Elucidating the Genetics of Craniofacial Shape Using a Data Driven Approach

David M Evans[1,2]

1 University of Queensland, Diamantina Institute, Translational Research Institute, Brisbane, Australia

2 Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK

Alterations in craniofacial size and shape are apparent in many monogenic diseases and syndromes, however remarkably little is known regarding the genetics of face shape within healthy populations. However, this may be set to change following publication of a study this month that combines unsupervised hierarchical spectral clustering and canonical correlation analysis to help identify common genetic variants associated with craniofacial shape.

Despite considerable effort, investigators using genome-wide association studies (GWAS) have identified only a handful of common genetic variants robustly associated with craniofacial shape[1-6]. This lack of progress is perhaps surprising given the remarkable similarity of monozygotic twins' faces and the high heritability of the traits involved[4]. However, this may change following a new study by Peter Claes and colleagues in this month's *Nature Genetics*[7].

Claes et al. used digital facial stereophotogrammetry to capture the three dimensional facial structure of 2329 unrelated individuals of European ancestry (Figure 1). The authors first aligned the three dimensional facial images of their subjects to 10000 quasi-landmarks and adjusted the measurements for potentially confounding variables (e.g. age, sex, height, weight, population structure, face size etc). They then partitioned individuals' face shapes into a series of five bifurcating levels (representing global to more specific facial features) spread over 63 segments in total using an unsupervised method called hierarchical spectral clustering[8]. After applying a statistical method to superimpose the images and remove size effects within each facial segment, the authors subsequently performed a GWAS of each of the 63 facial regions. Specifically, they tested the relationship between each SNP in their GWAS and an optimally weighted linear combination of principal component variables that summarized the three dimensional variation within each facial region whilst simultaneously controlling for an increased multiple testing burden.

The authors identified a total of 1932 SNPs at 38 loci significantly associated with various aspects of craniofacial shape, including 1821 SNPs at 15 loci that subsequently replicated in an independent sample of 1719 individuals of European ancestry. Importantly, at least eleven of these loci had been implicated in previous GWAS of face shape, including in a very large study of 23andMe volunteers that had analysed self-report measures of chin dimple and nose size[9]. Interestingly, most of the loci that Claes et al. identified were associated with variation around the lower half of the face, in particular the nose and the chin. Some variants were most strongly associated with highly localized areas of the face, whereas others showed association across multiple regions and had more global effects. The authors subsequently showed marked enrichment of H3K27ac signals in the vicinity of the peak SNPs in cranial neural crest cells as compared to >30 other cell types suggesting that the implicated variants may exert functional effects early in development.

A new way to analyse craniofacial data?

What is perhaps most striking about the *Claes et al* study is the large number of genetic loci identified at impressive levels of significance, using a relatively small number of subjects. We think that this success is likely a consequence of the authors' innovative analysis strategy. Previous GWAS of craniofacial shape have utilized comparatively unsophisticated analysis strategies, most often correlating SNP genotypes with simple linear Euclidean distances between facial landmarks or qualitatively graded features of the face[1-6]. However, a potential weakness of this "phenotype first" approach is that the genetic polymorphisms that underlie variation in craniofacial shape may in fact influence composites of these predefined phenotypes and hence show only weak association with the individual components that comprise these complex traits[3].

In contrast, the clustering method that Claes et al. adopt is unsupervised and produces a set of phenotypes that have been "learned" from the data rather than being based upon predefined landmarks. The hierarchical partitioning ensures that the authors can simultaneously focus on local aspects of craniofacial structure without ignoring larger more general aspects of the face at previous levels of the hierarchy, and indeed the pattern of association across the different facial segments provides clues as to how the individual variants might exert their effects. Second, the approach of Claes et al. uses canonical correlation analysis of principal components representing local facial structures to maximize the correlation between SNPs and many facial dimensions simultaneously. Again this strategy avoids loss of information that would occur by pre-specifying measures that may not be optimal in terms of uncovering genetic aetiology, although the method requires care to regenerate the same phenotypes when evaluating the evidence for replication in independent samples.

A new way forward?

The importance of the Claes et al study in our opinion lies not so much in the biological insights gleaned from identifying individual loci associated with face shape, but rather that the study provides a blueprint of the way the field might move forward in the future. Over the past couple of years progress in identifying common genetic variants associated with face shape has been slow. This has been due to a number of factors including (but not limited to) the relative paucity of three dimensional face data amongst large research cohorts, differences in imaging technologies and protocols, difficulties in sharing individual level data and meta-analysing summary statistics of comparable variables, and perhaps as Claes et al. suggest, non-optimal strategies for analysing the data. Our hope is that the success of their study will catalyse a new wave of analyses of existing face shape data as well as foster extensive collaboration and sharing of data within the craniofacial genetics community. We note that the methods that Claes et al. espouse may also prove useful in understanding the genetics of other high dimensional morphological traits that rely on imaging technologies.

References

1. Adhikari, K. *et al. Nat. Commun.* **7**, 11616 (2016).
2. Cole, J.B. *et al. PLoS Genet.* **12**, e1006174 (2016).
3. Lee, M.K. *et al. PLoS One* **12**, e0176566 (2017).

4. Liu, F. *et al. PLoS Genet* **8**, e1002932 (2012).
5. Paternoster, L. *et al. Am. J. Hum. Genet.* **90**, 478-85 (2012).
6. Shaffer, J.R. *et al. PLoS Genet.* **12**, e1006149 (2016).
7. Claes, P. *et al. Nat. Genet.* (2018).
8. Ng, A.Y., Jordan, M.I. & Weiss, Y. *Advances in Neural Information Processing Systems 14, Vols 1 and 2* **14**, 849-856 (2002).
9. Pickrell, J.K. *et al. Nat. Genet.* **48**, 709-17 (2016).

Figure 1. Flow chart illustrating a condensed version of the analysis protocol in *Claes et al*.