sPLINK: A Federated, Privacy-Preserving Tool as a Robust Alternative to Meta-Analysis in Genome-Wide Association Studies

Reza Nasirigerdeh[a], Reihaneh Torkzadehmahani[a], Julian Matschinske[a], Tobias Frisch[b], Markus List[a], Julian Späth[a], Stefan Weiß[c], Uwe Völker[c], Nina Kerstin Wenke[a], Tim Kacprowski[a,*], Jan Baumbach[a,b,*]

[a] *TUM School of Life Sciences, Technical University of Munich, Munich, Germany*
[b] *Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark*
[c] *Department of Functional Genomics, University Medicine Greifswald, Greifswald, Germany*

Abstract

Genome-wide association studies (GWAS) have been widely used to unravel connections between genetic variants and diseases. Larger sample sizes in GWAS can lead to discovering more associations and more accurate genetic predictors. However, sharing and combining distributed genomic data to increase the sample size is often challenging or even impossible due to privacy concerns and privacy protection laws such as the GDPR. While meta-analysis has been established an effective approach to combine summary statistics of several GWAS, its accuracy can be attenuated in the presence of cross-study heterogeneity. Here, we present *sPLINK* (*safe PLINK*), a user-friendly software tool set which performs federated, privacy-preserving GWAS on distributed datasets, preserving accuracy. *sPLINK* neither exchanges raw data nor does it rely on summary statistics. Instead, it performs model training in a federated manner, communicating only model parameters between cohorts and a central server. We verify that the federated results from *sPLINK* are the same as those from aggregated analysis conducted with *PLINK*. Moreover, we demonstrate that *sPLINK* is robust against imbalanced phenotype distributions across cohorts while existing meta-analysis tools considerably lose accuracy in such scenarios. Federated and user-friendly analysis with *sPLINK*, thus, has the potential to replace meta-analysis as the gold standard for collaborative GWAS.

Keywords: sPLINK, Federated Learning, Genome-Wide Association Studies, GWAS, Meta-Analysis, Distributed Machine Learning, Privacy

*Joint last authors

1. Introduction

Genome-wide association studies (GWAS) test millions of single nucleotide polymorphisms (SNPs) to identify possible associations between a specific variant and disease [1]. They have led to considerable achievements over the past decade including better comprehension of the genetic structure of complex diseases and the discovery of SNPs playing a role in particular traits or disorders [2, 3]. GWAS sample size is an important factor in detecting associations and larger sample sizes lead to identifying more associations and more accurate genetic predictors [2, 4].

PLINK [5] is a widely used open source software tool for GWAS. The major limitation of *PLINK* is that it can only perform association tests on local data. If multiple cohorts want to conduct a collaborative GWAS to take advantage of larger sample size, they can pool their data for a joint analysis (Figure 1); however, this is close to impossible due to privacy restrictions and data protection issues, especially concerning genetic and medical data. Hence, the field has established methods for meta-analysis of individual studies, where only the results and summary statistics of the individual analyses have to be exchanged [6] (Figure 1).

There are several software packages such as METAL [7], GWAMA [8] and PLINK [5] that implement different meta-analysis models including fixed or random effect models as well as Bayesian approaches and more [9]. Although meta-analysis approaches can protect the privacy, they suffer from two main constraints: first of all, they rely on detailed planning and agreement of cohorts on various study parameters such as meta-analysis model (e.g. fixed effect or random effect), meta-analysis tool (e.g. METAL or GWAMA), heterogeneity metric (e.g. Cochran's Q or the I^2 statistic), the covariates to be considered, etc [4]. Most importantly, the statistical power of meta-analysis can be adversely affected in the presence of cross-study heterogeneity, leading to inaccurate estimation of the joint results and yielding misleading conclusions [10, 11].

To address the aforementioned shortcomings, the association tests can be implemented in a federated fashion. That is, cohorts can collaboratively build the global association models by sharing the model parameters (e.g. hessian matrices) instead of private data (Figure 1). There are previous efforts in this regard [12, 13, 14, 15, 16] but they generally allows for only one association test (e.g. logistic regression or chi-square) and are not a user-friendly tool set like *PLINK* that can be easily deployed for GWAS.

In this paper, we present a novel federated tool set for GWAS called ***sPLINK (safe PLINK)***. Unlike *PLINK*, *sPLINK* is applicable to distributed data while protecting the data privacy. In *sPLINK*, each cohort installs a local client software that processes local data, extracts the local model parameters, and shares only those

