Exautomate: A user-friendly tool for region-based rare variant association analysis (RVAA)

Brent D. Davis¹*, Jacqueline S. Dron^{2,3}*, John F. Robinson², Robert A. Hegele^{2,3,4}, Dan J. Lizotte^{1,5}

¹ Department of Computer Science, Western University, London, Canada

² Robarts Research Institute, Schulich School of Medicine and Dentistry, Western University, London, Canada

³ Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, Canada

⁴ Department of Medicine, Schulich School of Medicine and Dentistry, Western University, London, Canada

⁵ Department of Epidemiology & Biostatistics, Schulich School of Medicine and Dentistry, Western University, London, Canada

* Authors contributed equally to this work.

Corresponding author details:

Jacqueline Dron, BSc
Robarts Research Institute, Western University
1151 Richmond Street North, Room 4299
London, Ontario, Canada, N6A 5B7
E-mail: jdron [at] uwo [dot] ca

Abstract

Region-based rare variant association analysis (RVAA) is a popular method to study rare genetic variation in large datasets, especially in the context of complex traits and diseases. Although this method shows great promise in increasing our understanding of the genetic architecture of complex phenotypes, performing a region-based RVAA can be challenging. The sequence kernel association test (SKAT) can be used to perform this analysis, but its inputs and modifiable parameters can be extremely overwhelming and may lead to results that are difficult to reproduce. We have developed a software package called “Exautomate” that contains the tools necessary to run a region-based RVAA using SKAT and is easy-to-use for any researcher, regardless of their previous bioinformatic experiences. In this report, we discuss the utilities of Exautomate and provide detailed examples of implementing our package. Importantly, we demonstrate a proof-of-principle analysis using a previously studied cohort of 313 familial hypercholesterolemia (FH) patients. Our results show an increased burden of rare variants in genes known to cause FH, thereby demonstrating a successful region-based RVAA using Exautomate. With our easy-to-use package, we hope researchers will be able to perform reproducible region-based RVAA to further our collective understanding behind the genetics of complex traits and diseases.

Introduction

Understanding the genetic architecture of complex traits and diseases is an active area of research, driven largely by massive next-generation sequencing (NGS) efforts and the availability of public repositories containing genotypic and phenotypic data on hundreds of thousands of individuals¹⁻³. While the previous obstacle in genetics research was acquiring such datasets, we are now faced with the challenge of figuring out how to effectively and appropriately study these data^{4,5}.

Bioinformatic analyses are necessary to help identify unknown genetic determinants and explore these datasets. Rare variant association analysis (RVAA) is an increasingly popular approach to study rare variants in this context⁶⁻¹⁰; however, their frequency in the population makes it difficult to attain large enough sample sizes to detect significant relationships between variants and disease¹¹⁻¹³. As such, researchers often perform region-based RVAA by grouping or collapsing rare variants together—typically by gene—to increase statistical power for association testing¹²⁻¹⁴.

One method frequently used for region-based RVAA is the sequence kernel association test (SKAT)¹⁵⁻¹⁸. While the method has been useful in revealing rare variants and genomic loci of interest in complex traits and diseases—such as cardiovascular disease, body-mass index, height, and neurodegenerative diseases¹⁹⁻²²—reproducing these results can be difficult. SKAT is challenging to implement for exome- and genome-scale analyses as it requires significant data preprocessing involving additional software dependencies and variables that can complicate reproducibility²³. On top of that, few published studies report the exact preprocessing steps and SKAT parameters used in their analysis. With an increase in the availability of genetic and phenotypic data, there has been a surge in exploratory analyses of the large NGS datasets, making SKAT a very popular tool. To assist with transparency in research and encouraging reproducibility of results, easily accessible and user-friendly bioinformatic tools are necessary.

We have created an open-access, modular script package built from pre-existing tools to automate data handling, processing, and perform a region-based RVAA using SKAT. As a proof of principle, we utilized publicly available data from the 1000 Genomes Project and data from a well-characterized lipid disorder in which the disease-causing genes are known, to test our script package. We also outline precautions needed when performing a region-based RVAA and adjusting SKAT parameters.

