2021.9.7-生信笔记-GWAS全基因组关联分析

全基因组关联分析/Genome-wide association study (GWAS)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/

定义

全基因组关联分析（Genome-wide association study）是指在人类全基因组范围内找出存在的序列变异，即单核苷酸多态性（single nucleotide polymorphisms, SNP），从中筛选出与疾病相关的SNPs。在全基因组范围内选择遗传变异进行基因分析，比较异常和对照组之间每个遗传变异及其频率的差异，统计分析每个变异与目标性状之间的关联性大小，选出最相关的遗传变异进行验证，并根据验证结果最终确认其与目标性状之间的相关性。

> 总结来说，GWAS是为了确定单核苷酸多态性(SNPs)与表型性状之间的关联。

关键词及其解释

single nucleotide polymorphism (SNP) 单核苷酸多态性

这是发生在基因组中一个特定位置的单个核苷酸（即A、C、G或T）的变异。一个SNP通常以两种不同的形式存在（例如，A与T）。这些不同的形式被称为等位基因。一个有两个等位基因的SNP有三种不同的基因型（例如，AA、AT和TT）。

SNP-heritability 单核苷酸多态性遗传力

这是一个性状变异表型的一部分，在分析中可以由所有的SNPs解释。

SNP-level missingness 单核苷酸多态性水平的丢失

这是样本中某个特定SNP信息缺失的个体数量。缺失程度高的SNP有可能导致偏差。

summary statistics 摘要统计学

这些是进行GWAS后得到的结果，包括染色体数目、SNP的位置、SNP(rs)-identifier、MAF、effect size (odds ratio/beta)、标准误差和P值等信息。GWAS的汇总统计资料通常可以免费获取或在研究者之间共享。

The Hardy–Weinberg (dis)equilibrium (HWE) law 哈迪-温伯格遗传不平衡定律

这涉及到等位基因和基因型频率之间的关系。它假定一个无限大的群体，没有选择、突变或迁移。该定律指出，基因型和等位基因频率在各代中是恒定的。违反HWE定律表明基因型频率与期望值有明显差异（例如，如果等位基因A的频率=0.20，等位基因T的频率=0.80；基因型AT的期望频率是 2*0.2*0.8=0.32），观察到的频率不应该有明显差异。

在GWAS中，一般认为偏离HWE是基因分型错误的结果。病例中的HWE阈值往往没有对照组那么严格，因为病例中违反HWE法则可能表明与疾病风险有真正的遗传关系。

clumping 分丛

在这一过程中，只有在每个LD block中最重要的SNPs（即具有最低的p值）才会被确定并选择，以供下一步分析。

这减少了剩余SNPs之间的相关性，同时保留了最有力的统计证据。

co-heritability 共同遗传力

一种用于衡量疾病之间遗传关系的度量单位。基于SNP的共遗传性是指疾病对（如精神分裂症和双相情感障碍）之间由SNP解释的协方差比例。

heterozygosity 杂合性

这是指携带一个特定SNP的两个不同的等位基因。一个个体的杂合率是杂合基因型的比例。一个个体内高水平的杂合率可能是样本质量低的表现，而低水平的杂合率可能是由于近亲繁殖。

individual-level missingness 个体水平的丢失

这是特定个体丢失的SNPs的数量。高水平的丢失可能意味着糟糕的DNA质量或是技术问题。

linkage disequilibium (LD) 连锁不平衡

这是对特定人群中同一染色体上不同位点的等位基因之间非随机关联的衡量。当SNP的等位基因的关联频率高于随机分配下的预期时，SNP就处于LD。LD涉及SNPs之间的相关模式。

minor allele frequency (MAF) 小等位基因频率

这是在特定位置上最不常出现的等位基因的频率。大多数研究在检测与低MAF的SNP的关联方面力量不足，因此排除了这些SNP。

population stratification 人群分层

这就是在一项研究中存在多个亚人群（如具有不同种族背景的个体）。由于等位基因频率在亚人群之间可能不同，人口分层可能导致假阳性关联和/或掩盖真正的关联。这方面的一个很好的例子是筷子基因，由于人口分层，一个SNP在用筷子吃饭的能力中占了近一半的变异。

