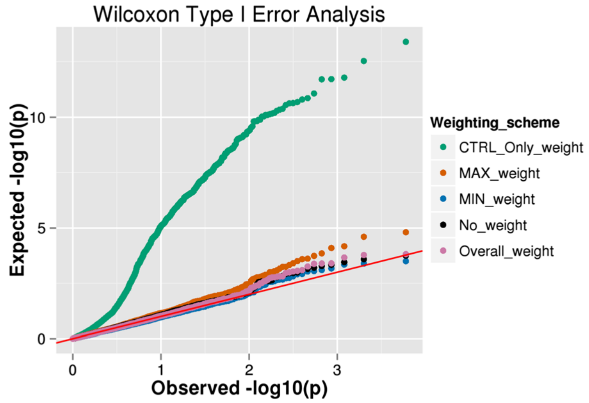


Fig. 6 Wilcoxon type I error analysis. A comprehensive quantile-quantile plot for the type I error Wilcoxon rank sum analysis showing the combined p-value distribution from the average gene, XL gene, and pathway simulations. The different colors are representative of the various weighting schemes in BioBin that were analyzed

Table 2 Type I error results. The Type I error simulation results displayed per BioBin weighting scheme tested, biological bin size assessed, and statistical analysis test

Statistical test	Bin size	Weighting scheme				
		Control only weight	Max weight	Min weight	No weight	Overall weight
Logistic	Gene	0.106	0.079	0.042	0.060	0.052
	XL Gene	0.343	0.096	0.042	0.062	0.057
	Pathway	0.847	0.090	0.040	0.066	0.059
Wilcoxon	Gene	0.064	0.061	0.052	0.060	0.057
	XL Gene	0.153	0.068	0.043	0.056	0.055
	Pathway	0.795	0.085	0.045	0.072	0.056

weighting yields a greatly inflated type I error, and further, Table 2 suggests that the amount of error is highly correlated with the specific bin structure. Further evidence of this can be seen in the supplemental section, where we show the QQ plots for each specific biological structure independently. In order to assess the role that the bin size plays in the false positive rate, we ran a logistic regression using the p-value of a simulated bin as the outcome and number of loci in a bin as the predictor. We chose logistic regression in this case because the outcome (p-value) is bounded between 0 and 1, but similar trends are seen using probit regression and ordinary least squares regressions as well (data not shown). The results, shown in Table 3, clearly indicate that for control weighting, an increase in the number of variants in a bin drastically increases the chance of a false positive finding. All other variant weighting strategies did not show any significant relationship between bin size and false positive rate. As discussed above, this trend is explained by the bias imposed when weighting variants using allele frequency thresholds calculated only from the control population, as no upper bound is imposed for case variant frequencies. However, when we impose bounds by weighting via the maximum, minimum or overall methods, we see a lower type I error rate.

Power analysis

The power analysis simulation results are shown in Figs. 7 and 8 for the logistic regression and Wilcoxon rank sum analysis, respectively. The most powerful BioBin weighting method was using the control weighting (*CTRL_ONLY_weight*). However, this weighting scheme has an inflated false positive rate, which is further magnified when the bin size is

Table 3 Correlation of bin size and significance. Using the control weighting, the larger bins result in a higher chance of a false positive finding, showing a correlation between bin size and p-value. All other weighting strategies have false positive rates independent of bin size

	Logistic		Wilcoxon	
	Beta (SE)	*p*-value	Beta (SE)	*p*-value
Control	−6.83e-3 (6.13e-4)	7.51e-29	−5.77e-3 (4.42e-4)	6.07e-39
Max	4.08e-5 (8.40e-5)	0.627	−4.75e-5 (8.39e-5)	0.572
Min	2.88e-5 (8.39e-5)	0.732	5.06e-5 (8.39e-5)	0.546
None	−2.43e-5 (8.38e-5)	0.772	−3.15e-5 (8.38e-5)	0.707
Overall	3.64e-5 (8.38e-5)	0.664	2.05e-5 (8.38e-5)	0.806