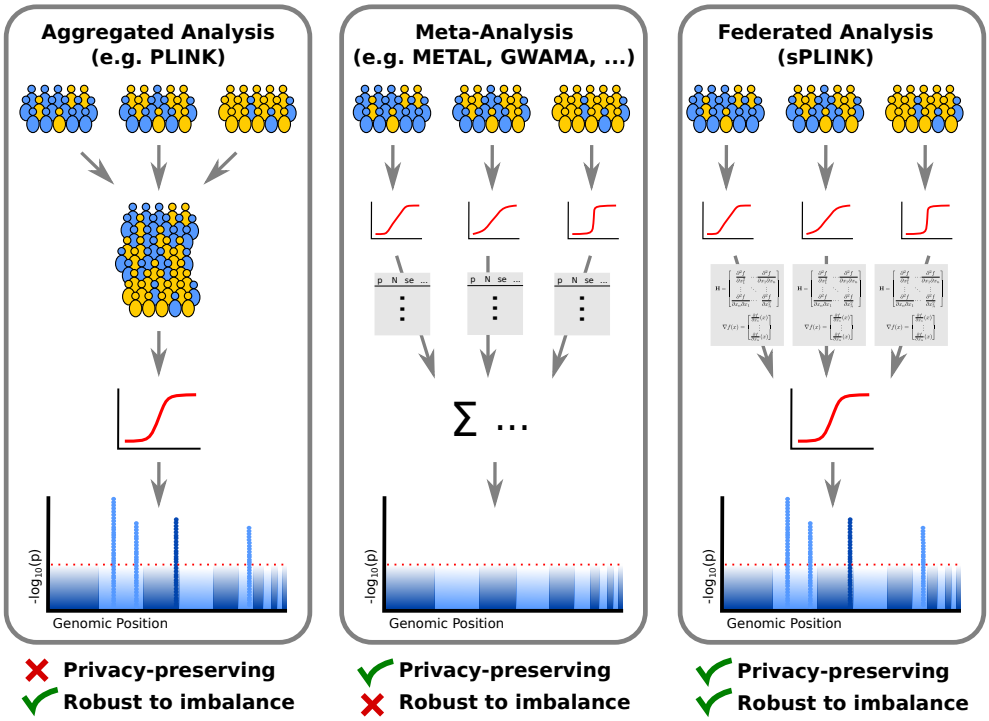


Figure 1: *sPLINK* vs. aggregated analysis and meta-analysis approaches: Aggregated analysis requires cohorts to pool their private data for a joint analysis. Meta-analysis approaches aggregate the summary statistics from the cohorts to estimate the combined p-values. In *sPLINK*, the cohorts extract the model parameters (e.g. hessian matrices) from the local data, which are aggregated to build a final model. *sPLINK* combines the advantages of aggregated analysis and meta-analysis, i.e. robustness against imbalanced data and preserving the privacy of cohorts' data. Yellow/blue color indicates case/control samples, respectively.

(but not the data) with the central server (Section 2 and Appendix B). Contrary to existing distributed tools for GWAS such as GLORE [12], *sPLINK* not only provides a user-friendly and easy-to-use web interface but also supports multiple association tests including logistic regression [17], linear regression [18], and chi-square (χ^2) [19] for GWAS (Appendix B).

The advantage of *sPLINK* over current meta-analysis approaches is two-fold: usability and robustness against heterogeneity. *sPLINK* is easier to use for collaborative GWAS compared to meta-analysis. In *sPLINK*, a coordinator initiates a collaborative study and invites the cohorts to join. The only decision the cohorts make is whether or not to contribute to the study. After accepting the invitation, the cohorts just join the study using their credentials and select the dataset they

want to employ in the study (Section 2 and Appendix A).

More importantly, *sPLINK* is robust against heterogeneous (i.e. imbalanced) phenotype distributions among the cohorts. It always gives the same results as aggregated analysis regardless of the heterogeneity of the phenotype distributions. In contrast, meta-analysis tools typically lose statistical power as the heterogeneity of phenotype distributions among the cohorts increases (Figure 2 and Section 3).

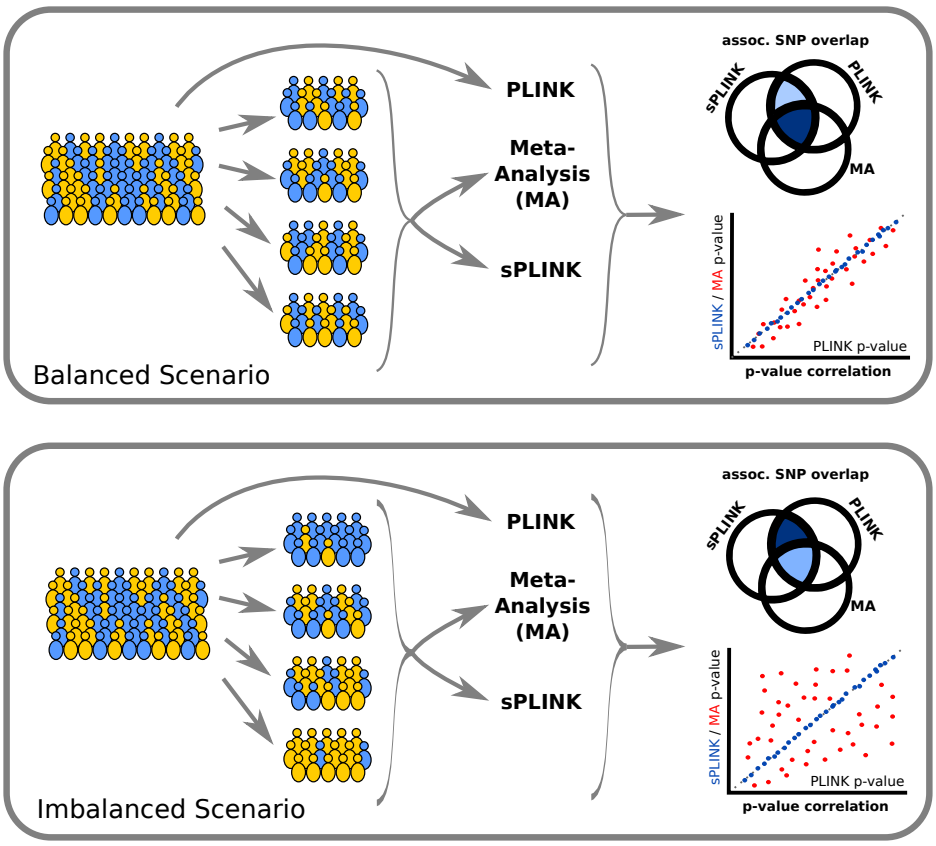


Figure 2: *sPLINK* vs. meta-analysis in different scenarios: *sPLINK* always obtains the same p-values as aggregated analysis and, thus, identifies the same set of SNPs as significant regardless of phenotype imbalance across cohorts. Meta-analysis gives comparable p-values in a balanced scenario whereas it loses accuracy in imbalanced scenarios, which leads to missing significant SNPs.

2. sPLINK

sPLINK implements a horizontal federate learning approach [20, 21, 22] to preserve the privacy of data. Federated learning is a type of distributed learning, where

multiple cohorts collaboratively learn a joint (global) model under the orchestration of a central server [23]. The cohorts never share their private data with the server or the other cohorts. Instead, they extract local parameters from their data (that cannot be used to infer any details about the individual samples in the data) and send them to the server. The server aggregates the local parameters from all cohorts to compute the global model parameters (or global results), which, in turn, are shared with all cohorts. Specifically, *sPLINK* works with distributed GWAS datasets, where samples are individuals and features are SNPs and categorical or quantitative phenotypic variables. While the samples are different across the cohorts, the feature space is the same since *sPLINK* only considers SNPs and phenotypic variables that are common among all datasets.