pruning 剪枝

这是一种选择处于近似连锁平衡的标记子集的方法。在PLINK中，这种方法使用染色体特定窗口（区域）内SNP之间的LD强度，并根据用户指定的LD阈值，只选择近似不相关的SNP。与clumping相反，pruning不考虑SNP的p值。

relatedness 相关性

这表明一对个体的遗传关系有多强。传统的GWAS假设所有的受试者都没有关系（即没有一对个体的关系比二级亲属更密切）。如果没有适当的校正，纳入亲属可能会导致SNP效应大小的标准误差的估计出现偏差。

> 请注意，分析家族数据的具体工具已经被开发出来。

sex discrepancy 性别差异

这是分配的性别和根据基因型确定的性别之间的差异。差异可能是由于实验室中的样本混淆。注意，只有在评估了性染色体（X和Y）上的SNPs后才能进行这种测试。

Polygenic risk score (PRS)

PRS 将多个 SNPs 的效应大小组合成一个单一的聚合评分，可用于预测疾病风险。 PRS 是一个个人层面的评分，计算基于一个人所携带的风险变量的数量，加权的 SNP 效应大小来自一个独立的大规模发现 GWAS。

因此，评分是一个特定个体的特定性状的总遗传风险的指标，可用于临床预测或筛查（例如：乳腺癌）。

但是，PRS判断的准确性还不足以应用到临床。不过，通过其预测疾病状态的能力，PRS 为我们了解精神病特征的遗传结构做出了贡献。 它已经被进一步用来研究从一个特定表型的 GWAS 获得的遗传效应大小是否可以用来预测另一个表型的风险

GWAS的输入数据

1. 全基因组WGS。可以检测全部的编码区和非编码区。但成本高。

   > 1000genome计划

2. WES。可以检测得到所有编码区的SNP，成本较高。
3. DNA chip（即DNA芯片，或称为SNP array）。只包含常见的10^6个SNP。成本低。

> 总体来说，GWAS要求输入数据的sample size必须足够大。

GWAS的输出结果

曼哈顿图

QQ-plot

GWAS的下游分析

Polygenic risk score (PRS) analyses

meta-analysis

软件工具

PLINK

plink的主要功能：数据处理，质量控制的基本统计，群体分层分析，单位点的基本关联分析，家系数据的传递不平衡检验，多点连锁分析，单倍体关联分析，拷贝数变异分析，Meta分析等等。

【PLINK version 1.07】可以从 http://zzz.bwh.harvard.edu/plink/ 下载。

> 除了 PLINK，还有许多其他可用于分析 SNP 数据的好选择：如 Genabel和 SNPTEST。
>
> 另外，GWAS的一些分析流程，也可以用R语言实现。

PLINK读取文件格式说明

PLINK可以读取文本格式的文件或二进制文件，但因为读取大型文本格式文件可能比较耗时，所以建议读取二进制文件。

文本格式的PLINK数据由两个文件组成：

> 一个文件包含个体及其基因型（genotypes）的信息【ped格式】
>
> > *.ped
> >
> > | FID | IID | PID | MID | Sex | P | rs1 | rs2 | rs3 |
> > |---|---|---|---|---|---|---|---|---|
> > | 1 | 1 | 0 | 0 | 2 | 1 | CT | AG | AA |
> > | 2 | 2 | 0 | 0 | 1 | 0 | CC | AA | AC |
> > | 3 | 3 | 0 | 0 | 1 | 1 | CC | AA | AC |
>
> 一个文件装有遗传标记（genetic markers）的资料【map格式】。

