bioRxiv preprint doi: https://doi.org/10.1101/2021.05.02.442370; this version posted May 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TITLE: An Omics Analysis Search and Information System (OASIS) for Enabling Biological Discovery in the Old Order Amish

AUTHORS:
James A Perry[1], Brady J Gaynor[1], Braxton D Mitchell[1,2], Jeffrey R O'Connell[1]

Affiliations:

1. Division of Endocrinology, Diabetes and Nutrition, and Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA
2. Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD, USA

Corresponding Author:

James A Perry, PhD
Division of Endocrinology, Diabetes and Nutrition
University of Maryland School of Medicine
670 W. Baltimore St., HSF-III 4053
Baltimore, MD 21201
Email: jperry@som.umaryland.edu

The authors have no competing interests and no conflicts of interests to declare.

ABSTRACT

The "Omics Analysis Search and Information System" (OASIS), developed at the University of Maryland School of Medicine, enables discovery by allowing researchers to mine results from genome wide association studies (GWAS). When interesting signals are found, the research can immediately ask follow-up questions and get answers in real-time. OASIS provides this unique capability with a web-based, scientist-friendly search system and a variety of real-time analysis tools (linkage disequilibrium calculations, conditional analysis, and direct variant comparison) plus on-demand visualizations (boxplots, histograms, LocusZoom & Haploview plots, and pedigree charts). Because OASIS uses a web-based user interface, an understanding of programming or the UNIX operating system is not required. The OASIS application has been used to enable discovery from whole-exome, whole-genome, metabolome, transcriptome and methylome association results for Old Order Amish studies at the University of Maryland School of Medicine.

KEYWORDS

Data mining, genomics, omics, search, GWAS, conditional analysis, annotation, bioinformatics

INTRODUCTION

Millions, perhaps billions, of dollars are being spent on whole genome sequencing (WGS), whole exome sequencing (WES), methylation sequencing (Methyl-seq), and transcriptome sequencing (RNA-seq) to allow association analysis to be run with hundreds of phenotypes related to disease that are collected from cohort studies and electronic health records. When significant associations are detected, investigators typically want "all known annotation and functional information" for the marker/probe. Many annotation resources have been developed to meet this need including Annovar [1], The Ensembl Variant Effect Predictor (VEP) [2], SnpEff [3], FAVOR[4], and WSGA [5]. Resources for variant function include the ENCODE Project [6-9] and the NIH Roadmap Epigenomics Program [10-12], which have mapped epigenetic and regulatory regions throughout the human genome, plus the GTEx Project [13-15] which provides association data to understand the effect of variants on gene expression. These resources are a wonderful set of assets, but they can quickly create a massive data mining challenge for the researcher attempting to assemble, search and assimilate thousands of annotations and predictions for the variants with meaningful association signals.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.02.442370; this version posted May 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

At the University of Maryland School of Medicine, we experienced this data assimilation challenge first-hand for our studies of the Old Order Amish. Managing the data volume from chip-based genotyping was difficult but the volume from WES and WGS for 6000 subjects was simply daunting. To address the challenge, we developed the “Omics Analysis Search and Information System” (OASIS).

METHODS

OASIS was designed to handle data from genotyping chips and WES for 6000 Amish subjects, producing 500,000 to 1 million variants, and WGS for 1100 Amish subjects that produced 10 million variants. From these genotype datasets, we generated association results for 1200 different Amish phenotypes. The genetic association results for our Amish studies were calculated using the MMAP [16] (Mixed Model Analysis for Pedigrees and Populations) software program. A schematic of the omics annotation and analysis pipelines is shown in Figure 1.

