Association analysis with haplotypes

Family based methods for the discovery of linkage and association depend on being able to calculate the probability of the child genotypes given the genotypes of the parents. These calculations are only meaningful if there is more than one allele that the children might inherit from the parents, ie the parents need to be heterozygous in order for them to be informative at a particular locus.

Since SNP only have two alleles the frequencies of heterozygote parents will generally be low. For example, with the highest possible minor allele frequency (MAF) of 50%, 50% of parents will be heterozygous and only 25% of families will have two heterozygous parents. With a MAF of 20%, 32% of parents will be heterozygous and only 10% of families will have both parents heterozygous. The frequency of heterozygotes can be increased by using sets of SNP in linkage disequilibrium that form haplotypes and that also have large numbers of alleles.

In order to test the viability of using haplotypes for the association analysis we ran a simulation using 1,000 genomes project data. SNP loci that are represented on the H3Africa SNP chip were extracted from 1,000 genomes project African populations. The BigLD R package (Kim et al., 2018) was used to identify blocks of linked SNP. We then used a custom Perl script to extract the genotypes at all H3Africa SNP chip loci in each block. Blocks greater than 8 SNP in length were broken into blocks of 4 SNP because long haplotype blocks had many alleles with frequencies that were too low for associations to be detectable with the numbers of samples that we are likely to collect. The script then obtained the frequency of each haplotype and hence the expected frequency of heterozygous individuals (F_e), $F_e = 1 - \sum_{h=1}^{h=n} \left(\frac{x_h}{2p}\right)^2$ where n is the number of different haplotypes in a haplotype block, x_h is the number of observations of the h^{th} haplotype and p is the population size. Supplementary Table 1 contains a list of haplotype blocks in each gene and some summary statistics for each block. Supplementary Table 2 contains a list of all distinct haplotypes and their frequencies.

Three hundred and seventy-eight haplotype blocks were identified in the 24 candidate genes with a mean of 16.4 (median 5) haplotype blocks per gene. One gene IL23A had no identifiable blocks as all SNP within the gene had minor allele frequencies of less than 5%. The mean number of haplotype alleles in a block was 8.6 (median 9), range 2 to 16 haplotype alleles per block (Fig 1A). Most haplotypes had low allele frequencies (mean 11.8%) (Fig 1 B) and consequently most individuals (mean 77%, median 80%) were heterozygous at any given haplotype block (Fig 1 C).

