Title

PAGER: A novel genotype encoding strategy for modeling deviations from additivity in complex trait association studies

Permalink

https://escholarship.org/uc/item/4bc6z13d

Journal

BioData Mining, 17(1)

ISSN

1756-0381

Authors

Freda, Philip J
Ghosh, Attri
Bhandary, Priyanka
et al.

Publication Date

2024

DOI

10.1186/s13040-024-00393-x

Peer reviewed

https://doi.org/10.1186/s13040-024-00393-x

BioData Mining

RESEARCH **Open Access**

PAGER: A novel genotype encoding strategy for modeling deviations from additivity in complex trait association studies

Philip J. Freda[1†], Attri Ghosh[1†], Priyanka Bhandary[1†], Nicholas Matsumoto[1], Apurva S. Chitre[2], Jiayan Zhou[3], Molly A. Hall[4], Abraham A. Palmer[2,5], Tayo Obafemi-Ajayi[6] and Jason H. Moore[1*]

[†]Philip J. Freda, Attri Ghosh and Priyanka Bhandary contributed equally to this work.

*Correspondence:
Jason H. Moore
jason.moore@csmc.edu

[1]Department of Computational Biomedicine, Cedars-Sinai Medical Center, 700 N. San Vincente Blvd., Pacific Design Center, Suite G540, West Hollywood, CA 90069, USA
[2]Department of Psychiatry, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093-0667, USA
[3]Department of Medicine, Stanford University School of Medicine, 291 Campus Dr., Li Ka Shing Building, Stanford, CA 94305, USA
[4]Department of Genetics and Institute for Biomedical Informatics, University of Pennsylvania, 3700 Hamilton Walk, Richards Building A301, Philadelphia, PA 19104, USA
[5]Institute for Genomic Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093-0667, USA
[6]Cooperative Engineering Program, Missouri State University, 901 S. National Ave, Springfield, MO 65897, USA

Abstract

Background The additive model of inheritance assumes that heterozygotes (Aa) are exactly intermediate in respect to homozygotes (AA and aa). While this model is commonly used in single-locus genetic association studies, significant deviations from additivity are well-documented and contribute to phenotypic variance across many traits and systems. This assumption can introduce type I and type II errors by overestimating or underestimating the effects of variants that deviate from additivity. Alternative genotype encoding strategies have been explored to account for different inheritance patterns, but they often incur significant computational or methodological costs. To address these challenges, we introduce PAGER (Phenotype Adjusted Genotype Encoding and Ranking), an efficient pre-processing method that encodes each genetic variant based on normalized mean phenotypic differences between diallelic genotype classes (AA, Aa, and aa). This approach more accurately reflects each variant's true inheritance model, improving model precision while minimizing the costs associated with alternative encoding strategies.

Results Through extensive benchmarking on SNPs simulated with both binary and continuous phenotypes, we demonstrate that PAGER accurately represents various inheritance patterns (including additive, dominant, recessive, and heterosis), achieves levels of statistical power that meet or exceed other encoding strategies, and attains computation speeds up to 55 times faster than a similar method, EDGE. We also apply PAGER to publicly available real-world data and identify a novel, relevant putative QTL associated with body mass index in rats (*Rattus norvegicus*) that is not detected with the additive model.

Conclusions Overall, we show that PAGER is an efficient genotype encoding approach that can uncover sources of missing heritability and reveal novel insights in the study of complex traits while incurring minimal costs.

Keywords Association studies, Case-control, Dominance, Genetics, Genotype encoding, GWAS, Heterosis, Inheritance, QTL analysis, Quantitative traits

Background

While the majority of diallelic single nucleotide polymorphisms (SNPs) likely follow an 'additive' model of inheritance, where the heterozygote's phenotype falls between those of the homozygotes [1, 2], there have been numerous observations of variants that deviate from this pattern, affecting a wide range of traits and systems [3–13]. These deviations collectively contribute significantly to explained phenotypic variance (V_P) in various systems, including humans [8, 11, 12]. Although detection of non-additive genetic variation is common, most analytical methods used for studying genotype-phenotype associations, such as quantitative trait locus (QTL) analysis and genome-wide association studies (GWAS), assume strict additivity at all genetic loci [2, 14–16]. However, if sources of non-additive genetic variation are left unaccounted for, the predictive accuracy of resulting models can be greatly reduced [17, 18]. This, in turn, can lead to misinterpretations of the effects of specific genetic variants via inadequate estimations of explained V_P [19] that could manifest as type I and type II errors.

In a diallelic system, genotype classes are typically encoded assuming the additive model of inheritance by the numbers 0, 1, and 2. Here, "0" represents homozygotes for the major allele (AA), "1" represents heterozygotes (Aa), and "2" represents homozygotes for the minor or alternate allele (aa). This numerical encoding effectively counts copies of a coded allele (usually the minor allele) present for an individual at a locus, facilitating the calculation of allele/genotype frequencies and the assessment of genetic relatedness among individuals. However, from a modeling perspective, the additive encoding presupposes that the phenotype (or disease risk) of heterozygotes is exactly intermediate in respect to the phenotypes of the two homozygote classes [14, 15]. Consequently, the regression models used in most GWAS and QTL analyses assume such additivity applies universally across all loci. However, the additive model is merely one example of the infinite phenotypic relationships that can occur between three genotypic classes. Other standard examples of inheritance, such as dominant and recessive models, arise when the effect of one allele partially or fully masks that of the other allele on the phenotype [14, 15]. These manifest as heterozygotes having phenotypes closer to one of the homozygote classes. Eye color in humans is mainly attributable to dominance effects [20]. Although not as common in the literature compared to dominant and recessive models, heterotic models of inheritance occur when the phenotypes of heterozygotes surpass or fall below (higher or lower fitness) those of either homozygote [14, 15]. Heterotic relationships are commonly referred to as either heterozygote advantage (also hybrid vigor) or heterozygote disadvantage depending on the phenotype and environment. Overall fitness regarding malarial resistance and the sickle-cell trait is a common example of heterosis in humans [21]. Additive, dominant, and recessive models serve as benchmarks for describing inheritance patterns in diallelic systems, with heterotic models also considered to a lesser extent. Yet, in theory, inheritance models encompass a spectrum wherein the relative phenotypic differences between genotypic classes can exhibit continuous variability, capable of being described by any combination of numeric values [14, 15].

Efforts have been made to model the inheritance patterns of individual SNPs in genetic analyses using alternative encoding strategies. However, these methods have not gained widespread popularity primarily because of two factors. Firstly, encoding methods may struggle in identifying significant genetic variants beyond what a standard additive model can detect. In other words, the strategic effort of encoding may not be justified, as alternative encoding methods may fail to identify significant associations or could even reduce power by assuming inferior models [22]. This is particularly true when using a single alternative inheritance model, such as dominant or recessive, which impose their own strict inheritance assumptions to all loci. Therefore, it may seem prudent to employ multiple models. However, this approach comes with the cost of correcting for multiple tests incurred by each additional model implemented. Secondly, encoding may require significant computational resources. Each SNP can theoretically exhibit its own unique inheritance pattern [14, 15], and modeling these across an entire dataset, and for multiple phenotypes, can be expensive as computational costs increase with both sample size and SNP count. Collectively, these challenges reduce the practical viability and utilization of alternative encoding approaches.

Despite the issues outlined above, a promising genotype encoding strategy recently developed is Elastic Data-driven Genetic Encoding (EDGE) [23, 24]. EDGE estimates the genotype encoding of heterozygotes dynamically using a data-driven pre-processing approach, alleviating the cumulative multiple testing burden associated with assuming multiple inheritance models. Although EDGE was developed with investigating genetic interactions (epistasis) as a central goal, it is extensible to univariate tests like GWAS and QTL analysis. EDGE is flexible in that heterozygote classes are recoded to reflect the accurate genotype-phenotype relationship of each SNP. This is achieved by building a specific linear or logistic model (depending on the phenotype of interest) for each SNP where the heterozygous encoding (SNP_{HET}: AA=0, Aa=1, aa=0) and the homozygous alternate encoding (SNP_{HA}: AA=0, Aa=0, aa=1) are both implemented. The vectors of these codominant dummy encodings are then used to construct the model shown in Eq. 1.

$$\begin{aligned} &E\left(Y|SNP_{Het}, SNP_{HA}, COV_i\right) = \\ &\beta_{Het}SNP_{Het} + \beta_{HA}SNP_{HA} + \sum_i \beta_{COV_i}COV_i \end{aligned} \tag{1}$$

Where $E(Y)$ is the expected value of the phenotype (Y) given the predictors in the model, SNP_{HET} is SNP vector using the heterozygous encoding, SNP_{HA} is the SNP vector using the homozygous alternate encoding, β_{Het} and β_{HA} are the regression coefficients for the heterozygous and homozygous alternate encodings, respectively, COV_i is the ith covariate (e.g., age or sex), and β_{COVi} is the corresponding regression coefficient. By fitting this regression model for each SNP using the phenotype as the dependent variable, EDGE estimates the effect sizes (beta coefficients) associated with the heterozygous (β_{Het}) and homozygous alternate (β_{HA}) genotype encodings. The ratio α captures the relative contribution of the heterozygous encoding to the phenotype compared to the homozygous alternate encoding, effectively quantifying deviations from the additive inheritance model (Eq. 2).

$$\alpha = \frac{\beta_{Het}}{\beta_{HA}} \tag{2}$$

This process allows EDGE to adjust the encoding of the heterozygous genotype based on its observed relationship with the phenotype. Specifically, the resulting α value replaces the heterozygous encoding for that SNP, while AA individuals remain scored as 0 and aa individuals are scored as 1. This recoding reflects the SNP's specific inheritance pattern as observed in the data. After recoding, univariate association tests can then be performed using the adjusted genotype encodings.

EDGE achieves conservative false-positive rates (FPR) and outperforms other inheritance models, including additive, dominant, and recessive, in terms of power [23, 24]. Nonetheless, EDGE's limitations are noteworthy. The computational time required to construct regression models to encode each SNP when considering large sample sizes and hundreds of thousands to millions of variants is significant. Furthermore, EDGE is designed for diallelic systems, limiting its applicability as many systems and variant types are multi-allelic. Additionally, EDGE is an integral component of a Python [25] software pipeline designed with a primary focus on investigating epistasis, making it somewhat challenging to apply EDGE to external datasets. Finally, while EDGE does eliminate the need for multiple encoding strategies, its use of regression models, and the necessary statistical tests required to determine α values from the phenotype, impose a distinct multiple testing burden additional to the conventional multiple test correction required in GWAS/QTL analyses. To mitigate this extra layer of testing, either significance thresholds can be made stricter, or alpha values can be estimated in a prior sample before applying the encodings to GWAS data. Both strategies significantly reduce the power to identify meaningful associations but are required.