The functional workflow of sPLINK is as follows: the coordinator initializes the project, creates a project token for each cohort, and sends the corresponding token to the cohort. Next, the cohorts join the project using their username, password, and token, contributing their dataset to the study. After the cohorts joined, the project is started automatically and the association test results are computed in a federated fashion. When the federated analysis is completed, the cohorts and coordinator can have access to the results (more details on the functional workflow of *sPLINK* can be found in Appendix A).

From a technical point of view, *sPLINK*'s architecture consists of three main software components: client, server, and web application (WebApp). The client package is installed on the local machine of each cohort with access to the private data. It computes the model parameters from the local data and sends the parameters to the server. The server and WebApp packages are installed on a central server. The former is responsible for aggregating the local parameters from all the cohorts to calculate the global parameters. The latter is employed to configure the parameters (e.g. association test) of the new study (Appendix A). *sPLINK* also provides a chunking capability to handle large datasets containing millions of SNPs. The chunk size (configured by the coordinator) specifies how many SNPs should be processed in parallel. Larger chunk sizes allow for more parallelism, and therefore less running time but require more computational resources (e.g. CPU and main memory) from the local machines of the cohorts and the server. See Appendix B for more details regarding the computational workflow of *sPLINK*. While we provide a readily usable web service running the server and WebApp packages[1], these packages can, of course, also be installed by a cohort on an own web server.

[1] *https://exbio.wzw.tum.de/splink*

Dataset	Samples	SNPs	Adjustments	Phenotype
SHIP	3699	$\sim 5M$	sex, age, smoking status, daily alcohol consumption	SLA, dichotomous (top 25th percentile, 934 cases, 2765 controls) SLA, quantitative, Mean±SD 1.23±0.3
COPDGene	5343	$\sim 600K$	sex, age, smoking status, pack years of smoking	COPD, dichotomous, (2811 cases, 2532 controls)

Table 1: Description of datasets: The *SHIP* (Study of Health in Pomerania) and *COPDGene* (Genetic Epidemiology of COPD) datasets are used to verify *sPLINK* and compare *sPLINK* with the existing meta-analysis tools. COPD: chronic obstructive pulmonary diseases, SLA: serum lipase activity

3. Results

We first verify *sPLINK* by comparing its results with those from aggregated analysis conducted with *PLINK* for all three association tests on a real GWAS dataset from the SHIP study [24]. We refer to this dataset as the *SHIP* dataset, which comprises the records of 3699 individuals with *serum lipase activity* as phenotype. The quantitative version represents the square root transformed serum lipase activity, while the dichotomous (binary) version indicates whether the serum lipase activity of an individual is above or below the upper 25th percentile. The *SHIP* dataset contains around 5 million SNPs and sex, age, smoking status (current-, ex-, or non-smoker), and daily alcohol consumption (in g/day) as confounding factors (Table 1).

We employ the binary phenotype for logistic regression and the chi-square test, and the quantitative phenotype for linear regression. We incorporate all four confounding factors in the regression models and no confounding factor in the chi-square test. We horizontally (sample-wise) split the dataset into four parts, simulating four different cohorts (with sample sizes of 1044, 1006, 941, and 708, respectively). *PLINK* computes the statistics for each association test using the whole dataset while *sPLINK* does it in a federated manner using the splits of the individual cohorts. To be consistent with *PLINK*, *sPLINK* calculates the same statistics as *PLINK* for the association tests (Appendix A).

We compute the difference between the p-values (p) as well as Pearson correlation coefficient (ρ) of p-values from *sPLINK* and *PLINK*. We use -$log_{10}(p)$ since the p-values are typically small and -$log_{10}(p)$ can be a better indicator of small p-value differences. According to Figure 3a-c, the p-value difference is zero for most of the

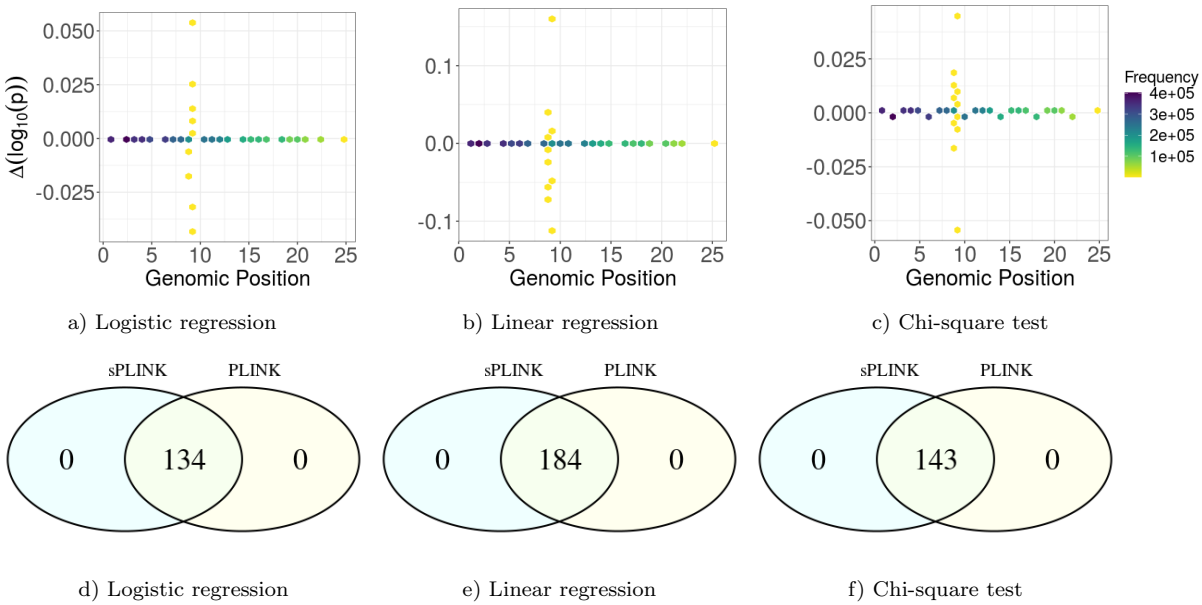


Figure 3: $\Delta log_{10}(p)$ between *sPLINK* and *PLINK* as well as the set of SNPs identified by *sPLINK* and *PLINK* as significant for logistic regression (a) and (d), linear regression (b) and (e), and chi-square test (c) and (f), respectively. For most of the SNPs, the difference is zero, indicating that *sPLINK* gives the same p-values as *PLINK*. The negligible difference between p-values for the other SNPs can be attributed to differences in floating point precision. *sPLINK* and *PLINK* also recognize the same set of SNPs as significant.

