Low-Dimensional Genotype Embeddings for Predictive Models

Syed Fahad Sultan
syedfahad.sultan@furman.edu
Furman University
Greenville, South Carolina, USA

Xingzhi Guo
xingzguo@cs.stonybrook.edu
Stony Brook University
Stony Brook, New York, USA

Steven Skiena
skiena@cs.stonybrook.edu
Stony Brook University
Stony Brook, New York, USA

Abstract

We develop methods for constructing low-dimensional vector representations (embeddings) of large-scale genotyping data, capable of reducing genotypes of hundreds of thousands of SNPs to 100-dimensional embeddings that retain substantial predictive power for inferring medical phenotypes. We demonstrate that embedding-based models yield an average F-score of 0.605 on a test of ten phenoypes (including BMI prediction, genetic relatedness, and depression) versus 0.339 for baseline models. Genotype embeddings also hold promise for creating sharing data while preserving subject anonymity: we show that they retain substantial predictive power even after anonymization by adding Gaussian noise to each dimension.

CCS Concepts: • **Applied computing** → **Bioinformatics**.

Keywords: genotype, embeddings, privacy-preserving

ACM Reference Format:
Syed Fahad Sultan, Xingzhi Guo, and Steven Skiena. 2022. Low-Dimensional Genotype Embeddings for Predictive Models. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22), August 7–10, 2022, Northbrook, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3535508.3545507

1 Introduction

At 3 billion base pairs, the high dimensionality of the human genome presents challenges in build predictive models. Genomic data, considering as a matrix, yields more SNPs (features) than there are subjects (rows) thus making it difficult to avoid overfitting. Each person is uniquely identifiable from their DNA, which makes it difficult to anonymize this data when sharing with biomedical research community.

ACM-BCB, August 07–10, 2022, Chicago, IL

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3535508.3545507

In this paper, we investigate the question of how best to construct compressed representations (embeddings) of genome-level data which are (1) compact enough to form useful features, even for modest sample sizes, and (2) provide representations that preserve privacy, so one cannot expand the representation to identify an individual.

Our major contributions in this paper are:

1. *Methods for Constructing Low-Dimensional Genotype Embeddings* – We investigate several design decisions in creating genotype embeddings. These include encoding heterozygosity, representing major/minor alleles, haplotype-sensitive chromosome partitioning, different matrix representations and dimensionality reduction techniques. We show that i) Hardy-Weinberg Equilibrium (HWE) [6] Principal Component Analysis (PCA) substantially outperforms traditional PCA on predicting a range of phenotypes ii) selection techniques based on genome-wise association studies (GWAS) are substantially more effective at capturing medical phenotypes but not ethnicity.
2. *Predicting Medical Phenotypes from Genotype Embeddings* – We study the power of genotype embeddings on a cohort of 20,000 people and over 5,000 different phenotypes from UK-Biobank [4]. We demonstrate that embedding-based models yield an average F-score of 0.605 on a test of ten phenoypes versus 0.339 for the baseline models.
3. *Privacy-Preserving Genotype Embeddings* – We demonstrate that privacy guarantees for the embeddings can be provided by adding Gaussian noise to the dimensions of the embedding such that (a) the nearest neighbor to the noisy representation is unlikely to be the original subject, while (b) preserving much of the predictive power of the uncorrupted embeddings.

2 Methods

2.1 Data

Our data is composed of 805,426 biallelic autosomal variants for of 19,831 subjects in UK-Biobank [4]. Each variant corresponds to a genomic locus and is composed of a number of alleles (ploidy). The calls in our data are diploid with two alleles at a call. Each of the two alleles correspond to the two chromosomes, one from each parent. Furthermore, each variant also has corresponding major/reference and minor/alternate alleles, as per a genome assembly. Here, we

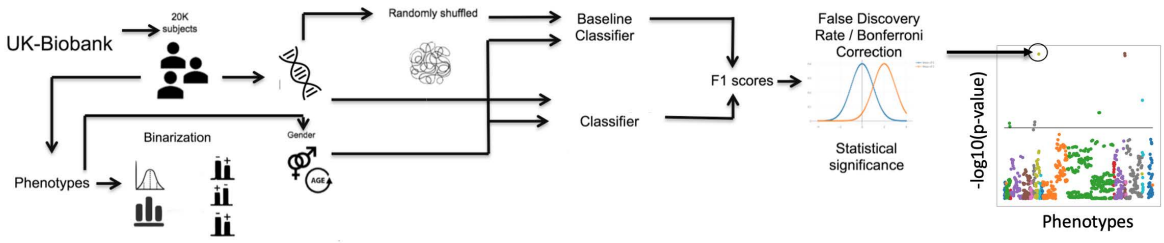

Figure 1. Schematic overview of the evaluation framework for classification

use the genome assembly GRCh37 as reference. For each subject, we had 4926 phenotypes including medical and socioeconomic information.

