**Title:** High-throughput experiment design for consortium-driven genome-wide association meta-analysis

**Author:** Guo-Bo Chen

**Affiliation:** Evergreen Landscape & Architecture Studio, Hangzhou, Zhejiang Province, China

**Running title:** high-throughput GWAMA

**Keywords:** meta-analysis; GWAS; experiment design; complex traits; summary statistics

**Postal address:** Xixi Road 562, Hangzhou, Zhejiang Province, 310026, China

**Email:** [chenguobo@gmail.com](mailto:chenguobo@gmail.com)

**Phone:** +86.158.6718.1509

**Abstract**

Given the success of genome-wide association studies (GWAS), one of its consequences is GWAS meta-analysis (GWAMA). A consortium-driven GWAMA has been one of the tools for gene discoveries and mining biological pathways. However, GWAMA is plagued for its logistic burden and subject to quality of summary statistics. In this study, based on **o**pen genome-wide **a**ssociation algori**th**m (OATH), an algorithm proposed recently, a novel experiment design of GWAMA is demonstrated. The new design of GWAMA is advanced in i) ***sufficiency***: the central hub of the GWAMA is able to evaluate all possible outcomes of a locus; ii) ***efficiency***: the central hub can generate customized summary statistics without additional logistic burden; iii) ***scalability***: the central hub can simultaneously conduct GWAMA for more than one trait of interest without additional cost. Compared with the conventional experimental design for GWAMA, the proposed design is typically high-throughput and should be useful to facilitate GWAMA in an even larger scale.

**Introduction**

One of the conventional utilities of meta-analysis is to synthesis the evidence distributed in literature. For instance, if we are interested in the heritability of inflammatory bowel disease, we can search the possible database for reported summary statistics relevant to their association; then synthesis them and draw conclusion about the heritability of the inflammatory bowel disease (Chen *et al.* 2014). Without loss of generality, the literature-driven meta-analysis can be considered as a ***retrospective*** study, often an irreversible process, under which hardly can the summary statistics be customised as inquired by a meta-analysis conductor.

Given the success of genome-wide association studies (GWASs), genome-wide association meta-analysis (GWAMA) becomes possible, and often leads to unrivalled sample size involving many cohorts, typically at a magnitude of now, such as **G**enetic **I**nvestigate of **An**thropometric **T**raits Consortium (GIANT, <http://www.broadinstitute.org/collaboration/giant/index.php/Main_Page>) (Lango Allen *et al.* 2010), and **S**ocial **S**cience **G**enetic **A**ssociation **C**onsortium (SSGAC, <http://ssgac.org/Home.php>) (Okbay *et al.* 2016). Compared with literature-searching/retrospective meta-analysis, a consortium-driven GWAMA resembles a ***prospective*** study. A consortium-driven GWAMA has a central centre (hub), which commands a specific analysis, and cohorts (nodes) provide the summary statistics as asked by the central hub. In other words, in a consortium-driven GWAMA the summary statistics are generated as inquired, a much active mode (otherwise specified, in the text below a GWAMA study refers to a consortium-driven GWAMA in this study). Theory and practice have already indicated that GWAMA is advanced in such as data unification (de Bakker *et al.* 2008; Winkler *et al.* 2014) and across-cohort quality control (QC) (Chen *et al.* 2016). However, under the current experiment design, GWAMA has been undermined and does not satisfy the principles:

I) ***Sufficiency***. The principle of sufficiency can be defined as a ratio between the realized outcome and all possible outcomes for a locus. For example, if a GWAMA includes cohorts, each of which has covariates, then each cohort can give possible linear regression models for a locus. For , possible models, and it sums to possible results in GWAMA for this locus. In contrast, is the realized outcomes in a real GWAMA. Under the current GWAMA design, a central hub often receives only one of the thirty-two results, given , for each cohort, and , leading to sufficiency as low as . Given and , . So the principle of sufficiency is poorly satisfied under the current GWAMA experiment design.

