Journal of Biology

BioMed Central

Minireview

Pitfalls in the phylogenomic evaluation of human disease-causing mutations

Andrew OM Wilkie

Address: Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK.
Email: awilkie@hammer.imm.ox.ac.uk

Published: 24 March 2009

Journal of Biology 2009, **8**:26 (doi:10.1186/jbiol127)

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/8/3/26

Abstract

A detailed sequence comparison of the MSX homeobox family sheds light on its evolution and identifies new conserved motifs. But in the absence of corroborative genetic data, phylogenomics alone can provide only limited insights into the pathogenicity of heterozygous missense substitutions in human genes.

The explosion in genome sequencing provides a rich resource for reconstructing the evolutionary origins of gene families. One proposed application for such phylogenomic information is to identify highly conserved sequences in human proteins suspected of being associated with disease, and to use this information to identify sequence variants in these regions as potential disease-causing mutations. A recent example of this approach is a study by Finnerty *et al*. of the MSX homeobox family published in *BMC Evolutionary Biology* [1]. MSX is of particular interest because it represents one of the most ancient families of animal homeodomain proteins, and mutations in both paralogous human genes, *MSX1* and *MSX2*, have been associated with craniofacial disorders [1,2]. The work by Finnerty *et al*. [1], which focuses on the *MSX1* sequence changes, provides a useful case study in the context of current initiatives to generate large amounts of genomic sequence data from complex diseases. These will yield thousands of rare sequence variants, causing headaches for interpretation of the pathogenicity of individual sequence changes. So, how successful has the analysis of MSX sequences been in aiding interpretation of human MSX1 sequence variations?

Human MSX: evolution and domain organization

The human genome contains two MSX paralogs, *MSX1* located at 4p16.2 and *MSX2* at 5q35.2. There is strong evidence that these genes arose from the second round of whole-genome duplication that took place at the base of the vertebrate radiation (the additional two copies expected from these duplication events have been lost in humans, but rodents retain an *Msx3* gene predicted to have split from *Msx1*/*Msx2* at the first duplication event).

Apart from the well-known homeodomain ('MH4'), Finnerty *et al*. [1] confirm and extend a recent analysis [3], finding six other highly conserved sequence elements within human MSX proteins, which they term MH1N, MH1C, MH2, MH3, MH5 and MH6 (Figure 1a). The elements MH1N and MH1C exhibit homology, suggesting that they arose from an ancient duplication of a Groucho-binding domain; MH1C has been secondarily lost in MSX2 (and independently in rodent Msx3), but is retained in MSX1. MH6 near the carboxyl terminus is a newly identified motif and represents a PIAS-binding domain. Finnerty *et al*. [1] convincingly demonstrate that use of phylogenetically deep

