An Effective Gene-Based Rare Variant Association Analysis Pipeline for Case-Control Studies of Disease

Canhong Wen
University of Science and Technology of China
Ruijia Li
University of Science and Technology of China https://orcid.org/0000-0002-3918-1553
Jiahui Cai
Shantou University Medical College
Haizhu Tan (linnanqia@126.com)
Shantou University Medical College

Research article

Keywords: rare variant, disease, bioinformatics, pipeline

Posted Date: December 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-116709/v1

RESEARCH

An effective gene-based rare variant association analysis pipeline for case–control studies of disease

Canhong Wen[1], Ruijia Li[2], Jiahui Cai[3] and Haizhu Tan[3*]

*Correspondence:
linnanqia@126.com
[3]Department of Preventive Medicine,
Shantou University Medical College, Shantou,
China
Full list of author information is available at the end of the article

Abstract

Background: In complex disease studies, genome-wide association study (GWAS) has been successfully used for identifying associated genetic risk loci. In fact, only a small fraction of the apparent heritability can be explained by common variants. Because research efforts have largely focused on common genetic variants, the missing heritability could be mostly due to rare genetic variants. Substantial research efforts have been devoted to developing software for genotype imputation and designing variant binning strategies and statistical methods for rare variant association testing of datasets on GWAS chips. However, few systematic pipelines have been proposed to identify rare disease-related genes.

Results: We present EGRVA, an Effective Gene-based Rare Variant Association analysis pipeline for genotype imputation, quality control, gene-based functional annotation, statistical analysis, and bioinformatics analysis of identified genes. As a complementary pipeline for rare variant analysis on GWAS chips, EGRVA is relatively straightforward and cost-efficient. Furthermore, we tested the EGRVA pipeline with the preterm birth (PTB) dataset from the GPN-PBR. We focused on the 6 genes identified by EGRVA: FLG, HRNR, PMS1, ATM, OR2AG1 and SLC22A25. We also explored the underlying biological interpretation of these potentially significant genes.

Conclusions: As a complementary pipeline for rare variant analysis on GWAS chips, EGRVA is relatively straightforward and costefficient. The application of the pipeline will contribute to the support of rare variants to explain the missing heritability by effectively discovering genes related to disease.

Keywords: rare variant; disease; bioinformatics; pipeline

Background

With the rapid development of DNA sequencing technology, an increasing number of susceptibility loci for some complex diseases have been revealed by GWAS ([1, 2, 3, 4]). These studies are mainly based on common genetic variants, typically with Minor Allele Frequency (MAF) $> 5\%$. Although susceptibility loci for many diseases has been identified, much of the estimated heritability for complex traits has not been explained. Even though a large GWAS meta-analysis has been performed for some diseases, much of the heritability remains unexplained ([5, 6]). For instance, in the genome-wide meta-analysis of Crohn's disease, although the number of confirmed susceptibility loci increased to 71, they explained less than a quarter of the heritability. ([7]).

For the problem of missing heritability, several explanations have been put forward ([8, 9]). Among them, rare variants attract attention because they are theoretically and empirically responsible for a fraction of the missing heritability. Theoretically, deleterious alleles are likely to be rare due to purification selection ([10, 11]). In fact, loss-of-function variants that block the generation of functional proteins are particularly rare ([12, 13]). There is also empirical evidence that rare variants are associated with complex diseases ([14, 15]). For example, many rare forms of common diseases and Mendelian diseases are caused by highly permeable rare variants ([16]). Therefore, the identification of rare variants associated with traits and diseases is becoming increasingly important.

With the gradual increase in research into rare variants, various studies have been proposed, but there are only a few works on pipeline development for Rare Variant Association Study (RVAS). Zuk et al. described a conceptual framework to address some key questions, including sample size required to detect association, relative merits of testing disruptive alleles and missense alleles, and frequency thresholds for filtering alleles ([17]). Seunggeung Lee et al. presented an analysis pipeline for RVAS and reviewed some existing cost-effective sequencing designs and genotyping platforms as well as statistical issues in RVAS with a focus on study designs and statistical tests ([18]). However, these pipelines provide limited functionality and are not a complete analysis pipeline for GWAS dataset. In addition, these pipelines lack methods to explore the potential biological interpretations of discovered "risky" variants.

