Identification of individuals by trait prediction using whole-genome sequencing data

姓名：程文文
学院：信息学院

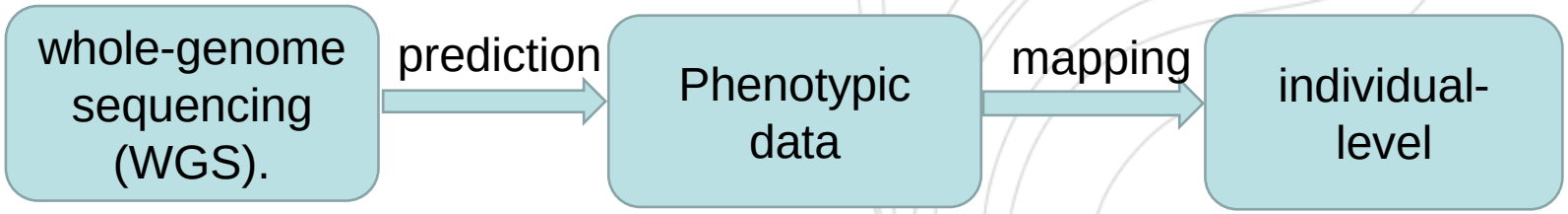
Proc Natl Acad Sci USA. 2017 Sep 19; 114(38): 10166–10171.

Contents

- 为了探索当前基于表型的基因组鉴定的能力，利用全基因组测序数据（WGS）、详细的表型分型、统计建模来预测1061名不同祖先参与者的生物特征。
- 开发了最大熵算法，其整合了多个预测来确定哪个基因组样本和表型测量来源于同一个体。
- 使用这种算法，我们能够在10个混合人群中重新识别8个以上个体的血缘；在10个非裔美国人或10欧洲人中平均能识别5个个体。

Ethical and Legal

- Large genetic databases online
- 23andMe
- Health Insurance Portability and Accountability Act (HIPAA)
- In this study :



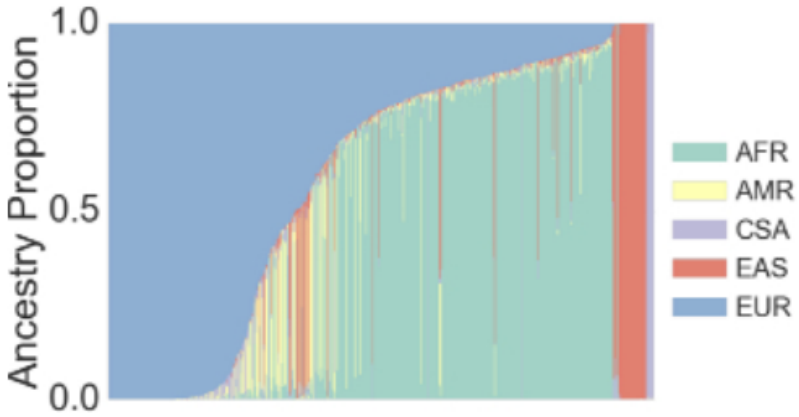
- Phenotype such as skin color, eye color, and ficial structure.

Study Population

A

	African	European	Latino	East Asian	South Asian	Other
Female	381	126	39	39	8	39
Male	188	147	24	24	10	36

B



C

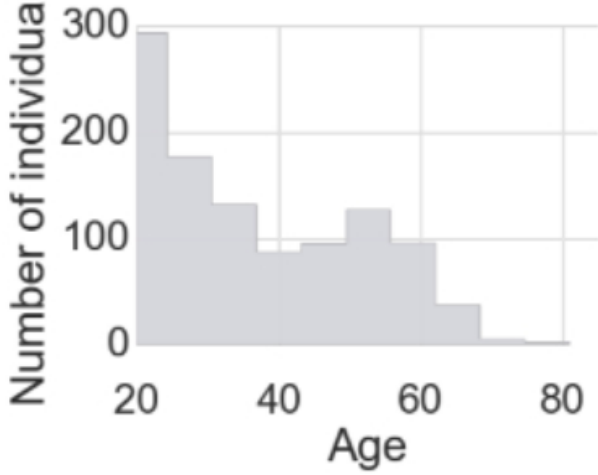


Fig. 1. Study overview. (*A*) Distribution of self-reported ethnicity in the study. (*B*) Inferred genomic ancestry proportions for each study participant. Ancestry components are African (AFR), Native American (AMR), Central South Asian (CSA), East Asian (EAS), and European (EUR). (*C*) Distribution of ages in the study.

