1 Utility Function

Supplementary Figure 1 illustrates how the utility function is calculated.

Supplementary Figure 1: Utility function. The family contains $M$ members, $m$ of which has arrived at the database. Addition of the first incoming member does not require removing any SNPs. For the other arrived members, certain part of the genomes are masked that are shown as bars with vertical and horizontal lines. The SNP positions that are common in every family member is represented as a black bar; the size of which is $V$. The formula shows how the utility ($U$) is calculated based on these numbers.

2 Outlier Pairwise Allele Counts in openSNP database

We label the alleles of an individual at a certain position with the number of non-reference alleles. The instances for which both alleles are the same as the reference genome is shown as 0, the positions wherein only one allele differ from the reference genome are denoted by 1, and finally the instances wherein an individual carry the two alternate alleles shall be denoted by 2. For two individuals, $i$ and $j$, we denote
the pairs of alleles with \([s_i, s_j]\). There are six possible combinations: both individuals have two reference alleles at the designated position: \([0,0]\), one individual has one reference allele and the other has two: \([1,0]\), the individuals are homozygous: \([2,0]\), the individuals are heterozygous: \([1,1]\), one individual has one alternate allele and the other one has two: \([2,1]\), and finally both individuals have two alternate alleles \([2,2]\). First of all, note that without applying our methodology to hide the positions, it might be possible to detect relatives based on pairwise allele counts. For example \([2,0]\) allele pair will almost always be zero in parent-offspring relations due to Mendel’s law.

We sampled 1000 random individuals from the openSNP database data, for each \([s_i, s_j]\), we record the minimum counts observed between any two unrelated individuals. We will require that at any point in the database, the count of no allele pairs are brought down to this recorded minimum threshold, thereby ensuring that the two individuals count resemble to the rest of the population. The boxplot of genotype counts is shown in Supplementary Figure 2. The outlier for different allele combinations are: \(o_{10} = 27K\), \(o_{11} = 27K\), \(o_{12} = 15K\).

![Boxplot](image)

**Supplementary Figure 2: Number of different allele pairs in the population.** \(n_{s_i s_j}\) is the number of genotype pairs where one individual has genotype \(s_i\) and the other individual has genotype \(s_j\), respectively.

### 3 General Optimization Framework

When the family members are very closely related a feasible solution might not exists that satisfy both privacy constraints. We therefore suggest relaxing one constraint and satisfying the other. The general framework of this solution is illustrated in Supplementary Figure 3.

### 4 Inference of Family B

Clustering techniques are used to group data points which are more similar to each other. Likewise, people who are genetically closer or more related can be grouped together via clustering methods. We infer Family A from openSNP genomic database using a hierarchical clustering approach. There were 1000 genomes available. The distance between two individuals or clusters is calculated using the genetic dissimilarity defined by Weir and Zheng [11] and average linkage is employed.

Supplementary Figure 4 represents the result of clustering graphically in a dendogram. The left axis denotes the height of the tree in terms of genetic dissimilarity and the right axis denotes the KING kinship coefficient. We extracted clusters that are extremely low hanging in the tree, these indicate clusters that have very low genetic dissimilarity thus, higher similarity. 53 such clusters are detected, these are candidate families in the openSNP database. Among these clusters, one cluster has 5 members, two clusters have 4 members, 6 clusters have 3 members and 44 clusters have 2 members. Supplementary...
Supplementary Figure 3: Protecting the privacy of a family when a new relative $i$ arrives at the database. There are two different ways to solve the problem: one is based on the relaxation of outlier constraints and the other is based on the relaxation of kinship constraints. For each method, if there is a feasible solution, person $i$ can be added to the database safely.

Figure 4 shows a small fraction of the dendrogram in which the data points representing two families are encircled. The family in cluster c1 consists of two members; whereas the family in cluster c2, which will be referred to as $f_B$ in the rest of this paper, consists of five members.

