Caring without sharing: Meta-analysis 2.0 for massive genome-wide association studies

Armin Pourshafeie¹, Carlos D. Bustamante^{2,3}, Snehit Prabhu^{2,#}

¹Department of Physics, Stanford University,

²Department of Biomedical Data Science, Stanford University,

³Chan Zuckerberg Biohub, San Francisco, CA

#E-mail: snehit@stanford.edu

October 5, 2018

Abstract

Genome-wide association studies have been effective at revealing the genetic architecture of simple traits. Extending this approach to more complex phenotypes has necessitated a massive increase in cohort size. To achieve sufficient power, participants are recruited across multiple collaborating institutions, leaving researchers with two choices: either collect all the raw data at a single institution or rely on meta-analyses to test for association. In this work, we present a third alternative. Here, we implement an entire GWAS workflow (quality control, population structure control, and association) in a fully decentralized setting. Our iterative approach (a) does not rely on consolidating the raw data at a single coordination center, and (b) does not hinge upon large sample size assumptions at each silo. As we show, our approach overcomes challenges faced by meta-studies when it comes to associating rare alleles and when case/control proportions are wildly imbalanced at each silo. We demonstrate the feasibility of our method in cohorts ranging in size from 2K (small) to 500K (large), and recruited across 2 to 10 collaborating institutions.

Introduction

Genome wide association studies (GWAS) are a popular approach to elucidate genetic architecture of human phenotypes. This design has led to the discovery of many novel loci underpinning a panoply of human traits (see Visscher et. al. for a recent review (1)). For traits driven by few variants with large effects, moderately sized cohorts have been sufficient to power discovery. However, the GWAS framework demands increasingly larger cohort sizes as the complexity of the trait grows. To achieve required statistical power today, large, multi-institutional consortia are assembled under a common data sharing agreement. Meta-analysis (2; 3; 4) or a combination of meta- and mega-analysis (centralized analysis) (5; 6; 7; 8) constitute two major approaches to conducting GWAS.

Each approach offers merits and shortcomings. For mega-analysis, collecting all the data at every analysis core is not only expensive and time consuming but also creates a security vulnerability at each institution that hosts a copy of the data. Conversely, the meta-analysis approach eliminates the need for data replication, but is more limited in flexibility. In particular, (a) subtle differences in models, assumptions, and quality control (QC) can introduce biases in the results (9), (b) the shared data and summary statistics might be inadequate for some types of inference (e.g. individual level population structure control, conditional or joint analysis, etc.), and (c) parameter estimates can be unreliable for rare variants or from centers contributing small sample sizes because the asymptotic properties of maximum likelihood estimation theory may not hold (10).

In this manuscript we develop a method that interpolates between centralized and meta-analysis methods. Like meta-studies, our paradigm leverages data fragmented across multiple silos, where each silo communicates aggregate level information with a central hub to perform the GWAS. Unlike meta-studies, in our paradigm the parameter estimates are computed iteratively (figure 1). To contrast these two paradigms, we henceforth refer to the one-time communication setting of traditional meta-analysis as "Meta 1.0", and to our iterative setting as "Meta 2.0" (see Figure1).

We show that all steps of a canonical case-control GWAS can be implemented in our framework, including quality control (QC) filters and population structure control by regressing out the top principal components (PCs). In particular, typical meta-analysis assumptions on the distribution of participant attributes in each silo (e.g. case-control status, ancestry, etc.) can be relaxed while simultaneously attaining near-perfect parameter estimates. For GWAS, our pipeline leverages an algorithm for performing PCA by communicating the LD-matrices (Algorithm 1), and subsequently a method for performing generalized regression (most commonly, linear and logistic regression) by

iteratively solving a regularized regression at each silo. This general, iterative approach, known as Alternating Directions Method of Multipliers (11; 12), is guaranteed to converge (13), has successfully been applied to many problems (14; 15; 16), and for GWAS we show that moderately high accuracies can be achieved with a few iterations (also see (14)). We show that our pipeline is accurate, scalable, practical, and a significant improvement over the meta 1.0 approach.

Tangentially, our approach also confers an increased level of privacy to study participants recruited across multiple centers. This is an area of active research: recent concerns about data-sharing have lead to a growing body of literature and methods that offer distributed analysis. UNICORN (17) provides fast access to external controls matched to the population structure. Privacy-first GWAS methods usually fall into one of two camps: a) Homomorphic Encryption techniques like (18; 19) which let the analyst perform computation on encrypted data, or b) Differential Privacy techniques which let the analyst report results in a manner that obfuscates participant re-identification. Both families of methods are active areas of research but not the primary focus of this work.

Results

We implemented an end-to-end canonical GWAS pipeline in our meta 2.0 framework and compared performance to both a mega-analysis and meta-study 1.0. We evaluated performance on datasets constructed from 2,270 European individuals in the POPRES dataset (20). Since POPRES does not have phenotype information, we simulate a case-control study by assigning random effect sizes to 10 SNPs according to a model similar to that of (21). Based on the overall risk, half of the individuals were assigned case status and the remaining half were assigned to the control group. See ?? for further details. Figure (2A) shows a Manhattan plot of the P values for the simulated data after correcting for the first 5 PCs. Figures 2B,C) show the projection of cases and controls onto the first 4 PCs. No geographical structure is observable that separates cases from controls (i.e. cases and controls are well matched).