Result 1

Predicting face and voice :

1. Represented face shape and texture variation using principal components (PC) analysis to define a low-dimensional representation of the face.
2. Next, we predicted each face PC separately using ridge regression with ancestry information from 1,000 genomic PCs , with sex, BMI, and age as covariates.

Fig. 2. Examples of real (*Left*) and predicted (*Right*) faces.

Assess the impact of covariates：

- we measured the per-pixel $R_{CV}^{\ 2}$ between observed and predicted faces.
- Because errors were anisotropic, we separated residuals for horizontal, vertical, and depth dimensions.

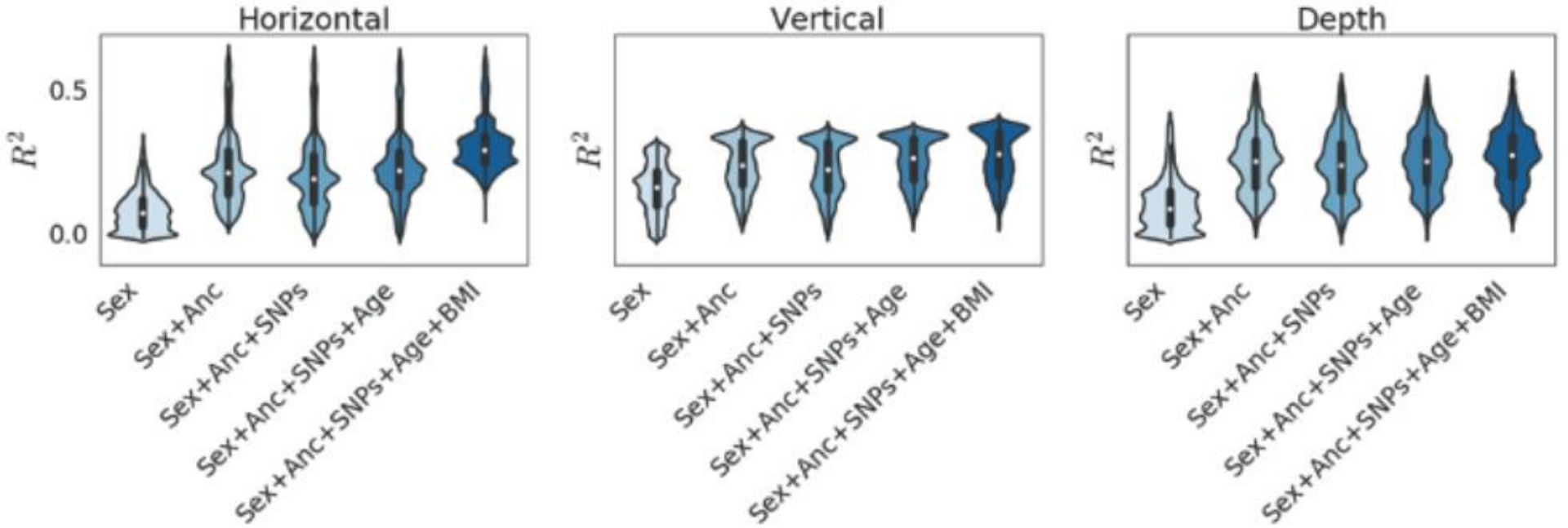


Fig. 3

- To further understand predictive accuracy for the full model,we mapped per-pixel accuracy onto the average facial scaffold, finding that most of the predictive accuracy was in facial regions that differed the most between African and European individuals.

