Application of haplotype pair analysis for the identification of hemizygous loci

B C Hendrickson, D Pruss, E Lyon, T Scholl

An expectation maximisation based prediction algorithm was created to identify unusual haplotypes in patient samples that may be caused by small intragenic deletions. In this approach, unphased SNP genotypes are compared to pairs of canonical haplotypes to identify potentially hemizygous regions. This method was successfully applied to identify five deletions in the 3′ region of BRCA1.

Most of the human SNP diversity at a given locus may be described as a set of “canonical” haplotypes, representing common haplotypes in a population. Of the few unusual haplotypes not in the canonical set, some are found in a genotype context that is similar to a genotype expected from a combination of a pair of canonical haplotypes, except that one or several polymorphic positions appear, unexpectedly, homozygous. These unusual haplotypes could represent hemizygous loci created by a change at one SNP in each haplotype. However, these could also be bona fide haplotypes that are too rare to be represented in the canonical set, resulting from recombinations between common haplotypes, or single base changes.

In the present study, we used unphased genotype data for 12 common biallelic BRCA1 SNPs (located from exons 4 to 16) generated during sequence based clinical mutation testing to obtain BRCA1 SNP haplotypes for 5911 anonymised samples, by applying an expectation maximisation (EM) algorithm similar to those described elsewhere. Ten canonical SNP haplotypes are known for BRCA1. Interestingly, two BRCA1 haplotypes account for the bulk of the genotypes, the consensus at 59% frequency, and the most common non-consensus haplotype at 21% (fig 1).

We identified 78 (1.3%) samples with non-canonical haplotypes, which in 62 cases were paired with a canonical haplotype. Among these samples, 14 were identified with rare haplotypes created by a change at the SNP in exon 16 (fig 1), suggestive of a possible deletion of this exon. This group was selected for molecular analysis because of its abundance, representing 18% of the unusual haplotypes. Also, additional sequence information available for one of these samples showed a heterozygous SNP near exon 17, placing a 3′ limit on the size of the putative deletion. This discovery increased the likelihood that the unusual haplotype in this sample, and perhaps others in this group, were the result of intragenic deletions, and not just recombinations.

Within the remaining samples, 42 contained rare haplotypes that appeared to arise from changes in one out of five SNP loci in exon 11, potentially indicative of a partial deletion of the exon. These samples were excluded from deletion testing because, to date, all clinically significant large deletions that have been characterised in BRCA1 are Alu mediated, whole exon deletions. An additional 17 samples were excluded from deletion testing because their haplotypes were defined by changes at two, non-adjacent haplotype defining SNPs, which could not be explained by a single deletion event. Several other samples were identified with haplotypes that may have been the result of intragenic deletions. However, the DNA from these samples was not available.

Long range PCR of the region between exons 14 and 18 showed that seven of the selected 14 samples produced smaller than the expected products. Restriction digests of the mutant fragments localised breakpoint regions, which were then characterised by sequencing (fig 2) to show five distinct novel deletions, ranging in size from 5629 to 7183 bases. All five deletions appear to be Alu mediated. Additional work is continuing to determine if the remaining seven samples may contain deletions with breakpoints outside the amplified region.

Our example shows that haplotype analysis is effective for identifying samples that may contain hemizygous regions. The use of haplotype analysis permitted rapid processing of a large sample set and selected the best candidate samples for biochemical analysis. Of the selected specimens in our target group, 50% were shown to contain deletions. Also, this example illustrates the potential increased information available from phased data that can be overlooked in unphased genotype data. Finally, this method may have broader application in large SNP based assays by identifying potentially hemizygous data that might confound downstream analyses.
Figure 2 Results of molecular analysis. Deletions were discovered in seven of the 14 samples. Three samples that claimed Latin American ancestry had the same deletion of exons 16 and 17 (deletion 1). Also, three additional deletions involving exons 16 and 17 (deletions 2-4), and one deletion of exons 15 and 16 (deletion 5, also reported elsewhere during the preparation of this manuscript) were identified. All numerical designations correspond to GenBank L78833. Top strands are genomic sequence at the upstream breakpoint region and lower strands are sequence from the downstream breakpoint region. Vertical lines indicate identity between patient data (middle sequence) and expected sequence. Grey boxes highlight the recombination regions.