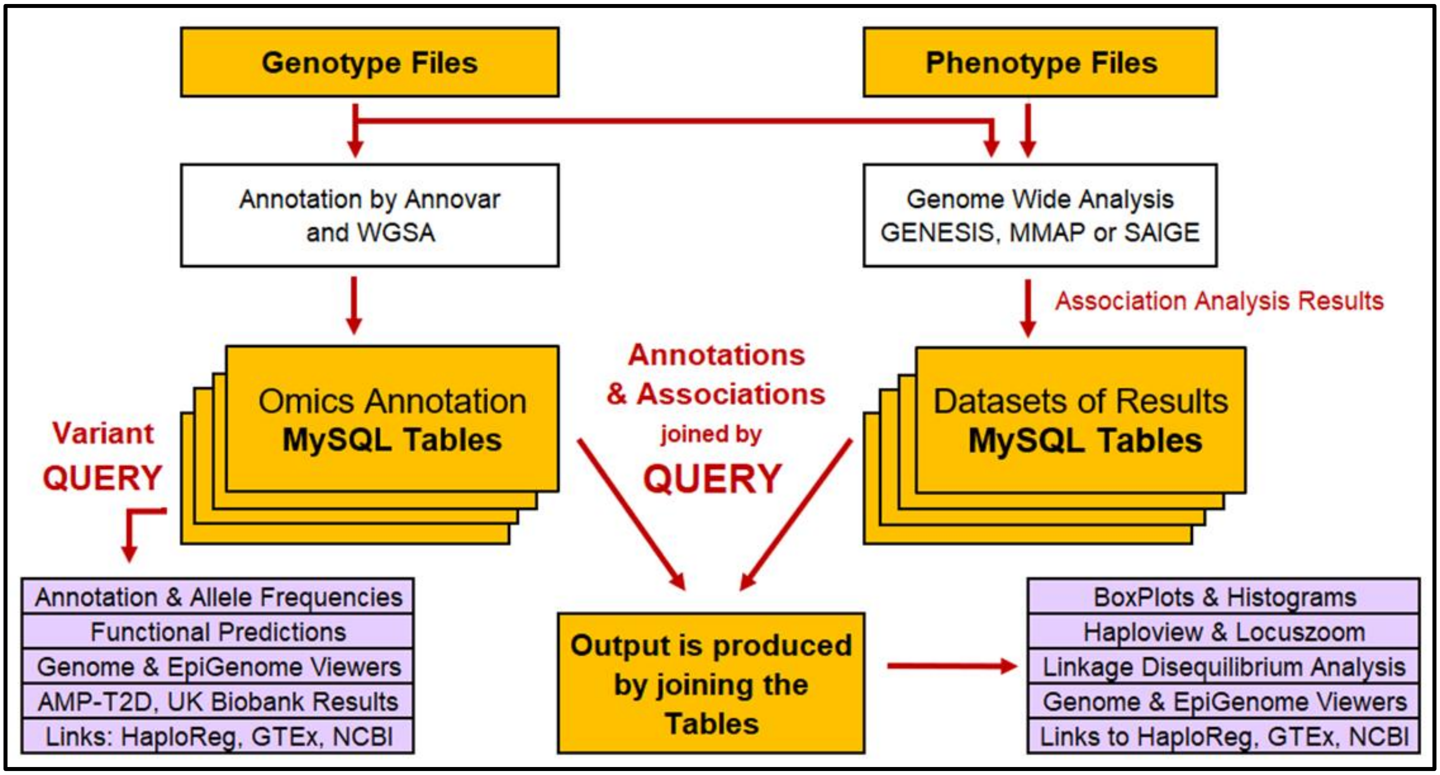

Genotype Files
Phenotype Files
Annotation by Annovar and WGSA
Genome Wide Analysis GENESIS, MMAP or SAIGE
Association Analysis Results
Omics Annotation MySQL Tables
Annotations & Associations joined by QUERY
Datasets of Results MySQL Tables
Variant QUERY
Annotation & Allele Frequencies
Functional Predictions
Genome & EpiGenome Viewers
AMP-T2D, UK Biobank Results
Links: HaploReg, GTEx, NCBI
Output is produced by joining the Tables
BoxPlots & Histograms
Haploview & Locuszoom
Linkage Disequilibrium Analysis
Genome & EpiGenome Viewers
Links to HaploReg, GTEx, NCBI

Figure 1. OASIS Annotation and Analysis Pipelines. OASIS files, tables and processes shown with gold backgrounds. Output from queries shown in purple. Processes external to OASIS shown in white.

