USING BIOBIN TO EXPLORE RARE VARIANT POPULATION STRATIFICATION*

CARRIE B. MOORE†
Center for Human Genetics Research, Vanderbilt University, 519 Light Hall
Nashville, TN 37232, USA
Email: carrie.c.buchanan@vanderbilt.edu

JOHN R. WALLACE‡
Center for Systems Genomics, Pennsylvania State University, 512 Wartik Laboratory
University Park, PA 16802, USA
Email: jrw32@psu.edu

ALEX T. FRASE
Center for Systems Genomics, Pennsylvania State University, 512 Wartik Laboratory
University Park, PA 16802, USA
Email: atf3@psu.edu

SARAH A. PENDERGRASS
Center for Systems Genomics, Pennsylvania State University, 512 Wartik Laboratory
University Park, PA 16802, USA
Email: sap29@psu.edu

MARYLYN D. RITCHIE
Center for Systems Genomics, Pennsylvania State University, 512 Wartik Laboratory
University Park, PA 16802, USA
Email: marylyn.ritchie@psu.edu

Rare variants (RVs) will likely explain additional heritability of many common complex diseases; however, the natural frequencies of rare variation across and between human populations are largely unknown. We have developed a powerful, flexible collapsing method called BioBin that utilizes prior biological knowledge using multiple publicly available database sources to direct analyses. Variants can be collapsed according to functional regions, evolutionary conserved regions, regulatory regions, genes, and/or pathways without the need for external files. We conducted an extensive comparison of rare variant burden differences (MAF < 0.03) between two ancestry groups from 1000 Genomes Project data, Yoruba (YRI) and European descent (CEU) individuals. We found that 56.86% of gene bins, 72.73% of intergenic bins, 69.45% of pathway bins, 32.36% of ORegAnno annotated bins, and 9.10% of evolutionary conserved regions (shared with primates) have statistically significant differences in RV burden. Ongoing efforts include examining additional regional characteristics using regulatory regions and protein binding domains. Our results show interesting variant differences between two ancestral populations and demonstrate that population stratification is a pervasive concern for sequence analyses.

* This project was funded by NIH grants LM010040, HL065962, CTSI: UL1 RR033184-01.
† This work is supported by F30AG041570 from the National Institute on Aging and Public Health Service award T32 GM07347 from the National Institute of General Medical Studies for the Medical-Scientist Training Program.
‡ This work is supported by a grant with the Pennsylvania Department of Health using Tobacco CURE Funds.

1. Introduction and Background

In the field of human genetics research, there has been increasing interest in the role of rare variation in complex human disease. This is in many ways a response to changing technology, but more importantly a response to the inability to completely explain heritability in common complex diseases and recognition of the true multifactorial mechanisms of genetic inheritance. It is believed that rare variants (RVs) likely have a larger effect size (compared to genome-wide association study (GWAS) findings) and can act alone, in concert with other RVs, or together with common variants. There is increasing evidence to support a role for RVs to contribute to common, complex disease. Recent studies on obesity, autism, schizophrenia, hypertriglyceridemia, hearing loss, complex I deficiency, age-related macular degeneration, kabuki syndrome, and type-1 diabetes implicate RVs with moderate effect sizes.[1–6]

Because of the frequency of RVs and thus the necessary sample size to gain reasonable power, association signals for RVs in a simple SNP-phenotype association study are harder to detect. Methods can be used to group the RVs and test for group association with disease status. Grouping, also known as binning or burden testing, better accounts for genetic heterogeneity and the possibility for multiple RVs to act in concert, which would have otherwise been overlooked in GWAS. Collapsing methods are popular for many reasons: to reduce the degrees of freedom in the statistical test, easy application to case-control studies (not limited to family transmission filtering), applicability to whole-genome data, and an accessible way to enrich association signals by combining RVs (often otherwise undetectable). Several collapsing methods have been published in the past five years.[2,7–14]

Our BioBin approach meets a critical need for an improved binning algorithm through the advantage of prior biological knowledge and potential cumulative effects of biologically aggregated RVs. BioBin requires the Library of Knowledge Integration (LOKI), which contains diverse prior knowledge from multiple collections of biological data. BioBin can be used to apply multiple levels of burden collapsing/testing, including: regulatory regions, evolutionary conserved regions, genes, and/or pathways without a need for an external feature file. Users can define the boundaries of a feature based on a specific hypothesis of interest; for example, is there a difference in RV burden in regions with known transcription factor binding sites between two groups? The adaptable design of BioBin and incorporation of prior biological knowledge provides the user with a flexible binning system and the opportunity to test a range of hypotheses.