To tackle these problems, we present an Effective Gene-based Rare Variant Association analysis (EGRVA) pipeline for case–control studies to identify rare disease-related genes. The EGRVA pipeline is for rare variants obtained from GWAS panels and combines novel methods with pre-existing tools. In this study, we present the EGRVA and show how to perform data pre-processing before imputation and how to use external functional annotation information to perform statistical tests to improve the statistical power. In addition, we validate the molecular genetics of some identified variants by using several well-known bioinformatics tools.

Methods

EGRVA accepts data in PLINK format as its input and involves the following steps: genotype imputation, Quality Control (QC), gene-based functional annotation, statistical analysis implemented with a Bayesian method for MIxture model based Rare variant Analysis on GEnes (MIRAGE), and further biological interpretations of the variants in the identified significant genes. The complete workflow is described in Figure 1.

Genotype imputation

Genotype imputation is an important step before rare variant genetic association studies because imputation accuracy increases the number of haplotypes, especially for rare variants ([19, 20, 21]). For imputation in our pipeline, we use the Michigan Imputation service with Minimac4 ([22]). To improve the accuracy of the imputation, we choose a large reference panel, the Haplotype Reference Consortium (HRC) to provide accurate genotype imputation at MAF as low as 0.1% ([23]). To further

guarantee the imputation effect and save imputation time, we included three data pre-processing steps before imputation. In the first step, a preliminary QC procedure is executed by PLINK software ([24]). The second step involves using LiftOver (http://genome.ucsc.edu/) to convert the genome coordinates into hg19 to satisfy the requirements of the Michigan imputation server. We chose SHAPEIT2 as the phasing tool because it combines the best features of SHAPEIT1 and IMPUTE2 ([25, 26]), which might improve efficiency and accuracy when estimating haplotypes from genotype data. The final step is to check the PLINK bim file against the HRC reference Single Nucleotide Polymorphism (SNP) list with the script in McCarthy Group Tools (https://www.well.ox.ac.uk/~wrayner/tools/). In this step, we update strand, position and ref/alt assignment, and remove SNPs with differing alleles, allele frequency differences greater than 0.2, and not in the reference panel. In addition, we set the R2 (squared Pearson correlation between imputed and experimental genotypes) to 0.3 to minimize the imputed file size when we start our imputation.

Quality Control

After imputation, BCFTOOLS is used to filter SNPs with R2 < 0.8 and change SNP ID to CHROM: POS: REF: ALT to avoid duplication. Next, we use PLINK software again for QC ([27]). Because this pipeline focuses on rare variants, SNPs with MAF $> 0.5\%$ and $< 1\%$ are retained. In addition, SNPs with call rate $> 10\%$ and samples with call rate $> 5\%$ and Hardy Weinberg Equilibrium P-value $<$ 1.00e-5 are eliminated. We also perform a multi-allele check to remove multiallelic SNPs.

Gene-based functional annotation

Functional variants, such as non-synonymous SNPs, are more likely to be rare and more prone to disrupt gene function ([28]). They are hypothesized to have greater expected impact on phenotypic development than other variants. Non-synonymous variants are exonic, lying in the coding regions of genes. They are able to disrupt the coding sequence of genes, thereby resulting in malformed and dysfunctional protein products. In this pipeline, we apply ANNOVAR to query the refGen database of functional effect prediction ([29]). Then, we select exonic non-synonymous variants and use three tools, including polyPhen2 HDIV scores ([30]), SIFT scores ([31]) and CADD scores ([32]), to classify the non-synonymous variants as damaging or non-damaging variants.

Statistical analysis

The single variant test method assesses rare variants with inadequate power ([16]). Instead of testing the effects based on a single variant, it is more powerful to run the association test based on multiple variants in a biologically relevant region. Many statistical methods have been proposed, such as CMC ([33]), the Variable Threshold approach ([34]) and C-alpha ([35]). Despite favorable performance, methods based on multiple variants feature unrealistic assumptions that all rare variants in a gene have a non-zero effect or zero effect together. To deal with this problem, Shengtong Han et al. develop a Bayesian method for MIRAGE ([36]). The key idea of this method is to model variants in a gene as a mixture of risk variants and non-risk

variants. Each variant has a prior probability of being a risk variant, which depends on the functional annotations of the variant. This prior probability reflects the sparsity of risk variants and better accounts for heterogeneity of variant effects within a gene.

