**tedtFAQ**

Q: The test dataset seems to be password-protected. What is the password?

A: You need to register before the deadline to obtain the password or request that from individual tracks that you are attending. Otherwise, please contact individual track organizers to check if they would like to provide the password.

**Track 2: Secure Parallel Genome Wide Association Studies using Homomorphic Encryption**

Q1: What are the evaluation criteria?

A1: The security level must be set at least 128 bit. The accuracy of the model (in term of p-value for variants, see Q2.2 & Q24) is the first consideration, followed by end-to-end speed (encryption, uploading, computation, downloading, and decryption), and total memory consumption. However, we will not consider solutions that cannot return an answer in 24 hours. Regarding the parameter settings, please refer to “5.4 TABLES of RECOMMENDED PARAMETERS” of the following white paper

<http://homomorphicencryption.org/white_papers/security_homomorphic_encryption_white_paper.pdf>

Q2: What is the hardware platform (#CPUs, #Cores, Memory size, Disk size) that will be used for this task?

A2: We plan to evaluate the algorithm on Amazon T2 Xlarge or equivalent VM, which has 4 vCPU, 16GB memory, disk size around 200GB.

**There are currently three ways to submit your solutions: (1) use UTHealth virtual machine (requires authentication) (2) use Docker container and push the image to the Docker hub (3) upload Docker image to securestor (please contact organizer if you prefer this solution).**

Q2.1 How are we going to pass input files into the docker image? We have a similar question for asking the output.

A2.1: I would suggest using the following -v syntax to include the input file. Please also write the output to the mounted volume as well. You can assume that covariates.csv will not change (so, it can be built into the image) and we will use different snpMap.txt for evaluation.

$ echo "hello world" > somefile.txt

$ docker run -it --rm -v $PWD/somefile.txt:/data/somefile.txt alpine cat /data/somefile.txt

hello world

Q2.2 How will we instrument the code we provide so you will track evaluation runtime?

A2.2: Please use the following form to

<https://uthtmc.az1.qualtrics.com/jfe/form/SV_dmWY7rVnEkFl7E1>

To instruct your output including security level, time/computational cost for encryption/computation/decryption/total. Note that we also need a file for **SNP statistics or p-values (see Q7 for more details)**, which correspond to the SNPs specified in snpMap.txt (or the test file). The evaluation is to access the wellness (e.g., precision, recall, etc.) of the estimated p-values (above or below certain cutoffs) on the encrypted model against the plaintext one (semi-parallel GWAS).

Q3: Data preparation: Can we assume that the pre-processing of the VCF files to generate snpMatrices can be done in the clear with the matrices then homomorphically encrypted?

A3: Sure, we can assume the conversion is done in the clear using our provided script or your own protocol bearing the same output.

Q4: What is the reference measurement for accuracy?

A4: The reference measurement will be the accuracy in the plaintext using the same approximation procedure as we provided.

Q5: Will the VCF files contain snps from a single chromosome, as per data set provided?

A5: For this competition, we can assume VCF files contain SNPs from a single chromosome.

Q6: Are we supposed to construct a solution with logistic regression? Because the reference article refers to both linear and logistic regression. In addition the challenge description seems to talk about a specific semi-parallel algorithm, "We will use an alternative semi-parallel algorithm for this competition and rely on an approximation to reduce the necessary rounds of computation.". Is there a specific algorithm that you are referring to and that hence should be used in the solution?

A6: Yes, We are supposed to use logistic regression in this task with this specific semi-parallel algorithm. The reason is that if you use logistic regression directly, you can only build one model at a time for each variant (SNP), which takes too long for a GWAS study involving thousands of variants. That is why we are referring to this paper that has a semi-parallel algorithm, allowing us to calculate logistic regression based models for many variants at the same time. You should refer to (fast\_covariates.r) code to see how they do it.