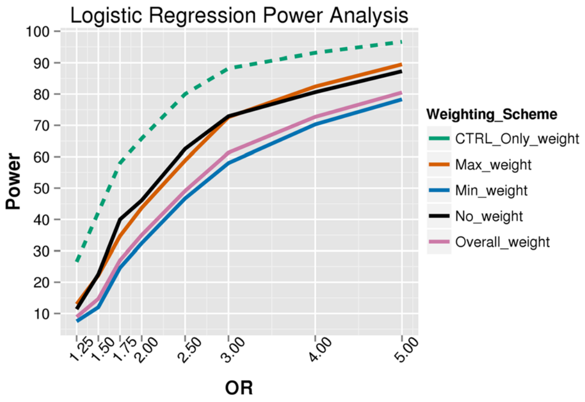


Fig. 7 Logistic regression power analysis. Statistical power estimates for multiple weighting methods assessed at varying odds ratios using logistic regression analysis. Results are based on 25 kb bins containing 50 (±10) variants with 10 of these representing causal variants. The different colors represent the various weighting schemes in BioBin that were analyzed. Control only weighting is presented with a dashed line to draw attention to the inflated type I error (presented in Fig. 6 and Additional files 2: Figure S1 and Additional file 3: Figure S2) of this weighting method

increased, as seen in Figs. 5 and 6 and Additional file 2: Figure S1. The most powerful BioBin weight with a more controlled type I error is the maximum weight, which has greater power with a Wilcoxon test than with logistic regression (however, this strategy has the second highest type I error rate around 8–9 %). To further evaluate sensitivity in the context of bin size for the minimum, maximum and overall weighting methods, we performed additional power analyses with 1000 simulations of a 100 kb sized gene bin, containing 200 total variants (standard deviation = 5) with 10 of these being causal or disease sites, and varying the OR of the causal variants from 1.25 to 5 (results not shown). Power of the 100 kb (200 variant) gene bins was assessed using logistic regression with a significance criteria of 0.05 and compared with that of the average 25 kb (50 variant) gene bins. Results of this comparison show consistently decreased power in the larger gene bins likely due to noise introduced by the addition of neutral variants, while maintaining the same number of causal variants as in the 25 kb bins. While future work will aim at testing

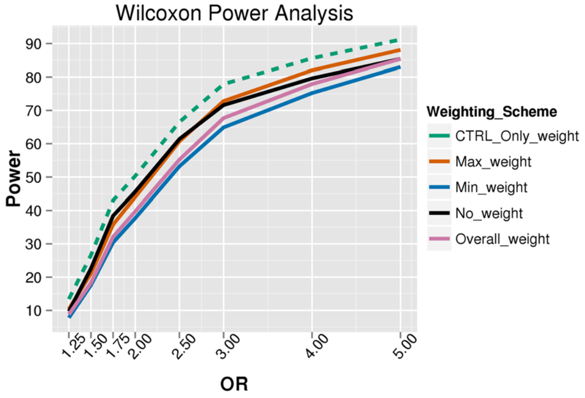


Fig. 8 Wilcoxon power analysis. Statistical power estimates for multiple weighting methods assessed at varying odds ratios using the Wilcoxon rank sum test. Results are based on 25 kb bins containing 50 (±10) variants with 10 of these representing causal variants. The different colors represent the various weighting schemes analyzed in BioBin. Control only weighting is presented with a dashed line to draw attention to the inflated type I error (presented in Fig. 5 and Additional file 2: Figure S1) of this weighting method

these observations by evaluating varied proportions of casual to neutral variants, early indications implicate this ratio of variants as the primary driver in sensitivity for a constant effect size (OR).

In the present simulations, the no loci weighting option in BioBin presents as statistically more powerful than both overall and minimum weighting. We believe this to be a result of the specific simulation parameters chosen for this analysis, and would likely be altered by the number of binned loci, the allele frequencies of the variants, the direction of the variant effect, and the sample size. Additional simulations were performed in an attempt to demonstrate the influence of chosen parameters on our simulation analyses. We performed comparable power analyses to those noted above, but restricted the selection of variants to only those having a MAF below 5 %, thereby causing all selected disease sites to be binned, and increased the number of casual variants to 20 (standard deviation = 5). The results of this analysis show that simulations without loci weighting had the lowest power across all tested ORs (1.25, 1.75, 2.5, 4 and 5) when compared with all other weighting methods. These results suggest that weighting approaches may have a larger influence on power when the selected disease sites are rare since different results were observed when disease sites with probabilities inversely proportional to the MAF are chosen. Overall, the power results are heavily influenced by simulation methodology, and future work will aim at performing a thorough sweep of simulation parameters and weighting methods in BioBin.

