Flexibly encoded GWAS identifies novel nonadditive SNPs in individuals of African and European ancestry

Jiayan Zhou[1,2], Lindsay Guare[3], Andre Luis Garao Rico[1], Tomas Gonzalez Zarzar[1], Nicole Palmiero[1], Themistocles L. Assimes[2], Shefali Setia Verma[4], Molly Ann Hall[1,5,6]

1. Department of Veterinary and Biomedical Sciences, College of Agricultural Sciences, The Pennsylvania State University, University Park, PA 16802, USA
2. Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94304, USA
3. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
4. Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
5. The Huck Institutes of the Life Science, The Pennsylvania State University, University Park, PA 16802, USA
6. Penn State Cancer Institute, The Pennsylvania State University, University Park, PA 16802, USA

Abstract
Most genome-wide association studies (GWASs) assume an additive inheritance model, where heterozygous genotypes (HET) are coded with half the risk of homozygous alternate genotypes (HA), leading to less explained nonadditive genetic effects for complex diseases. Yet, growing evidence indicates that with flexible modeling, many single-nucleotide polymorphisms (SNPs) show nonadditive effects, including dominant and recessive, which will be missed using only the additive model. We developed Elastic Data-Driven Encoding (EDGE) to determine the HET to HA ratio of risk. Simulation results demonstrated that EDGE outperformed traditional methods across all simulated models for power while maintaining a conserved false positive rate. This research lays the necessary groundwork for integrating nonadditive genetic effects into GWAS workflows to identify novel disease-risk SNPs, which may ultimately improve polygenic risk prediction in diverse populations and springboard future applications to thousands of disease phenotypes and other omic domains to improve disease-prediction capability.

Introduction

Although genome-wide association studies (GWASs) have identified hundreds of thousands of genotype-phenotype associations, the majority of explained variance for most complex diseases remains hidden. Since 2008, the majority of GWAS apply the additive genetic model, which assumes the heterozygous genotypes have half the risk of homozygous alternate[1], thus limiting the discovery of the single-nucleotide polymorphisms (SNPs) with nonadditive inheritance patterns. Other genetic models, including the dominant and recessive, were used in the early days of GWAS along with additive encoding[2], but only the additive encoding was widely adopted as a single common approach for GWAS to reduce the testing burden[3–7]. As expected, it is more powerful to study the SNP under distinctive inheritance patterns using corresponded genetic model[7]. Using the typical additive encoding is insufficient to identify the alleles with recessive effects[7], even for common alleles[8]. Thus, using traditional encoding approaches to model genetic disease risk can limit the efficacy of GWAS using a multi-encoding-adjusted genome-wide significance at 1×10^{-8} (accounts for 5 tests for one SNP). Many other genetic models have recently been established to incorporate the additional risk of genetics in general; for example, the codominant encoding, which is a dummy encoding approach, allows heterozygous and homozygous alternate to bear full risk in a single genetic model[9,10], and dominance deviance (DOMDEV) is another common method in which the deviation of dominance from additivity is determined[11]. However, the codominant encoding is unable to provide summarized statistics for post-GWAS analysis, including meta-analysis and polygenic risk score calculation.

In the current study, we illustrated the use of Elastic Data-Driven Encoding (EDGE)[12] in identifying the genotype-phenotype associations with SNPs functioning as an additive and nonadditive inheritance patterns. Previously, EDGE was applied to a gene-gene interaction in helping with identifying SNPs' interactions with nonadditive inheritance patterns for common diseases[12], but hasn't been incorporated with a single variant association test. EGDE provided an opportunity to allow the assignment of a unique, flexible, and data-informed risk to the heterozygous genotype. It has broken down current barriers to modeling and identifying nonadditive SNPs in the interaction analysis and to counting their effects in post-GWAS analysis. Results demonstrate the advantage of using EDGE to detect the SNP-trait associations beyond the traditional encodings while the inheritance patterns of alleles are unknown and to flexibly encode SNPs based on their unique inheritance models in large-scale, cross-ancestry GWAS.

Methods

Elastic Data-Driven Encoding (EDGE)

As previously described[12], EDGE assigns a flexible calculated heterozygous to homozygous alternate ratio of risk, as α, to the heterozygous genotype based on the inheritance model each SNP represents in the α-calculation dataset using a codominant (dummy) encoding with covariates.

$$E(Y|SNP_{Het}, SNP_{HA}, COV_i) = \beta_0 + \beta_{Het}SNP_{Het} + \beta_{HA}SNP_{HA} + \sum_i \beta_{cov_i}COV_i \qquad (1)$$

