Letters to the Editor

Am. J. Hum. Genet. 71:439–441, 2002

A Note on the Calculation of Empirical *P* Values from Monte Carlo Procedures

To the Editor:

It has become commonplace in the statistical analysis of genetic data to use Monte Carlo procedures to calculate empirical *P* values. The reasons for this include the following: (1) many test statistics do not have a standard asymptotic distribution; (2) even if a standard asymptotic distribution does exist, it may not be reliable in realistic sample sizes; and (3) calculation of the exact sampling distribution through exhaustive enumeration of all possible samples may be too computationally intensive to be feasible. In contrast, Monte Carlo methods can be used to obtain an empirical *P* value that approximates the exact *P* value without relying on asymptotic distributional theory or exhaustive enumeration. Examples of procedures for genetic analysis that use simulation methods to determine statistical significance are CLUMP (Sham and Curtis 1995), MCETDT (Zhao et al. 1999), and a new test of linkage for a second locus conditional on information from an already-known locus (Cordell et al. 2000).

In this letter, we would first like to draw attention to the fact that some currently available genetic-analysis programs (including some of our own) use a method of calculating empirical *P* values that is not strictly correct. Typically, the estimate of the *P* value is obtained as $\hat{p} = r/n$, where n is the number of replicate samples that have been simulated and r is the number of these replicates that produce a test statistic greater than or equal to that calculated for the actual data. However, Davison and Hinkley (1997) give the correct formula for obtaining an empirical *P* value as $(r+1)/(n+1)$. The reasoning is roughly as follows: if the null hypothesis is true, then the test statistics of the n replicates and the test statistic of the actual data are all realizations of the same random variable. These realizations can be ranked, and then the probability, under the null hypothesis, that the test statistic from the actual data has the observed rank or a higher rank is $(r+1)/(n+1)$, the proportion of all possible rankings of the realizations that fulfill this criterion.

It is perhaps worth explicitly making the point that this procedure utilizes the ranks, rather than the actual values, of the test statistics. Another approach to the Monte Carlo estimation of significance would be to use the simulated test statistics to estimate the shape of the probability distribution and then to calculate a *P* value from this, but the use of ranks renders the process distribution free and is used almost universally.

Given that the most accurate estimate of the *P* value is actually $(r+1)/(n+1)$, any procedure that uses r/n will tend to underestimate the *P* value if the null hypothesis is true—although, in most circumstances, to only a small degree. For example, if $r = 5$ and $n = 500$, then the correct estimate of the *P* value is $6/501 = 0.012$, rather than .01. The effect is greatest when r is small: for $r = 1$ and $n = 500$, the correct *P* value is $2/501 = .004$, rather than .002, and, for $r = 0$ and $n = 500$, the correct *P* value is $1/501 = .002$, rather than 0. It is straightforward to demonstrate this effect in practice. We wrote a small computer program to generate a random number, x, to represent a test statistic observed under the null hypothesis. It then generates n more random numbers, to obtain an empirical estimate of the *P* value associated with x, where r is the number of replicates obtained that are $\geqslant x$. We repeated this procedure 10^6 times, using a value of 500 for n and counting the number of times that we obtained an empirical *P* value $\leqslant .01$. When we used r/n to estimate the *P* value, we obtained a *P* value of .01 on 12,103 of 10^6 occasions, whereas, when we used $(r+1)/(n+1)$, this *P* value was obtained on 10,106 of 10^6 occasions. This confirms that use of r/n to estimate *P* values is anticonservative.

Using $(r+1)/(n+1)$ also avoids the problem of obtaining a *P* value of 0 when the observed test statistic is greater than those in any of the replicates. For n replicates, the minimum possible estimate of the *P* value becomes $1/(n+1)$. Thus, to obtain a very small *P* value, it will be necessary to simulate a large number of replicates. Another way of viewing this issue is as follows. Although use of $(r+1)/(n+1)$ produces an unbiased estimate of the true *P* value (in contrast to use of r/n), this procedure will consistently overestimate small *P* values but will underestimate large *P* values. In fact, the expectation of $(r+1)/(n+1)$ is $(np+1)/(n+1)$, so that the bias is $(1-P)/(n+1)$. Once again, when n is large, this overestimation is unlikely to be important.

It is helpful to provide some quantification of the effects that we describe. Typically, the true *P* value will be unknown, and judgments will need to be made on the basis of the observed values of *r* and *n*.