To evaluate our performance, we tackled two realistic scenarios. In scenario 1, we randomly split the data between 5 silos, simulating a fixed effect meta-study with a relatively small sample size per silo. In scenario 2, we split the data into only 2 silos based on case/control status, simulating a collaboration where cases and controls are recruited at different partner institutions. For each of these two scenarios, we compared performance using QC, PCA estimation, and PC-adjusted asso-

ciation analysis between our method and an equivalent centralized analysis. Lastly, to investigate the effects on extremely rare variants, we simulated a large, balanced, mega-cohort ($N = 500,000$) and spread the data evenly among 10 data silos (collaborating centers).

Step 1: Quality Control

The most important QC filters in a standard GWAS (call quality, missing-rate per individual, missing-rate per site, minor allele frequency, Hardy Weinberg equilibrium, and linkage disequilibrium) can be implemented efficiently under our paradigm. Call quality and missing-rate-per individual filters only require the communication of a threshold from the hub. Further, communicating (`#missing`, `#heterozygous`, `#homozygous ref.`) for each locus at each silo allows for computation of minor allele frequency (MAF), missing-per-loci and Hardy Weinberg equilibrium (HWE) at the hub. These values, along with the corresponding thresholds can be communicated back to perform MAF filters, missing-per-loci filters and HWE filters. Linkage disequilibrium (LD) pruning can also be implemented analogously, after communicating the local LD matrices from each silo to the center.

Table 1 demonstrates the comparison between our decentralized implementation and the centralized version of the equivalent filters as implemented by PLINK (22; 23). For each scenario, the table also includes the results of filtering based on local information at each silo (i.e. meta-analysis 1.0 approach). Furthermore, unlike the meta-study approaches, under our paradigm, relatedness of subjects across silos can be detected using locality sensitive hashing (24). Since our dataset does not contain related individuals and the performance of this approach has been previously analyzed by He, et. al. (24), we omit this filter.

	Scenario 1			Scenario 2	
Filter	Centralized	Meta 2.0	Meta 1.0	Meta 2.0	Meta 1.0
Total	447,245	447,245 (1.00)	447,245 (1.00)	447,245 (1.00)	447,245 (1.00)
Missing < .05	387,938	387,938 (1.00)	361,101 (0.93)	387,938 (1.00)	380,274 (0.98)
MAF > .05	296,886	296,886 (1.00)	268,476 (0.90)	296,886 (1.00)	288,450 (0.97)
HWE $P > 10^{-10}$	295,179	295,179 (1.00)	268,172 (0.90)	295,179 (1.00)	287,620 (0.97)
LD $r^2 < 0.2$	67,300	67,300 (1.00)	38,044 (0.52)	67,300 (1.00)	55,436 (0.80)

Table 1: **QC Comparison:** The most important QC filters can be exactly implemented under a decentralized framework with no loss of accuracy. The table above contrasts our pipeline with both a centralized approach and a meta-analysis approach. For both scenarios under consideration results identical to the centralized approach can be achieved under the meta 2.0 framework. In this table, the number in parentheses is a measure of concordance (Jaccard index) between the set of unfiltered loci and the set of loci from the centralized pipeline. All centralized filters were implemented using PLINK(23; 22) (see ?? for details).

Step 2: Population Structure Control

Correcting for the first few principal components (PC) has been shown to be an effective method of population structure control (25; 26). In this section we compare the results of a centralized and decentralized PCA on the POPRES dataset (20). First the data was filtered (MAF > 0.05, HWE with P value > 10^{-10} , and LD of $r^2 < 0.2$ using sliding windows of size 50 and strides of 25 loci). The remaining 67,300 loci were normalized and algorithm 1 was used to perform PCA on the decentralized datasets. Figure 3 provides a visual comparison of the projection of samples onto the first two PCs in scenario 1. Projection onto higher PCs for both scenarios can be seen in Supplementary Figure ?? and Supplementary Figure ?? for scenario 1, 2 respectively.

Communicating the LD matrices (SNP-SNP covariance matrix) constitutes the leading communication cost in algorithm 1. The size of this communication scales linearly in the number of silos and quadratically in the number of SNPs included and does not directly depend on the number of individuals in the study. In our dataset of 67,300 post-QC SNPs, communicating the LD-matrix required a one-time 18GB transmission per silo (using 32 bit floats and without exploiting symmetry of the covariance matrix). The eigenvalue decomposition was performed using Implicitly Restarted Arnoldi Methods of ARPACK accessed through scipy (27; 28). On a single core machine, the decomposition for the top 5 vectors (using 15 Lanczos vectors) was performed in under 3 mins. While

for the dataset under study, this is a significant time increase compared to local SVD decomposition, this algorithm does not scale with the number of individuals and presents a more competitive alternative for large-cohort GWAS studies.

Step 3: Association

The final step of a typical GWAS pipeline is performing association analysis. This step often follows imputation of array data. Thanks to third-party imputation servers such as Michigan Imputation Server (29) and Sanger Imputation Server (30), fast, scalable, and reliable imputation is possible without centralizing the data. Under our decentralized paradigm, imputation would be performed by such a third-party service. Since the results are expected to be the same for centralized and decentralized data, in our pipeline we will omit this step and proceed to the association analysis.

