Solutions to exam questions

1. **Consider two biallelic loci in a population of diploid chromosomes (i.e., population of haplotype pairs), the first locus with alleles A and a, and the second locus with alleles B and b. Let p_{AB} be the population frequency of the haplotype AB. Let $p_{A/B}$ be the population frequency of the "cross-haplotype" A/B (that is, A occuring in one chromosome and B in the other chromosome of the diploid). Show that**

$$p_{AB} + p_{A/B} = 2p_{AABB} + p_{AABb} + p_{AaBB} + \frac{1}{2}p_{AaBb},$$

where, in general, p_{uvxy} is the population frequency of haplotype pairs that have exactly the four alleles u, v, x, and y, that is, p_{AABb} is the total population frequency of the pairs (AB, Ab) and (Ab, AB).

Grading: **reasonable partial solutions with correct arguments yield 1–2 points; correct proof yields 3 points.**

Denote by $n_{ux/vy}$ the population count of the haplotype pair $\{ux, vy\}$. Denote by n_{AB} and $n_{A/B}$ the population counts $2np_{AB}$ and $2np_{A/B}$ of the haplotype AB and the cross-haplotype A/B, respectively. Let the number of diploids be n. We have

$$\begin{aligned} n_{AB} &= 2n_{AB/AB} + n_{AB/Ab} + n_{AB/aB} + n_{AB/ab}\,, \\ n_{A/B} &= 2n_{AB/AB} + n_{AB/Ab} + n_{AB/aB} + n_{Ab/aB}\,. \end{aligned}$$

On the other hand,

$$\begin{aligned} n_{AB/AB} &= np_{AABB} \\ n_{AB/Ab} &= np_{AABb} \\ n_{AB/aB} &= np_{AaBB} \\ n_{AB/ab} + n_{Ab/aB} &= np_{AaBb}\,. \end{aligned}$$

Thus,

$$p_{AB} + p_{A/B} = (n_{AB} + n_{A/B})/(2n) = 2p_{AABB} + p_{AABb} + p_{AaBB} + \frac{1}{2}p_{AaBb}\,,$$

as claimed.

2. **Suppose the haplotype frequencies p_h for $h \in H_0 = \{0,1\}^m$ and the genotyping error models $\tau_s(g_s; g'_s)$ for $g_s, g'_s \in \{0,1,2\}$, $s = 1, 2, \ldots, m$, are fixed and given. Assume HWE (as usual). Describe a method that given a trio genotype $\mathbf{g} = (g^{\mathrm{m}}, g^{\mathrm{f}}, g^{\mathrm{c}})$ outputs a most probable haplotype pair for each member of the trio; for instance, for the child, it outputs a pair $(i, j) \in H_0 \times H_0$ such that**

$$\Pr(\textbf{the child's haplotype pair is } (i,j) | \mathbf{g})$$

is maximized; analogously for the mother and the father. Analyze the time complexity of your method.

Grading: **reasonable and correct math expressions yield** $+1$ **point; correct analysis of the time complexity yields** $+1$ **point; correct and sufficient arguments for the time bound** $O(m4^m)$ **yield** $+1$ **point.**

Consider first the child's haplotype pair (i, j). To maximize $\Pr(i, j|\mathbf{g}) = \Pr(\mathbf{g}, i, j)/\Pr(\mathbf{g})$ it suffices to maximize the denominator

$$\begin{aligned}\Pr(\mathbf{g}, i, j) &= \sum_{i'}\sum_{j'} p_i p_j p_{i'} p_{j'} \tau(g^{\mathrm{m}}; i + i')\tau(g^{\mathrm{f}}; j + j')\tau(g^{\mathrm{c}}; i + j) \\ &= p_i p_j \tau(g^{\mathrm{c}}; i + j)\left(\sum_{i'} p_{i'}\tau(g^{\mathrm{m}}; i + i')\right)\left(\sum_{j'} p_{j'}\tau(g^{\mathrm{f}}; j + j')\right) \\ &= p_i p_j \tau(g^{\mathrm{c}}; i + j) a_i b_j \,,\end{aligned}$$

where a_i and b_j are defined in the obvious way. Now the a_i for all $i \in H_0$ and the b_j for all $j \in H_0$ can be computed in a straighforward manner in time $O(m(2^m)^2)$. Then for the maximing pair (i, j) can be found by simply looping through each of the 4^m pairs (i, j) and computing the product $p_i p_j \tau(g^{\mathrm{c}}; i + j) a_i b_j$ for each in time $O(m)$. The total running time is thus $O(m4^m)$.

