如何用 PrediXcan 建立受 SNP 调控的基因表达和性状的关系

从 GWAS 的结果中找到具有潜在功能性的基因一直的遗传学研究的重点。以往的经验告诉人们，离最显著 SNP 最近的基因的易感性最大，但越来越多的证据表明这种经验并不十分可靠。随着越来越多的 SNP 在非编码区被发现，并且通过远端或近端调控机制影响特定基因的表达，人们有理由相信那些**由 SNP 调控的基因表达改变是影响性状的一个重要机制。**因此，来自芝加哥大学的研究者们就开发了一个 gene-based 关联分析软件

——**PredicXcan**

一：PrediXcan 工作原理

作者认为基因表达水平受到三个因素的调控，其中主要的两个是遗传因素和疾病状态（图 1）。PrediXcan 的目的是建立起受遗传调控的基因表达与性状之间的关系。整个工作流程分为两步：（1）估算 SNP 调控的基因表达水平；（2）建立基因表达水平与性状之间的关联。第一步中，作者借助类似于机器学习的思想，利用 GTEx Project, GEUVADIS 和 DGN 数据库中基因型数据和基因表达数据做训练集，然后估算用户导入的基因型数据中缺失的表达数据。一旦得到表达数据，就可建立起基因表达与性状之间的关系。（图 2）

图 1 基因表达受到遗传，表型以及其他因素的调控

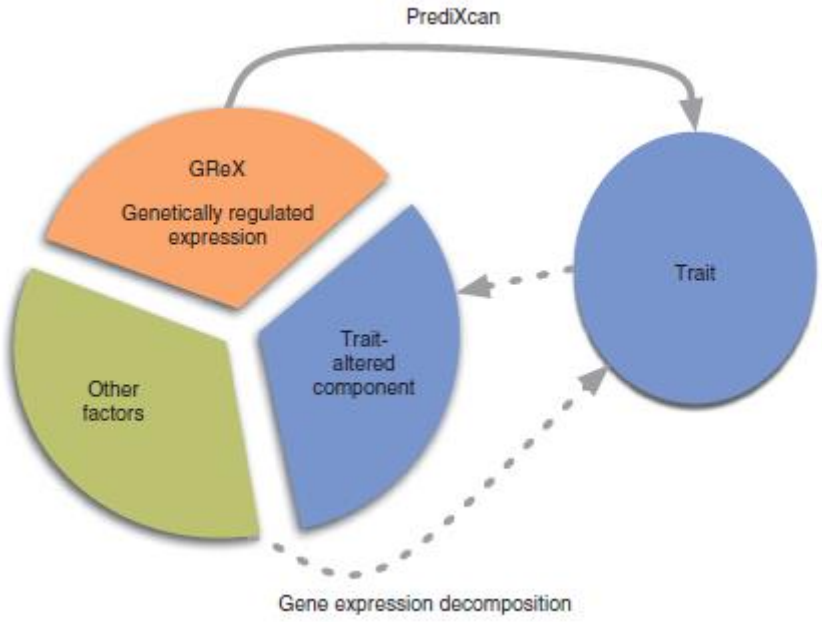
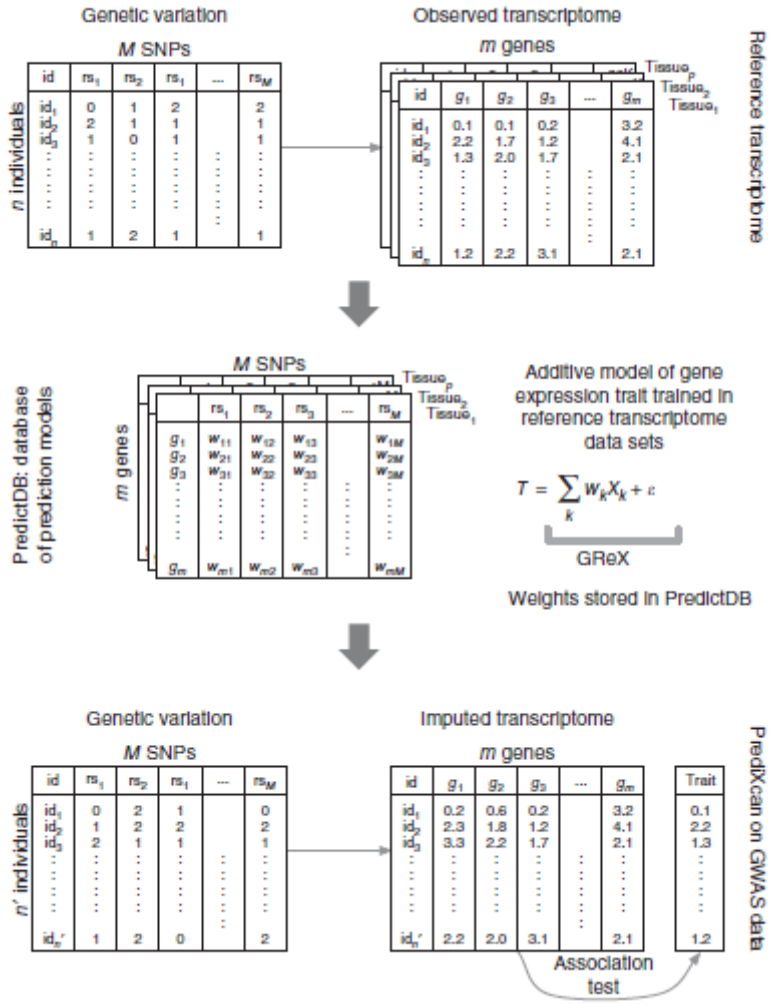


