The Journal of Open Source Software

HARE: A Python workflow for analyzing genomic feature enrichment in GWAS datasets

Olivia S. Smith [1], **Eucharist Kun** [1], **and Vagheesh M. Narasimhan** [1,2,3¶]

1 Department of Integrative Biology, The University of Texas at Austin, USA **2** Department of Statistics and Data Science, The University of Texas at Austin, USA **3** Department of Population Health, Dell Medical School ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 10 January 2024
Published: unpublished

Summary

Genomic datasets have grown by orders of magnitude in the last decade and thus require increasingly flexible and streamlined analysis workflows. One major line of research in genomics is gene enrichment analysis, which aims to quantitatively assess whether the genomic basis of a phenotype of interest has higher association with other genomic regions than would be expected by chance. HARE is a Python pipeline which annotates genome-wide significant positions, identifies overlap between these positions and a provided list of genomic features (here called 'elements') of interest, and determines enrichment likelihood through a resampling process of random length-matched regions. HARE is written in Python and is available for installation from source or via the Python Package Index (pip).

Statement of need

HARE is a computational pipeline run via command line for analyzing enrichment of element sets, such as the set of human-accelerated regions (HARs), in genomic regions associated with a phenotype of interest. Several genetic enrichment approaches are already available in the literature, such as gene set enrichment analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005), stratified linkage disequilibrium score regression (LDSC) (Finucane et al., 2015), and FUMA GWAS (Watanabe et al., 2017). However, these approaches are not translatable to studies examining overlaps in custom regions of the genome or when these regions are short, which is known to lead to bias in some approaches (Finucane et al., 2015). Our pipeline is flexible on either end of the enrichment analysis, both in terms of the regions it utilizes as well as the type of input genomic scan provided, allowing for a completely flexible model for genomic enrichment analysis.

Many tools also require time-consuming data reformatting which HARE minimizes by accepting a variety of summary statistic file formats. In contrast, HARE makes it possible to install the environment with conda and pip, allowing for an immediate run of the complete analysis pipeline. HARE has already been used to analyze over 200 traits in published work as well as ongoing studies presented at conferences (Kun, Javan, et al., 2023; Kun, Sohail, et al., 2023; Xu et al., 2023), exemplifying its usefulness in current and future research.

Dependencies

HARE requires Ensembl's Variant Effect Predictor (McLaren et al., 2016) with the appropriate cache (VEP Cache Documentation) as well as BEDTools (Quinlan & Hall, 2010). It also requires the use of `wget` for querying Ensembl BioMart (Martin et al., 2023). It is recommended that users install dependencies through a combination of conda using the `environment.yml` file

included in the repository and pip. Complete instructions for dependency and tool installation can be found in the repository README.

Workflow

A visual depiction of the HARE workflow is shown in Figure 1.