II) ***Efficiency***. The principle of efficiency is defined as a metrics for quantifying logistic cost, in particular for time cost and for storage cost. As in GWAMA the logistic burden is not ignorable, if a cohort is willing to further provide another round of summary statistics generated as requested by the central hub – which will satisfies sufficiency for GWAMA, it may lead to a prolonged process, an inflation for . If each cohort would like to provide summary statistics under all possible models, the cost in storage is proportional to , which is and given and , respectively. Neither time cost nor space storage makes the principle of efficiency managed under the current GWAMA design, especially when the principle of sufficiency should be met.

III) ***Scalability***. Scalability is defined as a metrics that is relevant to . In GWAMA, it is often that a pair of phenotypes switches their roles as the target trait and a covariate, such as height and body mass index (BMI). When height is the target trait, BMI is likely to be the covariate, and vice versa. So, for height GWAMA, , and for BMI GWAMA, . Without loss of generality, if a pair of traits are considered, under the current GWAMA design, the total cost , the scalability metrics is linear to the number of traits of question in GWAMA. With the increased number of traits of interest in GWAMA, if the realized can be smaller than the summation of the each single , then the principle of scalability is satisfied. So, for a well-designed GWAMA, the total cost, both time and storage, should be less than proportional to the traits in question.

All these three principles that underpowered in a GWAMA will be improved in a high-throughput design of GWAMA. In this study, a conventional GWAMA will be considered as a prospective design by using **o**pen genome-wide **a**ssociation study algori**th**m (**OATH**) (Niu and Chen 2016). The goal is to satisfy the principle of sufficiency, maximal to be 1 when , of a central hub in GWAMA, and enables high-throughput GWAMA design. For clarity, H-GWAMA refers to the high-through consortium-driven GWAMA as proposed, and C-GWAMA refers to the conventional consortium-driven GWAMA.

**Methods**

In general, C-GWAMA involves two steps: step 1) each cohort generates GWAS summary statistics and sends to the central hub; step 2) the central hub run meta-analysis for the received summary statistics. Both steps are modified in H-GWAMA: in step 1, a cohort , given cohorts involved in H-GWAMA, generates naïve summary statistics (NSS) and sends to the central hub; in step 2, the central hub synthesises cohort-level NSS for meta-analysis and conducts meta-analysis. A schematic comparison between the design of the C-GWAMA and the H-GWAMA is illustrated in **Figure 1**. In the text below C-GWAMA and H-GWAMA will be compared step by step, and their sufficiency and efficiency will be compared (**Table 1**). As a proof-of-principle demonstration, a single-locus additive model for a quantitative trait is considered.

**C-GWAMA Step 1:** **a cohort** **generates summary statistics for a locus**

In C-GWAMA, the cohort runs a genome-wide association model that has covariates, written as

[1]

in which is the phenotype of interest, is the coding for the target locus, is the covariate, and is the residual. is the grand mean, is the partial regression coefficient of the marker, and is the partial regression coefficient for the covariate. encodes the biallelic locus, which has alleles and of frequency and , respectively; the variance of the locus is . Often, it is upon the cohort analyst to decide what should be included in , and it often includes such as eigenvectors and other relevant covariates – some covariates may be the trait of interest to another C-GWAMA in the consortium. The cohort analyst sends GWAS summary statistics, such as and , to the central hub (**Figure 1A**).

Without loss of generality, because one round of communication is required between the central hub and the cohort. If the central hub asks summary statistics generated by an alternative model, it inflates . because cohorts send the summary statistic unit ) to the central hub. So, .

**H-GWAMA step 1: a cohort generates naïve summary statistics for a locus**

Rather than implementing Eq 1, in H-GWAMA a cohort analyst conducts single-marker regressions (**Figure 1B**). The trait and the covariates included in Eq 1 have its simple linear regression implemented against the marker

[2]

in which is the simple linear regression for the target trait, and is a simple linear regression for a covariate. As each regression in the regression array is a simple linear regression, we call their regression coefficients naïve summary statistics (NSS, denoted as ). For the target marker, it generates summary statistics .

