Note on the handling of SNPs that have been genotyped (forced selection) and SNPs that have problematic assay design (forced non-selection)

(1) Forced selection

If some SNPs have been genotyped in the study sample prior to the selection of tag SNPs, then in a sense these are already selected as tag SNPs. The tag SNP selection procedure therefore needs to be "forced" to include these SNPs. This is done by modifying certain elements of the asymmetric matrix of "similarities" that determine whether SNP *i* is able to tag SNP *j* (R[*ij*], for all *i* and *j*). Now let SNP *d* be one of the already genotyped SNPs. We can set R[*id*] = 0 for $i \neq d$, and R[*dd*] = 1. This guarantees that SNP *d* can be tagged by itself but not the others, thereby ensuring the selection of SNP *d* as a tag SNP.

(2) Forced non-selection

Some SNPs may be difficult to assay under a genotyping platform due to local sequence features. These "non-assayable" SNPs should not be selected as tag SNPs but rather should be tagged by other SNPs if possible.

2.1 "Non-assayable" SNPs that cannot be tagged by any "assayable" SNP will remain untagged no matter which SNPs are selected as tags. They are therefore listed and excluded from further consideration.

2.2 Since some "non-assayable" SNPs can be tagged only by certain SNPs, and these SNPs might not be selected as tag SNPs by the clustering algorithm, there is a chance that these "non-assayable" SNPs might be kept untagged at the end of the clustering procedure. With regard to this, we have a few measures (as mentioned below) to make certain that, for each "non-assayable" SNP, there is at least one tag SNP companion as long as it really exists.

2.2.1 Firstly, if there are already-genotyped SNPs, they are checked to see if they can tag the "non-assayable" SNPs. The "non-assayable" SNPs that are not already tagged constitute the set of untagged "non-assayable" SNPs.

2.2.2 Each "assayable" (but not already genotyped) SNP is checked for the number of untagged "non-assayable" SNPs that it is able to tag; the one with the largest number is marked out as a SNP for forced selection. The "non-assayable" SNPs that can be tagged by this SNP are then removed from the set of untagged "non-assayable" SNPs.

2.2.3 If there remain one or more untagged "non-assayable" SNPs, then step 2.2.2 is repeated until no untagged "non-assayable" SNP (with possible tag SNP companion) is left.

2.3 The SNPs selected in steps 2.2.2 and 2.2.3 will be treated in the same way as SNPs that have already been genotyped, i.e. forced selection by Procedure (1).

2.4 Let *n* be a "non-assayable" SNP, we set R[*ni*] = 0 for all *i* including *i*=*n*. This guarantees that "non-assayable" SNPs will not be selected as tag SNPs.

After the Procedures (1) and (2), the clustering algorithm for selecting tag SNPs can proceed.