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

而二进制格式的PLINK数据则由三个文件组成：

一个二进制文件包含个体标识符（individual identifiers (IDs) ）和基因型（genotypes）【bed格式】

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

例如，在躁郁症的研究中，bed格式的文件应包含患者和健康人群（对照组）的基因分型结果（genotyping results）

一个文本文件包含有个体信息（information on the individuals）【fam格式】

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

例如，在躁郁症的研究中，fam格式的文件应当包含与研究对象有关的数据，如研究对象与其他参与者的家庭关系，性别和临床诊断结果

一个文本文件装有遗传标记（genetic markers）【bim格式】

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

> 而bim格式的文件应当有关SNPs物理位置的信息。

如果要利用协变量（covariates）进行分析，则通常会需要第四个文件，这个文件包含每个个体的协变量的值。

Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

> GWAS中的协变量（covariates）
>
> 一般定义：在实验的设计中，协变量是一个独立变量（解释变量），不为实验者所操纵，但仍影响实验结果。
>
> 但是在GWAS模型中，因子和变量最后都变为了协变量。

标识符说明：

Legend			
FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

PLINK基本命令

PLINK 是一个命令行程序。通常，PLINK会从工作目录中文件夹中加载数据文件，并将结果文件保存在这个目录中。

> 注意：在提示符后，通过键入 PLINK 关键字来指示 PLINK 的使用。如果没有在标准目录中安装 PLINK，则必须在命令前输入安装 PLINK 的目录的路径，例如/usr/local/bin/PLINK。

PRSice

用于进行PRS分析。

R

fusion包

用于在R内进行GWAS流程。

流程

质量控制QC

在进行GWAS之前对基因型数据应用严格的质量控制（QC）程序，包括使用适当的方法来考虑种族异质性。

> 练习数据集：国际HapMap项目的一个具有二元结果测量的数据集（N = 207）
>
> > 只包括有北欧和西欧血统的犹他州居民（CEU）
>
> (http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-05_phaseIII/plink_format/; Gibbs et al., 2003)
>
> 由于HapMap数据的样本量相对较小，这些模拟中的遗传效应大小被设定为大于通常在复杂性状的遗传研究中观察到的值。值得注意的是，要检测复杂性状的遗传风险因素，需要更大的样本量（例如，至少在几千人的数量级，但可能甚至是几万或几十万）。带有模拟表型性状的HapMap数据可在https://github.com/MareesAT/GWA_tutorial/（1_QC_GWAS.zip）获得。

质量控制步骤及简表

这里提供了七个质量控制步骤的参考，但是QC中的阈值可能因研究的具体特点而不同。

这七个步骤包括根据以下情况过滤掉SNP和个体。

1. 个体和SNP缺失 individual and SNP missingness
2. 受试者的分配和遗传性别不一致（见性别差异）inconsistencies in assigned and genetic sex of subjects (see sex discrepancy)
3. 小等位基因频率（MAF）minor allele frequency (MAF)
4. 偏离Hardy-Weinberg平衡（HWE）deviations from Hardy–Weinberg equilibrium (HWE)
5. 杂合率 heterozygosity rate
6. 相关性 relatedness
7. 种族离群（见人口分层）ethnic outliers (see population stratification)

步骤	命令与功能	阈值设置与详解
	【--geno】Excludes SNPs that are missing in a large proportion of the	We recommend to first filter SNPs and

