XPASS

The XPASS package implement the XPASS approach for generating PRS in a target population by integrating multi-ethnic datasets.

Installation

```
#install.packages("devtools")
devtools::install_github("YangLabHKUST/XPASS")
```

Quick start

We illustrate the usage of XPASS using the GWAS summary statistics of BMI from UKB and BBJ. For demonstration, we use the easily accessible 1000 Genomes project genotypes as reference panels. **However, because these reference panels only contain 377 EAS amples and 417 EUR samples, optimal prediction accuracy of XPASS is not expected. In practice, it is suggested to use larger datasets as reference panels (n>2000). Therefore, we strongly suggest users to use their own reference panels with sufficiently large sample sizes.**

Data preparation

The datasets involved in the following example can be downloaded from here.

Input files of XPASS includ:

- summay statistics file of the target population
- summay statistics file of the auxiliary population
- reference panel of the target population in plink 1 format
- reference panel of the auxiliary population in plink 1 format
- covariates file associated with the target population reference panel (optional)
- covariates file associated with the auxiliary population reference panel (optional)

The XPASS format GWAS summary statistics file has 5 fields:

- SNP: SNP rsid
- N: sample size
- Z: Z-scores
- A1: effect allele
- A2: other allele.

Here, we use the BMI GWAS from BBJ male as the target training set and BMI GWAS from UKB as the auxiliary training set

```
$ head BMI_bbj_male_3M_format_3MImp.txt

SNP N Z A1  A2
rs117086422 159095  -1.20413423957762 T C
rs28612348  159095  -1.25827042089233 T C
rs4475691 159095  -1.19842287777303 T C
rs950122  159095  -1.2014434974188  C G
rs3905286 159095  -1.27046106136441 T C
rs28407778  159095  -1.26746342605063 A G
rs4246505 159095  -1.24706211285128 A G
rs4626817 159095  -1.26366297625074 A G
rs11507767  159095  -1.28611566069053 G A
```

```
$ head height_ukb_3M_format.txt

SNP N Z A1  A2
rs117086422 429312  1.42436004338939  T C
rs28612348  429312  1.48706291417224  T C
rs4475691 429312  1.53977372135067  T C
rs950122  429312  1.37958155329171  C G
rs3905286 429312  1.77045946243262  T C
rs28407778  429312  1.9908370573435 A G
rs4246505 429312  1.90922505355565  A G
rs4626817 429312  1.53216668392479  A G
rs11507767  429312  1.55873328059033  G A
```

We keep the BMI GWAS from BBJ female as the external validation dataset:

```
$ head BMI_bbj_male_3M_format_3MImp.txt

SNP N Z A1  A2
rs117086422 72390 0.897252679561414 T C
rs28612348  72390 0.904738461538462 T C
rs4475691 72390 0.89177374486687  T C
rs950122  72390 0.891523178807947 C G
rs3905286 72390 0.827441738675046 T C
rs28407778  72390 0.816801884323476 A G
rs4246505 72390 0.812148186935463 A G
rs4626817 72390 0.811624558188245 A G
rs11507767  72390 0.811100929441026 G A
```

