Imprinted genes have few and small introns

Sir — Neumann and colleagues[1] have recently concluded that imprinted genes often contain particular repeat sequences. Others have shown that imprinted genes tend to have unusual sex-specific recombination rates[2]. Here we add to the list of unusual properties of imprinted genes by showing that they tend to have few and typically small introns for their size.

From descriptions of the genomic structure of 16 imprinted genes, we have determined, for each, the total size of exons, the total size of introns and the number of introns/exons. From these data we calculated: the average size of introns, the number of introns per kb of exon and the ratio of the total intron size to the total exon size (Table 1). Mann-Whitney U tests were performed to compare the imprinted genes with a control set of genes. The control set totalled 90 randomly selected genes from man and mouse with the same ratio of human to murine as found in the imprinted set (Table 2).

The intron–exon structure of imprinted genes is very different from the control set (Table 2). Controlling for total exon size, the total amount of intronic DNA in imprinted genes is one fifth that of the control set (Fig. 1). This highly significant difference ($P < 0.0001$) is due to two separate effects. First, the average intron size in imprinted genes is significantly smaller than in the control group. Second, the average number of introns per kb of exon is significantly lower for imprinted genes than the control set. The average size of individual exons in the control set is significantly smaller, by approximately one third, than the average for the imprinted genes. In contrast, the average total exon size (approximately the size of the cDNAs) of imprinted genes is no different from that of the control set and so can be ruled out as a confounding variable.

The conclusion that imprinted genes are different from the control set is highly robust to alterations in the analysis. The current data[3–5] we consider includes *hCGβ* as an imprinted gene, although this remains to be proven. However, if removed from the study, the qualitative results remain unaltered (meaning, after control for Type I Error, the statistics

Table 1 Gene structure parameters of the murine and human imprinted genes.

gene	number of exons	total exon size	total intron size	average exon size	average intron size	number of introns/kb	total intron size/total exon size	ref.
Igf2 (mouse)[a]	4	4,209	7,270	1,052	2423	0.71	1.73	see [a]
IGF2 (human)[b]	4	5,270	6,430	1,318	2140	0.57	1.22	see [b]
p57KIP2 (human)	4	1,313	328	817	272	2.28	0.62	c
Igf2r (mouse)	48	8,891	81,768	185	1740	5.3	9.2	26
H19 (mouse)	5	2,248	270	450	68	1.77	0.12	27
H19 (human)	5	2,332	347	466	86.75	1.72	0.15	28
CGb (human)	3	716	585	239	293	2.8	0.82	29
SNRPN (human)†	8	1300	4,800	163	686	5.4	3.69	30
Xist (mouse)	6	15,000	4,100	2,500	820	0.33	0.27	31
Insulin 1 (mouse)	2	435	118	218	118	2.29	0.27	32
Insulin 2 (mouse)	3	435	606	145	303	4.6	1.393	32
U2afbp-rs (mouse)	1	2,950	-	2,950	-	0	0	33
Mas-oncogene (mouse)	1	1,217	-	1217	-	0	0	e
IPW (human)	3	2,075	2925	692	1463	0.96	1.41	34
Znf127 (mouse)	1	2,700	-	2,700	-	0	0	f
Mash2[d] (mouse)	2	1,526	100	763			0.07	35
Averages for the imprinted genes (±SEM)	6.5 (±3.24)	3,297 (±1083)	9,282 (±7279)	1,027 (±266)	729 (±220)	1.92 (±0.52)	1.37 (±0.66)	
Averages for the control set (±SEM)	12.01 (±1.34)	2,816 (±232)	23,230 (±3552)	378 (±43)	2,396(±303)	3.99 (±0.25)	7.64 (±0.88)	
Significance level from Mann-Whitney	-	-	-	-	$P = 0.0025$[h]	$P = 0.0018$	$P<0.0001$	
Significance after controls for Type I Error	-	-	-	-	$0.01>P>0.001$	$0.01>P>0.001$	$P<0.0001$	

The averages for the parameters are given below both for the imprinted genes (combining mouse and human genes) and for the control set (fifty from mouse, forty from human). For each of three parameters that we are interested in the Mann-Whitney U test was employed to compare the set of randomly selected control genes (n=90) and the imprinted genes (n=14) under the null hypothesis that there was no difference between the two groups. The resulting P values thus generated are given below. These P values are then corrected using the Bonferroni adjustment (for Type I Error, assuming three tests have been performed) and the resulting levels of significance are given. To avoid non-independence we utilised only the human sequences for *Igf2* and *H19* (hence n=14). Usage of the mouse genes instead makes no meaningful difference.

