Supplemental Materials

Secure discovery of genetic relatives across large-scale and distributed genomic datasets

Supplemental Tables

Supplemental Table S1: Symbols, parameters and default values.

Supplemental Table S2: SF-Relate effectively detects relatives across diverse datasets.

Supplemental Table S3: SF-Relate's recall remains high across different subpopulations.

Supplemental Table S4: SF-Relate achieves high recall across various parameter settings.

Supplemental Table S5: SF-Relate outperforms KING and achieves comparable results to more advanced methods, PC-Relate and RAFFI.

Supplemental Table S6: Assignment of UK Biobank assessment centers to geographic areas.

Supplemental Figures

Supplemental Figure S1: SF-Relate's runtime scales linearly in database dimension in practice.

Supplemental Figure S2: SF-Relate is more efficient than our alternative MPC-based implementation.

Supplemental Figure S3: Increasing the subsampling ratio s allows almost perfect precision.

Supplemental Figure S4: SF-Relate's hashing and micro-bucketing strategies effectively assign close relatives to the same bucket.

Supplemental Figure S5: SF-Relate and PC-Relate exclude spurious 4th-degree relatives detected by KING, when compared to RAFFI as the ground-truth.

Supplemental Figure S6: SF-Relate's alternative output mode accurately reports all computed kinship coefficients.

Supplemental Figure S7: SF-Relate's alternative output modes compute accurate individuals-level statistics with customizable thresholds.

Supplemental Figure S8: Relative detection based on Homer et al.'s attack results in near-random performance.

Supplemental Figure S9: SF-Relate's workflow.

Supplemental Figure S10: LSH for Hamming similarity retains more high-probability pairs (candidate IBD segments) than LSH for Jaccard similarity (MinHash).

Supplemental Figure S11: KING can be accurately estimated on subsampled subsets of SNPs.

Supplemental Notes

Supplemental Note S1: Secure MHE protocols for secure kinship evaluation

Symbol	**Definition**	**Default**
ϕ	KING kinship coefficient	—
d	relatedness degree	—
θ	KING coefficient cutoff	2^{-d-1}
cMLen	subchromosome length	8cM
cMStep	overlap between subchromosomes	4cM
TargetLen	# SNPs after random projection	80
w'	size of projection window	—
k	# SNPs in k-SNPs	8
n	# local individuals	—
m	# SNPs	—
τ	table ratio	$1 \sim 128$
s	subsampling rate (for SNPs)	$0 \sim 1$
N	# buckets/table size	$\tau \cdot n$
S	bucket capacity	1
ℓ	# LSH values in each index	4
L	# repetitions of hashing (*Step 1: Hashing and bucketing*)	3
M	# (subsampled) SNPs	$s \cdot m$
C	bucket capacity used in micro-bucketing	1
B	Number of values in each ciphertext	8192

Supplemental Table S1: **Symbols, parameters and default values.** The *Default* column indicates the optimal values for the parameters across all datasets in our experiments. For the encoding pipeline (see Methods), we divide each chromosome into subchromosomes of length in equal genetic distances cMLen, with adjacent segments overlapping in a fixed distance cMStep. Empirically, 3cM to 20cM are reasonable values, with 8cM being the best for both UK Biobank and All of Us datasets. After that, we randomly project each subchromosome vector down to a vector of a fixed length TargetLen. The projection randomly selects one variant out of every window of size w' SNPs, with probability proportional to their minor allele frequencies. Empirically TargetLen is set to 80 to ensure enough SNPs are chosen. To make each k-SNP correspond to roughly the same genetic distance, w' is chosen differently for each subchromosome, and it is computed as the ratio between the actual number of SNPs in the subchromosome and TargetLen. The parameter k in k-SNP can be between 5 and 30, and 8 is the best value for all datasets. The repetition parameter L specifies the number of times parties should repeat *Step 1: Hashing and bucketing* (Methods). To ensure resulting merged table is highly utilized ($>$ 99.9% non-dummies), L can be chosen on the fly independently by each party.

Dataset	**Recall** (%, counts)				
	Relatedness degree				**Overall**
	0th	**1st**	**2nd**	**3rd**	
Geographically Distributed	100.0% 22/22	100.0% 3370/3370	100.0% 1483/1483	99.5% 10591/10640	99.7% 15515/15466
Single African Ancestry	100.0% 104/104	100.0% 1278/1278	100.0% 850/850	98.8% 530/542	99.6% 3262/3274
Locally-Phased UKB	100.0% 16/16	99.5% 3999/4017	98.2% 1598/1626	93.2% 3604/3863	96.7% 9217/9522
Locally-Phased AoU	100.0% 14/14	100.0% 209/209	97.8% 91/93	87.0% 134/154	95.3% 434/468

Supplemental Table S2: **SF-Relate effectively detects relatives across diverse datasets.** We showcase the robustness of SF-Relate's in two scenarios. (1) *Geographically Distributed*: a collaborative setting involving six centers collecting patient data within their respective geographic region (see Supplemental Table S6). The dataset comprises 100K individuals of white British ethnicity sourced from the UK Biobank. (2) *Single African Ancestry*: a collaborative setting with two parties each holding 10K samples of African ancestry from All of Us. Additionally, we confirm that using locally-phased data does not affect SF-Relate's ability to identify relatives. We consider two examples, each involving 20K patients split between two parties and sourced either from the UK Biobank (*Locally-Phased UKB*) or All of Us (*Locally-Phased AoU*). We test two prominent phasing tools that we use independently on each site. We apply SHAPEIT 5 (`phase_common`) on UKB and Eagle v2.4.1 with the option `--maxMissingPerSnp 0.5` on AoU. Across all settings, SF-Relate's recall remains consistently high, all above 87%. We observe a possible minor reduction in the third-degree recall on locally phased AoU, likely due to phasing errors associated with the small dataset size. We expect local phasing to be more accurate in a larger cohort or when large public reference panels are used.