Based on the pairwise kinship coefficients of family members, we inferred the pedigree. One possibility is that the family comprises person $B$, the mother, the father, the maternal grandmother, and the paternal grandfather. Another possible family structure contains person $B$, the mother, the father, the maternal aunt and the paternal uncle. This ambiguity stems from the similar kinship coefficient interval between parent-offspring and siblings. We assume the first possibility holds and the family contains person $B$, his mother, father, maternal grandmother and his paternal grandfather.
5 Additional Results for the Arrival Sequences where Person A or Person B Arrives First

5.1 Results by Solving the Optimization Problem by Satisfying Outlier Constraints

5.1.1 Arrival Sequences where Person A Arrives First

Supplementary Figure 5 displays the results for family $f_A$ when the optimization problem is solved based on satisfying the outlier constraints and relaxation of the kinship constraints. We observe that if the parents of person A arrive in the second step, it is impossible to hide the relationship. When more distant relatives such as a sister or a maternal aunt arrives at the database after the arrival of person A, the kinship coefficient cannot be decreased to the level of two unrelated individuals, but the relationship can be hidden such that they are identified as more distant relatives: person A – aunt relationship can be deciphered as a 3rd degree relationship and sister – person A relation can be decreased to a 2nd degree relative relationship. In the third level of the tree, the parents of person A may not be added to the database since there is no feasible solution. Additionally, the sister cannot be added, thus first-degree members of person A are not allowed to be added after the aunt. At this stage of the tree, only the maternal aunt can be added safely, if the second added member is the sister. This scenario achieves 56% utility. In this solution, person A and the sister can be inferred as second degree relatives and the relationships of the maternal aunt to person A and the sister are not revealed. In the 4th level of the
tree, none of the family members can be added without hurting family’s privacy.

Supplementary Figure 5: The first arrived member is person A. The possible arrival sequences of family $f_A$ is shown where person A is the first arrived member. The optimization model is solved by relaxing kinship privacy constraints and satisfying outlier constraints.

5.2 Results by Solving the Optimization Problem by Satisfying Kinship Constraints

5.2.1 Arrival Sequences where Person B Arrives First

Supplementary Figure 6 illustrates all possible arrival sequences when person B is the first member to arrive at the database. The corresponding solutions are based on solving the model by relaxing the outlier constraints. Parents can be added at the second level but the outlier values will be very low, rising the risk of privacy impairment. In this case, it is not possible to add a third family member because the model is infeasible for the given outlier constraints. If the second added family member following person B is a second degree relative such as paternal grandfather or maternal aunt, then adding this member is feasible. This addition results in approximately 8σ leakage in $o_{11}$ value and 38.1% utility. A small amount of difference is observed in the utility and outlier values pertaining to the addition of maternal grandmother and paternal grandfather as the second members. At the third level of the tree, two addition sequences are obtained: {person B - paternal uncle - maternal aunt} and {person B - maternal aunt - paternal uncle}. These sequences show that the second degree relatives of the person B can be successfully added at any level of the tree when the outlier constraints are relaxed.
Supplementary Figure 6: The first arrived member is person B. The solutions for different arrival sequences of family $f_B$ is shown where person B is the first arrived member. The optimization model is solved by relaxing the outlier constraints and satisfying kinship constraints.

6 Additional Results by Solving the Optimization Problem by Satisfying Outlier Constraints

6.1 Results for Family A

6.1.1 Arrival Sequences where Sister Arrives First

We consider the arrival sequences where the sister arrives first to the database. The results for this set of sequences is displayed in Supplementary Figure 7. As she is the first member of the family in the database, her genome is stored without any withholding. We observe that the parents of the sister cannot be added to the database without revealing the parent-offspring relationship whereas the maternal aunt or person A can safely be added to the database if she or he arrives the second. If person A is the second added member, the maximum kinship value between the family members is 0.14. The relationship between person A and the sister, can be at most displayed as a second degree relative. In this arrival sequence $\{\text{Sister - Person A}\}$, the only subsequent member that can be added successfully is the maternal aunt. With the addition of maternal aunt, the $\Phi$ value in the family does not change, and the kinship value between siblings remains the same. There is no need to remove any SNPs from maternal aunt because she and her nephews are inferred as unrelated individuals due the low kinship values between them. This scenario achieves a maximum utility of 58% in the tree. If the data is shared without any removal for these three people, the utility value would be 60% for the whole family.

If the second added member is the maternal aunt, the $\Phi$ value increases to 0.054, where the sister and the aunt are denoted as third-degree relatives. However, after the maternal aunt, no family members can be added in the third order to the database.

Successful arrival orders for the sequences where the sister arrives first:

$\{\text{Sister - Person A - Maternal Aunt}\}$
$\{\text{Sister - Maternal Aunt}\}$
Supplementary Figure 7: The first arrived member is the sister. The possible arrival sequences of family $f_A$ is shown where the sister is the first arrived member. The optimization model is solved by relaxing kinship privacy constraints and satisfying outlier constraints.

6.1.2 Arrival Sequences where Maternal Aunt Arrives First

Supplementary Figure 8: The first arrived member is the maternal aunt. Solutions for arrival sequences for $f_A$ where the maternal aunt arrives first.