To address the issues of speed and extensibility while also providing a robust genotype encoding strategy, we present Phenotype-Adjusted Genotype Encoding and Ranking (PAGER). PAGER employs a straightforward and computationally efficient data-driven mathematical approach to compute the mean normalized relative phenotypic differences between genotypic classes on a SNP-by-SNP basis. Like EDGE, PAGER is a flexible, dynamic pre-processing step that reduces the burden of applying multiple encoding strategies and can be used for both binary and continuous phenotypes as well as for epistasis investigation. While PAGER shares EDGE's advantage of eliminating the need for multiple encoding models, it also incurs the cost of deriving encodings from the phenotype, necessitating an additional layer of multiple test correction. However, PAGER offers significant improvements over EDGE, including faster computation speeds, easier applicability to real-world datasets, and extensibility to multi-allelic systems and variants. This paper highlights PAGER's advantages by demonstrating its efficiency, with computation speeds up to 55 times faster than EDGE, while maintaining similar or better performance across all relevant metrics, including a lack of false positive inflation. We also conduct exploratory GWAS on publicly available data from the model system, *Rattus Norvegicus* and compare the results of using three different models—additive, recessive, and dominant—to those obtained with only EDGE and PAGER. The results show that both PAGER and EDGE uncover significant genotype-phenotype associations not detected by the standard additive encoding alone. PAGER also detects all putative QTL identified by the three uniform inheritance models (i.e., those that assume the same pattern of inheritance at all loci - additive, dominant, recessive, etc.) tested. These findings underscore the advantages of using flexible and dynamic encoding strategies despite their associated correction costs.

Methods

The PAGER method

PAGER uses the relative difference between the mean phenotype of each genotypic class (AA, Aa, and aa) per SNP to construct encodings using the following formulae adapted from Cohen's *d* [26]

$$PAGER\ Encoding_{AA} = \frac{(\bar{x}_{AA} - \bar{x}_{AA})}{\sigma_{Phenotype}} = 0\ (anchor) \tag{3}$$

$$PAGER\ Encoding_{Aa} = \frac{(\overline{x}_{Aa} - \overline{x}_{AA})}{\sigma\ _{Phenotype}}$$

$$PAGER\ Encoding_{aa} = \frac{(\overline{x}_{aa} - \overline{x}_{AA})}{\sigma\ _{Phenotype}}$$

where $\overline{x}_{AA}$, $\overline{x}_{Aa}$, and $\overline{x}_{aa}$ are the mean phenotype values (proportion of cases [$\hat{p}$] in case/control studies) for the AA (0), Aa (1), and aa (2) genotype classes per SNP, respectively and $\sigma_{Phenotype}$ is the standard deviation of the entire phenotype vector. However, unlike Cohen's *d*, our goal is to calculate the mean relative difference between genotypic classes and not the effect sizes between means. Thus, the process is simplified by removing the $\sigma_{Phenotype}$ term from each calculation as this term is redundant:

$$PAGER\ Encoding_{AA} = \bar{x}_{AA} - \bar{x}_{AA} = 0\ (anchor) \tag{4}$$

$$PAGER\ Encoding_{Aa} = \overline{x}_{Aa} - \overline{x}_{AA}$$

$$PAGER\ Encoding_{aa} = \overline{x}_{aa} - \overline{x}_{AA}$$

The three new encodings are min/max normalized between 0 and 1 for clarity and interpretability. Note, the AA genotype encoding for each SNP remains 0 before normalization as it serves as the anchor point for relative difference calculations. The normalized PAGER encodings replace the original 0 (AA), 1 (Aa), and 2 (aa) encodings for each SNP in the dataset. PAGER is fully extensible to any programming language, to genetic systems beyond diallelic, and for constructing multi-locus genotypes (MLG) for investigating epistasis. In the case of epistasis, PAGER can be used to generate new features that numerically describe the interaction between *n* loci. The only universal requirement to extend PAGER to multi-allelic systems and epistasis detection is that one genotype or MLG be chosen as the anchor and all relative differences then be calculated and normalized. In the case where only two genotypes exist at a particular SNP, PAGER encodes one as the anchor (0) and the other as 1. If only one genotype exists, PAGER ignores that SNP and moves to the next in the dataset. However, instances like these should be removed from analyses as low or zero-variance SNPs reduce power and/or could lead to type I and type II errors [16, 27].

Simulated datasets

For our various tests to compare PAGER to EDGE and to assess PAGER's power and ability to characterize multiple inheritance models, we generate true-positive main effect SNPs with both binary and continuous phenotypes using the Biallelic Model Simulator (BAMS) within the Pandas-Genomics Python [25] package [28]. This package is also where EDGE can be natively used. To avoid overfitting of EDGE and PAGER, two sets of data are simulated for (1) calculating EDGE alpha values and PAGER encoding values (training) and (2) applying these values to derive test statistics (validation). All other uniform encodings are also applied to the validation sets for comparison purposes. SNPs are simulated under eight distinct inheritance patterns: additive, subadditive, superadditive, dominant, recessive, heterosis, underdominant, and overdominant (File S1). Each set of simulations (training and validation) include varying sample sizes (2000, 5000, 10000, 15000, 20000, 25000, 50000), minor allele frequencies (MAF) (0.1, 0.2, 0.3, 0.4, 0.5), penetrance differences (PEN_DIFF) (0.1, 0.25, 0.33, 0.4), representing the difference between maximum and minimum probabilities (noise) in the penetrance table, and case/control ratios of 0.25 and 0.5 for binary phenotypes. For each of the eight inheritance patterns, we simulate 140,000 SNPs with a continuous phenotype (140 unique combinations with 1,000 replicates each) and 280,000 SNPs with a binary phenotype (280 unique combinations with 1,000 replicates each). Thus, in total, we simulate 1,120,000 continuous phenotype SNPs and 2,240,000 binary phenotype SNPs in both training and validation sets. These simulated SNPs are utilized for performing inheritance model and power experiments.

Computation time comparisons

To assess the computational efficiency of PAGER and compare it to EDGE, we measure the total time to calculate encoding values from single SNPs from the training set for each method using 100 replicates (for tractability due to GPU testing), across all combinations of test variables (sample size, MAF, PEN_DIFF, and case/control ratio (binary only)) and inheritance models for both binary and continuous phenotypes. We perform these calculations using a CPU with an Intel® Xeon® Gold 6342 Processor (2.8 GHz). Since PAGER can leverage GPU integration, we also measure PAGER's compute time using the same high performance computing system's GPU (NVIDIA® Tesla® V100 SXM2–32GB). We compare computation times for EDGE, PAGER CPU, and PAGER GPU by calculating the speed factor increase of PAGER over EDGE (EDGE computation time/PAGER computation time). We observe that computation times only vary significantly as sample size increases, hence, we average computation times across all levels of case/control ratio (binary only), PEN_DIFF, MAF, and inheritance model for both phenotypes (File S2). Instances where computation times are more than three standard deviations from the mean in PAGER GPU tests are removed as these represent instances of GPU initialization and are extreme outliers. This results in 223,580 and 111,980 replicate SNPs with a binary phenotype and a continuous phenotype, respectively.

We also measure total computation time between EDGE and PAGER (CPU and GPU) when sample size is kept constant but SNP number increases (1, 10, 100, 1,000, 10,000, and 100,000). Because BAMS can only generate single SNPs with an association to a phenotype, we use SNPs randomly selected from a publicly available real-world dataset from an outbred, related rat (*R. norvegicus*) population of males and females derived from eight inbred founders (Heterogenous Stock [29]) and used in a previously published GWAS investigating obesity-related traits [30, 31]. We measure the time to derive EDGE and PAGER values from 3,166 individuals for all datasets of increasing SNP size. We use the continuous phenotype of body mass index (BMI) including the animal's tail

(BMI_Tail) for this analysis and perform the same comparisons with the same hardware as the previous experiment where sample size varies.

Assessing PAGER's proficiency at characterizing a range of theoretical inheritance models

To determine if PAGER's encoding strategy can suitably capture various inheritance models (File S1), we illustrate the distributions of the normalized PAGER encodings for each genotypic class for both phenotypes under the eight inheritance patterns used to simulate our training datasets with box plots in ggplot2 [32] in R [33]. Box plots are created by aggregating PAGER values across all variations of case/control ratio (binary only), MAF, PEN_DIFF, and sample size. We compare these box plots to the theoretical genotype relationships of all eight inheritance models (File S1) used to generate simulated data.

Heterozygote encoding values of PAGER and EDGE are also compared to determine if PAGER can more accurately describe inheritance models compared to EDGE. Mean heterozygote encoding values for EDGE and PAGER under each inheritance model across all experimental parameters, where MAF is set to 0.1, and where MAF and PEN_DIFF are both set to 0.1 are compared using Wilcoxon rank sum tests in R and descriptive statistics including means, variances, standard deviations, and ranges are also calculated (File S2). We separate out instances where MAF is set to 0.1 and where MAF and PEN_DIFF are both set to 0.1 to investigate how both encodings handle more difficult SNP simulations where allele frequencies and noise levels are more extreme.

Evaluating PAGER's ability to detect genome-wide significance

To assess PAGER's overall signal potential compared to other encoding methods, we calculate power (rate of success in identifying a significant association between a simulated true-positive SNP and the phenotype) from p-values derived from univariate regression (logistic for binary phenotype and linear for continuous phenotype) on SNPs using the *association_study* function within the CLARITE Python package [34, 35]. Regression is performed on encoded true-positive main effect SNPs simulated from the validation set using all eight inheritance models, all experimental variables, and both phenotypes within BAMS. Encoding strategies include the standard additive model, inherent encoding, EDGE, and PAGER. The inherent encoding is the encoding method derived from the respective inheritance model being analyzed. For example, if SNPs are generated using a dominant inheritance model, the inherent encodings are 0, 1, and 1 for the AA, Aa, and aa genotypes, respectively (see File S1 for all inherent encodings). Since the additive encoding is already inherent to the additive inheritance model, additive-generated SNPs are only tested with three encodings - additive, EDGE, and PAGER. We assess mean power from the validation set across all levels of PEN_DIFF and with PEN_DIFF equal to 0.1 by case-control ratio (binary only), MAF, and sample size. Power is obtained by counting instances where the CLARITE-derived regression p-value for each encoding is less than the genome-wide significance threshold of 5×10^{-8} under each inheritance model for both phenotypes. This count is divided by the total number of SNPs in each variable combination to obtain each encoding's power. PEN_DIFF of 0.1 is separately illustrated because it exhibits the highest variance in performance, allowing better comparison across encodings as power fluctuates significantly at this level (File S2). The p-value cutoff of 5×10^{-8} ($-\log_{10}p = 7.3$) is selected as it represents a realistic level of experiment-wide significance for a simulated human GWAS [36]. This cutoff was also used for the initial EDGE testing [23, 24], providing a basis of comparison.