SNPs. We also observe that the maximum difference is 0.162 for a SNP in the linear regression. *sPLINK* and *PLINK* report 4.441×10^{-16} and 3.058×10^{-16} as p-values for the SNP, respectively. This negligible difference can be attributed to inconsistencies in floating point precision.

The correlation coefficient of p-values from *sPLINK* and *PLINK* for all three tests is $0.\overline{99}$, which is consistent with the results of p-value difference from Figure 3a-c. We also investigated the overlap of significantly associated SNPs between *sPLINK* and *PLINK*. We consider a SNP as significant if its p-value is less than 5×10^{-8}. *PLINK* and *sPLINK* recognize the same set of SNPs as significant (Figure 3d-f). Notably, the identified SNPs have also been implicated in a previous analysis of this dataset [25]. These results indicate that p-values computed by *sPLINK* in a federated manner are the same as those calculated by *PLINK* on the aggregated data (ignoring negligible floating point precision error). In other words, the federated computation in *sPLINK* preserves the accuracy of the results of the association tests.

Next, we compare *sPLINK* with some existing meta-analysis tools, namely *PLINK*,

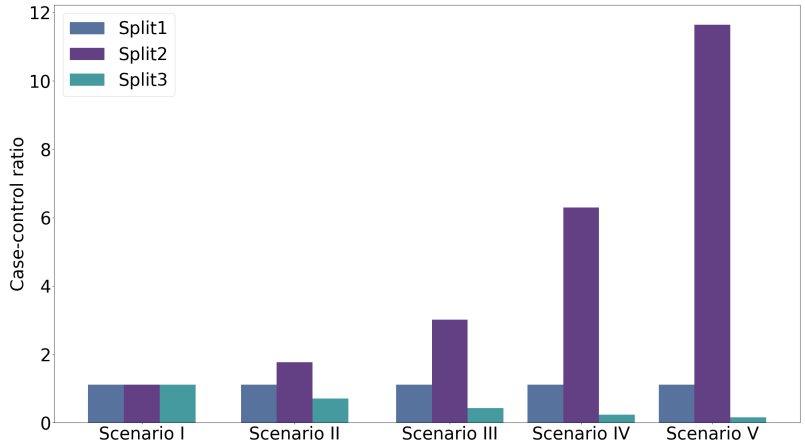


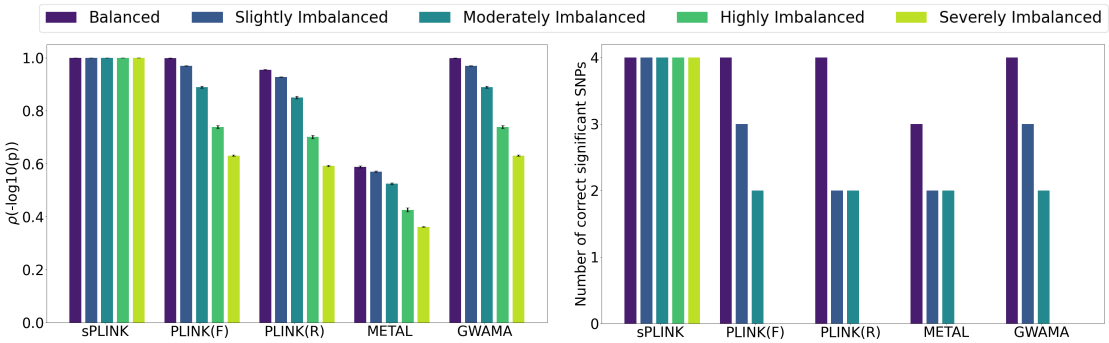
Figure 4: Scenarios used to compare *sPLINK* with the meta-analysis tools: *Scenario I* is *Balanced* scenario, where case-control ratio is the same (1.11) for all splits. As the scenarios become more imbalanced (from *Scenario II* to *Scenario V*), the case-control ratio in split2/split3 increases/decreases but it remains the same in split1.

METAL, and *GWAMA*. To do so, we leverage the *COPDGene* dataset (non-hispanic white ethnic group) [26], which has an equal distribution of case and control samples unlike the *SHIP* dataset (Table 1). *COPDGene* contains 5343 samples (ignoring 1327 samples with missing phenotype value) with binary phenotypes indicating COPD (chronic obstructive pulmonary disease) and around 600K SNPs. We included sex, age, smoking status, and packs years of smoking as confounding factors [27] (Table 1).

To simulate cross-study heterogeneity [28], we consider five different scenarios: *Scenario I* (*Balanced*), *Scenario II* (*Slightly Imbalanced*), *Scenario III* (*Moderately Imbalanced*), *Scenario IV* (*Highly Imbalanced*), and *Scenario V* (*Severely Imbalanced*) (Figure 4). In each scenario, we partition the dataset into three splits with the same sample size of 1781. The scenarios are different from each other in terms of the number of case and control samples in each split. In the balanced scenario, all three splits have the same case-control ratio while in the other scenarios, the case-control ratio of the splits differ from each other.

We obtain the summary statistics (e.g. minor allele, odds ratio, standard error, etc.) for each split to conduct meta-analyses. The results are then compared to the federated analysis employing *sPLINK*. Figure 5a shows the Pearson correlation coefficient of -$log_{10}(p)$ between each tool and the aggregated analysis for all five scenarios. Figure 5b depicts the number of SNPs correctly identified as significant

by the tools (true positives). We repeat each scenario five times, where in each run the dataset is randomly partitioned into three splits (the same sample size) ensuring the case-control ratio as depicted in Figure 4.



a) The Pearson correlation coefficient

b) Correctly identified significant SNPs

Figure 5: (a) The Pearson correlation coefficient (ρ) of -$log_{10}(p)$ between each tool and aggregated analysis and (b) the number of SNPs correctly identified as significant (true positives) by each tool. The height of each bar in (a) and (b) represents the average and median across the five runs, respectively. Error bars indicate the standard deviation. F and R stand for fixed-effect and random-effect, respectively.