The covariates files should not include row names and column names. The rows should be exactly corresponding to the individuals in the .fam file of reference genotypes. Each column corresponds to one covariate. A column of one should not be included in the file.

```
$ head 1000G.EAS.QC.hm3.ind.pc5.txt

-0.0242863 0.0206888 -0.0028171 0.0343263 0.0211044
-0.025051 0.0275325 -0.0332156 -0.0233166 0.0588989
-0.0198603 0.0286747 0.008464 -0.0215478 0.0189802
-0.0117008 0.0251493 -0.0228276 -0.0670019 0.0186454
-0.0246233 0.0281415 -0.0418795 0.0059947 0.0196607
-0.0270411 0.0127586 -0.0376545 0.0182809 0.0344744
-0.0193744 0.0367263 -0.0176168 -0.0307868 0.0254125
-0.0212299 0.022658 0.0225698 -0.0249273 0.0144324
-0.0159054 0.00558952 -0.00609582 -0.033497 0.0518336
-0.0258843 0.0476758 0.0073353 0.0164056 0.0072118
```

```
$ head 1000G.EUR.QC.hm3.ind.pc20.txt

0.0290072 0.0717627 -0.0314029 0.0317316 0.0618357 0.0385132 0.122857
-0.0289581 -0.0114267 -0.0205926 -0.0466983 0.0836711 0.00690379 0.0345008
-0.0179313 0.0109661 -0.0214763 0.0014544 0.0182944 0.0399625
0.0378568 0.0499758 -0.00811504 0.0363021 -0.0579984 0.0422509 0.141279
-0.0167868 -0.0181999 0.0165593 -0.0304088 0.0423324 0.0226001 0.00843853
0.0212477 -0.0666462 -0.0787379 0.0136196 0.108933 0.0801246
0.0417318 0.0732938 -0.0409508 0.0176872 0.0801957 0.0124742 0.0252689
-0.0444921 -0.00305238 0.0035535 -0.0070929 0.0240194 -0.00543616 0.0272464
-0.0048309 -0.0207223 0.0415044 0.0494025 0.0213837 -0.0021093
0.0348071 0.0715395 -0.0266058 0.00280025 -0.0166164 0.0440144 0.135709
0.017364 0.0276564 -0.00286321 0.0314583 0.00299185 0.0792055 0.012042
-0.0337269 -0.00999033 0.0186435 -0.0699027 0.00791191 -0.0131168
0.0444763 0.0713348 0.00162451 -0.0107805 0.0868909 0.0218014 0.0314216
-0.0429928 -0.0137937 0.00913544 -0.062828 0.0555199 0.0378234 -0.0162297
0.000344947 -0.0164497 0.0523967 -0.0861731 -0.038893 0.028166
0.0300713 0.0522082 -0.0116997 -0.029994 0.0539977 0.0162067 -0.0522507
0.0022036 0.0255471 0.0129012 0.0371803 0.0701241 0.0314957 0.00870374
0.00566795 0.116672 0.0267188 0.0451948 0.00288655 -0.0427304
0.0339424 0.0718272 0.000614329 -0.0183555 0.0162768 -0.0600599 0.00343218
0.0149793 -0.0236301 -0.0267658 0.0387814 -0.00624387 -0.0364751 0.00486515
-0.0341221 0.0415286 -0.0274807 -0.013188 0.0695243 0.0495376
0.0423378 0.0656126 -0.0331807 -0.0361484 -0.0155739 0.0557459 0.00428138
0.0840953 -0.034451 0.0753096 -0.0180153 0.0595412 -0.0367107 -0.0285888
-0.0986386 0.00845412 -0.00388558 -0.0641134 -0.05815 0.0242433
0.0429698 0.0431213 -0.0357636 -0.00270477 -0.0567958 0.0892002 0.0980711
0.0468321 -0.0359592 0.011195 -0.000235849 -0.0192522 -0.00271491 0.0381155
-0.00845755 -0.0171629 -0.026532 -0.0415778 0.0274635 0.122063
0.0290721 0.0541744 -0.0238317 0.0254426 0.0986334 0.0706142 0.0977585
0.00427919 -0.0381976 0.0020029 -0.0161052 -0.016666 -0.00627125 -0.00490556
-0.0410802 0.0125096 -0.0175252 0.0320359 0.00866061 0.0736791
```

Run XPASS

Once the imput files are formatted, XPASS will automatically process the datasets, including SNPs overlapping and allele matching.
Run XPASS with the following comand:

```
# library(devtools)
# install_github("https://github.com/YangLabHKUST/XPASS")
library(XPASS)
library(data.table)
library(RhpcBLASctl)
blas_set_num_threads(30)

# reference genotypes for EAS (prefix of plink file bim/bed/fam)
ref_EAS <- "1000G.EAS.QC.hm3.ind"

# covariates of EAS reference genotypes
cov_EAS <- "1000G.EAS.QC.hm3.ind.pc5.txt"

# reference genotypes for EUR (prefix of plink file bim/bed/fam)
ref_EUR <- "1000G.EUR.QC.hm3.ind"

# covariates of EUR reference genotypes
cov_EUR <- "1000G.EUR.QC.hm3.ind.pc20.txt"

# genotype file of test data (plink prefix).
# Note: for demonstration, we assume that the genotypes of prediction target
are used as the reference panel of target population.
# In practice, one can also use genotypes from other sources as reference
panel.
BMI_test <- "1000G.EAS.QC.hm3.ind"


# sumstats of height
BMI_bbj_male <- "BMI_bbj_male_3M_format_3MImp.txt"  # target
BMI_ukb <- "BMI_ukb_sumstat_format_all.txt"  # auxiliary

BMI_bbj_female <- "BMI_bbj_female_3M_format_3MImp.txt"  # external validation

fit_bbj <-XPASS(file_z1 = BMI_bbj_male,file_z2 = BMI_ukb,file_ref1 = ref_EAS,
                file_ref2 = ref_EUR,
                file_cov1 = cov_EAS,file_cov2 = cov_EUR,
                file_predGeno = BMI_test,
                compPRS=T,
                pop = "EAS",sd_method="LD_block",compPosMean = T,
                file_out = "BMI_bbj_ukb_ref_TGP")

Summary statistics file 1: BMI_bbj_male_3M_format_3MImp.txt
```

```
Summary statistics file 2: BMI_ukb_sumstat_format_all.txt
Reference file 1: 1000G.EAS.QC.hm3.ind
Reference file 2: 1000G.EUR.QC.hm3.ind
Covariates file 1: 1000G.EAS.QC.hm3.ind.pc5.txt
Covariates file 2: 1000G.EUR.QC.hm3.ind.pc20.txt
Reading data from summary statisitcs...
3506148 and 3777871 SNPs found in summary statistics files 1 and 2.
Reading SNP info from reference panels...
1209411 and 1313833 SNPs found in reference panel 1 and 2.
746454 SNPs are matched in all files.
0 SNPs are removed because of ambiguity; 746454 SNPs remained.
Calculating kinship matrix from the both reference panels...
127749 SNPs in the second reference panel are alligned for alleles according to
the first.
14337 SNPs have different minor alleles in population 1, z-scores are corrected
according to reference panel.
14332 SNPs have different minor alleles in population 2, z-scores are corrected
according to reference panel.
Assigning SNPs to LD Blocks...
Calculate PVE...
              h1          h2          h12         rho
[1,] 0.165790395 0.247603545 0.129399058 0.63866483
[2,] 0.007631346 0.008542654 0.006856057 0.02002658
...
Predicting PRS from test genotypes...
Done.
```