We use MIRAGE for the association analysis and performed it with the MIRAGE R-package. First, we separately count the number of rare variant alleles at each locus in the case group and the control group based on the results of ANNOVAR. Then, we group each variant into different proportions of risk variants based on functional annotation. Deleterious groups have a higher risk proportion. Finally, we calculate each gene's Bayes Factor (BF) to assess its association with the disease by comparing the likelihood of the full model (risk gene) to the likelihood of the null model (non-risk gene) ([37]). BF is similar to the likelihood ratio test. For example, a BF > 1 provides evidence for rejecting the null model and for the presence of a filler effect.

Bioinformatic analyses

After statistical analysis, a series of bioinformatics tools are used with the genes identified. The main tools include the NCBI database (https://www.ncbi.nlm.nih.gov/gene/), GeneCards database (https://www.genecards.org/), Uniprot (https://www.uniprot.org/), Human Protein Atlas database (https://www.proteinatlas.org/), online Mendelian Inheritance in the Man database (https://www.omim.org/), human disease database (https://www.malacards.org/) and human biological pathway unification PathCards (https://pathcards.genecards.org/). These tools can help in understanding the variants in the genes from genomics, biochemistry and other channels.

Results

EGRVA identifies putative risk genes in the PTB dataset

We used EGRVA with the PTB dataset to demonstrate that this pipeline can successfully identify risky genes. The dataset used for the analyse was obtained from the database of Genotype and Phenotype (dbGaP) found at (http://www.ncbi.nlm.nih.gov/gap). In total, 743 singleton pregnancies with spontaneous preterm birth (from 20 to $<$ 34 weeks' gestation) were defined as the case group and 752 pregnancies (from 39 to $<$ 42 weeks'gestation) with spontaneous onset of labor were defined as the control group. In total, data for 76 pregnancies without genotypes were removed. We focused on 702 samples in the case group and 717 in the control group, with 868,278 SNPs retained.

QC was performed before imputation, and two samples with a call rate $> 5\%$ were deleted. Overall, limiting the R2 > 0.8, 1,417 samples and 67,975 variants were retained after imputation. QC was implemented again. 1,417 samples and 64,939 variants remained after QC. Before the association study, we used ANNOVAR to identify 196 exonic non-synonymous variants. Then we set the variants with PolyPhen2 HDIV score > 0.975, CADD score > 20, or SIFT score < 0.05 as group 2 (high-risk group). Others were the low-risk group. In total, 1,417 samples with 196 exonic non-synonymous variants were analyzed using the MIRAGE method. Genes with BF > 10 and posterior probability (post.prob) > 0.8 simultaneously had top potential. Table 1 shows the results of the 6 genes identified by BF > 10 and post.prob > 0.8.

Bioinformatics analyses of the identified genes

PTB is a polygenic disease. Rare variants in identified genes were investigated for their involvement in innate immunity and inflammation. Below is a detailed analysis.

The expression of Filaggrin (FLG), with three SNPs (rs7537147, rs12073613, rs113652604), is highest in skin and is absent from tissues related to mother-and-child interactions such as uterus, placenta, or mammary glands ([38]). Hornerin (HRNR), with two SNPs (rs6659183, rs138421943), is abundant in barrier organs, such as uterine cervix and placenta. It can regulate protein phosphorylation and inflammatory and immune reactions ([39]). Furthermore, abundant HRNR can protect organs such as central nervous system and female gonads against the potentially damaging effects of an inflammatory immune response. Post-meiotic segregation increased 1 (PMS1), with three SNPs (rs1145231, rs1145232, rs1145234), is involved in the repair of errors that occur during DNA replication. The DNA mismatch repair (MMR) system has a significant relationship with human fertility ([40, 41]). ATM, with SNPs rs1800056 and rs3218673, belongs to the PI3/PI4-kinase family. It can activate p53. Uterine-specific p53 deficiency confers premature uterine senescence and promotes preterm birth in mice ([42]). For OR2AG1, with SNPs rs74057919 and rs74057920, it is found that some tocolytics acting via GPCR signal-mediated pathways are involved in myometrium relaxation and contraction ([43]). Previous studies showed that the organic anion transporter (OAT) subfamily constitutes approximately half of the solute carrier 22 (SLC22) transporter family, including SLC22A25, with SNPs rs17157907 and rs35722529 ([44]). OAT members are expressed in many tissues such as kidney, liver, and placenta ([45]).