First, any methodology that utilizes *r*/*n* to estimate the *P* value will tend to underestimate the actual *P* value by a factor of ~$r/(r+1)$. Often, it will be possible to recalculate the true estimate of the *P* value, but, in some situations, the estimated *P* value may not be stated explicitly (e.g., when multiple tests are applied and only corrected *P* values are provided). In any event, if $r \geqslant 4$, then the bias in the estimate of the *P* value will not be likely to lead to any serious error in interpretation. If $r < 4$, then perhaps results should be treated with some suspicion and a larger number of simulations should be performed.

Second, if *r* is small, then small *P* values will tend to be overestimated, and potentially important results could be missed. Obviously, if one uses $n = 19$, observes $r = 0$, and estimates a *P* value of $(r+1)/(n+1) = .05$, then the true *P* value might be as low as 10^{-6} or 10^{-12}. One would hope that any researcher obtaining $r = 0$ would want to repeat the procedure using larger *n*. The question obviously arises of what value of *r* is "enough"—that is, what value should one observe to be reasonably confident that one is not wildly overestimating the *P* value? For given values of the true *P* value and of *n,* we can use a binomial distribution to calculate the probability that a value of *r* will be obtained such that $(r+1)/(n+1)$ will overestimate *P* by a factor of ⩾2. The following examples are chosen such that, for a true *P* value of .01, the stated values of *r* and *n* will yield an estimate of ⩾.02: with $n = 149$, $P_{r\geqslant 2} = 0.44$; with $n = 249$, $P_{r\geqslant 4} = .24$; with $n = 449$, $P_{r\geqslant 8} = .085$; and with $n = 549$, $P_{r\geqslant 10} = .052$. As it turns out, the probabilities associated with these values of *r* remain very similar, albeit not identical, if different *P* values are used, along with appropriate values for *n* chosen to yield an overestimate by a factor of 2. For example, the corresponding probabilities of *r* exceeding the threshold values of 2, 4, 8, and 10, if the true *P* value is .00001, are .44, .24, .087, and .054, respectively. It should perhaps be emphasized that this tendency to overestimate small *P* values is not purely a consequence of using $(r+1)/(n+1)$ rather than *r*/*n* as an estimate: use of *r*/*n* would give corresponding probabilities (with a true *P* value of .00001) of .26, .14, .051, and .031. From these observations, we can construct the general rule that, if one observes $r = 2$, then there is a strong possibility that one may be overestimating the *P* value by a factor of ⩾2, whereas, if one observes $r \geqslant 10$, then such a large overestimate is fairly unlikely.

Finally, although we have said that use of *r*/*n* rather than $(r+1)/(n+1)$ is anticonservative to only a small degree—which would be unlikely to have an important effect on interpretation (at least provided $r \geqslant 4$)—there is one situation in which even a small bias could be important: when the power of different methods is being compared. We have noted that the true *P* value associated with $r = 5$ and $n = 500$ is .012, rather than .01. This means that a Monte Carlo method that used *r*/*n* to estimate the *P* value might find 20% more observations significant at a level of .01 compared with an accurate method. One might be concerned that, if one performed a power study comparing two such methods, the Monte Carlo method might be found to be considerably more powerful than the other method, such a finding being an artifact of the anticonservative nature of the Monte Carlo method. In fact, we have carried out extensive simulations and have found this not to be the case. We simulated affected sib-pair samples with allele-sharing probabilities increased above the null hypothesis value of 0.5 and measured the power of a Monte Carlo method using *r*/*n* compared to the power of an exact binomial method to detect this deviation. Once again, we found that, at least for values of $r \geqslant 4$, the power of the two methods was very similar and that the theoretically anticonservative nature of the Monte Carlo test did not, after all, have important practical implications. The reason for this seems to be that the Monte Carlo test does not measure significance, but only estimates it, and that the effect of the anticonservative bias is almost exactly counterbalanced by the tendency to overestimate small *P* values.

We therefore draw the following conclusions. First, taking *r*/*n* rather than $(r+1)/(n+1)$ as an estimate of the *P* value is essentially incorrect and should not be used. However, in practice, doing so is unlikely to have any serious implications either in individual tests or in power comparisons between methods, at least when $r \geqslant 4$. Second, Monte Carlo methods provide an estimate, rather than a measure, of the *P* value. This implies that they tend to overestimate *P* values that are, in reality, small, and, hence, they may have less power than other methods. This effect decreases as *r* increases and becomes fairly unimportant when $r \geqslant 10$. We therefore recommend that, for all applications, enough replicates are obtained to ensure that $r \geqslant 10$.