While BioBin was specifically developed to investigate RV burden in traditional genetic trait studies, this tool is useful for exploring the natural distribution of RVs in ancestral populations. Rapid population growth and weak purifying selection has allowed ancestral populations to accumulate low frequency variants, many of which are deleterious and potentially causal to human disease.[15,16] These RVs exhibit ancestral heterogeneity and can be completely unique to a single population. To demonstrate the magnitude of population stratification in RVs, Tennessen et al. identified more than 500,000 single nucleotide variants (SNVs) using 15,585 protein-coding genes from 2,440 individuals. Of these SNVs, 86% had a MAF < 0.5% and 82% were population specific (European American or African American).[16] Others have documented differences between ancestral populations using gene drug targets[15] and ENCODE data.[9,17] A thorough

understanding of the distribution of RVs across populations will help uncover unknown demographic and evolutionary forces acting on the genome. Since RVs are likely essential in understanding the etiology of common complex traits, it is also critical to understand population stratification for the sake of sequence data analysis. The magnitude of population stratification (and consequential inflation of type I error) is not yet known and adequate methods to correct for stratification have not been developed.[18,19]

Herein we present the methodology of BioBin and the structure of LOKI that provides the prior knowledge for assignment of bins in BioBin. We have tested BioBin using data simulations specifying RVs and applied BioBin to European descent (CEU) and Yoruba (YRI) individuals from 1000 Genomes Project Phase I data. Our tests show BioBin is a flexible and effective method for biological knowledge directed binning of RV data and highlight the importance of investigating RV distribution differences across diverse populations.

2. Methods

2.1. *General framework*

The rare variant analysis occurs in two steps: first, BioBin generates bins based on user-defined parameters and information from LOKI; second, the user applies an appropriate statistical association test. To bin, the user can change options in the configuration file to select certain database sources, adjust feature types, and/or configure the minor allele frequency (MAF) binning threshold. The MAF binning threshold determines the allele frequency limit under which variants are binned. For example, if the threshold is 0.03, a locus with MAF 0.04 would not be included in a bin but a locus with a MAF of 0.029 would be included. The minor allele at a given locus is determined from the second most frequent allele in the control group. For a biallelic locus, this is always the rarer allele. For a triallelic locus, the MAF reported by BioBin is calculated from the second most frequent allele, but all rare alleles are binned. Common alleles (including loci with low frequency variants above the binning threshold) are not binned and are not considered in this analysis, but could be combined with RV bins in subsequent statistical analyses. An example of major and MAF inclusion/exclusion from a single group is shown in Table 1.

Table 1. Variant binning with a MAF binning threshold < 0.05

Major Allele (AF)	Minor Allele(s) (AF)	MAF	Variants Binned
C: 0.97	T: 0.03	0.03	T
T: 0.80	A: 0.16, G: 0.04	0.16	
G: 0.95	C: 0.03, T: 0.02	0.03	C, T

Although the major and minor alleles are designated by frequency in the control group, RVs in the case group also contribute to the variants binned. To simplify, “rareness” is calculated separately for cases and controls. If a variant is considered rare (allele frequency less than the MAF bin threshold) in either group, it will contribute to the bin. In this way, we are not only accumulating risk variants (higher frequency in cases than controls) but also potentially protective variants (lower frequency in cases than controls). This reduces the number of false positive bins and reduces the correlation between bin size and significance.

2.2. *Software*

2.2.1. *BioBin*

BioBin is a standalone command line application written in C++ that uses a prebuilt LOKI database. Source distributions are available for Mac and linux operating systems and require minimal prerequisites to compile. Included in the distribution are tools that allow the user to create and update the LOKI database by downloading information directly from source websites. The computational requirements for BioBin are quite modest; for example, during testing, a whole-genome analysis including 185 people took just over two hours using a single core on a cluster (Intel Xeon X5675 3.06 GHz processor). However, because the vast amount of data included in the analysis must be stored in memory, the requirements for memory usage can be high; the aforementioned whole-genome analysis required approximately 13 GB of memory to complete. Even with large datasets, BioBin can be run quickly without access to expensive and specialized computer hardware or a computing cluster. The number of rare variants is the primary driver of memory usage.