The sampling variance of the locus should be provided by each cohort. Under Hardy-Weinberg equilibrium, so can be approximated by if is not provided as allele frequency is provided as summary statistics. As only one round of communication is required between the central hub and the cohorts, so H-GWAMA is equal to C-GWAMA in time cost . . If each cohort has 5 covariates, given 10 cohorts included but only for C-GWAMA. However, this disadvantage in will give H-GWAMA huge advantage in satisfying the principle of sufficiency, as described below.

In addition, the cohort analyst should provide a covariance matrix, which is actually a correlation matrix when the target trait and the covariates are standardized

[3]

All these information provided are based on summary statistics, and should not compromise the privacy for each cohort.

**C-GWAMA step 2:** meta-analysis based on GWAS summary statistics

The central hub synthesises the summary statistics, which are generated via Eq 1, provided by each cohort together, the effect of the locus can be estimated

in which .

In addition, can be equivalently estimated in generalized least square (GLS) regression

[4]

in which is a vector for the effects of the locus received, is the intercept of the GLS model and provides the GWAMA estimate of the locus effect, is a diagonal matrix, which has its corresponding element been adjusted by (**Figure 1A**). When there are overlapping samples between cohorts, the summary statistics will be correlated and lead to inflated type I error rate. One way to correct for overlapping sample is to fill the corresponding element in with the correlation estimated for the pair of GWAS summary statistics (Chen *et al.* 2016).

The lacking of flexibility is obviously for C-GWAMA because it only conducts GLS under the received summary statistics, the sufficiency metric is as low as . So, the principle of sufficiency is certainly undermined, and the curse of “garbagy in garbagy out” applies if any cohorts choose a controversial model, such as including a heritable covariate as discussed (Aschard *et al.* 2015).

**H-GWAMA step 2: meta-analysis based on naïve summary statistics**

We still use GLS regressions to estimate , but now each element at the left side in Eq 4 is estimated via OATH (**Figure 1B**). Rather than asking each cohort to provide required summary statistics, using OATH the central hub can synthesize a multiple regression that has any combination of covariates [see Appendix for its framework, and its technique details and demonstration may be found in (Niu and Chen, 2016)]. Given the and , there are ways to estimate the locus effect (**Figure 1B**). Of note, under C-GWAMA the customization of only possible given the individual-level data; in other words, if the central hub of a C-GWAMA wants to have those customized results, it will inflate due to increased communication rounds between the hub and the cohorts, and also inflate required to store the received customized summary statistics. As mentioned above, given cohorts and covariates included in each cohort, is 320, and for ; in contrast, , for , and , for , under H-GWAMA.

Using OATH, the vector at the left side of the GLS model becomes , and the model can be written as

[5]

in which is a vector containing the synthesized effect via OATH (**Figure 1B**). In particular, if every element in are estimated via the same model, say (estimate locus effect with adjustment for BMI and eigenvector 1), we call it homogeneous GWAMA GLS model; if any not all elements are estimated under the same model, we call it heterogeneous GWAMA models. Given cohorts and covariates included in each cohort, there are forms for Eq 5; given , there are for Eq 5. As long as the computational facility permits, H-GWAMA can evaluate all possible forms for Eq 5 and fully satisfy the principle of sufficiency.

It is obviously that in the received NSS (Eq 2) a target trait and a covariate is actually equivalent. In H-GWAMA, a covariate, for example BMI, can be switched to the position as a target trait, and the target trait, for example height, can be switched to the position as a covariate. Under this equivalency, we conduct two traits in H-GWAMA, while and will not be doubled. Then the principle of scalability can be satisfied. In contrast, in conventional GWAS, the total is likely doubled if the central hub is going to conduct GWAMA on a pair of traits.

**Simulation results**

H-GWAMA has never been considered in practice, so we only implemented a proof-of-principle simulation to demonstrate its application. 10 cohorts of the same genetic origin were simulated, and each cohort had 10,000 loci, which had their reference allele frequencies between 0.1~0.5. Sample size for each cohort was sampled from a uniform distribution between 500 and 1,000. No genetic effects were assigned – heritability was zero. For each cohort, NSS were generated for the target phenotype and for each of the five eigenvectors that were associated to the top five largest eigenvalues. With or without adjustment for any of the five covariates, it totalled to models for each locus. In H-GWAMA, each cohort only need to provides six naïve summary statistics for each locus and the matrix to the central hub. Of note, matrix is same for each locus. So, .