B. V. North,[1] D. Curtis,[1] and P. C. Sham[2]
[1]*Joint Academic Department of Psychological Medicine, St Bartholomew's and Royal London School of Medicine and Dentistry, and* [2]*Department of Psychological Medicine, Institute of Psychiatry, London*

References

Cordell HJ, Wedig GC, Jacobs KB, Elston RC (2000) Multilocus linkage tests based on affected relative pairs. Am J Hum Genet 66:1273–1286

Davison AC, Hinkley DV (1997) Bootstrap methods and their

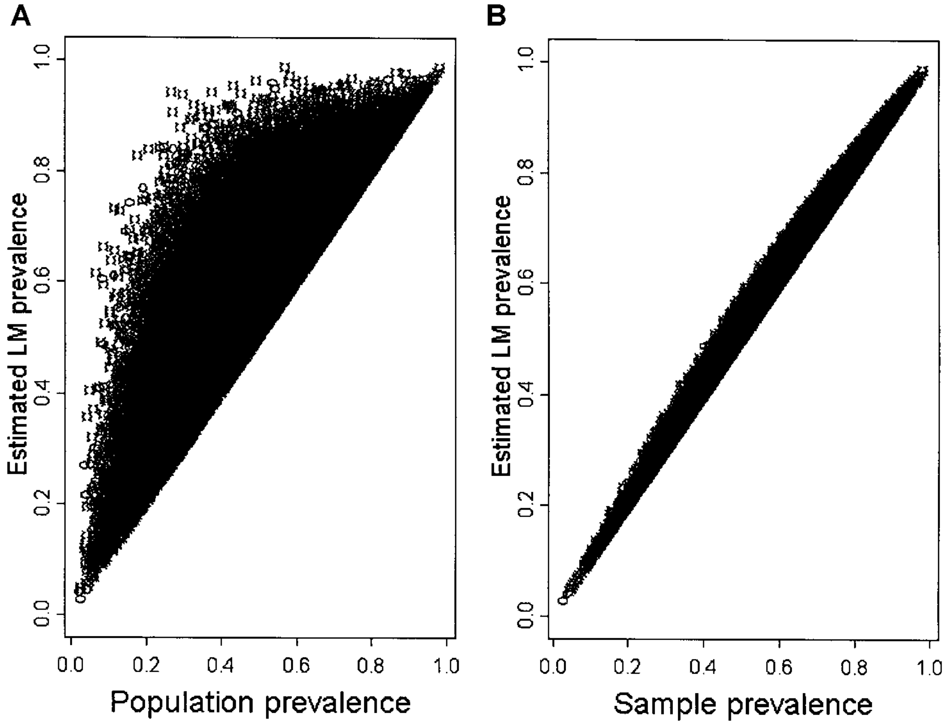

Figure 1 Scatter plots of the estimated prevalence of a disease, based on the Li-Mantel (LM) estimator. *A*, $\hat{p}_{LM}$ vs. p_P. *B*, $\hat{p}_{LM}$ vs. p_A. Data are based on 100,000 simulations of a scenario described in the text.

application. Cambridge University Press, Cambridge, United Kingdom

Sham PC, Curtis D (1995) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann Hum Genet 59:97–105

Zhao JH, Sham PC, Curtis D (1999) A program for the Monte Carlo evaluation of significance of the extended transmission/disequilibrium test. Am J Hum Genet 64:1484–1485

Address for correspondence and reprints: Dr. D. Curtis, Joint Academic Department of Psychological Medicine, St Bartholomew's and Royal London School of Medicine and Dentistry, 3rd Floor, Alexandra Wing, Turner Street, London E1 1BB, United Kingdom. E-mail: dcurtis@hgmp.mrc.ac.uk

0002-9297/2002/7102-0025$15.00

Am. J. Hum. Genet. 71:441–442, 2002

Response to Epstein et al.