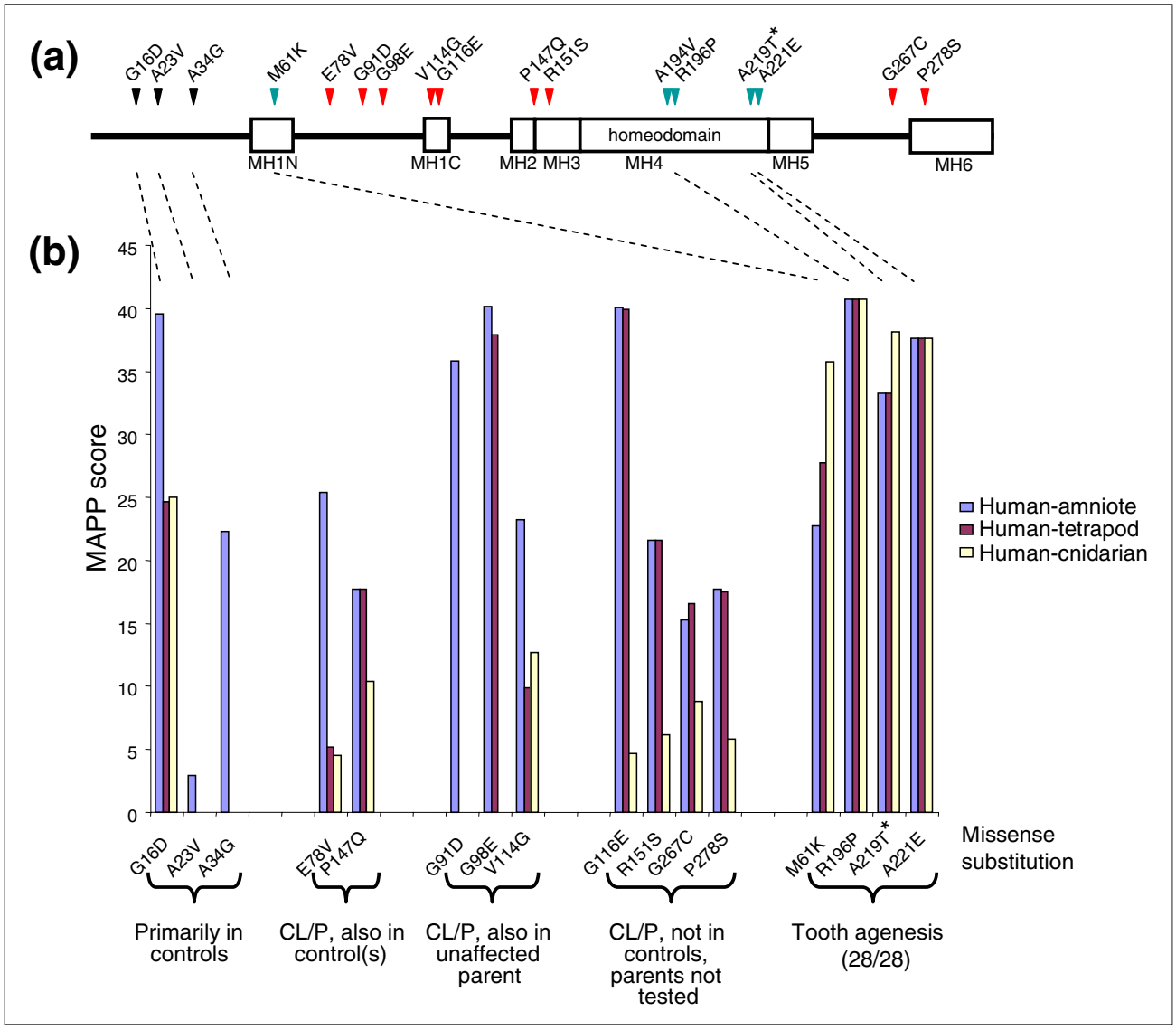

Figure 1
MSX1 structure and MAPP evaluation of sequence changes. **(a)** Cartoon of protein [1] showing relative positions of seven conserved motifs (boxes) and missense substitutions (arrowheads), colored according to whether they have been identified primarily in control samples (black), tooth agenesis (blue) or CL/P (red). The asterisk indicates that the A219T substitution is only associated with the phenotype in homozygotes. **(b)** MAPP scores for each substitution, arranged according to evidence for pathogenicity. Dashed lines linking to (a) indicate relative position in the protein for substitutions found in control and tooth-agenesis samples. Higher MAPP scores indicate a reduced likelihood that a substitution would be tolerated. Note that the A194V substitution [4] was not included in the MAPP analysis [1].

sequence comparisons can aid alignment of the more poorly conserved regions of the MSX proteins.

Having undertaken this alignment, sequence changes found in human MSX1 in samples from patients with either tooth agenesis or cleft lip/palate (CL/P) were mapped in relation to the conserved sequence elements, to help predict the severity of their functional effects [1]. Here I will focus on the missense changes, as these are the most difficult to interpret, and ask to what extent these efforts have succeeded.

Sequence variation in human MSX1: Mendelian tooth agenesis

Previous linkage studies of segregating Mendelian traits followed by candidate gene sequencing revealed sequence

changes in *MSX1* that are undoubtedly pathogenic; they show highly significant statistical association with disease (by segregation through a family) and are associated with a consistent phenotypic pattern of presentation and high penetrance. These heterozygous *MSX1* mutations characteristically cause agenesis (loss) of elements of the secondary dentition, especially the second premolars and third molars. The phenotype of these missense mutations can be deduced to be due to haploinsufficiency because dominant mutations that obviously confer loss of function (complete gene deletions, nonsense and frameshift mutations) give an identical phenotype.

The positions of the five MSX1 missense mutations that fit this category (M61K, A194V, R196P, A219T, A221E) as mapped onto the conserved sequence elements identified by Finnerty *et al.* [1] are shown in Figure 1a (blue arrowheads). They are all located within the most highly conserved regions of the protein (one in the MH1N domain and four in the homeodomain) and collectively exhibit very high disease penetrance for tooth agenesis (29 of 31 with the relevant mutant genotype); none of these individuals had CL/P. So far, so good: the molecular predictions appear to agree with the genetics. However, two important caveats should be noted. First, in the report of the A194V mutation [4], only one of the three heterozygotes studied had any dental abnormality, indicating that this particular substitution is associated with incomplete penetrance [4]. Second, in the report of the A219T mutation, only homozygous individuals exhibited dental abnormalities (five of five individuals); none of the eight heterozygotes identified had any dental manifestations [5]. This suggests that the particular missense alleles A194V and A219T confer only partial loss of function, to different extents - that is, they are hypomorphs - and it illustrates an important limitation to the type of *in silico* analysis carried out by Finnerty *et al*. [1]. Simply demonstrating that a sequence change is likely to be disruptive is an insufficient criterion for disease causation, as it does not predict whether (and in what proportion of individuals) that change will produce a disease phenotype when present in the heterozygous state. Only empirical genetic analysis can answer that question.