[a]Murine *Igf2* has three different promoters and makes three different mRNA[21] The values given here are the averages for these three putative transcripts.

[b]Human *Igf2* has at least six different transcripts[22] but one is adult specific and not imprinted so is hence not included. For the remainder we consider only the three transcripts utilising the complete *Igf2* coding region[23]. We take an average value of these three (intron sizes obtained from ref. 23).

[c]From EMBL — accession number: D64137.

[d]The published sequence of this gene may be missing two exons (R. Nicholls pers. comm.)

[e]The human gene is intronless[24] but not yet shown to be imprinted. The mouse gene is known to be imprinted, is of known cDNA size[25] (that which we employ), but not yet genomically characterised. We assume by comparison with the human sequence that it is intronless (note, the number of introns/exons in genes tends to be highly conserved within mammals).

[f]R. Nicholls personal communication.

[g]Estimate from graphical representation of gene structure.

[h]For this comparison n in the control set is 88 and n in the imprint set is 11 (intronless genes are excluded).

Table 2 The comparison between i) mouse and human genes and ii) human autosomal and X-linked genes.

	number of exons	total exon size	total intron size	average exon size	average intron size	number of introns/kb	total intron size/total exon size
Average for mice (n = 50) (±SEM)	9.36 (±1.08)	2339 (±210)	13084 (±2321)	399 (±60)	1847 (±307)	3.96 (±.35)	5.45 (±.79)
Average for human (n = 40) (±SEM)	15.32 (±2.63)	3411 (±436)	35914 (±6993)	353 (±60)	3086 (±551)	4.02 (±.35)	10.37 (±1.6)
Significance from Mann-Whitney U test					P=0.025	P=0.75	P=0.012
Significance after correction					ns	ns	0.05>P>0.01
Average for human X (n = 35) (±SEM)	8.54 (±0.99)	2802 (±365)	37418 (±8061)	514 (±107)	5634 (±1351)	3.34 (±0.4)	13.12 (±2.37)
Average for human A (n = 40) (±SEM)	15.32 (±2.63)	3411 (±436)	35914 (±6993)	353 (±60)	3086 (±551)	4.02 (±.35)	10.37 (±1.6)
Significance from Mann-Whitney U test					P=0.11	P=0.13	P=0.60
Significance after correction					ns	ns	ns

For both pairs of tests the derived P values following the Mann-Whitney U test are corrected for Type I Error by application of the Bonferroni adjustment assuming three tests have been performed. We find that mouse genes have significantly smaller introns than human genes (P<0.05). This effect is sensitive, however, to control for Type I Error. In contrast, and in accord with a previous report[7], mouse genes have a significantly lower ratio of the total intron size to the total exon size (P<0.05) and this remains significant after control for Type I Error. Mouse and human genes have the same number of introns per kb of exon. To counter the effect of differences between mouse and human genes, the proportion of human to mouse genes in the control data set is matched to that in the imprinted set.

remain significant). Likewise, the two insulin genes can be treated as non-independent points. Most conservatively, when *Ins1* is removed from the analysis, the qualitative results again remain unaltered. If the three non-protein coding genes are removed (namely *Xist*, *H19* and *IPW*) then, again, the qualitative results remain unaltered. Indeed, if these three and *hCGβ* are removed the statistics all remain significant.

Our list of imprinted genes does not include any that are polymorphically imprinted; therefore *WT1* was not included. Although, this gene does not fit our observed pattern (it has large introns) its incorporation into the data set does not qualitatively affect the results. Interestingly, the only other gene that does not fit the pattern — *Igf2r* — is not imprinted in all species.

We conclude that imprinted genes tend to have few and small introns; not only less total intronic DNA, but also fewer introns per kb of exon. We have also compared a collection of 35 human X-linked genes (not pseudo-autosomal) with 40 human autosomal sequences and found no significant differences (Table 2). Thus, the limited intronic content of imprinted genes does not appear to be simply a consequence of haploid expression.

