

doses of IR, perhaps as a result of haploinsufficiency of *Atm*. Longevity after IR was decreased in *Atm*^{+/-} mice, yet these animals died from similar causes (tumours and/or infections) as *Atm*^{+/+} mice. Premature greying was observed over the entire body of *Atm*^{+/-} mice, whereas *Atm*^{+/+} mice displayed minimal, localized greying. Cell lines from *Atm*^{+/-} mice² as well as *Atm*^{+/-} thymocytes⁵ showed intermediate levels of checkpoint function after IR compared with cell lines from *Atm*^{+/+} or *Atm*^{-/-} mice. Other cancer susceptibility genes in which haploinsufficiency has been proposed to have a role are those encoding p53, TGFβ and p27, which show effects that may be the result of cell-cycle functions of these molecules⁶⁻⁸. Our data support the hypothesis that heterozygosity of some cancer susceptibility genes may result in phenotypes caused by dosage reduction, perhaps due to decreases in cell-cycle checkpoint function.

We do not know if these findings are relevant to human AT carriers. There are conflicting data on the role of *ATM* heterozygosity in cancer risk⁹⁻¹³. We observed no differences in tumour spectra between irradiated *Atm*^{+/-} and *Atm*^{+/-}

mice. In addition, there does not appear to be an increase in *ATM* mutations in patients with extreme reactions to radiation therapy¹⁴. Consistent with this is the fact that *Atm*^{+/-} mice showed no evidence of increased acute radiation toxicity, but did display premature greying and decreased survival at higher sublethal doses of 4 Gy (400 Rad). By way of comparison, a standard diagnostic chest x-ray is 0.004 Rad (1 Rad = 1/100 Gy), and the maximum permitted annual dose to radiation workers is 5 Rad. If these data are extrapolated to humans, then human carriers of *ATM* mutations may display increased sensitivity to relatively high sublethal doses of IR.

Acknowledgements

We thank D. Larson and T. Hernandez for technical assistance, S. Hoogstraten-Miller for veterinary care and S. Abshire and N. Grey for mouse caretaking.

Carolee Barlow^{1,4}, Michael A. Eckhaus², Alejandro A. Schäffer³ & Anthony Wynshaw-Boris^{1,5}

¹Genetic Disease Research Branch, National Human Genome Research Institute, ²Veterinary

Resources Program, Office of Research Services, Office of the Director, National Institutes of Health, Bethesda, Maryland 20892, USA.

³Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland 21224, USA. ⁴Present address: The Salk Institute for Biological Sciences, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. ⁵Present address: UCSD School of Medicine, 9500 Gilman Drive, La Jolla, California 92093, USA. Correspondence should be addressed to A.W.-B. (e-mail: tonywb@nhgri.nih.gov).

1. Shiloh, Y. *Annu. Rev. Genet.* **31**, 635-662 (1997).
2. Barlow, C. *et al. Cell* **86**, 159-171 (1996).
3. Elson, A. *et al. Proc. Natl. Acad. Sci. USA* **93**, 13084-13089 (1996).
4. Xu, Y. *et al. Genes Dev.* **10**, 2411-2422 (1996).
5. Xu, Y. & Baltimore, D. *Genes Dev.* **10**, 2401-2410 (1996).
6. Bouffler, S.D., Kemp, C.J., Balmain, A. & Cox, R. *Cancer Res.* **55**, 3883-3889 (1995).
7. Tang, B. *et al. Nature Med.* **4**, 802-807 (1998).
8. Fero, M.L., Randel, E., Gurley, K.E., Roberts, J.M. & Kemp, C.J. *Nature* **396**, 177-180 (1998).
9. FitzGerald, M.G. *et al. Nature Genet.* **15**, 307-310 (1997).
10. Vorechovsky, I. *et al. Cancer Res.* **56**, 4130-4133 (1996).
11. Chen, J., Birkholtz, G.G., Lindblom, P., Rubio, C. & Lindblom, A. *Cancer Res.* **58**, 1376-1379 (1998).
12. Bishop, D.T. & Hopper, J. *Nature Genet.* **15**, 226 (1997).
13. Athma, P., Rappaport, R. & Swift, M. *Cancer Genet. Cytogenet.* **92**, 130-134 (1996).
14. Ramsay, J., Birell, G. & Lavin, M. *Lancet* **347**, 1627 (1996).
15. Becker, R.A., Chambers, J.M. & Wilks, A.R. *The New S Language* (Wadsworth and Brooks/Cole, Pacific Grove, California, 1988).