Dataset	**Race/ethnicity label**	**Relatedness degree**	**Recall** (counts, %)	
AoU-20K	Black or African American	3	39/44	88.6%
	None Indicated	3	16/17	94.1%
	White	3	77/78	98.7%
UKB-200K	Any other mixed background	3	9/10	90.0%
	Any other white background	3	77/86	89.5%
	British	3	8033/8499	95.1%
	Caribbean	3	22/22	100.0%
	Indian	3	27/32	84.4%
	Irish	3	209/217	96.3%
	Other ethnic group	3	14/17	82.4%
	Pakistani	3	18/20	90.0%
	Prefer not to answer	3	28/29	96.6%

Supplemental Table S3: **SF-Relate's recall remains high across different subpopulations.** We apply SF-Relate to the multi-ethnicity datasets with 20K individuals from AoU and 200K individuals from UKB, each split between two parties. The recall for the 0th, 1st and 2nd degrees is nearly 100% for all subpopulations in both datasets, and thus excluded from the table. We separately calculate and report the third-degree recall for each subpopulation with at least 20 third-degree samples.

cMLen	5.0								
k	6.0			8.0			10.0		
ℓ	3.0	4.0	5.0	3.0	4.0	5.0	3.0	4.0	5.0
Recall (%)	**84.2**	**96.6**	**97.5**	**96.1**	**97.5**	**98.3**	**97.6**	**98.0**	**98.3**

cMLen	8.0								
k	6.0			8.0			10.0		
ℓ	3.0	4.0	5.0	3.0	4.0	5.0	3.0	4.0	5.0
Recall (%)	**82.7**	**96.0**	**97.3**	**96.0**	**97.9**	**98.1**	**97.7**	**98.0**	**98.5**

cMLen	11.0								
k	6.0			8.0			10.0		
ℓ	3.0	4.0	5.0	3.0	4.0	5.0	3.0	4.0	5.0
Recall (%)	**77.3**	**95.7**	**97.2**	**96.0**	**97.3**	**97.5**	**97.4**	**97.5**	**97.9**

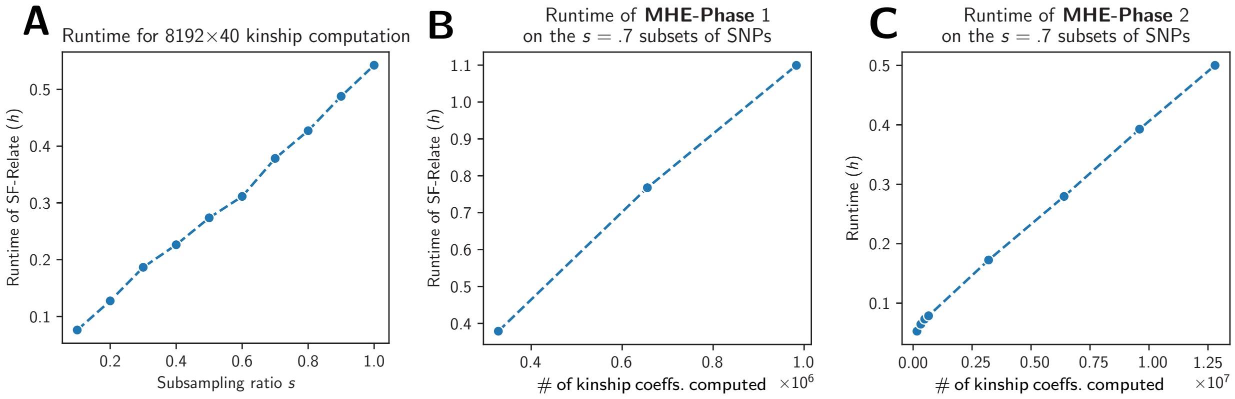
Supplemental Table S4: **SF-Relate achieves high recall across various parameter settings.** On UKB-200K, we select a combination of reasonable parameters and compute the third-degree recall achieved by SF-Relate. The overlap between chromosomes cMSTEP is set to equal half of the subchromosome length cMLEN (see Supplemental Table S1). We vary the values of k, the number of SNPs in k-SNPs and ℓ the number of LSH values in each index, and maintain the remaining parameters at their default values provided in Supplemental Table S1. SF-Relate consistently achieves near-perfect recall ($> 95\%$), except on the lower end of the parameters, when there is not enough entropy among the subchromosomes for the LSH to separate samples into buckets.

Degree combinations	**A**	**B**	**C**	D	**E**	F	G	H	**I**	J	**K**
SF-Relate	**1st**	**2nd**	**3rd**	3rd	**4th**	4th	U	U	**U**	U	**U**
KING	**1st**	**2nd**	**3rd**	3rd	**4th**	4th	3rd	4th	**4th**	U	**U**
RAFFI	**1st**	**2nd**	**3rd**	4th	**4th**	U	3rd	4th	**U**	U	**U**
PC-Relate	**1st**	**2nd**	**3rd**	3rd	**4th**	U	3rd	4th	**U**	4th	**U**
Number of individuals	**508**	**173**	**863**	48	**93**	23	22	23	**8410**	23	**9716**

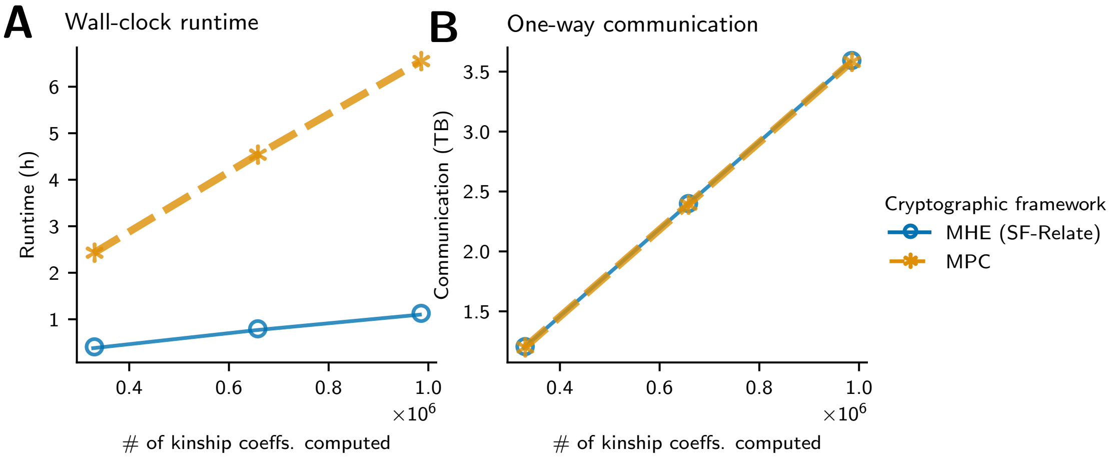
Supplemental Table S5: **SF-Relate outperforms KING and achieves comparable results to more advanced methods, PC-Relate and RAFFI.** On a subset of 20K samples from UKB-200K, we compute the maximum relatedness degree for each individual using the different methods, and report the number of individuals falling into each classification type (when this count exceeds 20). The abbreviation U represents *Unrelated.* **Underlined columns** correspond to the set of individuals with identical classifications across all methods. **Bold-faced columns** denote substantial sets containing at least 100 individuals. Most methods have consistent results in the largest categories, except in column I. Whereas SF-Relate and KING agree on most columns, SF-Relate outperforms KING by aligning with the more accurate methods in column I, effectively excluding the thousands of likely-spurious 4th-degree pairs identified by KING.