According to Figure 5a, the correlation of p-values between *sPLINK* and the aggregated analysis is ~ 1.0 for all five scenarios, implying that *sPLINK* gives the same p-values as aggregated analysis regardless of the phenotype distribution imbalance in the cohorts. In contrast, the correlation coefficient for the meta-analysis tools shrinks with increasing heterogeneity, indicating loss of accuracy.

Figure 5b illustrates that *sPLINK* correctly identifies all four significant SNPs in all scenarios. In the balanced scenario, almost all meta-analysis tools perform well and recognize all significant SNPs. An exception is METAL, which misses one of them. However, they miss more and more significant SNPs as the phenotype heterogeneity across splits increases. In the *Highly Imbalanced* and *Severely Imbalanced* scenarios, the meta-analysis tools cannot recognize any significant SNP. We also checked the number of SNPs wrongly identified as significant by the tools (false positives). *sPLINK* has no false positive in any of the scenarios and the meta-analysis tools introduce zero or one false positive depending on the scenario.

In summary, we showed that *sPLINK* is robust against the heterogeneity of phenotype distributions across the cohorts. That is, *sPLINK* always delivers the same p-values as aggregated analysis and correctly identifies all significant SNPs independent of the phenotype distributions across the separate datasets. In contrast,

the meta-analysis tools lose their statistical power in heterogeneous scenarios, which leads to missing some or all significant SNPs.

4. Conclusion and Future Development

We introduced *sPLINK*, a user-friendly federated tool set for GWAS. *sPLINK* preserves the privacy of the cohorts' data without sacrificing the accuracy of the test results. It supports multiple association tests including chi-square, linear regression, and logistic regression. *sPLINK* is consistent with *PLINK* in terms of the input data formats it supports and the results it reports to the user. We compared *sPLINK* to aggregated analysis with *PLINK* as well as meta-analysis with *METAL*, *GWAMA*, and *PLINK*. While *sPLINK* is robust against the heterogeneity of phenotype distributions across separate datasets, the statistical power of the meta-analysis tools is attenuated in imbalanced scenarios. We argued that *sPLINK* is easier to use for collaborative GWAS compared to meta-analysis approaches thanks to its straightforward functional workflow.

The future development of *sPLINK* can go in several directions. We plan to implement the federated version of more association tests [29] or more machine learning algorithms including random forest [30] or deep neural networks (DNN) [31] leveraged by the GWAS community in *sPLINK*. We will also investigate *sPLINK*'s potential to tackle other open challenges in GWAS such as trans-ethnicity [32], where the samples in the distributed datasets are from different ethnic groups.

In summary, *sPLINK* is a novel and robust alternative to meta-analysis, which performs collaborative GWAS in a federated and privacy-preserving manner. It has the potential to immensely impact the statistical genetics community by addressing current challenges in GWAS including cross-study heterogeneity and, thus, to replace meta-analysis as the gold standard for collaborative GWAS.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078. This reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B). SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany (www.community-medicine.de), which is funded by the Federal Ministry of Education and Research (grants no.

01ZZ9603, 01ZZ0103, and 01ZZ0403), the Siemens AG, the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A). ExomeChip data have been supported by the Federal Ministry of Education and Research (grant no. 03Z1CN22) and the Federal State of Mecklenburg-West Pomerania

References

[1] M. Fareed, M. Afzal, Single nucleotide polymorphism in genome-wide association of human population: A tool for broad spectrum service, Egyptian Journal of Medical Human Genetics 14 (2013) 123 – 134.

[2] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, J. Yang, 10 years of gwas discovery: biology, function, and translation, The American Journal of Human Genetics 101 (2017) 5–22.

[3] P. M. Visscher, M. A. Brown, M. I. McCarthy, J. Yang, Five years of gwas discovery, The American Journal of Human Genetics 90 (2012) 7–24.

[4] R. De, W. S. Bush, J. H. Moore, Bioinformatics challenges in genome-wide association studies (gwas), in: Clinical Bioinformatics, Springer, 2014, pp. 63–81.

[5] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al., Plink: a tool set for whole-genome association and population-based linkage analyses, The American journal of human genetics 81 (2007) 559–575.

[6] E. Evangelou, J. P. Ioannidis, Meta-analysis methods for genome-wide association studies and beyond, Nature Reviews Genetics 14 (2013) 379–389.

[7] C. J. Willer, Y. Li, G. R. Abecasis, Metal: fast and efficient meta-analysis of genomewide association scans, Bioinformatics 26 (2010) 2190–2191.

[8] R. Mägi, A. P. Morris, Gwama: software for genome-wide association meta-analysis, BMC bioinformatics 11 (2010) 288.

[9] K. L. Lunetta, Methods for meta-analysis of genetic data, Current protocols in human genetics 77 (2013) 1–24.

[10] R. M. Cantor, K. Lange, J. S. Sinsheimer, Prioritizing gwas results: a review of statistical methods and recommendations for their application, The American Journal of Human Genetics 86 (2010) 6–22.

[11] R. de Vlaming, A. Okbay, C. A. Rietveld, M. Johannesson, P. K. Magnusson, A. G. Uitterlinden, F. J. van Rooij, A. Hofman, P. J. Groenen, A. R. Thurik, et al., Meta-gwas accuracy and power (metagap) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies, PLoS genetics 13 (2017).

[12] Y. Wu, X. Jiang, J. Kim, L. Ohno-Machado, Grid binary logistic regression (glore): building shared models without sharing data, Journal of the American Medical Informatics Association 19 (2012) 758–764.

[13] W. Jiang, P. Li, S. Wang, Y. Wu, M. Xue, L. Ohno-Machado, X. Jiang, Webglore: a web service for grid logistic regression, Bioinformatics 29 (2013) 3238–3240.

