GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Letter to the Editor

Boning up on mutations: assessing the significance of candidate disease-causing DNA sequence variation

R. Dalgleish

Department of Genetics, University of Leicester, Leicester, United Kingdom

Corresponding author: R. Dalgleish
E-mail: raymond.dalgleish@le.ac.uk

Genet. Mol. Res. 10 (3): 1518-1521 (2011)
Received June 10, 2011
Published August 1, 2011
DOI 10.4238/vol10-3gmr1353

Dear Editor,

I write to you regarding the paper by Yang et al. (2011) entitled "Mutation characteristics in type I collagen genes in Chinese patients with osteogenesis imperfecta", which was recently published in Genetics and Molecular Research.

The impression given by the authors is that they have identified disease-causing DNA sequence variants in *COL1A1* and *COL1A2*, which result in osteogenesis imperfecta (OI) in five of the probands studied. However, the data presented in the paper do not support such a proposition and I will discuss each variant in turn. However, it is important to first clarify the identity of the relevant reference DNA sequences and variant reporting conventions.

Reference sequences and reporting conventions

The authors state that the reference sequences used were NC_00001 and NC_007405, respectively, for *COL1A1* and *COL1A2*. The former sequence does not exist in GenBank, and the latter is that of the mitochondrial genome of the algal species *Thalassiosira pseudonana*. The correct genomic DNA reference sequences for the two genes are NG_007400.1 and NG_007405.1. The newly emerging Locus Reference Genomic (LRG) (Dalgleish et al., 2010) reference sequence records LRG_1 and LRG_2 are also valid alternatives.

The authors state that "*Nucleotide changes at the genomic DNA level were numbered from the first base of the transcript*". This is a confusing statement on two levels. First, all variants are actually reported in terms of cDNA (c.) coordinates, not genomic DNA. Secondly, the convention for "c." coordinates is that base 1 is the A of the ATG codon at which translation begins, rather than the first base of the transcript.

It is also stated that "*Amino acid changes were numbered from the first glycine of the triple helix*". It is clear from the data presented in Table 4 that amino acid numbering actually begins, as it should, with the methionine at the beginning of the primary translation product. Numbering amino acids from the start of the collagen triple helix is a legacy system, which is deprecated.

COL1A1:c.97G>A (Family B: type I OI)

The c.97G>A variant in *COL1A1* actually results in the amino acid variant p.(Glu33Lys) and not in p.(Asp33Asn). There are Asp amino acids at positions 32 and 34, and it is evident from the DNA sequence (GAARACAGTAAGT) presented in Figure 2A that the actual variant in Family B is c.100G>A, yielding the amino acid variant p.(Asp34Asn).

Irrespective of the actual identity of the DNA-level or protein-level variant found in Family B, there is no convincing evidence presented that it is disease-causing. The authors state that "*Patients' unaffected parents were also studied in some instances to ascertain the presence of novel sequence changes*". However, no information is provided about whether this variant was found in either parent in this family. Such information would at least clarify whether the proposed disease-causing variant was inherited in a Mendelian fashion or if it arose *de novo* in the proband or in the germline of one parent. The authors go on to state that fifty normal chromosomes were studied and the variant was not found. Such a finding does not, in itself, prove that the detected variant is disease-causing.

There is speculation in the discussion that the sequence change might have an effect on splicing at the junction between exon 1 and intron 1, which would be consistent with the phenotype of OI I in the proband. However, analysis of the mutant and wild-type sequences using the SplicePort interactive splice-site analysis tool (Dogan et al., 2007) yields scores of 1.23254 and 1.02714, respectively, suggesting that the mutant allele encodes a more effective splice donor site.

It could be argued that the single-base substitution has an effect on splicing by disrupting an exonic splice enhancer (ESE) site (Lin and Fu, 2007). Analysis using ESEfinder 3.0 (Cartegni et al., 2003) indicates loss of a binding site for SF2/ASF at the exon/intron boundary (GACAgta), but the significance of this would have to be tested to see if it is functional.

An approach to determining the significance of amino acid changes is to consider sequence conservation among different species. Alignment of the amino acid sequence of the α1 chain of type I collagen from 14 vertebrate species indicates that the amino acid corresponding to position 34 in the human protein is absolutely conserved. This may indeed be evidence that the p.(Asp34Asn) variant is disease-causing, but more investigation would be required to substantiate this view.