Future work

We have performed a preliminary study on incorporating select burden and dispersion-based statistical tests as well as multiple phenotype analysis capabilities into the framework of BioBin [12]. Future work will include comprehensive testing of burden and dispersion methods as well as dissemination of an updated BioBin software package, BioBin 2.2.0, with these additional features.

Conclusions

Overall, BioBin is a powerful and versatile tool for the knowledge-guided biological binning and analysis of low frequency variants in sequence data. BioBin uses a diverse repository of data from a multitude of public sources, and thereby circumvents the necessity of manually curating biologically important data for variant collapsing. BioBin provides users with a flexible and customizable framework to analyze sequence data and uncover novel associations with complex traits.

Additional files

Additional file 1: Script for generating reference sequence. Python script used to generate a reference sequence file for input into SeqSIMLA2 simulation software. The allele frequency file used in the script was obtained by parsing the protein coding regions of the autosomes in the 1000 Genomes Project VCF file. Additional specifications include the number of reference samples to generate and the number of markers to include in the reference file. (DOCX 14 kb)

Additional file 2: Figure S1. Logistic regression type I error per biological feature. QQ plots for the type I error logistic regression analysis showing the p-value distribution from the average gene (**a**), XL gene (**b**), and pathway (**c**) simulations. The different colors represent various BioBin weighting schemes analyzed. (PNG 170 kb)

Additional file 3: Figure S2. Wilcoxon type I error per biological feature. QQ plots for the type I error Wilcoxon Rank Sum analysis showing the *p*-value distribution from the average gene (**a**), XL gene (**b**), and pathway (**c**) simulations. The different colors represent various BioBin weighting schemes analyzed. (PNG 147 kb)

Funding
NIH grants LM010040, HL065962 and the Pennsylvania Department of Health Tobacco CURE Funds were used in the design, analysis, interpretation of data, and in writing the manuscript.

Availability of data and materials
Project name: BioBin
Project home page: http://ritchielab.com/software/biobin-download
Operating system: Linux
Programming language: C++
Other requirements: Boost Libraries for C++, version 1.42 or later; SQLite, version 3.5.4 or later; Python, version 2.7; suds for Python, version 0.4 or later; apsw for Python; Please see manual for most up to date requirements: http://ritchielab.com/files/RL_software/biobin-manual-2.2.pdf
License: GPL, version 3
Any restrictions to use by non-academics: GPL, version 3
Simulation data: Available upon request. The custom script used for generation of reference sequence files is included in the supplemental material.

Authors' contributions
Programming for BioBin was performed by JRW; writing of the code for LOKI was performed by ATF. CBM and MDR have made substantial contributions to conception and design of this software. Simulation analyses were performed by AOB. CBM, AOB, JRW, ATF and MDR participated in drafting and revising the manuscript and have given final approval of the version to be published.

Competing interests
The authors declare that they have no competing interests.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Duke University Medical Center, Duke General Surgery, Durham, NC 27710, USA. [2]Department of Biochemistry and Molecular Biology, Center for Systems Genomics, The Pennsylvania State University, University Park, PA 16802, USA. [3]Biomedical and Translational Informatics, Geisinger Health System, Danville, PA 17821, USA.

