**Allele Frequency / Antigen Frequency**

- Terminology not well standardised
  - allele frequency \( g_i \) = frequency of the gene in the theoretical gene pool of a population: \( \sum g_i = 1 \)
  - “antigen” frequency \( f_i \) = observed frequency of a „trait“ in the individuals of a population: \( \sum f_i = 1 \)
  - For a population in Hardy-Weinberg-Equilibrium the following holds: \( f_i = g_i \)
- For codominant traits, \( f_i \) can be estimated by counting.
- Counting \( g_i \) is hampered by individuals with only one „marker“ without proof of homozygosity (missing data).

**Haplotype Frequency Estimation (HFE)**

- Missing data is a common problem in statistics.
- Types of missing information in HFE
  - phase of alleles
  - typing ambiguities
  - undetectable alleles
- Dealing with missing information requires some assumptions (here: Hardy-Weinberg-Equilibrium).
- General approach: Maximum-likelihood and expectation-maximisation-algorithm.

**Comparing Allele / Antigen Frequencies**

- Methodology
  - Comparing two Bernoulli-Distributions
    - \( \chi^2 \)-test
    - Fisher’s exact test
  - Correcting for multiple comparisons
    - Bonferroni correction
    - Holm–Bonferroni method
- Main Caveats
  - Depending on the underlying raw data, for allele frequency data the additional variance induced by the estimation process has to be taken into account.
  - Are your control frequencies really suitable?
  - Did you fix the whole study approach including any hypothesis in writing before starting the study?
Comparing Haplotype Frequencies

Complicated field with many frequently overlooked pitfalls!
• The number of haplotypes in a population is too big to use a hypothesis-free “fishing approach” in combination with the Holm-Bonferroni method.
• Additional variance induced by the estimation process generally relevant.
• Genetic distances must be used with care due to the high fraction of “noise” at low frequencies.
• Bootstrapping based methods can be used to overcome these difficulties.

Haplotype Frequencies
Application in HSC Donor Search
• The majority of HSC donors on match lists have ambiguous HLA data (serology, multiple allele codes, missing data for relevant loci).
• Known haplotypes can reduce this ambiguity by many orders of magnitude (see example: $10^{11}$).
• Haplotype frequencies can be used to calculate matching probabilities for ambiguous loci.
• Predictions are excellent with some caution necessary at the low end.

Match List Example

Match List with Probabilities