【step1】Missingness of SNPs and individuals 个体和SNP缺失	subjects. In this step, SNPs with low genotype calls are removed.排除在很大一部分受试者中缺失的SNPs。在这一步骤中，基因型调用量低的SNP被删除。 【--mind】Excludes individuals who have high rates of genotype missingness. In this step, individual with low genotype calls are removed.排除那些基因型缺失率高的个体。在这一步骤中，基因型调用率低的个体被删除。	individuals based on a relaxed threshold (0.2; >20%), as this will filter out SNPs and individuals with very high levels of missingness. Then a filter with a more stringent threshold can be applied (0.02).我们建议首先根据一个宽松的阈值（0.2；>20%）过滤SNP和个体，因为这将过滤掉具有非常高遗漏率的SNP和个体。然后，可以采用更严格的阈值（0.02）的过滤器。（注意，在进行个体筛选之前，应当先进行SNPs的筛选）
【step2】Sex discrepancy性别差异	【--check-sex】Checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates.检查数据集中记录的个体性别与基于X染色体杂合度/同源性率的性别之间的差异。	Can indicate sample mix-ups. If many subjects have this discrepancy, the data should be checked carefully. Males should have an X chromosome homozygosity estimate >0.8 and females should have a value <0.2.这一步骤可能表明样本混淆。如果许多受试者有这种差异，应仔细检查数据。男性的X染色体同源性估计值应大于0.8，女性的值应小于0.2。
【step3】Minor allele frequency (MAF)小等位基因频率	【--maf】Includes only SNPs above the set MAF threshold. 只包括超过设定的 MAF 阈值的 SNPs	低MAF的SNP很罕见，因此缺乏检测SNP-表型关联的能力。这些SNP也更容易出现基因分型错误。MAF阈值应取决于你的样本大小，较大的样本可以使用较低的MAF阈值。分别来说，对于大样本（N = 100000）与中样本（N = 10000），通常使用0.01和0.05作为MAF阈值。
【step4】Hardy–Weinberg equilibrium (HWE)哈迪-温伯格平衡	【--hwe】Excludes markers which deviate from Hardy–Weinberg equilibrium.不包括偏离Hardy-Weinberg平衡的标记。	是基因分型错误的常见指标，也可能表明进化的选择。对于二元性状，我们建议这样排除：HWE p值在病例中<1e-10，在对照中<1e-6。不太严格的病例阈值可以避免在选择下丢弃疾病相关的SNPs（见在线教程 https://github.com/MareesAT/GWA_tutorial/对于定量性状，我们建议HWE p值<1e-6。
【step5】Heterozygosity 杂合性	教程代码： https://github.com/MareesAT/GWA_tutorial/ 为了排除杂合率高或低的个体	偏差可能表明样本污染、近亲繁殖。我们建议删除那些偏离样本杂合率平均值±3SD的个体。
【step6】Relatedness相关性	【--genome】Calculates identity by descent (IBD) of all sample pairs.计算所有样本对的血统特征（IBD）。【--min】设置阈值，并创建一个亲缘关系超过所选阈值的个体列表。意味着可以检测到例如pi-hat>0.2（即二级亲属）的	【--genome】在这个分析中使用独立的SNPs（修剪），并且只限于常染色体的分析。 【--min】隐性关联性会干扰关联分析。如果你有一个基于家庭的样本（如父母-后代），你不需要删除相关的配对，但统计分析应该考虑到家庭关联性。然而，

	受试者的关系。	对于基于人群的样本，我们建议使用0.2的pi-hat阈值。
【step7】Population stratification人口分层	【--genome】计算所有样本对的血统特征（IBD）。【--cluster --mds-plot *k*】Produces a k-dimensional representation of any substructure in the data, based on IBS.基于IBS，产生数据中任何子结构的k维表示。	【--genome】计算所有样本对的血统特征（IBD）。【--cluster --mds-plot *k*】K是维度的数量，需要定义（通常是10）。这是质量控制的一个重要步骤，包括多个程序，但为了完整起见，我们在表中简要地提到了这一步骤。这个步骤将在 "控制人口分层 "一节中详细描述。

质量控制教程

https://github.com/MareesAT/GWA_tutorial/

它提供了数据质量控制和可视化的潜在偏见来源的脚本。这些脚本在人类基因组单体型图(HapMap)数据的 CEU 组上执行 QC，可以应用于其他数据集，但基于家庭的数据集和涉及多个不同族裔群体的数据集除外。

一般来说，如果一个样本包括多个族群(例如，非洲人、亚洲人和欧洲人)，建议分别对每个族群进行关联测试（tests of association in each of the ethnic groups separately），并使用合适的方法，例如荟萃分析meta-analysis，将结果结合起来。

step7：控制人口分层

GWAS 系统性偏差的一个重要来源是人口分层。研究表明，在一个单一种族人口中，即使是微妙程度的人口分层也可能存在。因此，检测和控制人口分层的存在是一个必不可少的 QC 步骤。