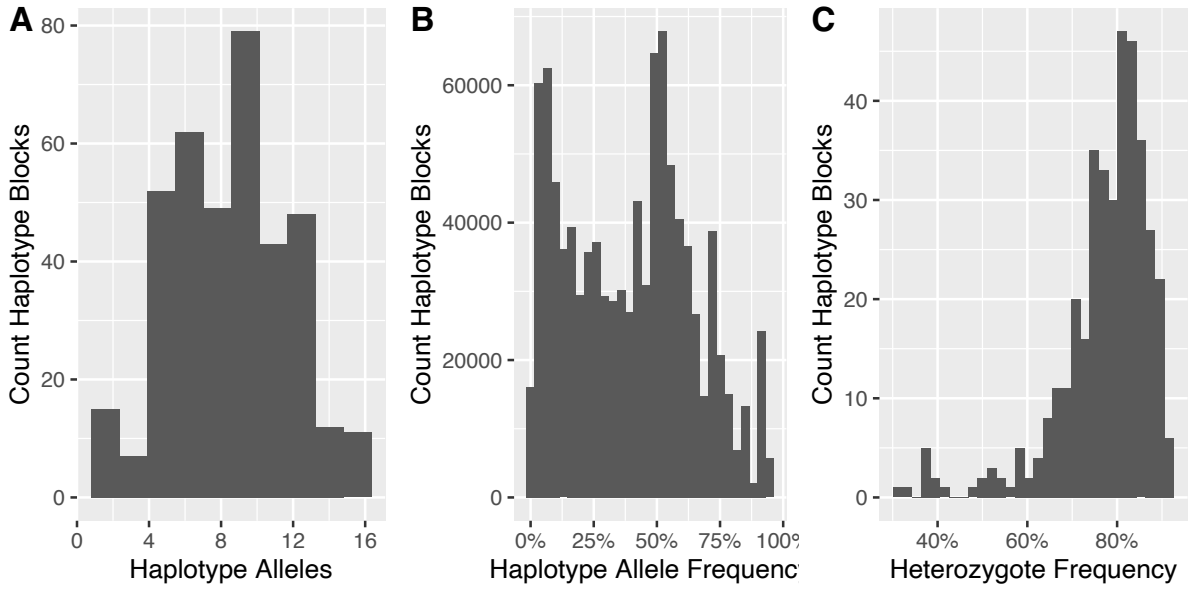

Figure 1. Counts and frequencies of haplotype alleles and heterozygotes in 1000 Genomes Project African samples at candidate gene loci. **A** Counts of different haplotype alleles per block. **B** Histogram of haplotype allele frequencies **C** Histogram of heterozygote frequencies. Most haplotype blocks had high frequencies of heterozygotes.

A power analysis was conducted using the powerpkg library in R (Weeks, Daniel E, 2012), that implements the power calculations for the TDT of Abel and Muller-Myhsok (Abel & Müller-Myhsok, 1998) for three levels of haplotype relative risk (HRR) (1.5, 2.0, 3.0). For each level of relative risk, the number of families required to have an 80% chance of detecting a significant association at the 95% confidence level was calculated after correcting for the 378 haplotype loci being tested. The calculation was repeated for allele frequencies between 0.05 and 0.95 (Fig 2A). It was assumed that the marker haplotype frequency was the same as the functional variant frequency and that LD between marker haplotype and functional variant was 1. These assumptions are unlikely to be realistic and therefore the predicted numbers of families required will be minima.

The power analysis showed that 250 families would provide at least 80% power to detect a HRR of 2 for haplotypes with allele frequencies between 16% and 72% (Fig 2A). 62% of chromosomes have haplotypes with an allele frequency in this range (Fig 2B). However, at 46 out of the 378 loci no haplotypes had allele frequencies in this range (Fig 2C) and therefore only large HRR (HRR > 2) are likely to be detectable at these loci. Despite this at 80% of haplotype blocks over 50% of chromosomes had a haplotype with a frequency in the informative range (16-72%) (Fig 2D). In summary 250 families will provide at least 80% power to detect HRR > 2 at the 80% of blocks that have haplotypes with allele frequencies between 16 and 72%.