$$\alpha = \frac{\beta_{Het}}{\beta_{HA}} \qquad (2)$$

Simulated datasets

To ensure that EDGE assigns the expected heterozygous genotype value across different types of underlying inheritance models, we simulated main effect SNPs with the following genetic models using the Biallelic Model Simulator (BAMS)[12] within the pandas-genomics: null (NULL), recessive (REC), sub-additive (SUB), additive (ADD), super-additive (SUP), and dominant (DOM) (1,000 simulated datasets for each model type) across varying minor allele frequencies (MAFs) Figure 1). Within each simulation, one SNP demonstrated a main effect, while another demonstrated a null effect with no interaction. To avoid overfitting, we simulated two sets of data for 1) EDGE's alpha calculation and 2) the application of alpha value to derive p-values and effect estimates. The existing encodings, including the additive, recessive, dominant, and dominance deviance (DOMDEV), were also applied to the second set of data for average p-value and power comparisons to EDGE. A variety of parameters were considered for mimicking the possible data structures with a combination of the following: MAF: 0.05, 0.1, 0.2, 0.3, and 0.4; sample size: 2000, 5000, 10000, 50000, and 100000; case-control ratio: 1:1 and 1:3, and penetrance difference (the difference between the minimum and maximum probabilities in the penetrance table): 0.05, 0.1, 0.175, 0.25, 0.33, and 0.4.

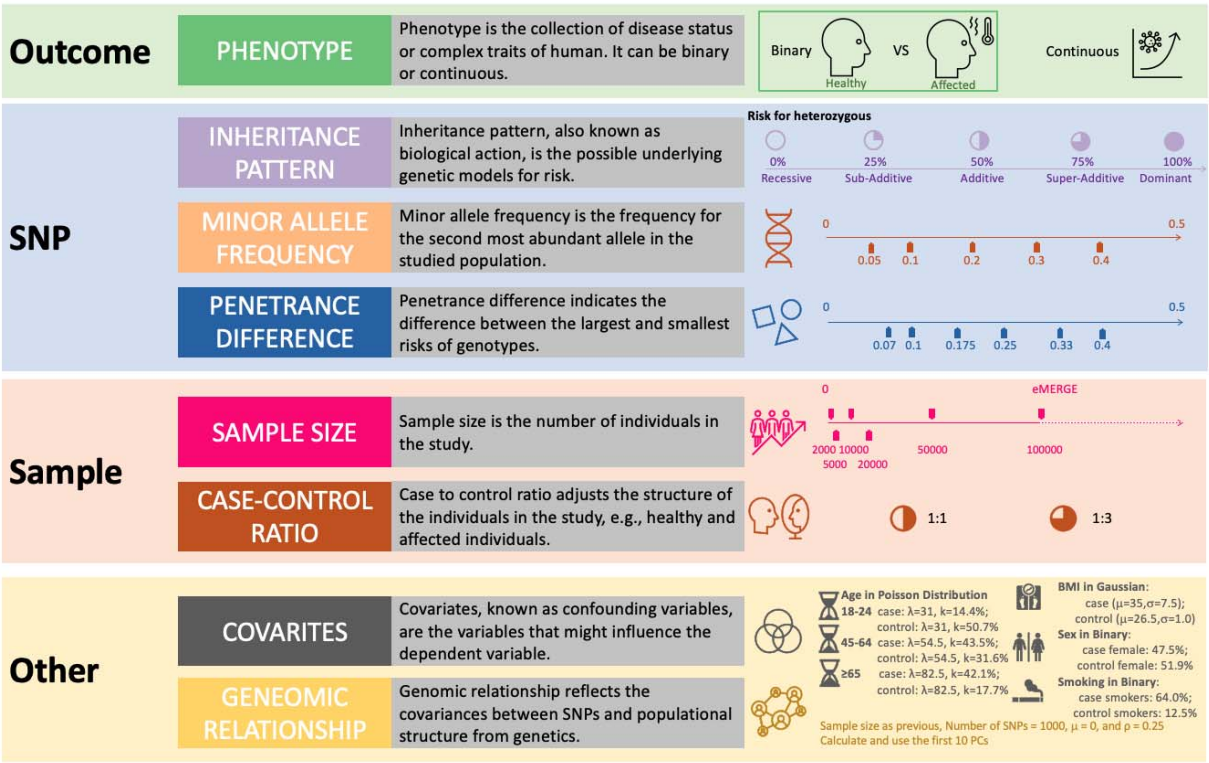

Figure 1. Workflow of the simulation study to understand the performance of the EDGE encoding compared to other encoding methods.

In addition, we simulated two continuous confounder variables and two binary variables to assess potential impact of these different types of covariates for cases and controls separately. The simulated covariates and distributions were referenced from the National Diabetes Statistics Report, 2020[13]. The age was simulated for three groups, including 18 to 44 (Poisson distribution; case: λ=31, k=14.4%, control: λ=31, k=50.7%), 45 to 64 (Poisson distribution; case: λ=54.5, k=43.5%, control: λ=54.5, k=31.6%), and ≥ 65 (Poisson distribution; case: λ=82.5, k=42.1%,

control: λ=82.5, k=17.7%). The body mass index (BMI) was simulated by considering the Gaussian distribution for case (μ=35,σ=7.5) and control (μ=26.5,σ=1.0) groups. Two binary covariates were generated as the sex (case female: 47.5%; control female: 51.9%) and Smoking status (case smokers: 64.0%; control smokers: 12.5%) by referencing the binomial distribution.
To presume the population structure derived from the genotyping data, we simulated a genomic relationship matrix (G-matrix), a variance-covariance matrix, with all covariance at 0.25 for 1000 SNPs and corresponding sample size for each simulation. The principal component analysis (PCA) was performed to calculate the first ten principal components (PCs), and those PCs were used as covariates to represent the relationships between individuals.

