What's in *your* DNA? Balancing privacy and utility when releasing genetic data

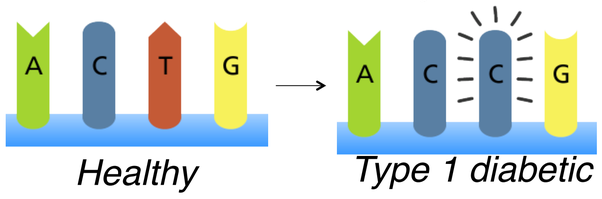
Carlos Rojo[1], Daniel Aranki[2], Ruzena Bajcsy[2]
1. San José City College, 2. UC Berkeley

Abstract

Genetic datasets are a rich information source that continue to guide biomedical research. However, genetic datasets are extremely sensitive, as they can be used to reveal information like ethnicity and disease risk. Here we explore the implementation of user-specified personalized privacy to balance the need to release genetic data with high utility (able to predict disease risk) with the desire to keep certain information private (such as ancestry).

Motivation and Goal

- Single variations (SNPs) reveal disease risk and ancestry.



- Increasingly being used in the clinic.
- Recent privacy breaches show that even releasing summary statistics is not private.
- How can we share genetic data while maintaining privacy?

Goal: user-defined personal privacy on genetic data

"Which of the following would you like to hide when releasing your data?"

- ☐ Alzheimer's Risk
- ☐ Type I Diabetes Risk
- ☑ Ancestry

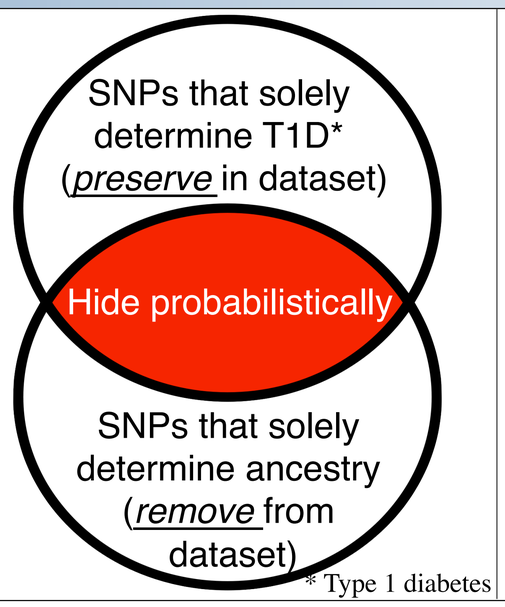
Diabetes Classifier

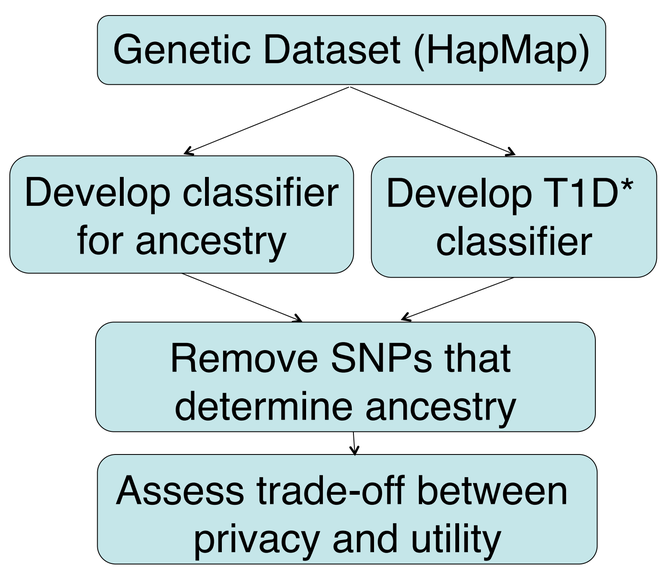
- Previous work has identified SNPs that are associated with Type 1 diabetes.
- A model to predict T1D based on genetic data established by Winkler, C., Krumsiek, J. et al. (1)
- Can use this weighted model to classify Type 1 diabetes risk based on genetic data of individual.

$$\log_e\left(\frac{p}{1-p}\right) = \beta_o + \beta_1(SNP_1) + \beta_2(SNP_2) + \ldots + \beta_n(SNP_n)$$

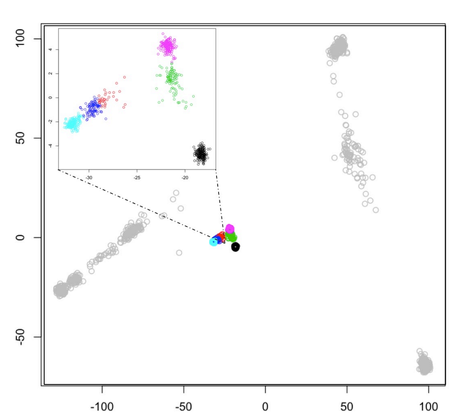
p = probability of having Type 1 diabetes
B_0 = baseline Type 1 diabetes risk
B_i = proportion that each SNP contributes to Type 1 diabetes risk
SNP_n = Sequence at the SNP. That is, whether indiv. has disease allele

