ALGORITHM

In general, within the framework of our algorithm, we want to achieve a compromise between the computational performance (as measured with the maximal bit size) and linkage power or accuracy of haplotype reconstruction. Our basic assumption is that the sub-pedigrees informative for multipoint linkage analysis or haplotypes reconstruction should contain closely related genotyped people, *i. e.* the total relationship between sub-pedigree members should be as large as possible for a given bit size. The genotyped persons are considered to be the subjects of interest (SOI).

Sub-pedigree building

Sub-pedigree is formed around central person, who is selected from {SOI}. Let closest relatives of a person are his/her parents and children. Denote the maximum bit size of sub-pedigree specified by user as Max_bit.

The algorithm includes two parallel procedures: sub-pedigree extension and estimation of its bit size.

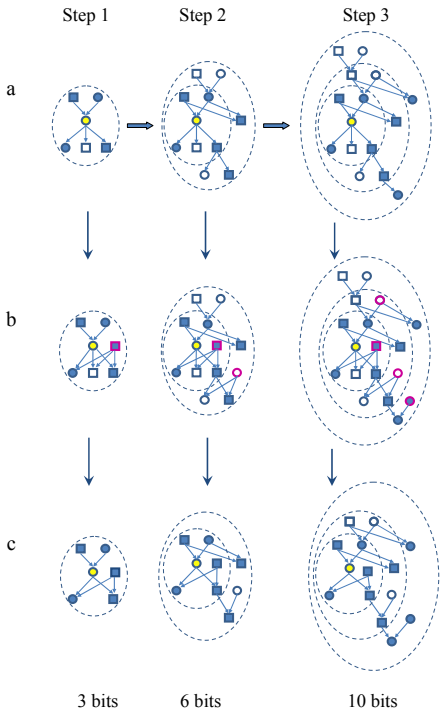

Fig. 1. Sup-pedigree building. Central person is indicated by yellow. Filled symbols denote genotyped pedigree members belonging to SOI, open symbols indicate unmeasured pedigree members who do not belong to SOI.
a) Step by step increasing of sub-pedigree size. Sub-pedigree consists of the central person, her parents and three children on the first step. Two grandparents of the central person, her brother and two grandchildren were added at the second step. Two grand grandparents of the central person, her aunt and a grand grandchild were added at the third step.
b) Pedigree structure reconstruction (marked by red). Husband of the central person is added after the first step. Husband of the central person and wife of her son are added after the second step. Husband of the central person, her son's wife and her grandson's wife are added after the third step.
c) Elimination of uninformative individuals. Unmeasured son of the central person is eliminated from the sub-pedigree after the first step. Unmeasured son of the central person and unmeasured parents of her mother are eliminated from the sub-pedigree after the second step. Unmeasured son of the central person and unmeasured grand grandparents are eliminated from the sub-pedigree after the second step.

Sub-pedigree extension

Closest relatives of the pedigree members, included in the sub-pedigree at the previous steps, are added to the sub-pedigree (Fig. 1a).

Bit size estimation

Bit size is estimated after every step of sub-pedigree extension. First of all we reconstruct sub-pedigree structure by adding of the second parent to each person having only one parent in the sub-pedigree (Fig. 1b). Then we eliminate all uninformative individuals (unmeasured unmarried offspring and unmeasured sub-pedigree founders having single child) from sub-pedigree (Fig. 1c). Bit size (BS) of modified sub pedigree is calculated as twice offspring number minus number of sub-pedigree founders.

Procedure of sub-pedigree is stopped when BS ≥ Max_bit. If BS = Max_bit, output sub-pedigree is the sub-pedigree obtained on the last step after its reconstruction and elimination of uninformative individuals. If BS > Max_bit, output sub-pedigree is the sub-pedigree obtained on the last but one step after its reconstruction and elimination of uninformative individuals. Several closest relatives may be added to some sub-pedigree members at the last step, if a bit size of resulting sub-pedigree does not outnumber Max_bit. Optimal selection of these relatives is defined by maximum increasing of total relationship of resulting sub-pedigree.

To decrease the running time and to prevent the loose sub-pedigree with large number of unmeasured individuals, we introduced special parameter which restricts number of "empty" steps, where size of sub-pedigree is not increased.