Sequence variation in human MSX1: cleft lip/palate in case-control studies

In contrast to the demonstrated Mendelian inheritance of MSX1 defects in tooth agenesis, the association of mutations in MSX1 with human CL/P are based on genetic data that are much less robust for each individual sequence variant. Prompted by the clefting phenotype in *Msx1*$^{-/-}$ mice [6] and by the occurrence of CL/P associated with a heterozygous S105X mutation in four out of twelve members of a family segregating tooth agenesis [7], several groups have undertaken DNA sequencing of large numbers of CL/P cases and compared these with control samples. These studies yielded rare heterozygous missense changes in around 1% of cases [8], prompting claims that MSX1 mutations are an important 'cause' of CL/P. Importantly, none of the variants identified resides within the MH1N or homeodomain regions harboring the tooth-agenesis mutations; rather, they locate to other regions of the protein, some in the remaining conserved motifs described above, and some outside them (Figure 1a, red arrowheads).

On the basis of the MSX phylogenomic analysis, Finnerty *et al*. [1] attempted to analyze the pathogenicity of each of these variants individually, as judged by the degree of sequence conservation at their location and thus the potential effect of the mutation on protein structure and function. Several considerations indicate that this exercise will be problematic. In contrast to the tooth-agenesis mutations, none of the CL/P variants has presented in a pedigree showing clear Mendelian inheritance: at best, some familial clustering is observed, suggesting a more complex causation involving multiple genetic and/or environmental factors. In cases with available parental samples, one parent has always been found to harbor the same variant, even when they are unaffected themselves. Most of these variants have been identified in only single CL/P cases, making the task of obtaining a statistically robust distinction from controls formidably difficult (if 1 variant is found in 100 affected cases, it must be absent from 1,900 controls to obtain a *P*-value for the difference of 0.05). In the two instances where the variants have been discovered in multiple affected samples (E78V and P147Q), they have also turned up in control sample(s) from ethnically matched populations.

The most direct way to estimate the penetrance of CL/P associated with these variants would be to trace them back through the proband's family and ask what proportion of heterozygotes was affected. However, few such cascaded family studies have been undertaken. Where they have been performed (for example, in the cases of the G116E [8] and P147Q [9] variants) the correlation with phenotype has been poor, with the variant absent in some affected family members and present in some unaffected members. In this difficult context, can phylogenomic analysis help to sort out which of these sequence changes may be conferring a higher liability for CL/P than others?

Interpreting pathogenicity from sequence conservation and protein motifs

Finnerty *et al*. [1] examined the impact of amino acid changes in human MSX1 using the multivariate analysis of

protein polymorphism (MAPP) program [10]. This evaluates pathogenicity on the basis of both sequence conservation at the substituted position and the comparative physicochemical properties of the wild-type and substituted amino acids. MAPP analysis was performed at three different depths of sequence conservation, human-amniote, human-tetrapod and human-cnidarian [1]. Although MAPP is not the only method for undertaking this type of analysis, it is unlikely that choice of a different algorithm would have substantially altered the conclusions.

Combining the MAPP analysis with the location of sequence changes relative to the conserved elements, Finnerty *et al.* [1] concluded that several of the CL/P variants were likely to be disease alleles. They further proposed that the different MSX1 mutant phenotypes are related to whether the sequence changes occur in regions functionally redundant with MSX2. From this viewpoint, mutations in the highly conserved MH1N and MH4 regions cause 'mild' phenotypes because MSX2 can partially replace these roles; by contrast, mutations outside these regions (the amino terminus excepted) cause 'strong' CL/P phenotypes because they affect the nonredundant functions of MSX1 (for example, those involving the MH1C domain). The authors further proposed that the CL/P variants are acting as dominant-negatives. Although ingenious, this explanation is not entirely convincing. From the genetic evidence the CL/P variants are not dominant negative - they are neither simple dominants, nor associated with the same phenotype as loss-of-function mutation. Nor do they preferentially occur in the conserved regions with functions supposedly distinct to MSX1 (Figure 1a). An equally plausible interpretation is that the location of the CL/P variants simply reflects avoidance of the most highly conserved parts of MSX1, and that they represent a bunch of susceptibility alleles of varying degrees of weakness, which sometimes act in concert with other genetic/environmental factors to disrupt palatogenesis.

