**Supplementary Information for: An Autoencoder-Based Deep Learning Method for Genotype Imputation**

Meng Song1, Jonathan Greenbaum2, Joseph Luttrell IV1, Weihua Zhou3, Chong Wu4, Zhe Luo2, Chuan Qiu2, Lan Juan Zhao2, Kuan-Jui Su2, Qing Tian2, Hui Shen2, Huixiao Hong5, Ping Gong6, Xinghua Shi7, Hong-Wen Deng2\*, Chaoyang Zhang1\*

1School of Computing Sciences and Computer Engineering, University of Southern Mississippi, Hattiesburg, MS, USA

2Tulane Center of Biomedical Informatics and Genomics, School of Medicine, Tulane University, New Orleans, LA, USA

3College of Computing, Michigan Technological University, Houghton, MI, USA

4Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

5Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR 72079, USA

6Environmental Laboratory, U.S. Army Engineer Research and Development Center, Vicksburg, MS, USA

7Department of Computer & Information Sciences, Temple University, Philadelphia, PA 19122, USA

**Authors for correspondence:**
Dr. Chaoyang Zhang, chaoyang.zhang@usm.edu , Dr. Hong-Wen Deng, hdeng2@tulane.edu

**Disclaimer:** This article reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration and U.S. Army Corps of Engineers.

1. **Supplementary for Evaluation Metrics**

We evaluated our model in terms of evaluation metrics including the concordance rate (CR), the Hellinger score, the scaled Euclidean norm (SEN) score, and the imputation quality score (IQS) (Stahl et al., 2021). In the following equations for different metrics, $x$ is the row vector of the original values, $y$ is the corresponding row vector of the imputed values, and $N$ is the number of samples.

CR is the ratio of correctly imputed SNPs out of all SNPs (Stahl et al., 2021):

$CR=\frac{M}{G} $⑴

where M is the number of matching SNPs, and G represents the total number of SNPs.

Hellinger score is a measure of the distance between two probability distributions (Roshyara et al., 2014):

$H\left(g\right)=1-\sqrt{1-\sum\_{i=1}^{3}\sqrt{f\_{x}^{i}(g)f\_{y}^{i}(g)}}$⑵

where $f\_{x}(g)$ and $f\_{y}(g)$ are the trinomial probability distributions of the true genotype at g and the imputed genotype for each sample, respectively. The original true genotype probability of $f\_{x}\left(g\right)$ can be defined as

$f\_{x}(g))=(p\_{0}, p\_{1}, p\_{2})=(p\_{0}, p\_{1}, 1-(p\_{0}+p\_{1}) )$⑶

where

$\left(p\_{0}, p\_{1}, 1-(p\_{0}+p\_{1}) \right)=\left\{\begin{array}{c}\left(\left(1-ε\right)^{2}, 2ε\left(1-ε\right), ε^{2} \right) if g=0\\\left(ε\left(1-ε\right), \left(1-ε\right)^{2}+ε^{2}, ε\left(1-ε\right) \right) if g=1\\\left(ε^{2}, 2ε\left(1-ε\right), \left(1-ε\right)^{2} \right) if g=2\end{array}\right.$⑷

and parameter $ε$ is the probability of genotyping errors ($ε=0$).

Similarly, the imputed true genotype probability of $f\_{y}(g)$ can be defined as:

$f\_{y}(g))=(\tilde{p\_{0}}, \tilde{p\_{1}}, \tilde{p\_{2}})=(\tilde{p\_{0}}, \tilde{p\_{1}}, 1-(\tilde{p\_{0}}+\tilde{p\_{1}}) $⑸

SEN score calculates the scaled Euclidean distance between the imputed dosage and true dosage per SNP per sample (Roshyara et al., 2014):

$S=(1-\frac{(M^{obs}-M^{imp})^{2}}{4})$⑹

where $M^{obs}$ and $M^{imp}$ are defined as follows:

$M^{obs}=0\*p\_{0}+1\*p\_{1}+2\*\left(1-\left(p\_{0}+p\_{1}\right)\right)=2-(2p\_{0}+p\_{1})$⑺

$M^{imp}=0\*\tilde{p\_{0}}+1\*\tilde{p\_{1}}+2\*\left(1-\left(\tilde{p\_{0}}+\tilde{p\_{1}}\right)\right)=2-(2\tilde{p\_{0}}+\tilde{p\_{1}})$⑻

IQS adjusts for chance agreement to assess the imputation accuracy with an emphasis on rare variants and is calculated based on the observed proportion of agreement ($P\_{o}$) and the chance agreement ($P\_{c}$) as follows (Lin et al., 2010):

$IQS=\frac{P\_{o}-P\_{c}}{1-P\_{c}}$⑼

where $P\_{o}$ and $P\_{c} $are defined as:

$P\_{o}=\frac{\sum\_{i}^{}n\_{ii}}{n\_{..}}$⑽

$P\_{c}=\frac{\sum\_{i}^{}n\_{i.}\*n\_{.i}}{n\_{..}^{2}}$⑾

For each SNP across N samples, each cell of $n\_{ij}$ in a $3×3$ table (e.g., imputed genotypes for AA, AB, and BB, and true genotypes for AA, AB, and BB) represents the number of samples with imputed genotype i and true genotype j, and $n\_{i.}$ and $n\_{.i}$ are the marginal frequencies that would occur if genotypes are called randomly with the same marginal rates.

**Reference**

Browning, B. L., and Browning, S. R. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am J Hum Genet* 84, 210–223. doi: 10.1016/j.ajhg.2009.01.005.

Lin, P., Hartz, S. M., Zhang, Z., Saccone, S. F., Wang, J., Tischfield, J. A., et al. (2010). A New Statistic to Evaluate Imputation Reliability. *PLOS ONE* 5, e9697. doi: 10.1371/journal.pone.0009697.

Roshyara, N. R., Kirsten, H., Horn, K., Ahnert, P., and Scholz, M. (2014). Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genetics* 15, 88. doi: 10.1186/s12863-014-0088-5.

Stahl, K., Gola, D., and König, I. R. (2021). Assessment of Imputation Quality: Comparison of Phasing and Imputation Algorithms in Real Data. *Frontiers in Genetics* 12. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2021.724037 [Accessed July 24, 2022].