Our “Exautomate” package is user-friendly, designed with genetic researchers in mind, and generates a detailed methods-log to be utilized in publishing efforts. Our automated

analysis package is our standardization attempt to ensure consistent, reproducible region-based RVAA results, and to further the validity of this modern, exploratory genetic method.

Methods & Recommended Script Usage

Operating system compatibility

Our packages have been tested on the following platforms: 1) Windows Subsystem for Linux (WSL, Bash on Windows) I; 2) Ubuntu 16.04 and 18.04; and 3) Mac OS X 10.13.1.

Software dependencies and installation

The Exautomate package is presented as a series of bash scripts that can be found on Github (<https://github.com/exautomate/Exautomate-Core>). An installer script [`Installer.sh` or `mac-installer.sh`] is available to download and install the following dependencies: bedtools²⁴, BWA²⁵, Genome Analysis Toolkit (GATK) including Picard²⁶, SAMtools²⁷, tabix²⁸, and VCFtools²⁹. The following programming languages are required: C, Java, Perl, Python, and R.

These dependencies are installed using a combination of browserless downloads, wget, and the apt-get package manager. The included Mac OS X installer requires homebrew. When using R, the R-specific libraries of SKAT¹⁵, ggplot2³⁰, and reshape2³¹ are installed inside `RunSkat.R` if they have not been installed previously. At the time of this publication, all packages listed are open-source and freely available. The authors of PLINK³² and ANNOVAR³³ would prefer that their tools be registered before being downloaded and are therefore not included in our installers. Details on how to download ANNOVAR and PLINK for use by Exautomate can be found in the **Supplemental Materials (Methods)**.

Performing the region-based RVAA

Options '1' and '2' from the Exautomate main menu allow the user to: 1) perform SKAT or SKAT-O on a pre-merged .vcf file containing variant data on both controls and cases, or 2) merge a .vcf file containing variant data on controls with a .vcf file containing variant data on cases, from which the resultant .vcf file is used for SKAT or SKAT-O. Exautomate is set up such that the user inputs the necessary pieces of information (i.e. number of cases vs. controls, file names, kernel option, SKAT method) at the start of the workflow. There are two subsequent instances where the user will need to interact with the terminal before script completion. First, the user must encode the newly generated .fam file with control/case information by assigning the "phenotype" column variable as 1 (unaffected) or 2 (affected). Second, the user must modify the newly generated .SetID file; the "sets" in the first column of the file cannot be greater than 50

characters. Exautomate is set up under the assumption that the user wishes to group variants into gene sets, and therefore generates a gene-based .SetID file from ANNOVAR output. This step in the workflow is where the user may alter their sets as required. Considerations for preparing .SetID files can be found in the **Supplemental Materials (Additional Information)**.

Retrieval of 1000 Genomes data

Option '3' from the Exautomate main menu allows the user to download data from the 1000 Genomes FTP site. There are options to filter the downloaded data based on ethnicity and genomic sites of interest, as specified by a .bed file. An example of using this option can be found in **Supplemental Materials (Methods)**. An important note on this option is that it does not download information related to the sex chromosomes or mitochondrial DNA, and it does not take relatedness of the 1000 Genomes participants into account.

Creating a synthetic dataset

Option '4' from the Exautomate main menu allows the user to perform SKAT or SKAT-O analysis with a synthetic dataset in the form of a .sim file, generated using PLINK. This option was largely used to test the functionality of the Exautomate package and has been included as an extra feature for users.

Proof-of-principle demonstration analysis

The European subset of the 1000 Genomes cohort (N=503) was retrieved using option '3' of Exautomate and was filtered to contain sites covered by the LipidSeq targeted NGS panel³⁴. Patients diagnosed with familial hypercholesterolemia (FH) (N=313), which is defined as having severely elevated low-density lipoprotein (LDL) cholesterol levels, were used as a case cohort. Since these patients have been genetically diagnosed in a previous study³⁵ and the FH phenotype itself is well characterized, we have an a priori expectation of what the analysis should reveal, making it an ideal cohort for a proof of principle.