One can evaluate the limitations of MAPP analysis in this type of situation by regrouping the results of the analysis of Finnerty *et al.* (given in Figure 7 of [1]) according to the phenotype with which the sequence change has been associated, and according to the strength of the genetic evidence supporting the association (Figure 1b). The only consistent feature in these three analyses is that the four tooth-agenesis mutations examined have high MAPP scores, indicating that the amino acid position affected is highly conserved and the altered residue is therefore likely to be deleterious. Note, however, that the recessive A219T substitution is indistinguishable at all three evolutionary levels from the other, dominantly transmitted, changes. Turning to the other missense substitutions, no trends are apparent. There is substantial variation in MAPP scores both within and between categories, and the most consistently high set of scores concern a sequence change, G16D, that was observed in controls [8] rather than CL/P samples (Figure 1a, black arrowhead). This inconsistency indicates that the ability of MAPP analysis to predict the penetrance of different heterozygous sequence changes associated with CL/P is likely to be poor.

There's no substitute for good genetic studies!

Ultimately, the interpretation of the data on MSX1 mutations in CL/P [1] is undermined by the key consideration that we cannot easily know what the consequence of a missense change - that might be obviously pathogenic in the homozygous state - will be in the heterozygous state. We need a framework in order to make such interpretations, as we have in the case of the Mendelian condition of tooth agenesis. Here, we can conclude, by genetic and comparative arguments, that certain mutations cause complete or partial loss of function. Such a framework is currently missing for these CL/P variants.

So, are phylogenomic comparisons of no use in interpreting disease-associated mutations? Of course this is not the case; I frequently use such evaluations in my own work on Mendelian mutations. But the difficulties become much greater when attempting to understand the significance of rare variants in common complex disease. Ultimately, the only sure way to interpret the disease burden associated with these CL/P variants will be to undertake much larger case-control studies, and to ensure that thoroughly cascaded family follow-up is performed on those rare sequence changes that are encountered.

Acknowledgements

Work in the author's laboratory is funded by Wellcome Trust Programme and MRC Project Grants.

References

1. Finnerty JR, Mazza ME, Jezewski PA: **Domain duplication, divergence, and loss events in vertebrate Msx paralogs reveal phylogenomically informed disease markers.** *BMC Evol Biol* 2009, **9:**18.
2. Mavrogiannis LA, Taylor IB, Davies SJ, Ramos FJ, Olivares JL, Wilkie AOM: **Enlarged parietal foramina caused by mutations in the homeobox genes *ALX4* and *MSX2*: from genotype to phenotype.** *Eur J Hum Genet* 2006, **14:**151-158.
3. Takahashi H, Kamiya A, Ishiguro A, Suzuki AC, Saitou N, Toyoda A, Aruga J: **Conservation and diversification of Msx protein in metazoan evolution.** *Mol Biol Evol* 2008, **25:**69-82.
4. Mostowska A, Biedziak B, Trzeciak WH: **A novel c.581C>T transition localized in a highly conserved homeobox sequence of *MSX1*: is it responsible for oligodontia?** *J Appl Genet* 2006, **47:**159-164.
5. Chishti MS, Muhammad D, Haider M, Ahmad W: **A novel missense mutation in MSX1 underlies autosomal recessive oligodontia with associated dental anomalies in Pakistani families.** *J Hum Genet* 2006, **51:**872-878.

6. Satokata I, Maas R: ***Msx1* deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth development.** *Nat Genet* 1994, **6:**348-356.
7. van den Boogaard MJ, Dorland M, Beemer FA, van Amstel HK: **MSX1 mutation is associated with orofacial clefting and tooth agenesis in humans.** *Nat Genet* 2000, **24:**342-343.
8. Jezewski PA, Vieira AR, Nishimura C, Ludwig B, Johnson M, O'Brien SE, Daack-Hirsch S, Schultz RE, Weber A, Nepomucena B, Romitti PA, Christensen K, Orioli IM, Castilla EE, Machida J, Natsume N, Murray JC: **Complete sequencing shows a role for MSX1 in non-syndromic cleft lip and palate.** *J Med Genet* 2003, **40:**399-407.
9. Vieira AR, Avila JR, Daack-Hirsch S, Dragan E, Félix TM, Rahimov F, Harrington J, Schultz RR, Watanabe Y, Johnson M, Fang J, O'Brien SE, Orioli IM, Castilla EE, FitzPatrick DR, Jiang R, Marazita ML, Murray JC: **Medical sequencing of candidate genes for nonsyndromic cleft lip and palate.** *PLoS Genet* 2005, **1:**e64.
10. Stone EA, Sidow A: **Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity.** *Genome Res* 2005, **15:**978-986.