XPASS output

XPASS returns a list of results, the some key are:

- H: a table of estimated heritabilities, co-heritability and genetic correlation (first row) and their corresponding standard erros (second row).

```
> fit_bbj$H
              h1          h2          h12         rho
[1,] 0.165790395 0.247603545 0.129399058 0.63866483
[2,] 0.007631346 0.008542654 0.006856057 0.02002658
```

- mu: a data frame storing the posterior means computed by LDpred-inf using only the target dataset (mu1) and only the auxiliary dataset (mu2), and the posterior mean computed by XPASS (mu_XPASS). SNPs information is also returned: A1 is the effect allele, A2 is the other allele.

```
> head(fit_bbj$mu)
  CHR       SNP    POS A1 A2           mu1           mu2       mu_XPASS
1   1 rs4475691 846808  T  C -1.271443e-04 -0.0001837393 -0.0003248579
2   1 rs7537756 854250  G  A -4.778206e-05 -0.0002169705 -0.0002979870
3   1 rs3748592 880238  A  G -3.201406e-04 -0.0008911591 -0.0007477507
4   1 rs2340582 882803  A  G -3.396992e-04 -0.0009056487 -0.0008076512
5   1 rs4246503 884815  A  G -3.318260e-04 -0.0009265197 -0.0008148705
6   1 rs3748597 888659  T  C -3.327131e-04 -0.0009142760 -0.0008113772
```

- PRS (if file_predGeno provided and compPRS=T): a data frame storing the PRS generated using mu1, mu2 and mu_XPASS, respectively.

```
> head(fit_bbj$PRS)
      FID     IID         PRS1       PRS2  PRS_XPASS
1 HG00403 HG00403  0.14310343 0.92040953 0.55960909
2 HG00404 HG00404 -0.19478307 0.63289820 0.03773899
3 HG00406 HG00406 -0.09440555 0.69099450 0.14173516
4 HG00407 HG00407  0.05533466 0.07036618 0.05091082
5 HG00409 HG00409  0.25063980 1.54556843 0.83527097
6 HG00410 HG00410  0.13454734 0.23143305 0.10214319

# One can also compute PRS after fitting the model:
> PRS <- predict_XPASS(fit_bbj$mu,ref_EAS)
> head(PRS)
      FID     IID         PRS1       PRS2  PRS_XPASS
1 HG00403 HG00403  0.14310343 0.92040953 0.55960909
2 HG00404 HG00404 -0.19478307 0.63289820 0.03773899
3 HG00406 HG00406 -0.09440555 0.69099450 0.14173516
4 HG00407 HG00407  0.05533466 0.07036618 0.05091082
5 HG00409 HG00409  0.25063980 1.54556843 0.83527097
6 HG00410 HG00410  0.13454734 0.23143305 0.10214319
```

XPASS will also write above outputs into the files with `file_out` prefix, if provided.