Although OATH could synthesize NSS into any of the 32 models possible using OATH and lead to GLS models for each locus in meta-analysis, to make it computationally realistic only the homogenous GLS models was considered. It was reduced to 32 GLS models only for each locus in meta-analysis.

As a comparison to H-GWAMA, C-GWAMA was mimicked to implement these 32 possible models. Every cohort conducted 32 GWAS models for each locus using PLINK (Purcell *et al.* 2007), and these 32 GWAS summary statistics were sent to the central hub. So, , which was five times larger than .

As OATH could use NSS to synthesize partial regression coefficients as those observed in individual-level data, the 32 homogenous GWAMA results were expected to be identical between H-GWAMA and C-GWAMA. As illustrated in **Figure 2**, the meta-analysis results were identical between H-GWAMA and C-GWAMA for the estimated effects and z-scores for 10,000 loci across 32 GLS models (**Figure 2**). Furthermore, as expected too, the type I error rate was well controlled at for each of the 32 models (**Figure 3**). Of note, to run the same multiple regression models, say a linear model with five covariates, as OATH runs on summary statistics, it was much faster than conventional linear regression that is based on the individual-level data.

**Discussion**

GWAMA has been one of the major tools for gene-discovery especially given the success of GWAS. However, it also suffers from low efficiency and logistic issues in its implementation. In the proposed experimental design, GWAMA can be significantly improved. In particular, the proposed H-GWAMA can increase **sufficiency.** H-GWAMA could completely satisfy the principle of sufficiency, while C-GWAMA was a special case to H-GWAMA. Without OATH, even though C-GWAMA wants to meet the principle of sufficiency, increasing the sufficiency from to 1, its and will increase dramatically. In contrast, H-GWAMA keeps and almost unchanged given realization of the principle of sufficiency for GWAMA.

If we treat *p*-value as a statistic for decision-making, we are likely to have its distribution conditional on all possible summary statistics that can be generated in GWAMA. In theory, H-GWAMA is able to access all possible models, towards the sufficiency of 1. When the principle of sufficiency is satisfied, H-GWAMA may provide a realization for a prior distribution for *p*-values under the Bayesian doctrine (Bayes 1763). This prior is conditional on all covariates that are under consideration by each cohort.

Within a trait-specific consortium, new research interest is often raised. For example, for GIANT, height and body mass index (BMI) are often equally interested and switched their roles as the target trait or a covariate. Under the H-GWAMA, if both height and BMI are included in NSS, H-GWAMA can directly conduct two traits GWAMA simultaneously without adding additional workload for the cohorts. If BMI is not included as a covariate at the first place, when each cohort provide NSS for BMI the central hub can easily build multiple regression model using OATH. It is another way to reduce the cost in time and space.

One possible challenges of the proposed H-GWAMA is that the central hub is demanded to have more computational resource if the principle of sufficiency wants to be guaranteed. For example, as showed in simulation, the all possible results for a locus is , and it is rocketed to results, in which is the number of loci, a loading that is far beyond the computation capability. Nevertheless, as promised by quantum computing (Monz *et al.* 2016), the computational capability will bring out the possibility of the proposed design for H-GWAMA and bring out more reliable results.

**Conflict of interest**

The author declares no conflict of interest.

**Acknowledgements**

No specific funding was received for this work.

**References**

Aschard, H., B. J. Vilhjalmsson, A. D. Joshi, A. L. Price, and P. Kraft, 2015 Adjusting for Heritable Covariates Can Bias Effect Estimates in Genome-Wide Association Studies. Am. J. Hum. Genet. 96: 329–339.

de Bakker, P. I. W., M. A. R. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri *et al.*, 2008 Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. 17: R122-8.

Bayes, T., 1763 An Essay towards solving a problem in the doctrine of chances. Biometrika 45: 370–418.

Chen, G.-B., S. H. Lee, M. J. A. Brion, G. W. Montgomery, N. R. Wray *et al.*, 2014 Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. Hum. Mol. Genet. 23: 4710–4720.