COL1A2:c.87T>C (Families A and E: type I OI)

Although this variant is correctly described at the DNA level, the correct HGVS nomenclature-compliant description for a silent variant at the protein level is p.(=). As the

authors note, this variant has been observed before and it is recorded in dbSNP (Sherry et al., 2001) with the identifier rs1801182. The dbSNP entry indicates that the frequency of both alleles has been determined in a cohort of 45 unrelated Han Chinese individuals with the less common C allele having a frequency of 0.289. Hence, it is not unexpected that this variant has been identified in Families A and E. However, it is surprising that the authors did not detect the variant in their own cohort of normal Han Chinese individuals. No pathogenicity has ever been ascribed to the variant, and analysis using ESEfinder 3.0 indicates that no ESE site is altered. Consequently, it is highly unlikely that this variant is the cause of the disease phenotype in either family.

The legend for Figure 2B erroneously indicates that this variant is in *COL1A1* rather than in *COL1A2*.

COL1A1:c.1209T>A (Family C: type I OI)

As with the previous variant, the correct HGVS nomenclature-compliant description for a silent variant at the protein level is p.(=). Analysis of the sequences of the two alleles using ESEfinder reveals minor differences in predicted binding sites for SF/ASF, SC35 and SRp40 but no binding sites are either created or eliminated by the base substitution. Consequently, it is highly unlikely that this variant is the cause of the disease phenotype in this family.

COL1A1:c.3702C>T (Family D: type II OI)

Once again, the correct HGVS nomenclature-compliant description for a silent variant at the protein level is p.(=). Analysis of the sequences of the two alleles using ESEfinder reveals the loss of a predicted binding site for SF2/ASF (IgM BRCA1) and gain of a predicted site, SRp40; however, the significance of these changes would need to be determined by testing for effect on function. In the absence of any other supporting evidence, it is unlikely that this variant is the cause of the disease phenotype in this family.

The proband in Family D is described in Table 3 as being 24 years old, but this is at odds with the perinatal lethal nature of OI type II.

Sequence variants in intron 31 of *COL1A1* and intron 30 of *COL1A2*

The authors present data in Figure 2 illustrating sequence variation in intron 31 of *COL1A1* (NG_007400.1:c.2127+128G>C) and intron 30 of *COL1A2* (NG_007405.1:c.1764+162G>A). The data are interesting, but appear to be measures of observed variant frequencies in the patient and control cohorts, rather than actual mutation rates as stated. For the *COL1A1* variant, the frequency is given as "33.3% (5/15)" but it is not clear how the number "15" is derived. There are 8 OI families in the study and the 16 chromosomes of the probands ought to have been tested and compared with the 50 normal control chromosomes. In addition, the relatively small numbers of cases and controls mean that the differences in allele frequencies may not be significant, especially if the cases and controls were not carefully matched, but no information is presented in this regard. Similar questions apply to the data presented for *COL1A2*, although the difference in allele frequencies is much more pronounced and probably warrants further investigation. In respect of the

COL1A2 data, it is not clear what is meant by "(~0%)". If there are 50 controls, the lowest non-zero frequency for the minor allele would be 0.01 (1%).

SUMMARY

The data presented by the authors do not support the claim that they have identified the causative mutations in five cases of osteogenesis imperfecta. The disease-causing variants may be harbored by the *COL1A1* and *COL1A2* genes, particularly in the familial cases, but the genes known to harbor recessive variants leading to OI need to be considered, too. The authors mention the *CRTAP* and *LEPRE1* genes, but, in addition, *FKBP10*, *PLOD2*, *PPIB*, *SERPINF1*, *SERPINH1*, and *SP7* need to be screened for disease causing-variants.

REFERENCES

Cartegni L, Wang J, Zhu Z, Zhang MQ, et al. (2003). ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* 31: 3568-3571.

Dalgleish R, Flicek P, Cunningham F, Astashyn A, et al. (2010). Locus reference genomic sequences: an improved basis for describing human DNA variants. *Genome Med.* 2: 24.

Dogan RI, Getoor L, Wilbur WJ and Mount SM (2007). SplicePort - an interactive splice-site analysis tool. *Nucleic Acids Res.* 35: W285-W291.

Lin S and Fu XD (2007). SR proteins and related factors in alternative splicing. *Adv. Exp. Med. Biol.* 623: 107-122.

Sherry ST, Ward MH, Kholodov M, Baker J, et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29: 308-311.

Yang Z, Ke ZF, Zeng C, Wang Z, et al. (2011). Mutation characteristics in type I collagen genes in Chinese patients with osteogenesis imperfecta. *Genet. Mol. Res.* 10: 177-185.