Q7: What are the necessary outputs of the computation for each SNP (e.g beta coefficient and standard error? Test statistic? p-value? or all of the above)?

A7: **test statistics or p-values** are what we are seeking after. **(sorry for this update as we realized that some p-values are very very small, affecting the precision that a HE scheme might achieve, so test statistics are also acceptable)** If you choose to output test statistics, we will use wald z-test, so the test statistics is b (estimated parameter for the SNP) divided by err (standard deviation of the estimated parameter). We will perform the following in plaintext using the decrypted value of (b / err), pvalue\_parallel = 2 \* pnorm(-abs(b / err))

Q8: How is accuracy evaluated and on what scale or to what degree of precision. For p-values for example, is an answer of P\_est=0.04 when the truth is P\_true=0.05 considered much less accurate than an answer of P\_est=0.04e-100 when the truth is P\_true=0.05e-100? Or is the accuracy evaluated in log space, in which case an answer of P\_est=0.50 when the truth is P\_true=0.05 is considered much more accurate than an answer of P\_est=0.05e-100 when the truth is P\_true=0.05e-102?

A8: No, we will actually evaluate the performance at several different cut-off values. For example, we will set a threshold value (e.g., 0.01, or smaller) and calculate type-I error (false positive) and type-II error (false negative) based on the answer and the truth (generated by the plaintext semi-parallel algorithm). We might use F1 score as a single index to summarize the performance.

Q9: Will both linear and logistic regressions be performed?

A9: Logistic regression is sufficient for this GWAS task

Q10: Will there be missing data and, if so, how should missing data handled? In the provided example the missing data is mean filled.

A10: It will be mean filled if there is missing value in the test data.

Q11: What is the input format for the genotype data? The example vcf files do not have genotype entries and the snp matrix only contains 0/1 entries (whereas real vcf files can contain either 0/1/2 genotype calls or real value probabilities).

A11: Our raw training data are in the format of VCF but we only consider SNPs (a conversion program is provided in "buildMat.py"). You can assume our inputs are just 0/1 entries (based on our conversion program) but this would not limit the generalizability of protocol to handle 0/1/2 genotype calls. You can assume the input will be provided in the format like "snpMat.txt"

Q12: In the paper they compute the values in two steps, first they construct a logistic regression model not taking the SNP values into account. Afterwards they compute the coefficients of the logistic regression corresponding to the SNP values in a semi-parallelised way. For the competition what is assumed to be the input? Do we assume the first logistic regression model is also constructed homomorphically? Do we assume we know the coefficients of the logistic regression model without the SNPs and should we only evaluate this homomorphically to get the p-values? Or do we even start from the p-values?

A12: The input would be the VCF converted matrix (e.g., snpMat.txt), which includes the SNP information of each genome. That is, if patient A has a SNP at position XXXX, that location will be 1. You need to follow the two-step approach to compute the model and output p-value in the end. Both the logistic regression (without considering SNP values) and the semi-parallel coefficient (test statistics or p-value) computation for SNP values should be conducted homomorphically.

Q13: In the evaluation criteria, the scalability seems to be very important factor. Could you provide the range of the scalability for data size?

A13: We would test about 200-300 individuals and 10k-20k SNPs for this task.

Q14: We have a new HE framework that has not been published yet but it might give a boost to the parallelism, can we use the new framework to develop a solution?

A14: It would be great to build the competition technology based on an established framework. If the paper has not yet been published, it would be nice to provide two solutions based on an established framework and this new framework

Q15: In this task, as far as we understood, the main goal is to compute the p-value and the standard errors (or variation).

Referring the page 8 of the paper "GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies",

we can calculate p-value = <z,s>/ <s,s> and variation = 1/<s,s>, where <,> denotes a weighted inner product. What we believe is that the tuple (<z,s>, <s,s>) gives exactly the same information with the tuple (p-value, variation).

In this point of view, is it acceptable to output encryptions of <z,s> and <s,s>, rather than p-value and variation themselves?