We performed association analysis using logistic regression with the standardized value of each locus and the first 5 PCs as covariates. P values for each locus were computed using a likelihood ratio test (LRT) comparing the model containing only the first 5 PCs as covariates against the model that contains the SNP as well as the first 5 PCs. While the paradigm allows for computing P values using both Wald's test and LRT, LRT was chosen as it tends to have higher power than Wald test (see Supplementary Figure ??), particularly for rare variants in case-control studies (31). Additionally, LRT only requires each silo to communicate a likelihood value, where Wald test requires communication of the entire information matrix. Qualitatively, our results are similar for both tests (see **Supplement Figures** for all corresponding plots with Wald test).

In order to further probe our performance on very rare loci, we simulated an additional cohort with $N = 500,000$. To simulate the cohort, an allele was created by randomly combining two haplotypes with 5 normally distributed nuisance random variables. These nuisance parameters account for 20% of overall variance in the burden and each loci accounts for $\frac{25 \times r}{2 \times \text{MAC}}$ of the variance, where MAC is the minor allele count and r is a uniformly drawn random variable. Disease status was assigned to create a balanced study based on the overall burden.

Figure 2 A) shows a Manhattan plot for the P values of the POPRES cohort when the association is performed after centralizing the data. When the data is randomly distributed between the silos (scenario 1), a fixed-effect, meta-study approach can be taken to obtain parameter estimates and P values. Figure 4 A shows the concordance between parameter estimates of an inverse-variance fixed effect approach with those of a centralized study. Figure 4 B shows the corresponding plot for the P values. In this case, errors accumulate from, the difference in quality control, PC's included in

the regression and the uncertainty in the regression. In all cases, a cut off of 40 has been applied to $-\log_{10}(\text{P value})$. The meta-study approach appears to be slightly under-powered for highly significant loci. Figures 4 **C,D** compare the results of our fully decentralized pipeline where the regression parameters have been computed using an iterative approach. We observe a nearly perfect agreement between this decentralized approach and the centralized approach. The loss of power is further exaggerated when alleles under consideration have a low allele frequency. To probe this scenario we utilize our $N = 500,000$ simulated dataset, described earlier. Figure 5 **A** compares the loss in power (Centralized logP - Decentralized logP) between an inverse variance meta-analysis and our approach (meta 2.0). While meta-analysis leads to a better initial estimate in all cases, within the first 5 iterations our approach out-performs these methods. Within 20 iterations, our approach effectively matches the power of the centralized analysis. As expected, when the minor allele count increases, both approaches perform better.

Scenario 2 challenges regression based meta-studies as each silo exclusively contains cases or controls, hence, local regression parameters cannot be computed. As figures 4 **E,F** show, our pipeline remains robust to this local label imbalance.

The decentralized regression is performed using alternating direction method of multipliers (32). Unless otherwise is specified, for each locus 100 iterations of the algorithm was performed. Each iteration requires the silo to communicate a local estimate of a regularized regression, and receive an updated estimate of parameter values (i.e. a back and forward communication of size equal to the number of covariates). Figure 5 **B, C** show the improvement in accuracy of the coefficient estimates (compared to the centralized estimates) as a function of number of iterations (data from chromosome 17, scenario 1 and 2 respectively). As before, in both cases good accuracy can be achieved in tens of iterations. The convergence for significant and non-significant loci is similar for scenario 1. In scenario 2, while the convergence rates are similar, significant loci lag behind non-significant loci. This is attributed to initializing the algorithm with 0, which is a better estimate for non-significant loci. Supplementary Figure ?? shows a similar analysis as a function of iterations and MAF. The centralized regressions were performed using sklearn's LogisticRegression ($C = 10^4$)(33). In our experiments, a serialized implementation of this approach (i.e. all local updates performed in serial and the global update performed after local updates from all silos), took approximately $3.7\times$ and $1.5\times$ longer than the centralized approach for scenario 1, 2 respectively.

Discussion

The ability to perform full-fidelity association studies in a decentralized setting has logistical, security, and policy benefits when compared to centralizing the data. We believe this approach can lead to higher data security over the current paradigm where the data is often gathered by multiple institutions and subsequently centralized and reproduced at any institution approved for accessing the data. Our paradigm restricts the security and logistical risks to the few original data hosting centers. Securing these centers is cheaper, easier, and thereby more effective than securing every server where the accumulated data might be copied to. Furthermore, the reduction in the number of individuals with access to the raw data can reduce the privacy risks caused by human-error. Lastly, our algorithms only share aggregate level information, and unless the centers are case-only or control-only, they do not provide aggregate information by phenotype status. As a result, they act as a barrier against abuse and strictly improves on the privacy of sharing the raw data.

In contrast to meta-studies, which enjoy an increased level of the aforementioned benefits, our pipeline does not rely on large cohort asymptotic behavior at each silo, makes no assumptions regarding the distribution of phenotype status across silos, or uniformity of samples by population demography. While these parameters can effect the convergence rate of the association analysis, as demonstrated, under realistic scenarios (including complete separation of cases and controls) the convergence to relevant accuracies is fast. Furthermore, similar to centralized studies, our pipeline is flexible and can be extended to include new interaction effects, as well as other models (e.g. linear regression). The flexibility in choice of model, and weak requirements on the distribution of silos make our pipeline appealing for the future of GWAS.

