The Journal of Open Source Software

TarGene: A Nextflow pipeline for the estimation of genetic effects on human traits via semi-parametric methods.

Olivier Labayle[1,2], **Breeshey Roskams-Hieter**[1,2], **Joshua Slaughter**[1,2], **Kelsey Tetley-Campbell**[1,2], **Mark J. van der Laan**[4], **Chris P. Ponting**[1], **Ava Khamseh**[1,2,4], **and Sjoerd Viktor Beentjes**[1,3,4]

1 MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom. **2** School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom **3** School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh EH9 3FD, United Kingdom **4** Division of Biostatistics, University of California, Berkeley, CA, USA

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Abhishek Tiwari

Reviewers:

- @ZeyuanSong
- @kescobo

Submitted: 14 July 2025
Published: unpublished

Summary

Genetic variants are the foundation of biological diversity, they play a crucial role in the adaptability, survival, and evolution of populations. Discovering which and how genetic variants affect human traits is an ongoing challenge with applications in healthcare and medicine. In some cases, genetic variants have an obvious effect because they change the coding sequence of a gene and thus its function. In the vast majority of cases however, variants occur in sequences of unknown function and could impact human traits or disease mechanisms in complex ways. TarGene is a Nextflow pipeline leveraging highly flexible machine-learning methods and semi-parametric estimation theory to capture these complex genetic dependencies including higher-order interactions.

Statement of Need

All currently existing software for the estimation of genetic effects are based on parametric distributions, additionally assuming linearity of the relationship between variants and traits (Purcell et al., 2007, pp. yang2011gcta, loh2018mixed, zhou2018efficiently). If these assumptions are violated, the reported effect sizes will be biased and error rates inflated. In particular, this can lead to inflated false discovery rates and suboptimal allocation of computational resources and research funding. Some recently published software also account for more complex relationships but do not offer the full modelling flexibility provided by TarGene. REGENIE fits a two-stage whole-genome model for each phenotype of interest but still assumes linearity and normality (Mbatchou et al., 2021). DeepNull is a semi-parametric method which models non-linear covariate effects but also assumes genetic effects to be linear and does not allow complex interactions between covariates and genetic variants (McCaw et al., 2022). KnockoffGWAS (Sesia et al., 2021) is non-parametric but does not estimate effect sizes, instead it aims at controlling the false discovery rate of variant selection in a genome-wide manner. In comparison, TarGene is the only method able to model arbitrarily complex genetic effects while preserving the validity of statistical inference. It does so by leveraging Targeted Learning (Van der Laan et al., 2011), a framework combining methods from causal inference, machine learning and semi-parametric statistical theory. The estimation process works as follows. In a first step, flexible machine-learning algorithms are fitted to the data. In the second, targeting step, TarGene regularises the estimate of the quantity of interest in a theoretically optimal way.

Features

TarGene is Nextflow pipeline which can be run as follow:

```
nextflow run https://github.com/TARGENE/targene-pipeline/ \
  -r TARGENE_VERSION \
  -c CONFIG_FILE \
  -resume
```

where the `CONFIG_FILE` provides the list of problem-specific parameters (data, arguments, options). Below we list some important features of TarGene. For detailed explanations, please refer to the online documentation.

Scalability

Machine learning methods are computationally intensive, however statistical genetics analyses need to scale to hundreds of thousands of variants and thousands of traits. For this reason, TarGene leverages Nextflow (Di Tommaso et al., 2017), a pipeline management system that can parallelize independent estimation tasks across HPC platforms.

Databases

TarGene works with standard formats, plink `.bed` and `.bgen` formats for genotypes, `.csv` or `.arrow` format for human traits. Furthermore, TarGene has direct support for two large scale biomedical databases, the UK Biobank (Bycroft et al., 2018) and the All of Us cohort (Us Research Program Investigators, 2019). The example considers the UK Biobank for which genotypes and traits are provided via `BED_FILES` and `TRAITS_DATASET` respectively. Because the UK Biobank has a non-standard format, the `UKB_CONFIG` provides trait definition rules. The following is an illustration for the body mass index phenotype, but the default is to consider all 768 traits as defined by geneAtlas (Canela-Xandri et al., 2018).

```
traits:
  - fields:
      - "21001"
    phenotypes:
      - name: "Body mass index (BMI)"
```

Study Designs

TarGene supports traditional study designs in population genetics, that is, genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS). Because TarGene has a focus on complex effects, interactions (e.g. gene-gene, gene-environment, gene-gene-environment) can also be investigated up to any order.

Estimators

In TarGene we default to using Targeted Maximum-Likelihood Estimation (Van der Laan & Rose, 2018) and XGBoost (Chen & Guestrin, 2016) as the machine-learning model. We have selected this default because it was the best performing estimator in simulations across a variety of genetics tasks (Labayle et al., 2025). In the presence of computational restrictions, tradeoffs can be made and lighter models can be used.

Acknowledgements

Olivier Labayle was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. Breeshey Roskams-Hieter was supported by the Health Data Research UK & The Alan Turing Institute Wellcome PhD Programme in Health Data Science (Grant Ref: 218529/Z/19/Z). Mark van der Laan is supported by NIH grant R01AI074345. Chris P. Ponting was funded by the MRC (MC_UU_00007/15). Ava Khamseh was supported by the XDF Programme from the University of Edinburgh and Medical Research Council (MC_UU_00009/2), and by a Langmuir Talent Development Fellowship from the Institute of Genetics and Cancer, and a philanthropic donation from Hugh and Josseline Langmuir.

References

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., & others. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Canela-Xandri, O., Rawlik, K., & Tenesa, A. (2018). An atlas of genetic associations in UK biobank. *Nature Genetics*, *50*(11), 1593–1599. https://doi.org/10.1038/s41588-018-0248-z

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Labayle, O., Roskams-Hieter, B., Slaughter, J., Tetley-Campbell, K., Laan, M. J. van der, Ponting, C. P., Beentjes, S. V., & Khamseh, A. (2025). Semiparametric efficient estimation of small genetic effects in large-scale population cohorts. *Biostatistics*, *26*(1), kxaf030. https://doi.org/10.1093/biostatistics/kxaf030

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., & others. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*(7), 1097–1103. https://doi.org/10.1038/s41588-021-00870-7

McCaw, Z. R., Colthurst, T., Yun, T., Furlotte, N. A., Carroll, A., Alipanahi, B., McLean, C. Y., & Hormozdiari, F. (2022). DeepNull models non-linear covariate effects to improve phenotypic prediction and association power. *Nature Communications*, *13*(1), 241. https://doi.org/10.1038/s41467-021-27930-0

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., & others. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Sesia, M., Bates, S., Candès, E., Marchini, J., & Sabatti, C. (2021). False discovery rate control in genome-wide association studies with population structure. *Proceedings of the National Academy of Sciences*, *118*(40), e2105841118. https://doi.org/10.1073/pnas.2105841118

Us Research Program Investigators, A. of. (2019). The "all of us" research program. *New England Journal of Medicine*, *381*(7), 668–676. https://doi.org/10.1056/nejmsr1809937

Van der Laan, M. J., & Rose, S. (2018). *Targeted learning in data science*. Springer. https://doi.org/10.1007/978-3-319-65304-4

Van der Laan, M. J., Rose, S., & others. (2011). *Targeted learning: Causal inference for observational and experimental data* (Vol. 4). Springer. https://doi.org/10.1007/978-1-4419-9782-1