A15: Actually, we only care about p-value. It but we definitely prefer solutions that output only p-value

Q16: Is it acceptable to encrypt the matrix X in different forms? (e.g., we initially have both Enc(X) and Enc(X^t))

A16: Sure, the binary X (snpMat.txt) is just your input, you can encrypt in whatever ways to facilitate your computation/algorithm

Q17: Can the imputation and normalization be done in the clear, prior to encryption?

A17: Yes, we can assume the imputation and normalization are done in the clear, prior to encryption. Indeed, we will test your algorithms on normalized and imputed clear text inputs

Q18: How many covariates are expected to be used in the challenge? 3, as in the example provided?

A18: The number of covariates will be on the same scale to the training dataset.

Q19: What are we submitting for evaluation? Code? Executables? A virtual machine? Something else?

A19: We will prefer to have code and executables. Like the last year, we will provision the executables in a virtual machine with test data to evaluate the performance.

Q20: Can a team submit multiple solutions to the same task?

A20: Up to two solutions are allowed for each team.

Q21: Do you need access to source code?

A21: We might need to check the code to ensure the executables are faithfully following the competition's guideline. A detailed document describing the method, parameter, running configuration will help us significantly.

Q22: Will there be required public releases of designs and software?

A22: No, this will be completely voluntary for track 2.

Q23: How much pre-computation is allowed in solutions?

A23: We expect only encryption on the input data (covariates and SNPs) are conducted locally, the rest should be performed on the server.

Q24: The database provided by the competition has less than 250 samples, and the samples in the second half of the dataset have no covariates (they are all NaN). In the first half of the database, only 40% of the patients have the condition. This makes the recall sometimes relatively low as the program may be biased to mark the patients as not having the condition. In particular, even when the program predicts that no-one has the condition, the accuracy is still pretty high.

A24: The goal is not to compare the accuracy/recall of predictions (for the condition). The goal is to evaluate how well the estimated parameters are (in terms of p-values). This is for GWAS purpose to find which SNPs are strongly associated with the condition. We will evaluate on the type-1 and type-II errors at a cutoff (e.g., 0.01, or smaller) regarding the plaintext output vs. encrypted computational output (for example, are the significant SNPs are still significant in these two versions?). Here, we refer to the plaintext output by the output of provided semi-parallel GWAS program (instead of the actual logistic regression based algorithm, which needs one model for each SNP, making the computation infeasible).

Q25: Is there another dataset we could use to test our program?

A25: Probably not at this stage, but we will definitely test the program with a reserved dataset.

Q26: Do you plan to use recall to evaluate the programs? (It seems that the main point is to pinpoint patients who may have the condition, so just using accuracy without recall may not be a great way of evaluating the results.)

A26: We are not using accuracy/recall of the prediction for condition as an evaluation metric. The evaluation is about p-values of the estimated parameters. The provided program has this output from the semi-parallelization program “pval = as.numeric(as.matrix(pvalue\_parallel))”.

Q27: The semi-parallel algorithm does not do a good job in ranking top SNPs as it assumes the same covariants for all SNPs (and ignoring the impact of them on intercepts), why are we using it since GWAS is to find most significant SNPs?

A27: It is true that semi-parallel program does not excel at ranking but it does a good job in pruning irrelevant SNPs at a very high efficiency. At a threshold of 0.01, less than 10% of the SNPs will remain for further computation after this initial step, which allows us to use more expensive model (e.g., logistic regression for one SNP at a time) to calculate their actual p-values.

**Track 3: Secure search of DNA segments in large genome databases**

Q1: Is any leakage allowed on the database or query?

A1: No leakage allowed for the query (e.g., if matched segment is found or not). Minimum for the database, e.g., the number of similar genomes, and the length of matched segment, etc.

Q2: What is the exact answer expected? Is it the length of the matched segment, its index and the number of matches found?

