**Title**: A randomized algorithm with controlled privacy leaking for identifying overlapping relative across cohorts

**Motivation**: More and more individuals are sequenced and deposited into various datasets. Due to ethical restraint, it is not allowed to directly access individual information, including genotypic information, across datasets, but in practice, many studies requires knowledge of how samples, or their relative, are included in another datasets.

**Results**: using randomized algorithm, we presented an cryptographic solution for detecting overlapping samples/relatives across the cohorts with little compromise of individual privacy. We investigate i) the requirement of iteration for randomization and the compromise of privacy; ii) the choice of optimal SNP markers, in particular in the presence of multi-ethnicity. We demonstrated its application of the randomized algorithm in UKB for detecting first-degree relatives.

The general idea of using randomness in this study falls in randomized matrix theory and applications1.

**Key elements:**

,the genotype matrix for the cohort.

, the relationship matrix

,the orthonormal matrix for

,the approximation for .

**Question 1**: how the randomized genetic relationship matrix approximates the true relationship?

**Question 2**: how to determine the dimension of the low-rank approximation matrix? Power calculation function for linear regression.

**Question 3**: cross ethnicity problem—how to choose columns (loci) for ?

**Question 4**: If is different between a pair of cohorts, say using summary statistics from GWAS?

**1 Randomized estimator for genetic relationship**

Given the genotypic matrix , with dimension of for samples and , the conventional genomic relationship can be expressed as , it takes to finish the computation. is the standardized form for .

Alternatively, we can construct a much computationally faster but approximation for .

**2 Randomized algorithm is as below**

1 Generate matrix, from a normal distribution .

2 **,** . Of note, is an matrix. , in which is the column for the matrix, and on average each locus explains of the variation. Accordingly, using Mailman algorithm, it can further reduce the computational cost2.

3 .

It is easy to see that , an correlation matrix.Obviously, the greater is the more precisely approaches . So . It seems obvious that a big is anticipated to have unbased estimate . However, the computational cost for is , too big is not computational economic. Upon how big is, approaches under tolerance .

The steps above makes the so-called *randomization* kernel of many modern matrix algorithm1.

The proposed randomized algorithm resembles the previously proposed Gencrypt method3, but the presented one is more robust to various technique error and is able to detect not only overlapping individuals but their relatives.

For each PPS, , in which is the column for the matrix, and on average each locus explains of the variation.

For an individual a pair of PPS, say and , has .

Each PPS can be seen as a trait with because it does not have any sampling variance. For a pair of individuals, individual and individual , when both and have been standardized, their covariance for the PPS , in which is the relatedness scores in terms of identity by state4. Depending on the relatedness between a pair of individuals, for monozygous twins or to a duplicated sample, for first-degree relatives such as parent and offspring or full sibs. In general, for -degree of relatives, .

The theory presented above provides a theoretical basis for detecting overlapping samples using PPS other than sharing individual level genotypes. Assuming that each individual has independent PPS ( having elements), for individual and , we can regress on ,

**(Equation 12)**

in which is the grand mean, is the regression coefficient, and is the residual. . if individual is not correlated with individual , for first-degree relatives, and if individual and are genetically same, say an overlapping sample or homozygous twins. The sampling variance of is . Under the null distribution for no related or overlapping samples, . The residual accounts the discordant genotypes, including missing genotypes and genotyping or imputation errors. For current GWAS data, after quality control, the discordant rate is often smaller than 1%.

If now we have cohorts for which the individual genotypes of which cannot be disclosed to the central analysis hub, overlap between cohorts can be identified if PPS are supplied. By regressing their PPS to each other the overlapping individuals could be detected if . Assuming there are samples in each cohort, a total of regressions need to be carried out as defined in Equation 12. If we want to control the experiment-wise type I error rate under the null hypothesis and type II error rate (with power) for , the required number of pseudo profile scores for each individual is

**(Equation 13)**

in which and are scores under the given *p*-values at the subscripts. To accommodate technical errors, such as missing genotypes and genotype error, a cutoff of 0.95 for is adopted for detection of overlapping samples, and for detecting first-degree relative.

In Chen’s 2017 EJHG paper5, page 144, it reads

“We use WTCCC data as an illustration to detect 2934 shared controls between any two of the diseases by PPSR. Among 330K not palindromic loci, we randomly picked M=100, 200, and 500 SNPs, to generate pseudo profile scores. It generated 21 cohort-pair comparisons, leading to the summation for 488 587 090 total individual-pair tests. To have an experiment-wise type I error rate=0.01, type II error rate=0.05 (power=0.95) for detecting overlapping individuals, we needed to generated at least 57 PPSs.”