Comparing false positive rates across encodings

To assess PAGER's inflation in terms of false positive rates (FPR) and if it is comparable to other encodings, we simulate 1,000 null effect (true-negative) SNPs with BAMS with both binary and continuous phenotypes across all combinations of experimental parameters (total of 280,000 binary-phenotype SNPs and 140,000 continuous-phenotype SNPs) and encode with additive, EDGE, and PAGER. As with the other analyses, EDGE and PAGER values are derived from a training set and then applied to a validation set. These encoded null effect SNPs are regressed with the *association_study* function in CLARITE used in the power experiment. The mean FPR for each combination of sample size and MAF is calculated by counting times each encoding produces a p-value ≤ 0.05 and dividing by the number of combinations of PEN_DIFF and case/control ratio (binary only) (two levels of case/control by four levels of PEN_DIFF by 1,000 iterations = 8,000 for binary and 4,000 for continuous).

Application to real-world data

To explore if EDGE and PAGER have the capabilities to reveal genome-wide significant variants that elude single models and to compare GWAS results between encodings and methods, we use all LD-pruned SNPs (128,401 variants) and phenotype data from the rat GWAS [30, 31] used for the computation time experiment. We selected this dataset because the controlled genetic and environmental factors in the breeding design reduce confounding variables commonly present in human studies, allowing for a clearer assessment of the effectiveness of both PAGER and EDGE. We choose BMI_Tail, body weight (BW), retroperitoneal fat (RetroFat) as our phenotypes of interest as they have the largest sample size (n = 3,166), relatively high heritability, and numerous putative QTLs detected in the original GWAS study. Here, we perform GWAS by applying three inheritance model encodings: additive (AA = 0, Aa = 1, and aa = 2), dominant (AA = 0, Aa = 1, and aa = 1), and recessive (AA = 1, Aa = 1, and aa = 0) and compare these GWAS results to those when only applying EDGE and PAGER. These three inheritance encodings are chosen as they are likely to collectively reflect most inheritance relationships that exist in natural and laboratory populations [1, 3, 4, 37, 38]. Thus, we compare three distinct GWAS methods: tri-encoding, EDGE, and PAGER across all three phenotypes.

Prior to performing GWAS, GEMMA software [39] is used to produce a genetic relatedness matrix (GRM) to account for relatedness between individual rats based on allele count (i.e., the standard additive encoding (0, 1, and 2)). Although we apply different genotype encodings in the fixed effects of our GWAS models, the GRM is computed once using the standard additive encoding to capture the overall genetic relatedness among individuals. This approach allows us to control for population structure and kinship in the mixed-model association analysis, ensuring that any associations detected are not confounded by relatedness. GWAS are performed using encoded genotypes from each method in GEMMA, with the generated GRM, using the leave-one-chromosome out (LOCO) method to avoid proximal genetic contamination [40, 41]. The genome-wide significance threshold across experiments is determined by Bonferroni correction ($0.05/128{,}401 = 3.89 \times 10^{-7}$; $-\log_{10}p = 6.41$). Note: this is stricter than the significance threshold used for the original GWAS, -$\log_{10}p=5.6$ [30, 31]. For the tri-encoding GWAS, this threshold is further divided by three (1.30×10^{-7}; -$\log_{10}p=6.89$) to account for the multiple testing of three models. Since both EDGE and PAGER use the phenotype to derive genotype encodings, there is an increased risk of overfitting. To address this and for accurate comparison to the tri-encoding GWAS, the Bonferroni threshold is halved (1.95×10^{-7}; -$\log_{10}p=6.71$) for the EDGE GWAS and the PAGER GWAS. GWAS QTL independence for each encoding strategy is determined by performing conditional analysis using custom R scripts adapted from the original GWAS study [30] in which top signal SNPs are used as covariates for subsequent GWAS to control for QTL on the same chromosome. Putative QTLs are identified, validated, and compared by calculating LD intervals and assessing overlap between encodings. LD intervals are calculated from original genotype files in PLINK [42] by scanning upstream and downstream of the peak parker for SNPs with a correlation coefficient (r^2) greater than or equal to 0.6. Beyond these points, the LD interval ends. If peak markers of putative QTL from one encoding strategy have overlapping LD intervals on at least one end with peak markers of putative QTL from another encoding strategy, these QTL are given the same identification number for comparison purposes.

Gene set enrichment analysis (GSEA) is performed to retrieve gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for novel QTL identified by PAGER, EDGE, or any alternative encoding model other than additive. This allows us to determine if novel loci detected by alternate encoding strategies are associated with terms and pathways that reveal relevant biological insights related to BMI, metabolism, and/or obesity. LD intervals (as determined for the QTL comparison and identification step) of these loci are submitted to the Rat Genome Database (https://rgd.mcw.edu/), assembly Rnor 6.0, using custom R scripts to retrieve gene models for enrichment. GO term and KEGG pathway enrichments are performed in R using BioConductor [43] and clusterProfiler [44] packages with FDR correction (corrected p-value and q-value cutoffs=0.05).

Results

PAGER demonstrates progressive speed improvements over EDGE

Both CPU and GPU integrations of PAGER consistently show speed increases over EDGE in all combinations of case/control ratio (binary only), sample size, MAF, PENN_DIFF, and inheritance model with the factor of the speed-up increasing with sample size (Fig. 1; File S2). PAGER GPU speed increases over EDGE range from 11.7x at a sample size of 2,000 to 52.7X at 50,000 individuals with a binary phenotype (Fig. 1A) and 10.9x to 54.6x with a continuous phenotype (Fig. 1B). A similar trend of increasing speed up as sample size increases is also observed for PAGER CPU where speed increases range from 7x to 30x with a binary phenotype (Fig. 1A) and 6.4x to 29.5x with a continuous phenotype (Fig. 1B).

Computation time increases do not scale with SNP number as observed with sample size (File S2). However, a considerable increase is observed with the GPU at 100,000 SNPs (4.4X at 10,000 SNPs to 6.4X at 100,000 SNPs; File S2). This speed increase is likely observed because, in general, VRAM is faster than system RAM. This, coupled with VRAM's proximity to the GPU cores, loading the dataset in VRAM can improve the performance and efficiency of GPU processing applications which can be observed as data

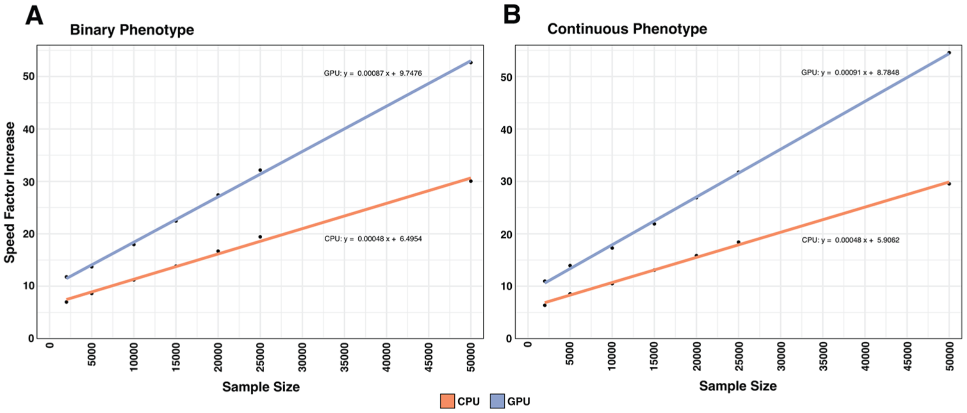

Fig. 1 Mean computation time comparisons, as speed factor increases of PAGER over EDGE, across all experimental variables as sample size increases for (**A**) binary and (**B**) continuous phenotypes. CPU speed factors are in orange and GPU in blue

size increases. Mean SNP-wise speed increases of PAGER over EDGE are 4.4X and 4.8X for the CPU and GPU, respectively (File S2).

On average, from the increasing sample size experiment, the times to encode a single discrete SNP are 0.072 s, 4.5×10^{-3} seconds, and 2.7×10^{-3} seconds for EDGE, PAGER CPU, and PAGER GPU, respectively as sample size increases from 2,000 to 50,000 (File S2). For continuous SNPs, mean encoding times are 0.067 s for EDGE, 4.4×10^{-3} seconds for PAGER CPU, and 2.6×10^{-3} seconds for PAGER GPU (File S2). Total encoding times from the SNP-wise experiment as SNPs increase from 1 to 100,000 range from 0.02 s to 1,919 s for EDGE, 4.4×10^{-3} seconds to 437 s for PAGER CPU, and 4.2×10^{-3} seconds to 300 s for PAGER GPU (File S2).

PAGER values accurately capture phenotypic patterns of theoretical inheritance models

The box plots of aggregated PAGER values across experimental parameters for training data generated using each of the eight inheritance models are highly comparable to expected phenotype distributions per genotypic class of theoretical models (File S1; Fig. 2). This is observed for both binary (black boxplots) and continuous (white boxplots) phenotypes.

When comparing mean heterozygote encodings of simulated SNPs, PAGER and EDGE values are similar for most inheritance patterns across MAF and noise conditions (Table 1). However, in heterotic models, EDGE heterozygote encodings deviate significantly from expected values. Also, compared to EDGE, PAGER exhibits much smaller variances for heterozygote values overall (File S2). This is primarily due to PAGER's normalization process. Additionally, PAGER is less affected by changes in MAF and noise in additive, superadditive, dominant, heterosis, underdominant, and overdominant inheritance models in SNPs with a binary phenotype (Table 1; File S2). In continuous SNPs, PAGER values are more consistent across MAF and noise levels in subadditive, superadditive, dominant, heterosis, underdominant and overdominant models. EDGE exhibits better consistency in recessive inheritance SNPs in both phenotypes (Table 1; File S2).

PAGER achieves competitive power and a conservative false positive rate

Across all experimental parameters and inheritance models, the power of EDGE and PAGER are highly similar (Fig. 3; File S2). Across all levels of PEN_DIFF and inheritance

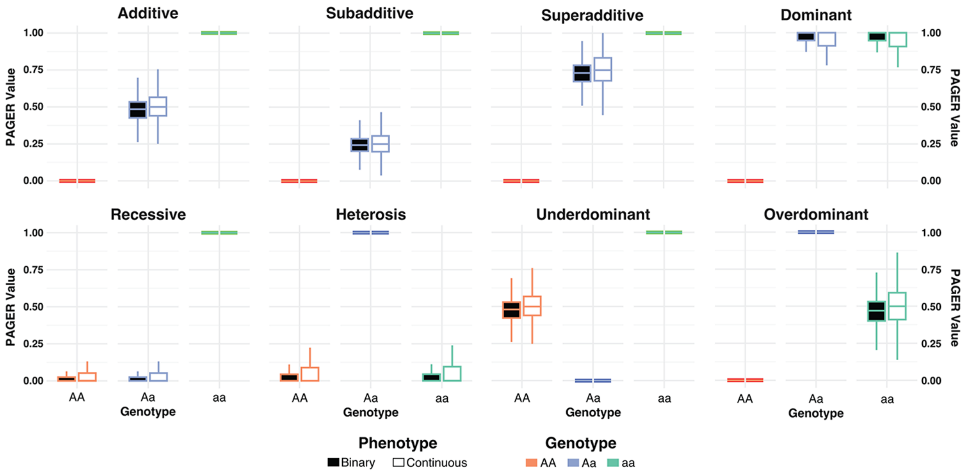

Fig. 2 Boxplots of mean PAGER values across all experimental variables for training data simulated with the eight theoretical inheritance models used in this study. AA, Aa, and aa genotypic classes are depicted in red, blue, and green lines and boxplots, respectively. Boxplots filled in with black represent SNPs with a binary phenotype while white-filled boxplots represent SNPs with a continuous phenotype