Annotations for all genomic variants were collected from Annovar [1] and WGSA [5] and loaded into OASIS’ omics annotation tables of the MySQL database, shown on the left side of Figure 1. Next, association results from a genome wide association analysis were loaded into the MySQL tables shown on right side of Figure 1. The association analysis, which is performed external to OASIS, can be generated using any statistical analysis package. OASIS supports multiple file formats for association results including formats from Encore[17], GENESIS[18], and MMAP[16]. OASIS allows for two broad types of searches: 1) “query-by-variants” which searches just the functional annotations to help a user understand what variants are present in their study cohort (e.g. loss-of-function variants, variants reported as pathogenic by ClinVar) and the allele frequencies regardless of whether the study has phenotypes that associate with those variants, and 2) “query-by-trait” which provides a combined search of association results and functional annotations with criteria specified for both.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.02.442370; this version posted May 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

OASIS Features: When performing either type of query, the OASIS user interacts with a web-based form to specify the search criteria plus the desired annotation and functional prediction information to be displayed. Upon form submission, OASIS issues a call to the MySQL database and, in the case of query-by-trait, performs a relational database join (see Figure 1) to generate the desired report. Along with the association results (e.g. p-value, effect size) and annotation for each variant, OASIS provides hyperlinks for each variant/gene to NCBI databases [19] (dbSNP, Gene, Variation Viewer), Broad (HaploReg [20], GTEx [14]), and Open Targets[21], and the International Mouse Phenotype Consortium [22]. Users can also perform ‘on the fly’ analyses of individual level data for immediate follow-up on questions or hypothesis generated when reviewing variant associations. These analyses, performed in real-time with system calls to MMAP, include options for direct variant comparison and conditional analysis. The conditional analysis combines multiple variants and phenotypic covariates, all specified by the user via an OASIS web form, into a customized model to determine if an association is driven by single or multiple variants. Also included are linkage disequilibrium (LD) calculations relative to a user-specified reference variant as well as buttons for exploring LD with LocusZoom [23] and Haploview [24] (Figures 2 & 3).

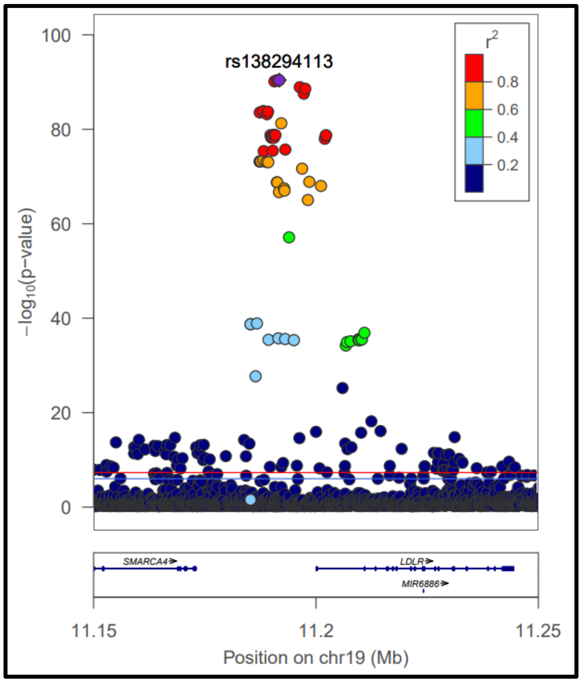


Figure 2. LocusZoom plot of LDL associations in promoter region upstream from *LDLR*.