To further evaluate the performance of EDGE for the rare variants, we simulated and ran EDGE with SNPs with MAFs at 0.025, 0.01, 0.005, and 0.001, varying the sample size (from 2000 to 100000) with case to control ratio at 1:3, four preset covariates (age, sex, BMI, and smoking status), and six inheritance patterns (recessive, sub-additive, additive, super-additive, dominant, and heterozygous) for a thousand of replicates. The non-convergence rate was calculated by computing the ratio between the number of models that were not converged and the total number of replicates that could be finished.

Genome-wide association studies (GWAS)
We conducted GWAS analyses with a binary outcome using the logistic regression for simulated datasets under EDGE and other possible traditional encoding schemes with simulated covariates in age, BMI, sex, and smoking status, including the additive, recessive, dominant, and dominance deviance (DOMDEV), separately. Three different schemes were applied to the GWAS using simulated data, including genotyping only, genotyping with four designed covariates, and genotyping with four designed covariates and the first ten presumable PCs. The power was obtained for each combination of parameters. The false positive rate (FPR) was calculated from the simulation with null effect SNPs under every combination of encodings and parameters.

Variable importance on power and α calculation
Pairwise comparisons were performed to compare the power of each encoding using the simulated data. We only considered the simulated data with penetrance difference at 0.1 and sample size from 5000 to 50000 with all possible MAFs, case-control ratio, and all inheritance models. Overfitted random forest analyses were conducted to rank the importance of parameters regarding the power and α calculation using the *RandomForest* package in R v4.2.0. Variable importance was computed using the percentage of increase in mean squared error (MSE) and expressed relative to the maximum.

Results
EDGE detected common and rare variants for the greatest number of simulated models
A total of 2.1 million SNPs were simulated in the study to assess the EDGE performance compared with other traditional encodings. We first found density distribution of α peaks corresponded to the simulated inheritance models (REC $\cong$ 0, SUB $\cong$ 0.25, ADD $\cong$ 0.5, SUP $\cong$ 0.75, and DOM $\cong$ 1; Figure 2), demonstrating the efficacy of EDGE α to infer inheritance models for additive and nonadditive SNPs. We then contrasted the average power and significance of each encoding for SNPs with distinct desired inheritance patterns and MAF at 30% (Figure 3).

The MAF at 30% was considered by referencing the first reported associations for T2D with the E23K variant in *KCNJ11* (MAF= ~30% for Europeans)[19] and AMD with variants in *CFH* (MAF = 24.5% for Europeans)[20]. EDGE detected signals at genome-wide significance (5×10^{-8}) for the most simulated models. No method identified the SNPs with recessive inheritance models; however, EDGE identified all other SNP inheritance models, unlike any other method, including additive. EDGE was also the top method for every inheritance model. While the codominant encoding was a close second to EDGE here, our previous studies showed that the codominant encoding suffers from model non-convergence for lower MAFs[12]. Regarding inflation, EDGE demonstrated a conserved false positive rate of 5.4%, which falls within Bradley's liberal criteria of 2.5%-7.5%[21] (Figure 4).

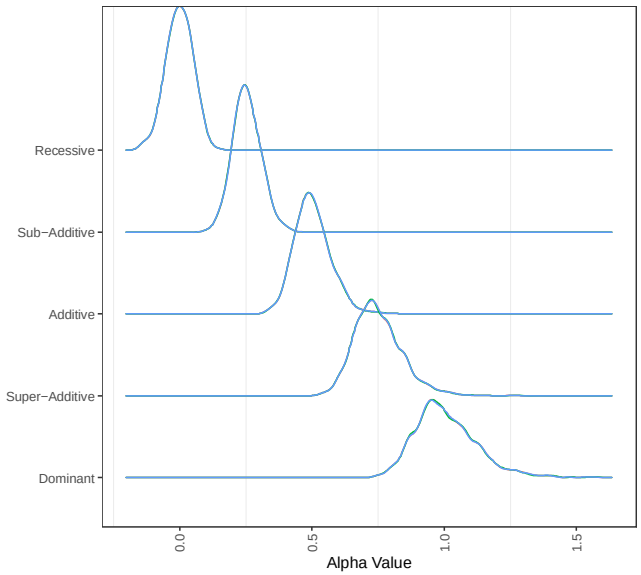

Figure 2. Distribution of EGDE's alpha values for SNPs with different inheritance patterns.