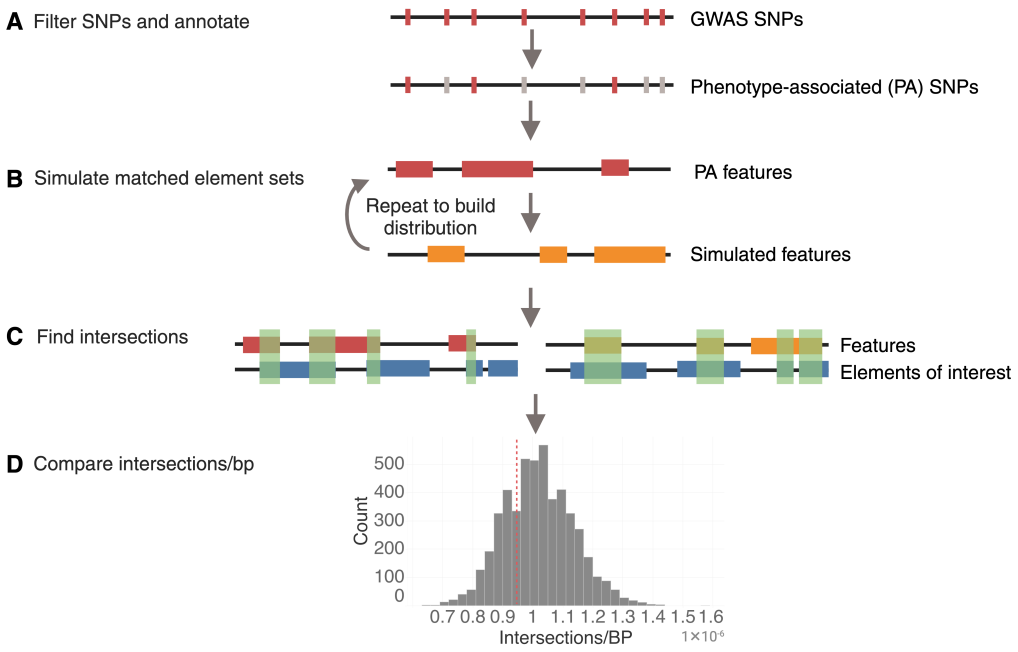

Figure 1: HARE visual workflow. **(A)** Summary statistics from a genome-wide association study (GWAS) are filtered to genome-wide significant single nucleotide polymorphisms (SNPs) based on p-value threshold. These are the phenotype-associated (PA) SNPs. **(B)** For each PA SNP, the nearest genetic feature within a given bp distance and its location are annotated. This is called the element set. **(C)** Simulated element sets built from random regions matched on length to the PA features are generated. **(D)** Intersections per bp between each element set, both PA and simulated, and the elements of interest (e.g. human accelerated regions) and an empirical p-value are computed.

Annotation

The first portion of the workflow involves identifying and annotating the phenotype-associated (PA) features to be tested for enrichment (Figure 1A-B). First, genome-wide significant single nucleotide polymorphisms (SNPs) from the GWAS summary statistics are identified using a p-value threshold (e.g. p < 1 10-8). Using Ensembl's Variant Effect Predictor (VEP), the nearest gene within an upstream/downstream buffer distance (bp) is annotated. Users can specify what annotation biotypes are allowed – protein coding, all protein biotypes, or regulatory features. We then use Ensembl's BioMart to locate the chromosome, start, and stop positions of all of these features. These features compose the phenotype-associated (PA) element set for enrichment analysis.

Simulation-based enrichment analysis

Once the element set is defined, we determine parameters for creation of a matched set of genetic regions in the genome (Figure 1C). We split the element set into equally-sized bins and compute the average length within each bin. Using these parameters, we simulate a large number of element sets with BEDTools random (Quinlan & Hall, 2010) which then

comprise a background distribution for testing enrichment. We recommend that users use a minimum of 1,000 simulations and element sets of at least 40 features for analysis. With BEDTools intersect, we calculate the intersections per base pair of sequence between the phenotype-associated features and the comparison set (the elements of interest, e.g. HARs). To determine enrichment, we compute an empirically-computed p-value comparing the observed intersections/bp with the distribution obtained from the random re-sampling (one tailed, Figure 1D).

Prerank

For additional enrichment analysis, tools such as GSEA (Mootha et al., 2003; Subramanian et al., 2005) and WebGestalt (Liao et al., 2019; Wang et al., 2013, 2017; Zhang et al., 2005) require ranked lists of genes. HARE is able to create this ranked list using any selection scan or dataset containing genomic positions and associated p-values. It annotates genes within an upstream/downstream buffer distance (filtering on top hits or based on p-value threshold is allowed) and builds a score for each gene. This score is the average -log10(p-value) for all positions associated with the gene. This gene:score ranked list is provided and can be used with these additional enrichment analysis tools.

Example case

In Kun, Javan, et al. (2023), we used HARE to analyze enrichment of human accelerated regions (HARs) within gene sets associated with skeletal phenotypes. We performed automated image processing of DXA x-rays from over 30,000 individuals from the UK Biobank to generate measurements of bone lengths. We then performed genome wide association studies (GWAS) for 23 image-derived phenotypes such as hip width to shoulder width, arm to leg, and tibia to femur ratios Using HARE, we identified genome-wide significant SNPs associated with each phenotype as well as a number of other phenotypes with publicly-available GWAS summary statistics and created phenotype-associated gene sets. Through comparing these gene sets to human accelerated regions, we found enrichment in 8 phenotypes (Figure 2).