These shortcomings in addition to concerns about data-sharing have lead to a growing body of literature and methods. UNICORN (17) provides fast access to external controls matched to the population structure - however it relies on knowledge about the population under study and is limited to control samples. Concerns about privacy of data sharing has brought about cryptographic GWAS approaches (18; 19). These methods allow for secure use of untrusted compute resources (e.g. cloud computation) but struggle for practicality and scalability. In particular, (18) is limited to simple computations and does not scale well. A more recent work (19) improves scalability and flexibility of analysis but still relies on a two step approach to limit the number of logistic regressions. These cryptographic approaches present a promising path forward but they are inefficient under a paradigm where the data is owned by each silo and the computations performed on the silos (and

the hub) are not malicious. As a result, the homomorphic encryption literature is only tangentially relevant to our work. On the other hand, concerns about leakage of information from the reported results have given rise to differential privacy-preserving algorithms in genetics (34). With the expanding cohort sizes, the potential for leakage of individual-level information from the results decreases. The last broad area of relevant work is the decentralized/batch-based general machine learning algorithms that have become an ever-more-active area of research in order to meet the challenges presented by large datasets. Our work is inspired by, and can be expanded upon based on insights from these approaches (14; 35; 36; 37; 38; 39; 40; 41).

We see many opportunities to expand the current paradigm and pipeline. In this work we assume a minimal privacy threat model and focus on efficiency. One extension to our methods and paradigm would be to adopt an honest-but-curious threat model for all parties and aim to prevent leakage of individual level information from the data owning silos to any other party involved. Since for all steps of the pipeline the central hub simply sums the results of all silos (or takes averages), when more than 2 silos contribute, multi-party, secure sum protocols (42; 43) can be used to decrease the overall chance of information leakage. An alternative approach would be to replace the pipeline with differential privacy preserving algorithms. Indeed, many of the current components have differential privacy preserving counter parts (see (44; 45)). Yet another angle for expansion is to work in the regime where the number of silos and the cohort size are similar.

Methods

Here we present the algorithms used to implement the fully decentralized GWAS pipeline.

Notations

Unless otherwise is specified, we adopt the notations from table 2. Throughout, superscripts enumerate the silos.

Data Filtering

A typical GWAS QC consists of many filters. The most important filters are easy to implement under the meta 2.0 framework. Missing-per-individual, quality-of-call filters and other individual level filters can be implemented as usual after a threshold is communicate from the central hub. After communicating the number of individuals per silo (n_i), MAF, HWE, and missing-per-snp

N	Total number of individuals	$n^{(i)}$	Number of individuals in silo i
K	Number of collaborating silos	G	ind $\times$ SNP Normalized genotype matrix
$G^{(i)}$	The genotype matrix at silo i	X	Matrix of covariates
Y	Vector of responses. i.e. phenotype		

Table 2: Table of symbols.

statistics for each loci can be computed using (# missing, # het., # homo. ref) per silo. Specifically, missing-per-snp and MAF statistics can be computed as follows:

$$\text{missing-per-snp} = \frac{1}{N} \sum_i^K \text{\#missing}^{(i)}$$

$$\text{MAF} = \frac{1}{2} \min \left\{ \frac{\sum_i^K \text{\#het}^{(i)} + 2\text{\#homo. ref}^{(i)}}{N - \sum_i^K \text{\#missing}^{(i)}}, 2 - \frac{\sum_i^K \text{\#het}^{(i)} + 2\text{\#homo. ref}^{(i)}}{N - \sum_i^K \text{\#missing}^{(i)}} \right\}$$

The HWE statistics is also a function of the total number of homozygous (ref and alt) and heterozygous individuals. For our experiments, we used the mid-p adjustment (46) as implemented by PLINK (22). After computing these statistics, filtering can be performed by simply communicating a threshold from the center to each silo.

The final filter we implemented is the LD-pruning filter. This filter is often used to choose sites for down-stream PCA analysis. In our implementation, we used a weighted averaging method similar to the one used for computing missing-per-snp statistics. However, whereas all the above filters require one back-and-forward communication of $\mathcal{O}(1)$ in size per locus, this filter requires a back-and-forward communication of order $\mathcal{O}(w^2)$ per stride, where w is the LD-window size and an stride is the distance the pruning window moves per iteration. As a result, detection of LD moderately far away is difficult without a heuristic. Furthermore, whereas previous filters can be computed in chunks, and the computation on chunks can be parallelized, unless window size and stride size are the same, such optimizations are not possible under our current implementation. Nevertheless, these implementation inefficiencies are solely to reproduce the results provided by PLINK and can be avoided by real-world applicable heuristics.

Finally, we note that, as demonstrated by He et. al., it is possible to identify and filter for relatedness under the meta 2.0 paradigm.

PCA

Ignoring population stratification can lead to false discoveries or a reduction in power (47; 48). Principal component analysis (PCA) is a common approach to reduce the adverse effects of stratification (47; 26). One common approach is to regress out the top PCs of the genotype matrix (26; 25). In this section, we discuss an implementation of such an approach.

