Graph-Based Imputation

Input: 1) Target locus set, e.g. ABCQR
2) ambiguous HLA data, possibly with missing loci (as compared to target locus set)

Output: 1) multilocus genotype list: list of unambiguous haplotype pairs

Handling ambiguous HLA input data

1) Convert ambiguous HLA into a list of unambiguous unphased multilocus genotypes
2) Convert each unambiguous unphased multilocus genotype into a list of unambiguous haplotype pairs representing each of its (2^{n-1}) phases (n = number of input loci)
3) Use unambiguous haplotype pairs (possibly with missing loci as compared to target locus set) as input to "Plan A" imputation method, using an initial ε-value, e.g. $\emptyset.\emptyset1$.
4) If "Plan A" imputation method fails when using the initial ε value, try again using a lower ε value, e.g. $\emptyset.\emptyset\emptyset1$. If that fails, try a few more times, using lower and lower ε values (e.g. $\emptyset.\emptyset\emptyset\emptyset1$, $\emptyset.\emptyset\emptyset\emptyset\emptyset1$). If that fails, try again using $\varepsilon=\emptyset$.
5) Only if "Plan A" imputation method yields no genotype frequencies (via haplotype pairs) for the target locus set, use "Plan B" imputation method to generate diaplotype candidates and frequency estimates for all input haplotypes which could not be imputed using "Plan A".

"Plan A" method, e.g. Find top k ABCQR haplotypes for unambiguous ABR haplotype, having frequency $>\varepsilon$

"Plan B" method, e.g. for ABR haplotype not found in graph
Consider all ABR direct subcomponents

P(AB) P(R) P(CQ)
P(AR) P(B) P(CQ)
P(BR) P(A) P(CQ)

← If no answer (0 frequency) break subcomponents to next level and try again

For subcomponent freq, use value from graph, if found, otherwise $\emptyset$

for allele freq, use value from graph if found, otherwise use α_1

for missing input, use top k candidates from population (taking no info from input HLA)

Priors

$\alpha_\emptyset$ = unobserved ABCQR haplotype frequency

α_1 = unobserved allele frequency

Populate graph using $\tilde{P}_i$, where (for haplotype)

$$\tilde{P}_i = \frac{P_i + \alpha_\emptyset}{Z}$$

Z = normalization factor, e.g. 1.5

So, $Z = \sum \tilde{P}_i$

For unobserved haplotypes,

$$\tilde{P}_i = \frac{\alpha_\emptyset}{Z}$$