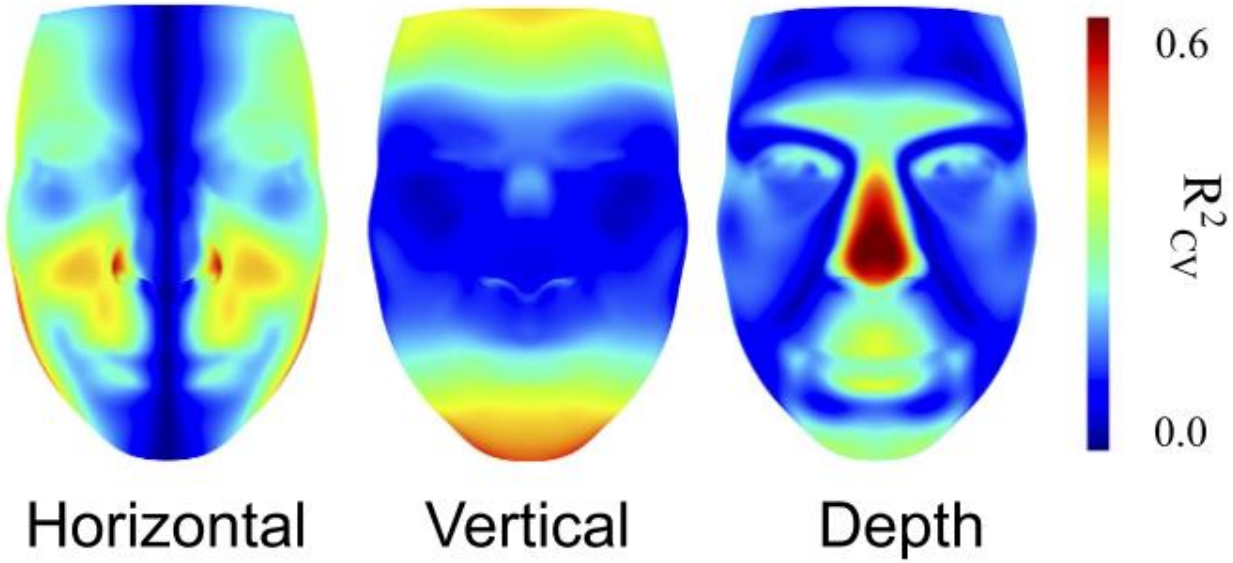


Fig. 4. Per-pixel R^2_{CV} in face shape for the full model, across three shape axes.

- Besides genomic prediction, our method for reidentification used predictions from image and voice embeddings.

Table 1. Prediction from images and voice samples

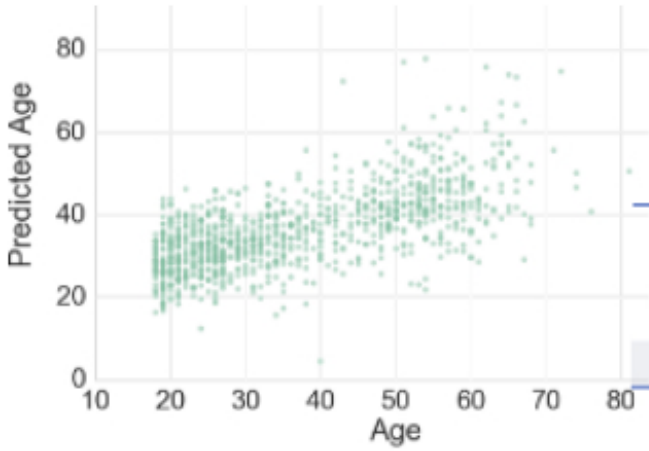
Source trait	Age	Sex	AFR	EUR	EAS	AMR	CSA
Shape	0.82	0.79	0.84	0.78	0.57	0.16	0.11
Color	0.75	0.84	0.89	0.84	0.62	0.24	0.24
Voice	0.62	0.70	0.67	0.38	0.14	0.03	0.02

R^2_{CV} values for age, sex, and five components of genetic ancestry from face shape (shape), face color (color), and voice.

Result 2

Predicting Age from WGS Data:

A：Predicted vs. true age.
R^2_{CV} for models using features including telomere length and X/Y copy numbers.



A

Age

Features	R^2_{CV}
Telomeres	0.29
X/Y Copy	0.31
Both	0.44

Height, Weight, and BMI Prediction:

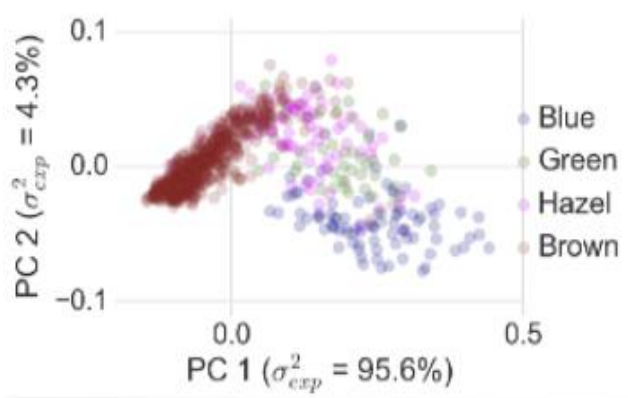
B：Predictive performance for height, weight, and BMI using covariate sets composed from predicted age and/or sex, 1,000 genomic PCs, and previously reported SNPs.