2.2.2. *Library of Knowledge Integration (LOKI) Database*

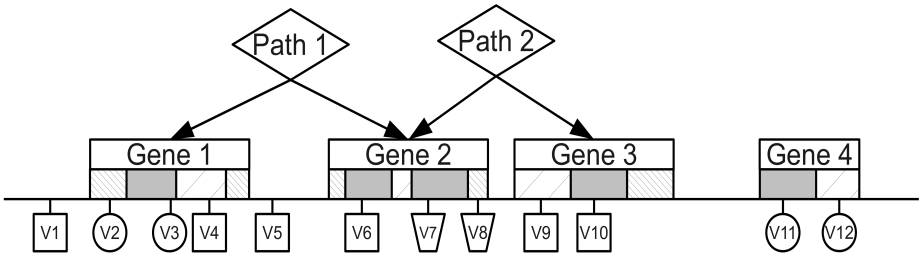
Harnessing prior biological knowledge is a powerful way to inform collapsing feature boundaries. BioBin relies on the Library of Knowledge Integration (LOKI) for database integration and boundary definitions. LOKI contains resources such as: the National Center for Biotechnology (NCBI) dbSNP and gene Entrez database information,[20] Kyoto Encyclopedia of Genes and Genomes (KEGG),[21] Reactome,[22] Gene Ontology (GO),[23] Protein families database (Pfam),[24] NetPath - signal transduction pathways,[25] Molecular INTeraction database (MINT),[26] Biological General Repository for Interaction Datasets (BioGrid),[27] Pharmacogenomics Knowledge Base (PharmGKB),[28] Open Regulatory Annotation Database (ORegAnno),[29] and information from UCSC Genome Browser about evolutionary conserved regions.[30]

LOKI is used as a means to provide a standardized interface and terminology to disparate sources each containing individual means of representing data. The three main concepts used in LOKI are *positions*, *regions* and *groups*. The term *position* refers to single nucleotide polymorphisms (SNPs), single nucleotide variants (SNVs) or RVs. The definition of *region* can be applied to a broader scope of biology. Any segment with a start and stop position can be defined as a region, including genes, copy number variants (CNVs), insertions and deletions, and evolutionary conserved regions (ECRs). *Sources* are databases (such as those listed above) that contain *groups* of interconnected information, thus organizing the data in some way.

LOKI is implemented in SQLite, a relational database management system, which does not require a dedicated database server. The user must download and run installer scripts (python) and allow for 10-12 GB of data from the various sources. The updater script will automatically process and combine this information into a single database file (~ 6.7 GB range). A system running LOKI should have at least 50 GB of disk storage available. LOKI runs locally wherever needed.

2.3. *Binning approach*

We chose NCBI dbSNP and NCBI Entrez Gene as our primary sources of position and regional information due the quality and reliability of the data, and clearly defined database schema. Intergenic regions are bins generated by BioBin to catch variants that do not fit into the user-defined feature types. For example, if one were testing RV burden differences between cases and controls across genes, all variants in genes would be collapsed into respective gene bins, and variants outside of gene boundaries would be binned corresponding to the intergenic regions. BioBin provides an option to generate intergenic bins of a user-specified size to catch intergenic variants. Figure 1 shows an example of RV binning strategies; different knowledge applied to the same variants produces alternate bins.



Legend	
Gene Information	
	Exon
	Intron
	Regulatory
Variant Information	
	Protective
	Detrimental
	Neutral (Unk)

Sample Binning Strategies

Gene Burden Analysis

Gene 1	Gene 2	Gene 3	Gene 4	Intergenic
V2, V3, V4	V6, V7, V8	V9, V10	V11, V12	V1, V5

Pathway Burden Analysis

Path 1	Path 2	Gene 4	Intergenic
V2, V3, V4, V6, V7, V8	V6, V7, V8, V9, V10	V11, V12	V1, V5

Functional Pathway Burden Analysis

Path 1 (+)	Path 1 (-)	Path 1 (N)	Path 2 (-)	Path 2 (N)	Gene 4 (+)	Intergenic (N)
V2, V3	V7, V8	V4, V6	V7, V8	V9, V10	V11, V12	V1, V5

Figure 1. Binning strategies for three example burden analyses.

