多位点单倍体型频率估计

单核苷酸多态性（single nucleotide polymorphism，SNP）主要是指在基因组水平上由单个核苷酸的变异所引起的 DNA 序列的多态性，是人类可遗传的变异中最常见的一种。而单倍体型 （Haplotype）是指同一条染色体上紧密相连的一些位点（或基因）的组合。不同位点间排列距离越近，在减数分裂时发生重组的可能性就越小，以同一单倍体型的整体传递给子代的可能性也就越大。

SNP 单体型就是指不同 SNPs 位点上核苷酸碱基的线性排列，每一种线性排列称为一种 SBP 单体型。例如：如果在某一段 DNA 片段上发现了 10 个 SNPs，理论上可能存在 1024（2^{10}）种单体型，但由于连锁不平衡的存在，实际发现的单体型的数目远远少于理论上的值。人类绝大多数疾病都存在遗传基础，而其中许多疾病如高血压和冠心病等并不是由单个 SNP 位点的突变引起的，而是由若干个位点上的 SNPs 联合作用导致的，亦即 SNP 单体型是影响疾病发生的遗传单位。

单倍体型分析可以帮助研究者进一步验证在连锁分析中所取得的数据，排除假阳性；通过多个位点的单倍体型排列，还可以使某些多态性不够充分的单个位点的信息得以应用，多位点单体型分析能够发现单体型-疾病表型之间明显强于单个位点-疾病表型的关联。单倍体型分析含有较大的信息量，而且充分考虑多个位点连锁不平衡的信息，被证明是一条便捷、有效的候选基因分析方法。

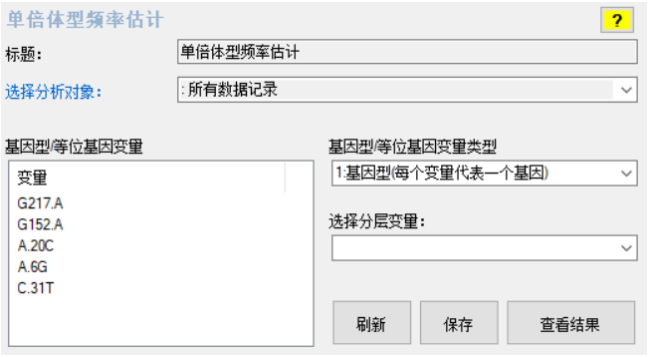
研究者在处理 SNP 单体型数据时，常常需要估计单倍体型的频率。易侕的多位点单倍体型频率估计模块采用 R 里的最大似然估计法估计单倍体型频率。

例 1，某研究选取血管紧张素原（angiotensinogen, AGT）基因启动子-217，-152，-20，-6， 内含子 1 的+31，共 5 个位点，对 100 例正常人进行了 SNP 位点基因型检测，下载数据：

http://www.empowerstats.com/empowerStats/exdata/hardyweinberg_data.xls

其中 5 个 SNP 位点的基因型分别为：G217.A 为 AA，AG 和 GG；G152.A 为 AA，AG 和 GG，A.20C 为 AA，AC 和 CC，A.6G 为 AA，AG 和 GG，C.31T 为 CC，CT 和 TT，均编码为 0、1、2。试求得可能存在的单体型，并计算单体型概率。

输入界面：



如果原始数据中是基因型变量，则应该在“基因型/等位基因变量类型”下选择“1:基因型(每个变量代表一个基因)”；如果是等位基因变量，则应该选择“2:等位基因(按顺序 2 个变量代表一个基因)”。

输出结果：

单倍体型频率估计

#	单倍体型编码	G217.A	G152.A	A.20C	A.6G	C.31T	频率
1	1	1	1	1	1	1	0.5447129
2	3	1	1	1	2	1	0.1263794
3	2	1	1	1	1	2	0.0972353
4	9	1	2	1	1	2	0.0443125
5	17	2	1	2	2	2	0.0374048
6	16	2	1	2	2	1	0.0312797
7	5	1	1	2	1	1	0.0273825
8	11	2	1	1	1	1	0.0163130
9	15	2	1	2	1	2	0.0155047
10	14	2	1	2	1	1	0.0144343
11	13	2	1	1	2	2	0.0107744
12	4	1	1	1	2	2	0.0102960
13	7	1	1	2	2	1	0.0088381
14	8	1	2	1	1	1	0.0056600
15	6	1	1	2	1	2	0.0051558
16	12	2	1	1	1	2	0.0042889
17	10	1	2	1	2	2	0.0000275

对数似然值	似然比检验（卡方值）	自由度	P 值
-282.745709850361	87.0198780987903	11	6.38378239159465e-14

结果解释：

输出的第一个表格中列出了可能出现的 17 种单倍体型，并且针对单倍体型出现的概率进行了从大到小的排列。表中的数字 1 或 2 代表每个 SNP 的两种野生型与变异型，须结合变量编码来判断分别代表什么。如 G217-A 原编码 0、1、2 分别代表 AA，AG 和 GG，野生型为 A，变异型为 G；C+31T 原编码 0、1、2 分别代表 CC，CT 和 TT，野生型为 C，变异型为 T。最有可能出现的两种单倍体型为：1-1-1-1-1 即为 A-A-A-A-C，1-1-1-2-1 即为 A-A-A-G-C。第二个表格是对 5 个位点是否存在连锁不平衡关系进行假设检验，结果：对数似然值为-282.7，卡方值为 87.0，自由度为 11，P 值为 $6.38*10^{-14}$。因此，5 个位点存在较强的连锁不平衡关系。