To the Editor:

We entirely agree with the statement in a recent article by Epstein et al. (2002) that the likelihood used in our example 1 (Burton et al. 2000) fails because it inappropriately assumes marginal independence: marginal dependence is introduced because the unobserved determinants of stratum-specific risk are shared by siblings. However, that is the whole point. It is an analysis of this type that is carried out whenever (as is usual) such heterogeneity is ignored. The reality is that, despite advances in both biology and biostatistics, we are a long way from being able to claim that the modeling of unobserved heterogeneity is "solved," and, until we can, the relevant interpretational problems (Burton et al. 2000) remain real.

There is one important area where our interpretation does differ from that of Epstein et al. This is in our contention that when heterogeneity is ignored, the resultant ascertainment-adjusted estimates reflect parameters in the ascertained sample rather than those in the original population. Epstein et al. state that the estimates "generally do not reflect the true values in either the original population or the ascertained subpopulation" (2002, p. 886). We do not agree. Relationships B1, B2, and B3 in Appendix B of the article by Epstein et al. (2002) all represent weighted means for the prevalence in stratum k (p_k). Because the weights under B2 (which generate the disease prevalence in the ascertained subpopulation [p_A]) are different from those under B3 (which generate the Li-Mantel estimate [$\hat{p}_{LM}$]), we agree that the latter does not provide a consistent estimate of the former (see also Olson and

Table 1

Frequency of SNPs and Standardized Linkage-Disequilibrium Coefficients

SNP	FREQUENCY ($n = 192$ Chromosomes)	$D' = D/D_{\max}$[a]					
		235Thr	11535A	11608T	12058A	12194C	12429T
Met235Thr	.42	…	…	…	…	…	…
C11535A	.31	−.803	…	…	…	…	…
C11608T	.33	−.667	1.000	…	…	…	…
G12058A	.06	.568	−.733	−.750	…	…	…
A12194C	.06	.734	−.754	−.769	.911	…	…
C12429T	.07	.629	−.771	−.786	1.000	1.000	…
T12822C	.39	.816	−.829	−.720	.863	.874	.766

[a] D' is the standardized coefficient of linkage disequilibrium; D is the classical coefficient of linkage disequilibrium; and $D_{\max}$ is the maximum D value that is possible given the allele frequencies.

Cordell 2000). However, the word "reflect" does not imply a "consistent estimator," and we did not use the latter term; in fact, we used phrases such as "good approximations." It is easy to see that the ratio of the weight under B3 for any given stratum to that under B2 for the same stratum must lie between 1:1 and 2:3, the latter ratio being attained only as p_k tends to 0. This means that the estimates under the two weighting systems are unlikely to be seriously discrepant.

To illustrate, we generate 100,000 simulated data sets, each equivalent to the general case considered in Appendix B of the article by Epstein et al. (2002), which itself corresponds to example 1 given by Burton et al. (2000). For each of four strata (k = 1,…,4), p_k is the stratum-specific prevalence of disease and π_k is the proportion of the original population in that stratum. In each simulation, each p_k and each π_k are randomly sampled to take any real value between 0 and 1, with uniform probability. Each π_k is then normalized (divided by $\sum_{k=1}^{4} \pi_k$) so that, after normalization, $\sum_{k=1}^{4} \pi_k = 1$ in every simulated data set. We then used the expressions B1, B2, and B3, given by Epstein et al. (2000), to obtain the prevalence in the original population (p_P), p_A, and $\hat{p}_{LM}$, respectively. Figure 1*A* illustrates the resultant relationship between $\hat{p}_{LM}$ and p_P, and figure 1*B* illustrates that between $\hat{p}_{LM}$ and p_A. The latter is a straight line with a gradient of 1.004 and a correlation of 0.996. The maximum discrepancy between p_A and $\hat{p}_{LM}$ across all 100,000 simulations is 0.085 (corresponding to $p_A = 0.401$ and $\hat{p}_{LM} = 0.486$). In 95% of simulations, the difference is <0.042. In contrast, the relationship between p_P and $\hat{p}_{LM}$ is much weaker. The maximum discrepancy across the 100,000 simulations is 0.67 (corresponding to $p_P = 0.27$ and $\hat{p}_{LM} = 0.94$); and, in 38% of simulations, the absolute discrepancy is >0.10. Consequently, we remain faithful to our contention that, unless something formal is done to address an unobserved heterogeneity in risk that is shared by family members and therefore introduces marginal dependence, $\hat{p}_{LM}$ reflects the marginal distribution of prevalence in the sample, not the general population. The extent to which this important conclusion may be extrapolated to other scenarios and to analyses based on statistics other than the Li-Mantel estimator warrants further study.