Our finding corresponds with a previous analysis showing that genes in GC-rich isochores (where imprinted genes tend to be localized[6]) on average have a low ratio of intronic to coding DNA[7]. This correlation is not understood but, may be a consequence of increased recombination rates in GC rich regions[7]. Imprinted genes appear to have higher recombination rates in males than females (the converse of the usual)[2]. We are unaware of reports that the absolute rate is higher for imprinted genes.

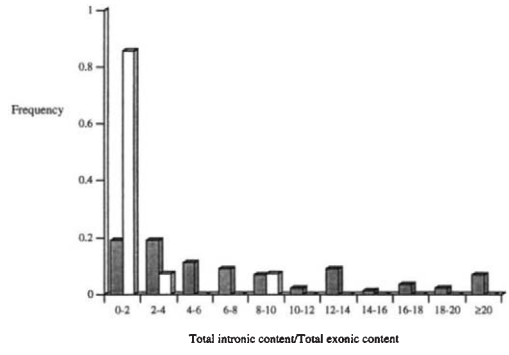

Fig. 1 The distributions of the ratio of the total intron size to the total exon size in the control set (n = 90) and in the set of imprinted genes (n = 14). The Mann-Whitney test shows the distributions of the two sets of genes to be highly significantly different (P < 0.0001). On the average, imprinted genes have between 1/5 and 1/6 the intronic DNA per unit length of exon that is found in the control set. The imprinted genes are shown in white, the control set in grey.

We speculate that the scarcity of intronic DNA in imprinted genes may be a consequence of selection favouring the maximization of the effective dosage of gene products. The evolution of imprinting may be a consequence of a 'genomic conflict'. The 'conflict theory' of imprinting[8–11] follows as a simple extension of classical parent-offspring conflict. In species with multiple paternity, within or between broods, paternally inherited genes are less related to fellow progeny of the same mother than are maternally inherited genes. It follows that paternally derived genes will be under selection to extract more resources from the mother than it is in her interest to give. The maternally derived genome will, by equal measure, be under selection to inhibit this to some degree. The activity of these genes could directly affect embryonic growth or do so indirectly through, for example, altering suckling behaviour[11].

From over 14 theories of the evolution of mammalian genomic imprinting[12], the conflict theory is unique in predicting that if paternally expressed genes directly affect embryonic growth (rather than suckling

behaviour and so on) they should be growth enhancing, while if maternally expressed, and directly affecting growth, they should suppress this effect. At least seven imprinted genes are thought to affect growth[13]. *Igf2*, *Mas* and *Ins2* are all paternally expressed and enhance growth rates. By analogy *Ins1* is expected to be a growth enhancer. *Igf2r*, *H19* and $p57^{KIP2}$ are all expressed off the maternally derived genome and reduce growth rates. In addition, *Mash2* is expressed off the maternally derived genome and is thought to affect growth. The effect is however ambiguous — deletion of *Mash2* is associated with placental failure, but maternal duplication of the region[14] is associated with growth retardation (this could be due to linked genes). Ignoring *Mash2*, the correspondence between the direction of growth effects and direction of imprinting can be analysed employing the G-test of independence with Williams' correction for a 2×2 table. This reveals $G_{adj}=7.8$, $0.01 > P > 0.001$ showing that there is a significant non-independence between the direction of imprinting and the direction of growth effects. This effect is robust to alterations in assumptions, including the addition of *Mash2* as a growth enhancer, the removal of *H19* and the removal of one insulin gene.