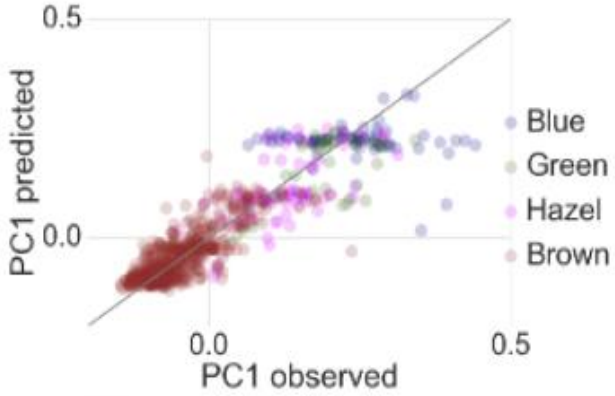
B

Height, weight, BMI R^2_{CV}

Features	Height	Weight	BMI
Age	0.02	0.04	0.07
Age + Sex	0.44	0.05	0.09
Age + Sex + Genomic PCs	0.50	0.15	0.17
Age + Sex + Reported SNPs			0.17
Age + Sex + Genomic PCs + Reported SNPs	0.53	0.15	0.17

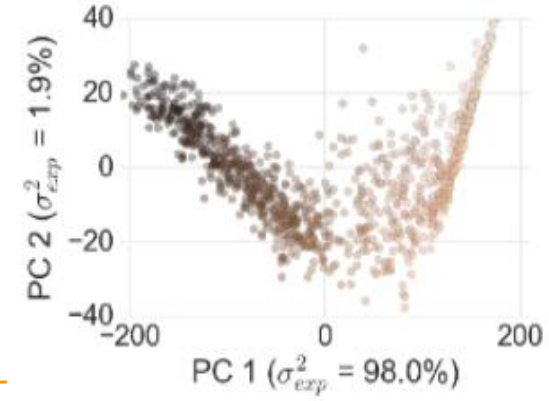
Eye Color and Skin Color Prediction:

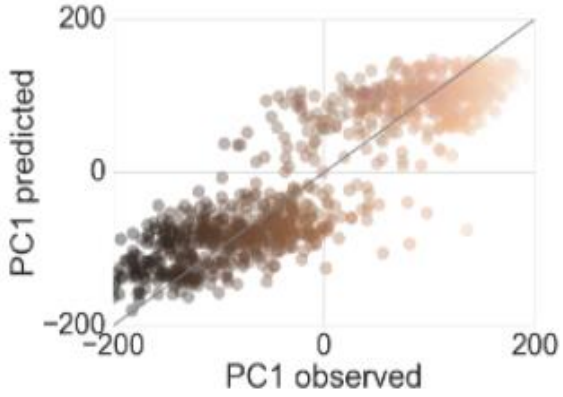




C

Eye color	R^2_{CV}		
Features	**R**	**G**	**B**
Genomic PCs	0.74	0.67	0.58
Reported SNPs	0.74	0.79	0.78
Genomic PCs + Reported SNPs	0.80	0.82	0.80





D

Skin color	R^2_{CV}		
Features	**R**	**G**	**B**
Genomic PCs	0.71	0.76	0.78
Reported SNPs	0.68	0.71	0.75
Genomic PCs + Reported SNPs	0.77	0.79	0.81

Result 3

Linking Genomes to Phenotypic Profiles:

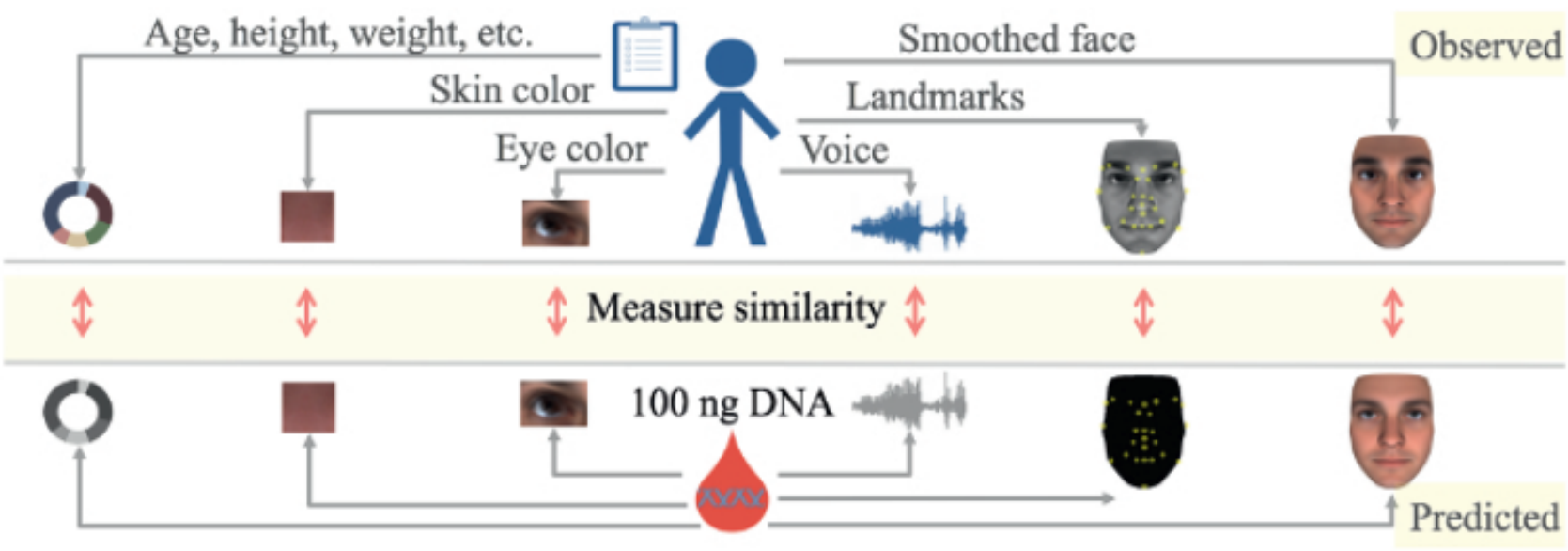


Fig. 6. Overview of the experimental approach. A DNA sample and a variety of phenotypes are collected for each individual. We used predictive modeling to derive a common embedding for phenotypes and the genomic sample as detailed in *SI Appendix*, Table S14. The concordance between genomic and phenotypic embeddings are used to match an individual's phenotypic profile to the DNA sample.