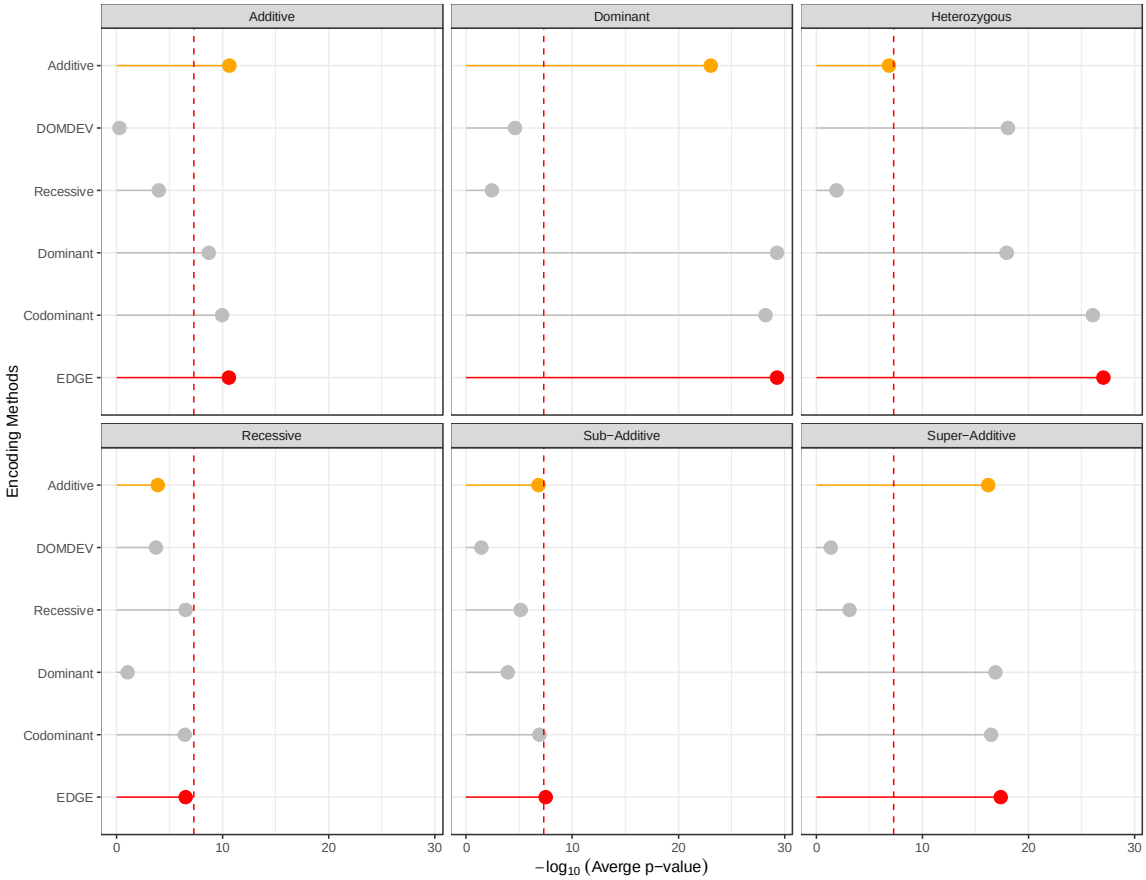

Figure 3. SNPs with MAF at 30% were simulated for 12500 cases and 37500 controls. The red dashed line represents the significance threshold at the genome-wide scale.

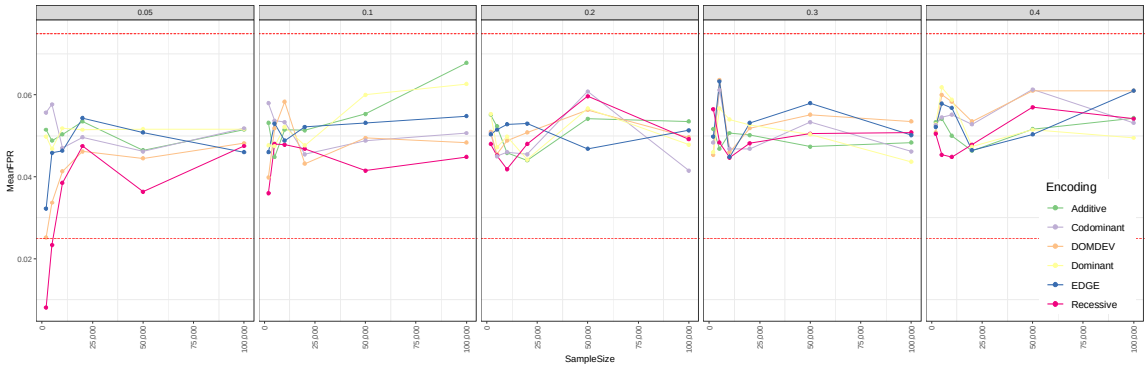

Figure 4. Inflation of using EDGE and other five encodings. Dash line represents Bradley's liberal criteria of 2.5%-7.5%.

We also observed that the power increases with sample size and MAF increases. Using the EDGE or codominant encodings could achieve the desired power of analysis at 80% more rapidly as the increasing the sample size than other traditional encodings for SNPs with the lowest MAF at 5% (Figure 5). We compared the difference in median power across all combinations of parameters between EDGE and additive using the Kruskal-Wallis χ^2 tests.

EDGE's power was significantly higher than additive (p: 8.7×10^{-5}; Figure 6). Especially, the EDGE could help to identify the SNPs with recessive inheritance patterns and heterozygous inheritance patterns that additive encoding fails to (Figure 7). Other traditional encodings suffer from the deficient power to detect SNPs with one or more inheritance patterns, such as the SNPs with recessive and sub-additive using dominant or DOMDEV encodings with limited samples. EDGE encoding has the robust power to identify SNPs under any inheritance patterns, especially the sub-additive and recessive ones that other encodings cannot with desired power. At low MAF and sample size, EDGE has a higher analysis power than other encodings to capture the associations between the SNPs and outcome. The additive encoding requires more than one hundred thousand samples to attain the desired power for SNPs with lower MAF.