To highlight usage of the Exautomate package, we document each step of the workflow—from installation to output—for our proof-of-principle analysis using data from the 1000 Genomes and the FH patient cohort in the **Supplemental Materials (Methods)**. We strongly encourage users to consult this document first for a complete understanding of the tool, as we outline important considerations for each stage of the workflow. We also provide a general overview in **Figure 1** regarding the flow of information between file types.

Results

Region-based RVAA on FH patient cohort

The .vcf files from all FH patients and 1000 Genomes controls were merged together and filtered on the following parameters: biallelic sites, minor allele frequency $\leq 1\%$ (based on the gnomAD database³⁶), sequence ontology (insertion, deletion, missense, splice acceptor, splice donor, nonsense), and *in silico* predictions (CADD Phred^{37,38} ≥ 10). This filtered .vcf was used as the input .vcf for option '1' of Exautomate. We selected options for a linear weighted, gene-based SKAT-O analysis and a Bonferroni adjustment of P-values. From start to finish—including manually editing the .fam and .SetID files—Exautomate ran for 5 minutes and 13 seconds on an 8-cores @ 2.33GHz, 64GB RAM Ubuntu 18.04 Server machine.

The gene sets and their adjusted P-values are detailed in **Table S1**, while the distribution of adjusted P-values is illustrated in **Figure 2**. Overall, 19 genes met statistical significance when using an α -threshold of 0.05 (**Table 1**). Of importance, neither the SKAT nor SKAT-O analysis indicate which study group (i.e. the cases or the controls) has the increased burden of rare variants driving the statistical association; therefore, additional downstream analysis is required to determine if the increase of rare variants is specific to the cases or controls.

Discussion

Through the development and implementation of our open-access, user-friendly Exautomate package, we have made possible the ability to conduct a region-based RVAA following a reliable, reproducible, and transparent method. As a proof-of-principle, we utilized Exautomate to perform a region-based RVAA on a previously studied disease cohort of 313 FH patients with an a priori understanding of the genetic factors causing their phenotypes³⁹. As one of the most commonly inherited types of metabolic disease, the molecular basis and mechanisms leading to FH are extremely well characterized³⁹.

After performing optimally adjusted SKAT-O, our resultant output suggested a reliable analysis. The gene with the most significant prevalence of rare variants in cases compared to controls was *LDLR*, encoding the LDL receptor (LDLR). The LDLR is the primary receptor responsible for the removal of LDL particles from the blood; an extreme accumulation of LDL particles leads to an extreme elevation in LDL cholesterol levels, which is the main phenotypic feature of FH⁴⁰. Given that mutations in *LDLR* account for >90% of FH cases⁴¹—with over 2000 mutations reported to cause FH³⁹—it is unsurprising that our analysis revealed *LDLR* to have the greatest prevalence of rare variants. It should be noted that in the previously described FH

study, 105 unique *LDLR* variants were reported to explain the FH phenotype of 53.7% of patients, while our region-based RVAA only utilized 70 variants for analysis. This difference is because our analysis only considers single-nucleotide variants, and if there is a single missing allele call at any position, that entire genomic coordinate will be excluded from analysis. These are points that should be considered when running Exautomate on any dataset.

Two other genes known to cause ~8% and ~2% of FH cases include *APOB* and *PCSK9*, encoding apolipoprotein (apo) B and proprotein convertase subtilisin/kexin type 9 (PCSK9) respectively⁴¹; both genes were present on our list of significant results. Apo B is the main protein constituent of LDL particles and serves as the primary ligand for LDLR binding, allowing for the clearance of LDL particles from circulation⁴². Disruptions to the LDLR-binding site causes disruptions in the uptake of LDL, leading to elevations in circulating levels of LDL cholesterol³⁹. Conversely, PCSK9 is a circulating protein that directly interacts with LDLR. When bound to LDLR, PCSK9 prevents the cell-surface recycling of the receptor following its internalization, and instead targets it for lysosomal degradation⁴³. Rare gain-of-function variants in PCSK9 direct more receptors towards degradation⁴⁴; fewer available LDLRs leads to a decrease in LDL particle clearance and an increase in LDL cholesterol levels.

