HDR: A statistical two-step approach successfully identifies disease genes in autosomal recessive families

Atsuko Imai, Masakazu Kohda, Akihiro Nakaya, Yasushi Sakata, Kei Murayama, Akira Ohtake, Mark Lathrop, Yasushi Okazaki, Jurg Ott

Supplementary Information

Random HDR levels

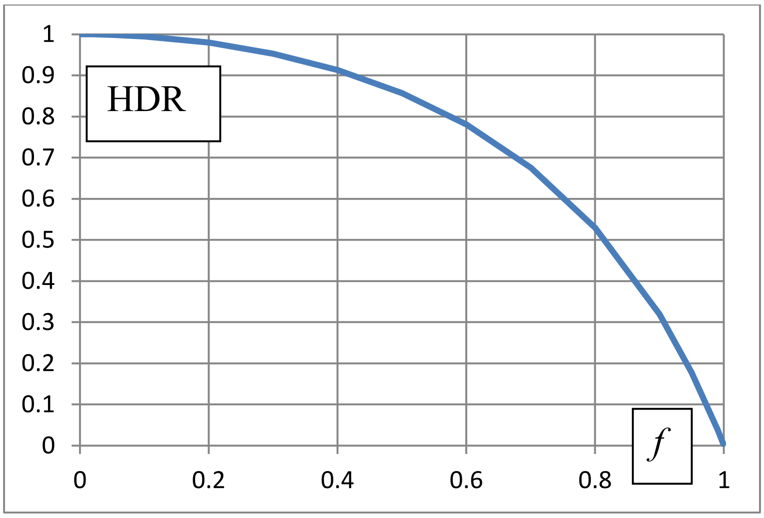
It is of interest to know what levels of HDR we can expect in the absence of disease variants. Consider a variant with two alleles, *A* and *R*, with alternate (non-wild) allele frequency, P(*A*) = *f*. Assuming Hardy-Weinberg equilibrium, homozygotes *A*/*A* are expected with frequency f^2. For two unrelated individuals, we expect pairs of genotypes at this variant as given in the following table:

Individual 1	Individual 2	
	homozygous	not homozygous
homozygous	f^4	$f^2(1-f^2)$
not homozygous	$f^2(1-f^2)$	N/A

The expected value for our Hamming distance ratio is then equal to

$$\text{HDR} = 2 \times f^2(1-f^2)/[2 \times f^2(1-f^2) + f^4] = 1 - f^2/(2-f^2).$$

The graph below of HDR as a function of *f* shows that for a wide range of allele frequencies, $0 < f < 0.8$, the expected value of HDR exceeds 0.50.



We verified these predictions in our data and found that the majority of control-control HDR values exceeded 0.50.

Effects of sequencing errors

As suggested by one of the reviewers, we looked at the effects of errors on HDR. We adopted the following simple error model [1], where e is a small genotype error:

Individual 1	*Individual 2*	
	homozygous	not homozygous
homozygous	$p_1 - e$	$p_2 + e/2$
not homozygous	$p_3 + e/2$	N/A

The Hamming distance ratio then becomes

$$\mathrm{HDR}_e = (p_2 + p_3 + e)/(p_1 + p_2 + p_3) = \mathrm{HDR}_0 + e/(p_1 + p_2 + p_3).$$

Thus, HDR increases or decreases depending on the sign of the error e (note that this is not always the case – in linkage analysis, for example, some misclassification errors always lead to an upward bias of the recombination fraction estimate [2]). Presumably, random errors will be positive for some variants and negative at others, so the net effect on HDR is likely to be small, as anticipated by the reviewer.

Supplementary references

1. Gordon, D., Finch, S.J., Nothnagel, M. & Ott, J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum. Hered.* **54**, 22-33 (2002).
2. Ott, J. Linkage analysis with misclassification at one locus. *Clin. Genet.* **12**, 119-124. (1977).