2.4. *Statistical analysis*

BioBin is a bioinformatics tool used to create new feature sets that can be analyzed in subsequent statistical analyses. We believe that statistical tests can and should be chosen according to the hypothesis being tested, the question of interest, or the type of data being tested. There are explicit situations that require the use of regression analysis (linear, logistic, polytomous), Fisher's exact test, permutation of unique statistical test, etc. For this reason, no specific statistical test is implemented into BioBin. Unless otherwise noted, the results presented herein were calculated using a Wilcoxon 2-sample rank sum test implemented in the R statistical package.[31] There was no need for adding covariates to the model and Wilcoxon provides simple implementation and interpretation. All individuals (CEU and YRI) are ranked according to number of variants they individually contribute to a bin (variants must be under binning threshold). Using a simple model, we assume the genotypes are independent and normally distributed.

2.5. *Data simulation strategy to assess type I error*

To test BioBin, genetic data was simulated using a forward time simulator, simuPOP.[32] We used a constant distribution for the selection coefficient and a mutation rate of 1.8×10^{-8} per nucleotide per generation. The population sizes were N_e = 8100, 8100, 7500, and 10000 with 5000 generations, 10 generations, and 370 generations respectively. A 10kb region and 50kb of genetic data were simulated using the standard parameters in the simuRareVariants.py script for simuPOP. This script simulates introduction and evolution of RVs and can allow complex fitness and selection modeling (http://simupop.sourceforge.net/cookbook/).

To generate a sample data set evaluating type I error, all individual's genotypes were generated by randomly choosing two haplotypes from a haplotype pool. This process was repeated for three different scenarios: 1) sample size of 1000 individuals (500 cases and 500 controls) on a 10kb region, 2) sample size of 4000 individuals (2000 cases and 2000 controls) on a 10kb region, 3) sample size of 4000 individuals (2000 cases and 2000 controls) on a 50kb region. Phenotypes were randomly assigned to each individual to test the null hypothesis of no association between variants and disease status. The type I error was calculated as the proportion of the 10,000 replicates with a p-value <= 0.05. In this case, an error rate above 5% would indicate a higher false-positive test and an error rate lower than 5% would indicate a conservative test.

2.6. *1000 Genomes Project data: CEU and YRI comparison*

In a recent resequencing study of 202 drug targets, Nelson et al. reported the abundance of rare variants to be approximately 1 every 17 bases and most often population specific.[15] To further investigate population stratification, we used 1000 Genomes Project data. The project was started in 2008 with the mission to provide deep characterization of variation in the human genome. As of October 2011, the sequencing project includes whole-genome sequence data for 1094 individuals, and aims to sequence 2,500 individuals by its completion.[33]

We conducted a pairwise comparison of RV burden differences between two ancestry groups (YRI and CEU) of the 1000 Genomes Project (October 2011 release ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/). The data includes 87 CEU samples and 88 YRI samples. We implemented a minimum bin size of two variants and set the binning threshold to 0.03. We performed the following feature specific analyses:

A. Gene and intergenic regions
B. Pathways
C. Regulatory regions
D. Evolutionary conserved regions (ECRs)

The NCBI Entrez source provided gene start and stop positions to form gene bin boundaries for the gene and intergenic region analyses (A). Intergenic bins (50kb) were generated to "catch" variants not collapsed into other source-informed bins; in this case, intergenic bins collapsed variants not binned into gene region bins. For the pathway-based analysis (B), pathway and group information came from many LOKI sources and collapsed variants from all genes/regions in a specific pathway together in a bin. The regulatory region analyses (C) bin boundaries used in this analysis were from ORegAnno, a database of regulatory region annotations. For the evolutionary conserved region analysis (D), boundaries were calculated from PhastCons score output

downloaded from UCSC Genome Browser (http://genome.ucsc.edu/). There are three groups of ECRS available within the UCSC Genome Browser, the first group is derived from multiple alignments of 45 vertebrate genomes to the human genome, the second group is a set of placental mammals (32 placental Mammal genomes) aligned to the human genome, and the third group is a set of nine primates aligned with the human genome (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/). For each group, we calculated segments of the genome with 70% identity, a minimum length of 100bp, and allowed for 50bp gaps. These ECRs were clustered in bands according to the PhastCons output, which corresponded to an average of 13 ECRs per band. This was necessary since a single ECR is not large variable enough to generate a viable bin. In this paper, reported p-values have been corrected for multiple testing using a Bonferroni correction (number of generated bins in each analysis).