有多种方法可以用来检测人口分层。在本教程中，仅展示一个在PLINK中被合并的方法：【多维缩放the multidimensional scaling (MDS) approach】

多维缩放 MDS

这种方法计算整个基因组内任何一对个体之间共享的等位基因的平均比例，以产生每个个体的遗传变异的定量指数（成分）quantitative indices (components)。可以将个体成分得分绘制出来，以探索是否有一些个体群体在遗传上比预期的更相似。

> 例如，在一项包括来自亚洲和欧洲的受试者的遗传研究中，MDS分析将显示亚洲人在遗传上比欧洲人更相似。

为了调查生成的成分分数（generated component scores）在哪些个体上偏离了样本的目标人群，绘制被调查样本的分数和已知种族结构的人群（如HapMap/1KG数据）是有帮助的。这个步骤被称为锚定。这使研究者能够获得其样本的民族信息并确定可能的民族异常值。

> 在https://github.com/MareesAT/GWA_tutorial/（2_Population_stratification.zip）中提供了一个脚本，以1KG项目的数据为锚对自己的数据进行MDS（http://www.1000genomes.org/）。

注意，根据MDS分析，属于离群值的个体应从进一步的分析中剔除。排除这些个体后，必须进行新的MDS分析，其主要成分需要作为关联测试的协变量，以校正人群中任何剩余的人群分层。

> 需要包括多少成分取决于人群结构和样本大小，但精神病遗传学界普遍接受包括多达10个成分。

在完成 QC 和 MDS 之后，数据就可以进行后续的关联测试了。

关联测试 tests of association

常用的SNPs和感兴趣的表型特征之间的关联测试，同时控制潜在的混杂因素。

> 或者称之为【二元性状与数量性状相关性的统计学检验】STATISTICAL TESTS OF ASSOCIATION FOR BINARY AND QUANTITATIVE TRAITS

根据感兴趣的性状或疾病的预期遗传模型和所研究的表型性状的性质，可以选择适当的统计测试。

> 在附带的教程中，提供了适合二元性状（如酒精依赖患者与健康对照）或定量性状（如每周饮用的酒精饮料数量）的各种关联类型的脚本。https://github.com/MareesAT/GWA_tutorial/ (3_Association_GWAS.zip)
>
> PLINK提供一个自由度（1 df）的等位基因测试，其中性状值，或二元性状的对数，作为风险等位基因（小等位基因[a]与大等位基因[A]）数量的函数而线性增加或减少。此外，还有非加性试验，例如，基因型关联试验（2 df：aa vs. Aa vs AA），显性基因作用试验（1 df：[aa & Aa] vs AA），以及隐性基因作用试验，（1 df：aa vs [Aa & AA]）。然而，非加性检验non-additive tests没有得到广泛的应用，因为在实践中检测非加性的统计能力很低
>
> 更复杂的分析（如Cox回归分析和治愈模型）可以通过使用PLINK中基于R的 "插件 "功能进行。

二元结果测量 Binary outcome measure

> 在PLINK中，SNP和二元结果（值1=未受影响，值2=受影响；0和-9代表缺失；前面代表PLINK中的默认选项，可以改变）之间的关联可以用选项【--assoc】或【--logistic】来测试。
>
> PLINK中的【--assoc】选项执行的是X2的关联测试，不允许包含协变量。
>
> 使用【--logistic】选项，将进行Logistic回归分析，允许包含协变量。
>
> 【--logistic】选项比【--assoc】选项更灵活，但是它的代价是增加了计算时间（包含协变量的检测似乎都会增加计算时间）。