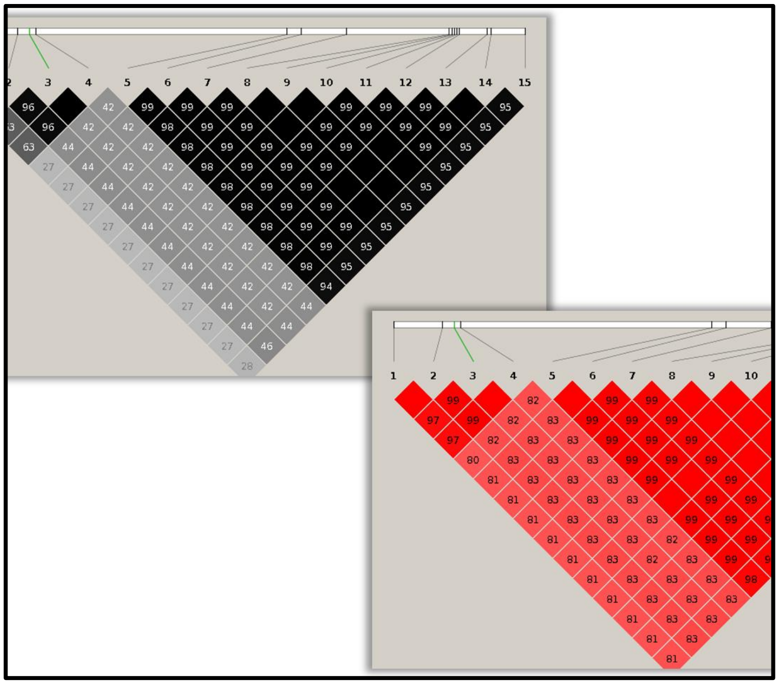


Figure 3. Haploview plot of linkage disequilibrium in Amish for region in *APOB*. r^2 (black), D-prime (red)

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.02.442370; this version posted May 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Boxplots and histograms (Figure 3) showing trait values by genotype and split by sex are available with a single click. The boxplots are particularly useful as they also provide sample sizes for each genotype and variation within the genotype to allow the user to evaluate at a glance the robustness of the association result.

