# Learning PRS

## PRSについて

https://onlinelibrary.wiley.com/doi/10.1002/mpr.1608 より引用

## 6 PRS ANALYSIS

### 6.1 Computing a PRS

Single variant association analysis has been the primary method in GWAS but requires very large sample sizes to detect more than a handful of SNPs for many complex traits (Gratten, Wray, Keller, & Visscher, 2014; Visscher, Brown, McCarthy, & Yang, 2012). In contrast, PRS analysis does not aim to identify individual SNPs but instead aggregates genetic risk across the genome in a single individual polygenic score for a trait of interest (Purcell et al., 2009; see Figure 4 for a simplified example). In this approach, a large discovery sample is required to reliably determine how much each SNP is expected to contribute to the polygenic score (“weights”) of a specific trait. Subsequently, in an independent target sample, which can be more modest in size (Dudbridge, 2013), polygenic scores can be calculated based on genetic DNA profiles and these weights (see below for details on the calculations). As a rule of thumb, a target sample around 2,000 subjects provides sufficient power to detect a significant proportion of variance explained. Furthermore, the discovery and target samples should have the same number of subjects until the target sample includes 2,000 subjects. If more samples are available, additional subjects should be included in the discovery sample to maximize the accuracy of the estimation of the effect sizes (Dudbridge, 2013). Although PRS is not powerful enough to predict disease risk on the individual level (Wray et al., 2013), it has been successfully used to show significant associations both within and across traits. For example, a PRS analysis of schizophrenia showed for the first time that an aggregate measure of the genetic risk to develop schizophrenia, estimated based on the effects of common SNPs (from the discovery sample) that showed nominally significant associations with disease risk, was significantly associated with schizophrenia risk in an independent (target) sample. The significant association was found despite the fact that the available sample sizes were too small to detect genome-wide significant SNPs (Purcell et al., 2009). In addition, GWAS for schizophrenia (the discovery sample) has been used to significantly predict the risk in target samples with various phenotypes, such as bipolar disorder, level of creativity, and even risk of immune disorders (Power et al., 2015; Purcell et al., 2009; Stringer et al., 2014; Wray et al., 2013).

To conduct PRS analysis, trait-specific weights (beta's for continuous traits and the log of the odds ratios for binary traits) are obtained from a discovery GWAS. In the target sample, a PRS is calculated for each individual based on the weighted sum of the number of risk alleles that he or she carries multiplied by the trait-specific weights. For many complex traits, SNP effect sizes are publicly available (e.g., see https://www.med.unc.edu/pgc/downloads).

Although in principle all common SNPs could be used in a PRS analysis, it is customary to first clump (see clumping) the GWAS results before computing risk scores. p value thresholds are typically used to remove SNPs that show little or no statistical evidence for association (e.g., only keep SNPs with p values <0.5 or <0.1. Usually, multiple PRS analyses will be performed, with varying thresholds for the p values.

### 6.2 Conducting polygenic risk prediction analyses

Once PRS have been calculated for all subjects in the target sample, the scores can be used in a (logistic) regression analysis to predict any trait that is expected to show genetic overlap with the trait of interest. The prediction accuracy can be expressed with the (pseudo-)R2 measure of the regression analysis. It is important to include at least a few MDS components as covariates in the regression analysis to control for population stratification. To estimate how much variation is explained by the PRS, the R2 of a model that includes only the covariates (e.g., MDS components) and the R2 of a model that includes covariates + PRS will be compared. The increase in R2 due to the PRS indicates the increase in prediction accuracy explained by genetic risk factors.

The prediction accuracy of PRS depends mostly on the (co-)heritability of the analysed traits, the number of SNPs, and the size of the discovery sample. The size of the target sample only affects the reliability of R2 and typically a few thousand of subjects in the target sample are sufficient to achieve a significant R2 if the (co-)heritability of the trait(s) of interest and the sample size of the discovery sample used are sufficiently large. For an R script to perform power calculations for your own PRS analysis, we refer to the POLYGENE script on https://sites.google.com/site/fdudbridge/software (Dudbridge, 2013).

A convenient program to perform PRS analysis is PRSice (see http://prsice.info; Euesden, Lewis, & O'Reilly, 2015). It takes care of clumping, p value thresholds, MDS components, and plots attractive graphs. We refer to https://github.com/MareesAT/GWA_tutorial/ (4_PRS.doc) for a tutorial on how to perform your own PRS analysis using PRSice. Other programs for the application of PRS are, for example, PLINK (--score) and LDpred (Purcell et al., 2007; Vilhjalmsson et al., 2015).