Received: 4 January 2016 Accepted: 18 June 2016
Published online: 30 August 2016

References

1. Cruchaga C, Chakraverty S, Mayo K, Vallania FLM, Mitra RD, Faber K, et al. Rare variants in APP, PSEN1 and PSEN2 increase risk for AD in late-onset Alzheimer's disease families. PLoS ONE. 2012;7, e31039.
2. Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer/'s disease. Nature. 2014;505:550–4.
3. Schulte EC, Fukumori A, Mollenhauer B, Hor H, Arzberger T, Perneczky R, et al. Rare variants in β-Amyloid precursor protein (APP) and Parkinson's disease. Eur J Hum Genet. 2015;23:1328–33.
4. Ramachandrappa S, Raimondo A, Cali AMG, Keogh JM, Henning E, Saeed S, et al. Rare variants in single-minded 1 (SIM1) are associated with severe obesity. J Clin Invest. 2013;123:3042–50.
5. Bronzetti E, Artico M, Forte F, Pagliarella G, Felici LM, D'Ambrosio A, et al. A possible role of BDNF in prostate cancer detection. Oncol Rep. 2008;19:969–74.
6. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. 2014;46:736–41.
7. Witte JS. Rare genetic variants and treatment response: sample size and analysis issues. Stat Med. 2012;31:3041–50.
8. Stitziel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. 2011;12:227.
9. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet. 2012;21:R1–9.
10. Moore CB, Wallace JR, Frase AT, Pendergrass SA, Ritchie MD. BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. BMC Med Genomics. 2013;6:S6.
11. Moore CB, Wallace JR, Wolfe DJ, Frase AT, Pendergrass SA, Weiss KM, et al. Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. PLoS Genet. 2013;9, e1003959.
12. Basile AO, Wallace JR, Peissig P, McCarty CA, Brilliant M, Ritchie MD. Knowledge driven binning and phewas analysis in marshfield personalized medicine research project using Biobin. Pac Symp Biocomput Pac Symp Biocomput. 2016;21:249–60.
13. Kim D, Li R, Dudek SM, Wallace JR, Ritchie MD. Binning somatic mutations based on biological knowledge for predicting survival: an application in renal cell carcinoma. Pac Symp Biocomput Pac Symp Biocomput. 2015;96–107.
14. Consortium T 1000 GP. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
15. Rasmussen-Torvik LJ, Stallings SC, Gordon AS, Almoguera B, Basford MA, Bielinski SJ, et al. Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. Clin Pharmacol Ther. 2014;96:482–9.

16. Pendergrass SA, Frase A, Wallace J, Wolfe D, Katiyar N, Moore C, et al. Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. BioData Min. 2013;6:25.
17. Resource NCBI. Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2013;41:D8–20.
18. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109–14.
19. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, et al. Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers. 2012;4:1180–211.
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
21. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30.
22. Kandasamy K, Mohan S, Raju R, Keerthikumar S, Kumar GSS, Venugopal AK, et al. NetPath: a public resource of curated signal transduction pathways. Genome Biol. 2010;11:R3.
23. Bush WS, Dudek SM, Ritchie MD. Biofilter: A Knowledge-Integration System for the Multi-Locus Analysis of Genome-Wide Association Studies. Pac Symp Biocomput Pac Symp Biocomput. 2009;368–79.
24. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. In: Haines JL, Korf BR, Morton CC, Seidman CE, Seidman JG, Smith DR, editors. Curr. Protoc. Hum. Genet. [Internet]. Hoboken: John Wiley & Sons, Inc; 2013. p. 7.20.1–7.20.41. [cited 2015 Oct 27]. Available from: http://doi.wiley.com/10.1002/0471142905.hg0720s76.
25. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11:863–74.
26. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
27. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5, e1000384.
28. Lemire M. Defining rare variants by their frequencies in controls may increase type I error. Nat Genet. 2011;43:391–2.
29. Pearson RD. Bias due to selection of rare variants using frequency in controls. Nat Genet. 2011;43:392–3. author reply 394–5.
30. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence Kernel association test. Am J Hum Genet. 2011;89:82–93.
31. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostat Oxf Engl. 2012; 13:762–75.
32. Chung R-H, Tsai W-Y, Hsieh C-H, Hung K-Y, Hsiung CA, Hauser ER. SeqSIMLA2: simulating correlated quantitative traits accounting for shared environmental effects in user-specified pedigree structure. Genet Epidemiol. 2015;39:20–4.
33. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
34. Fuchs G, Voichek Y, Benjamin S, Gilad S, Amit I, Oren M. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome Biol. 2014;15:R69.
35. Dering C, König IR, Ramsey LB, Relling MV, Yang W, Ziegler A. A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required. Front Genet [Internet]. 2014;5:323. [cited 2015 Jul 13]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4164031/.
36. Bacanu S-A, Nelson MR, Whittaker JC. Comparison of statistical tests for association between rare variants and binary traits. PLoS ONE. 2012;7, e42530.
37. Clarke GM, Rivas MA, Morris AP. A flexible approach for the analysis of rare variants allowing for a mixture of effects on binary or quantitative traits. PLoS Genet. 2013;9, e1003694.
38. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83:311–21.
39. Asimit JL, Day-Williams AG, Morris AP, Zeggini E. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered. 2012;73:84–94.
40. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010;34:188–93.
41. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95:5–23.