**Generating individual risk scores using LDpred**

**Required input-**

Plink binary files (.bed .bim .fam) containing the genotypes of samples to score

**Input File Formats-**

For the .fam files, sex cannot be left blank in column 5. A phenotype is also required in column 6

For the .bim files, rsids are required (not chr:pos) in col2 and dbgap allele encodings (A,C,T,G) in cols 5,6

**SNP content**- Use of imputed datasets is encouraged.

Scores have been generated using the hapmap3 SNP lists to maximize compatibility with the various genotyping platforms in use across the consortium. If there is insufficient overlap, ldpred will return a warning and not generate the scores.

**Racial groupings-**

We have generated LDpred weights using LD patterns from the following populations

AFR- African

AMR- Hispanic

ASN- East Asian (Chinese / Japanese)

EUR- Northern European

TRANS ETHNIC – All of the above

**We request you run the specific population that is closest to your dataset and the Trans ethnic score. Therefore, if your cohort is Chinese please run the ASN and TRANS ETHNIC**

**Running LDpred-**

Download the weight files from the dropbox, there are 8 per population

In this example I’m running the European weights against my Illumina 610 data with the prefix Human610-Quadv1HRC.rsid.bmi – substitute this with your plink file names

ldpred score --gf Human610-Quadv1HRC.rsid.bmi --rf EUR-Locke-BMI-HH3.weight --out EUR-LOK-BMI-score --pf Human610-Quadv1HRC.rsid.bmi.fam --pf-format FAM --rf-format LDPRED --summary-file predictions-eur-bmi

By not specifying a fraction of SNPs, scores will be calculated for all. LDpred returns a Pearson’s R2 for the different fractions – note however that for continuous traits these calculations are not necessarily accurate and we suggest carrying over all fractions to the steps below.

**Calculating the AUC**

I use the R library pROC

**For binary phenotypes** we just run the PRS against the phenotype

testroc<- roc(Yourdatafilename$phenotype, Yourdatafilename $PRS)

plot(testroc, print.auc=TRUE)

For the fraction with the highest AUC, report the regression coefficient:

coef(glm( as.factor(scale(dichoPheno))~PRS,family=binomial(link='logit'),data=DATA))

**For continuous traits** code the top 1% and 5% of the phenotype distribution as cases and the remainder as controls (I use the percentile(range, 0.99) function in excel)

Return the logs odds for the logistic regression of the PRS from the fraction with the highest AUC against the dichotomized phenotype.

coef(glm( as.factor(scale(dichoPheno))~PRS,family=binomial(link='logit'),data=DATA))

Also return the beta of the linear regression model of the continuous trait against the PRS from the fraction with the highest AUC

Requires R package lm.beta

library(lm.beta)

coef(lm.beta(lm(PHENO~PRS, data =DATA)))

Return the following data to John Connolly at CHOP:

1) Brief demographic description of cohort, number of cases, race, pediatric / adult, source of phenotype (self report, EMR, physician diagnosed etc)

2) The weight files that were used to generate the scores and the best fraction.

3) Derivative data, plots, AUCs and regression coefficients



library(standardize)