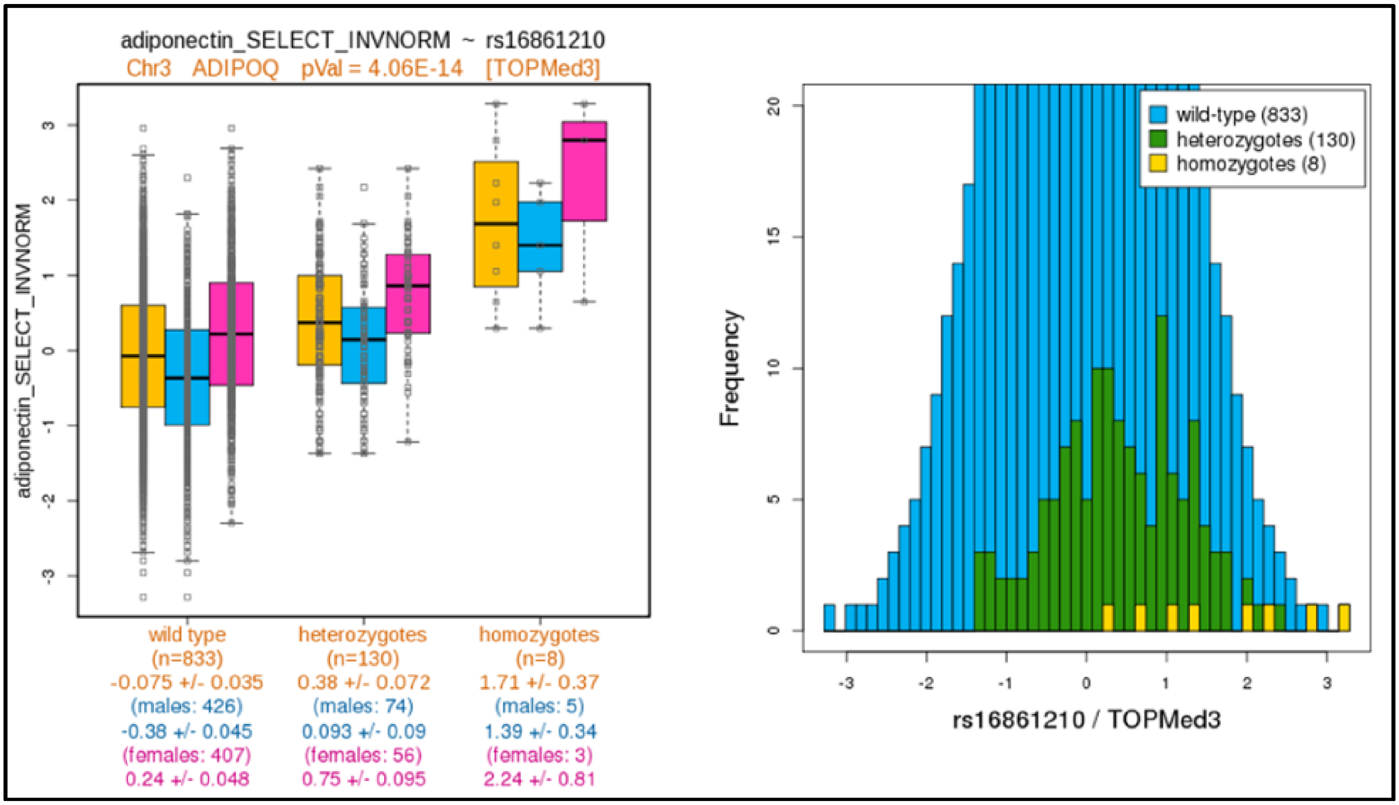
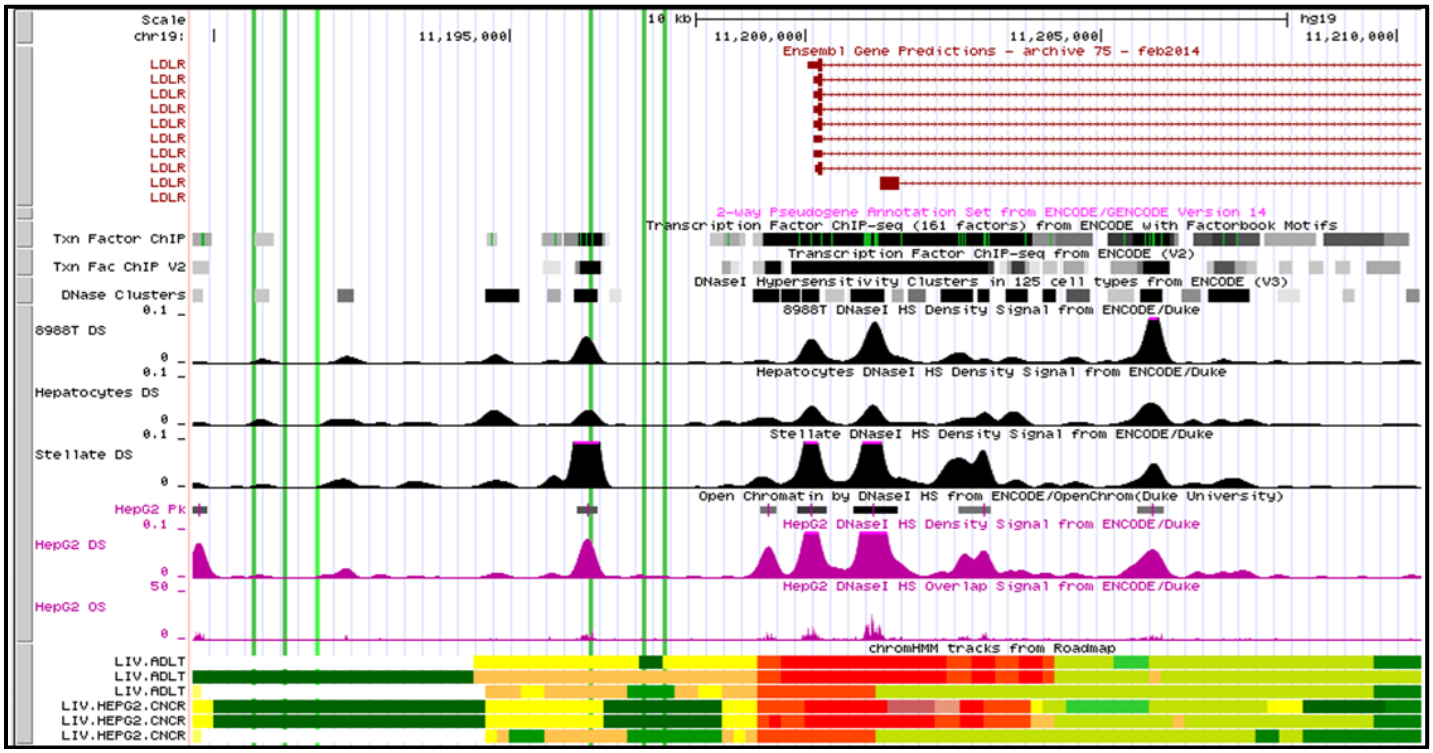