2.2 Embeddings Methods

Our embedding methods are defined by a mix of matrix operations, sequence preprocessing, and genome partitioning methods. We outline the space of design decisions below.

2.2.1 Genome Sequence Processing In order to embed the genotype data into lower dimensions, we first need to convert it into a numerical representation. Almost all encodings are based primarily on two properties of the variants: 1) If the two alleles are identical or not (homozygous / heterozygous) 2) If the alleles match the reference or not (major/minor allele). Because of these two binary properties, most encodings are at least two bits. In our experiments, we focus on two encodings: i) SparSNP [2] encoding both cases of heterozygosity ii) count of minor alleles.

2.2.2 Base selection In evaluating various embedding techniques on the genotypic data, we try the following feature selection strategies to remove any irrelevant bases.
Entropy selected array – We removed SNPs that were likely to be erroneous reads, and those which were found to be either the same for most subjects (entropy < 0.1) or had too much variance (entropy > 0.9) .
GWAS Catalog selected array – Genome-Wide Association Studies (GWAS) catalog [3] is a public available database of SNP-trait associations, composing of many individually published genome-wide association studies. Here we use GWAS catalog as a feature selection tool for our models, retaining only SNPs found in previously published work in the GWAS catalog. We use v1.0 of the catalog with results from 4188 publications reporting 214,295 associations on 133,651 SNPs with 4662 phenotyes/disease/traits.
Full SNP array – In addition to employing the above outlined strategies for base selection, in our experiments, we also used the full SNP array for comparison.

2.2.3 Chromosome partitioning
Haplotype-based partitioning – The haplotype boundaries are identified using recombination hotspots in terms of a unity for measuring genetic linkage (centimorgans). Chromosomes were partitioned into haplotypes with a threshold of ≥ 100 centimorgan using values from Phase II HapMap [5].
Equal partitioning – Principal Component Analysis (PCA) using Singular Value Decomposition (SVD) is applied on each individual $m \times n$ matrix where n is the number of SNPs in the chromosome. For each chromosome (25 in total, chromosome 1-22, MT, X, Y), k top principal components are concatenated together to make the final $m \times (25 * k)$ dimensional embeddings.

2.2.4 Matrices In addition to embedding the $n \times m$ genotype matrix directly, we alternatively tried converting it into a dense intermediary $n \times n$ matrix first and then embedding that into $m \times d$ dimensions. We tried the following two intermediary representations:
Pairwise Hamming Distance matrix – We use hamming distance to compute pairwise distance between all subjects based on the binary representations of their genomic sequence encoding and base selection. This $n \times n$ matrix is then reduced using a variant of PCA, and the top k principal components are concatenated to form the embeddings.
Genetic relatedness matrix (GRM) – The genetic relationship matrix (GRM) G encodes genetic correlation between each pair of samples. It is defined by $G = MM^T$ where M is a standardized version of the genotype matrix.

2.2.5 Matrix compression For the final matrix compression, we use PCA [1] and Hardy-Weinberg normalized PCA (HWE-PCA) [6], a specialized version of PCA for statistical genetics based on projecting samples to a small number of ancestry coordinates.

3 Evaluation

In this study, we restrict ourselves to the problem of binary classification for each of the 5,034 phenotypes. Without binarization of the ordinal variables, evaluation of ordinal and nominal variables would have been inconsistent and incomparable. Since age and sex can be a strong confounding factor for most phenotypes, we corrected for age and sex by using them as features to both our genotype and baseline classifier. For our baseline model, we used a Logistic Regression classifier with the features age, sex and randomly shuffled

Table 1. A comparative evaluation of different genotype embeddings revealed that haplotype and whole-genome partitioning, full SNP array and HWE PCA improve classification scores for ethnicity. On the other hand, chromosome level partitioning and GWAS SNPs are better suited for prediction of medical conditions. Hamming Distance matrix and/or removing high/low entropy SNPs had no affect. Minor Allele Count was used for encoding of whole genome embeddings and SparSNP [2] for rest.