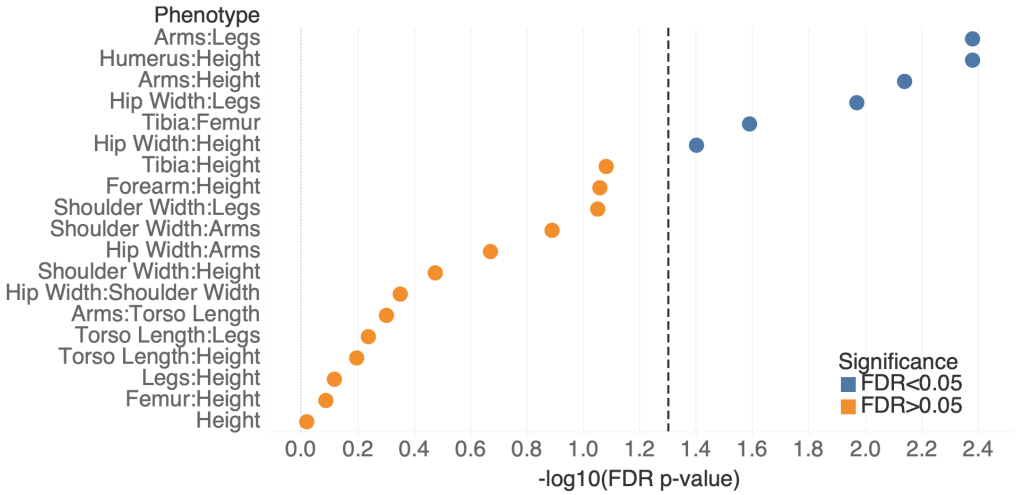

Figure 2: HARE example cases showing p-values of enrichment for overlap between skeletal, dermatological, endocrine, neurological, cancer, metabolic, autoimmune, and gastrointestinal phenotypes and human accelerated regions (HARs) as compared to randomly sampled gene sets of comparable length distribution. Traits with FDR-corrected p-values of less than 0.05 are shown in orange and traits above the threshold are shown in blue.

Acknowledgements

The authors would like to thank Liaoyi Xu for providing additional user testing and feedback. O.S.S. was supported by NSF Graduate Research Fellowship (DGE 2137420). GPU and compute resources were supported by a Director's Discretionary Award from the Texas Advanced Computing Cluster.

References

paper.bib

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., Consortium, R., Psychiatric Genomics Consortium, S. W. G. of the, Consortium, T. R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., … Price, A. L. (2015). "Partitioning heritability by functional annotation using genome-wide association summary statistics". *Nat Genet*, *47*, 1228–1235. https://doi.org/10.1038/ng.3404

Kun, E., Javan, E. M., Smith, O. S., Gulamali, F., Fuente, J. de la, Flynn, B. I., Vajrala, K., Trutner, Z., Jayakumar, P., Tucker-Drob, E. M., Sohail, M., Singh, T., & Narasimhan, V. M. (2023). "The genetic architecture and evolution of the human skeletal form". *Science*, *381*, eadf8009. https://doi.org/10.1126/science.adf8009

Kun, E., Sohail, M., & Narasimhan, V. M. (2023). *"A timeline of human evolution: Leveraging GWAS and comparative genomic data to contextualize human-evolved diseases and morphological traits"*. https://www.ashg.org/wp-content/uploads/2023/10/ASHG2023-PosterAbstracts.pdf

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. (2019). "WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs". *Nucleic Acids Research*, *47*, W199–205. https://doi.org/10.1093/nar/gkz401

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Bignell, A., Boddu, S., Lins, P. R. B., Brooks, L., Ramaraju, S. B., Charkhchi, M., Cockburn, A., Fiorretto, L. D. R., … Flicek, P. (2023). "Ensembl 2023". *Nucleic Acids Res.*, *51*, D933–D941. https://doi.org/10.1093/nar/gkac958

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). "The ensembl variant effect predictor". *Genome Biol*, *17*, 122. https://doi.org/10.1186/s13059-016-0974-4

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., … Groop, L. C. (2003). "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.". *Nat Genet*, *34*, 267–273. https://doi.org/10.1038/ng1180

Quinlan, A. R., & Hall, I. M. (2010). "BEDTools: A flexible suite of utilities for comparing genomic features". *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". *PNAS*, *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). "WEB-based GEne SeT AnaLysis toolkit (WebGestalt): Update 2013". *Nucleic Acids Research*, *41*, W77–83. https://doi.org/10.1093/nar/gkt439

Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. (2017). "WebGestalt 2017: A more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit". *Nucleic Acids Research*, *45*, W130–W137. https://doi.org/10.1093/nar/gkx356

Watanabe, K., Taskesen, E., Bochoven, A. van, & Posthuma, D. (2017). "Functional mapping and annotation of genetic associations with FUMA". *Nat Commun*, *8*, 1826. https://doi.org/10.1038/s41467-017-01261-5

Xu, L., Kun, E., Brasil, M. F., Singh, T., & Narasimhan, V. M. (2023). *"Deep learning to understand the genetic architecture and evolution of the human pelvis"*. https://www.ashg.org/wp-content/uploads/2023/10/ASHG2023-PlatformAbstracts.pdf

Zhang, B., Kirov, S., & Snoddy, J. (2005). "WebGestalt: An integrated system for exploring gene sets in various biological contexts". *Nucleic Acids Research*, *33*, W741–748. https://doi.org/10.1093/nar/gki475