Figure 3. Boxplot and Histogram available from OASIS for any variant and phenotype combination. Boxplot on left is inverse normalized adiponectin by genotype for males (blue), females (pink) and both sexes (gold). Histogram on right shows frequency of phenotype values for each genotype including wild-type (blue), heterozygotes (green) and homozygotes (gold).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.02.442370; this version posted May 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Genomic visualizations for multiple variants can be achieved by selecting desired variants in the OASIS report and then clicking the "UCSC" button. OASIS then sends the coordinates and color code information for each selected variant to the UCSC genome browser [25] along with the desired regulatory tracks for the tissues selected by the user.For example, in Figure 4, the user selected six variants that are upstream of the LDLR gene. One of the six lies directly in a DNase hypersensitivity site and also in a predicted enhancer region (yellow) for adult liver tissue, based on the chromatin HMM models from Roadmap [11]. This information can help with developing hypothesis on which variant is causal and what the functional mechanism might be.



Scale
chr19:
10 kb
hg19
11,195,000
11,200,000
11,205,000
11,210,000
Ensembl Gene Predictions - archive 75 - feb2014
LDLR
LDLR
LDLR
LDLR
LDLR
LDLR
LDLR
LDLR
LDLR
LDLR
2-way Pseudogene Annotation Set from ENCODE/GENCODE Version 14
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE (V2)
Txn Fac ChIP V2
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
8988T DNaseI HS Density Signal from ENCODE/Duke
0.1
8988T DS
0
Hepatocytes DNaseI HS Density Signal from ENCODE/Duke
0.1
Hepatocytes DS
0
Stellate DNaseI HS Density Signal from ENCODE/Duke
0.1
Stellate DS
0
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
HepG2 Pk
HepG2 DNaseI HS Density Signal from ENCODE/Duke
0.1
HepG2 DS
0
HepG2 DNaseI HS Overlap Signal from ENCODE/Duke
50
HepG2 OS
0
chromHMM tracks from Roadmap
LIV.ADLT
LIV.ADLT
LIV.ADLT
LIV.HEPG2.CNCR
LIV.HEPG2.CNCR
LIV.HEPG2.CNCR

Figure 4. UCSC Genome Browser with user-specified variants (green lines) and tissue tracks (liver tissues)

Complete pedigree data is available for our Amish studies. This allows OASIS to generate pedigree charts by genotype for any variant and include the numeric phenotype values for each subject as shown in Figure 5. The pedigree chart feature is available with one additional click after drawing a boxplot. This feature can be used for any cohort with a pedigree data file.