				Ethnicity			Medical Conditions		
Base Selection	**Partitioning**	**Matrices**	**Matrix Compression**	**Asian or Asian British**	**Black or Black British**	**Other ethnic group**	**Diabetes**	**Cancer**	**Bipolar Disorder**
Full SNPs	Per chromosome	Hamming	SVD/PCA	0.71	0.38	0.17	**0.07**	0.1	0
Entropy	Haplotype	Genotype	SVD/PCA	**0.77**	0.54	**0.2**	0.05	0.1	0
Full SNPs	Haplotype	Genotype	SVD/PCA	0.75	**0.56**	**0.2**	0.05	0.1	0
Entropy	Per chromosome	Genotype	SVD/PCA	0.64	0	0.1	0.06	0.1	0
Full SNPs	Per chromosome	Genotype	SVD/PCA	0.62	0	0.09	0.06	**0.11**	**0.1**
GWAS SNPs	Whole genome	GRM	SVD/PCA	0.71	0.54	0.3	0.29	0.24	0.12
Full SNPs	Whole genome	GRM	HWE PCA	**0.97**	**0.61**	**0.54**	0.1	0.09	0.05
GWAS SNPs	Whole genome	GRM	HWE PCA	0.93	**0.61**	0.4	**0.32**	**0.54**	**0.13**

Table 2. Genotype-phenotype linkages identified with strong statistical significance (**; p-value ≤0.001, bonferroni corrected) across different phenotype categories.

Description	**Model recall**	**Baseline recall**	**Model precision**	**Baseline precision**	**Model acc**	**Baseline acc**	**Model f1**	**Baseline f1**
Body Mass Index	0.737	0.504	0.579	0.377	0.704	0.507	0.649	0.431
Body Weight Percentage	0.734	0.49	0.579	0.369	0.701	0.496	0.648	0.421
Cholesterol and other meds	0.954	0.436	0.123	0.037	0.746	0.555	0.218	0.068
Genetic kinship	0.956	0.458	0.123	0.039	0.747	0.562	0.219	0.072
Genetic relatedness	0.957	0.389	0.122	0.033	0.745	0.558	0.217	0.061
Trauma	0.999	0.47	0.999	0.438	0.999	0.464	0.999	0.453
Depression	0.638	0.475	0.408	0.324	0.574	0.498	0.498	0.385
Assessment Center	0.998	0.512	0.793	0.518	0.871	0.528	0.884	0.515
Ethnic: White vs. rest	0.886	0.522	0.997	0.456	0.95	0.525	0.938	0.486
Ethnic: Black vs. rest	0.999	0.514	0.999	0.479	0.999	0.505	0.999	0.496

genomic features, to ensure parity in number of features and their respective distributions. In order to evaluate how impressive the results of our genotype classifier were from baseline, we report statistical significance scores.

We create 20 different splits for subjects into 20 different sets of training and testing subjects, using the train:test ratio of 75:25. For each split, we trained classifiers on the training subjects and tested on the testing subjects, computing f1 score for each one. Using the two f1-score distributions, each with 20 constituent data points, one for our model and one for the baseline, we performed simple t-tests. We use this p-value and its associated t-test statistic as the two primary measures of quality of our models. We used Bonferroni correction to correct for family-wise error rates.

4 Results

4.1 Design choices in genotype embeddings

Table 1 present the results of our experiments in terms of prediction quality (F1 Scores) with respect to the different embedding design choices. Our results show that using GWAS SNPs for base selection significantly improves prediction of medical conditions but not ethnicity. On the contrary, HWE normalized PCA significantly improves ethnicity prediction over regular PCA. The Hamming distance pairwise comparison fails to significantly improve genotype embeddings. Partitioning on recombination hotspots (haplotypes) improves prediction of ethnicity as compared to equi-partitioning of chromosomes. Partitioning on chromosomes improves prediction of medical conditions. Removing SNPs with extremely high or extremely low entropy has little effect on prediction scores.