Paul R. Burton,[1,2] Lyle J. Palmer,[2,3,4]
Kevin J. Keen,[4] Jane M. Olson,[4]
and Robert C. Elston[4]

[1]*Genetic Epidemiology Unit, Department of Epidemiology and Public Health, University of Leicester, Leicester, United Kingdom;* [2]*Division of Biostatistics and Genetic Epidemiology, TVW Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, Perth, Australia;* [3]*Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston; and* [4]*Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland*

References

Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, Elston RC (2000) Ascertainment adjustment: where does it take us? (erratum 67:672 [2001]) Am J Hum Genet 67:1505–1514

Epstein MP, Lin X, Boehnke M (2002) Ascertainment-adjusted parameter estimates revisited. Am J Hum Genet 70:886–895

Olson JM, Cordell HJ (2000) Ascertainment bias in the estimate of sibling genetic risk parameters. Genet Epidemiol 18: 217–235

Address for correspondence and reprints: Dr. Paul Burton, Genetic Epidemiology Unit, Department of Epidemiology and Public Health, 22–28 Princess Road West, Leicester, LE1 6TP, United Kingdom. E-mail: pb51@leicester.ac.uk

0002-9297/2002/7102-0026$15.00

Am. J. Hum. Genet. 71:443–444, 2002

SNPs at the 3′ End of the Angiotensinogen Gene Define Two Haplotypes Associated with the Common 235Met Variant

To the Editor:

Nakajima et al. (2002) have provided a valuable SNP-based haplotype map of the angiotensinogen gene, *AGT* (MIM 106150), in both Japanese and white American populations. Several linkage and association studies have supported the hypothesis that angiotensinogen plays a role in the pathogenesis of essential hypertension (MIM 145500) and preeclampsia (MIM 189800) (Jeunemaitre et al. 1992; Ward et al. 1993; Hata et al. 1994). An exon 2 SNP that results in a Thr-Met polymorphism at codon 235 has been associated with variation in plasma-angiotensinogen concentrations and with hypertensive disorders. It is not clear whether this is due to either a deleterious effect of a 235Thr-bearing allele or a protective effect of the 235Met allele. In nonpregnant subjects, 235Met is associated with lower concentrations of plasma angiotensinogen, although, in normotensive pregnant subjects, this pattern was reversed in the population that we studied (Jeunemaitre et al. 1992; Morgan et al. 2000). We have genotyped polymorphisms in both the 5′ flanking region (corresponding to G−217A, A−20C, and A−6G) and exon 2 (Thr174Met [C3889T] and Thr235Met [C4072T]) of *AGT,* in 96 healthy white Europeans from Nottingham, United Kingdom, who were recruited sequentially from a blood-donor clinic (Morgan et al. 1996). This investigation demonstrated patterns—both of linkage disequilibrium and of haplotype frequencies—similar to those described for the Utah population that Nakajima et al. (2002) studied. We have confirmed their observation that a single haplotype carrying 235Met (4072T) accounts for more than half of the genes in white Europeans (58% in the Nottingham population). Variants in both the 5′ flanking region and exon 2 were found only on alleles bearing 235Thr.

We have also screened the 3′ end of the gene—including the 3′ UTR of exon 5—and the 1,350 bases of flanking region between exon 5 and the *AGT* dinucleotide repeat polymorphism (Kotelevstev et al. 1991) (EMBL Nucleotide Sequence Database accession number AJ277498). We used SSCP analysis and direct sequencing, to characterize six SNPs (table 1). Three polymorphisms identified in this region have not, to our knowledge, previously been described; the remaining three correspond to SNPs 40 (C11535A), 41 (C11608T), and 42 (G12058A), as observed by Nakajima et al. (2002).