[14] Y. Li, X. Jiang, S. Wang, H. Xiong, L. Ohno-Machado, Vertical grid logistic regression (vertigo), Journal of the American Medical Informatics Association 23 (2015) 570–579.

[15] S. Wang, X. Jiang, Y. Wu, L. Cui, S. Cheng, L. Ohno-Machado, Expectation propagation logistic regression (explorer): distributed privacy-preserving online model learning, Journal of biomedical informatics 46 (2013) 480–496.

[16] S. D. Constable, Y. Tang, S. Wang, X. Jiang, S. Chapin, Privacy-preserving gwas analysis on federated genomic datasets, in: BMC medical informatics and decision making, volume 15, BioMed Central, p. S2.

[17] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, volume 1, Springer series in statistics New York, 2001.

[18] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, volume 1, Springer series in statistics New York, 2001.

[19] M. L. McHugh, The chi-square test of independence, Biochemia medica: Biochemia medica 23 (2013) 143–149.

[20] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, et al., Communication-efficient learning of deep networks from decentralized data, arXiv preprint arXiv:1602.05629 (2016).

[21] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, D. Bacon, Federated learning: Strategies for improving communication efficiency, arXiv preprint arXiv:1610.05492 (2016).

[22] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–19.

[23] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, arXiv preprint arXiv:1912.04977 (2019).

[24] H. Völzke, D. Alte, C. O. Schmidt, D. Radke, R. Lorbeer, N. Friedrich, N. Aumann, K. Lau, M. Piontek, G. Born, et al., Cohort profile: the study of health in pomerania, International journal of epidemiology 40 (2011) 294–307.

[25] F. U. Weiss, C. Schurmann, A. Guenther, F. Ernst, A. Teumer, J. Mayerle, P. Simon, H. Völzke, D. Radke, A. Greinacher, et al., Fucosyltransferase 2 (fut2) non-secretor status and blood group b are associated with elevated serum lipase activity in asymptomatic subjects, and an increased risk for chronic pancreatitis: a genetic association study, Gut 64 (2015) 646–656.

[26] COPDGene, `http://www.copdgene.org/`, (accessed Mar 22, 2020).

[27] S. G. Pillai, D. Ge, G. Zhu, X. Kong, K. V. Shianna, A. C. Need, S. Feng, C. P. Hersh, P. Bakke, A. Gulsvik, et al., A genome-wide association study in chronic obstructive pulmonary disease (copd): identification of two major susceptibility loci, PLoS genetics 5 (2009).

[28] Y.-F. Pei, Q. Tian, L. Zhang, H.-W. Deng, Exploring the major sources and extent of heterogeneity in a genome-wide association meta-analysis, Annals of human genetics 80 (2016) 113–122.

[29] EPACTS, `https://genome.sph.umich.edu/wiki/EPACTS`, (accessed Mar 22, 2020).

[30] T.-T. Nguyen, J. Z. Huang, Q. Wu, T. T. Nguyen, M. J. Li, Genome-wide association data classification and snps selection using two-stage quality-based random forests, in: BMC genomics, volume 16, Springer, p. S5.

[31] P. Bellot, G. de los Campos, M. Pérez-Enciso, Can deep learning improve genomic prediction of complex human traits?, Genetics 210 (2018) 809–819.

[32] Y. R. Li, B. J. Keating, Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations, Genome medicine 6 (2014) 91.

[33] PLINK data formats, `http://zzz.bwh.harvard.edu/plink/data.shtml`, (accessed Mar 22, 2020).

Appendix A. *sPLINK*'s Functional Workflow

The *functional workflow* of *sPLINK* (Figure A.6) is comprised of six major steps:

1. **Project Creation**: The coordinator creates the project (new study) through the WebApp interface. To this end, he/she specifies project name, association test name, chunk size, and the list of confounding features (only for regression tests). He/She also generates a unique project token for each cohort.
2. **Cohort Invitation**: The coordinator sends the project token to each cohort via a safe channel for inviting them to the project.
3. **Cohort Joining**: The cohorts use their corresponding username, password, and project token to join the project.
4. **Dataset Selection**: After the cohorts successfully joined the project, they can see the project parameters configured by the coordinator (e.g. association test name and chunk size). They need to choose the dataset they want to employ in the study. To be consistent with *PLINK*, *sPLINK* supports *.bed* (value of SNPs), *.fam* (sample IDs as well as gender and phenotype values), *.bim* (chromosome number, name, and base-pair distance of each SNP), *.cov* (value of confounding factors), and *.pheno* (phenotype values that should be used instead of those in *.fam* file) file formats as specified in the *PLINK* manual [33]. For linear regression, phenotype values must be quantitative while for logistic regression and chi-square, phenotype values have to be binary (control/case are encoded as 1/2).
5. **Federated Computation**: In *sPLINK*, the association test results are computed by the client package (running on the local machines of cohorts) and server package (running in the central server) in a federated manner. The computation is iterative and consists of four general steps:
   (a) **Get Global Parameters**: All clients obtain the required global model parameter values from the server.
   (b) **Compute Local Parameters**: Each client computes the local model parameter values using the local data and global parameter values from the server.
   (c) **Post Local Parameters**: The local parameter values are shared with the server.
   (d) **Aggregate Local Parameters**: When the server receives the local parameters from all clients, it starts the aggregation process. This involves simple logical operations such as equality check or mathematical operations like addition, subtraction, or multiplication of scalar values or matrices. The results of the aggregation process are the global model parameter values.

6. **Result Downloading**: The final results are automatically downloaded for the cohorts but the coordinator needs to download them manually through the web interface. Similar to *PLINK*, *sPLINK* reports minor allele name (*A1*) and p-value (*P*) for all three association tests, chi-square (*CHISQ*), odds ratio (*OR*), minor allele frequency in cases (*F_A*), and minor allele frequency in controls (*F_U*) for chi-square test, and the number of non-missing samples (*NMISS*), beta (*BETA*), and t-statistic (*STAT*) for linear and logistic regression tests.