If we assume that the conflict theory is correct — as supported by duplications and knockouts — the dosage of the products of these genes is important to the net effect on growth (refs 15,16). Consider then a mutant paternally expressed gene that could somehow slightly increase its net dosage (or that of another paternally expressed product). The theory predicts that such a mutant allele would spread and also provide the conditions for the spread of a mutant maternally expressed antagonist that had increased dosage to re-suppress the paternal allele. This process could potentially reiterate, possibly indefinitely, what might be referred to as an 'arms race' between maternally and paternally derived genes. The selection favouring increased dosage of products would not result in selection for the loss of imprinting, only for increased product levels from the active alleles. This arms race may be manifested in terms of selection promoting a variety of features: i) If transcription of intronic DNA is time consuming, the arms race may lead to the condition where transcription rate is limiting and selection may then act to favour reduced intronic sequences so as to accelerate the rate of expression per gene. ii) Selection for increased dosage per unit time might favour the activation of duplicated versions of the genes concerned[17] (compare refs 18,19). If some duplications are generated by retroposition then imprinted genes as a class should be expected to have fewer introns per kb than the control set. In addition, iii) if paternal interference with maternal genes (and vice versa) can be effected through altering patterns of RNA splicing, possibly leading to mRNA degradation or inviability, then avoidance of such suppression, through loss of introns or intronic residues, may be expected.

Although these possibilities, which are not mutually exclusive, are theoretically feasible they remain to be established. The first model is given some credence by the observation of substantial effects on growth of deletions and duplications and more directly, of abundant transcripts, for example in placental tissues, of *H19*, *Igf2*, $p57^{KIP2}$ and others. It is unclear whether this model is of general applicability. Consistent with the first model would be the finding that non-imprinted genes that are very rapidly transcribed also have few and/or small introns. Perhaps this explains why genes in GC% rich isochores tend to have relatively little intronic DNA.

The possibility that imprinted genes may commonly be members of multigene families can be anecdotally supported (for example the multiple insulin genes and the multiple copies of *hCGβ*). In addition, this model would predict that imprinted genes should more commonly be intronless than other genes. Applying the *G* test of independence with the Williams' control to the data sets that we have (3/14 imprinted genes 2/90 in the control set are intronless), it is found that the frequency of intronless genes is higher than expected

($G_{adj} = 5.09$, $P < 0.05$). There may be an alternative interpretation of this finding. If only a limited number of chromosomal domains allow the creation of an imprint, then by chance, genes that can retropose more often than others are more likely to become imprinted as they are more likely to end up in one of these few domains.

The third possibility predicts that imprinted genes may affect, and be affected by, splicing. This prediction is consistent with the existence of imprinted genes which potentially affect splicing (*U2afbp-rs* and *SNRPN*) and with the control of splicing witnessed for some transcripts of *Igf2* (ref. 20).

The hypothesised connection between an intron paucity and maternal-fetal conflict can be falsified by showing that the imprinted genes have always had small introns, not that they lost them. Limited data suggest that the intronic content was lost. For both insulin and *Igf2* there is an increase in GC% found in mammals that is not found in the non-imprinted *Igf1* suggesting that the increased GC% of imprinted genes is correlated with the assumption of imprint status[6]. If intron size and GC% are causally related then the imprinted genes are expected to have lost their intronic DNA with the rise in GC%. As expected, *Igf1* has huge introns.

Laurence D. Hurst
Gilean McVean
Department of Genetics, Downing Street, Cambridge, CB2 3EH, UK

Tom Moore
Department of Development and Signalling, Babraham Institute, Cambridge, CB2 4AT, UK