UK Biobank Assessment Center	**Geographic Area**	**Number of Samples**
Edinburgh Glasgow	Scotland	7051
Middlesborough Newcastle	Northeast England	14185
Liverpool Bury Stockport Manchester Leeds	Northwest England	25096
Birmingham Stoke Sheffield Nottingham Wrexham	Southeast England	23362
Barts Hounslow Croydon Reading Oxford	Central England	16836
Swansea Cardiff Bristol	Wales	13470
Total Number of Samples		100,000

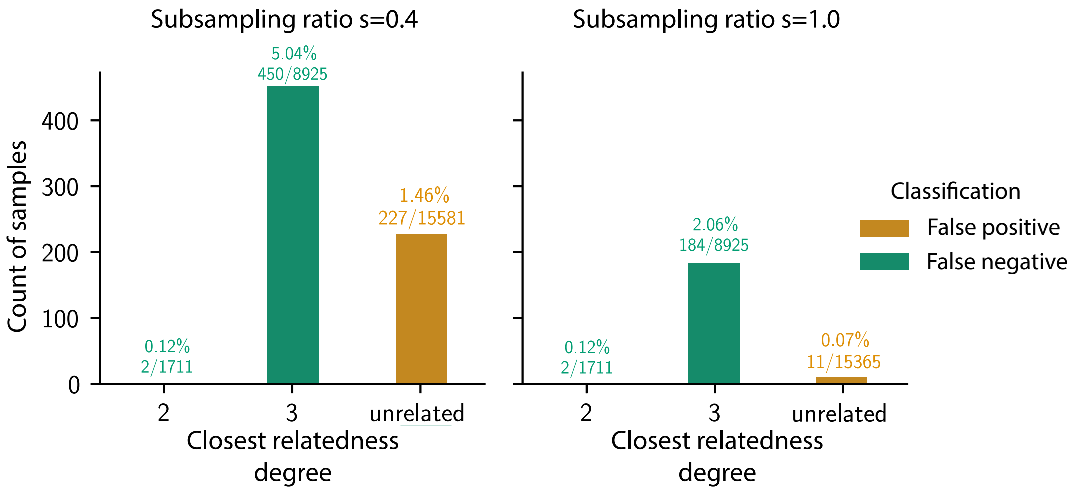
Supplemental Table S6: **Assignment of UK Biobank assessment centers to geographic areas.** To simulate a federated study, we use a subset of 100K white British individuals from UK Biobank dataset. We organized the 22 data collection centers (Data-Field 54) into six study groups according to their geographic locations within the UK.



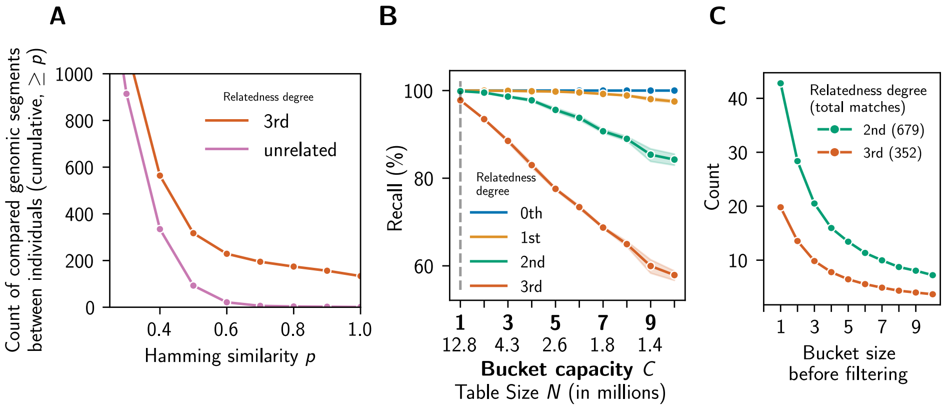
Supplemental Figure S1: **SF-Relate's runtime scales linearly in database dimension in practice.** In **(A)**, we perform a fixed number of kinship computations $N = 327,680$ (i.e., 40 blocks of 8192) while varying the number of SNPs by varying the subsampling rate s. In **(B)** and **(C)**, we increase the number of kinship computations N while keeping the subsampling rate s at .7. We report the runtime of the MHE-Phase 1 and MHE-Phase 2 in (B) and (C) separately.