Similarly, for the mother's haplotype (i, i') the quantity to be maximized can be written as

$$\begin{aligned}\Pr(\mathbf{g}, i, i') &= \sum_{j}\sum_{j'} p_i p_j p_{i'} p_{j'} \tau(g^{\mathrm{m}}; i + i')\tau(g^{\mathrm{f}}; j + j')\tau(g^{\mathrm{c}}; i + j) \\ &= p_i p_{i'} \tau(g^{\mathrm{m}}; i + i')\left(\sum_{j} p_j \tau(g^{\mathrm{c}}; i + i')\left(\sum_{j'} p_{j'}\tau(g^{\mathrm{f}}; j + j')\right)\right) \\ &= p_i p_{i'} \tau(g^{\mathrm{m}}; i + i') c_i \,,\end{aligned}$$

where the c_i for each $i \in H_0$ can be computed by evaluating the two nested sums in two phases in time $O(m4^m)$. Finally, the maximing (i, i') can be found as in the case of the child, taking time $O(m4^m)$ in total.

The case of the father is analogous to the case of the mother.

2. Present a pseudo code for an algorithm that for a given list of n **haplotypes over** m **SNPs with no missing data, outputs a tagging set according to the Patil et al. definition with** $\beta = 0.80$**. What is the asymptotic running time of your algorithm in terms of** n **and** m**?**

Grading: **basic understanding of the key concepts demonstrated yields** $+1$ **point; correct algorithm with sufficient description yields** $+1$ **point; correct and rather tight analysis of the runtime yields** $+1$ **point.**

Recall a subset S of the m markers is called a tagging set if the markers in S suffice for uniquely distinguishing at least 80% of the haplotypes and the size of S is the smallest possible.

The algorithm visits all subsets $S \subseteq \{1, \ldots, m\}$ in increasing order of the size $k = |S|$ and terminates as soon as S uniquely distinguishes at least 80% of the haplotypes, that is, the following subroutine returns `true`. Denote by $h[j]$ the jth input haplotype and by $h_S[j]$ its allele sequence along the markers in S.

1. Initialize a data structure that stores integer pairs (i, d) indexed by allele sequencies of length k. Initialize a global variable u to 0.

2. For $j := 1, 2, \ldots, n$:

   (a) if there is no pair indexed by the projection $h_S[j]$, then add the pair $(j, 1)$ and increase u by 1; else fetch the pair (i, d) and do the following: if $h[j]$ differs from $h[i]$ or $d = 0$, then replace the pair with $(j, 0)$ and decrease u by d; else replace the pair with $(j, d + 1)$ and increase u by 1.

3. If $u \geq 0.80n$, then return `true`.

Note that for each "tag haplotype" the value d is the number of uniquely distinguished haplotypes seen so far.

The running time of the algorithm, for a fixed S, is obviously $O(nm)$. So the total time requirement is $O(nm2^m)$ in the worst case.

3. When the number of (biallelic) markers m is large, the number of potential haplotypes, 2^m, may prohibit the application of the presented techniques for haplotype inference. There is a popular heuristic approach, called *partition ligation*, to address this issue. The idea is to infer the possible candidate haplotypes separately for the first and the last $m/2$ markers; usually the number of haplotypes with nonzero frequency estimates is much below $2^{m/2}$ for each of the two parts, say ℓ and r for the first and the last part, respectively. Then the haplotype frequencies over the complete set of m markers is inferred under the supposition that non-zero frequencies are hold only by haplotypes that belong to the $k := \ell r \ll 2^m$ possible combinations of the already inferred shorter haplotypes.

Consider unrelated genotype and the model extended by a deletion haplotype. Show that each iteration in the EM algorithm for estimating the frequencies of the k possible haplotypes can be computed in time $O(k(\ell+r)m)$ (per observed unrelated genotype).

***Grading:* sufficient argumentation for the time bound $O(k^2m)$ yields 1 point; sufficient argumentation for the bound $O(k(\ell + r)m)$ yields 3 points.**

By Proposition 2 of the lecture notes of Week 3 the task is to compute, for any fixed genotype g, the values

$$I_h(g) = \sum_{h'} p_h p_{h'} \tau(g; h, h')$$

for all h, and their sum

$$L(g) = \sum_{h} I_h(g)\,;$$

here h and h' range through all valid haplotypes over the m markers, that is, the deletion haplotype and the $k = \ell r$ combined haplotypes. Indeed, for any other haplotype h the frequency p_h vanishes by the assumption. Clearly, computing the values $I_h(g)$ determines the time complexity. Now, a straightforward algorithm computes each $I_h(g)$ separately in time $O((k + 1)m)$, hence in time $O((k + 1)^2m) = O(k^2m)$ in total.