1. Neumann, B., Kubicka, P. & Barlow, D.P. *Nature Genet.* **9**, 12–13 (1995).
2. Paldi, A., Gyapay, G. & Jami, J. *Curr. Biol.* **5**, 1030–1035 (1995).
3. Haig, D. *Prenatal Diagnosis* **13**, 151 (1993).
4. Degroot, N. *et al. Prenatal Diagnosis* **13**, 1159–1160 (1993).
5. Goshen, R. *et al. Am. J. Obst. Gyn.* **170**, 700–701 (1994).
6. Ellsworth, D.L., Hewettemmett, D. & Li, W.H. *Mol. Biol. Evol.* **11**, 875–885 (1994).
7. Duret, L., Mouchiroud, D. & Gautier, C. J. *Mol. Evol.* **40**, 308–317 (1995).
8. Haig, D. & Westoby, M. *Am. Nat.* **134**, 147–155 (1989).
9. Haig, D. & Graham, C. *Cell* **64**, 1045–6 (1991).
10. Haig, D. & Westoby, M. *Phil. Trans R. Soc. Lond.* **B333**, 1–13 (1991).
11. Moore, T. & Haig, D. *Trends Genet.* **7**, 45–49 (1991).
12. Hurst, L.D. in *Genomic Imprinting* (eds Reik, W. & Surani, A.) (Oxford University Press, Oxford, In the press).
13. Hatada, I. & Mukai, T. *Nature Genet.* **11**, 204–206 (1995).
14. Guillemot, F. *et al. Nature Genet.* **9**, 235–242 (1995).
15. Efstratiadis, A. *Curr. Opin. Genet. Devel.* **4**, 265–280 (1994).
16. Solter, D. *Annu. Rev. Genet.* **22**, 127–146 (1988).
17. Haig, D. Q. *Rev. Biol.* **68**, 495–532 (1993).
18. Hurst, L.D. *Genetics* **130**, 229–30 (1992).
19. Hurst, L.D. *Genetics* (In the press).
20. Nielsen, F.C. *et al. Nature* **377**, 358–362 (1995).
21. Rotwein, P. & Hall, L.J. *DNA Cell Biol.* **9**, 725–735 (1990).
22. Holthuizen, P. *et al. Biochem.* Biophys. Acta **1087**, 341–343 (1990).
23. Dull, T.J. *et al. Nature* **310**, 777–781 (1984).
24. Young, D. *et al. Cell* **45**, 711–719 (1986).
25. Metzger, R. *et al. Febs Letters* **357**, 27–32 (1995).
26. Szebenyi, G. & Rotwein, P. *Genomics* **19**, 120–129 (1994).
27. Pachnis, V., Brannan, C.I. & Tilghman, S.M. *EMBO J.* **7**, 673–681 (1988).
28. Brannan, C.I. *et al. Mol. Cell. Biol.* **10**, 28–36 (1990).
29. Policastro, P. *et al. J. Biol. Chem.* **258**, 11492–11499 (1983).
30. Ozcelik, T. *et al. Nature Genet.* **2**, 265–269 (1992).
31. Brockdorff, N. *et al. Cell* **71**, 515–526 (1992).
32. Wentworth, B.M. *et al. J. Mol. Evol.* **23**, 305–312 (1986).
33. Hayashizaki, Y. *et al. Nature Genet.* **6**, 33–40 (1994).
34. Wevrick, R., Kerns, J.A. & Francke, U. *Hum. Mol. Genet.* **3**, 1877–1882 (1994).
35. Guillemot, F. *et al. Nature* **371**, 333–336 (1994).

Human choroideremia protein contains a FAD-binding domain

Sir — Patients with choroideremia, an X-linked form of late-onset retinal degeneration, carry mutations in the *CHM* gene that has been positionally cloned[1,2]. The *CHM* gene encodes the large subunit of geranylgeranyl transferase II, an enzyme that attaches geranylgeranyl groups to C-terminal cysteines of Rab GTP-binding proteins[3,4]. This subunit is not required for the catalytic activity but binds the Rab protein and presents it to the catalytic component[5]. After geranylgeranylation, it remains tightly associated with the Rab protein and is thought to deliver it to another protein which is directly responsible for the insertion of Rab into membranes; therefore the CHM gene product and other closely related proteins are called Rab escort proteins, or REPs[5,6]. Proteins homologous to REPs have also been identified as GDP dissociation inhibitors (GDIs) that prevent the GDP/GTP exchange and release GDP-bound Rab from membranes, thus playing a role in Rab recycling[7–10]

REPs and GDIs are highly conserved from mammals to yeast[11]. However, analysis of their amino acid sequences have so far failed to provide any clues for the activity of these proteins. I report here that REPs and GDIs contain a putative FAD-binding domain.

A comparison of the CHM protein sequence to the non-redundant protein sequence database (National Center for Biotechnology Information, NIH) with the BLASTP program[12] revealed, in addition to the high similarity to other REPs and GDIs, a limited similarity to several dehydrogenases (the probability of finding the alignment with the putative dTDP-4-dehydrorhamnose reductase from *Mycoplasma genitalium* was about about 0.08). Even though this similarity was not highly statistically significant, the respective regions of the dehydrogenases included their most conserved segment, the N-terminal FAD-binding motif[13]. A further database search for conserved sequence motifs with the MoST program[14] detected a highly significant relationship (see legend to Fig. 1) between these