Conclusions

We described a straightforward rare variant analysis pipeline called EGRVA for analyzing GWAS data and validated it with a dataset for PTB. The EGRVA pipeline is easy to implement and is sensitive to risky genes. We verified the effectiveness of the proposed pipeline by bioinformatics analyses of the associated genes with the PTB dataset. The EGRVA pipeline can be an effective supplement to next-generation sequencing of rare variants. However, several limitations still exist in the EGRVA pipeline. Indeed, environmental effects related to disease are not considered, and the sample size of the current GWAS dataset is still not large enough.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The data in this paper are from the public database platform and have been approved for publication.

Availability of data and materials
The PTB dataset that support the findings of this study was obtained from the database of Genotype and Phenotype (dbGaP) found at (http://www.ncbi.nlm.nih.gov/gap) (accession number phs000714.v1.p1). Samples were provided by the NICHD-funded Genomic and Proteomic Network for Preterm Birth Research (GPN-PBR). EGRVA is freely available on https://github.com/ruijiali/EGRVA.

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

Funding
This work has been supported by the National Natural Science Foundation of China [11801540 to C.H.W.]; the Natural Science Foundation of Guangdong [2017A030310572 to C.H.W.]; the Fundamental Research Funds for the Central Universities [WK2040170015,WK2040000016 to C.H.W]; the National Key Research and Development Program of China [2018YFC1315400]; the Science and Technology Planning Project of Guangdong Province [2017A010101030]; and the third Medical technology projects of Shantou in 2018.

Authors' contributions
All authors contributed to the article. Author's contribution concept: HZT CHW RJL. Analysis data: HZT RJL. Auxiliary analysis: HZT JHC. Writing a paper: HZT CHW RJL.

Acknowledgements
Not applicable

Author details
[1]Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, China. [2]School of Data Science, University of Science and Technology of China, Hefei, China. [3]Department of Preventive Medicine, Shantou University Medical College, Shantou, China.

References

1. Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J.: Five years of gwas discovery. The American Journal of Human Genetics **90**(1), 7–24 (2012). https://doi.org/10.1016/j.ajhg.2011.11.029
2. Hindorff, L.A.: A catalog of published genome-wide association studies. http://www. genome. gov/26525384 (2009)
3. Lee, J.C., Parkes, M.: Genome-wide association studies and crohn disease. Briefings in functional genomics **10**(2), 71–76 (2011). https://doi.org/10.1093/bfgp/elr009
4. Ferreira, M.A., Gamazon, E.R., Al-Ejeh, F., Aittomäki, K., Andrulis, I.L., Anton-Culver, H., Arason, A., Arndt, V., Aronson, K.J., Arun, B.K., *et al.*: Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nature communications **10**(1), 1–18 (2019). https://doi.org/10.1038/s41467-018-08053-5
5. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segre, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., *et al.*: Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics **44**(9), 981 (2012). https://doi.org/10.1038/ng.2383
6. Slatkin, M.: Epigenetic inheritance and the missing heritability problem. Genetics **182**(3), 845–850 (2009). https://doi.org/10.1534/genetics.109.102798
7. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., *et al.*: Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. Nature genetics **42**(12), 1118 (2010). https://doi.org/10.1038/ng.717
8. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., Nadeau, J.H.: Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics **11**(6), 446–450 (2010). https://doi.org/10.1038/nrg2809
9. Zuk, O., Hechter, E., Sunyaev, S.R., Lander, E.S.: The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences **109**(4), 1193–1198 (2012). https://doi.org/10.1073/pnas.1119675109
10. Kryukov, G.V., Pennacchio, L.A., Sunyaev, S.R.: Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. The American Journal of Human Genetics **80**(4), 727–739 (2007). https://doi.org/10.1086/513473
11. Pritchard, J.K.: Are rare variants responsible for susceptibility to complex diseases? The American Journal of Human Genetics **69**(1), 124–137 (2001). https://doi.org/10.1086/321272
12. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., *et al.*: A systematic survey of loss-of-function variants in human protein-coding genes. Science **335**(6070), 823–828 (2012). https://doi.org/10.1126/science.1215040
13. Consortium, .G.P., *et al.*: An integrated map of genetic variation from 1,092 human genomes. Nature **491**(7422), 56 (2012). https://doi.org/10.1038/nature11632
14. Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Masson, G., Agnarsson, B.A., Benediktsdottir, K.R., Sigurdsson, A., Magnusson, O.T., Gudjonsson, S.A., Magnusdottir, D.N., *et al.*: A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nature genetics **44**(12), 1326 (2012). https://doi.org/10.1038/ng.2437
15. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., *et al.*: Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease. Nature genetics **43**(11), 1066 (2011). https://doi.org/10.1038/ng.952
16. Gibson, G.: Rare and common variants: twenty arguments. Nature Reviews Genetics **13**(2), 135–145 (2012). https://doi.org/10.1038/nrg3118
17. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., Lander, E.S.: Searching for missing heritability: designing rare variant association studies. Proceedings of the National Academy of Sciences **111**(4), 455–464 (2014). https://doi.org/10.1073/pnas.1322563111
18. Lee, S., Abecasis, G.R., Boehnke, M., Lin, X.: Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics **95**(1), 5–23 (2014). https://doi.org/10.1016/j.ajhg.2014.06.009
19. Browning, B.L., Browning, S.R.: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics **84**(2), 210–223 (2009). https://doi.org/10.1016/j.ajhg.2009.01.005