Understanding and correctly interpreting our results required an intimate understanding of our study cohorts and NGS panel. For example, we were left to consider why *APOB* and *PCSK9* did not rank as the second and third most significant hits from our RVAA, respectively. One of the biggest considerations of this proof-of-principle analysis was that each cohort was sequenced using different methods. Regarding our second most significant gene output, *CEL*, we observed that the majority of rare variants appeared in our FH cohort. Some individuals might interpret this to mean that *CEL* is related to FH; however, through prolonged use and familiarity of our LipidSeq panel, we know that *CEL* is often met with sequencing artifacts due to a neighboring pseudogene, which we have observed to harbour a large number of variants⁴⁵. Since our control cohort was sequenced using a different method, these artifacts do not appear in the control dataset, explaining the apparently artefactual statistical association of *CEL* with FH in this analysis. Had our control dataset been sequenced with LipidSeq, we anticipate this would have corrected the issue. It may be the case that a few of our significant gene hits are false positives due to this cohort-sequencing bias. When applying a region-based RVAA to any dataset of interest, it is imperative to understand possible differences in sequencing methods, be familiar with the nature of genes, and consider inherent characteristics of both case and control cohorts—this will help in remaining mindful of the high false positive rate and will assist with correctly interpreting results.

We recommend performing a region-based RVAA using our Exautomate package as an early-stage analysis for exploratory or observational purposes. Given the considerations discussed above, and reports of this analysis having a higher false positive rate and other limitations^{13,46,47}, potentially interesting results should be followed up with more stringent approaches including segregation analyses and functional studies. An attractive application for a region-based RVAA may be to serve as a guide for gene exploration: if the objective is to identify a previously unreported gene related to a disease or phenotype of interest, the significant output may be used as a starting point to help narrow the focus on potential genes of interest. This could be particularly helpful when dealing with exome data, with over 20,000 possible genes to consider.

The potential for variable results is large, given the number of parameters that could be adjusted prior to performing a region-based RVAA. When developing Exautomate, we took into consideration the importance of reproducible and transparent results, so we created a detailed method log output to assist in reporting. Particularly with SKAT and SKAT-O, even slight parameter adjustments at the outset can produce significant differences in output. In a substantial search of published articles reporting the use of SKAT or SKAT-O for different phenotypes—including cardiovascular disease, body-mass index, height, amyotrophic lateral sclerosis, red blood cell traits, Alzheimer’s disease, Parkinson’s disease, lipid traits, and blood pressure—there are virtually no statements or reporting of the specific parameters used, other than perhaps the variant frequency threshold^{19-22,48-54}. Unfortunately, this trend of minimal methodological information neither instills confidence in the research, nor does it facilitate replications of results. As a powerful analysis, researchers using SKAT-related tools must provide the appropriate information in published works; one might argue that these methodological details are more important than the results themselves.

While the possibility for false positives remains and efforts to explore potential biologically relevant results requires careful consideration and subsequent analysis, a region-based RVAA still remains an attractive method to set the stage for in-depth studies of rare variants influencing complex phenotypes. In order to successfully perform these analyses, researchers must have easily accessible tools that support the idea of transparency and reproducibility in research. With our Exautomate package and its implementation of SKAT and SKAT-O, we hope that researchers will utilize this tool to assist in their efforts to publish well-documented methods, correctly interpret results, and make new discoveries that will continue to add to our growing understanding of the genetic architecture underlying complex traits and disease.