Pedigree splitting for haplotype reconstruction (PedStr_H program)

The aim of this program is splitting of a large pedigree into overlapping sub-pedigrees designed to reconstruct haplotypes for all genotyped pedigree members. We build sub-pedigrees around each genotyped person step by step including his/her closest relatives in the set of sub-pedigree members and controlling the sub-pedigree bit size

A set of these sub-pedigrees is sufficient to reconstruct haplotypes for all $i \in \{SOI\}$. However, some elements of this set are redundant and thus not necessary for reconstruction of haplotypes: some sub-pedigrees overlap and many individuals belong to several sub-pedigrees. To decrease a set of sub-pedigrees we estimate the sub-pedigree information for reconstruction of haplotype for every individual *i* in a given sub-pedigree. We use a sum of relationship coefficients between individual *i* and all members of a given sub-pedigree belonging to {SOI} as a measure of this information. We suppose that any sub-pedigree with maximum value of this information for individual *i* provides approximately equal amount of information for his/her

hyplotyping, because the most informative close relatives have to be included in any equally sized sub-pedigree with maximum value of information. We selected a minimum set of overlapping sub-pedigrees which guarantees inclusion of all $i \in \{SOI\}$ with maximum value of this information

Overlapping set of sub-pedigrees

Define optimal set of bit size limited overlapping sub-pedigrees as F.

1. $F \leftarrow 0$.
2. For each element $i \in \{SOI\}$:
   - construct sub-pedigree Q_i in accordance with algorithm described in "Sub-pedigree building";
   - define $\{SOI_i\}$ as a subset of $\{SOI\}$ belonging to this sub-pedigree;
   - for each element j of $\{SOI_i\}$ calculate $R_i(j)$ defined as a sum of relationship coefficients between j and every element of $\{SOI_i \backslash j\}$.
3. Construct matrix **R**

$$\begin{bmatrix} R_1(1) & R_1(2) & \cdots & R_1(f) & \cdots & R_1(N) \\ R_2(1) & R_2(2) & \cdots & R_2(f) & \cdots & R_2(N) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ R_m(1) & R_m(2) & \cdots & R_m(f) & \cdots & R_m(N) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ R_N(1) & R_N(2) & \cdots & R_N(f) & \cdots & R_N(N) \end{bmatrix},$$

   where each row corresponds to sub-pedigree (number of sub-pedigrees, N, equals to number of elements in $\{SOI\}$) and each column corresponds to each element of $\{SOI_i\}$. If the element of $\{SOI\}$ with number p is absent in $\{SOI_i\}$, $R_i(p) = 0$.
4. For each column f define the maximum value of $R_i(f)$, $i = 1, N$, and name it as R^f_{max}.
5. Construct matrix **T**

$$\begin{bmatrix} T_1(1) & T_1(2) & \cdots & T_1(f) & \cdots & T_1(N) \\ T_2(1) & T_2(2) & \cdots & T_2(f) & \cdots & T_2(N) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ T_m(1) & T_m(2) & \cdots & T_m(f) & \cdots & T_m(N) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ T_N(1) & T_N(2) & \cdots & T_N(f) & \cdots & T_N(N) \end{bmatrix}$$

   where $T_i(f) = 1$, if $R_i(f) = R^f_{max}$; otherwise $T_i(f) = 0$.
6. Identify the row m with maximum number of non-zero elements of matrix **T**.
7. $F \leftarrow F \cup Q_m$.
8. For every column f where $T_m(f) = 1$ define $T_j(f) = 0$ for all $j = 1, N$ except m.
9. For row m define $T_m(i) = 0$ for all $i = 1, N$.

Repeat steps 6-9 until $\boldsymbol{T}$ is zero matrix.

If there are more than one row with the same maximum number of non-zero elements on the step 6, the row, whose non-zero elements are more frequent in residuary rows, is selected. According to this algorithm, for every individual $f \in \{\text{SOI}\}$ there is one and only one resultant sub-pedigree m where $T_m(f) = 1$. This sub-pedigree is optimal for reconstruction of haplotype of f. In result of proposed algorithm a set of sub-pedigrees and a set of individuals haplotyped on the base of each of them are defined.

Pedigree splitting for IBD computation and multipoint linkage analysis (PedStr_L program)

The aim of this program is to split a large pedigree into non-overlapping sub-pedigrees for IBD computation and multipoint linkage analysis using Lander-Green-Kruglyak algorithm. In the first step we build sub-pedigrees around each genotyped person in the same way as for haplotype reconstruction. The sum of relationships between pairs of genotyped sub-pedigree members (total relationship) is calculated for all sub-pedigrees. We choose the sub-pedigree with maximum total relationship and exclude the members of this sub-pedigree from a set of SOI. We repeat this procedure until a set of SOI is empty.

If there is more than one sub-pedigree with equally high total relationship at some step, we select the one, which leaves the highest total relationship between remaining elements of SOI after elimination the members of this sub-pedigree.

The algorithm results in a set of sub-pedigrees, where every element of SOI is included as measured individual only in one sub-pedigree (non-overlapping set of sub-pedigrees). If an individual belonging to SOI was included in a sub-pedigree, it may be included in other sub-pedigree only as an unmeasured individual.