© 1999 Nature America Inc. • http://genetics.nature.com

Loss of information due to ambiguous haplotyping of SNPs

A string of tightly linked diallelic loci (SNPs), with number of loci equal to k, might appear equivalent to a single locus with 2^k alleles, because 2^k distinct haplotypes can be formed. These haplotypes, however, cannot always be identified in an individual, even with parental genotyping (Fig. 1); that is, one cannot always determine which SNP alleles occur together on a chromosome. Thus, the information available from k SNPs may be less than that from one 2^k-allele locus. This potential loss of information, which we demonstrate in the absence of linkage disequilibrium (LD), can be expected to increase sample size requirements and reduce power of SNPs for association and linkage studies of genetic diseases^{1,2}.

To quantify the extent of haplotype (phase) ambiguity, we assume k diallelic loci, Hardy-Weinberg equilibrium and linkage equilibrium. We denote the alleles at the ith locus as A_i and B_i, with frequencies p_i and q_i=1-p_i, respectively. The total number of possible genotypes (ignoring

linkage phase) is 3^k, because there are three possibilities (A_iA_i, A_iB_i, B_iB_i) at each locus.

An 'ambiguous individual' is one whose haplotypes cannot be inferred with certainty. For example, an A₁B₁, B₂B₂, A₃B₃ individual may be haplotyped as

A₁B₂A₃/B₁B₂B₃ or A₁B₂B₃/B₁B₂A₃, and is thus ambiguous. An individual is ambiguous if, and only if, s/he is heterozygous at two or more loci. The number of ambiguous k-locus genotypes is found by subtracting from 3^k the number of genotypes homozygous at all k loci (2^k) or at exactly k-1 loci (k2^{k-1}), yielding 3^k-2^k-k2^{k-1} (for example, for k=3, there are 7 ambiguous genotypes: A₁B₁A₂B₂A₃B₃; A₁B₁A₂B₂B₃B₃; A₁B₁A₂A₃B₂B₃; A₁B₁B₂A₂A₃B₃; A₁B₁B₂B₂A₃B₃; B₁B₁A₂B₂A₃B₃; A₁B₁A₂B₂A₃B₃).

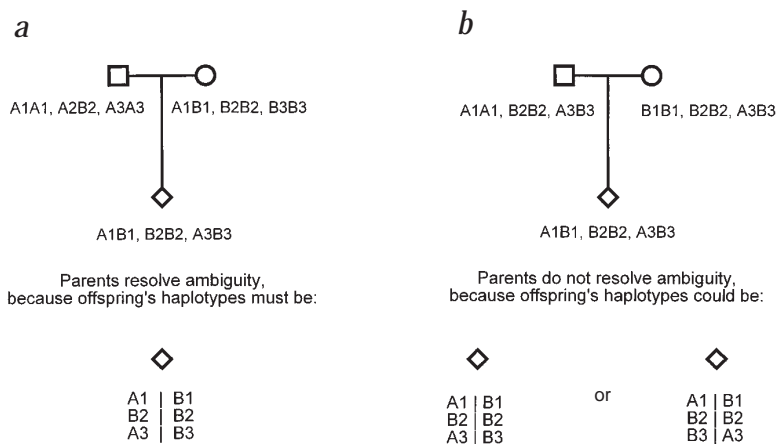


Fig. 1 Examples of an unambiguous (a) and an ambiguous (b) triad.

Table 1 • Counts and population frequencies of ambiguous individuals and triads, for k SNPs

k	Individuals						Triads					
	Counts			Population frequencies ^a			Counts			Population frequencies ^b		
	Ambiguous	Total	Ratio	p _i =0.5 ^c	p _i =0.25	p _i =0.1	Ambiguous	Total	Ratio	p _i =0.5	p _i =0.25	p _i =0.1
2	1	9	0.11	0.25	0.14	0.03	7	100	0.07	0.11	0.05	0.01
3	7	27	0.26	0.50	0.32	0.09	163	10 ³	0.16	0.24	0.11	0.02
4	33	81	0.41	0.69	0.48	0.15	2,575	10 ⁴	0.26	0.35	0.18	0.03
6	473	729	0.65	0.89	0.73	0.30	4.22×10 ⁵	10 ⁶	0.42	0.53	0.31	0.06
10	5.29×10 ⁴	5.90×10 ⁴	0.90	0.99	0.94	0.56	6.41×10 ⁹	10 ¹⁰	0.64	0.73	0.51	0.12
20	3.48×10 ⁹	3.49×10 ⁹	1.00	1.00	1.00	0.90	8.78×10 ¹⁹	10 ²⁰	0.88	0.93	0.77	0.27

^aPopulation frequencies of ambiguous individuals. ^bPopulation frequencies of triads in which phase of offspring cannot be determined unambiguously from parents. ^cp_i gives frequency of allele A at each locus.

The population frequency of ambiguous individuals is found by subtracting from unity the probability of being homozygous at all k loci or at exactly k-1 loci:

$$1 - \prod_{i=1}^k (1-2p_iq_i) - \sum_{\ell=1}^{k-1} 2p_{\ell}q_{\ell} \prod_{i \neq \ell} (1-2p_iq_i)$$

When all p_i equal a common value p, this becomes 1-(1-2pq)^k-k(2pq)(1-2pq)^{k-1}. The number of ambiguous genotypes and their probabilities increase rapidly with k (Table 1, left).

