RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

SEQMINER: An R-Package to Facilitate the Functional Interpretation of Sequence-Based Associations

Xiaowei Zhan[1†] and Dajiang J. Liu[2,3†*]

[1]*Department of Clinical Sciences, Quantitative Biomedical Research Center, Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America;* [2]*Institute for Personalized Medicine, College of Medicine, Pennsylvania State University, Pennsylvania, Hershey, United States of America;* [3]*Division of Biostatistics, Department of Public Health Sciences, College of Medicine, Pennsylvania State University, Hershey, Pennsylvania, United States of America*

Received 20 May 2015; Revised 1 July 2015; accepted revised manuscript 17 July 2015.
Published online 23 September 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21918

ABSTRACT: **Next-generation sequencing has enabled the study of a comprehensive catalogue of genetic variants for their impact on various complex diseases. Numerous consortia studies of complex traits have publically released their summary association statistics, which have become an invaluable resource for learning the underlying biology, understanding the genetic architecture, and guiding clinical translations. There is great interest in the field in developing novel statistical methods for analyzing and interpreting results from these genotype-phenotype association studies. One popular platform for method development and data analysis is R. In order to enable these analyses in R, it is necessary to develop packages that can efficiently query files of summary association statistics, explore the linkage disequilibrium structure between variants, and integrate various bioinformatics databases. The complexity and scale of sequence datasets and databases pose significant computational challenges for method developers. To address these challenges and facilitate method development, we developed the R package SEQMINER for annotating and querying files of sequence variants (e.g., VCF/BCF files) and summary association statistics (e.g., METAL/RAREMETAL files), and for integrating bioinformatics databases. SEQMINER provides an infrastructure where novel methods can be distributed and applied to analyzing sequence datasets in practice. We illustrate the performance of SEQMINER using datasets from the 1000 Genomes Project. We show that SEQMINER is highly efficient and easy to use. It will greatly accelerate the process of applying statistical innovations to analyze and interpret sequence-based associations. The R package, its source code and documentations are available from http://cran.r-project.org/web/packages/seqminer and http://seqminer.genomic.codes/.**
Genet Epidemiol 39:619–623, 2015. Published 2015 Wiley Periodicals, Inc.*

KEY WORDS: **next-generation sequencing; information retrieval; statistical genetics; genome annotation; R**

Introduction

Next-generation sequencing has made it possible to assess the impact of a full spectrum of functional variants on complex human diseases. Many large-scale genetic studies are being performed to gain insights from genotype-phenotype association data, to learn the underlying biology, and to enhance genomics-guided clinical translations.

One area of particular interest in statistical genetics is in developing novel methods to enhance the analysis and interpretation of summary association statistics. Currently, it is a standard practice for consortia studies to release summary association statistics publically. These datasets have become an invaluable resource. Numerous methods have been developed to leverage this information to conduct fine mapping [Kichaev et al., 2014], perform gene-level association tests [Lee et al., 2013; Liu et al., 2014], and infer causal relationships between biomarkers and diseases [Burgess et al., 2014; Do et al., 2013]. There is an ever-increasing need for new methods and tools to more effectively use these datasets.

The statistical package R is a popular platform for data analysis and methodology development. In order to facilitate the development of new methods in R for the functional interpretation of genotype-phenotype associations, we developed the R package SEQMINER with a variety of useful features: first, SEQMINER supports variant annotation and data integration for whole-genome datasets of summary association statistics. Second, SEQMINER allows efficient random access and queries for sequence datasets, files of summary association statistics, and files for the correlation coefficients between summary association statistics. Retrieved information can be automatically parsed and made ready for downstream analysis. Finally, SEQMINER is self-contained and optimized for statistical genetics analyses of summary association statistics. Many commonly used features in statistical genetics can be performed using a minimum of steps, which considerably expedites the analyses and method development.

We evaluated the performance of SEQMINER using datasets from the 1000 Genomes Project [Consortium et al.,

Supporting Information is available in the online issue at wileyonlinelibrary.com.