Let G denote the normalized genotype matrix, with $G_{i,j}$ represents the z-transformed frequency of the i^{th} individual at the j^{th} SNP. $G = U\Sigma V^{\top}$ represents the singular value decomposition of G . Let $G^{(i)}$ represent the corresponding matrix at silo i with all $G^{(i)}$'s normalized using the global mean and standard deviation. Algorithm 1 describes the steps to perform a decentralized PCA. In this algorithm computing the full eigen-spectrum for $G^{\top}G$ can be prohibitively difficult. In real-world settings, PCA-based structure correction only relies on the top few eigen-components. Power iteration based methods, such as Lanczos algorithm, can be used to efficiently compute the top eigenvalue, eigenvector pairs. For this work we have used the implicitly Restarted Arnoldi Method (28) through the Scipy.sparse.linalg.eigsh interface (27).

Algorithm 1: Decentralized Principal Component Analysis

Data: $G^{(i)}, n^{(i)} \forall k = 0 \dots K - 1$; desired number of singular values: q

Result: Top q singular values (Σ_q), left and right singular vectors (U_q, V_q):

for $i \leftarrow 0$ **to** $K - 1$ **do in parallel**

$M^{(k)} \leftarrow G^{(k)\top} G^{(k)}$

Communicate $M^{(k)}$ to the center

end

$M \leftarrow \sum_{k=1}^{K-1} M^{(k)}$; // computed at the center

$V_q \Sigma_q V_q^{\top} \leftarrow \text{EignVal}(M, q)$; // Eigenvalue decomposition for top q EVs

Communicate Σ_q^{-1}, V_q back to all collaborating silos

for $i \leftarrow 0$ **to** $K - 1$ **do in parallel**

$U_q^{(i)} \leftarrow G^{(i)} V_q \Sigma_q^{-1}$

end

Algorithm 1 requires a one time communication of the gene-by-gene covariance matrix ($G^{(i)\top} G^{(i)}$) from each silo to the center as well as a backward communication of Σ_q^{-1}, V_q . Typically we are interested in the top few PC's, therefore, the backward communication to the silos is generally small; however, the covariance matrix can be relatively large ($\sim \frac{L^2}{2}$ where L is the number of SNPs passing

quality control for PCA). For 100k SNPs this would take an approximately 10GB communication per silo (using 32 bit floats and symmetry of the matrix). If the number of silos is very large, this communication can become overwhelming. In such a case, batch based approaches (41) or projection on the dimensions learned from a public dataset are viable alternatives.

Association

Regression serves as one of the most common approaches in performing association in a GWAS. Under the maximum likelihood estimation (MLE) paradigm, performing a regression can be reduced to solving the following optimization problem:

$$\beta^* = \arg \min_{\beta} \sum_i^K \ell(X^{(i)}, Y^{(i)}, \beta) \quad [1]$$

Where, $X^{(i)}, Y^{(i)}$ are the matrix of covariates and vector of outcomes at silo i respectively, and $\ell(X, Y, \beta)$ is the negative log-likelihood function. We use the alternating directions method of multipliers' (ADMM)(14) framework to separate this problem by silos. To gain intuition, we explore the approach of inverse-variance fixed-effect meta-studies, where the MLE of each silo is provided ($z^{(i)} = \arg \min_x \ell(X^{(i)}, Y^{(i)}, x)$) and the meta-study estimate takes a weighted average of the local estimates:

$$\beta_{FE} = \sum_i^K w^{(i)} z^{(i)} \quad [2]$$

where $w^{(i)}$ are a set of weights (inverse variance of the local estimate). With analogy to the fixed-effect approach the optimization problem of equation 1 can be written as:

$$\begin{aligned} \beta^* = \arg \min_{\beta} \sum_i^K \ell(X^{(i)}, Y^{(i)}, z^{(i)}) \\ \text{subject to: } z^{(i)} = \beta \quad \forall i \end{aligned} \quad [3]$$

The constraint in equation 3 can be imposed using an augmented Lagrange(49; 50):

$$L_{\rho}(\beta, \lambda, z^{(i)}) = \sum_i^K \ell(X^{(i)}, Y^{(i)}, z^{(i)}) + \lambda^{\top} (z^{(i)} - \beta) + \frac{\rho}{2} \|z^{(i)} - \beta\|_2^2 \quad [4]$$

Here, λ is an estimate for the usual Lagrange multiplier and ρ is constant. This leads to the iterative solution

$$z_{t+1}^{(i)} = \arg \min_z \ell(X^{(i)}, Y^{(i)}, z^{(i)}) + \lambda_t^{(i)\top} (z^{(i)} - \beta_t) + \frac{\rho}{2} \|z^{(i)} - \beta_t\|_2^2 \quad [5]$$

$$\beta_{t+1} = \frac{1}{K} \sum_i^K \left(z_{t+1}^{(i)} + \lambda_t^{(i)} \right) \quad [6]$$

$$\lambda_{t+1}^{(i)} = \lambda_t^{(i)} + z_{t+1}^{(i)} - \beta_{t+1}$$

In words, at each iteration, all silos update their local estimates ($z^{(i)}$) in parallel. These estimates are then communicated to the central hub where a consensus estimate is computed by averaging all local estimates. At the end of each iteration, this consensus parameter estimate is communicated back to each silo along with a silo specific dual variable $\lambda^{(i)}$.