One approach to resolving the phase ambiguity is to genotype parents as well. We define the resulting configuration of offspring plus parents as an 'ambiguous triad' if knowing the offspring and parental genotypes does not enable identification of the haplotypes of the offspring (Fig. 1).

A triad is ambiguous if, and only if, parents and offspring are all heterozygous at one or more locus, and the offspring is heterozygous at two or more loci. Possible configurations are as follows: we define 'configuration I' at the ith locus as one where offspring and both parents are heterozygous (this configuration has probability 2p_i²q_i²); configuration II has offspring heterozygous but one or more parent homozygous (probability 2p_iq_i-2p_i²q_i²); and configuration III has offspring homozygous (probability 1-2p_iq_i). An unambiguous triad has either no locus in configuration I, or exactly one locus in I, but the remaining k-1 in III. Subtracting these probabilities from unity yields the population frequency of triads with ambiguous offspring haplotypes:

$$1 - \prod_{i=1}^k (1-2p_i^2q_i^2) - \sum_{\ell=1}^{k-1} (2p_{\ell}^2q_{\ell}^2) \prod_{i \neq \ell} (1-2p_iq_i)$$

or, when all p_i equal a common value: 1-(1-2p²q²)^k-k(2p²q²)(1-2pq)^{k-1}. Similar reasoning reveals that the number of total possible triads (parental order irrelevant) is 10^k, of which 10^k-9^k-k6^{k-1} have ambiguous offspring haplotypes.

Counts and population frequencies of ambiguous individuals and triads for selected values of k and selected allele frequencies are given (Table 1). We note: (i) the extent of ambiguity increases with increasing numbers of loci, approaching unity as k increases; (ii) ambiguity is more severe when loci are more polymorphic; and (iii) triads are less ambiguous than individuals. Knowledge of parental genotypes does not, however, resolve all ambiguity in the offspring (for example, more than one-third of triads can be ambiguous with as few as four loci). Typing additional relatives (for example, grandparents or an additional sibling) would further reduce ambiguity, but the patterns noted will continue to hold for any family configuration.

Thus, a k-locus diallelic system is not equivalent to one locus with 2^k alleles in terms of information. This information loss is analogous to that in dominant or recessive systems in which multiple genotypes also result in the same phenotype. Statistical methods such as the E-M algorithm may be used to partially resolve the ambiguity^{3,4}. Alternatively, for very tightly linked SNPs, the phase ambiguity may be resolvable experimentally by long-range allele-specific PCR (refs 5,6), although at greater cost; this approach may be more effective than collecting triads as a way to resolve phase ambiguity.

The above results assume linkage equilibrium. Currently, the extent and magni-

tude of LD in the human genome are not well characterized^{7,8}, so its full impact cannot yet be assessed. LD, when it exists, will mitigate ambiguity partially by making some haplotypes *a priori* more likely than others. On the other hand, in the limit of complete disequilibrium, k SNPs approach a single two-allele locus, so some information will be lost. In conclusion, k SNPs can and generally will provide substantial information, but this information will be less than for a single 2^k-allele locus with corresponding allele frequencies, with or without LD.

Acknowledgements

We thank H. Göring, J. Hoh, F. Collins and K. Weiss for helpful comments. This work was supported in part by grants MH-48858, DK-31813, DK-31775, HG-00376 and HL-35018.

Susan E. Hodge¹, Michael Boehnke² & M. Anne Spence³

¹Department of Psychiatry, Columbia University, NY State Psychiatric Institute, Unit 24, 1051 Riverside Drive, New York, New York 10032, USA. ²Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan 48109, USA. ³Department of Pediatrics, UC Irvine Medical Center, Orange, California 92868, USA. Correspondence should be addressed to S.E.H. (e-mail: seh2@columbia.edu).

- Risch, N. & Merikangas, K. *Science* **273**, 1516-1517 (1996).
- Collins, F.S., Guyer, M.S. & Chakravarti, A. *Science* **278**, 1580-1581 (1997).
- Excoffier, L. & Slatkin, M. *Mol. Biol. Evol.* **12**, 921-927 (1995).
- Long, J.C., Williams, R.C. & Urbanek, M. *Am. J. Hum. Genet.* **56**, 799-810 (1995).
- Nickerson, D.A. et al. *Nature Genet.* **19**, 233-241 (1998).
- Clark, A.G. et al. *Am. J. Hum. Genet.* **63**, 595-612 (1998).
- Pennisi, E. *Science* **281**, 1787-1789 (1998).
- Terwilliger, J.D. & Weiss, K.M. *Curr. Opin. Biotechnol.* **9**, 578-594 (1998).