Table 1 Mean heterozygote encoding values of EDGE and PAGER

Phenotype	Inheritance Model	Expected Value	All		MAF = 0.1		MAF/ PEN_DIFF = 0.1	
			EDGE	PAGER	EDGE	PAGER	EDGE	PAGER
Binary	Additive	0.50	0.5142	0.4915	0.5666	0.5262	0.7477	0.5548
	Subadditive	0.25	0.2652	0.2519	0.2751	0.2822	0.2958	0.3280
	Superadditive	0.75	0.7453	0.7327	0.7231	0.7494	0.6677	0.7346
	Dominant	1.00	0.9796	0.9563	0.7856	0.9233	0.0650	0.8626
	Recessive	0.00	0.0062	0.0257	0.0230	0.0386	0.0741	0.0973
	Heterosis	1.00	107.65	0.9991	2.4454	0.9959	-2.0471	0.9840
	Underdominant	0.00	-1.0590	0.0019	-1.3057	0.0076	-0.9913	0.0283
	Overdominant	1.00	2.1533	0.9964	2.9815	0.9870	2.2889	0.9560
Continuous	Additive	0.50	0.5035	0.5161	0.5343	0.5448	0.5646	0.5712
	Subadditive	0.25	0.2525	0.2706	0.3325	0.3132	0.5134	0.3813
	Superadditive	0.75	0.7062	0.7488	0.6884	0.7450	0.2705	0.7051
	Dominant	1.00	1.0949	0.9296	0.9978	0.8785	0.9092	0.8873
	Recessive	0.00	-0.0092	0.0543	-0.0078	0.0886	-0.0009	0.2022
	Heterosis	1.00	-23.000	0.9948	-3.1527	0.9813	-0.9586	0.9368
	Underdominant	0.00	-0.9541	0.0087	-0.7623	0.0260	0.7380	0.0869
	Overdominant	1.00	2.1302	0.9868	1.1540	0.9602	0.9102	0.8873

Mean heterozygote encoding values of EDGE and PAGER derived from binary and continuous phenotype SNPs from the training set across all parameters, at MAF set at 0.1, and at MAF and PEN_DIFF both set at 0.1. The 'Expected Value' column contains heterozygote encoding values expected under each theoretical model

model and in both phenotypes, mean power levels of EDGE and PAGER are slightly less than that of inherent with the difference decreasing as sample size and MAF increase (Fig. 3A and B; File S2). Additive tends to achieve the lowest levels of power, except in SNPs simulated with an additive inheritance model (Files S1 and S2). Additive encoding has the greatest difficulty with recessive and heterotic inheritance models. Regardless of phenotype or encoding, we observe that power increases with sample size and MAF. Additionally, higher power is observed in the binary phenotype for all encodings (Fig. 3; File S2). We also observe these trends in instances where PEN_DIFF is set at 0.1 (Fig. 3C and D) and within each inheritance model (Files S1 and S2).

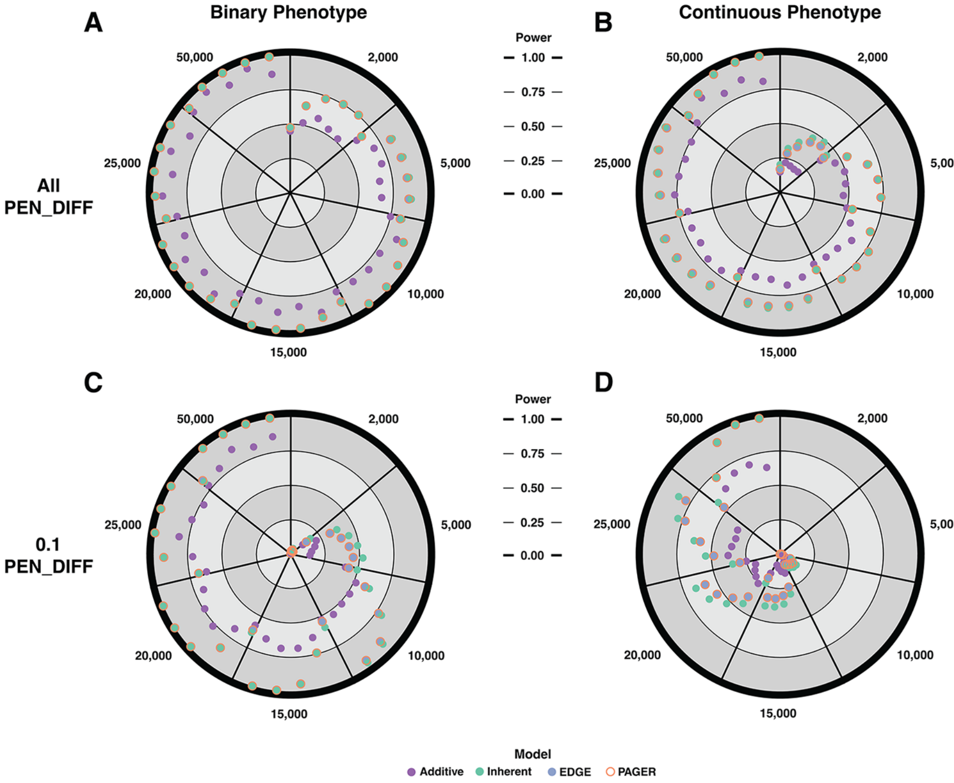

Fig. 3 Polar plots of mean power for each encoding as sample size and minor allele frequency (MAF) increases for SNPs with binary (**A** and **C**) and continuous phenotypes (**B** and **D**). A and B show mean power across all levels of PEN_DIFF while C and D show mean power with PEN_DIFF set at 0.1. Large radial sectors, separated by lines extending from the center, represent increases in sample size in a clockwise direction. Within each radial sector, MAF increases clockwise from 0.1 to 0.5 in 0.1 increments. Power increases from the center of the circle to the edge from 0 to 1 in increments of 0.25, depicted by concentric rings of alternating shades of gray

PAGER achieves a conservative FPR at all levels of sample size and MAF, for both phenotypes, similar to that of additive encoding and EDGE (~5%), within Bradley's liberal criteria [45] of 2.5–7.5% (Fig. 4; File S2). FPR does not seem to scale (negatively or positively) as MAF or samples size increase.

PAGER and EDGE identify putative QTL not detected by additive encoding

There are 13 unique putative QTL denoted by LD regions identified by the three GWAS across all phenotypes (Figs. 5 and 6; Table 2). Separated by phenotype, we observe three, eight, and six QTL for BMI_TAIL, BW, and RetroFat, respectively. Two QTL, QTL 3 and 5, are pleiotropic with two phenotypes (BMI_TAIL/BW and BW/RetroFat, respectively). QTL 1, which is detected at least once in all phenotypes by all encoding methods, is pleiotropic with all three phenotypes (Figs. 5 and 6; Table 2). SNPs in the LD interval of QTL 1 are in close proximity to putative loci with high signal identified by the original GWAS study for multiple phenotypes [30, 31].

QTL 3 is identified as a putative QTL by only recessive, EDGE, and PAGER (Figs. 5 and 6; Table 2). While recessive and PAGER detect this QTL for both BMI_TAIL and BW, EDGE only detects it for BMI_TAIL. Furthermore, no SNPs within QTL 3's LD intervals are implicated in the original GWAS [30, 31]. Apart from EDGE not detecting QTL 3 in BW, EDGE and PAGER detect all the putative QTL that the standard models in the tri-encoding GWAS collectively detect across all three phenotypes (Figs. 5 and 6;

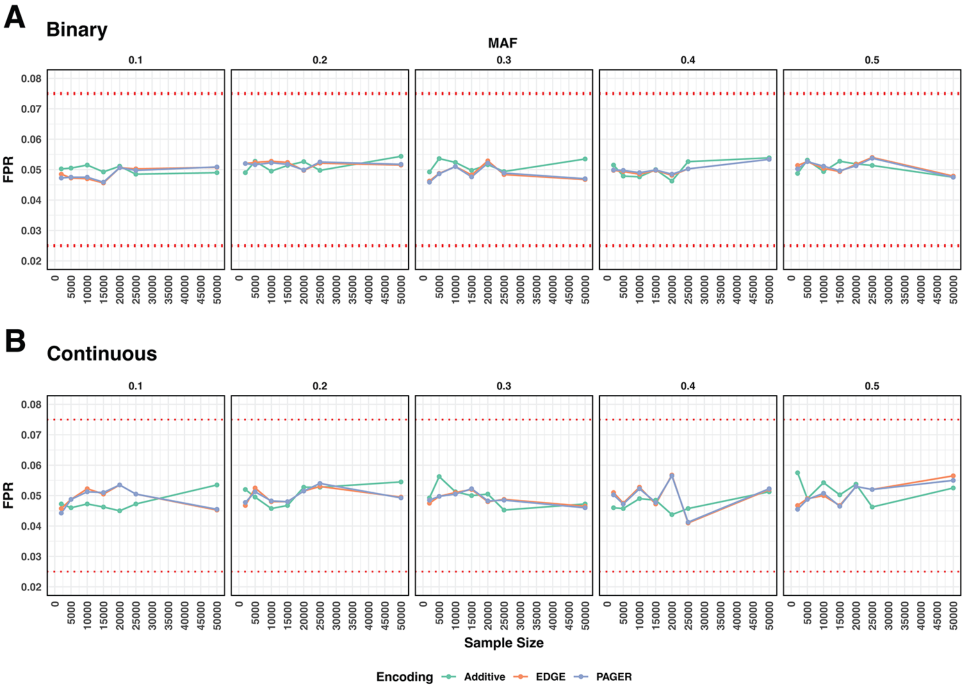

Fig. 4 Mean false positive rate (FPR) for (**A**) binary and (**B**) continuous phenotypes as sample size increases for additive encoding, EDGE, and PAGER. MAF increases left to right from sub-plot to sub-plot. Each point in each subplot is derived from simulated null effect (true-negative) SNPs for each combination of sample size and MAF. Bradley's liberal criteria (2.5–7.5%) is denoted with horizontal dotted red lines

Table 2). It is important to note that additive encoding does detect significant putative QTL that are also detected by recessive and dominant (Fig. 5; Table 2). However, recessive SNPs at QTL 3 do not reach significance under additive encoding (Fig. 5; Table 2). QTL 2, 5 (RetroFat only), 6, 10, and 13 are only detected by the additive encoding in the tri-encoding GWAS (Fig. 5; Table 2). No QTL, as defined by LD intervals, are exclusively detected by dominant encoding (Fig. 5; Table 2). Due to our increased significance threshold compared to the original GWAS study, we do not detect four QTL (one for BMI_TAIL, two for BW, and one for RetroFat) that were detected in that study [30, 31] using any encoding (File S3).