量化的结果衡量Quantitative outcome measure

> 在PLINK中，可以用选项【--assoc】和【--linear】来测试SNPs和定量结果指标之间的关联。
>
> 当PLINK检测到定量结果指标（即除1、2、0或缺失以外的值）时，【--assoc】选项会自动将其视为定量结果指标，执行通常的学生t检验的渐进版本来比较两个平均值。这个选项不允许使用协变量。
>
> PLINK中的【--linear】以每个单独的SNP作为预测因子进行线性回归分析。与【--logistic】选项类似，【--linear】选项可以使用协变量，而且比【--assoc】选项要慢一些。

多重检验校正 Correction for multiple testing

现代基因分型阵列可以同时对400万个标记进行基因分型，这就产生了大量的测试，从而产生了相当大的多重测试负担。SNP的归因可能会进一步增加测试的关联数量。各种模拟表明，对于欧洲人群的研究，广泛使用的全基因组意义阈值为5×10^{-8}，充分控制了整个基因组中独立SNP的数量，而不考虑研究的实际SNP密度。当测试非洲人群时，由于这些个体的遗传多样性更大，需要更严格的阈值（可能接近1.0×10^{-8}）。

> 广泛应用的多重检验校正方法有：
>
> Bonferroni correction
>
> > The Bonferroni correction用0.05/n 公式计算调整后的 p 值阈值，n 是测试的 SNPs 数量。
> >
> > 但是，由于连锁不平衡LD，许多 SNPs 是相关的，因此根据定义，这些SNPs不是独立的。因此，这种校正方法往往过于保守，导致假阴性结果的比例增加。
>
> Benjamini–Hochberg false discovery rate (FDR)
>
> > 假设SNP是独立的，FDR控制所有信号中假阳性的预期比例，其FDR值低于固定的阈值。
> >
> > 这比Bonferroni校正要保守得多。应该注意的是，控制FDR并不意味着任何统计学意义的概念；它只是一种将假阳性的预期比例最小化的方法，例如用于后续分析。此外，这种方法也有其局限性，因为SNP和P值不是独立的，而这是FDR方法的一个假设。
>
> permutation testing

因此，PLINK提供了一个选项【--adjust】，生成的输出中显示了未经调整的p值，以及用各种多重检验校正方法校正的p值。

最后，可以用互换法（permutation methods）来处理多重检验的负担。

> 为了计算基于互换的P值，结果测量标签被随机互换多次（如1,000-1,000,000），这就有效地消除了结果测量和基因型之间的任何真实关联。然后对所有被处理过的数据集进行统计测试。这就提供了测试统计量的经验分布和在无关联的零假设下的P值。
>
> 随后，将从观察数据中得到的原始检验统计量或P值与P值的经验分布进行比较，以确定经验调整的P值。
>
> 为了使用这种方法，可以将两个PLINK选项【--assoc】和【--mperm】结合起来，生成两个p值：
>
> > EMP1是经验P值（未校正），EMP2是为多重检验校正的经验P值。
> >
> > 这个过程是计算密集型的，尤其是在需要许多次排列组合的情况下，这对于准确计算非常小的p值是必要的。

PRS分析

PRS分析的目的不是识别单个SNP，而是将整个基因组的遗传风险聚集在一个感兴趣的性状的单个多基因分数中。

> 而GWAS一般侧重于Single variant association analysis，但需要非常大的样本量来检测许多复杂性状的少数SNPs。

在这种方法中，需要大量的 discovery sample 来可靠地确定每个SNP对一个特定性状的多基因评分（"权重"）的贡献程度。

不过，在一个独立的target sample中，其样本规模可以更小，多基因分数可以根据遗传DNA图谱和这些权重来计算（关于计算的细节见下文）。作为一个经验法则，2000个左右的target sample可以提供足够的力量来检测出相当比例的变异解释。此外，discovery sample和target sample应该有相同数量的受试者，直到target sample包括2000名受试者。如果有更多的样本，应该在discovery sample中包括更多的受试者，以最大限度地提高效应大小估计的准确性。