To speed up the algorithm we make use of the factorization

$$\tau(g; h, h') = \prod_{s=1}^{m} \tau_s(g_s; h_s, h'_s)\,.$$

In fact, we will only need the factorization

$$\tau(g; h, h') = \tau_L(g_L; h_L, h'_L)\tau_R(g_R; h_R, h'_R)$$

with

$$\tau_L(g_L; h_L, h'_L) = \prod_{s\in L} \tau_s(g_s; h_s, h'_s)\,, \qquad \tau_R(g_R; h_R, h'_R) = \prod_{s\in R} \tau_s(g_s; h_s, h'_s)\,,$$

where L denotes the first $m/2$ and R the last $m/2$ of the m markers, and the subscripts indicate the corresponding left and right parts of the genotype g and the haplotypes h and h'.

Suppose for a moment that deletion haplotypes are ignored. Then each haplotype h is a combination of its left part i and right part j; write ij for h and $i'j'$ for h'. We have

$$\begin{aligned} I_{ij}(g) &= \sum_{i'}\sum_{j'} p_{ij}p_{i'j'}\tau_L(g_L; i, i')\tau_R(g_R; j, j') \\ &= p_{ij}\sum_{i'}\tau_L(g_L; i, i')\left(\sum_{j'} p_{i'j'}\tau_R(g_R; j, j')\right). \end{aligned}$$

Observe that for fixed j and i' the sum over j', denoted by $a_{i'j}(g)$, can be computed in time $O(rm)$; thus all the values $a_{i'j}(g)$ in time $O(\ell r^2 m)$ in total. Then for fixed i and j the remaining sum over i' can be computed in time $O(\ell m)$; thus for all i and j in time $O(\ell^2 rm)$ in total. Combining gives a running time of $O(\ell r(\ell + r)m) = O(k(\ell + r)m)$.

Finally, consider the possibility of deletion haplotypes. If $h = ij$ is the deletion haplotype, then $I_h(g)$ can be computed by looping through all combinations $i'j'$ where either both i' and j' are deletion haplotypes or neither of them is ("partial deletion haplotypes" are not valid in our model). This takes time $O(km)$. If i and j are not deletion haplotypes, we need to add to the previsouly computed $I_h(g)$ the term corresponding to the case when both i' and j' are deletion haplotypes. For each fixed ij this addition takes time $O(m)$, hence time $O(km)$ in total.

4. Read the article

- **Stefansson H et al., A common inversion under selection in Europeans. Nature Genetics 37, 129 - 137 (2005)**

and answer the questions below. A link to the article can be found on the web page of the course; if you do not manage to download the article, contact the instructor for getting a copy.

(a) Which one of the two haplotype lineages H1 and H2 is more homogeneous and which one more diverse?

(b) What observations or experiments speak for non-neutrality (i.e., positive or negative selection) of the inversion polymorphism? (Use at most 100 words in your answer.)

***Grading:* correct and relevant answers yield (a) +1 point, (b) +1–2 points.**

(a) H1 is more diverse and H2 more homogeneous.

(b) Mother carrying H2 have higher recombination rate and more children (however, homozygous carries of H2 have fewer children than heterozygous carries). A simulation study under the neural model produced results that deviate from the observations, thus supporting non-neutrality.

5. Read the article

- **Sindi SS, Raphael BJ, Identification and frequency estimation of inversion polymorphisms from haplotype data. J Comput Biol 17, 517 531 (2010)**

and answer the questions below. A link to the article can be found on the web page of the course; if you do not manage to download the article, contact the instructor for getting a copy.

(a) What is the key difference or advancement of the proposed method as compared to an earlier method by Bansal et al. (2007) cited in the article?

(b) How good is the method at estimating the population frequency of an inversion? (Describe strengths and weaknesses.)

***Grading:* correct and relevant answers yield (a) +1 point, (b) +1–2 points.**

(a) At a putative inversion boundary, Bansal et al. measure the *total* LD between the haplotypes inside the inversion region and outside the inversion region. In particular, Bansal et al. do *not* consider grouping the haplotypes into two groups: one for inverted and the other one for normal haplotypes. Sindi and Raphael implement such grouping and then measure LD only between the corresponding groups; technically, this leads to a mixture model and an EM algorithm.

(b) In a simulation study, the predicted inversion frequencies are accurate when the inversion is 500 Kb or large and its true frequency is between 20% and 80%. However, for rare ($< 10\%$) or very common ($> 90\%$) inversion the predicted frequencies are notably inaccurate; this is probably due to the small sample size (in the HapMap data). Comparison to known inversion polymorphims shows that the predicted frequencies are in good agreement with estimates obtained by other (non-computational) methods on small samples of chromosomes.