†Both the authors contributed equally to the work.

*Correspondence to: Dr. Dajiang J. Liu, 500 University Dr. Hershey, PA 17033, USA. E-mail: dajiang.liu@psu.edu; Dr. Xiaowei Zhan, 5323 Harry Hines Blvd., Dallas, Texas 75390.

2012]. We show that SEQMINER is highly efficient for annotating and querying summary association statistics. It can be a valuable tool to facilitate the analyses and interpretation of sequence-based association analyses.

Methods

Method Overview

Major uses of summary association statistics can include the following: (1) fine-mapping GWAS loci and identifying causal variants; (2) performing gene-level (or pathway-level) association tests; and (3) performing Mendelian randomizations to study the causal impact of risk factors on diseases.

These applications require knowledge of the functional annotations of DNA sequence variants and linkage disequilibrium information between sequence variants, and often require joint analyses of summary association statistics between multiple studies and multiple traits. Given the ever-increasing scale of sequence datasets and the complexity of bioinformatics databases, it is necessary to develop software packages in R that can effectively and efficiently integrate these datasets and retrieve information of interest.

SEQMINER implements features to:

1. annotate sequence variants and summary association statistics;
2. efficiently retrieve summary association statistics and their variance-covariance information specific to genetic regions of interest; and
3. collate summary association statistics and their covariance information from multiple studies.

Detailed descriptions for these functionalities are given below.

Annotate Sequence Datasets and Summary Association Statistics

Annotation information is necessary for the analysis and interpretation of sequence-based association. For example, annotation information is needed to determine the analysis unit in gene-level association tests (i.e., nonsynonymous variants within a gene). Bioinformatics databases such as functional prediction scores were shown to be helpful in prioritizing causal variants and improving power for detecting genotype-phenotype associations [Price et al., 2010].

SEQMINER implements comprehensive features for annotating sequence variants and summary association statistics. The package is designed to take generic tab-delimited files as input, which saves ad hoc efforts from users for preparing intermediate input files. Supported file formats include sequence variant genotypes in VCF/BCF format [Danecek et al., 2011] or files of summary association statistics in METAL [Willer et al., 2010] or RAREMETAL format [Feng et al., 2014; Liu et al., 2014].

To integrate information, SEQMINER preprocesses and augments the input sequence dataset with annotation information. The integrated dataset is stored in the same format as the input dataset, which remains compatible with external tools and can be indexed by tabix [Li, 2011]. Subsequent queries can be performed directly on the integrated dataset.

Two types of variant annotations are supported, gene-based annotation and region-based annotation. For gene-based annotation, sequence variants are annotated by their induced changes on the amino acid. A variety of gene/transcript definitions are supported, including UCSC KnownGenes, RefSeq, and GENCODE. In region-based annotation, genomic regions of interest (e.g., transcription factor binding sites, known GWAS signals, etc.) are listed by chromosomal position in the BED files. Sequence variants are annotated by whether they overlap with these genomic regions. Using the functionality of region-based annotation, SEQMINER can integrate numerous bioinformatics databases, e.g., PolyPhen2, SIFT, GERP scores, etc.

Finally, SEQMINER allows for the efficient generation of integrated datasets, combining sequence variants with annotation information. This feature is critical for many statistical genetics analyses where annotation information needs to be re-used repeatedly.

Efficient Retrieval of Summary Association Statistics and Sequence Variants

SEQMINER allows efficient queries for tabix-indexed sequence datasets (either preprocessed or generic). Built-in functions in SEQMINER implement a variety of frequently used queries, including extracting summary association statistics by genomic position, gene names, or annotation types. For example, it takes only one command to extract genotype (GT) information for synonymous variants (NS) in a given gene, as well as the allele frequency (AF), allele count (AC), and the positions (CHROM, POS) of the variants

```
readVCFToListByGene (fileName, geneFile, geneName=
  "CFH", annoType="Synonymous", vcfColumn=
  c("CHROM", "POS"), vcfInfo = c("AF", "AC"), vcfIndv=
  c("GT")).
```

In this function, *vcfColumn* allows users to specify descriptive columns in VCF files to be extracted. These columns include CHROM, POS, and ID. The option vcfInfo allows users to choose fields in the INFO field from VCF files. These fields may include (but are not limited to) allele frequencies (AF), annotation information (ANNO), etc. Finally, vcfIndv allows users to specify fields defined in the FORMAT column to be extracted. Examples include genotypes (GT), genotype likelihoods (GL), allelic depth (AD), etc.

