Suppose there is a LINE1 insertion to the reference genome. Due to historical mutations, some haplotypes may harbor additional SNPs on the LINE1 insertion. Strictly speaking, these further mutated LINE1 insertions are new alleles. However, when we study the frequency of this LINE1 insertion, we much prefer to group all LINE1 insertions into one group. This leads to the allele grouping problem below.

Suppose $n$ individuals have $m$ distinct alleles: $a_1, a_2, \ldots, a_m$ with $m \le n$. We write $a \sim b$ if allele $a$ and $b$ are similar according to a predefined similarity measurement[1]. $f(a_i)$ is the frequency of allele $a_i$. Without losing generality, let $f(a_1) \ge f(a_2) \ge \cdots \ge f(a_m)$. The allele grouping problem is to find a partition of alleles $\mathcal{A} = \{A_1, \ldots, A_g\}$ such that for $a \in A_i$, there exists $b \in A_i$ satisfying $a \sim b$.

Algorithm 1 is a simple greedy solution to the allele grouping problem. After grouping, the frequency of an allele group $A$ is defined as $f(A) = \sum_{a \in A} f(a)$. The sum of minor allele frequencies (sMAF) of $\mathcal{A}$ is $1 - \max_{A \in \mathcal{A}} f(A)$, which, unlike MAF, can be larger than 0.5.

**Algorithm 1:** A greedy solution to the allele partition problem

```
Input: A list of alleles $a_1, \ldots, a_m$ with $f(a_1) \ge \ldots \ge f(a_m)$
Output: Partition $\mathcal{A}$
$\mathcal{A} \leftarrow \{\}$
for $i \leftarrow 1$ to $m$ do
    for $A \in \mathcal{A}$ in the descending frequency order do
        if $\exists a \in A$ such that $a_i \sim a$ then
            $A \leftarrow A \cup \{a_i\}$                 ▷ Add $a_i$ to an existing group
            break
    if $a_i$ has not been added then
        $\mathcal{A} \leftarrow \mathcal{A} \cup \{\{a_i\}\}$       ▷ Create a new allele group
```

[1] E.g. the length difference between $a$ and $b$ is smaller 50bp. Alignment- or path-based measurement would be more accurate.