Chen, G.-B., S. H. Lee, M. R. Robinson, M. Trzaskowski, Z. Zhu *et al.*, 2016 Across-cohort QC analyses of GWAS summary statistics from complex traits. Eur. J. Hum. Genet. doi:10.1038/ejhg.2016.106.

Fisher, R. A., 1921 On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. London 222: 309–368.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon *et al.*, 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–8.

Monz, T., D. Nigg, E. A. Martinez, M. F. Brandl, P. Schindler *et al.*, 2016 Realization of a scalable Shor algorithm. Science 351: 1068–1070.

Niu, Y.-F., and G.-B. Chen, 2016 Reproducing Genome-wide Association Studies Using Naïve Summary Statistics. G3 Submitted.

Okbay, A., J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers *et al.*, 2016 Genome-wide association study identifies 74 loci associated with educational attainment. Nature 533: 539–542.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–75.

Winkler, T. W., F. R. Day, D. C. Croteau-Chonka, A. R. Wood, A. E. Locke *et al.*, 2014 Quality control and conduct of genome-wide association meta-analyses. Nat. Protoc. 9: 1192–212.

**Figure legends**

**Figure 1 Schematic illustration for the C-GWAMA and the proposed T-GWAMA.** A) Under the C-GWAMA, each cohort sends GWAS summary statistics, which are generated under an *ad-hoc* model, to the central hub, and the central hub runs meta-analysis using generalized least square regression. B) Under the proposed H-GWAMA, each cohort sends naïve summary statistics to the central hub; using **o**pen genome-wide **a**ssociation algori**th**m (OATH) the central hub can synthesize the summary statistics most proper for GWAMA; and the central hub uses generalized least square regression to run the GWAMA. The left side of the dashed line represents step 1, and the right side of the dashed line represent step 2. For more details, please refer to the main text.

**Figure 2 The comparison for H-GWAMA and C-GWAMA for 32 homogenous meta-analysis models.** Only 1 of the 100 simulation replications was used. The grey squares represent covariates, and with or without filling indicates the inclusion or exclusion of the covariate for estimating summary statistics that are used in GSL meta-analysis. Each panel showed consistency for the estimated effects in each GSL model. The x-axis represents the GSL estimate based on plink exact models, and the y-axis represents the GSL estimate based on OATH models. The red-points panels are for the estimated beta effects, and the blue-points panels are for z scores.

**Figure 3 Type I error rates (alpha=0.05) for 32 GWAMA models.** The five squares below each bar represent five covariates, from top to bottom indicating eigenvector 1 to eigenvector 5. A filled squared means its corresponding covariate is included.

**Table 1 Comparison between H-GWAMA and C-GWAMA for the analysis of a single locus**

|  |  |  |
| --- | --- | --- |
|  | H-GWAMA | C-GWAMA |
| Sufficiency |  |  |
| Efficiency | and | and |
| Scalability | and |  |

**Notes:** Assuming there are cohorts involved in GWAMA and each cohort has covariates in the linear model for the analysis of the target trait. is the upper bound of possible outcome for a locus in GWAMA, and , in which . See the main text for more details.

**Appendix: The synthesis of a complicated model from the NSS using OATH**

We have a vector , which is estimated from the simple regression. In addition, we have 2 matrices and the other symmetric matrix . Therefore,

The highlighted elements are provided by each cohort as in the matrix.

Of note, because .

In addition, there is the second matrix, , a diagonal matrix taking the diagonal elements of , and

When all variables, the target trait and the covariates, are standardized, is a diagonal matrix with element 1. The joint estimate is

[A1]

The sampling variance of the is

[A2]

Given equations A1 and A2, the *p*-values can be calculated for each partial regression coefficient in a corresponding manner. Equations A1 and A2 can be tailored to include only a subset of covariates by introducing a design matrix , which represents the presence or absence of a covariate. For example, given 5 covariates, possible ways to composite an alternative model, based on the NSS. The implication of Equation A1 and A2 is that, given a couple of naïve regression models, it can find exactly the same result as that found using the individual-level data. Obviously, given NSS, OATH is not necessary to have other information, such as individual-level data, to reproduce and recover the underreported results.