Acknowledgements

We would like to thank the individuals that independently tested the Exautomate package and provided feedback on areas requiring improvement: Allison Dilllott, Arden Lawson, and Julieta Lazarte.

References

1. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
2. Hebbiring, S. Genomic and Phenomic Research in the 21st Century. *Trends Genet* **35**, 29-41 (2019).
3. Hirata, M. *et al.* Overview of BioBank Japan follow-up data in 32 diseases. *J Epidemiol* **27**, S22-S28 (2017).
4. Schmidt, B. & Hildebrandt, A. Next-generation sequencing: big data meets high performance computing. *Drug Discov Today* **22**, 712-717 (2017).
5. Bien, S.A. *et al.* The Future of Genomic Studies Must Be Globally Representative: Perspectives from PAGE. *Annu Rev Genomics Hum Genet* (2019).
6. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869-72 (2004).
7. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* **37**, 161-5 (2005).
8. Cohen, J.C., Boerwinkle, E., Mosley, T.H., Jr. & Hobbs, H.H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264-72 (2006).
9. Romeo, S. *et al.* Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet* **39**, 513-6 (2007).
10. Ji, W. *et al.* Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* **40**, 592-599 (2008).
11. Michailidou, K. Meta-Analysis of Common and Rare Variants. *Methods Mol Biol* **1793**, 73-88 (2018).
12. Auer, P.L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* **7**, 16 (2015).
13. Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5-23 (2014).
14. Ionita-Laza, I., Cho, M.H., Laird, N.M. Statistical Challenges in Sequence-Based Association Studies with Population- and Family-Based Design. *Statistics in Biosciences* **5**, 54-70 (2013).
15. 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).
16. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).
17. Derkach, A., Lawless, J.F. & Sun, L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet Epidemiol* **37**, 110-21 (2013).
18. Hsieh, A.R. *et al.* A non-threshold region-specific method for detecting rare variants in complex diseases. *PLoS One* **12**, e0188566 (2017).
19. Pirim, D. *et al.* Resequencing of LPL in African Blacks and associations with lipoprotein-lipid levels. *Eur J Hum Genet* **23**, 1244-53 (2015).
20. Turcot, V. *et al.* Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nat Genet* **50**, 26-41 (2018).
21. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186-190 (2017).

22. Nho, K. *et al.* Integration of bioinformatics and imaging informatics for identifying rare PSEN1 variants in Alzheimer's disease. *BMC Med Genomics* **9 Suppl 1**, 30 (2016).
23. Mangul, S. *et al.* Challenges and recommendations to improve installability and archival stability of omics computational tools. *bioRxiv* (2019).
24. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
26. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
27. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-93 (2011).
28. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718-9 (2011).
29. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).
30. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York* (2016).
31. Wickham, H. Reshaping Data with the reshape Package. *Journal of Statistical Software* **21**, 1-20 (2007).
32. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
33. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
34. Johansen, C.T. *et al.* LipidSeq: a next-generation clinical resequencing panel for monogenic dyslipidemias. *J Lipid Res* **55**, 765-72 (2014).
35. Wang, J. *et al.* Polygenic Versus Monogenic Causes of Hypercholesterolemia Ascertained Clinically. *Arterioscler Thromb Vasc Biol* **36**, 2439-2445 (2016).
36. Karczewski, K.J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210 (2019).
37. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
38. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019).
39. Defesche, J.C. *et al.* Familial hypercholesterolaemia. *Nat Rev Dis Primers* **3**, 17093 (2017).
40. Berberich, A.J. & Hegele, R.A. The complex molecular genetics of familial hypercholesterolaemia. *Nat Rev Cardiol* **16**, 9-20 (2019).
41. Iacocca, M.A. *et al.* ClinVar database of global familial hypercholesterolemia-associated DNA variants. *Hum Mutat* **39**, 1631-1640 (2018).
42. Hegele, R.A. Plasma lipoproteins: genetic influences and clinical implications. *Nat Rev Genet* **10**, 109-21 (2009).
43. Burke, A.C., Dron, J.S., Hegele, R.A. & Huff, M.W. PCSK9: Regulation and Target for Drug Development for Dyslipidemia. *Annu Rev Pharmacol Toxicol* **57**, 223-244 (2017).
44. Dron, J.S. & Hegele, R.A. Complexity of mechanisms among human proprotein convertase subtilisin-kexin type 9 variants. *Curr Opin Lipidol* **28**, 161-169 (2017).
45. Fjeld, K. *et al.* A recombined allele of the lipase gene CEL and its pseudogene CELP confers susceptibility to chronic pancreatitis. *Nat Genet* **47**, 518-522 (2015).