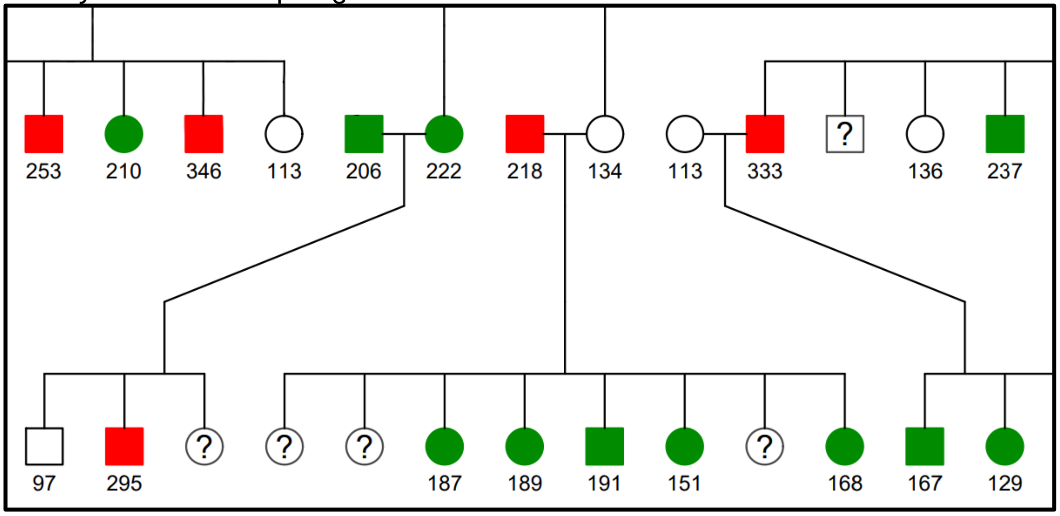


Figure 5. OASIS generated pedigree chart for a selected variant showing subjects by genotype. Homozygous-reference (white), heterozygous (green), homozygous-alternate (red), unknown genotype (question mark), male subjects (squares), female subjects (circles), connecting lines indicate parent-child-sibling relationships. Numeric values are subject measurements for the selected phenotype.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.02.442370; this version posted May 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Under the hood: OASIS is a web-based application that runs on a Linux/Apache [26] web server with 12 dual-core, 128Gb memory, and 8TB disk storage. The user interface is form-based and is displayed via a web browser using HTML5, CSS3 and JavaScript [27]. Visualizations (boxplots, bar charts, pedigree charts) are written in R [28]. The core code is written in Perl [29] using Perl libraries CGI, DPI and TMPL for preparing and processing requests and responses to/from the Apache web server and a MySQL [30] database. The Perl code also generates the customized links to online resources (NCBI, GTEx, etc.) and, if the "UCSC button" is clicked, will integrate user-selected variants into customized calls to the UCSC genome browser [25] to provide an understanding of how the variants are related to other genomic features with tissue-specific tracks in the UCSC genome browser. The tracks displayed in the UCSC genome browser are specified through OASIS's own customized setup files for a dozen tissues and/or diseases (e.g. liver, heart, diabetes) for the user to choose. Open source packages, such as LocusZoom [23] and Haploview [24], are installed on the web server and called by the core code using selections and options provided by the user via web page forms.

Data storage for OASIS is maintained in multiple formats for efficiency. These include MySQL database tables, standard comma-separated-value (csv) files and compressed binary genotype/omics files. MySQL database tables, with multiple indexes, are used to store the annotated omic variants/probes and the association results. Current prototypes have omics annotation tables with as many as 468 million variants. The association results tables contain the variant/probe identifier, used as the join field for MySQL queries joining to the omics annotation table, plus results fields from the association analysis. Association results tables are split by study for efficiency.

Standard csv files are used to store subject-level phenotype data needed for the real-time generation of boxplots, bar charts and LD calculations, all based on the limited set of subjects for any given phenotype. Phenotype descriptions, sample annotation files (with age, sex, ancestry, etc. for each sample), and pedigrees for the Amish studies are also stored in standard csv files. Genotype/omics files, also needed for boxplots and LD calculations, are converted to highly compressed binary formats created by MMAP. Genotype files for WGS are divided by chromosome for efficiency.

RESULTS and DISCUSSION

OASIS for our Amish cohort currently contains 2 billion "searchable" association results generated from 27 million genotypes. Phenotypes include 1200 traits from the Amish research studies, 200 traits from metabolomics datasets, methylation M-values for 450,000 probes and expression levels for 15,000 RNA-Seq transcripts. Phenotypes may be continuous (LDL, BMI) or binary traits such as "diabetes status". In the case of eQTL or mQTL analysis, the expression or methylation level becomes the trait. Metabolomic, eQTL and mQTL association results have also been successfully implemented in OASIS.