Note that in equation 6, after the first iteration, the update for the consensus estimate simplifies to $\beta_{t+1} = \frac{1}{K} \sum_i^K z_{t+1}^{(i)}$. We contrast this equation to the fixed-effect, meta-study update of equation 2. Whereas fixed-effect updates use a weighted averaging that takes the variance of each local estimate (and therefore the sample size at each silo) into account, in our approach all the local estimates are simply averaged. The effects of sample size and the overall likelihood are taken into account when the local estimates are produced from the regularized regression of equation 5. For logistic regression, the likelihood function is given by $\ell(X, Y, \beta) = \sum_{i=1}^N Y_i \log(\hat{Y}_i) + (1 - Y_i) \log(1 - \hat{Y}_i)$ where $\hat{Y}_i$ is the probability of a positive outcome for the i^{th} sample (i.e. $\hat{Y}_i = \frac{1}{1 + \exp(-X_i \beta)}$). We used L-BFGS-B (51) with warm-start for the PCA estimates and default scipy(27) parameters to perform the update in equation 5.

P-value computation

P values for the significance of each parameter can be constructed using both the likelihood ratio test (LRT) and Wald test. In this work we focus on LRT since it is both more efficient to compute and more powerful(31). To compute the P-value for a loci, each silo simply computes the local log-likelihood using the most recent global estimate. The local log-likelihood is then communicated and summed at the central hub to compute the overall log-likelihood. This global value can then be compared to the global log-likelihood of the model without the locus (which only needs to be computed once for the entire study).

To compute Wald test P values, the global parameter can be used to construct the local Fisher information matrix at each silo ($\mathcal{I}^{(i)}$). The overall information matrix can be assembled as the

sum of these local matrices ($\mathcal{I} = \sum_i^K \mathcal{I}^{(i)}$). The Wald test P values can be computed using the overall information matrix at the central hub. Unfortunately, for a model with p parameters, this computation requires a communication of volume $\mathcal{O}(p^2)$ per locus.

Efficient implementation

For simplicity, we have presented all steps in the context of a single locus. In many cases, the network latency can be higher than a single computation. Fortunately, for most computations all loci are considered independently, therefore, it is possible to compute the updates on many loci and report them in groups. For instance, to compute the global allele frequencies, a vector of a few thousand local frequencies can be computed and reported at a time. Similarly, the ADMM updates on multiple loci can be computed and reported as a group. The silos can move on to compute the next block while awaiting the global updates.

Code and Data Availability: POPRES data set is available through dbGaP. The preprocessing scripts and all other codes are available at <https://github.com/apoursh/caringWithoutSharing>.

Acknowledgements: AP and CDB are supported by Genome Sequencing Project grant 1U01HG009080, Chan Zuckerberg Biohub. CDB and SP are supported by Clinical Genome Research Project grant U41HG009649. AP, CDB, SP are supported by CERSI grant U01FD004979.

References

- 1 P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, “10 years of gwas discovery: biology, function, and translation,” *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, 2017.
- 2 A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, *et al.*, “Defining the role of common variation in the genomic and biological architecture of adult human height,” *Nature genetics*, vol. 46, no. 11, p. 1173, 2014.
- 3 A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, *et al.*, “Genetic studies of body mass index yield new insights for obesity biology,” *Nature*, vol. 518, no. 7538, p. 197, 2015.
- 4 M. Nikpay, A. Goel, H.-H. Won, L. M. Hall, C. Willenborg, S. Kanoni, D. Saleheen, T. Kyriakou, C. P. Nelson, J. C. Hopewell, *et al.*, “A comprehensive 1000 genomes–based genome-wide

association meta-analysis of coronary artery disease,” *Nature genetics*, vol. 47, no. 10, p. 1121, 2015.

5 R. J. Strawbridge, J. Dupuis, I. Prokopenko, A. Barker, E. Ahlqvist, D. Rybin, J. R. Petrie, M. E. Travers, N. Bouatia-Naji, A. S. Dimas, *et al.*, “Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes,” *Diabetes*, vol. 60, no. 10, pp. 2624–2634, 2011.

6 B. F. Voight, L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, R. P. Welch, E. Zeggini, C. Huth, Y. S. Aulchenko, G. Thorleifsson, *et al.*, “Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis,” *Nature genetics*, vol. 42, no. 7, p. 579, 2010.

7 S. Ripke, N. R. Wray, C. M. Lewis, S. P. Hamilton, M. M. Weissman, G. Breen, E. M. Byrne, D. H. Blackwood, D. I. Boomsma, S. Cichon, *et al.*, “A mega-analysis of genome-wide association studies for major depressive disorder,” *Molecular psychiatry*, vol. 18, no. 4, p. 497, 2013.

8 M. A. Rivas, M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burt, *et al.*, “Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease,” *Nature genetics*, vol. 43, no. 11, p. 1066, 2011.

9 T. Greco, A. Zangrillo, G. Biondi-Zoccai, and G. Landoni, “Meta-analysis: pitfalls and hints,” *Heart, lung and vessels*, vol. 5, no. 4, p. 219, 2013.