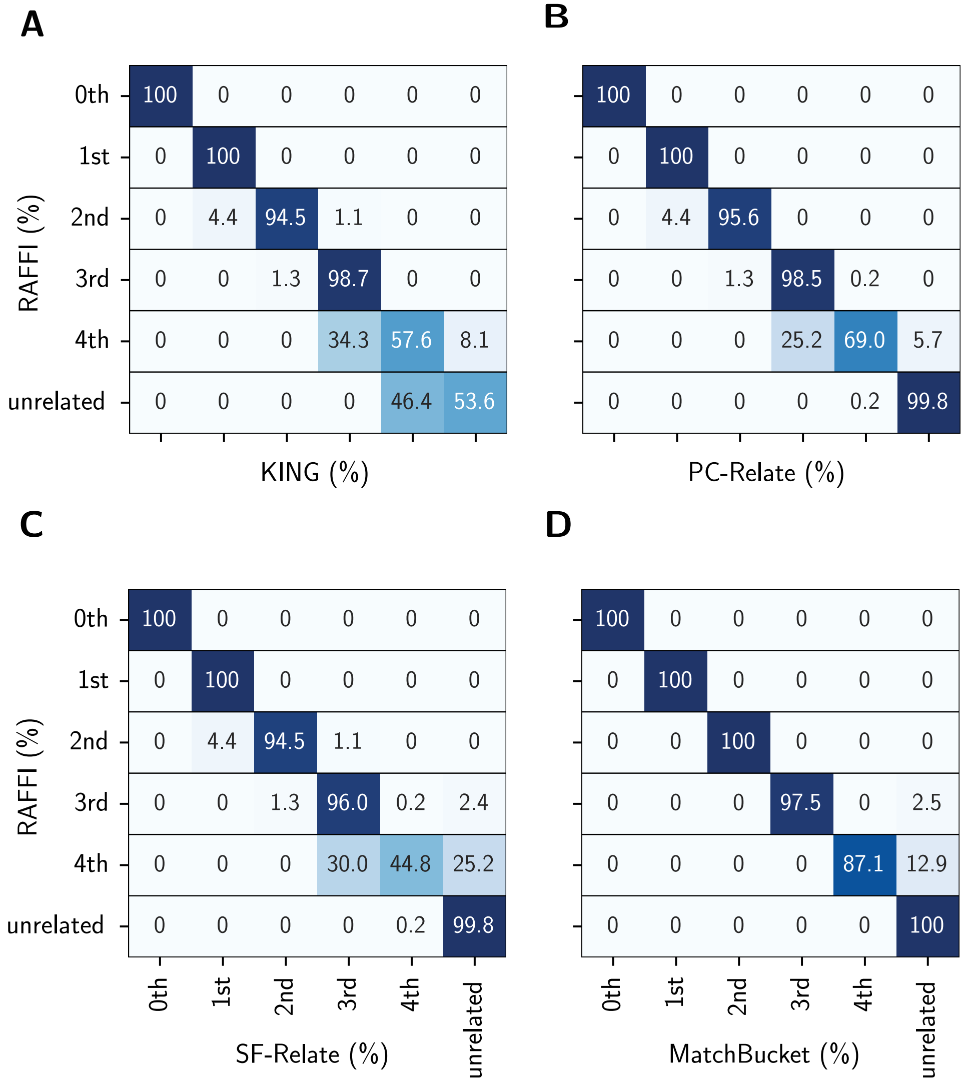
Supplemental Figure S2: **SF-Relate is more efficient than our alternative MPC-based implementation.** To illustrate the effectiveness of our MHE approach, we developed an alternative approach based on secure multiparty computation (MPC) for pairwise kinship coefficient calculation. We optimized the MPC protocol based on a prior work (Cho et al. 2018) and evaluated its performance in the same environment as SF-Relate (Methods). We report the runtime **(A)** and communication costs **(B)** of both methods for varying numbers of kinship coefficients. We observed approximately 8 times faster runtimes for SF-Relate and comparable communication costs; while MPC protocols rely more heavily on interaction, a 16-fold expansion of data size due to encryption in the MHE approach leads to a comparable degree of data exchange.



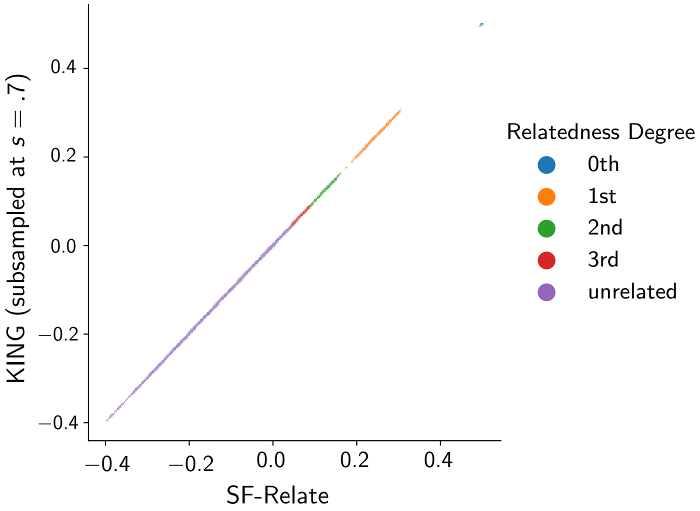
Supplemental Figure S3: **Increasing the subsampling ratio s allows almost perfect precision.** We perform SF-Relate on UKB-200K under the alternative subsampling (sketching) ratio $s = 1$, and show counts of individuals based on the their closest relations and SF-Relate's detection results. Using the full set of SNPs in SF-Relate allows significant reductions in false positives and false negatives due to the sketching noise, with a moderate increase of runtime from 14.5 to 21 hours.



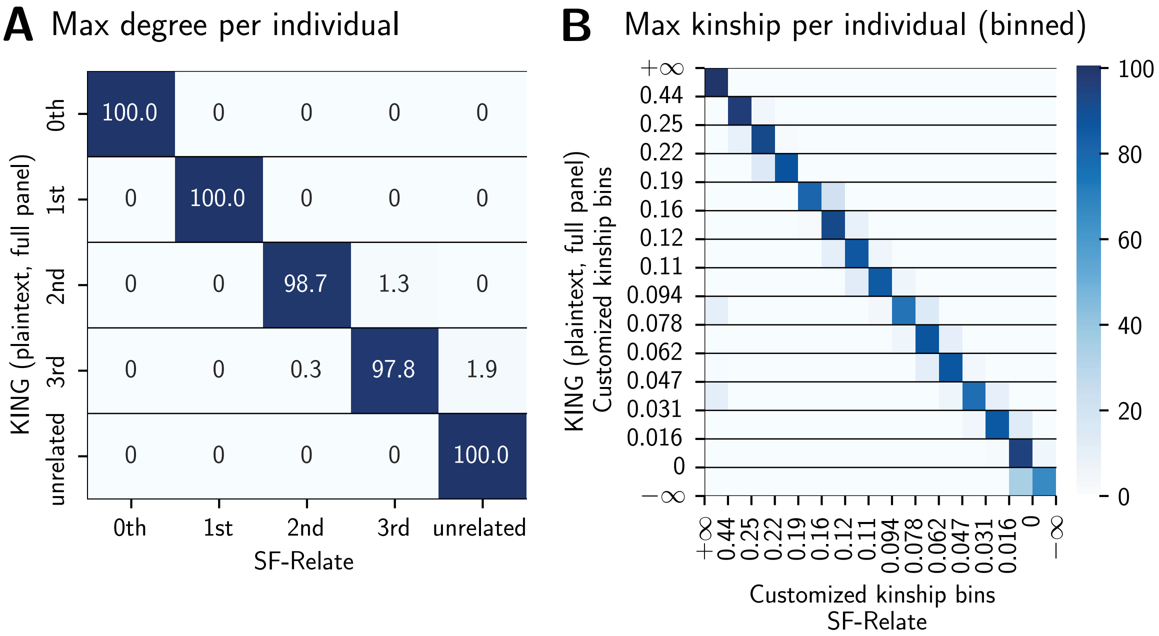
Supplemental Figure S4: **SF-Relate's hashing and micro-bucketing strategies effectively assign close relatives to the same bucket.** (**A**) Hamming LSH with SF-Relate's k-SNP encoding scheme enables separation of pairs of genomic segments with high Hamming similarity (likely IBD) between close relatives from those pairs between unrelated individuals. (**B**) Setting the bucket capacity $C = 1$ achieves the highest recall in relative detection compared to larger values of C. For comparison, we adjust the table size N accordingly to keep the number of comparisons $NC^2 = 1.28$ million constant. The recall for each relatedness degree is the fraction of relative pairs of that degree that are assigned to the same bucket. (**C**) Close relatives (of 2nd and 3rd degrees) who are assigned to the same bucket are most often found in size-1 buckets before trimming. All results are based on UKB-200K.