3. Results

3.1. *Type I error calculation*

It is important to investigate the level of type I error that might be present in any novel approach. Thus, using the script simuRareVariants.py from the simuPOP simulation algorithm, we simulated a 10kb genomic region with 31 RVs and 50kb genomic region with 154 RVs using the parameters described above in the methods section. Overall, 10,000 individuals were simulated, each with two haplotypes. We created populations by sampling the haplotypes, and generated 10,000 replicates of 1000 or 4000 individuals with balanced numbers of cases and controls. The threshold for significance was $p \leq 0.05$. We calculated the type I error rate as the number of replicates with Wilcoxon p-value less than or equal to 0.05 divided by the total number of replicates. The Wilcoxon 2-sample rank sum test seems to control the type I error in BioBin, but the false positive rate nominally increases as the sample size or bin size increases (see Table 2).

Table 2. Type I error calculation results.

Population Size	Simulated Region Size	Type I Error Rate
1000	10kb	0.0479
4000	10kb	0.0533
4000	50kb	0.0564

3.2. *1000 Genomes Project data: CEU and YRI comparison*

We tested BioBin using whole-genome ancestral data from 1000 Genomes Project using 87 CEU and 88 YRI individuals. There are considerably more variants in the YRI samples than in the CEU samples. Table 3 provides the total number of variants according to the Phase I generation of 1000 Genomes Project data for both populations. Of note, while there is only one more YRI individual compared to the number of CEU individuals, there is almost a 7 million variant difference between the two groups. Figure 2 shows a density function, which indicates the density of variants at each MAF; overall, there is a higher density of low frequency variants in YRI.

Table 1. 1000 Genomes Project Phase I data characteristics for CEU and YRI

Population	Number of Variants	Number of People
CEU	11,198,921	87
YRI	18,022,152	88

Using a MAF binning threshold of 0.03, we binned genes and intergenic regions, pathways, regulatory regions, and evolutionary conserved regions as described above in the methods. The top result from each feature in these four analyses (labeled A-D) is shown in Table 4.[34]

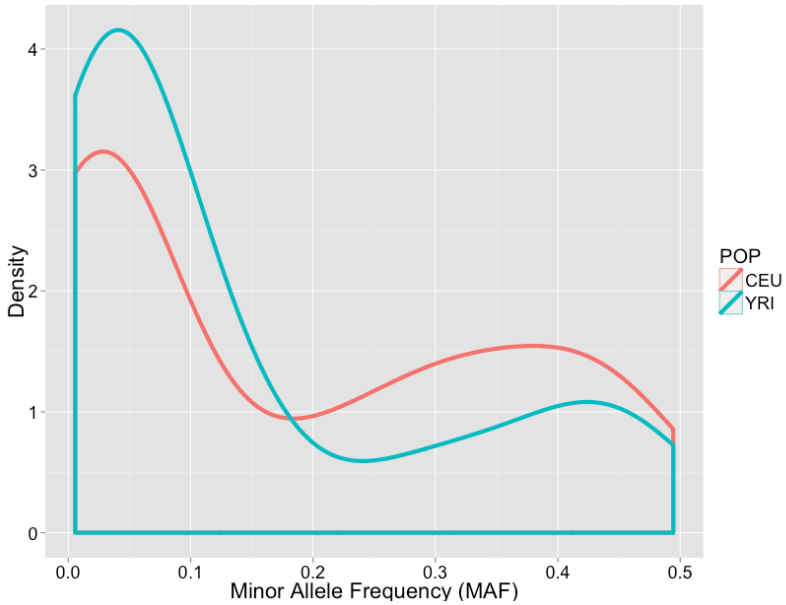


Figure 2. Minor allele frequency density distribution for CEU (red) and YRI (green)

Table 2. Top result from each feature across the four analyses (A-D)

	Feature	Top Bin	Adj. p-val	Annotation/ Location	Function
A	Genes	*CTXN2*	7.18×10^{-29}	Chr5:48483867-48495951	Cortexin 2-Integral to membranes
	Intergenic regions	chr15.638	5.13×10^{-28}	Chr15:31900000-31950000	3' to OTUD7A, a protease that cleaves ubiquitin
B	Pathways	PF11057	1.76×10^{-29}	Cortexin protein family	Expressed in kidney and brain, involved in intra and extracellular signaling
C	ORegAnno	OREG0003872	1.83×10^{-32}	Chr5:142124712-142125230	Transcription Factor Binding site, expressed in the heart
D	ECR-vertebrates	Chr5:33951654-33951791	3.24×10^{-33}	SLC45A2	Melanocyte differentiation antigen. Substance transport for melanin biosynthesis.
	ECR-placental Mammals	Chr5:33951651-33951791	3.24×10^{-33}	SLC45A2	
	ECR-primates	Chr15:48426444-48426724	1.94×10^{-33}	SLC24A5	Cation exchanger involved in pigmentation, melanosome ion transport