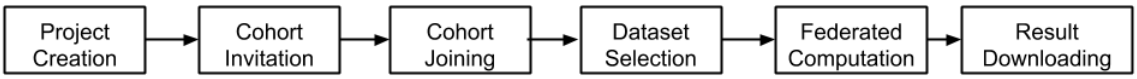


Figure A.6: Functional workflow of *sPLINK*: the coordinator creates a new project and invites a set of cohorts to join the project. The cohorts join the project and select the dataset. The project is started automatically, when all cohorts joined. After the computation is done, the cohorts and coordinator can access the results.

Appendix B. *sPLINK*'s Computational Workflow

The *computational workflow* of *sPLINK* involves six steps in common among all association tests as well as a couple of steps specific to each association test (Figure B.7). In the following, we provide an overview of each step.

1. **Init**: Each client i opens the files of the dataset selected by the cohort to be employed in the study and creates its phenotype vector (Y_i) and feature matrix (X_i), which includes the value of SNPs and confounding factors. It is worth noting that there is a separate feature matrix for each SNP but the phenotype vector is the same for all SNPs. Assume a dataset containing three SNPs named *SNP1*, *SNP2*, and *SNP3* and *age* and *sex* as confounding features. There will be three different feature matrices, one feature matrix per SNP. For instance, the feature matrix of *SNP1* has three columns including *SNP1*, *age*, and *sex* values. Phenotype vector and feature matrix are the private data of the cohorts. They cannot be shared with the coordinator or the other cohorts. The aggregation process in the server just makes sure that all clients successfully initialized the project.
2. **SNP Name**: Each client extracts the SNP names from its *.bim* file. In the aggregation process, the server computes the intersection of all SNP names. Only shared SNPs are considered in the computation of the association test results.

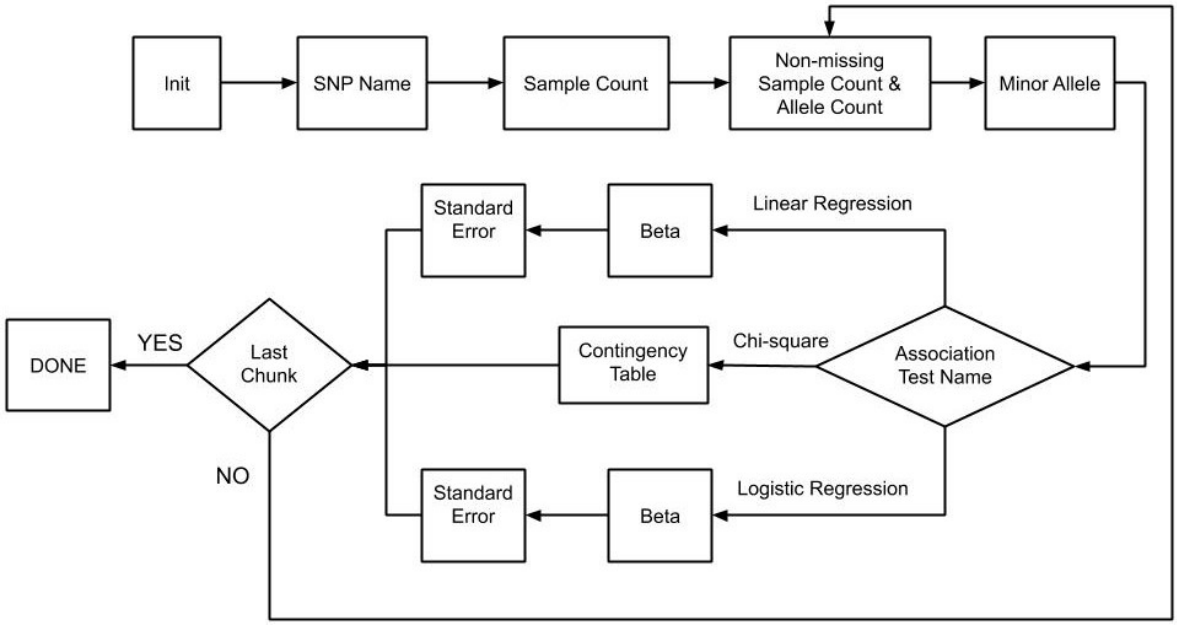


Figure B.7: Computational workflow of sPLINK: six steps are common among all three association tests: (1) *Init* in which the clients open the required files and prepare the dataset for computation; (2) *SNP Name*, where the clients share the SNP names from .bim file with the server; (3) *Sample Count* in which each client sends the number of the samples in the dataset considering those with missing values to the server; (4) *Non-missing Sample Count & Allele Count*, where the clients share the sample count ignoring those with missing values as well as the frequency of each allele with the server; (5) *Minor Allele* in which the clients update the minor allele name based on global minor allele and update the mapping of the SNP values accordingly; after *Minor Allele* step, there are association test specific steps: *Beta* and *Standard Error* steps for regression tests and *Contingency Table* for chi-square test. Logistic and linear regression compute beta and standard error in different ways, especially beta step is iterative for logistic regression. The regression tests compute p-values from standard error values while chi-square does it using the contingency table. The last step is *Done*, where the results from federated computation are shared with the cohorts. Notice that the results from *Non-missing Sample Count & Allele Count*, *Minor Allele*, and association test specific steps are per SNP and they are computed chunk by chunk.

3. **Sample Count**: Each client i calculates its local sample count n_i (number of samples in its dataset including missing samples, which is the size of vector Y_i). The server adds up the local sample counts from K clients to compute the global sample count: $n = \sum_{i=1}^{i=K} n_i$
4. **Non-missing Sample Count & Allele Count**: In this step, SNPs are

split into chunks which can be processed in parallel. The chunking capability is provided to handle very large datasets containing millions of SNPs. The clients compute the sample count after filtering out the missing values (value of -9 is considered as missing). Likewise, they calculate the local allele count by counting the number of alleles in each SNP. In the aggregation process, the server adds up the local non-missing sample count from the clients to compute the global non-missing sample count. Next, it does the same for the local allele count and calculates the global allele count. Finally, the server determines the global minor allele based on the value of the global allele count.