A2: The answer should be the number of records in the genome database containing a matched segment with the query that is longer than the given threshold, and the lengths of these matched segments. You may also output IDs of these genomes, which do not contain any sensitive information.

Q3: “ The solution should be non-interactive” means interactivity will result in a reduced score or in a null score ?

A3: We prefer solutions without interaction. But solutions with interactions will be considered (with a reduced score) if it result in significantly more efficient solutions.

Q4: evaluation formula ? f(security,accuracy,speed,size,interactivity)= ?

A4: We do not have a specific formula for the evaluation. The general criteria will be security > Interactivity > accuracy > speed > size

Q5: is the threshold static (i.e. fixed before preprocessing the data) or is it dynamic (i.e. chosen by the client at runtime) ?

The threshold is fixed before preprocessing of the data, but may change over time. The default threshold for the competition is 1000. Participating teams may optimize their algorithms based on this default value.

Q6: Must the solution be completely non-interactive, or is interactive preprocessing (which is independent of the database and the queries) permitted?

The solution must be non-interactive.

Q7: I am not sure what is “interactive preprocessing”, if it is independent of the database and queries? Can you please elaborate?

The solution must be non-interactive. Please refer to the dataset after registration.

Q8: How do you define accuracy? I assume there are at least two types of accuracy, boundary prediction and the number of records.

Both the number of records correctly matched and correct boundary setting will be taken into account. The average accuracy of boundaries will be used as one of the main criteria. The final accuracies will be computed by a reasonable comparison of the results to the non-secure algorithm’s output. The exact details of the accuracy computation is currently left unclear because the teams may aim at training algorithms to maximize the accuracy calculation. This is not our intention in the competition.

Q9: “There is no x\_j \in X (j \neq i) such that genotypes of z and x\_j exactly match”. Did you mean

exactly match from a to b? The definition seems different from the cited paper by Durbin.

Yes. In this context, we require only exact matches. The Durbin paper allows a small number of mismatches, which we do not require in this task.

Q10: “The solution should be non-interactive”. Does this mean that it is only allowed to have one round of interaction during the protocol?

A small number (e.g., two) of rounds is allowed. But fewer rounds are preferred (i.e., the protocol with two rounds will be considered to be worse than the protocol with one round unless its efficiency is greater of magnitudes).

Q11: What is the requirement on security model? Is there a single server, or multiple servers (some of which might be corrupted)? Are the server(s) semi-honest or malicious?

There is a single server from the data owner and a single server (workstation) for the data querier. Each of them may be semi-honest (honest but curious).​

(Q12 and A12 from email conversation)

Q12: What is being output to the client and what is output to the server?

A12: The server learns nothing. The client learns the genome id and the length of each set-maximal match.

Q13: Is it acceptable that the client learns matches' position (start and end) ?

Can we expect the client to do some computations ? To what extend ?

For exemple : Can the server return local maximum and the client infer set-maximum ?

Server->Client: Where 1 is a match

Entry1 : 111110000

Entry2 : 011100000

Entry3 : 000110000

Client can infer set-max at postion 0 is 111110000 -> 5

Q13: What is the hardware platform that will be used for this task? Is it also (as for Task 2) an equiv. To Amazon T2 XLarge and can we know if there will be AVX or AVX2 support ?

Q15: Sometimes there are several matches that match the definition of a set-max starting at i and ending at j. PBWT considers that those matches are all set-max that should be reported. Should we report all of them as well ? (ie: consider that the inclusion property of a set-max is strict : “This is not a set-maximal if it is strictly included in another match”) Or should we report only one ? (ie the inclusion is large : “This is not a set-maximal if it is included or equal to another match”).

Q16: Is it possible to generate matches.1.txt from panel.macs and query.1.txt file using <https://github.com/richarddurbin/pbwt> ? If yes, what command line arguments should we use to generate it?

Q17: Can we use SGX?

Q18: What are the network specifications between the client and the server ?