Across phenotypes, PAGER has the most instances (13/17 with 5/13 shared with EDGE) of having the highest QTL signal, in terms of $-\log_{10}p$, compared to other encodings (Table 2; File S3). QTL 3 in BMI_TAIL reaches the highest significance with recessive while QTL 6, 11, and 12 reach the highest significance with additive encoding.

GSEA identifies 150 GO terms and 16 KEGG pathways that are significantly enriched among gene models within the LD interval of QTL 3 (File S3). Most of these enriched GO terms are associated with two genes: *Apc* and *Wnt8a*.

Discussion

PAGER achieves competitive performance and scalable efficiency in simulations

PAGER achieves encoding speeds up to 55 times faster than EDGE when leveraging GPU integration and operating on a sample size of 50,000 (Fig. 1; File S2). Significant speed increases are also observed at lower sample sizes and when utilizing a CPU. These speed increases are significant and highlight that genotype encoding by PAGER will not significantly burden large-scale single locus analyses. Additionally, PAGER accurately

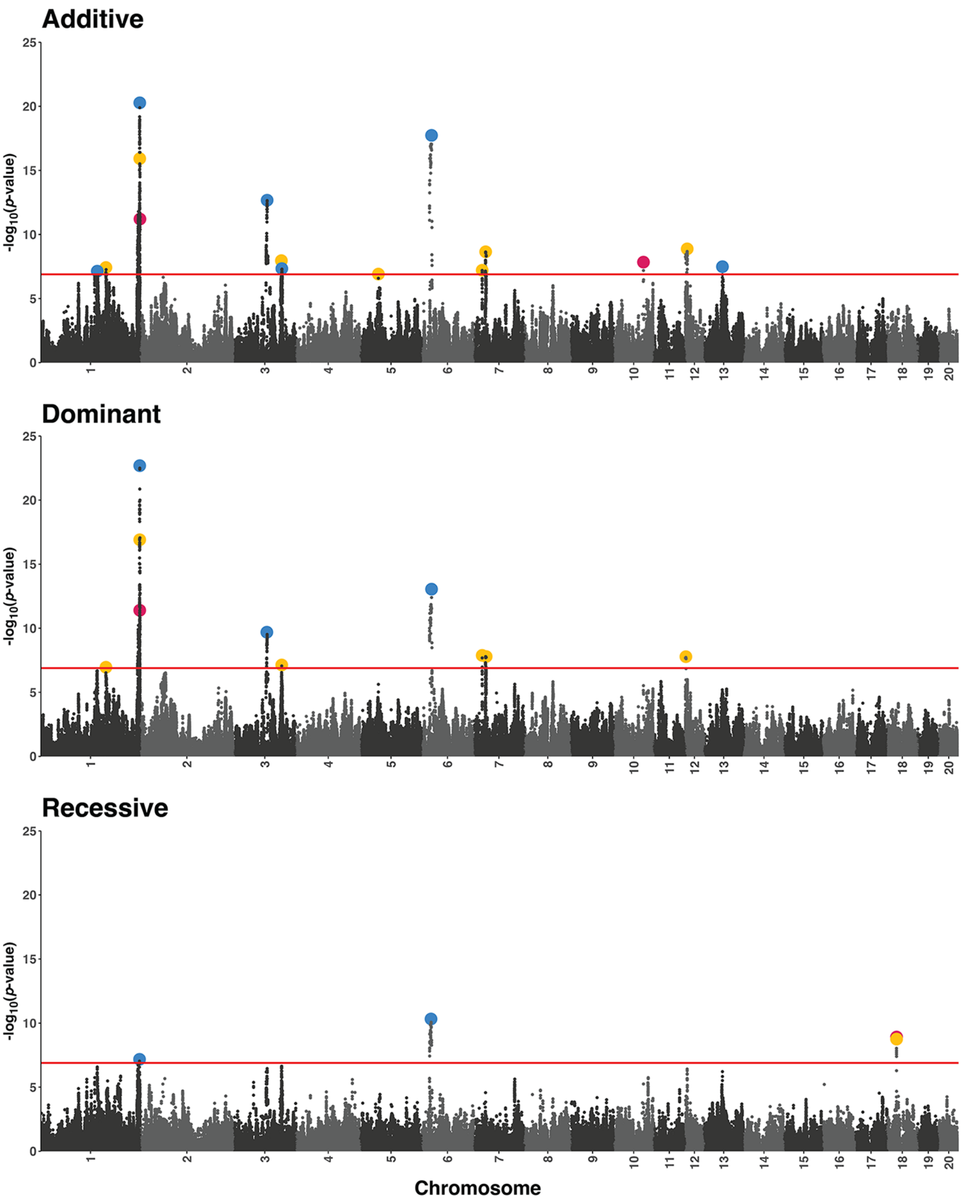

Fig. 5 Porcupine plots (aggregated Manhattan plots) representing –log$_{10}$p-values for each SNP for all three phenotypes for the tri-encoding GWAS. Inheritance model encoding names are above their respective plot. Chromosome number increases along the x-axis. Red horizontal lines denote the experiment-wide significance threshold (-log$_{10}$p = 6.89). Red, yellow, and blue circles represent putative QTL hits for BMI_TAIL, BW, and RetroFat, respectively

describes eight theoretical inheritance models used for SNP simulations (Fig. 2; File S1), underscoring its efficiency and versatility. Comparisons of heterozygote encodings between EDGE and PAGER reveal that both PAGER and EDGE generate heterozygote values close to theoretical levels for both phenotypes, even at a low MAF and high noise, which lead to the most challenging SNPs (File S2). However, EDGE has difficulty describing heterotic inheritance models (heterosis, underdominant, and overdominant). This arises because EDGE uses two anchors (AA and aa) while PAGER only uses one (AA), allowing PAGER to be more flexible and accurately describe any inheritance model. EDGE's lack of flexibility in this regard can make post-analysis interpretations of SNP inheritance models challenging. Despite this, EDGE's power to detect significance

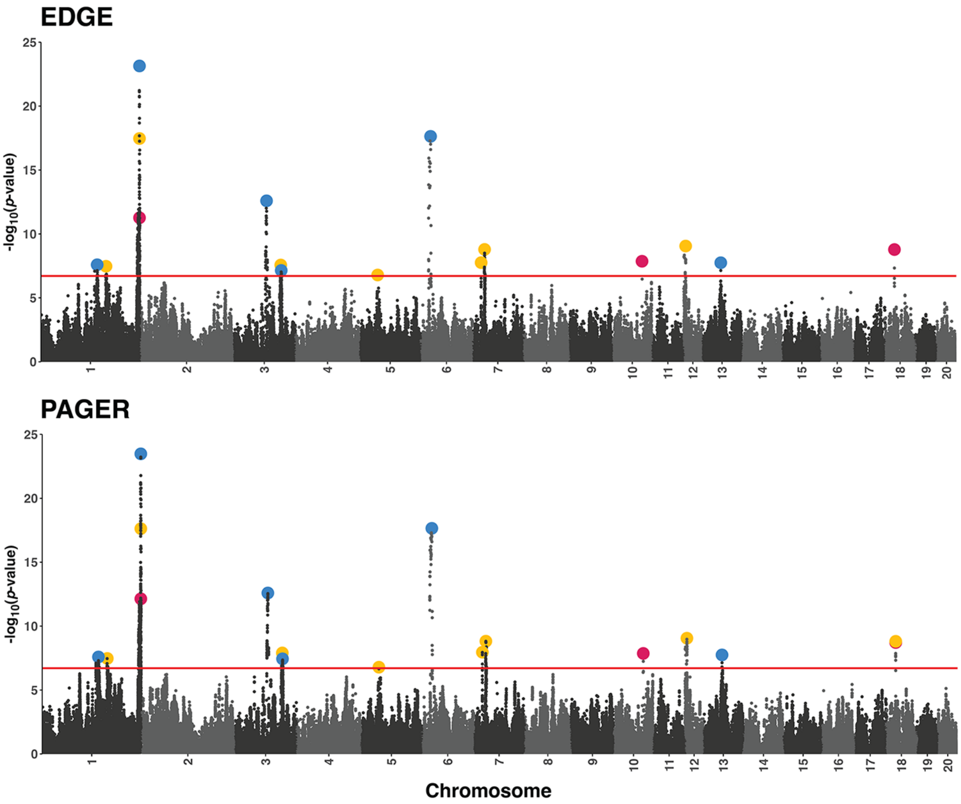

Fig. 6 Porcupine plots (aggregated Manhattan plots) representing –log$_{10}$p-values for each SNP for all three phenotypes for EDGE and PAGER GWAS. The name of each method is above their respective plot. Chromosome number increases along the x-axis. Red horizontal lines denote the experiment-wide significance threshold (-log$_{10}$p = 6.71). Red, yellow, and blue circles represent putative QTL hits for BMI_TAIL, BW, and RetroFat, respectively

in SNPs simulated with these heterotic models are highly comparable to PAGER (Fig. 3; Files S1 and S2). This suggests that, from a modeling perspective, the heterozygote values derived by EDGE for heterotic SNPs are still informative. Consequently, EDGE should still be effective in detecting significant associations in SNPs following heterotic inheritance patterns.

Although PAGER and EDGE do not achieve the highest levels of power, they compensate with their flexibility. Since both methods derive encoding values from a training set, applying these encodings to an external validation set results in some power loss. Nevertheless, EDGE and PAGER significantly outperform the additive encoding in SNPs simulated from non-additive inheritance models, with the greatest performance differences observed in recessive and heterotic models and at the highest noise level (PEN_DIFF = 0.1). This finding suggests that the additive model can underperform in real-world situations, potentially leading to type II errors when SNPs follow alternative inheritance models.