It also requires only one command to retrieve summary association statistics and covariance information between variants from files in RAREMETAL format.

```
rvmeta.readDataByRange(scoreTestFiles, covFiles,
  tabixRanges)
```

Extracted information will be automatically parsed and stored in standard R objects (e.g., list, matrix) for

downstream statistical analysis. By leveraging the programming environment in R, the queries can be flexibly refined.

Collate Summary Association Statistics from Multiple Studies

There is considerable interest in the field in interrogating genetic variants with pleiotropic effects [Giambartolomei et al., 2014; Hu et al., 2013; Lee et al., 2013; Tang and Lin, 2014, 2013], examining if genetic effects vary across cohorts/ethnic groups [Wen and Stephens, 2014], performing meta-analyses that combine results from multiple studies [Liu et al., 2014] or implementing Mendelian randomization experiments by joint analyses of genetic associations with risk factors and disease outcomes [Do et al., 2013; Voight et al., 2012]. These research questions all require joint analysis of multiple sets of summary association statistics and their covariance information. Multiple studies may not have the same set of genetic variants genotyped (particularly for sequencing studies, where different variant sites are called in each study). It can be a nontrivial task to randomly access a large number of files of summary association statistics and covariance matrices, efficiently retrieve information specific to a genetic region of interest, and collate variant sites between studies. A great amount of ad hoc scripting may be needed.

To address this research need, SEQMINER is designed to read and process multiple files of summary association statistics. Loaded data will be automatically parsed, stored in standard R objects and made ready by downstream analyses. This functionality has been extensively used to implement methods for meta-analyses of gene-level association tests. Since its release, SEQMINER has been used in several large-scale meta-analyses of complex traits, including lipid levels, anthropometric traits, smoking and drinking addictions, etc.

Algorithmic Optimization

We implemented a series of algorithmic optimizations to improve the performance of SEQMINER: first, SEQMINER supports directly reading/writing compressed and tabix-indexed files. To support efficient random information retrieval from large data files, we incorporated and extended the tabix library into SEQMINER. Tabix proceeds by indexing blocks of compressed data files (bgzip) format. Using the binning index and linear index, the tabix library allows the quick location of the sequence data from disks that overlap the query interval. This design allows the retrieval of sequence information at a time complexity of $O(\log(N))$. The original tabix library only allows storing all retrieved information as strings, which cannot be directly analyzed in R. In SEQMINER, we extended tabix and implemented features to randomly access files in METAL/RAREMETAL format. Retrieved information is automatically parsed, converted to the appropriate data types (as strings, floating numbers, etc.) and made available for analysis as standard R objects, e.g., list or data frames.

Second, SEQMINER implements novel data structures for storing bioinformatics databases to speed up annotations and query. To support efficient region-based variant annotations, we used a segment tree data structure, enabling queries of $O(\log(N))$time complexity and construction of $O(N\log(N))$time and space complexity where N is the number of regions. Specifically, we constructed a red-black tree and stored every region (start position, end position, and region names) in a tree leaf. These regions are ordered internally by their start positions and then end positions. As the tree is balanced, querying genomic variants requires at most $\log_2(N)$ comparisons. This data structure has high performance in practice, and it enables the online annotation of range-based databases such as transcription factor binding sites.

Lastly, SEQMINER provides a user-friendly R interface that is easy for new users and utilizes C++ for all computationally intensive or I/O intensive queries. Algorithms described above are carefully implemented in C++ and will be compiled and optimized during package installation. These combine the high flexibility of R with the high performance of C++.