*以下にChatGPTによる日本語訳を示す*

# 6 PRS（ポリジェニックリスクスコア）解析

## 6.1 PRSの計算

単一変異関連解析はGWAS（ゲノムワイド関連解析）での主要な方法であり、多くの複雑な特性について多くのSNPを検出するには非常に大規模なサンプルサイズが必要です（Gratten, Wray, Keller, & Visscher, 2014; Visscher, Brown, McCarthy, & Yang, 2012）。対照的に、PRS解析は個々のSNPを特定することを目指さず、代わりに遺伝的リスクを特定の興味のある特性の一つの個体のポリジェニックスコアに集約します（Purcell et al., 2009; 図4を参照してください、簡略化された例を示します）。このアプローチでは、各SNPが特定の特性のポリジェニックスコアにどれだけ寄与するかを確実に決定するために大規模な発見サンプルが必要です（「ウェイト」と呼ばれます）。その後、より控えめなサイズであることができる独立したターゲットサンプル（Dudbridge, 2013）で、遺伝的DNAプロファイルとこれらのウェイトに基づいてポリジェニックスコアが計算されます（計算の詳細については以下を参照）。一般的なルールとして、2,000人前後のターゲットサンプルは、説明される分散のかなりの部分を検出するために十分なパワーを提供します。さらに、発見サンプルとターゲットサンプルは、ターゲットサンプルが2,000人を含むまで同じ数の被験者を持つべきです。利用可能なサンプルがもっとある場合、効果サイズの推定の精度を最大限に高めるために、発見サンプルに追加の被験者を含めるべきです（Dudbridge, 2013）。PRSは個人レベルで疾患リスクを予測するには強力ではありませんが（Wray et al., 2013）、特性内および特性間の有意な関連性を示すために成功を収めています。たとえば、統合失調症のPRS解析は、共通のSNPの効果に基づいて推定された統合失調症の遺伝的リスクの集約尺度が、独立した（ターゲット）サンプルで統合失調症リスクと有意に関連していることを初めて示しました。有意な関連性は、ゲノム全体で有意なSNPを検出するには利用可能なサンプルサイズが小さすぎたにもかかわらず見つかりました（Purcell et al., 2009）。さらに、統合失調症のためのGWAS（発見サンプル）は、双極性障害、創造性のレベル、さらには免疫障害のリスクなど、さまざまな特性を持つターゲットサンプルでリスクを有意に予測するのに使用されています（Power et al., 2015; Purcell et al., 2009; Stringer et al., 2014; Wray et al., 2013）。

PRS解析を実施するために、特性固有のウェイト（連続特性のベータおよび2値特性のオッズ比の対数）は発見GWASから取得されます。ターゲットサンプルでは、各個体に対して、運び物のリスクアレルの数に特性固有のウェイトを掛けた重みつき合計に基づいてPRSが計算されます。多くの複雑な特性について、SNPの効果サイズは一般に公に利用可能です（例：https://www.med.unc.edu/pgc/downloads を参照）。

原則として、PRS解析にはすべての一般的なSNPを使用できますが、通常は計算前にGWASの結果をクランプ（クランプを参照）することが一般的です。p値のしきい値は、関連性のほとんどないまたはまったくないSNPを削除するために通常使用されます（例：p値<0.5または<0.1のSNPのみを保持）。通常、p値のさまざまなしきい値を使用して複数のPRS解析が実行されます。

## 6.2 ポリジェニックリスク予測解析の実施

ターゲットサンプルのすべての被験者に対してPRSが計算されたら、スコアは（ロジスティック）回帰分析で予測対象の特性を予測するために使用できます。予測の精度は、回帰分析の（疑似）R2尺度で表現できます。人口分化をコントロールするために回帰分析には少なくともいくつかのMDSコンポーネントを共変量として含めることが重要です。PRSによって説明される変動の量を推定するには、共変量のみを含むモデル（MDSコンポーネントなど）のR2と、共変量+PRSを含むモデルのR2を比較します。PRSによるR2の増加は、遺伝的リスク要因による予測精度の増加を示します。

PRSの予測精度は、主に分析される特性の（共）遺伝度、SNPの数、および発見サンプルのサイズに依存します。ターゲットサンプルのサイズはR2の信頼性にのみ影響を与え、興味のある特性の（共）遺伝度と使用される発見サンプルのサイズが十分に大きい場合、ターゲットサンプル内の数千人の被験者は通常、有意なR2を達成するのに十分です。独自のPRS解析のパワー計算を実行するためのRスクリプトについては、https://sites.google.com/site/fdudbridge/software の「POLYGENEスクリプト」を参照してください（Dudbridge, 2013）。

PRS解析を実施するための便利なプログラムはPRSiceです（http://prsice.info を参照、Euesden, Lewis, & O'Reilly, 2015）。クランプ、p値のしきい値、MDSコンポーネント、魅力的なグラフのプロットを処理します。PRSiceを使用して独自のPRS解析を実行する方法については、https://github.com/MareesAT/GWA_tutorial/(4_PRS.doc)を参照してください。PRSの適用に関する他のプログラムには、PLINK（--score）およびLDpred（Purcell et al., 2007; Vilhjalmsson et al., 2015）などがあります。