While respective inherent models outperform EDGE and PAGER, especially in scenarios with lower sample sizes and MAFs, employing multiple inherent models in a GWAS or QTL analysis will significantly increase the burden of multiple testing (as observed in our real-world experiment). Therefore, it is more prudent to use EDGE or PAGER to flexibly and dynamically assign each SNP an encoding that closely matches the actual inheritance model at that locus. This approach reduces the burden of multiple testing while maintaining reasonable power. Since we observe that EDGE and PAGER

Table 2 Putative QTL detected in tri-encoding, EDGE, and PAGER GWAS

QTL	Position (chr.bp)	LD Interval Start	LD Interval Stop	Add	Dom	Rec	EDGE	PAGER	MAF
BMI_TAIL									
1	chr1.281474527	chr1.280924333	chr1.282109150		11.40				0.42
	chr1.281489331	chr1.280924333	chr1.282109150					**12.14**	0.42
	chr1.281509176	chr1.280924333	chr1.282109150				11.26		0.43
	chr1.282070632	chr1.280924333	chr1.282561614	11.20					0.46
2	chr10.84021443	chr10.83611968	chr10.84896193	7.84			**7.86**	**7.86**	0.46
3	chr18.27355039	chr18.26640423	chr18.27355039			**8.91**	8.78	8.72	0.31
Bodyweight									
4	chr1.185170500	chr1.184483617	chr1.187801594	7.42	6.95				0.26
	chr1.185730317	chr1.184483617	chr1.187728266				**7.47**	**7.47**	0.26
1	chr1.281420356	chr1.280924333	chr1.282109150		16.91		17.47		0.42
	chr1.281489331	chr1.280924333	chr1.282109150	15.92				**17.63**	0.42
5	chr3.136021511	chr3.132304192	chr3.137146532	7.96			7.57		0.23
	chr3.136707086	chr3.134362543	chr3.137146532					**7.90**	0.25
	chr3.136975356	chr3.136201348	chr3.138849452		7.13				0.49
6	chr5.50933779	chr5.49236189	chr5.50933779	**6.91**			6.79	6.79	0.22
7	chr7.24971798	chr7.24895116	chr7.25205985	7.21	7.88			**7.96**	0.073
	chr7.25030336	chr7.24895116	chr7.25205985				7.76		0.074
8	chr7.34874677	chr7.34156704	chr7.36522260				8.79		0.12
	chr7.34917462	chr7.34132390	chr7.36522260	8.65				**8.82**	0.12
	chr7.36417727	chr7.34156704	chr7.36522260		7.79				0.12
9	chr12.1985457	chr12.850324	chr12.5747642		7.78				0.18
	chr12.5747404	chr12.4956491	chr12.6054210	8.88			**9.06**	**9.06**	0.40
3	chr18.27355039	chr18.26640423	chr18.27355039			8.75		**8.81**	0.31
RetroFat									
10	chr1.160493164	chr1.157254290	chr1.162850128	7.14			**7.59**	**7.59**	0.47
1	chr1.280876316	chr1.279904638	chr1.281355424			7.17			0.45
	chr1.281420356	chr1.280924333	chr1.282109150		22.69		23.14		0.42
	chr1.281509176	chr1.280924333	chr1.282109150	20.27				**23.47**	0.43
11	chr3.94614669	chr3.92341346	chr3.97685154		9.69				0.42
	chr3.95378412	chr3.92341346	chr3.97685154	**12.67**			12.59	12.59	0.42
5	chr3.137286589	chr3.136201348	chr3.138849452	7.34				**7.44**	0.48
	chr3.137335822	chr3.136201348	chr3.138849452				7.14		0.48
12	chr6.27047239	chr6.23069244	chr6.28634206			10.32			0.34
	chr6.28560776	chr6.27047239	chr6.29668903	**17.74**	13.05		17.63	17.65	0.35
13	chr13.53419980	chr13.52504586	chr13.54655489	7.49			**7.75**	**7.75**	0.47

QTL are separated by phenotype and position of peak marker (chromosome and base pair location), LD start and stop intervals, -log10p-values per inheritance encoding, and minor allele frequencies (MAFs) are displayed in columns. Bolded QTL numbers denote pleiotropy (association with more than one phenotype). QTL that have multiple SNP identifier/location rows denote SNPs with overlapping LD windows. Displayed -log10p-values denote that the encoding method identified that SNP as a peak marker for the respective QTL. Bolded -log10p-values represent the highest signal and most significant encoding observed at that QTL

power losses diminish as sample sizes increases, we recommend using training and validation splits in large-scale studies. However, for studies with inherently lower power due to sample size limitations, increasing the significance threshold statistically acknowledges this constraint while still penalizing EDGE and PAGER. While we acknowledge that selecting between penalization strategies is not ideal, our suggestion offers a pragmatic approach for leveraging EDGE and PAGER in practical applications and balancing between power availability and the potential of overfitting.

PAGER's performance on these benchmarking metrics shows that the approach does not incur the significant costs or penalties, such as time loss, often associated with other genotype encodings. PAGER distinguishes itself through its streamlined mathematical approach and flexibility, which facilitate its application to a broad range of systems and phenotypes. Most encoding strategies are tailored to specific species, models, and/or phenotype categories (such as binary or continuous traits). Additionally, it is common for studies to implement a limited selection of inherent inheritance models tailored to particular phenotypes of interest [9, 10, 13]. In contrast, PAGER's architecture supports its use in any genetic framework (including polyploid systems) or phenotype (binary or continuous) and can be implemented for both univariate analyses and investigation of epistasis by simple extension of the algorithm. Moreover, PAGER's ability to incorporate and describe any theoretical inheritance model (Fig. 2; File S1), significantly simplifies the analytical process by removing the need to apply multiple encodings, and thus, the associated multiple testing burden. Finally, PAGER's enhanced processing speed significantly diminishes computational costs associated with genotype encoding. Through these advantages, PAGER emerges as a highly versatile and efficient tool for simplifying and expediting the encoding process for a diverse spectrum of genetic investigations.

PAGER and EDGE reveal a biologically relevant and novel putative QTL

In their 2019 review on the benefits and limitations of GWAS, Tam et al. use an iceberg metaphor to contrast current knowledge with potential future discoveries by GWAS [16]. The tip of the iceberg, visible above water, symbolizes our existing understanding of GWAS, including the reliance on the additive inheritance model. The larger submerged portion represents the untapped future potential of GWAS, including the implementation of alternative inheritance models. This exploratory study builds on that premise and demonstrates that three alternative encoding strategies - recessive, EDGE, and PAGER - identify a novel putative QTL (QTL 3) that additive did not in two phenotypes (EDGE in only one phenotype – BMI_TAIL).

LD intervals around putative SNPs for QTL 3 do not overlap with any QTL found in the original GWAS study [30, 31]. Thus, QTL 3 is novel for this population of rats and these phenotypes. Gene models in the LD interval of QTL 3 (chr18.26640423 - chr18.27355039) show enrichment in 150 specific GO terms (five MF, 143 BP, and two CC) and 16 KEGG pathways (File S3). KEGG pathways are related to Wnt signaling and certain diseases including cancers (including colorectal cancer and gastric cancer), Alzheimer's disease, and Cushing syndrome (File S3). GO terms are associated with primarily two genes: *Wnt8a* and *Apc*.

Wnt8a participates in the Wnt signaling and thus is likely involved in roles including cell fate determination, cell migration, cell polarity, neuron differentiation, and organogenesis [46]. Indeed, many of the enriched biological functions for QTL 3 are associated with these roles (File S3). *Apc*, just upstream of *Wnt8a*, is an APC (adenomatous polyposis coli) regulator of the Wnt signaling pathway. *Apc*, in addition to being involved in growth, development, and cell differentiation, is also a tumor suppressor gene linked to certain cancers including colorectal and brain cancers [47, 48]. In humans and mice, obesity and obesity-driven inflammation, precursors to colorectal cancers, are linked to overactivation of the Wnt signaling pathway [49, 50]. In turn, *Apc* negatively regulates increased Wnt signaling [49–51]. Our results could indicate that mutations in *Wnt8a* and *Apc* are associated with some of the variation in BMI and bodyweight we observe in this population of rats. Additional studies are required to validate this claim.

Nevertheless, EDGE and PAGER effectively highlight a novel genetic locus, and areas for scientific investigation, that were not identified in the initial GWAS.

Encoding methods Differ in Peak QTL marker positions and significance

Variation in the base pair positions of peak markers of the same QTL are observed depending on the encoding method employed (Table 2; File S3). Because of this, some encoder-specific LD intervals do not have the same start and end points yet do overlap at least at one end. This likely occurs as each encoding method implicates (or prefers) peak markers that conform better to the inheritance pattern(s) modeled. These preferences result in fluctuations of the signal of each encoding's peak marker and explain the varying significance levels observed across encoding methods, including dominant and recessive (Table 2; File S3).

PAGER can model each SNP to describe the inheritance pattern observed from average phenotypes more accurately than EDGE (primarily in heterotic models) and other inheritance encodings (Table 1: File S2), making it more sensitive to variation. Thus, PAGER can accurately subsume all theoretical inheritance models. In theory, it follows that PAGER should achieve the highest significance level for every QTL. However, it is likely that additive encoding inflates the signal of some SNPs that do not conform to strict additivity [16, 52], even those exhibiting moderate to high deviations from additivity. This is likely true for dominant and recessive encodings as well. Indeed, according to PAGER heterozygote values, no SNP at a peak marker is observed to follow a purely additive model of inheritance in which the heterozygote is completely intermediate (i.e., 0.5; File S3). Despite this, additive achieves higher significance than any other model in three QTL.

An explanation for why significance values differ and reach higher levels in different models may involve how EDGE and PAGER dynamically derive encoding values. Modeling SNPs closer to their true inheritance pattern likely results in QTL signals closer to their 'true' significance level and chromosomal position as EDGE and PAGER account for and quantify deviations from additivity on an SNP-by-SNP basis. Although it could be argued that this is a type of overfitting, higher significance values observed in other models point to an alternative explanation. It is also possible that additive, recessive, and dominant models, which are applied uniformly to all SNPs, may introduce some error, either increasing or decreasing main effect signals and potentially resulting in type I and type II errors in marginally significant SNPs [16, 52]. Another way to frame this is that uniform inheritance models enforce their own inherent biases across the entire dataset, introducing error.

An alternative explanation for the varying levels of significance across models is that PAGER and/or EDGE interact with LD structures differently compared to uniform models. This interaction may increase or decrease the significance of nearby loci by effectively shortening or lengthening the virtual LD windows. While our steps for controlling proximal contamination and ensuring QTL independence address some of these issues, there could be additional aspects to explore. Yet another alternative, exclusively regarding PAGER, is the potential violation of the assumption of normality in phenotypic distributions for each genotypic class. However, this is unlikely as MAFs are greater than 0.20 in every QTL where an encoding other than PAGER achieves a higher signal (Table 2; File S3). Although this does not guarantee normality, it implies that adequate sample sizes exist for each genotypic class in these QTL, hence not limiting PAGER's efficacy. Despite these differences observed in significance levels, PAGER and additive encodings share the same peak marker in 80% (12/15) of instances where both encodings identify the same QTL (Table 2; File S3). Alternatively, additive and EDGE and PAGER and EDGE only share peak markers in 53.3% (8/15) and 56.25% (9/16) of instances, respectively (Table 2; File S3). Thus, in most circumstances, additive and PAGER encodings highlight the same peak marker underlying these complex phenotypes and, in that way, are more comparable. This implies that PAGER may generalize better to complex traits across phenotypes and systems compared to EDGE. However, additional experimentation and biological validation is required to bolster this claim.