Next, we evaluated the prevalence of significant RV differences between CEU and YRI data. Using the Bonferroni corrected threshold of significance specific for each analysis, we calculated the proportion of bins that were significant. The results are shown in Figure 3.[34]

The height of each bar represents the total number of bins in each feature type; the dark blue indicates the proportion of significant bins. For example, 9.10% of the bins generated from ECR-primate multiple alignment comparison was significant after correction for multiple testing (which accounted for all tests performed in analysis D).

There are a surprising number of significant bins in each feature, but this can be explained by the difference in total number of variants between CEU and YRI. The total number of variants binned by BioBin using a MAF-binning threshold of 0.03 was 16,145,128 variants. Of these, 65.5% were private to YRI ancestral population.

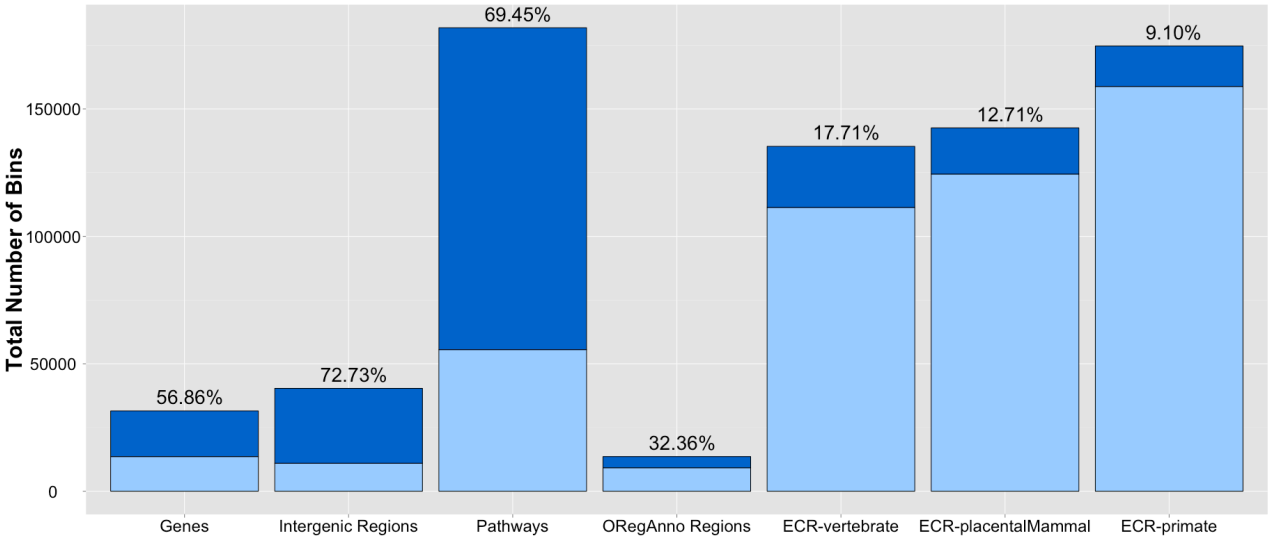


Figure 3. CEU-YRI pairwise comparison. Dark blue indicates the proportion of significant bins.

4. Discussion

4.1. *Type I error calculation*

As shown in Table 2, the Wilcoxon 2-sample rank sum test is slightly anticonservative in large population sizes and seems to worsen when more RVs are binned together. This is interesting since Li et al. reported that increasing the number of variants binned in a type I error simulation decreased the type I error rate using a collapsing approach and a Pearson χ^2 statistical test and others have reported conservative type I error rates using asymptotic statistical tests on relatively small sample sizes.[8,35]These methods were tested on simulated data with controlled RV allele frequencies and used different statistical tests, but highlights the importance and perhaps limitations of simulation testing. Although, the type I error seems to be well-controlled in this experiment, further investigation should be done to assess strictly how the RV allele frequency distribution affects type I error, calculate the type I error using additional sample population sizes and alternative statistical tests, and examine if the number of variants in a bin consistently inflate the false positive rate.