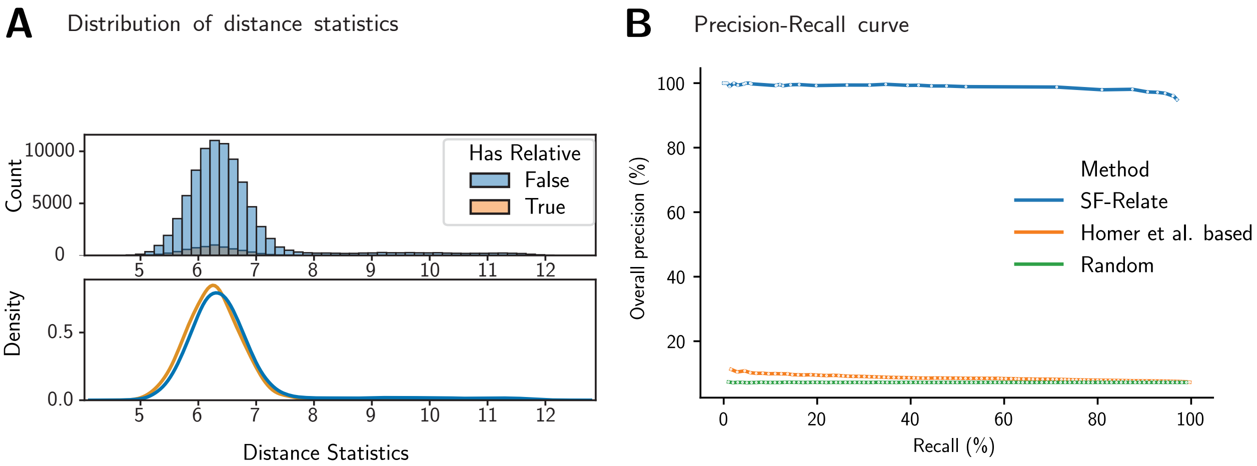
Supplemental Figure S5: **SF-Relate and PC-Relate exclude spurious 4th-degree relatives detected by KING, when compared to RAFFI as the ground-truth.** On a subset with 20K samples from UKB-200K, we present the confusion matrices assessing the relatedness classification accuracy of KING **(A)**, PC-Relate **(B)** and SF-Relate **(C)**, comparing them with the output of RAFFI as the ground-truth. *MatchBucket* **(D)** denotes the (non-private) hybrid approach where RAFFI is performed in plaintext on pairs in SF-Relate's corresponding buckets and serves as a reference. Both SF-Relate and PC-Relate label RAFFI-unrelated individuals as unrelated, unlike KING, which labels them as 4th-degree relatives. This suggests that both methods avoid the spurious relationships identified by KING.



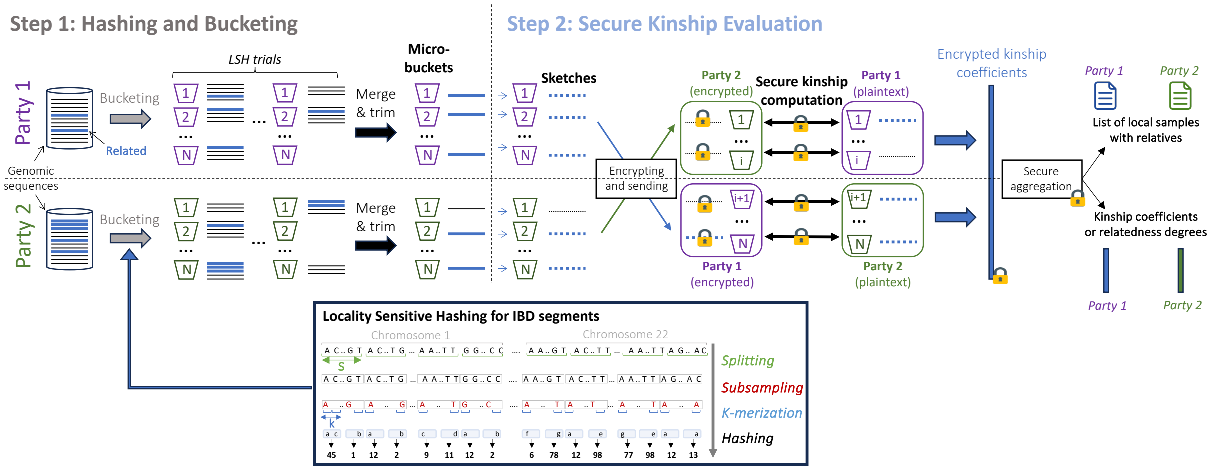
Supplemental Figure S6: **SF-Relate's alternative output mode accurately reports all computed kinship coefficients.** On a 10% random subset of kinship coefficients that SF-Relate computed on UKB-200K, we evaluate the accuracy of the alternative setting of SF-Relate where the full list of kinship coefficients is revealed. As shown in the plot, SF-Relate's output accurately matches with KING, indicating that the MHE noise is small with respect to the kinship coefficients being computed on the subsampled set of SNPs.



