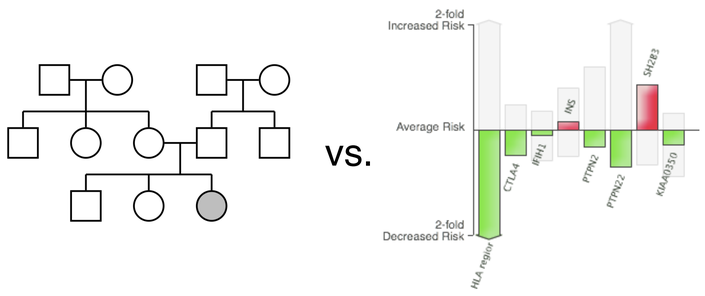
Prediction of complex multifactorial disease
Comparing family history and genetics

C.B. Do[1], J.M. Macpherson[1], D.A. Hinds[1], B.T. Naughton[1], U. Francke[1,2], N. Eriksson[1].

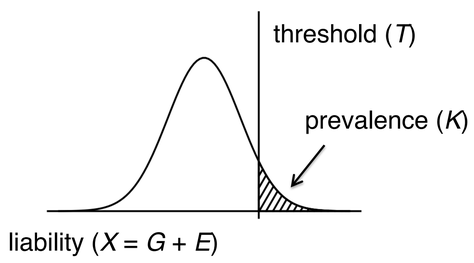
[1]23andMe, Inc, Mountain View, CA; [2]Stanford University School of Medicine, Stanford, CA.

Introduction

Although the use of **family history** and **SNP-based risk assessment** is well understood for simple Mendelian disorders, to date, little is known regarding the **relative performance** of these methods for **complex polygenic diseases**.

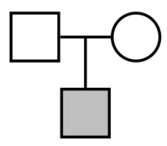


In this study, we used the standard **liability-threshold model** from quantitative genetic theory to analyze the influence of **disease prevalence (K)** and **heritability (h_L^2)** on the predictive accuracy of family history and SNP-based models.



Methods

Consider any arbitrary family structure, such as the trio shown below, where one member is specially designated as the index individual:



and where the correlations in genetic and environmental liabilities for individuals in the family are assumed to be fully specified. For example, in the family above, we might assume:

$$\begin{bmatrix} G_1 \\ G_2 \\ G_3 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, h_L^2 \cdot \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix} \right) \qquad \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, (1 - h_L^2) \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)$$

We can calculate the joint probability of any pattern of disease occurrence in a family using multivariate integration (e.g., $P(D_1 = 1, D_2 = 0, D_3 = 0)$). Using Bayes' rule, we can then calculate various conditional probabilities of disease associated with each family history pattern:

family history pattern $((D_2, D_3))$	risk $(P(D_1 = 1 \mid D_2, D_3))$	frequency among cases $(P(D_2, D_3 \mid D_1 = 1))$	frequency among controls $(P(D_2, D_3 \mid D_1 = 0))$
(0,0)	0.1806	0.4064	0.6145
(0,1)	0.3152	0.2364	0.1712
(1,0)	0.3152	0.2364	0.1712
(1,1)	0.4833	0.1208	0.0431

which in turn allow us to compute various estimates of predictive accuracy (e.g., AUC, sensitivity, PPV).