图 2 PrediXcan 工作流程



（二）如何使用 PrediXcan

2.1：文件准备

运行 PrediXcan 需要输入三个文件：**转录组预测模型**文件，**基因型**文件和**样本信息**文件。

下面一一介绍。

转录组预测模型文件：该文件不用自己制作，去 PredictDB 网站下载即可：

http://predictdb.org/。大家可以根据自己的需要选择不同的组织数据。

基因型文件：该文件每一行表示一个 SNP，包含的信息分别为：chromosome rsid position allele1 allele2 MAF，后面的每一列的内容是每一个样本在该 SNP allele2 的 dosage，最好是每一条染色体分开制作文件。

样本信息文件：直接将 PLINK 的 fam 文件导入即可。

2.2：基因表达预测

该步骤需要用到 PrediXcan 的“predict”功能，代码如下：

$./PrediXcan.py --predict --dosages genotype/ --dosages_prefix chr --samples samples.txt --weights model/DGN-HapMap-2015/DGN-WB_0.5.db --output_prefix results/DGN-HapMap

这一步中，我们在 PrediXcan.py 脚本存放的目录运行程序，假设我们的基因型文件的名称前缀是“chr”,样本信息文件的名称为“samples.txt”且存放在基因型文件同一目录下。该步骤会生成一个后缀为“predicted_expression.txt”的文件，存放估算的基因表达水平，可直接用于下一步。

2.3：基因表达与性状的关联分析

该步骤需要制作一个额外的表型文件，前两列分别是 FID 和 IID。从第三列起可以存放表型，数据类型可以是分类变量也可以是连续变量，如果是分类变量，0 表示 unaffected，1 表示 affected。默认缺失值是 NA。如果有多个表型列，可以用参数—mpheno 指定要分析的表型位于那一列，如—mpheno 1 则表示将文件中第三列作为要分析的表型。

代码如下：

$./PrediXcan.py --assoc --pheno My_pheno.txt --mpheno 1 --pred_exp

results/TW_Brain_Frontal_predicted_expression.txt --logistic --output_prefix

results/DGN-HapMap

最后奉上 PrediXcan 在 GitHub 上的下载地址

https://github.com/hakyimlab/PrediXcan。