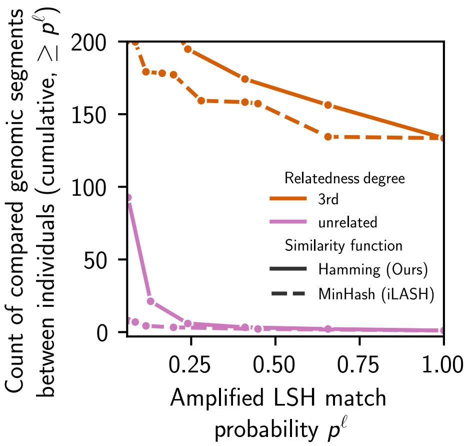
Supplemental Figure S7: SF-Relate's alternative output modes compute accurate individuals-level statistics with customizable thresholds. On a subset of 10% kinship coefficients computed in UKB-200K's hash tables, we evaluate SF-Relate's alternative output modes that support the comparison of the computed kinship coefficients with a sequence of thresholds. Numbers in the cells or their colors indicate the corresponding recall rates. In **(A)**, we choose the thresholds to be the recommended kinship cutoff by KING (Manichaikul et al. 2010). In **(B)**, the thresholds defining the refined bins are marked on the axes. The predictions perfectly align with the ground-truth KING predictions computed on the full SNP panel on more than 99.9% and 85% individuals in **(A)** and **(B)**, respectively. More than 99.9% of the predictions in **(B)** are accurate, with deviations at most shifted by a single bin. This result indicates that both modes produce reliable assessment of the maximum kinship level, and highlight the utility of SF-Relate in various workflows.



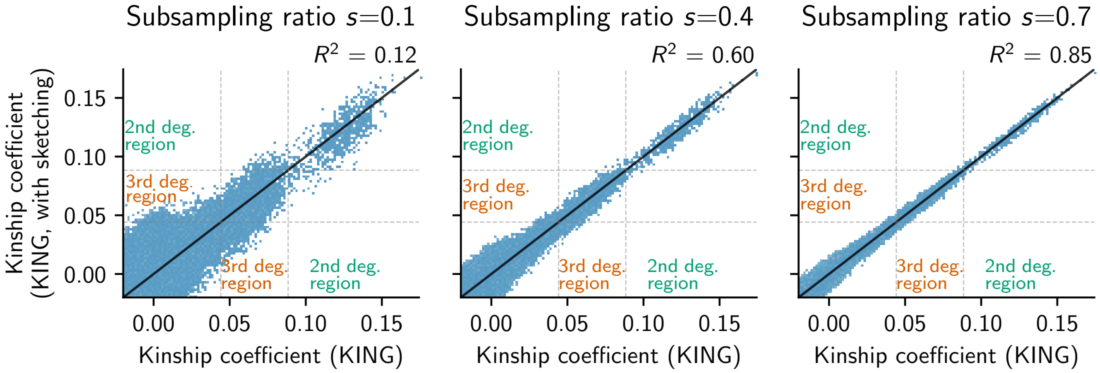
Supplemental Figure S8: **Relative detection based on Homer et al.'s attack results in near-random performance.** Homer et al.'s attack (2008) predicts whether an individual contributes to a genetic dataset by statistically testing whether the individual's genotype count vector is closer to the allele frequencies in the dataset than some alternative reference frequencies. The same attack can potentially reveal the presence of a relative in a dataset due to shared genetic sequences. We evaluate the efficacy of this approach on the UKB-200K dataset split among two parties. In (**A**), we compute (1) the L^1 distance of every sample's genotype count vectors to the mean genotype count vector from the local dataset (excluding relatives), representing the background statistic, and (2) the distance between the sample genotype count vector and the mean count vector in the other party's dataset, representing the target statistic. We then subtract the two distances to see which dataset the sample is closer to. The figure shows that this estimator does not effectively separate samples that have a relative in the other dataset from samples that do not. In (**B**), we show the precision-recall curves for various methods for detecting 3rd-degree or closer relatives in UKB-200K. For Homer et al., we vary the distance threshold used as a cut-off to determine whether a sample is closer to the target dataset. For SF-Relate, we plot the curve by varying the table size parameter (τ in Methods). Homer et al.'s approach obtains precision comparable to the level of random guessing, whereas SF-Relate achieves near-perfect precision.



Supplemental Figure S9: **SF-Relate's workflow**. In Step 1, the parties perform multiple trials of LSH to bucket samples before merging and trimming to obtain buckets of size 1 (micro-buckets). In Step 2, each sample is sketched and securely compared against the other party's sample in the same bucket to evaluate kinship (MHE-Phase 1). Finally, parties securely aggregate the results to obtain per-sample output (MHE-Phase 2).



Supplemental Figure S10: **The LSH for Hamming similarity retains more high-probability pairs (candidate IBD segments) than LSH for Jaccard similarity (MinHash).** On the UKB-200K dataset, we count the number of subchromosome pairs between individuals based on the probability that the pairs would produce the same LSH index with an LSH amplification with $\ell = 4$. A higher probability between segments indicate higher similarity and thus more likely to be in IBD. The raw similarity score $0 \leq p \leq 1$ between vectors is transformed to the amplified LSH match probability p^{ℓ} by the amplification described in *Step 1: Hashing and bucketing* (Methods). Each curve shows the count of segments with probability higher than p^{ℓ} averaged over pairs of related samples in the respective relatedness degree classes. The plot shows that Hamming LSH identifies more segments with high matching probability after the LSH amplification, suggesting its ability to detect more IBD segments.



Supplemental Figure S11: **KING can be accurately estimated on subsampled subsets of SNPs.** A subsampling ratio at $s = .7$ achieves reliable relatedness degree classification, where accurate kinship is not necessary. We apply SF-Relate on UKB-200K and compute the kinship coefficients on the list of micro-bucket pairs, under subsets of SNPs with different subsampling ratios. The x axis shows the kinship on the full set of SNPs. Points near the degree boundaries have a higher chance of classification error. The scatterplots show that sketching introduces noise to the kinship, but reliable relatedness degree classification is possible when points are not too close to the thresholds.