It is unclear why additive detects some SNPs with notable deviations from additivity while not detecting QTL 3 on chromosome 18 (Table 2; File S3). For example, additive encoding detects QTL 8 and 13, both of which have large deviations from additivity, according to PAGER values and respective detections from the dominant encoding (QTL8; File S3). It may be that these SNPs still contribute significantly to the additive genetic variation of the traits (BW and RetroFat), which allows the additive model to detect them [14]. In addition to EDGE and PAGER, QTL 3 is also identified by the recessive encoding GWAS (Fig. 5; File S3), reinforcing its potential as a genuine true positive as the PAGER values for this locus point to a recessive model of inheritance. However, this QTL is not detected by the additive encoding. Interestingly, our simulation experiments reveal that the additive model's power to detect recessive-simulated SNPs is significantly lower compared to its performance with dominant-simulated SNPs (Files S1 and S2). The representations of the recessive and dominant inheritance models in File S1 provide insight on this observation. The slopes and linear relationships of the additive and dominant models are better aligned compared to those of the additive and recessive models. This suggests that the dominant model (and the dominant encoding) more closely resembles the additive model than the recessive model does. Indeed, both QTL 8 and 13 follow highly dominant inheritance patterns, according to PAGER encoding values. This discrepancy between the additive and recessive models makes detecting recessive SNPs more challenging for the additive encoding and explains why additive fails to identify QTL 3 in the GWAS. Additional validation experiments are required to elucidate QTL 3's role, if any, in obesity and metabolism in this system. However, it is important to note that we provide evidence that sole use of the additive model in GWAS significantly reduces the power to detect putative QTL following a recessive model of inheritance.

Notably, while the dominant model did not identify any unique QTLs beyond those detected by the additive model, it yielded more significant p-values for QTL 1 across all phenotypes (Table 2; File S3). This is likely due to the peak markers at this locus exhibiting strong deviations from additivity, according to PAGER values, that align more with subadditive models of inheritance (Aa range = 0.191–0.218; File S3). While QTL 1 exerts substantial main effects on the phenotypes tested in this rat population, the increased significance with the dominant model suggests it could detect this locus where the additive model might fail if the effects were more marginal. However, it must be noted that the dominant model highlights different peak markers than the additive model, which show greater deviations from additivity (File S3). This supports the notion that different encoding strategies may favor variants aligning more closely with their model assumptions. Interestingly, both dominant and additive encodings identify the same peak marker for QTL 7, yet the dominant model yields a more significant *p*-value despite this SNP exhibiting an inheritance pattern that is nearly additive (Aa = 0.445), according to PAGER values (File S3). The reason for this is unclear but highlights that uniform model assumptions may lead to the introduction of noise and error in single-locus analyses.

The inability of the additive encoding to detect QTL 3 in BMI_TAIL and BW raises concerns about its effectiveness in identifying QTLs following heterotic models of inheritance. Indeed, additive power values in SNPs simulated from heterotic models are low (Files S1 and S2). In this study, we employ three uniform encoding strategies in the tri-encoding GWAS, additive, dominant, and recessive, which successfully identify concordant QTLs with EDGE and PAGER. However, in other systems and phenotypes, SNPs exhibiting strong effects can follow heterotic patterns (e.g., heterosis, underdominance, and overdominance). This is especially true in economically significant domesticated plants [5, 7] and animals [53, 54]. It remains uncertain whether the additive encoding, or dominant and recessive, can adequately capture all, most, or any of these variants. Had such variants been present in our study, additional uniform encoding models might have been necessary for their detection, further increasing the multiple testing burden of the tri-encoding GWAS. This underscores the potential of using dynamic tools like PAGER and EDGE, which are designed to detect any significant SNP association, regardless of the inheritance model. Further experimentation using real-world data is needed to evaluate the effectiveness of the additive encoding in capturing significant heterotic signals and to further explore PAGER's capabilities in this context.

Limitations of PAGER

Every model has assumptions, and therefore limitations. Indeed, the assumption underlying the additive inheritance model (and all uniform models) could be considered the most unrealistic when compared to EDGE and PAGER. It assumes that for every SNP, across all systems and phenotypes, heterozygotes exhibit traits that are precisely intermediate between those of homozygotes. PAGER, on the other hand, uses the relative differences between mean phenotype values of each genotypic class. This dynamic nature makes PAGER a powerful tool, but it can lose power to accurately describe inheritance models when phenotype distributions deviate from normality. This is an issue only in continuous phenotypes, as the mean phenotype per genotypic class is the proportion of cases in case/control studies. However, skewed phenotypic distributions can affect any encoding strategy [27, 55]. We expect that with large sample sizes and typical MAFs, significant deviations from normality will not be common. However, if this is not the case, we suggest transforming phenotypes or editing the PAGER formulae by replacing the mean phenotype value per genotypic class with the median and comparing performance between the two approaches. In some highly skewed distributions, the median may be more descriptive than the mean [56] and better capture the central tendency of the data. It is important to prune data of SNPs that have low MAFs to further reduce instances of skewed phenotype distributions. Future work will focus on the impact of skewness on PAGER accuracy. Small sample sizes and low allele frequencies can also lower PAGER's ability to detect true significance. Yet, as stated above concerning skewed genotype distributions, these issues negatively impact all encoding strategies [27, 57]. Indeed, we observe how low sample sizes and MAFs negatively affect power in our simulated data experiments for all methods including additive and inherent (Fig. 3: Files S1 and S2).

Although PAGER does not generate statistical models and perform hypothesis tests like EDGE, it does use the phenotype to derive SNP encodings. Often termed 'double-dipping,' this can lead to significant overfitting [58]. Even though PAGER does achieve higher significance in most QTL compared to other encodings, these signals are not largely inflated nor is the trend universal (i.e., other encodings are observed to achieve higher signal for some QTL). Despite these observations, PAGER, along with EDGE, should be penalized due to the costs associated with more accurately modeling each SNP's inheritance pattern using the phenotype. For our application of PAGER to real-world data, we adjusted the significance threshold by halving the Bonferroni cutoff to make it comparable to the tri-encoding GWAS, which utilizes the same population of rats and divides the Bonferroni cutoff by three. As we have touched upon previously and demonstrated in our simulation experiments, in non-exploratory studies with large sample sizes, using training and validation splits can provide a viable and potentially more robust alternative. Regardless of the penalty selected, their implementation can prevent the detection of some variants with substantial main effects. Despite this, we demonstrate that PAGER not only captures the same genetic associations as multiple uniform inheritance models, including a novel putative QTL, but also achieves greater efficiency, as fewer tests and corresponding corrections are required compared to when multiple models are applied. While no approach to genotype encoding is flawless, PAGER is expected to perform efficiently in the vast majority of cases, especially within robust experimental designs featuring large sample sizes. When selecting phenotypes for PAGER encoding, researchers should choose traits that are directly relevant to their research questions and for which they have high-quality phenotype data. Ensuring adequate sample size and appropriate phenotype distribution (e.g., avoiding highly skewed data) will enhance the accuracy of the genotype encodings and the reliability of results.

Conclusions

We have effectively demonstrated that PAGER is a fast, robust, and efficient genotype encoding strategy that, along with EDGE, detects biologically relevant genotype-phenotype associations that the standard additive encoding alone does not. We also provide evidence that it is more prudent to use dynamic SNP-wise encoding strategies, like EDGE and PAGER, than employing multiple uniform inheritance encodings. PAGER is designed to replicate findings from additive, dominant, and recessive models while also identifying novel associations that these standard models may miss. The variation in findings across different models, as observed in Table 2, underscores the importance of considering multiple inheritance patterns in association analyses. By capturing a wide range of inheritance patterns within a single analysis, PAGER reduces the need to apply multiple models separately. This flexibility allows researchers to detect SNPs with diverse effect types without increasing the multiple testing burden, thereby enhancing the discovery of meaningful genetic associations. Taken together, these results suggest PAGER's potential in uncovering previously unrecognized variants, and thus missing heritability, influencing complex traits. Our approach facilitates a more nuanced understanding of quantitative genetics, surpassing traditional methods that follow uniform models of inheritance. Importantly, PAGER has the capability of opening new avenues for research into the genetic basis of traits and diseases, offering promising directions for future genetic studies and the discovery of novel putative loci and therapeutic targets.

PAGER is easy to implement in any programming language, faster than competitive methods, is extensible to any genetic system and phenotype, and can be easily adapted to investigate epistasis. In future work, we aim to apply PAGER to other systems and phenotypes and explore its capability to detect pairwise and higher order epistatic interactions. Additionally, we aim to incorporate PAGER as a feature encoding and feature engineering operator in automated machine learning (autoML) workflows. PAGER's speed and extensibility make it an ideal addition to any autoML method, and we expect that it can even be applied to non-biological categorical features.

We hope this work motivates others to implement PAGER in their respective fields and that its use leads to new discoveries. Python and R code to implement PAGER is publicly available at: https://github.com/EpistasisLab/PAGER.

Abbreviations

APC	Adenomatous Polyposis Coli
BAMS	Biallelic Model Simulator
BMI	Body Mass Index
BMI_Tail	Body Mass Index including the animal's tail
bp	base pair(s)
BW	Body Weight
CLARITE	CLeaning to Analysis: Reproducibility-based Interface for Traits and Exposures
CPU	Central Processing Unit(s)
EDGE	Elastic Data-driven Genetic Encoding
EVD	Extreme Value Distribution(s)
FDR	False Discovery Rate(s)
FPR	False Positive Rate(s)
GEMMA	Genome-wide Efficient Mixed Model Association
GO	Gene Ontology(ies)
GPU	Graphics Processing Unit(s)
GRM	Genetic Relatedness Matrix
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-Wide Association Study(ies)
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage Disequilibrium
LOCO	Leave One Chromosome Out
MAF	Minor Allele Frequency(ies)
MLG	Multi-Locus Genotype(s)
PAGER	Phenotype-Adjusted Genotype Encoding and Ranking
PEN_DIFF	Penetrance Difference(s)
QTL	Quantitative Trait Locus(i)
RAM	Random-Access Memory
RetroFat	Retroperitoneal Fat
SNP	Single Nucleotide Polymorphism
V_P	Phenotypic Variance
VRAM	Video Random-Access Memory

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-024-00393-x.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7

Supplementary Material 8

Supplementary Material 9

Supplementary Material 10

Supplementary Material 11

Supplementary Material 12

Supplementary Material 13

Supplementary Material 14

Supplementary Material 15

Supplementary Material 16

Supplementary Material 17

Supplementary Material 18

Supplementary Material 19

Acknowledgements
Not applicable.

Author contributions
PJF, AG, and PB contributed equally to the manuscript. PJF wrote the manuscript, developed the PAGER algorithm, designed experiments, created figures and tables, curated results, and performed numerous analyses including statistical tests and LD calculations. AG wrote and implemented EDGE and PAGER scripts and designed, coded, and performed all benchmarking experiments. PB designed and performed initial GWAS, genetic relatedness calculations, conditional analyses, and GSEA. NM aided in coding various benchmarking scripts and implementing EDGE to real-world data. ASC, JZ, MH provided analytical and manuscript feedback and shared code to perform various analyses including GWAS, conditional analysis, data simulation, EDGE implementation, and benchmarking. AAP, TOA, and JHM served as mentors, providing guidance and assistance, and editing the final manuscript. JHM also provided funding for the research. All authors read and approved the final manuscript.

Funding
The authors gratefully acknowledge the support from the NIH under grants R01 LM010098 and R01 AG066833 awarded to JHM and grant P50 DA037844 awarded to AAP.