4.2. *1000 Genomes Project data: CEU and YRI comparison*

Using 1000 Genomes Project whole-genome data, we used BioBin to identify features (genes, intergenic regions, pathways, regulatory regions, and ECRs) with significant differences in rare RV burden between two ancestral populations. A population-genetics approach retains natural qualities of data (compared to simulated data) and incorporates case/control status according to ancestry group. Comparable approaches have been used by other groups.[9,17]

We compared multiple feature types between two ancestral populations from 1000 Genomes Project to highlight a known issue in genomic studies, population stratification. BioBin explored RV burden differences between CEU and YRI ancestral populations. In each RV burden test, there were a considerable number of statistically significant bins (after Bonferroni multiple testing correction). Table 4 shows the most significant bins for each feature type. The gene burden top result and the pathway burden top result corresponded to a Cortexin-2 gene and Cortexin pathway respectively. According to PFAM, this group of proteins is important for intracellular and extracellular signaling in the kidney and brain (http://pfam.sanger.ac.uk/family/PF11057). To our knowledge, Cortexin-2 has not been acknowledged in ancestry comparison studies. However, another protein in the Cortexin family was identified as a candidate gene for non-diabetic forms of end-stage renal disease in African Americans.[36] This is interesting since studies with admixed populations could contain a higher incidence of false positives due to RV population stratification and mixed ancestry.

We could not find biological interpretation for the significant intergenic RV burden differences on chromosome 15 or the transcription factor-binding site on chromosome 5. However, the ECR analyses highlighted *SLC45A2* and *SLC24A5*; both participate in pigmentation.

Mutation rates vary across the genome. They can vary according to specific sequence contexts, within regions on a chromosome, and between chromosomes.[37] While mutation rates are commonly studied between orthologous sequences, polymorphisms collapsed by regions within species can also provide interesting insight into evolutionary history and mutation. BioBin does not provide detailed sequence output to investigate mutation rate variation between CEU and YRI, but it does provide some information about higher rates of variation in regions (genes, intergenic regions, pathways, regulatory regions, and ECRs) and between chromosomes. The results in Figure 3 show an interesting trend between functional regions of the genome and variant tolerance. Approximately 57% of the gene bins had significant differences in RV burden, whereas approximately 73% of the intergenic region bins had significant differences in RV burden. There is some weak evidence that genes undergo adaptive evolution, which explains why regions in the genome with potential for highly deleterious mutations evolve lower mutation rates. There are two potential explanations: 1) additional level of repair of DNA damage in transcriptional active regions by transcription coupled repair (TCR), 2) approximately 3% of the genome is subject to negative selection, however it is estimated that functionally dense regions contain up to 20% sites under selection.[37,38] In this analysis, gene bins are inclusive of intronic regions, thus it would be interesting to break down the gene bins into intronic and exonic bins to see how the variant tolerance differs between coding and noncoding regions.

There are far fewer regulatory region bins, but there appears to be smaller proportion of significant differences between CEU and YRI compared to genes or intergenic regions. Again, perhaps mutations are less tolerated in these regions and we see overall less variability. ECRs have been long known to be conserved among species, and in this analysis they are also the features least likely to have variation between CEU and YRI. There is some debate about selection and functional significance in these regions. It is unknown what factors have the largest effect on mutation rates,[37] but it is possible that consistently low mutation rates in these sections have generated conserved regions throughout evolution.[38]

We found that over 65% of the variant loci in dataset were fixed in CEU individuals. This is not surprising since it is well known that individuals of African descent have more variation than individuals of other ancestral groups (see Table 3). This difference in rare variation is driving the high percentages seen in Figure 3. We should further investigate the effects of stratification in other ethnicities, and evaluate correction methods such as PCA and mixed models.[18,19]

5. Conclusion

There is a global health, scientific, and financial motivation for understanding the genetic etiology of common complex disease. It is imperative to consider genetic variants beyond common single nucleotide polymorphisms, as RVs may have larger phenotypic effects and can help us better comprehend the biology of a disease process. BioBin is a novel collapsing method that uses allele frequency data and biological information to bin RVs. It is unique because it is packaged with LOKI and is not coupled with any statistical method. Access to integrated biological knowledge (pathways, groups, interactions, ECRs, regulatory regions, etc.) is valuable to researchers that do not want to spend considerable effort to combine this knowledge manually. Freedom from implemented statistical methods provides users with the ability to apply association tests most appropriate for their data analysis. In general, for any given bin, statistical tests from other published collapsing methods can be applied to BioBin output. However, these other methods do not incorporate feature selection; therefore, the user must provide boundaries for each bin.