Results

Performance for Annotating Summary Association Statistics

We evaluated the performance of SEQMINER for annotating and querying summary association statistics using datasets from the 1000 Genomes phase 1 project. The call set consisted of 1,092 individuals genotyped at ~39 million variants. We simulated phenotypes under the null hypothesis of no genotype-phenotype associations. Summary association statistics were generated using RVTESTS [https://github.com/zhanxw/rvtests]. The output from RVTESTS is automatically bgzip-compressed and tabix-indexed. The file for single variant association statistics after compression is 6.2 GB, and the file for the covariance matrix between single variant score statistics after compression is 266 GB.

We evaluated the performance for annotating summary association statistics. The whole dataset was annotated as a plain text file using the function *seqminer::annotatePlain*. The annotation for the whole genome dataset took ~2 CPU hours with ~63 MB RAM. The software is thus highly scalable for very large datasets.

Performance of Retrieving Summary Association Statistics

SEQMINER supports random access and retrieval of summary association statistics that are stored in tab-delimited files and organized by chromosomal positions. In particular, it supports both METAL and RAREMETAL formats.

First, we used SEQMINER to query tabix-indexed files of summary association statistics (as described in section Performance for Annotating Summary Association Statistics). For the file containing single variant association

Table 1. Comparison of querying files of summary association statistics

Function	Task	Time complexity	Memory complexity
seqminer::rvmeta.readDataByRange	Retrieve summary association statistics from 100 randomly chosen regions	0.32 sec	550 KB
seqminer::rvmeta.readDataByRange	Retrieve summary association statistics from 100 randomly chosen regions. Also retrieve covariance matrix between these summary association statistics	1.12 sec	1.3 MB
data.table::fread	Read entire file of summary association statistics into memory	7.8 min	103 MB
data.table::fread	Read entire file of summary association statistics and their covariance matrix into memory	34.6 hr	263 GB

We compared the performance of SEQMINER for querying files of summary association statistics and files of correlations coefficients between summary association statistics.

statistics for 39 million variants, SEQMINER took 0.3 sec to retrieve summary association statistics using the function *seqminer::rvmeta.readDatabyRange*. Not surprisingly, reading the entire file into R (using either *data.table::fread* or *read.table*) and performing in-memory query took considerably longer and required a much larger memory footprint (Table 1). While SEQMINER's main advantage lies in its capability to randomly access files of summary association statistics, and it is often not necessary to read the entire file into the memory, we also compared the speed of SEQMINER in reading the entire file as a benchmark. Using SEQMINER's tabix.read.table function to read the entire file into memory in R took 7.5 min, which is ~4% faster than R's read.table command.

Second, the package is also very efficient when applied to retrieve correlation information between pairs of sequence variants. Retrieving correlation information between pairs of sequence variants for 100 randomly selected genes required only 1.13 sec. Reading all the covariance files into memory using *data.table::fread* and parsing them took over 35 hr, an unrealistic amount of time given the size of the file (266 GB).

Performance for Annotating and Querying VCF Files

As a companion feature, SEQMINER also supports annotating and querying VCF files of sequence variant genotype calls. To our knowledge, there is one R package VariantAnnotation that supports annotating and retrieving sequence variants from VCF/BCF files. VariantAnnotation relies on a variety of the Bioconductor packages to query and annotate VCF files. The package, however, was not designed to annotate and query summary association statistics.

Our package performs competitively in its shared features of annotating and querying VCF files. We compared the efficiency of VariantAnnotation and SEQMINER for annotating large-scale datasets and extracting genetic regions of interest (Tables 2 and 3): first, using 1,092 samples from the 1000 Genomes Project, we compared time and memory efficiency for annotating whole chromosome variants. In all scenarios examined, SEQMINER was >20× faster and required ~100-fold less memory than VariantAnnotation. We then benchmarked the extraction of nonsynonymous variants from 100 randomly selected genes. VariantAnnotation used 23.0 min and 1,095 MB memory, while SEQMINER used 1.3 min and 37 MB memory. In all scenarios considered, SEQMINER exhibited advantages in time and memory efficiency.