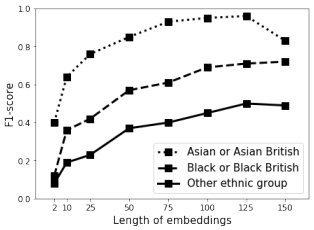

Figure 2. Effect of embedding dimensionality on prediction. Increasing the length of the embeddings improves the prediction of ethnicity across different categories, but the improvements taper off around $d = 100$.

4.2 Dimensionality of Embeddings

Figure 2 gives results for prediction of ethnicity using different lenghth of the genotype embeddings learned using GWAS SNPs and HWE-PCA. Increasing the length of the embeddings, upto a point, improves the prediction of ethnicity across different categories, tapering off around $d = 100$. The inflection point seems to be consistent across different ethnic labels.

4.3 Privacy Preserving Embeddings

Genotype embeddings for a subject can be anonymized by adding Gaussian noise $\mathbb{N}(\mu = 0, \sigma =$ standard deviation of each dimension) while still preserving information of the phenotypes. We measure the privacy preservation by comparing the nearest-neighbors in the embedding space before and after anonymization. Figure 3 (top) shows that beginning at around $\mathbb{N}(0, \sigma)$ of noise, a subject's 50 dimensional embeddings become indistinguishable from an average 2-5 nearest neighbors. At $\mathbb{N}(0, 1.2\sigma)$, this rank similarity decreases so as to make the subject anonymous among a group of 100. Meanwhile, for the same 50-length embeddings, it takes a higher amount of noise to adversely impact the prediction score of ethnicity. The longer the length of the embeddings, the more the noise needs to be added, as Figure 3(top) shows, around 1.5 σ of each dimension of Gaussian noise, the nearest neighbor to the noisy representation is unlikely to be the original subject, while Figure 3(bottom) demonstrates that at the noisy level of 1.5 σ, it preserves much of the information about the subject's broader ethnicity.

4.4 Phenotype prediction

A sample of phenotypes of interest and their corresponding prediction statistics are shown in Table 2. Our concise 100-dimensional embeddings were able to predict some key medical conditions such as mental health variables including depression and trauma as well as Body Mass Index (BMI) and Body weight percentage. If the subject had been prescribed any medication for cholesterol, blood pressure or diabetes, then that information was also preserved. Altogether, our results serve as evidence of preservation of wide diversity of traits and phenotypes in a drastically smaller number of dimensions in our genotype embeddings.

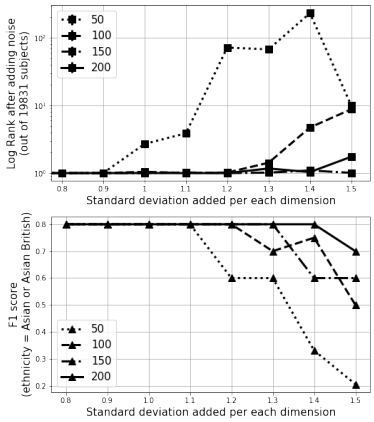

Figure 3. Anonymizing genotype embeddings by adding Gaussian noise to each dimension. The embeddings prove more robust to added noise as dimensionality increases. The nearest neighbor to the noisy representation becomes increasingly unlikely to be the original subject (top), while even at the noise level of 1.5 σ, it still preserves much of the predictive power of the embedding (bottom).

5 Conclusion

In this work, we construct low-dimensional embeddings of large-scale genotyping data, reducing genotypes of hundreds of thousands of SNPs to embeddings on the order of 10-100 dimensions. We demonstrate that these embeddings retain a lot of predictive power while also preserving privacy. For future work, we envision focusing on theoretical bounds for the pareto optimal embedding sizes as well as looking at finer resolution of genomic conservation beyond haplotypes.

References

[1] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.

[2] Gad Abraham, Adam Kowalczyk, Justin Zobel, and Michael Inouye. 2012. SparSNP: Fast and memory-efficient analysis of all SNPs for phenotype prediction. *BMC bioinformatics* 13, 1 (2012), 1–8.

[3] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* 47, D1 (2019), D1005–D1012.

[4] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 7726 (2018), 203–209.

[5] International HapMap Consortium et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 7164 (2007), 851.

[6] Sun Wei Guo and Elizabeth A Thompson. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* (1992), 361–372.