The frequencies of haplotypes that combine these six SNPs and Thr235Met were estimated by the expectation-maximization method, by use of Arlequin software. Strong linkage disequilibrium was observed between all polymorphisms, and five haplotypes accounted for 91% of those observed in this study (table 2). Interestingly, polymorphisms at C11535A and C11608T, which are in complete linkage disequilibrium with each other, defined two common haplotypes bearing 235Met. To our knowledge, these are the only SNPs that have, to date, been described in white Europeans, which split the common haplotype bearing 235Met; all other variants have been described in association with the 235Thr allele. C11535A lies within the 3′ UTR of exon 5; C11608T lies 30 bases downstream from the 3′ end of exon 5. Whether they have functional effects on angiotensinogen expression requires further investigation, but it is worth noting that in vitro experiments have demonstrated that there is enhancer activity in this region (Nibu et al. 1994*a,* 1994*b*). Given the interest in the Thr235Met polymorphism as a marker for hypertensive disorders, we recommend that genotyping at C11535A or C11608T be included in the haplotyping profile for angiotensinogen in linkage-disequilibrium studies.

Acknowledgment

We thank the Trustees of Nottingham University Hospital for their financial support of this project.

S. Plummer, L. Morgan, and N. Kalsheker
Clinical Chemistry Division
School of Clinical Laboratory Sciences
University Hospital
Nottingham
United Kingdom

Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

Arlequin, http://lgb.unige.ch/arlequin/ (for Arlequin program)

EMBL Nucleotide Sequence Database, http://www.ebi.ac.uk/embl/ (for *AGT* 3′ flanking region [accession number AJ277498])

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for EHT [MIM 145500], *AGT* [MIM 106150], and preeclampsia [MIM 189800])

References

Hata A, Namikawa C, Sasaki M, Sato K, Nakamura T, Tamura K, Lalouel J (1994) Angiotensinogen as a risk factor for essential hypertension in Japan. J Clin Invest 93:1285–1287

Jeunemaitre X, Soubrier F, Kotelevtsev Y, Lifton R, Williams C, Charru A, Hunt S, Hopkins P, Williams R, Lalouel J, Corvol P (1992) Molecular basis of human hypertension: role of angiotensinogen. Cell 71:169–180

Kotelevtsev Y, Clauser E, Corvol P, Soubrier F (1991) Dinucleo-

Table 2

Common Haplotypes Defined by the Met235Thr Polymorphism and Six SNPs at the 3′ End of *AGT*

	Allele at SNP[a]							
Haplotype	Met235Thr	C11535A	C11608A	G12058A	A12194C	C12429T	T12822C	Frequency
1	Met	C	C	G	A	C	T	.27
2	Met	A	A	G	A	C	T	.27
3	Thr	C	C	G	A	C	C	.28
4	Thr	C	C	A	C	T	C	.05
5	Thr	C	C	G	A	C	T	.04

[a] Nucleotides are numbered with respect to the transcription start site.

tide repeat polymorphism in the human angiotensinogen gene. Nucleic Acids Res 19:6978

Morgan L, Broughton-Pipkin F, Kalsheker N (1996) DNA polymorphisms and linkage disequilibrium in the angiotensinogen gene. Hum Genet 98:194–198

Morgan L, Crawshaw S, Baker P, Broughton Pipkin F, Kalsheker N (2000) Polymorphism in oestrogen response element associated with variation in plasma angiotensinogen concentrations in healthy pregnant women. J Hypertens 18: 553–557

Nakajima T, Jorde LB, Ishigami T, Umemura S, Emi M, Lalouel J-M, Inoue I (2002) Nucleotide diversity and haplotype structure of the human angiotensinogen gene in two populations. Am J Hum Genet 70:108–123

Nibu Y, Takahashi S, Tanimoto K, Murakami K, Fukamizu A (1994*a*) Identification of cell type-dependent enhancer core element located in the 3′-downstream region of the human angiotensinogen gene. J Biol Chem 269:28598–28605

Nibu Y, Tanimoto K, Takahashi S, Ono H, Murakami K, Fukamizu A (1994*b*) A cell type-dependent enhancer core element is located in exon 5 of the human angiotensinogen gene. Biochem Biophys Res Commun 205:1102–1108

Ward K, Hata A, Jeunemaitre X, Helin C, Nelson L, Namikawa C, Farrington PF, Ogasawara M, Suzumori K, Tomoda S, Berrebi S, Sasaki M, Corvol P, Lifton RP, Lalouel J-M (1993) A molecular variant of angiotensinogen associated with preeclampsia. Nat Genet 4:59–61

Address for correspondence and reprints: Dr. Linda Morgan, Division of Clinical Chemistry, University Hospital, Queen's Medical Centre, Nottingham, NG7 2UH, United Kingdom. E-mail: linda.morgan@nottingham.ac.uk

0002-9297/2002/7102-0027$15.00