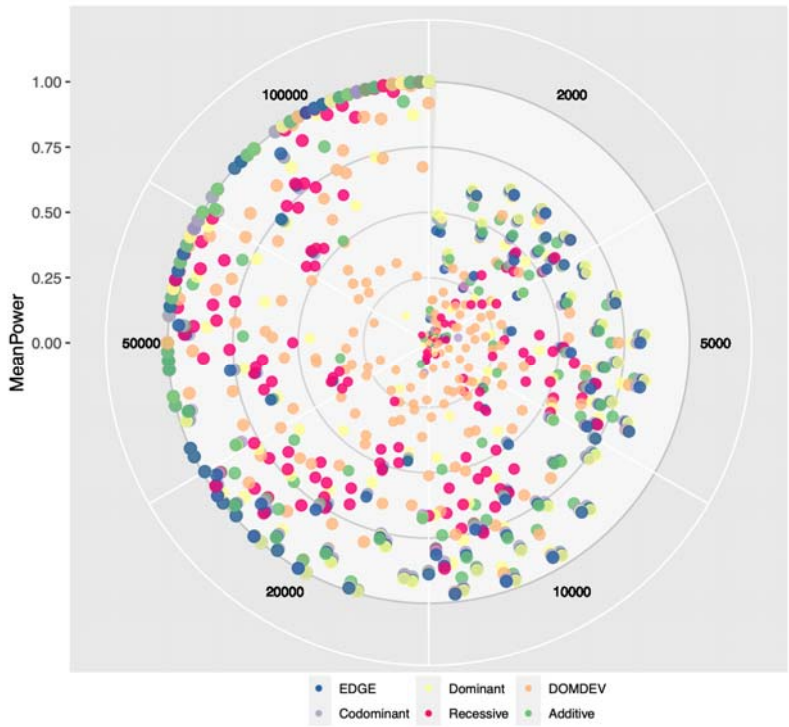

Figure 5. Power of EDGE across all simulated models and parameters with the existing encodings. Clockwise, the sample size increases from 2000 to 100000. In each sample size segment, the MAF increases left to right from 0.05 to 0.4. Power increases as we move from the center of the plot outward.

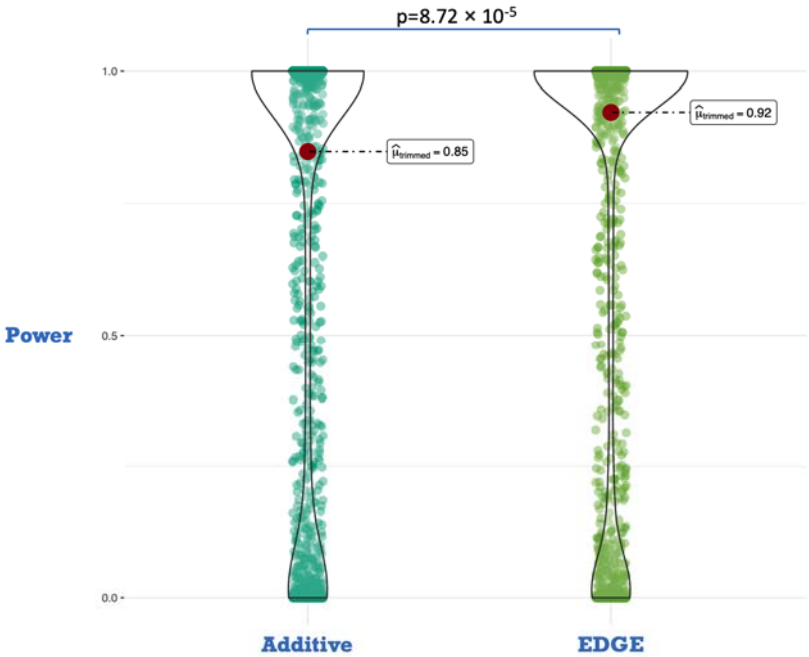

Figure 6. Power comparison between Additive and EDGE considering all possible inheritance models.