In this paper, we evaluated RV burden differences between CEU and YRI populations. Although population stratification is often considered in genomic analyses, to our knowledge, no previous studies have quantified the magnitude of RV burden differences across multiple features. From the ancestry comparison results, we learned RV burden differences among features showed a pattern consistent with current mutation rate theory but also highlighted the magnitude of RV stratification between CEU and YRI populations from 1000 Genomes Project data.

In summary, our results suggest that BioBin will be a useful tool to analyze sequence data. While no one can unequivocally guess the role RVs will play in uncovering hidden heritability for common complex disease, it seems that testing them in aggregate can provide valuable knowledge about the biology. Prerequisites for installation and running of BioBin and LOKI are documented in the manual, which is publicly available with the software and example statistical association scripts in R at https://ritchielab.psu.edu/ritchielab/software.

References

1. Johansen, C. T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet* **42**, 684–687 (2010).
2. Bhatia, G. *et al.* A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS.Comput Biol* **6**, e1000954 (2010).
3. Ionita-Laza, I., Buxbaum, J. D., Laird, N. M. & Lange, C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS.Genet* **7**, e1001289 (2011).
4. Haack, T. B. *et al.* Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nat Genet* **42**, 1131–1134 (2010).
5. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics* **42**, 790–793 (2010).
6. Raychaudhuri, S. *et al.* A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. *Nat Genet* **43**, 1232–1236 (2011).
7. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* **615**, 28–56 (2007).
8. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311–321 (2008).
9. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS.Genet* **5**, e1000384 (2009).
10. Han, F. & Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* **70**, 42–54 (2010).
11. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* **86**, 832–838 (2010).
12. Hoffmann, T. J., Marini, N. J. & Witte, J. S. Comprehensive approach to analyzing rare genetic variants. *PLoS.One.* **5**, e13584 (2010).
13. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
14. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res* **21**, 1529–1542 (2011).
15. Nelson, M. R. *et al.* An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* **337**, 100–104 (2012).
16. Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337**, 64–69 (2012).
17. Zhang, L., Pei, Y.-F., Li, J., Papasian, C. J. & Deng, H.-W. Efficient utilization of rare variants for detection of disease-related genomic regions. *PLoS ONE* **5**, e14288 (2010).
18. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463 (2010).
19. He, H. *et al.* Effect of population stratification analysis on false-positive rates for common and rare variants. *BMC Proceedings* **5**, S116 (2011).
20. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **39**, D38–D51 (2010).
21. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* **40**, D109–D114 (2011).
22. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* **39**, D691–D697 (2010).
23. Dimmer, E. C. *et al.* The UniProt-GO Annotation database in 2011. *Nucleic Acids Research* **40**, D565–D570 (2011).
24. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Research* **40**, D290–D301 (2011).
25. Kandasamy, K. *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* **11**, R3 (2010).
26. Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–861 (2012).
27. Stark, C. *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* **39**, D698–704 (2011).
28. McDonagh, E. M., Whirl-Carrillo, M., Garten, Y., Altman, R. B. & Klein, T. E. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark Med* **5**, 795–806 (2011).
29. Griffith, O. L. *et al.* ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research* **36**, D107–D113 (2007).
30. Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucl. Acids Res.* (2010).doi:10.1093/nar/gkq963
31. R Development Core Team *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing: Vienna, Austria, 2011).at <http://www.R-project.org>
32. Peng, B., Amos, C. I. & Kimmel, M. Forward-time simulations of human populations with complex diseases. *PLoS Genet.* **3**, e47 (2007).
33. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
34. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer New York: 2009).at <http://had.co.nz/ggplot2/book>
35. Daye, Z. J., Li, H. & Wei, Z. A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Res.* **40**, e60 (2012).
36. Bostrom, M. *et al.* Candidate genes for non-diabetic ESRD in African Americans: a genome-wide association study using pooled DNA. *Human Genetics* **128**, 195–204 (2010).
37. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics* **12**, 756–766 (2011).
38. Ellegren, H., Smith, N. G. & Webster, M. T. Mutation rate variation in the mammalian genome. *Current Opinion in Genetics & Development* **13**, 562–568 (2003).