10 D. C. Hoaglin, “We know less than we should about methods of meta-analysis,” *Research synthesis methods*, vol. 6, no. 3, pp. 287–289, 2015.

11 R. Glowinski and A. Marroco, “Sur l’approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité d’une classe de problèmes de dirichlet non linéaires,” *Revue française d’automatique, informatique, recherche opérationnelle. Analyse numérique*, vol. 9, no. R2, pp. 41–76, 1975.

12 D. Gabay and B. Mercier, *A dual algorithm for the solution of non linear variational problems via finite element approximation*. Institut de recherche d’informatique et d’automatique, 1975.

13 W. Deng and W. Yin, “On the global and linear convergence of the generalized alternating direction method of multipliers,” *Journal of Scientific Computing*, vol. 66, no. 3, pp. 889–916, 2016.

- 14 S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
- 15 B. Narasimhan, D. L. Rubin, S. M. Gross, M. Bendersky, and P. W. Lavori, “Software for distributed computation on medical databases: A demonstration project,” *Journal of Statistical Software*, vol. 77, no. i13, 2017.
- 16 D. Hallac, J. Leskovec, and S. Boyd, “Network lasso: Clustering and optimization in large graphs,” in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 387–396, ACM, 2015.
- 17 C. A. Bodea, B. M. Neale, S. Ripke, M. Barclay, L. Peyrin-Biroulet, M. Chamaillard, J.-F. Colombel, M. Cottone, A. Croft, R. D’Incà, *et al.*, “A method to exploit the structure of genetic ancestry space to enhance case-control studies,” *The American Journal of Human Genetics*, vol. 98, no. 5, pp. 857–868, 2016.
- 18 K. A. Jagadeesh, D. J. Wu, J. A. Birgmeier, D. Boneh, and G. Bejerano, “Deriving genomic diagnoses without revealing patient genomes,” *Science*, vol. 357, no. 6352, pp. 692–695, 2017.
- 19 H. Cho, D. J. Wu, and B. Berger, “Secure genome-wide association analysis using multiparty computation,” *Nature biotechnology*, vol. 36, no. 6, p. 547, 2018.
- 20 M. R. Nelson, K. Bryc, K. S. King, A. Indap, A. R. Boyko, J. Novembre, L. P. Briley, Y. Maruyama, D. M. Waterworth, G. Waeber, *et al.*, “The population reference sample, popres: a resource for population, disease, and pharmacological genetics research,” *The American Journal of Human Genetics*, vol. 83, no. 3, pp. 347–358, 2008.
- 21 A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny, “Human demographic history impacts genetic risk prediction across diverse populations,” *The American Journal of Human Genetics*, vol. 100, no. 4, pp. 635–649, 2017.
- 22 S. Purcell and C. Chang, “Plink 1.9,” Available at: <https://www.cog-genomics.org/plink2>, 2015.

- 23 C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, “Second-generation plink: rising to the challenge of larger and richer datasets,” *Gigascience*, vol. 4, no. 1, p. 7, 2015.
- 24 D. He, N. A. Furlotte, F. Hormozdiari, J. W. J. Joo, A. Wadia, R. Ostrovsky, A. Sahai, and E. Eskin, “Identifying genetic relatives without compromising privacy,” *Genome research*, vol. 24, no. 4, pp. 664–672, 2014.
- 25 A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, “Principal components analysis corrects for stratification in genome-wide association studies,” *Nature genetics*, vol. 38, no. 8, p. 904, 2006.
- 26 A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, “New approaches to population stratification in genome-wide association studies,” *Nature Reviews Genetics*, vol. 11, no. 7, p. 459, 2010.
- 27 E. Jones, T. Oliphant, P. Peterson, *et al.*, “SciPy: Open source scientific tools for Python,” 2001–.
- 28 R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK users’ guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, vol. 6. Siam, 1998.
- 29 *Michigan Imputation Server*, 2018. Available at: <https://imputationserver.sph.umich.edu/>.
- 30 *Sanger Imputation Server*, 2018. Available at: <https://imputation.sanger.ac.uk/>.
- 31 G. Xing, C.-Y. Lin, S. P. Wooding, and C. Xing, “Blindly using wald’s test can miss rare disease-causal variants in case-control association studies,” *Annals of human genetics*, vol. 76, no. 2, pp. 168–177, 2012.
- 32 S. Boyd, “Alternating direction method of multipliers,” in *Talk at NIPS Workshop on Optimization and Machine Learning*, 2011.
- 33 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