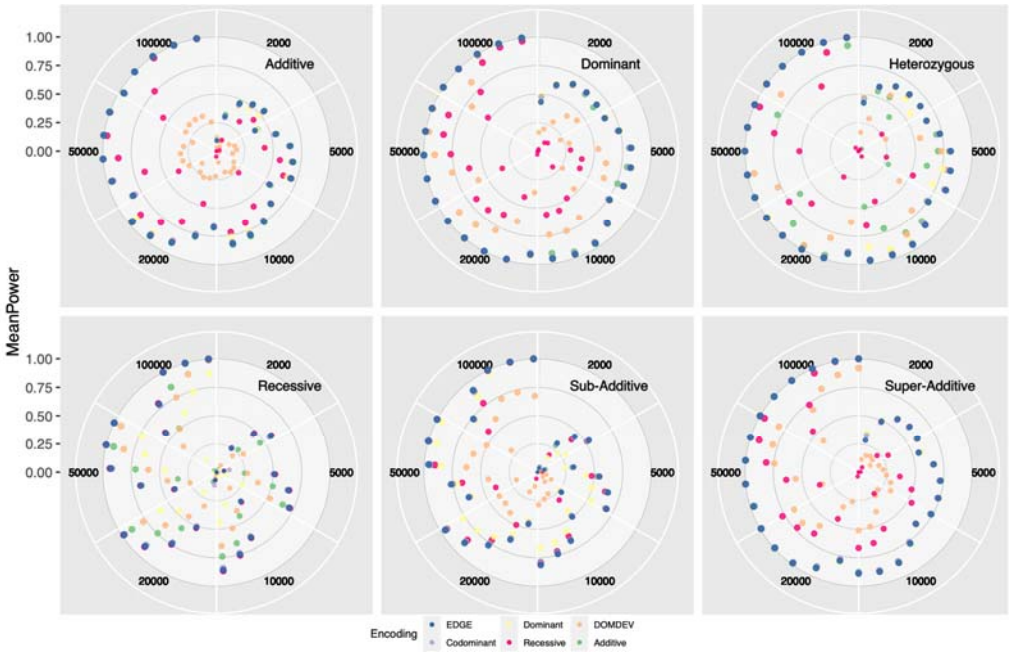

Figure 7. Power calculation by inheritance patterns.

We further estimated the convergence percentage for using EDGE in regression for SNPs as common and rare variants varying MAF and sample size. The convergence rate for EDGE is 99.35% while testing the common SNPs with a low MAF at 0.05 of 2000 participants (Figure

S1). We further simulated SNPs with extremely low MAF at 0.025, 0.01, 0.005, and 0.001 as rare variants at varied sample sizes with designed covariates. The convergence rate dropped below 50% for simulations with 2000 samples, while MAF achieved lower than 0.01. The rising of the participants might overcome the non-convergence of using EDGE for rare variants.

EDGE can find significance and retain the analysis power

We sought to validate the aptitude of using the EDGE strategy to identify the significant results with a small number of participants for SNP under a low MAF. EDGE can pinpoint more significant results than other encoding methods, particularly for SNPs with heterozygous, dominant, and super-additive inheritance patterns (Figure 8). The codominant encoding functions are similar to the EDGE encoding, which both have a higher analysis power and a calculated smaller p-value for the significant results. We restricted these analyses to only using codominant and EDGE encoding to compare the results' significance. EDGE raises the likelihood of revealing the significant results as more substantial than using codominant encoding with the smallest MAF (0.05, Kruskal-Wallis χ^2 = 34.24, p=4.87×10^{-9}) and smallest samples (2000, Kruskal-Wallis χ^2 = 72.27, p=1.88×10^{-17}) (Table 1). We also compared the performance of EDGE and codominant encodings using covariates and PCs. EDGE can consistently assign more significance to simulations with smaller MAF and sample sizes.

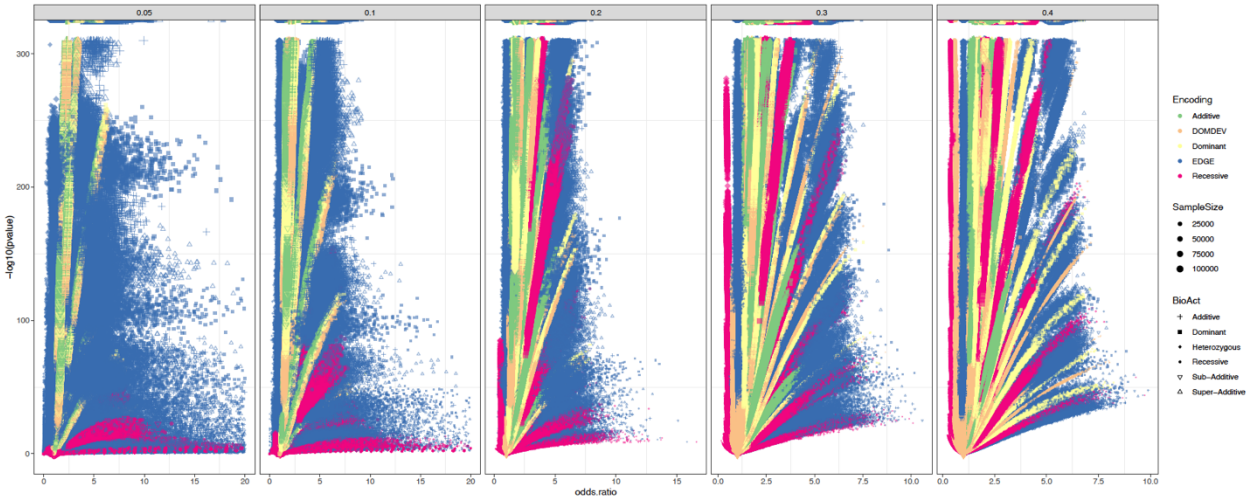

Figure 8. Volcano plots for showing the distribution of odds ratio and p-values.