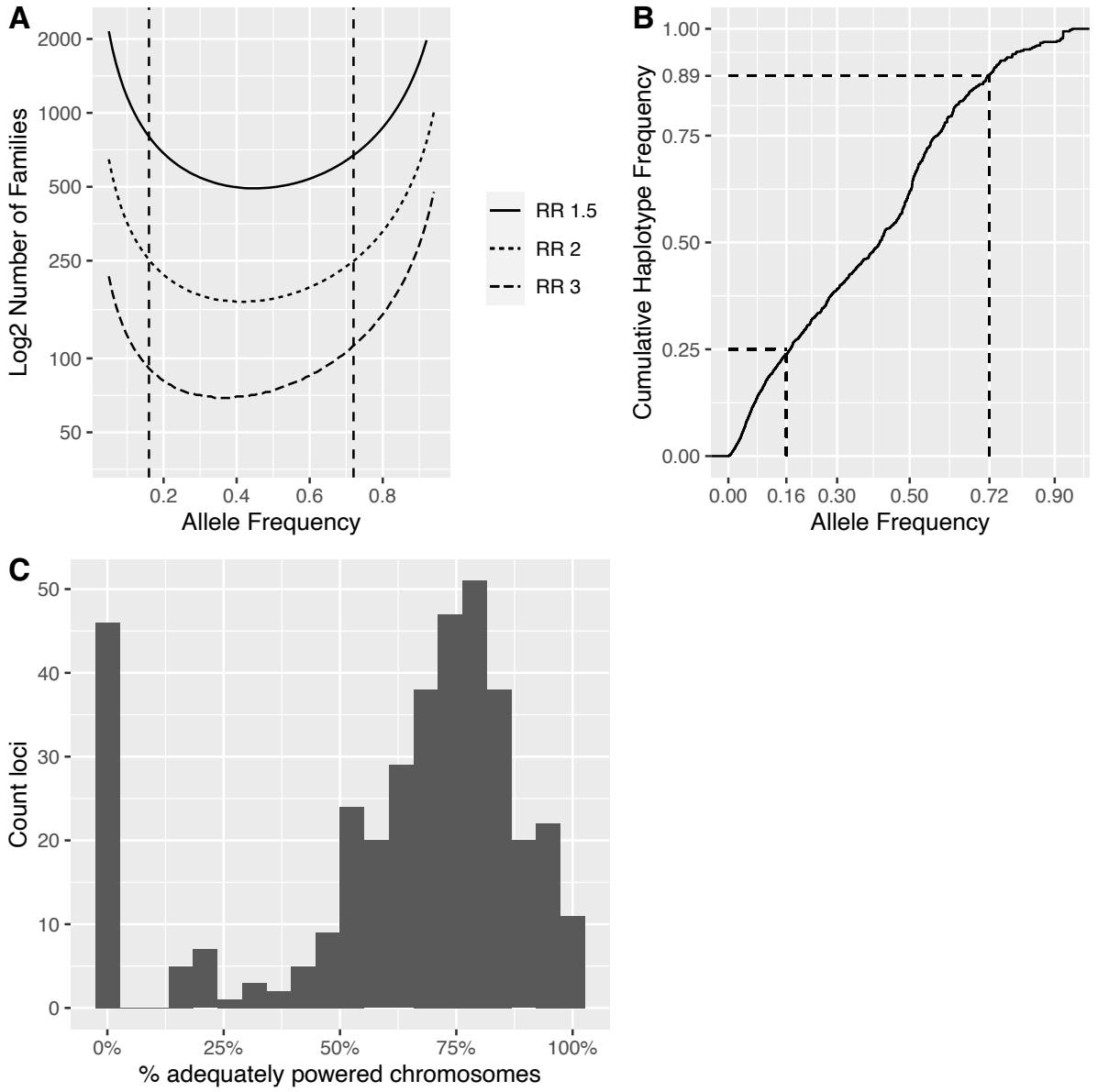

Figure 2. **A** Plot of numbers of families required to have 80% chance of detecting a significant association in the TDT with 3 different levels of genotype relative risk and allele frequencies between 0.05 and 0.95. Note that the curves are not symmetrical about a 50% allele frequency, at low allele frequencies most individuals carrying the allele will be heterozygous and therefore informative but as allele frequencies increase the frequency of homozygotes which are not informative also increase and power declines. Optimum allele frequency is around 40%. The dashed lines delimit the region in which 250 hundred families would provide at least 80% power to detect a haplotype relative risk of 2. **B**. Cumulative frequency of haplotypes by allele frequency, the dashed lines show the region of allele frequencies between 16 and 72% where 250 families would be sufficient to detect a haplotype relative risk (HRR) of 2 from plot A. 64% of haplotypes (0.89-0.25) have an allele frequency in this range. **C.** Counts of haplotype blocks (loci) by the percentage of haplotypes with allele frequencies between 16 and 72% and therefore powered to detect a HRR of 2 with 250 families. The median haplotype block had 71% of chromosomes with allele frequencies in 16%-72% range. 46 blocks had no haplotypes with allele frequencies in this range and therefore only large genetic effects (HRR > 2) are likely to be detectable at these loci.

References

Abel, L., & Müller-Myhsok, B. (1998). Maximum-Likelihood Expression of the Transmission/Disequilibrium Test and Power Considerations. *The American Journal of Human Genetics*, *63*(2), 664–667. https://doi.org/10.1086/301975

Kim, S. A., Cho, C.-S., Kim, S.-R., Bull, S. B., & Yoo, Y. J. (2018). A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics (Oxford, England)*, *34*(3), 388–397. https://doi.org/10.1093/bioinformatics/btx609

Weeks, Daniel E. (2012). *Power analyses for the affected sib pair and the TDT design* (1.5) [R].