The tree of arrival sequences where the maternal aunt is the first comer is shown in Supplementary Figure 8. Maternal aunt’s closest relative is the mother. With the addition of the mother, the kinship value between the mother and the maternal aunt decreases up to 0.133 which denotes the kinship value between second degree relatives. However, after adding these two very close members, adding another family members is impossible.

The second-degree relatives of the maternal aunt i.e. her nephews, can be added safely as the second person achieving approximately 39% utility in both cases. In these cases, the maximum kinship coefficient among any member of the family will be around only 0.05; this is very close to the maximum kinship value that two unrelated individual can have which is 0.04 (see Table 1 [2]). Thus, this arrival sequence is quite favorable. However, in the following step when a third family member arrives, there is no feasible
solution.

There are cases where two sequences involve the same people, but the order in these sequences affect the solution significantly. For example, although \{Sister - Person A - Maternal Aunt\} sequence can be added safely, \{Maternal Aunt - Sister - Person A\} cannot be added in the database in this order.

**Successful arrival orders for the sequences where the maternal aunt arrives first:**
- \{Maternal Aunt - Sister\}
- \{Maternal Aunt - Person A\}
- \{Maternal Aunt - Mother\}

### 6.1.3 Arrival Sequences where Mother Arrives First

Supplementary Figure 9: The sequence where the first arrived member is the mother.
Solutions for arrival sequences for $f_A$ where the mother arrives first for the case where kinship constraints are relaxed and the outlier constrains are satisfied.

Supplementary Figure 9 displays the solutions for the arrival sequences where the mother is the first arrived family member. In the left and right hand side of the tree shows the cases where the second arrived members is the offspring of the mother. We observe it is impossible to add a parent and her offspring concurrently in the database without leading a privacy breach. The middle branch in the tree shows that if it is the maternal aunt arrives subsequently, she can safely added to the database. However, after her addition, it is not possible to add any other member in the family. The reason behind this is once again difficulty of preserving privacy in parent-offspring relations as their genomes bear extensive similarity. This situation is observed in the previously presented arrival sequences, a parent and an offspring should not be added concurrently to the database.

**Successful arrival orders for the sequences where the mother arrives first:**
- \{Mother - Maternal Aunt\}

### 6.1.4 Arrival Sequences where the Father Arrives First

Father has two blood-relatives in family $f_A$; person A and the sister. Supplementary Figure 10 shows that if children arrive to the database, they cannot be added to the database without a privacy leak. $\Phi$ value will be relaxed to 0.193 when the sister is added, and it will relaxed to 0.179 if person A is the one to arrive. In both cases, parent-offspring relationship does not decrease at least by one degree. Thus successful addition is not possible.
Supplementary Figure 10: The first arrived member in $f_A$ is the father. The possible arrival sequences of family $f_A$ is shown, where it is the father that arrives first. The tree represents solutions obtained by satisfying outlier constraints and relaxing kinship constraints.

6.2 Results for Family B

6.2.1 Arrival Sequences where Maternal Grandmother Arrives First

Supplementary Figure 11: The first arrived member is the maternal grandmother. The possible arrival sequences of family $f_B$ is shown where the maternal grandmother is the first arrived member. The optimization model is solved by relaxing kinship privacy constraints and satisfying outlier constraints.

In this scenario, the grandmother is the first family member that arrives to the database. As shown in Supplementary Figure 11 only family member that can be added is person B. Provided that person B is the second added member, the paternal grandfather can be added as the third family member. The maximum kinship value in the family will be 0.09, with a high utility 58.8%. Adding parents to the database at any time point makes the model infeasible.

Successful arrival orders for the sequences where the maternal grandmother first:
{Maternal grandmother - Person B - Paternal Grandfather}

6.2.2 Arrival Sequences where Paternal Grandfather Arrives First

The sequence tree of the arrival sequences where paternal grandfather arrives first is given in Supplementary Figure 12. This tree is very similar to the arrival tree where maternal grandmother arrives first.
Supplementary Figure 12: The sequence where the first arrived member is the paternal grandfather. Solutions for arrival sequences for $f_B$ where the paternal grandfather arrives first for the case where kinship constraints are relaxed and the outlier constrains are satisfied.

(Supplementary Figure [11]). The only difference between the two solutions is the difference in utility values, which is around 0.1%, when the third person is added. Note that since the number of SNP positions is large, 0.1% of this number corresponds to large number of SNPs.