46. Jeng, X.J., Daye, Z.J., Lu, W. & Tzeng, J.Y. Rare Variants Association Analysis in Large-Scale Sequencing Studies at the Single Locus Level. *PLoS Comput Biol* **12**, e1004993 (2016).
47. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet* **11**, e1005165 (2015).
48. Gaastra, B. *et al.* Rare genetic variation in UNC13A may modify survival in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener* **17**, 593-599 (2016).
49. Chami, N. *et al.* Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. *Am J Hum Genet* **99**, 8-21 (2016).
50. Nettiksimmons, J., Tranah, G., Evans, D.S., Yokoyama, J.S. & Yaffe, K. Gene-based aggregate SNP associations between candidate AD genes and cognitive decline. *Age (Dordr)* **38**, 41 (2016).
51. Tran, N.T. *et al.* PCSK9 variation and association with blood pressure in African Americans: preliminary findings from the HyperGEN and REGARDS studies. *Front Genet* **6**, 136 (2015).
52. Lin, Y.C. *et al.* Identifying rare and common disease associated variants in genomic data using Parkinson's disease as a model. *J Biomed Sci* **21**, 88 (2014).
53. Pirim, D. *et al.* Lipoprotein lipase gene sequencing and plasma lipid profile. *J Lipid Res* **55**, 85-93 (2014).
54. Pirim, D. *et al.* Resequencing of the CETP gene in American whites and African blacks: Association of rare and common variants with HDL-cholesterol levels. *Metabolism* **65**, 36-47 (2016).

Figures and Tables

Figure 1. The flow of information between file formats from sequencing to a region-based RVAA. Our automation pipeline aims to reduce the need for user input and minimize potential sources of variability. Abbreviations: RVAA = rare variant association analysis.

Figure 2. Kernel density plot of SKAT-O output, after performing a region-based RVAA on FH patients (N=313) vs. controls from the 1000 Genomes cohort (N=503). The dashed line represents a P-value of 0.05. P-values have been adjusted using the Bonferroni correction method.

Table 1. Output from proof-of-principle SKAT-O analysis demonstrating genes with a significant burden of rare variants in one study cohort compared to the other.

Gene group	Adjusted P-value	Number of variants used in test
<i>LDLR</i> *	1.32E-52	70
<i>CEL</i>	1.04E-31	23
<i>APOB</i> *	6.66E-13	19
<i>CYP2D6</i>	2.84E-07	20
<i>ABCA1</i>	3.24E-06	13
<i>ABCG8</i>	0.000582935	9
<i>PYGM</i>	0.001242416	10
<i>LIPE</i>	0.002173174	9
<i>CPT2</i>	0.002959799	9
<i>ABCG5</i>	0.003864299	9
<i>LMNB2</i>	0.006346106	3
<i>POLD1</i>	0.008366644	5
<i>LMF1</i>	0.011220071	7
<i>NPC1L1</i>	0.011858161	8
<i>PCSK9</i> *	0.015456462	7
<i>SORT1</i>	0.023792097	6
<i>ABCC8</i>	0.030397477	7
<i>HNF4A</i>	0.031537345	4
<i>PNPLA2</i>	0.040399011	5

P-values have been adjusted using the Bonferroni correction method. “*” indicate genes known to cause autosomal dominant FH.