The standardization of genotypes can either use the allele frequency from each cohort, or from a reference sample. Throughout the study, we used the allele frequency calculated from UKB cohort as the reference, and using it as an approximation to standardize genotypes for all cohorts in comparison.

**Workflow for PPSR.** Given the statistical method for detecting overlapping samples as described above, the whole workflow for detecting can be split into three steps (**Supplementary Fig. 10**).

**In step 1, the required type I and type II error rates are defined and from that the required number of pseudo profiles to be generated.** The GWAMA central analyst selects consensus SNP markers across cohorts, and determines additive effects matrix that will be used to generate pseudo profile scores for each cohort. In order to avoid strand issues, the loci having palindromic loci (A/T alleles or G/C alleles) are excluded.

**In step 2, each cohort generates PPSR for each individual with the set of consensus markers and the marker weights received from the GWAMA coordinator.** After generate the PPS, they send them back to the coordinator. This will be a file that contains N rows and K columns with pseudo-profile scores.

**In step 3, the coordinator runs PPSR for each sample in a cohort on each PPS generated for another cohort.** The final product of running PPSR is to generate a matrix for a pair of cohorts, which have and samples respectively. For each pair of individuals in comparison, we take the one from cohort as the response variable and from cohort as the predictor variable in PPSR. In principle, swapping the response variable and the predictor variable do not affect the performance of PPSR. Each entry, the regression coefficient of PPSR, in the matrix represents genetic similarity for these pair of individuals in comparison. Once the regression coefficients are above the threshold, it indicates there are samples duplicated. The central analyst can then request each cohort that is implicated in containing samples that are also in other cohorts to drop those samples, without revealing where the duplication occurred.

**Privacy issues when using PPSR.** As the exchange of the PPS is within a meta-analysis facility, it is not as vulnerable as that of releasing the GWAS summary to the public domain as discussed in previous studies6–8. However, as PPS are generated from genotypes, it is worth to consider whether the PPS will reveal individual genotype information, or can be decoded from PPS. As a demonstration for the principle-of-proof, we consider to reverse Equation 11 to estimate genotypes. We consider the case where the additive effect matrix in Equation 11 is known, otherwise it is nearly impossible to recover genotype information. Given the workflow of PPSR, the analysts who coordinate the meta-analysis know the additive effect matrix, in Equation 11, and receive PPS from each cohort have the information to decode genotypes that are employed to generate PPS.

After reversing Equation 11, using the standard regression method, the genotype in each locus can be estimated as

**(Equation 14)**

In detail,

in which is the column in the additive effects matrix in Equation 11. Although , which is an unbiased estimate of the genotype, its sampling variance is . The sampling variance can be further written as because and is denoted as . The greater the ratio between and , the larger the sampling variance, and consequently the lower probability to construct the real genotype.

Without loss of generality, the accuracy of the estimated , a continuous variable, and , a discrete variable with values of 2, 1, and 0, can be measure using the squared correlation ()9,

**(Equation 15)**

in which and are:

, , and if is the reference allele, and , , and are weighted frequency given the distribution of . .

When the reference allele frequency follows a uniform distribution between , assuming that the loci follow Hardy-Weinberg proportions, , , and , in which follows a uniform distribution between and and .

and , , and .

If the reference allele frequency follows a uniform distribution between (0, 0.5), .

Given loci with MAF of 0.5, the expected frequencies for , , and are , , , and ., and . Plugging them in to the Equation 13 leads to .

Equation 13 can be rewritten as , in which if MAF is 0.5, and if MAF in nearly from a uniform distribution. From Equation 13, it is easy to calculate the ratio between the number of markers and the number of PPS given a controlled ,

**(Equation 16)**

For uniform distribution of MAF, if is set as the threshold, ; if , , and if , . In general, the higher the ratio between and , the less information can be inferred. We suggest may be sufficient.

References

1 Halko N, Martinsson PG, Tropp JA. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev* 2011; **53**: 217–288.

2 Liberty E, Zucker SW. The Mailman algorithm: A note on matrix-vector multiplication. *Inf Process Lett* 2009; **109**: 179–182.

3 Turchin MC, Hirschhorn JN. Gencrypt: one-way cryptographic hashes to detect overlapping individuals across samples. *Bioinformatics* 2012; **28**: 886–8.

4 Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet* 2010; **11**: 800–5.

5 Chen G-B, Lee SH, Robinson MR *et al.* Across-cohort QC analyses of GWAS summary statistics from complex traits. *Eur J Hum Genet* 2017; **25**: 137–46.

6 Homer N, Szelinger S, Redman M *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008; **4**: e1000167.

7 Visscher PM, Hill WG. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet* 2009; **5**: e1000628.

8 Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. *Nat Genet* 2009; **41**: 965–7.

9 Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013; **9**: e1003348.