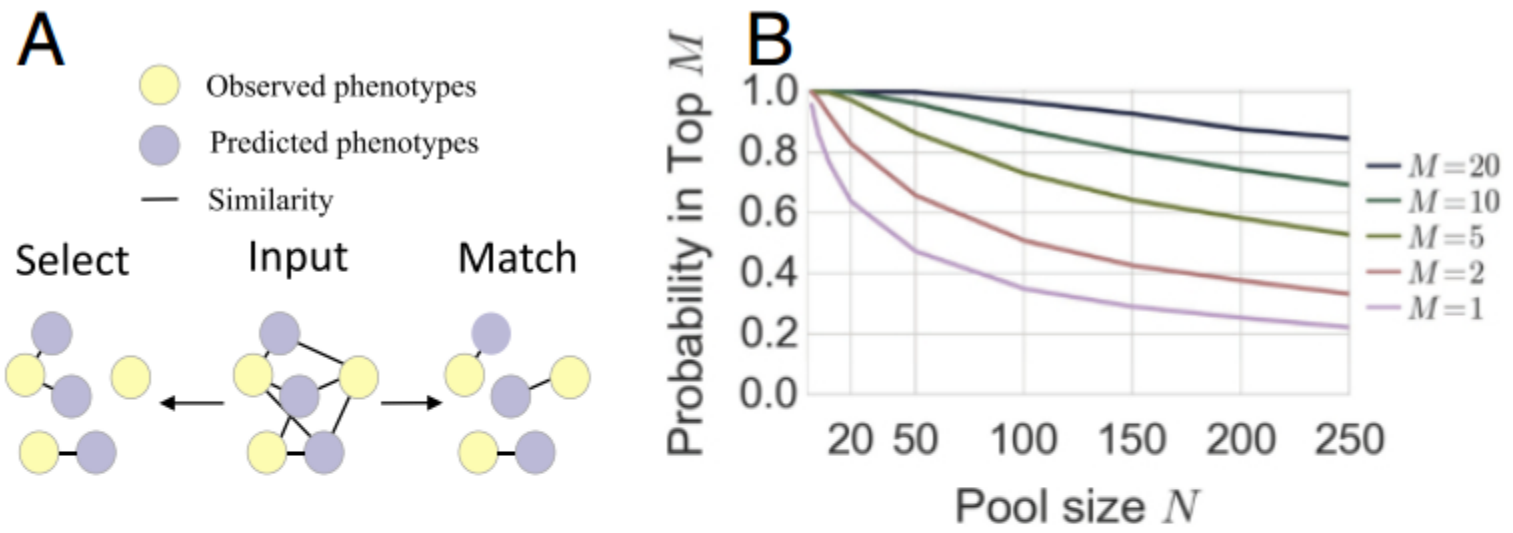


Fig. 7. Ranking individuals. (*A*) Schematic representation of the difference between select (best option chosen independently) and match (jointly optimal edge set chosen). Select corresponds to picking an individual out of a group of *N* individuals based on a genomic sample. Match corresponds to jointly matching a group of individuals to their genomes. (*B*) Ranking performance. The empirical probability that the true subject is ranked in the top *M* as a function of the pool size *N*.

Table 2. Top one accuracy in match and select

		Pool size	2	5	10	20	50
Together	Full	Match	0.97	0.92	0.83	0.7	0.53
		Select	0.93	0.83	0.74	0.62	0.45
	All Face + Demogr.	Match	0.98	0.91	0.82	0.7	0.53
		Select	0.93	0.83	0.73	0.61	0.45
	All Face + Add'l	Match	0.96	0.85	0.72	0.55	0.33
		Select	0.91	0.76	0.63	0.48	0.28
	All Face	Match	0.95	0.84	0.71	0.53	0.32
		Select	0.9	0.76	0.62	0.46	0.29
All Face	3D Face	Match	0.94	0.8	0.64	0.46	0.25
		Select	0.89	0.74	0.58	0.42	0.24
	Landmarks	Match	0.87	0.61	0.39	0.24	0.1
		Select	0.81	0.55	0.38	0.23	0.11
	Eyecolor	Match	0.85	0.55	0.33	0.19	0.075
		Select	0.8	0.52	0.33	0.19	0.085
	Skincolor	Match	0.81	0.5	0.29	0.16	0.065
		Select	0.79	0.47	0.29	0.16	0.07
Demogr.	Ethnicity	Match	0.9	0.71	0.54	0.41	0.27
		Select	0.87	0.66	0.5	0.36	0.25
	Age	Match	0.69	0.35	0.19	0.1	0.042
		Select	0.66	0.35	0.2	0.11	0.043
	Gender	Match	0.74	0.39	0.22	0.12	0.051
		Select	0.73	0.39	0.21	0.11	0.049
Add'l	Voice	Match	0.88	0.66	0.44	0.26	0.11
		Select	0.84	0.61	0.42	0.26	0.12
	Height/Weight/BMI	Match	0.77	0.46	0.27	0.14	0.061
		Select	0.74	0.43	0.26	0.15	0.065
	Random	Match	0.5	0.2	0.1	0.05	0.02
		Select	0.5	0.2	0.1	0.05	0.02

Reidentification accuracy in select and match averaged over all possible lineups formed for each CV fold of different pool sizes from 2 to 50 using the various phenotype sets

Summary

- 该研究打破了传统的GWAS、eQTL已知表型对基因进行差异表达分析的思想，而是利用WGS数据对表型进行预测，具有创新性思维。
- 但是还需要进一步优化模型提高预测的准确率。
- 启发我们在科研之路应善于应用逆向思维思考问题，综合学习。
- 是否可以应用到利用植物基因组数据对其表型预测？？

THANKS !