Table 1. Comparison of the difference in assigning the significance using codominant and EDGE encodings with and without introducing covariates.

Parameter	Scale	SNP				SNP + COVs			
		Kruskal-Wallis chi-squared	p-value	Median of the p-value from GWAS		Kruskal-Wallis chi-squared	p-value	Median of the p-value from GWAS	
				Codominant	EDGE			Codominant	EDGE
MAF	**0.05**	**34.24**	**4.87×10^{-9}**	**8.92×10^{-12}**	**8.04×10^{-12}**	**82.86**	**8.81×10^{-20}**	**9.03×10^{-12}**	**8.84×10^{-12}**
	0.1	3.70	0.054	5.76×10^{-23}	3.05×10^{-23}	0.021	0.89	5.70×10^{-23}	3.07×10^{-23}
	0.2	0.59	0.44	3.09×10^{-42}	3.72×10^{-42}	0.0017	0.97	3.09×10^{-42}	3.75×10^{-42}
	0.3	0.08	0.77	1.91×10^{-55}	2.39×10^{-54}	1.24	0.27	1.94×10^{-55}	2.39×10^{-54}
	0.4	***9.51***	***0.002***	***1.22×10^{-63}***	***1.68×10^{-62}***	***0.49***	***0.48***	***1.22×10^{-63}***	***1.70×10^{-62}***
Sample Size	**2000**	**72.27**	**1.88×10^{-17}**	**4.54×10^{-6}**	**4.06×10^{-6}**	**82.86**	**8.81×10^{-20}**	**4.55×10^{-6}**	**4.17×10^{-06}**
	5000	0.031	0.86	2.24×10^{-13}	1.08×10^{-13}	0.021	0.89	2.24×10^{-13}	1.07×10^{-13}
	10000	0.00072	0.98	2.04×10^{-25}	7.98×10^{-26}	0.0017	0.97	2.06×10^{-25}	7.87×10^{-26}
	20000	1.24	0.27	2.70×10^{-49}	1.81×10^{-49}	1.24	0.27	2.66×10^{-49}	1.81×10^{-49}
	50000	0.50	0.48	1.61×10^{-119}	5.84×10^{-119}	0.49	0.48	1.56×10^{-119}	5.72×10^{-119}
	100000	2.14	0.14	9.82×10^{-237}	6.87×10^{-235}	2.12	0.15	9.84×10^{-237}	7.26×10^{-235}

We constructed three models to testify to the stability of the power and α-calculation for EDGE, including SNP, SNP with covariates, and SNP with covariates and pretend PCs as the population structure. We then executed two overfitted random forest paradigms to uncover the variable importance for the analysis power and α calculation. The sample size and penetrance difference influenced analysis power, followed by encoding methods, inheritance patterns of SNP, and MAF. α calculation was drastically affected by the MAF of the SNPs but not by their inheritance patterns. Both calculations were not altered by the changes in the case-control ratio and the covariates and PCs.

We evaluated the distribution of α values under different combinations of MAF, penetrance difference, sample size, and case-to-control ratio. α value for the single SNPs was steadily estimated regardless of the numbers and types of covariates and population structures. We observed distinct density peaks aligned with the simulated inheritance patterns, signifying that the α values from EDGE reflect the inheritance patterns of SNPs. As the increases of MAF, sample size, and baseline risk, the density peaks were shrunk to the desired inheritance patterns of the SNPs and more divergent from each other for genetic models. The fewer samples with a lower MAF and baseline risk of SNPs would lead to the great variability of the calculated α values.

Discussion

We developed a novel encoding to flexibly encode each SNP according to the inheritance model a SNP exhibits for a given phenotype, thereby increasing the ability to identify novel, nonadditive SNP particularly, EDGE assists in revealing the inheritance patterns underlying SNP-disease associations with conservative analysis power while these inheritance patterns remain unknown on fore. These SNPs may be ignored by using additive encoding due to their potential nonadditive genetic architecture. Under the multi-encoding GWAS, the multiple-test correction would have yielded results insignificant.

We noticed several limitations in our research. The regression model suffered from poorly fitting as a lack of convergence for rare variants (MAF < 0.001) with few participants. However, EDGE can provide a robust capability to cover the finding of common variants in the GWAS setting with an extremely small sample size. We also found that the sample size is suboptimal for the populations with smaller sample sizes (East Asian, South Asian, and Admixed American) for current datasets in the study. The consolidation of other available datasets of diverse ancestry is required to understand these diseases for different ancestry subgroups. Our current strategy for ancestry-based stratification combines individuals of Admixed Americans into one dataset and uses principal components as covariates. This may not model the diversity and richness of genetic risk profiles in this population. Therefore, additional approaches must be evaluated to overcome this limitation and include the full spectrum of diversity in our research.

In conclusion, EDGE helps to identify the SNPs with additive and nonadditive inheritance patterns. These results demonstrate that EDGE breaks down current barriers to identifying and modeling additive and nonadditive SNPs in GWAS, even for gene-environment interactions and polygenic risk score (PRS) research, by developing the methodological and computational developments necessary to make flexible SNP modeling accessible and commonplace in the

human genetics community. These innovations will enable improved prediction accuracy for diseases and complex traits and a refined understanding of the genetic architecture of these diseases across diverse populations with the potential for future applications to thousands of phenotypes.