目前认为，PRS还不足以预测个人层面的疾病风险，但它已被成功用于显示性状内部和跨性状的重要关联。

计算PRS

【discovery GWAS实例】

Discovery GWAS

	Weight*	Risk Allele
SNP1	0.2	A
SNP2	-0.3	C
SNP3	0.1	G

Individual	Alleles SNP1	Alleles SNP2	Alleles SNP3
1	AT	AA	CG
2	AA	CA	GG
3	TT	AC	CG
4	TT	AA	GG
5	TA	CA	GC
6	AT	CA	CG
7	AA	AA	GG
8	AA	CC	CG
9	TA	CC	GC
10	AT	AA	CG

PRS:

Individual	SNP 1	SNP 2	SNP 3	**PRS**
1	0.2+0.0	0.0+0.0	0.0+0.1	**0.3**
2	0.2+0.2	-0.3+0.0	0.1+0.1	**0.3**
3	0.0+0.0	0.0-0.3	0.0+0.1	**-0.2**
4	0.0+0.0	0.0+0.0	0.1+0.1	**0.2**
5	0.0+0.2	-0.3+0.0	0.1+0.0	**0.0**
6	0.2+0.0	-0.3+0.0	0.0+0.1	**0.0**
7	0.2+0.2	0.0+0.0	+0.1+0.1	**0.6**
8	0.2+0.2	-0.3-0.3	0.0+0.1	**-0.1**
9	0.0+0.2	-0.3-0.3	0.1+0.0	**-0.3**
10	0.2+0.0	0.0+0.0	0.0+0.1	**0.3**

> 三个单核苷酸多态性(SNPs)聚集成单个个体多基因风险评分(PRS)的工作实例。
>
> *weight（权重）是beta（贝塔系数）或the log of the odds-ratio（比值比的对数），这取决于分析的是连续性特征还是二元性状。
>
> > beta用于衡量连续型特征；
> >
> > the log of teh odds-ratio用于衡量二元性状。

总的来说，为了进行PRS分析，从discovery GWAS中获得特异性状的权重（连续性性状的β值和二元性状的比值比对数）。

然后，在target sample中，根据每个人携带的风险等位基因的数量乘以特定性状的权重，用以计算每个人的PRS。

> 对于许多复杂的性状，其SNP效应大小是公开可用的。https://www.med.unc.edu/pgc/downloads

虽然原则上，所有常见SNPs都可用于PRS分析，但在计算风险分数（risk score）之前，通常会对GWAS的结果进行clumping。p值阈值通常用于去除显示出很少或没有统计学证据的SNPs（例如，只保留p值<0.5或<0.1的SNPs）。通常情况下，将进行多个PRS分析，P值的阈值也不同。

进行多基因风险预测分析 Conducting polygenic risk prediction analyses

一旦计算出目标样本中所有受试者的PRS，这些分数可用于（逻辑）回归分析，以预测任何预计与感兴趣的性状有遗传重叠的性状。

预测的准确性可以用回归分析的 $pseudo-R^2$ 指标来表示。重要的是在回归分析中至少包括几个MDS成分作为协变量，以控制种群分层的情况。

为了估计PRS能解释多少变化，将比较只包括协变量（如MDS成分）的模型的R^2和包括协变量+PRS的模型的R^2。由于PRS导致的R^2的增加表明遗传风险因素所解释的预测准确性的增加。

PRS的预测精度主要取决于所分析性状的（共）遗传性、SNP的数量和发现样本的大小。目标样本的大小只影响R^2的可靠性，如果感兴趣的性状的（共）遗传性和所使用的发现样本的样本量足够大，通常目标样本中的几千个受试者就足以达到显著的R^2。

> 关于为自己的PRS分析进行功率计算的R脚本，可以参考 https://sites.google.com/site/fdudbridge/software上的POLYGENE脚本。

用于进行PRS分析的程序是PRSice（http://prsice.info/），可以进行clumping、p值阈值、MDS成分等功能。

> 教程见：https://github.com/MareesAT/GWA_tutorial/（4_PRS.doc）

参考数据库

GTEx

> 提供了关于 SNPs 和基因表达之间关联的信息

FUMA

> 是功能注释常用的工具

GCTA

相关教程

教程合集

入门讲解