External validation using independent GWAS data

Because the WeGene data is currently not available online, we use the GWAS of female BMI from BBJ as the external validation dataset to approximate the prediction R^2. Specifically we use the following equation:

$$R^2 = corr(y, \hat{y})^2 = \left(\frac{cov(y,\hat{y})}{\sqrt{var(y)var(\hat{y})}} \right)^2 = \left(\frac{z^T \tilde{\mu}/\sqrt{n}}{\sqrt{\tilde{\mu}^T \Sigma \tilde{\mu}}} \right)^2,$$

where z is the z-score of external summsry statistics, n is its sample size, $\tilde{\mu}$ is the posterior mean of effect size at the standardized genotype scale, Σ is the LD reference panel.

```
> R2 <- evalR2_XPASS(fit_bbj$mu,BMI_bbj_female,ref_EAS)
> R2
      PRS1       PRS2  PRS_XPASS
0.02235596 0.01509865 0.02905665
```

While the reference panels have only limmited samples, XPASS still achieves 30% relative improvement compared to LDpred-inf in terms of R^2.

Prediction of Type 2 Diabetes

We provide an additional example with common disease Type 2 Diabetes (T2D). We use the GWAS of males from BBJ as traing set and that of females as testing data, UKB GWAS as auxiliary dataset, and 1000 Genomes as reference panels. Datasets are available here.

```
# reference genotypes for EAS (prefix of plink file bim/bed/fam)
ref_EAS <- "1000G.EAS.QC.hm3.ind"

# covariates of EAS reference genotypes
cov_EAS <- "1000G.EAS.QC.hm3.ind.pc5.txt"

# reference genotypes for EUR (prefix of plink file bim/bed/fam)
ref_EUR <- "1000G.EUR.QC.hm3.ind"

# covariates of EUR reference genotypes
cov_EUR <- "1000G.EUR.QC.hm3.ind.pc20.txt"

# genotype file of test data (plink prefix).
T2D_test <- "1000G.EAS.QC.hm3.ind"


# sumstats of height
T2D_bbj_male <- "BMI_bbj_male_3M_format_3MImp.txt"  # target
T2D_ukb <- "BMI_ukb_sumstat_format_all.txt"  # auxiliary

T2D_bbj_female <- "BMI_bbj_female_3M_format_3MImp.txt"  # external validation

fit_bbj <-XPASS(file_z1 = T2D_bbj_male,file_z2 = T2D_ukb,file_ref1 = ref_EAS,
                file_ref2 = ref_EUR,
                file_cov1 = cov_EAS,file_cov2 = cov_EUR,
                file_predGeno = T2D_test,
                compPRS=T,
                pop = "EAS",sd_method="LD_block",compPosMean = T,
                file_out = "T2D_bbj_ukb_ref_TGP")

Summary statistics file 1: T2D_bbj_male_3M_format.txt
Summary statistics file 2: T2D_UKB_summary_format.txt
```

```
Reference file 1: 1000G.EAS.QC.hm3.ind
Reference file 2: 1000G.EUR.QC.hm3.ind
Covariates file 1: 1000G.EAS.QC.hm3.ind.pc5.txt
Covariates file 2: 1000G.EUR.QC.hm3.ind.pc20.txt
Reading data from summary statisitcs...
3701030 and 11971734 SNPs found in summary statistics files 1 and 2.
Reading SNP info from reference panels...
1209411 and 1313833 SNPs found in reference panel 1 and 2.
754466 SNPs are matched in all files.
0 SNPs are removed because of ambiguity; 754466 SNPs remained.
Calculating kinship matrix from the both reference panels...
128611 SNPs in the second reference panel are alligned for alleles according to
the first.
233891 SNPs have different minor alleles in population 1, z-scores are
corrected according to reference panel.
460962 SNPs have different minor alleles in population 2, z-scores are
corrected according to reference panel.
Assigning SNPs to LD Blocks...
Calculate PVE...
              h1         h2         h12        rho
[1,] 0.096949968 0.04137762 0.041067706 0.64840131
[2,] 0.006767725 0.00166678 0.002257857 0.02400119
...
Predicting PRS from test genotypes...
Done.
```

We compute the prediction R^2 as

```
> R2 <- evalR2_XPASS(fit_bbj$mu,T2D_bbj_female,ref_EAS)
> R2
       PRS1        PRS2   PRS_XPASS
0.033219004 0.008530789 0.041027152
```

For T2D, XPASS achieves 23.5% relative improvement interms of prediction R^2.

Development

The XPASS package is developed by Mingxuan Cai (mcaiad@ust.hk).

Contact information

Please contact Mingxuan Cai (mcaiad@ust.hk) or Prof. Can Yang (macyang@ust.hk) if any enquiry.