Successful arrival orders for the sequences where the paternal grandfather arrives first:
{Paternal grandfather - Person B - Maternal grandmother}

6.2.3 Arrival Sequences where Mother Arrives First

Supplementary Figure [13] displays the arrival sequences and their solutions for the case where the mother is arrived first. This case is very constraining, after her arrival, no relatives can be added to the database. Although the model has a feasible solution, it is not enough to reduce the parent-offspring relationships by at least one degree.

Supplementary Figure 13: The first arrived member in $f_B$ is the mother. The possible arrival sequences of family $f_B$ is shown, where it is the mother that arrives first. The tree represents solutions obtained by satisfying outlier constraints and relaxing kinship constraints.

6.2.4 Arrival Sequences where Father Arrives First

Similar to the mother case, when the father arrives to the database first, no family member should be added to the database if the kinship privacy is to be preserved. The sequence tree of the father is given
in Supplementary Figure 14. None of the relationships between father and his relatives decreases at least one relationship degree.

**Supplementary Figure 14:** The first arrived member is the father. Solutions for arrival sequences for \( f_B \) where the father arrives first and where the outlier constraints are satisfied.

7 Results by Solving the Optimization Problem by Satisfying Kinship Constraints

7.1 Other Addition Sequences of Family A

7.1.1 Arrival Sequences where Sister Arrives First

Supplementary Figure 15 displays all possible arrival scenarios in which the sister is the first arrived member of the family. We observe that the solution is very similar to the case where person A arrives at the database as the first person (Figure 3). This is expected as the sister and person A have similar kinship distance to the family members. Differences arises in the outlier and the utility attained as the kinship estimates of the person A and the sister to other family member are not exactly the same. The most significant difference is observed in the outlier values of two sequences: \{Sister - Person A\} and \{Person A - sister\}. The outlier value for successful addition of the former sequence should be \( 2.25\sigma \) away and for the latter sequence it should be \( 2.75\sigma \) away from the \( o_{12} \) value. However, the total number of removed SNPs is equal. Indeed, the solution to this nonlinear optimization problem is not unique. As CPLEX returns one of the solutions randomly from all the possible solutions, they can end up different. Actually, the different \( o_{12} \) value is one of the alternative solutions.

**Successful arrival orders for the sequences where the sister arrives first:**

- \{Sister - Father\}
- \{Sister - Mother\}
- \{Sister - Maternal Aunt - Person A - Father\}
- \{Sister - Person A - Maternal Aunt - Father\}
- \{Sister - Person A - Mother - Maternal Aunt\}
- \{Sister - Person A - Mother - Father\}
- \{Sister - Person A - Father - Maternal Aunt\}
- \{Sister - Person A - Father - Mother\}
Supplementary Figure 15: The first arrived member is the sister. Solutions for arrival sequences for $f_A$ where the sister arrives first for the case where outlier constraints are relaxed and the kinship constraints are satisfied.

7.1.2 Arrival Sequences where Maternal Aunt Arrives First

Supplementary Figure 17 shows the arrival sequences and their corresponding solutions when the maternal aunt arrives at the databases as the first person. When the second added member is one of the nephews, four family members can be added. The three people sequences, \{Maternal Aunt - Person A - Sister\} and \{Maternal Aunt - Sister - Person A\}, sequences are similar as the last two people are nephews. However, here too the results of these sequences are slightly different from each other. The latter sequence achieves higher utility. These sequences can be extended with the arrival of the father but leads to major privacy leaks in the outlier values.

If the mother is the second arrived family member, sister and person A can successfully be added as the third member. However, if the nephews are added in the second and third place, it is observed that a fourth member can also be added to the database and this addition sequence ensures the maximum utility of 73.9% between all possible sequences.
Supplementary Figure 16: The first arrived member in \( f_A \) is the maternal aunt. The possible arrival sequences of family \( f_A \) is shown, where it is the maternal aunt that arrives first. The tree represents solutions obtained by relaxing outlier constraints and satisfying kinship constraints.