OASIS is widely used at the University of Maryland by the Amish Research team and their collaborators. During a recent 12-month, 35 Amish researchers performed 30,000 separate OASIS queries. All visualizations and real-time analyses are initiated from the web-based interface at the direction of the investigator who is not required to understand programming, command-line syntax or the Unix operating system. Visualizations available with a single click include boxplots, histograms, pedigree charts, Locuszoom and Haploview plots and UCSC Genome Browser displays for selected variants. Real-time analyses include conditional analysis between selected variants, LD calculations based on the cohort genotypes, and one-to-one comparisons for selected variants.

SOFTWARE AVAILABILITY

We are currently seeking funding to produce a production version that includes comprehensive web security. When upgraded, OASIS will be available for licensing free of charge for non-commercial use. Please visit the following link for updates on OASIS software availability. http://edn.som.umaryland.edu/OASIS/

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health [U01HL137181].

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.02.442370; this version posted May 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES:

1. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.
2. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.
3. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
4. Li, X., et al., *Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale.* Nat Genet, 2020. **52**(9): p. 969-983.
5. Liu, X., et al., *WGSA: an annotation pipeline for human genome sequencing studies*, in *J Med Genet*. 2016. p. 111-2.
6. Consortium, E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.
7. Consortium, E.P., *A user's guide to the encyclopedia of DNA elements (ENCODE).* PLoS Biol, 2011. **9**(4): p. e1001046.
8. Hong, E.L., et al., *Principles of metadata organization at the ENCODE data coordination center.* Database (Oxford), 2016. **2016**.
9. Raney, B.J., et al., *ENCODE whole-genome data in the UCSC genome browser (2011 update).* Nucleic Acids Res, 2011. **39**(Database issue): p. D871-5.
10. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.
11. Chadwick, L.H., *The NIH Roadmap Epigenomics Program data resource.* Epigenomics, 2012. **4**(3): p. 317-24.
12. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
13. Carithers, L.J. and H.M. Moore, *The genotype-tissue expression (GTEx) project*. 2015, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.
14. Keen, J.C. and H.M. Moore, *The Genotype-Tissue Expression (GTEx) Project: linking clinical data with molecular analysis to advance personalized medicine.* Journal of personalized medicine, 2015. **5**(1): p. 22-29.
15. Lonsdale, J., et al., *The genotype-tissue expression (GTEx) project.* Nature genetics, 2013. **45**(6): p. 580.
16. O'Connell, J.R. *MMAP: Mixed Model Analysis for Pedigrees and Populations*. Available from: https://mmap.github.io/
17. *Encore - web-based tool for running association tests with large-scale genetic sequencing data.*; Available from: https://encore.sph.umich.edu/sign-in.
18. *GENESIS - GENetic EStimation and Inference in Structured samples*. Available from: https://github.com/UW-GAC/GENESIS.
19. Coordinators, N.R., *Database resources of the national center for biotechnology information.* Nucleic acids research, 2017. **45**(Database issue): p. D12.
20. Ward, L.D. and M. Kellis, *HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease.* Nucleic acids research, 2015. **44**(D1): p. D877-D881.
21. Carvalho-Silva, D., et al., *Open Targets Platform: new developments and updates two years on.* Nucleic Acids Research, 2018. **47**(D1): p. D1056-D1065.
22. Dickinson, M.E., et al., *High-throughput discovery of novel developmental phenotypes.* Nature, 2016. **537**(7621): p. 508-514.
23. Pruim, R.J., et al., *LocusZoom: regional visualization of genome-wide association scan results.* Bioinformatics, 2010. **26**(18): p. 2336-2337.
24. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2004. **21**(2): p. 263-265.
25. Casper, J., et al., *The UCSC genome browser database: 2018 update.* Nucleic acids research, 2017. **46**(D1): p. D762-D769.
26. *The Apache HTTP Server Project*. Available from: https://httpd.apache.org/.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.02.442370; this version posted May 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

27. Carey, P.M., *New Perspectives on Html5, Css3, and Javascript, Loose-Leaf Version*. 2017: Cengage Learning.
28. *The R Project for Statistical Computing*. Available from: https://www.r-project.org/.
29. *Perl Programming Documentation*. Available from: http://perldoc.perl.org/.
30. *MySQL Community Edition*. Available from: https://www.mysql.com/products/community/.