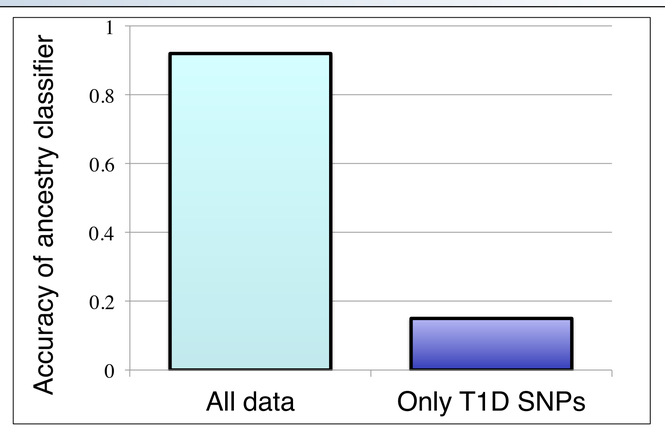
Analysis Overview





Results

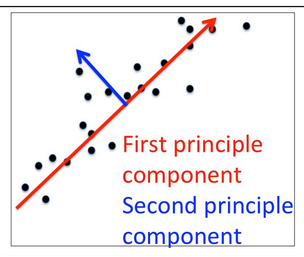


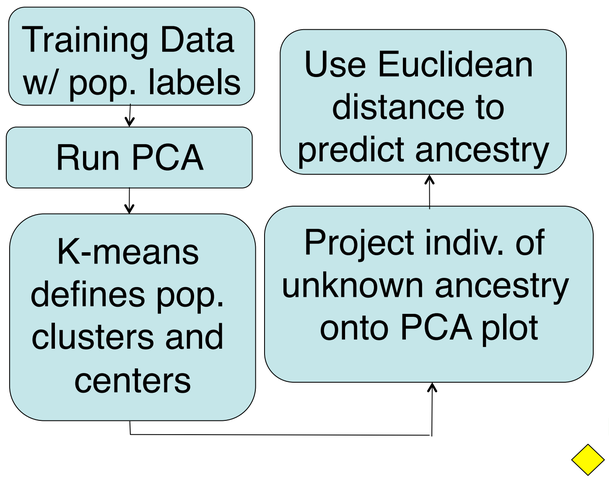


Left: Figure demonstrating how ability to classify ethnicity changes when you remove ancestry-determining SNPs (center colored clusters). Harder to distinguish ancestry via Euclidean distance (shown is what happens when you mask all SNPs except those that cause T1D and chromosome 6).
Right: Accuracy of ethnicity classifier after removing all SNPs except those that determine T1D risk. Utility has not changed (since no T1D SNPs were removed) but privacy is increased.

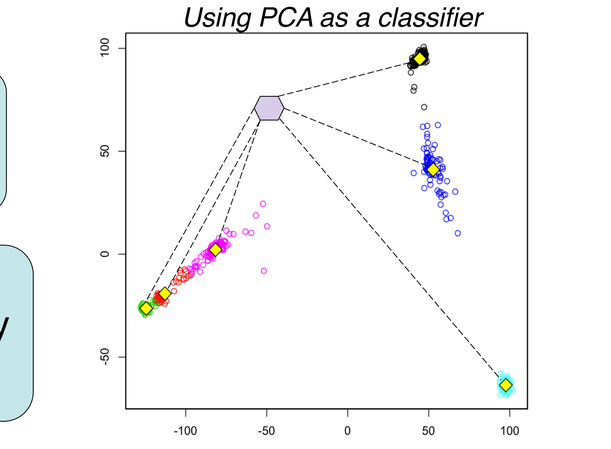
Ancestry Classifier

- Principle Components Analysis (PCA) highlights principle SNPs that explain variability within genetic dataset.
- PCA can be used to stratify individuals based on ancestry.



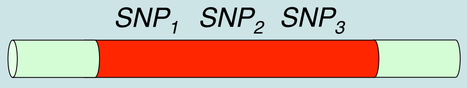


Using PCA as a classifier



Conclusions and Future Directions

- Whenever privatizing data, there can be a trade off between utility and privacy.
- Using genetic classifiers, personalized privacy on genetic data is possible. Here we demonstrate that you can effectively remove the prediction of ancestry while maintaining ability to diagnose a disease like Type 1 diabetes.
- Future work will attempt to increase the utility by allowing for prediction of risk for multiple diseases, while still privatizing ancestry determination.
- Future work will establish global guarantee against future classifiers.
- This work will also require addressing the issue of *linkage disequilibrium* (LD, or correlation between SNPs) to ensure privacy:



Removing SNP_2 from the dataset provides no privacy as it can be deduced from the sequence at SNP_1 and SNP_3. SNP_1, SNP_2, and SNP_3 are said to be in LD

The authors would like to thank the E3S center and its staff, Josephine Yuen, James Hake, and Lea Marlor for coordinating the RET summer program.

Contact Information
Carlos Rojo: Carlos.Rojo@sjcc.edu

Support Information
This work was funded by National Science Foundation Award EEC-1405547.

References:
(1) Winkler, C., Krumsiek, J. et al. Diabetologia (2014)