Successful arrival orders for the sequences where the maternal aunt arrives first:
- \{Maternal Aunt - Person A - Sister - Father\}
- \{Maternal Aunt - Sister - Person A - Father\}
- \{Maternal Aunt - Mother - Sister\}
- \{Maternal Aunt - Mother - Person A\}

7.1.3 Arrival Sequences where Mother Arrives First

When the mother is the first arrived family member in the database, person A, or the maternal aunt can be added as the subsequent member. The addition of the person A prevents further family members from being added as he is close to the other family members. On the other hand, the maternal aunt results with longer sequences: \{Mother - Maternal Aunt - Sister\} and \{Mother - Maternal Aunt - Person A\}. These sequences achieve approximately 56% utility. The latter sequence causes significantly lower levels of privacy leaks in outlier values but addition of a fourth member makes the model infeasible.
Successful arrival orders for the sequences where the mother arrives first:

\{Mother - Person A\}
\{Mother - Sister\}
\{Mother - Maternal Aunt - Sister\}
\{Mother - Maternal Aunt - Person A\}

7.1.4 Arrival Sequences where Father Arrives First

The father is only related to person A and the sister in the family. If the father is first arrived member to the database, addition of one of his offsprings results with very high privacy leaks in outlier $o_{11}$ values. The father’s tree is given in Supplementary Figure 18.

Successful arrival orders for the sequences where the father arrives first:

\{Father - Person A\}
\{Father - Sister\}

Supplementary Figure 18: The first arrived member is the father. Solutions for arrival sequences for $f_A$ where the father arrives first and kinship constraints are satisfied.
7.2 Other Addition Sequences of Family B

7.2.1 Arrival Sequences where Mother Arrives First

There are two relatives of the mother in $f_B$: maternal grandmother and person B, with both of whom she has parent-offspring relationships. The solutions from the optimization model for addition of these members require almost $0 \sigma_{11}$ values. Therefore, after the addition of second member, no one member should be added.

Supplementary Figure 19: The first arrived member in $f_B$ is the mother. The possible arrival sequences of family $f_B$ is shown, where it is the mother that arrives first. The tree represents solutions obtained by satisfying kinship constraints and relaxing outlier constraints.

Successful arrival orders for the sequences where the mother arrives first:
- {Mother - Person B}
- {Mother - Maternal grandmother}

7.2.2 Arrival Sequences where Father Arrives First

This scenario is very similar to the previous scenario in which the mother is the first arrived member. Father’s relatives are paternal grandfather and person B; the father has two parent-offspring relationships in the family. Addition of one of them requires masking a large number of SNPs, and a huge privacy leakage in terms of population statistics.

Supplementary Figure 20: The first arrived member is the father. Solutions for arrival sequences for $f_B$ where the father arrives first. The optimization model is solved by relaxing outlier privacy constraints and satisfying kinship constraints.
Successful arrival orders for the sequences where the father arrives first:
{Father - Person B}
{Father - Paternal grandfather}

7.2.3 Arrival Sequences where Maternal Grandmother Arrives First

Supplementary Figure 21 illustrates the possible arrival sequences in which the maternal grandmother is arrived at the database first. Mother or person B can be added as the subsequent person. Given that the mother is added second, a third family member cannot be added because the addition of the mother almost sets the \( o_{11} \) value to zero.

If the second incoming member is person B, arrival of him produces a feasible solution with 38.2% utility and 7.5\( \sigma \) decrease in \( o_{11} \) value. Following person B, it is not possible to add his parents. On the other hand, if \( o_{11} \) value decreases by 1.50\( \sigma \) and \( o_{10} \) decreases by 1.75\( \sigma \), the paternal grandfather can be added.

Supplementary Figure 21: The sequence where the first arrived member is the maternal grandmother. Solutions for arrival sequences for \( f_A \) where the maternal grandmother arrives first for the case where kinship constraints are satisfied and the outlier constrains are relaxed.

Successful arrival orders for the sequences where the maternal grandmother arrives first:
{Maternal grandmother - Mother}
{Maternal grandmother - Person B - Paternal grandfather}
### 7.2.4 Arrival Sequences where Paternal Grandfather Arrives First

Supplementary Figure 22: The first arrived member is the paternal grandfather. The possible arrival sequences of family $f_B$ is shown where the paternal grandfather is the first arrived member. The optimization model is solved by satisfying kinship privacy constraints and relaxing outlier constraints.

Supplementary Figure 22 displays the case where the paternal grandfather arrives first. In this scenario, we observe a tree very similar to Supplementary Figure 21. The longest sequence is \{Paternal Grandfather - Person B - Maternal Grandmother\}; which results in 57.6% utility and 9.25σ increase in $o_{11}$ and 1.75σ decrease in $o_{10}$ values. At the second level of tree, when person B is added, the utility value is 1% lower and 0.25σ more privacy leakage occurs compared to the previous scenario described in Section 7.2.3. This is because of high genomic similarity between the paternal grandfather and person B.

### References