Data availability
Rat genotypes, phenotypes, and GWAS summary statistics are available at https://library.ucsd.edu/dc/object/bb83725195. Python and R implementations of PAGER, and all custom shell, Python, and R scripts used for all analyses can be found at https://github.com/EpistasisLab/PAGER.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 7 June 2024 / Accepted: 30 September 2024

Published online: 11 October 2024

References

1. Hill WG, Goddard ME, Visscher PM. Data and Theory Point to mainly additive genetic variance for Complex traits. PLoS Genet. 2008;4:e1000008.
2. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. 2009;10:565–77.
3. Matsui T, Mullis MN, Roy KR, Hale JJ, Schell R, Levy SF, et al. The interplay of additivity, dominance, and epistasis on fitness in a diploid yeast cross. Nat Commun. 2022;13:1463.
4. Hallin J, Märtens K, Young AI, Zackrisson M, Salinas F, Parts L, et al. Powerful decomposition of complex traits in a diploid model. Nat Commun. 2016;7:13311.
5. Yang J, Mezmouk S, Baumgarten A, Buckler ES, Guill KE, McMullen MD, et al. Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. PLoS Genet. 2017;13:e1007019.
6. Wu X, Li R, Li Q, Bao H, Wu C. Comparative transcriptome analysis among parental inbred and crosses reveals the role of dominance gene expression in heterosis in Drosophila melanogaster. Sci Rep. 2016;6:21124.
7. Hua J, Xing Y, Wu W, Xu C, Sun X, Yu S, et al. Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. Proc Natl Acad Sci. 2003;100:2574–9.

8. Wermter A-K, Scherag A, Meyre D, Reichwald K, Durand E, Nguyen TT, et al. Preferential reciprocal transfer of paternal/maternal DLK1 alleles to obese children: first evidence of polar overdominance in humans. Eur J Hum Genet. 2008;16:1126–34.
9. Bonnafous F, Fievet G, Blanchet N, Boniface M-C, Carrère S, Gouzy J, et al. Comparison of GWAS models to identify non-additive genetic control of flowering time in sunflower hybrids. Theor Appl Genet. 2018;131:319–32.
10. Joo J, Kwak M, Ahn K, Zheng G. A robust genome-wide scan Statistic of the Wellcome Trust Case–Control Consortium. Biometrics. 2009;65:1115–22.
11. Hoggart CJ, Venturini G, Mangino M, Gomez F, Ascari G, Zhao JH, et al. Novel Approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body Mass Index. PLoS Genet. 2014;10:e1004508.
12. Tukiainen T, Pirinen M, Sarin A-P, Ladenvall C, Kettunen J, Lehtimäki T, et al. Chromosome X-Wide Association Study identifies loci for fasting insulin and height and evidence for incomplete dosage compensation. PLoS Genet. 2014;10:e1004127.
13. Liu H-M, Zheng J-P, Yang D, Liu Z-F, Li Z, Hu Z-Z, et al. Recessive/dominant model: alternative choice in case-control-based genome-wide association studies. PLoS ONE. 2021;16:e0254947.
14. Falconer DS. Introduction to quantitative Genetics. India: Pearson Education; 1996.
15. Lynch M, Walsh B. Genetics and Analysis of quantitative traits. Sunderland, MA: Sinauer; 1998.
16. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20:467–84.
17. Liu T, Luo C, Ma J, Wang Y, Shu D, Qu H, et al. Including dominance effects in the prediction model through locus-specific weights on heterozygous genotypes can greatly improve genomic predictive abilities. Heredity. 2022;128:154–8.
18. Costa-Neto G, Fritsche-Neto R, Crossa J. Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity. 2021;126:92–106.
19. Gaye A, Davis SK. Genetic model misspecification in genetic association studies. BMC Res Notes. 2017;10:569.
20. Mackey DA. What colour are your eyes? Teaching the genetics of eye colour & colour vision. Edridge Green lecture RCOphth Annual Congress Glasgow May 2019. Eye. 2022;36:704–15.
21. Hedrick PW. Population genetics of malaria resistance in humans. Heredity. 2011;107:283–304.
22. Palmer DS, Zhou W, Abbott L, Wigdor EM, Baya N, Churchhouse C, et al. Analysis of genetic dominance in the UK Biobank. Science. 2023;379:1341–8.
23. Hall MA, Wallace J, Lucas AM, Bradford Y, Verma SS, Müller-Myhsok B, et al. Novel EDGE encoding method enhances ability to identify genetic interactions. PLoS Genet. 2021;17:e1009534.
24. Zhou J, Guare L, Rico ALG, Zarzar TG, Palmiero N, Assimes TL et al. Flexibly encoded GWAS identifies novel non-additive SNPs in individuals of African and European ancestry [Internet]. medRxiv; 2023 [cited 2023 Oct 31]. p. 2023.06.01.23290857. https://www.medrxiv.org/content/https://doi.org/10.1101/2023.06.01.23290857v1
25. Van Rossum G, Drake FL. Python 3 reference Manual. Scotts Valley, CA: CreateSpace; 2009.
26. Cohen J. Statistical Power Analysis for the behavioral sciences. New York, NY, USA: Routledge Academic; 1988.
27. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nat Rev Methods Primers. 2021;1:1–21.
28. Hall MA. GitHub - HallLab/pandas-genomics: Pandas ExtensionDtypes for dealing with genomics data [Internet]. [cited 2024 Jan 4]. https://github.com/HallLab/pandas-genomics
29. Hansen C, Spuhler K. Development of the National Institutes of Health genetically heterogeneous Rat Stock. Alcoholism: Clin Experimental Res. 1984;8:477–9.
30. Chitre AS, Polesskaya O, Holl K, Gao J, Cheng R, Bimschleger H, et al. Genome-wide Association study in 3,173 outbred rats identifies multiple loci for Body Weight, Adiposity, and fasting glucose. Obesity. 2020;28:1964–73.
31. Chitre AS, Polesskaya O, Holl K, Gao J, Cheng R, Bimschleger H et al. Genome-Wide Association Study in 3,173 Outbred Rats for Body Weight, Adiposity, and Fasting Glucose [Internet]. Genes and Addiction: NIDA Center for GWAS in Outbred Rats. 2022 [cited 2022 Jul 18]. https://cgord.org/dataset/2
32. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. 2016 edition. New York, NY: Springer; 2016.
33. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing. 2022. https://www.R-project.org/
34. Passero K, He X, Zhou J, Mueller-Myhsok B, Kleber ME, Maerz W et al. Phenome-wide association studies on cardiovascular health and fatty acids considering phenotype quality control practices for epidemiological data. Biocomputing 2020 [Internet]. WORLD SCIENTIFIC; 2019 [cited 2024 Feb 26]. pp. 659–70. https://www.worldscientific.com/doi/abs/10.1142/9789811215636_0058
35. Lucas AM, Palmiero NE, McGuigan J, Passero K, Zhou J, Orie D et al. CLARITE Facilitates the Quality Control and Analysis Process for EWAS of Metabolic-Related Traits. Frontiers in Genetics [Internet]. 2019 [cited 2024 Feb 26];10. https://www.frontiersin.org/journals/genetics/articles/https://doi.org/10.3389/fgene.2019.01240
36. Fadista J, Manning AK, Florez JC, Groop L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet. 2016;24:1202–5.
37. Powell JE, Henders AK, McRae AF, Kim J, Hemani G, Martin NG, et al. Congruence of Additive and Non-additive effects on Gene expression estimated from pedigree and SNP Data. PLoS Genet. 2013;9:e1003502.
38. Azevedo CF, de Resende MDV, e Silva FF, Viana JMS, Valente MSF, Resende MFR, et al. Ridge, Lasso and bayesian additive-dominance genomic models. BMC Genet. 2015;16:105.
39. Zhou X, Stephens M. Genome-wide Efficient Mixed Model Analysis for Association Studies. Nat Genet. 2012;44:821–4.
40. Cheng R, Parker CC, Abney M, Palmer AA. Practical considerations regarding the use of genotype and Pedigree Data to Model Relatedness in the context of Genome-Wide Association Studies. G3 Genes|Genomes|Genetics. 2013;3:1861–7.
41. Gonzales NM, Seo J, Hernandez Cordero AI, St. Pierre CL, Gregory JS, Distler MG, et al. Genome wide association analysis in a mouse advanced intercross line. Nat Commun. 2018;9:5162.
42. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a Tool Set for whole-genome Association and Population-based linkage analyses. Am J Hum Genet. 2007;3:559–75.
43. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115–21.
44. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innov. 2021;2:100141.

45. Bradley JV, Robustness?. Br J Math Stat Psychol. 1978;31:144–52.
46. Komiya Y, Habas R. Wnt signal transduction pathways. Organogenesis. 2008;4:68–75.
47. Bhat RV, Baraban JM, Johnson RC, Eipper BA, Mains RE. High levels of expression of the tumor suppressor gene APC during development of the rat central nervous system. J Neurosci. 1994;14:3059–71.
48. De Filippo C, Caderni G, Bazzicalupo M, Briani C, Giannini A, Fazi M, et al. Mutations of the apc gene in experimental colorectal carcinogenesis induced by azoxymethane in F344 rats. Br J Cancer. 1998;77:2148–51.
49. Liu Z, Brooks RS, Ciappio ED, Kim SJ, Crott JW, Bennett G, et al. Diet-induced obesity elevates colonic TNF-α in mice and is accompanied by an activation of *wnt* signaling: a mechanism for obesity-associated colorectal cancer. J Nutr Biochem. 2012;23:1207–13.
50. Taketo MM. Shutting down wnt signal–activated cancer. Nat Genet. 2004;36:320–2.
51. Liu W, Crott JW, Lyu L, Pfalzer AC, Li J, Choi S-W, et al. Diet- and genetically-induced obesity produces alterations in the Microbiome, inflammation and wnt pathway in the intestine of Apc+/1638 N mice: comparisons and contrasts. J Cancer. 2016;7:1780–90.
52. Bush WS, Moore JH. Chapter 11: genome-wide Association studies. PLoS Comput Biol. 2012;8:e1002822.
53. Wu X-L, Zhao S. Editorial: Advances in Genomics of Crossbred Farm Animals. Front Genet [Internet]. 2021 [cited 2024 May 22];12. https://www.frontiersin.org/journals/genetics/articles/https://doi.org/10.3389/fgene.2021.709483/full
54. Xiao Q, Huang Z, Shen Y, Gan Y, Wang Y, Gong S, et al. Transcriptome analysis reveals the molecular mechanisms of heterosis on thermal resistance in hybrid abalone. BMC Genomics. 2021;22:650.
55. John M, Ankenbrand MJ, Artmann C, Freudenthal JA, Korte A, Grimm DG. Efficient permutation-based genome-wide association studies for normal and skewed phenotypic distributions. Bioinformatics. 2022;38:ii5–12.
56. McClave J, Sincich T. Statistics. 13th edition. Boston: Pearson; 2016.
57. Klein RJ. Power analysis for genome-wide association studies. BMC Genet. 2007;8:58.
58. Ball TM, Squeglia LM, Tapert SF, Paulus MP. Double dipping in machine learning: problems and solutions. Biol Psychiatry Cogn Neurosci Neuroimaging. 2020;5:261–3.

Publisher's note