20. Howie, B.N., Donnelly, P., Marchini, J.: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics **5**(6) (2009). https://doi.org/10.1371/journal.pgen.1000529
21. Li, Y., Willer, C.J., Ding, J., Scheet, P., Abecasis, G.R.: Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology **34**(8), 816–834 (2010). https://doi.org/10.1002/gepi.20533
22. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., *et al.*: Next-generation genotype imputation service and methods. Nature genetics **48**(10), 1284–1287 (2016). https://doi.org/10.1038/ng.3656
23. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., *et al.*: A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics **48**(10), 1279–1283 (2016). https://doi.org/10.1038/ng.3643
24. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J.: Second-generation plink: rising to the challenge of larger and richer datasets. Gigascience **4**(1), 13742–015 (2015). https://doi.org/10.1186/s13742-015-0047-8
25. Delaneau, O., Marchini, J., McVean, G.A., Donnelly, P., Lunter, G., Marchini, J.L., Myers, S., Gupta-Hinch, A., Iqbal, Z., Mathieson, I., *et al.*: Integrating sequence and array data to create an improved 1000 genomes project haplotype reference panel. Nature communications **5**, 3934 (2014). https://doi.org/10.1038/ncomms4934
26. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., *et al.*: Reference-based phasing using the haplotype reference consortium panel. Nature genetics **48**(11), 1443 (2016). https://doi.org/10.1038/ng.3679
27. Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., *et al.*: Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic epidemiology **34**(6), 591–602 (2010). https://doi.org/10.1002/gepi.20516
28. Carlton, V.E., Ireland, J.S., Useche, F., Faham, M.: Functional single nucleotide polymorphism-based association studies. Human genomics **2**(6), 391 (2006). https://doi.org/10.1186/1479-7364-2-6-391
29. Wang, K., Li, M., Hakonarson, H.: Annovar: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research **38**(16), 164–164 (2010). https://doi.org/10.1093/nar/gkq603
30. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., Shendure, J.: A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics **46**(3), 310 (2014). https://doi.org/10.1038/ng.2892
31. Ng, P.C., Henikoff, S.: Sift: Predicting amino acid changes that affect protein function. Nucleic acids research **31**(13), 3812–3814 (2003). https://doi.org/10.1093/nar/gkg509
32. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R.: A method and server for predicting damaging missense mutations. Nature methods **7**(4), 248–249 (2010). https://doi.org/10.1038/nmeth0410-248
33. Li, B., Leal, S.M.: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. The American Journal of Human Genetics **83**(3), 311–321 (2008). https://doi.org/10.1016/j.ajhg.2008.06.024
34. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.-J., Sunyaev, S.R.: Pooled association tests for rare variants in exon-resequencing studies. The American Journal of Human Genetics **86**(6), 832–838 (2010). https://doi.org/10.1016/j.ajhg.2010.04.005
35. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., Daly, M.J.: Testing for an unusual distribution of rare variants. PLoS genetics **7**(3) (2011). https://doi.org/10.1371/journal.pgen.1001322
36. Han, S., Knoblauch, N., Wang, G., Zhao, S., Liu, Y., Xie, Y., Sheng, W., Nguyen, H.T., He, X.: A bayesian method for rare variant analysis using functional annotations and its application to autism. bioRxiv, 828061 (2019). https://doi.org/10.1101/828061
37. Lodewyckx, T., Kim, W., Lee, M.D., Tuerlinckx, F., Kuppens, P., Wagenmakers, E.-J.: A tutorial on bayes factor estimation with the product space method. Journal of Mathematical Psychology **55**(5), 331–347 (2011). https://doi.org/10.1016/j.jmp.2011.06.001
38. Fu, S.-J., Shen, S.-L., Li, S.-Q., Hua, Y.-P., Hu, W.-J., Guo, B., Peng, B.-G.: Hornerin promotes tumor progression and is associated with poor prognosis in hepatocellular carcinoma. BMC cancer **18**(1), 815 (2018). https://doi.org/10.1186/s12885-018-4719-5
39. Gerstel, U., Latendorf, T., Bartels, J., Becker, A., Tholey, A., Schröder, J.-M.: Hornerin contains a linked series of ribosome-targeting peptide antibiotics. Scientific reports **8**(1), 1–15 (2018). https://doi.org/10.1038/s41598-018-34467-8
40. Hu, M.-h., Liu, S.-y., Wang, N., Wu, Y., Jin, F.: Impact of dna mismatch repair system alterations on human fertility and related treatments. Journal of Zhejiang University-SCIENCE B **17**(1), 10–20 (2016). https://doi.org/10.1631/jzus.B1500162
41. Brown, K.D., Rathi, A., Kamath, R., Beardsley, D.I., Zhan, Q., Mannino, J.L., Baskaran, R.: The mismatch repair system is required for s-phase checkpoint activation. Nature genetics **33**(1), 80–84 (2003). https://doi.org/10.1038/ng1052
42. Hirota, Y., Daikoku, T., Tranguch, S., Xie, H., Bradshaw, H.B., Dey, S.K.: Uterine-specific p53 deficiency confers premature uterine senescence and promotes preterm birth in mice. The Journal of clinical investigation **120**(3), 803–815 (2010). https://doi.org/10.1172/JCI40051
43. Salomonis, N., Cotte, N., Zambon, A.C., Pollard, K.S., Vranizan, K., Doniger, S.W., Dolganov, G., Conklin, B.R.: Identifying genetic networks underlying myometrial transition to labor. Genome biology **6**(2), 12 (2005). https://doi.org/10.1186/gb-2005-6-2-r12
44. Eraly, S.A., Monte, J.C., Nigam, S.K.: Novel slc22 transporter homologs in fly, worm, and human clarify the phylogeny of organic anion and cation transporters. Physiological genomics **18**(1), 12–24 (2004). https://doi.org/10.1152/physiolgenomics.00014.2004