- 34 S. Simmons, C. Sahinalp, and B. Berger, “Enabling privacy-preserving gwass in heterogeneous human populations,” *Cell systems*, vol. 3, no. 1, pp. 54–61, 2016.
- 35 P. Lubell-Doughtie and J. Sondag, “Practical distributed classification using the alternating direction method of multipliers algorithm,” in *Big Data, 2013 IEEE International Conference on*, pp. 773–776, IEEE, 2013.
- 36 B. Recht, C. Re, S. Wright, and F. Niu, “Hogwild: A lock-free approach to parallelizing stochastic gradient descent,” in *Advances in neural information processing systems*, pp. 693–701, 2011.
- 37 S. Gopal and Y. Yang, “Distributed training of large-scale logistic models,” in *ICML (2)*, pp. 289–297, 2013.
- 38 J. Chen, R. Monga, S. Bengio, and R. Jozefowicz, “Revisiting distributed synchronous sgd,” *arXiv preprint arXiv:1604.00981*, 2016.
- 39 O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, “Optimal distributed online prediction using mini-batches,” *Journal of Machine Learning Research*, vol. 13, no. Jan, pp. 165–202, 2012.
- 40 G. Mateos, J. A. Bazerque, and G. B. Giannakis, “Distributed sparse linear regression,” *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5262–5276, 2010.
- 41 N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan, “Averaging stochastic gradient descent on riemannian manifolds,” *arXiv preprint arXiv:1802.09128*, 2018.
- 42 R. Sheikh, B. Kumar, and D. K. Mishra, “A distributed k-secure sum protocol for secure multi-party computations,” *arXiv preprint arXiv:1003.4071*, 2010.
- 43 C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, “Tools for privacy preserving distributed data mining,” *ACM Sigkdd Explorations Newsletter*, vol. 4, no. 2, pp. 28–34, 2002.
- 44 C. Dwork, K. Talwar, A. Thakurta, and L. Zhang, “Analyze gauss: optimal bounds for privacy-preserving principal component analysis,” in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 11–20, ACM, 2014.
- 45 T. Zhang and Q. Zhu, “Dynamic differential privacy for admm-based distributed classification learning,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 172–187, 2017.

- 443 46 J. Graffelman and V. Moreno, “The mid p-value in exact tests for hardy-weinberg equilibrium,”
444 *Statistical applications in genetics and molecular biology*, vol. 12, no. 4, pp. 433–448, 2013.
- 445 47 C. Tian, P. K. Gregersen, and M. F. Seldin, “Accounting for ancestry: population substructure
446 and genome-wide association studies,” *Human molecular genetics*, vol. 17, no. R2, pp. R143–
447 R150, 2008.
- 448 48 C. D. Campbell, E. L. Ogburn, K. L. Lunetta, H. N. Lyon, M. L. Freedman, L. C. Groop,
449 D. Altshuler, K. G. Ardlie, and J. N. Hirschhorn, “Demonstrating stratification in a european
450 american population,” *Nature genetics*, vol. 37, no. 8, p. 868, 2005.
- 451 49 M. R. Hestenes, “Multiplier and gradient methods,” *Journal of optimization theory and applica-*
452 *tions*, vol. 4, no. 5, pp. 303–320, 1969.
- 453 50 M. J. Powell, “A method for nonlinear constraints in minimization problems,” *Optimization*,
454 pp. 283–298, 1969.
- 455 51 R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, “A limited memory algorithm for bound constrained
456 optimization,” *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

Figure 1: **Comparison of data-flows** In the mega-study (centralized) framework, the data is gathered and analyzed at a single analysis center. For a typical meta-analysis approach (meta 1.0), summary statistics from previously analyzed datasets in collected and analyzed at the center. In our proposed framework (meta 2.0), the data is analyzed at each silo and a summary statistics is computed and reported to the center. Consequently, a global consensus is achieved at the center and reported back to each silo. The consensus estimates can then be used to improve the local estimates at each silo and the process can be iterated

Figure 2: **Simulated dataset.** **A)** Manhattan plot of the simulated data. The P value are compute after centralizing the data and correcting for the first 5 PCs. The x-axis is SNP position and the y-axis is the significance on a $-\log_{10}$ scale. The dashed line marks the threshold for genome-wide significance ($P = .01$ after Bonferroni correction) ($n = 2,270$). **B)** Projection of cases and controls onto the first two PCs. **C)** Projection of cases and controls onto the 3rd and 4th PCs

Figure 3: **Principal components analysis comparison.** Projection of POPRES individuals onto the first two principal components. **A)** Projection using singular value decomposition with centralized data **B)** Projection using singular value decomposition on decentralized data of scenario 1. The results are similar for scenario 2 and therefore have been omitted. See Supplementary Figure ??, and Supplementary Figure ?? for projections onto higher PCs. Conventional meta-analyses cannot control for population structure using PCA of the global dataset

Figure 4: **Association analysis.** Concordance of effect sizes and $-\log_{10}(\text{P values})$ between centralized, meta 1.0, and meta 2.0 approach. In all figures a P value cut-off of 40 has been used. **A, B)** Concordance of a fixed effect approach with a centralized analysis in scenario 1. **C, D)** Concordance of our approach (100 ADMM iterations) with the centralized approach in scenario. **E,F)** Concordances between our method and a centralized approach in scenario 2 (100 iterations). In both scenarios our approach leads to results identical to a centralized analysis

Figure 5: **Convergence by iteration.** For all plots, the x-axis shows the number of iterations of the algorithm (relevant only for our approach). **A)** Convergence of our approach for alleles with very low minor allele frequencies (MAF) in a large simulated dataset ($N = 500,000$). The y-axis shows the difference in $\log_{10}P$ value between a centralized study and one of the two decentralized approaches (100 draws). The minor allele frequencies correspond to minor allele counts of 50, 100, 200, 300. **B)** Convergence of the estimated coefficient to the centralized value for our approach and the inverse-variance FE approach, in scenario 1. **C)** Convergence of the estimated coefficient to the centralized value for our approach in scenario 2.