Table 2. Comparison of time and memory complexity for annotating sequence variants

Tools	Chunk size	Time (second)	Memory (kilobytes)
SEQMINER	Entire chromosome	8,371	63,072
VariantAnnotation	5,000	144,403	8,364,784
	10,000	125,748	16,236,324
	20,000	116,078	30,078,896

We benchmarked the performance of *SEQMINER* and *VariantAnnotation* for annotating sequence variants in chromosome 1 from the 1000 Genomes Project phase 1 datasets. Annotation by *SEQMINER* was done using function annotateVCF. *VariantAnnotation* cannot analyze chromosome 1 in one batch due to memory constraints. We compared the performance of *VariantAnnotation* by dividing the chromosome 1 dataset into chunks and annotating each chunk separately. For measuring memory consumption, we recorded peak memory usage. Cumulative time is recorded for annotating the entire chromosome.

Table 3. Comparison of time and memory complexity for querying selected genes/ranges

Tool	Task	Time (seconds)	Memory (kilobytes)
SEQMINER	Extract 100 randomly	76	37,948
VariantAnnotation	Selected ranges	1,313	1,122,204
SEQMINER	Extract 100 randomly	462	59,736
VariantAnnotation	Selected genes	1,718	1,461,404

We compared the performance of *SEQMINER* and *VariantAnnotation* in extracting nonsynonymous variants from 100 randomly selected genes or ranges. Whole genome datasets from the 1000 Genomes Project phase 1 were used. To extract randomly selected genes, we used readVCFToListByGene function in *SEQMINER*. For *VariantAnnotation*, we first determined the genomic ranges for each gene and extract variants within these genomic ranges. We then predicted the function of retrieved variants and select the subset of variants that were nonsynonymous.

Conclusions

In summary, we implemented an efficient software package, SEQMINER, to facilitate the analysis and interpretation of genotype-phenotype association summary statistics. We showed that the tools can scale well to large datasets with millions of variants. SEQMINER provides a useful platform where new methods can be developed and distributed. Since its release, the software package has contributed to several large-scale meta-analyses of complex traits, including lipid levels, height, and body mass index, as well as smoking and alcohol addictions. We envision that the software package will continue to be extremely valuable for interpreting genotype-phenotype association results in the sequencing era.

Acknowledgments

We would like to thank Dr. Hyun Min Kang and Yanming Li for valuable input regarding the storage of bioinformatics databases, and Dr. Arthur Berg

and Ms. Jessie Norris for critical reading of the manuscript. The authors do not have conflict of interest to declare.

References

Burgess S, et al. 2014. Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. *PLoS ONE* 9(10):e108891.

Consortium GP, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.

Danecek P, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.

Do R, et al. 2013. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet* 45(11):1345–1352.

Feng S, et al. 2014. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics*.

Giambartolomei C, et al. 2014. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 10(5): e1004383.

Hu YJ, et al. 2013. Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *Am J Hum Genet* 93(2):236–248.

Kichaev G, et al. 2014. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* 10(10):e1004722.

Lee S, et al. 2013. General framework for meta-analysis of rare variants in sequencing association studies. *Am J Hum Genet* 93(1):42–53.

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27(5):718–719.

Liu DJ, et al. 2014. Meta-analysis of gene-level tests for rare variant association. *Nat Genet* 46(2):200–204.

Price AL, et al. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86(6):832–838.

Tang ZZ, Lin DY. 2013. MASS: meta-analysis of score statistics for sequencing studies. *Bioinformatics* 29(14):1803–1805.

Tang ZZ, Lin DY. 2014. Meta-analysis of sequencing studies with heterogeneous genetic associations. *Genet Epidemiol* 38(5):389–401.

Voight BF, et al. 2012. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* 380(9841):572–580.

Wen X, Stephens M. 2014. Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. 176–203.

Willer CJ, Li Y, Abecasis GR. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26(17):2190–2191.