Supplemental Note S1 Secure MHE protocols for secure kinship evaluation

We detail here SF-Relate protocols for computing and detecting kinship coefficients that are above a predefined threshold (MHE-Phase 1) and for aggregating these results per individual (MHE-Phase 2).

We denote a matrix $\boldsymbol{A}$ with N rows and M columns as $\boldsymbol{A}^{N\times M}$ and a vector $\boldsymbol{x}$ with B elements as $\boldsymbol{r}^{B\times 1}$. We use 0-based indexing, i.e. columns and rows are numbered starting from 0. $\mathbf{x}[a:b]$ denotes the subvector from row a to b (including a but excluding b). We omit a when it is 0, and omit b if it equals $N-1$. Similarly, for matrices, we use $\mathbf{A}[a:b,c:d]$ to denote the submatrix specified by the ranges. For a vector $\boldsymbol{x}^{N\times 1}=(x_0,\ldots,x_{N-1})$, we use $\boldsymbol{x}^2$ to denote the vector $(x_0^2,\ldots,x_{N-1}^2)$, i.e. the one where elements are squared. The notation $\mathsf{Sign}(E)$ denotes the indicator of the event E. When applied to a vector of events as in $\mathsf{Sign}(\boldsymbol{x}^{n\times 1}==1)$, it corresponds to the vector of indicators. In other words, $\mathsf{Sign}(\boldsymbol{x}==1)$ is equivalent to $(\mathsf{Sign}(\boldsymbol{x}[i]==1),\ldots,\mathsf{Sign}(\boldsymbol{x}[n-1]==1))$. Finally, $\mathbf{0}^{B\times 1}$ and $\mathbf{1}^{B\times 1}$ denotes the vector $(0,0,\ldots,0)^{B\times 1}$ and $(1,1,\ldots,1)^{B\times 1}$, respectively.

Every ciphertext under the CKKS encryption (Cheon et al. 2017) encrypts a vector with length B, the CKKS block length, which is typically a power of 2 like 8192. We denote the encryption of a vector $\boldsymbol{x}$ as $\boxed{\boldsymbol{x}}$. We simplify the function interfaces of the CKKS implementation in Lattigo (2022) as follows. Cryptographic keys are omitted in the function calls.

- Enc($\mathbf{x}$) takes in a plaintext vector $\mathbf{x}$ and returns an encrypted ciphertext $\boxed{\mathbf{x}}$.
- CollaborativeDecrypt($\boxed{\mathbf{x}}$) takes in an encrypted vector $\boxed{\mathbf{x}}$ and returns the plaintext result $\mathbf{x}$. Note that all parties would need to collaborate for this operation.
- Homomorphic Single-Instruction Multiple-Data (SIMD) operations, including coordinate-wise addition, subtraction and multiplication between any combination of ciphertexts encrypting vectors and plaintext vectors are denoted by $+,-,\cdot$, respectively.
- The Rotate($\boxed{\mathbf{m}}, i$) operation that that receives as input $\mathbf{m}^{B\times 1}=(m_0,\ldots m_{B-1})$ outputs a ciphertext encrypting $(m_i,m_{i+1},\ldots,m_B,m_0,m_1,\ldots,m_{i-1})$.

We also implemented the following helper functions:

- Sign($\boxed{\mathbf{v}}$) makes use of the Chebyshev polynomial interpolation and Newton's method to approximately compute the sign function, using combinations of the homomorphic SIMD operations $+,-$ and $\cdot$. It returns a ciphertext $\boxed{\boldsymbol{r}}$ such that if one decrypts it, the result equals the Boolean-valued indicator for $\mathsf{Sign}(\mathbf{v})$.
- Extract($\boxed{\boldsymbol{r}^{B\times 1}}, i$) returns a ciphertext encrypting the vector $(0,\ldots,0,\boldsymbol{r}[i],0,\ldots,0)^{B\times 1}$. We implement this by homomorphically multiplying the i-th basis vector with $\boxed{\boldsymbol{r}^{B\times 1}}$.
- InnerSum($\boxed{\mathbf{m}^{B\times 1}}$) transforms the vector $\mathbf{m}=(\mathbf{m}[0],\ldots,\mathbf{m}[B-1])$ into the vector $\mathbf{s}=(s,s,\ldots,s)^{B\times 1}$ where $s:=\sum_{i=0}^{B}\mathbf{m}[i]$. Adding a length-$B$ vector with its powers-of-2 rotated copy (see Rotate($\boxed{\boldsymbol{m}}, i$) above), for all the powers of 2 at most B, namely $2^1,2^2,\ldots,B$ achieves this. Hence, we implement it efficiently by iterating over $0,\ldots,\log B$ rotating $\boxed{\boldsymbol{m}}$ accordingly, and then homomorphically add the results together.

We display here the two protocols used in Methods, namely MHE-Phase 1 and MHE-Phase 2, for secure kinship evaluation between two parties.

MHE-Phase 1 Distributed relative detection

Input: Party 1 and Party 2 have the genotype counts matrices $\boldsymbol{A}^{N\times M}, \boldsymbol{D}^{N\times M} \in \{0,1,2\}^{N\times M}$, respectively. θ denotes the kinship detection threshold, and B denotes the CKKS block length. The rows in the matrices are organized in the order specified by SF-Relate's hash tables.