Supporting Information

All data are available upon application from each biobank/consortium.

The genetic analyses were performed by using publicly available software tools, including PLINK v2.0 (https://www.cog-genomics.org/plink/2.0/), pandas-genomics (BAMS) v0.11.0 (https://github.com/HallLab/pandas-genomics), CLARITE v2.2.0 (https://clarite-python.readthedocs.io/en/latest/), and PLATO v2.1.0 (https://ritchielab.org/plato). The sample codes for calculating the α from EDGE and applying the α are available at https://github.com/HallLab. The plots were generated by using R v4.2.0 and python v3.7.

Supplement

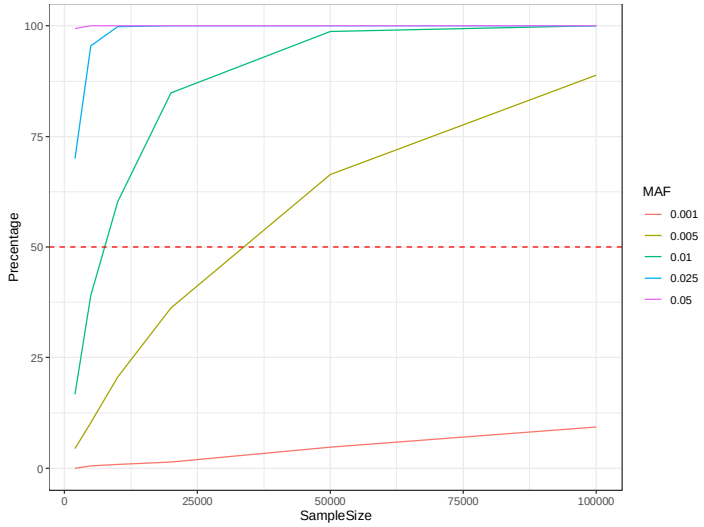

Figure S1. Convergence of using EDGE towards to the rare variants.

Acknowledgement
This work is supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under Project #PEN04275 and Accession #1018544, startup funds from the College of Agricultural Sciences, Pennsylvania State University (https://agsci.psu.edu/), and the Dr. Frances Keesler Graham Early Career Professorship from the Social Science Research Institute, Pennsylvania State University (https://ssri.psu.edu/) to MAH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions
J.Z. , L.G. , GR. AA. , GZ. T. , N.P. , T.L.A., S.S.V. and M.A.H. contributed to this project.

Competing interests
All authors have no competing interests.

Reference

1. Igo Jr., R. P., Kinzy, T. G. & Bailey, J. N. C. Genetic Risk Scores. *Curr. Protoc. Hum. Genet.* **176**, 100–106 (2019).
2. Arking, D. E. *et al.* A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat. Genet.* **38**, 644–651 (2006).
3. Bierut, L. J. *et al.* Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum. Mol. Genet.* **16**, 24–35 (2007).
4. Wallace, C. *et al.* Genome-wide Association Study Identifies Genes for Biomarkers of Cardiovascular Disease: Serum Urate and Dyslipidemia. *Am. J. Hum. Genet.* **82**, 139–149 (2008).
5. Uda, M. *et al.* Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1620–1625 (2008).
6. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLOS Comput. Biol.* **8**, e1002822 (2012).
7. Lettre, G., Lange, C. & Hirschhorn, J. N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet. Epidemiol.* **31**, 358–362 (2007).
8. Juran, B. D. & Lazaridis, K. N. Genomics in the post-GWAS era. *Semin. Liver Dis.* **31**, 215–222 (2011).
9. Lewis, C. & Lewis, C. M. Genetic association studies: Design, analysis and interpretation. *Brief. Bioinform.* **3**, 146–153 (2002).
10. Minelli, C., Thompson, J. R., Abrams, K. R., Thakkinstian, A. & Attia, J. The choice of a genetic model in the meta-analysis of molecular association studies. *Int. J. Epidemiol.* **34**, 1319–1328 (2005).
11. Clarke, G. M. *et al.* Basic statistical analysis in genetic case-control studies. *Nat. Protoc.* **6**, 121 (2011).
12. Hall, M. A. *et al.* Novel EDGE encoding method enhances ability to identify genetic interactions. *PLoS Genet.* **17**, e1009534 (2021).
13. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. *Natl. Diabetes Stat. Rep.* 2 (2020).
14. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med. 2013 1510* **15**, 761–771 (2013).
15. Stanaway, I. B. *et al.* The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet. Epidemiol.* **43**, 63–81 (2019).
16. Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.* **5**, 370 (2014).
17. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
18. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nat. 2021 5907845* **590**, 290–299 (2021).
19. Gloyn, A. L. *et al.* Large-scale association studies of variants in genes encoding the pancreatic β-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568–572 (2003).

20. Klein, R. J. *et al*. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (80-.).* **308**, 385–389 (2005).
21. Bradley, J. V. Robustness? *Br. J. Math. Stat. Psychol.* **31**, 144–152 (1978).