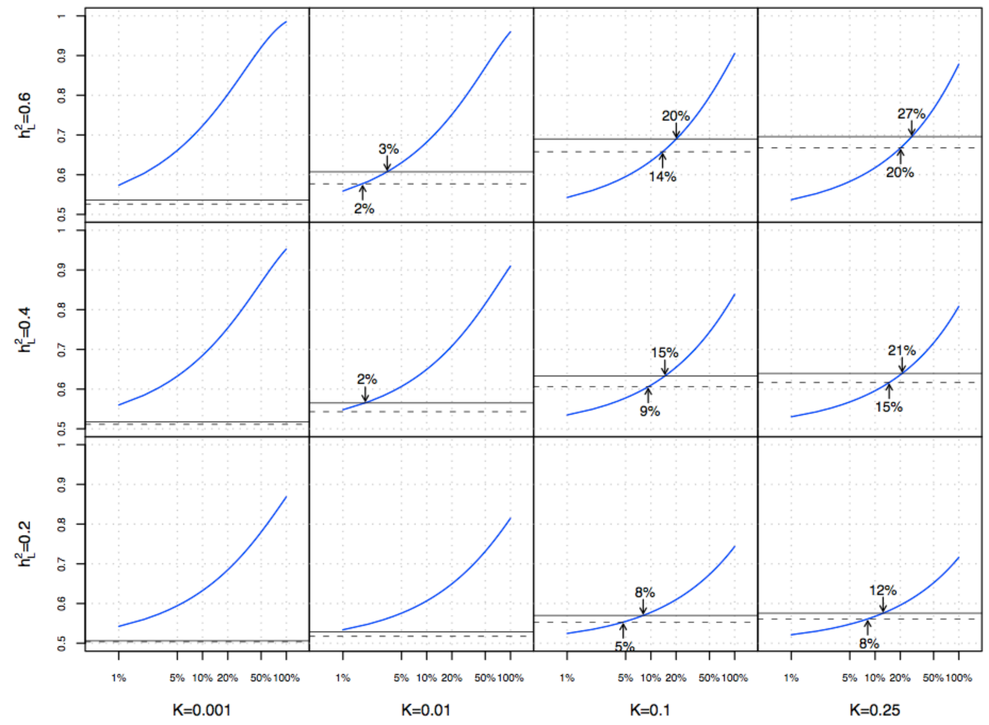
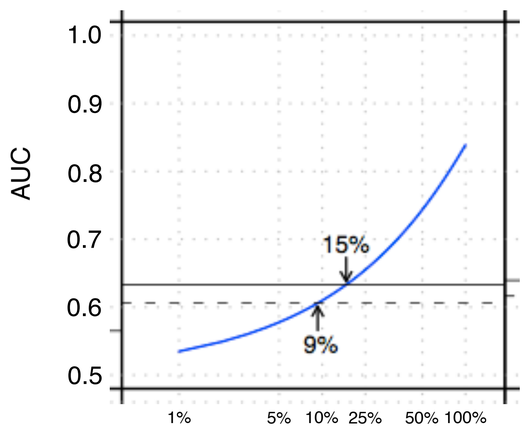


Figure 1. Risk prediction accuracy for family history and SNP-based models in a large 3-generation pedigree across a range of different disease **heritabilities** (h_L^2) and **prevalences** (K). Rows of the panels correspond to different disease heritabilities, and columns correspond to different disease prevalences.

Results

The following graph depicts the relationship between **predictive accuracy** as measured by AUC (vertical axis) and the **proportion of genetic variance explained by known SNP associations** (horizontal axis) for a disease of moderate heritability ($h_L^2 = 0.4$) and prevalence (K = 0.1) in a large 3-generation pedigree:



Proportion of genetic variance explained

<u>Understanding the graph</u>

1. The **solid blue line** represents the performance of SNP-based risk assessments.
2. The **solid black line** represents the predictive performance of an ideal family history-based risk assessment.
3. The **dotted black line** represents the predictive performance of a family history-based model that only distinguishes between 0, 1, or >1 first-degree relatives with disease.
4. The **intersection points labeled with percentages** show the proportion of genetic variance explained at which SNP-based models outperform family history.

<u>Observations (see Figure 1)</u>

1. **Family history is most discriminative for common conditions** (as the chance of having an affected relative is higher)**, whereas SNP-based models maintain high discriminative power for rarer conditions provided that enough of the genetic variance is explained.**
2. In most cases, the bulk of predictive accuracy of family history can be captured by a model taking into account only **first-degree relatives**.
3. For diseases with <1% prevalence, the crossover point occurs between 1-4% of the variance explained. **This is well within the detection limits of current GWAS, and in fact, a large fraction of the diseases studied to date have already crossed this line.**

Conclusion

<u>Limitations of our model</u>

- No highly penetrant mutations
- No age-of-onset information
- No non-additive effects
- Use of lifetime risk only
- Recall biases for family history in practice
- Difficulty of obtaining heritability estimates

<u>Take-home messages</u>

1. **The relative performance of family history and SNP-based models at predicting disease risk depends largely on the characteristics of the disease considered.**
2. **For diseases of low or moderate frequency (< 1% prevalence), current SNP-based risk assessments may be significantly more discriminative than family history.**