Output: A block-encrypted vector of the N detection results, $\boxed{\boldsymbol{r}_1^{B\times 1}} \ldots, \boxed{\boldsymbol{r}_{\lfloor N/B \rfloor}^{B\times 1}}$.

```
for $b = 0, \ldots, \lceil (\lfloor N/B \rfloor - 1)/2 \rceil$ do        ▷ remaining blocks computed in parallel by switching roles
    Party 1: $\boldsymbol{X}^{B\times M} \leftarrow \boldsymbol{A}[bB : (b+1)B, :]$
    Party 2: $\boldsymbol{Y}^{B\times M} \leftarrow \boldsymbol{D}[bB : (b+1)B, :]$
    Party 1: $\boldsymbol{x}_{sq}^{B\times 1} \leftarrow \sum_{i=1}^{M} (\boldsymbol{X}[:, i])^2$, and $\boldsymbol{h}_x^{B\times 1} = \sum_{i=1}^{M} \mathsf{Sign}(\boldsymbol{X}[:, i] == 1)$        ▷ $\mathsf{Sign}(v)$ is the indicator of $v \geq 0$
    Party 2: $\boldsymbol{y}_{sq}^{B\times 1} = \sum_{i=1}^{M} (\boldsymbol{Y}[:, i])^2$ and $\boldsymbol{h}_y^{B\times 1} = \sum_{i=1}^{M} \mathsf{Sign}(\boldsymbol{Y}[:, i] == 1)$
    Party 2: send $\boxed{\boldsymbol{y}_{sq}} \leftarrow \textsc{Enc}(\boldsymbol{y}_{sq})$ and $\boxed{\boldsymbol{h}_y} \leftarrow \textsc{Enc}(\boldsymbol{h}_y)$ to party 1.
    Party 1: $\boxed{\boldsymbol{p}^{B\times 1}} \leftarrow \textsc{Enc}(\boldsymbol{0}^{B\times 1})$.        ▷ initiate an encrypted vector of B zeros
    for $j = 0, \ldots, M-1$ do
        Party 2: send $\boxed{\boldsymbol{y}_j^{B\times 1}} \leftarrow \textsc{Enc}(\boldsymbol{Y}[:, j])$ to party 1
        Party 1: $\boldsymbol{x}_j{}^{B\times 1} \leftarrow (\boldsymbol{X}[:, j])$, and $\boxed{\boldsymbol{g}} \leftarrow \boxed{\boldsymbol{p}} + 2\boldsymbol{x}_j \cdot \boxed{\boldsymbol{y}_j}$        ▷ SIMD coordinate-wise multiplication
    Party 1: $\boxed{\boldsymbol{t}_1^{B\times 1}} \leftarrow \textsc{Sign}((2-4\theta)\boldsymbol{h}_x - (\boldsymbol{x}_{sq} + \boxed{\boldsymbol{y}_{sq}} - \boxed{\boldsymbol{g}}))$        ▷ $\textsc{Sign}(\boxed{v})$ is the (encrypted) indicator $\boxed{v \geq 0}$
    Party 1: $\boxed{\boldsymbol{t}_2^{B\times 1}} \leftarrow \textsc{Sign}((2-4\theta)\boxed{\boldsymbol{h}_y} - (\boldsymbol{x}_{sq} + \boxed{\boldsymbol{y}_{sq}} - \boxed{\boldsymbol{g}}))$
    Party 1: Save the batch $\boxed{\boldsymbol{r}_b} \leftarrow \boxed{\boldsymbol{t}_1} \cdot \boxed{\boldsymbol{t}_2}$ and share with party 2. ▷ Results are shared to execute MHE-Phase 2
```

MHE-Phase 2 Accumulation to per-sample output

Input: Party 1 has a list of block-encrypted boolean values $\boxed{\boldsymbol{R}^{N\times 1}} = (\boxed{\boldsymbol{r}_0^{B\times 1}}, \ldots, \boxed{\boldsymbol{r}_{\lfloor N/B \rfloor}^{B\times 1}})$, each storing B indicators of positive detection for close relations (i.e. the corresponding KING coefficient passes the threshold θ), and a list $D = (D_1, \ldots, D_N)$ storing the local IDs party 1 used in the comparisons.

Output: A list of (decrypted) boolean values $b_1, \ldots, b_n$, one per local ID, signifying whether each ID appears in at least one high-kinship comparison.

```
for $i = 0, \ldots, \lfloor n/B \rfloor - 1$ do
    $\boxed{\boldsymbol{o}^{B\times 1}} \leftarrow \textsc{Enc}(\boldsymbol{0}^{B\times 1})$        ▷ $\boldsymbol{0}^{B\times 1}$ means the vector of B zeros
    for $\mathsf{id} = iB+1, \ldots, \min((i+1)B-1, n)$ do        ▷ Iterate by blocks of size B
        $\boxed{\boldsymbol{e}^{B\times 1}} \leftarrow \textsc{Enc}(\boldsymbol{0}^{B\times 1})$
        $\mathsf{locs}_{\mathsf{id}} \leftarrow$ list of locations $j$ in $1, \ldots, N$ where $D_j = \mathsf{id}$
        for $j \in \mathsf{locs}_{\mathsf{id}}$ do
            $\boxed{\boldsymbol{u}^{B\times 1}} \leftarrow \textsc{Extract}(\boxed{\boldsymbol{r}_{\lfloor j/B \rfloor}}, j \bmod B)$ ▷ Extract the boolean result through multiplication with a one-hot encoded plaintext vector
            $\boxed{\boldsymbol{e}} \leftarrow \boxed{\boldsymbol{e}} + \boxed{\boldsymbol{u}}$
        $\boxed{\boldsymbol{e}} \leftarrow \textsc{InnerSum}(\boxed{\boldsymbol{e}})$        ▷ $\textsc{InnerSum}(\boxed{\mathbf{m}})$ transforms $\boxed{(\mathbf{m}_1, \ldots, \mathbf{m}_B)}$ into $\boxed{(\sum_{i=1}^{B} \mathbf{m}_i) \cdot \mathbf{1}^{B\times 1}}$
        $\boxed{\boldsymbol{o}} \leftarrow \boxed{\boldsymbol{o}} + \textsc{Extract}(\boxed{\boldsymbol{e}}, \mathsf{id} \bmod B)$
    $\boxed{\boldsymbol{f}} \leftarrow \textsc{Sign}(\boxed{\boldsymbol{o}} - 0.5 \cdot \mathbf{1}^{B\times 1})$        ▷ Compute whether the result is (much) larger than 0
    $b_{i*B}, \ldots, b_{\min((i+1)*B, n)} \leftarrow \textsc{CollaborativeDecrypt}(\boxed{\boldsymbol{f}})$        ▷ Decrypt ciphertext collaboratively
```

References

Cheon JH, Kim A, Kim M, and Song Y. 2017. Homomorphic encryption for arithmetic of approximate numbers. In Takagi T and Peyrin T., editors, *Advances in Cryptology – ASIACRYPT 2017*, pages 409–437.

Cho H, Wu DJ, and Berger B. 2018. Secure genome-wide association analysis using multiparty computation. *Nature Biotechnology*, **36**(6):547–551.

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, and Craig DW. 2008. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, **4**(8):e1000167.

Lattigo 2022. Lattigo v4. Online: https://github.com/tuneinsight/lattigo. EPFL-LDS, Tune Insight SA.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, and Chen WM. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**(22):2867–2873.