5. **Minor Allele**: The clients compare their local minor allele with the global minor allele. If they are the same, they do nothing. Otherwise, they update the mapping of SNP values read from .bed file. Each SNP value can be 0,1,2 or 3 (missing value). These values are encoded based on the minor allele name. If the minor allele is changed, the value of the SNP needs to be swapped if it is 0 or 2. Thus, if a client's minor allele is different from global minor allele, it inverses the mapping of SNP values ($0 \rightarrow 2$ and $2 \rightarrow 0$). The aggregation in the server makes sure that all clients successfully completed this step.
6. **Association Test Specific Steps**: In the following, we elaborate the steps specific to each association test.

   Chi-square: The only test-specific step for chi-square test is *Contingency Table*, where each client i computes its local contingency table containing minor allele frequency for cases (p_i), minor allele frequency for controls (r_i), major allele frequency for cases (q_i), and major allele frequency for controls (s_i). The server adds up the locally computed contingency tables from K clients to compute the global (observed) contingency table (Table B.2). It also calculates an expected contingency table based on the observed contingency table (Table B.3).
   Given the observed contingency table (O) and the expected contingency table (E), the server computes odds ratio (OR), χ^2, and p-value (P) as follows:

$$OR = \frac{p \times s}{q \times r} \tag{B.1}$$

$$\chi^2 = \sum \frac{(E - O)^2}{E} \tag{B.2}$$

$$P = 1 - F_t(\chi^2, 1) \tag{B.3}$$

	Minor Allele	Major Allele	Total
Case	$p = \sum_{i=1}^{i=K} p_i$	$q = \sum_{i=1}^{i=K} q_i$	$p + q$
Control	$r = \sum_{i=1}^{i=K} r_i$	$s = \sum_{i=1}^{i=K} s_i$	$r + s$
Total	p + r	q + s	n

Table B.2: Global (observed) contingency table

	Minor Allele	Major Allele
Case	$\frac{(p+q)\times(p+r)}{n}$	$\frac{(p+q)\times(q+s)}{n}$
Control	$\frac{(r+s)\times(p+r)}{n}$	$\frac{(r+s)\times(q+s)}{n}$

Table B.3: Expected contingency table

, where F_t is the cumulative distribution function (CDF) of χ^2 distribution (degree of freedom is 1).

Linear Regression: *Beta* and *Standard Error* are two steps specific to linear regression test. In the *Beta* step, each client i computes $X_i^T X_i$ and $X_i^T Y_i$, where X_i^T is the transpose of X_i. In the aggregation process, the server performs the following calculations (K is the number of clients):

$$X^T X = \sum_{i=1}^{i=K} X_i^T X_i \tag{B.4}$$

$$X^T Y = \sum_{i=1}^{i=K} X_i^T Y_i \tag{B.5}$$

$$\beta = (X^T X)^{-1}(X^T Y) \tag{B.6}$$

, where $()^{-1}$ indicates the inverse matrix.

In *Standard Error* step, each client i calculates the local sum square error (SSE_i) by having the global β vector.

$$\hat{Y}_i = X_i\beta \tag{B.7}$$

$$SSE_i = \sum(Y_i - \hat{Y}_i)^2 \tag{B.8}$$

and then the server calculates the global standard error vector (SE) as follows:

$$VAR = (\frac{\sum_{i=1}^{i=K} SSE_i}{n-m-1})(X^TX)^{-1} \tag{B.9}$$

$$SE = \sqrt{\text{diag}(VAR)} \tag{B.10}$$

, where n is global non-missing sample count, m is the number of features (1 + number of confounding factors), and *diag* is the main diagonal of the matrix. Given the standard error vector, the server computes the T *statistic* (T) and *p-value* (P) as follows:

$$T = \frac{\beta}{SE} \tag{B.11}$$

$$DF = n - m - 1 \tag{B.12}$$

$$P = 2 \times (1 - F_t(|T|, DF)) \tag{B.13}$$

, in which DF is degree of freedom and F_t is the CDF of T distribution.

Logistic Regression: Similar to linear regression, logistic regression has two specific steps: *Beta* and *Standard Error*. However, the *Beta* step is iterative in logistic regression (maximum number of iterations is specified by the coordinator and its default value is 20). In each iteration, each client i computes local gradient (∇_i), hessian matrix (H_i) and log-likelihood (L_i) as follows:

$$\hat{Y}_i = \frac{1}{1 + e^{-X_i\beta}} \tag{B.14}$$

$$\nabla_i = X_i^T(Y_i - \hat{Y}_i) \tag{B.15}$$

$$H_i = (X_i^T \circ (\hat{Y}_i \circ (1 - \hat{Y}_i))^T)X_i \tag{B.16}$$

$$L_i = \sum(Y_i \circ \log \hat{Y}_i + (1 - Y_i) \circ \log(1 - \hat{Y}_i)) \tag{B.17}$$

, where β is the global beta vector from the previous iteration and $\circ$ indicates element-wise multiplication.
The server aggregates the local gradients, hessian matrices and log-likelihood values from K cohorts as follows:

$$\nabla = \sum_{i=1}^{i=K} \nabla_i \tag{B.18}$$

$$H = \sum_{i=1}^{i=K} H_i \tag{B.19}$$

$$L = \sum_{i=1}^{i=K} L_i \tag{B.20}$$

Then, it updates the β values accordingly:

$$\beta_{new} = \beta_{old} + H^{-1}\nabla \tag{B.21}$$

, where β_{old} is the β value from the previous iteration. The server also compares the newly computed log-likelihood value (L) with the one from previous iteration(L_{old}). If their difference is less than a pre-specified threshold, β values converged, and therefore, it stops updating beta.
In the *Standard Error* step, the server shares the global β values with the clients. Each client i computes its local hessian matrix (H_i) using the global β. The server gets the local hessian matrices from K cohorts and applies the following formula to obtain the global standard error vector (SE):

$$SE = \sqrt{\text{diag}\Big(\big(\sum_{i=1}^{i=K} H_i \big)^{-1} \Big)} \tag{B.22}$$

Having standard error values, the server calculates T statistics (T) and p-value (P) as follows:

$$T = \frac{\beta}{SE} \tag{B.23}$$

$$P = 1 - F_t(|T|^2, 1) \tag{B.24}$$

, where F_t is CDF of χ^2 distribution (degree of freedom is 1).

7. **DONE**: The computation of association test results have been completed for all chunks and the results are shared with all cohorts.