45. Nigam, S.K., Bush, K.T., Martovetsky, G., Ahn, S.-Y., Liu, H.C., Richard, E., Bhatnagar, V., Wu, W.: The organic anion transporter (oat) family: a systems biology perspective. Physiological reviews **95**(1), 83–123 (2015). https://doi.org/10.1152/physrev.00025.2013

Figures

Figure 1 The complete Effective Gene-based Rare Variant Association analysis (EGRVA) pipeline for case–control studies of disease.

Tables

Table 1 The 6 genes identified by MIRAGE (BF $>$ 10 and post.prob $>$ 0.8).

Gene	Bayes Factor	Post.prob	Chromosome	Included SNPs
FLG	4704.92089	1	1	rs7537147 rs12073613 rs113652604
HRNR	482.1306102	1	1	rs6659183 rs138421943
PMS1	101.3975234	1	2	rs1145231 rs1145232 rs1145234
ATM	19.90539749	1	11	rs1800056 rs3218673
OR2AG1	19.81828387	1	11	rs74057919 rs74057920
SLC22A25	10.5674771	1	11	rs17157907 rs35722529

Figures

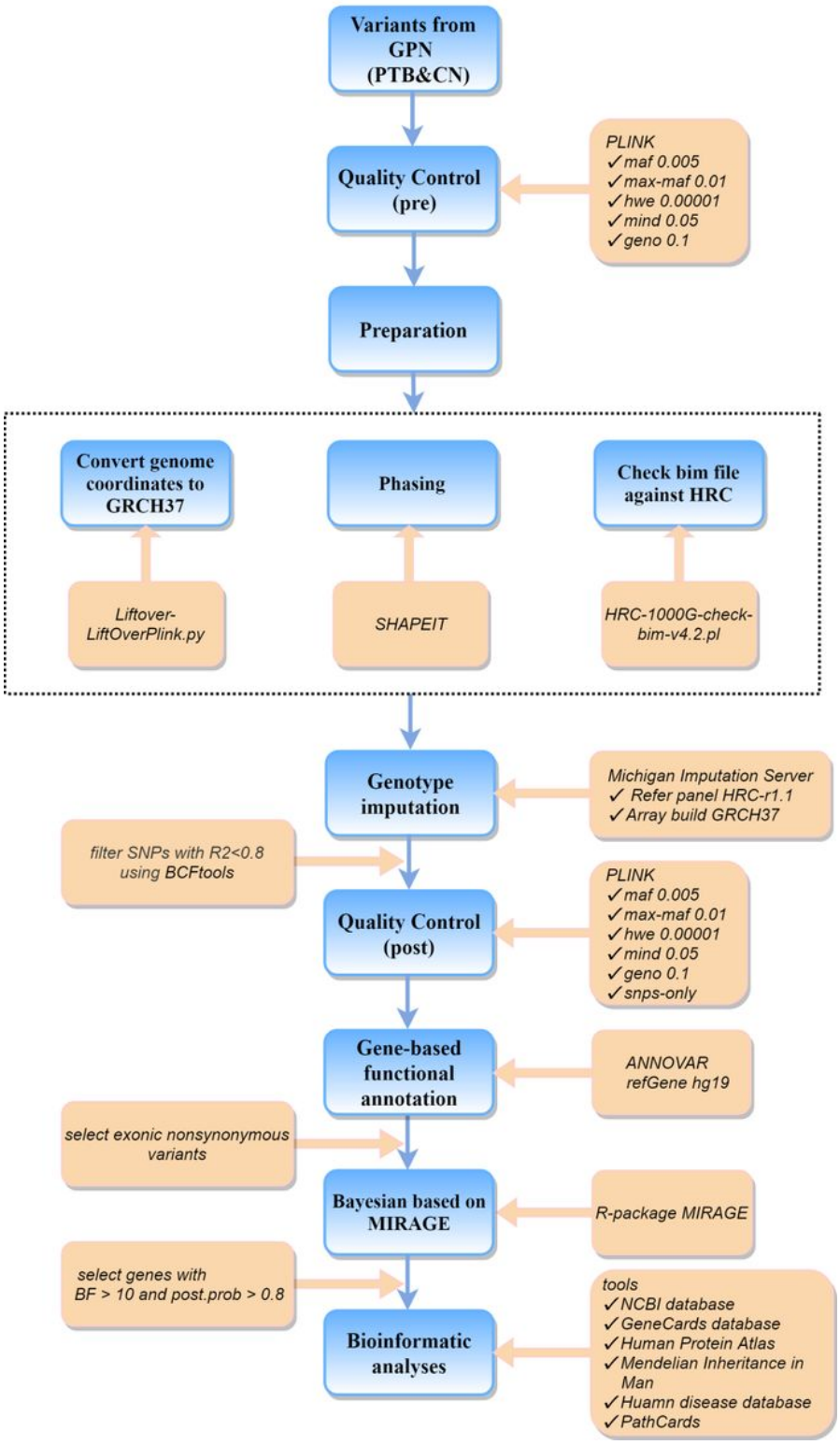

Figure 1

The complete Effective Gene-based Rare Variant Association analysis (EGRVA) pipeline for case{control studies of disease.