Journal of
Genetics and Genome Research

Identification of the Single Nucleotide Polymorphisms Affecting Normal Phenotypic Variability in Human Craniofacial Morphology Using Candidate Gene Approach

Mark Barash[1,2]*, Philipp E Bayer[3,4] and Angela van Daal[2]

[1]School of Mathematical and Physical Sciences, Centre for Forensic Sciences, Faculty of Science, University of Technology Sydney, Sydney, NSW, Australia

[2]Faculty of Health Sciences and Medicine, Bond University, Gold Coast, QLD, Australia

[3]School of Agriculture and Food Sciences, The University of Queensland, Brisbane, QLD, Australia

[4]Australian Centre for Plant Functional Genomics, The University of Queensland, Brisbane, QLD, Australia

*Corresponding author: *Mark Barash, Centre for Forensic Sciences, School of Mathematical and Physical Sciences, Faculty of Science, University of Technology Sydney, Building 4, Thomas St, Broadway NSW 2007, Sydney, Australia*

Abstract

There is a remarkable variety of human facial appearances, almost exclusively the result of genetic differences, as exemplified by the striking resemblance of identical twins. Despite intensive research on the genetics of craniofacial morphology using animal models and human craniofacial syndromes, the genetic variation that underpins normal human facial appearance is still largely elusive.

As a part of efforts on detecting genomic variants affecting normal craniofacial appearance, we have implemented a targeted candidate gene approach by selecting 1,319 single nucleotide polymorphisms (SNPs) in over 170 candidate genes and intergenic regions. This list has been further expanded with additional 4,732 tag polymorphisms, representing extended haplotype. All the markers were genotyped in 587 DNA samples using a massively parallel sequencing approach. We used 3-dimentional (3D) facial scans and direct cranial measurements to calculate 104 craniofacial anthropometric distances, which were analysed for associations with 2,332 polymorphisms.

An application of a Bonferroni - corrected genome - wide significance threshold produced associations between six SNPs and five craniofacial traits. Specifically, significant associations of nasal width with rs8035124 (15q26.1), cephalic index with rs16830498 (2q23.3), nasal index with rs37369 (5q13.2), transverse nasal prominence angle with rs59037879 (10p11.23) and rs10512572 (17q24.3), and a composite trait represented by a principal component with rs37369 (5p13.2) and rs390345 (14q31.3) were observed.

Due to over-conservative nature of the Bonferroni correction, we also report the associations that reached the traditional genome-wide p-value threshold ($< 5.00E-08$) as suggestive. Based on the genome-wide significance threshold, 8 craniofacial phenotypes demonstrated significant and mostly novel associations with 33 intergenic and extragenic SNPs, potentially involved in gene regulation.

This study identified a large number of genetic variants associated with normal craniofacial morphology variation, including confirmation of the two previously reported genes. These results enhance our understanding of the craniofacial genetics affecting normal craniofacial appearance and will be of particular value for clinical diagnostics and forensic molecular phenotyping.

Keywords

SNPs, Craniofacial, Facial appearance, Embryogenetics, Forensic DNA phenotyping

List of Abbreviations

3D: 3-Dimentional; AIMs: Ancestry Informative Markers; ASW: African Ancestry in Southwest USA; BAM: Binary Alignment Map; BMI: Body Mass Index; CAU: Caucasian; CHB: Han Chinese in Beijing, China; ENGINES: ENtire Genome INterface for Exploring Snps; EVT: Externally Visible Characteristic; DVI: Disaster Victim Identification; FDP: Forensic DNA Phenotyping; GWAS: Genome Wide Association Studies; HWE: Hardy-Weinberg Equilibrium; JPT: Japanese in Tokyo, Japan; LD: Linkage Disequilibrium; lncRNAs: Long Non-coding RNAs; LINE-1: Long Interspersed Nuclear Element 1; MAF: Minor

Allele Frequency; ME: Measurement Error; MD: Mean Difference; OMIM: Online Mendelian Inheritance in Man; ORFs: Open Reading Frames; pfSNP: Potentially Functional SNP; PCA: Principal Component Analysis; RGB: Red, Green, Blue (colours); SNP: Single-nucleotide Polymorphism; SNAP: SNP Annotation and Proxy Search; STR: Short Tandem Repeat; TF: Transcription Factor; VCF: Variant Call Format; YRI: Yoruba in Ibadan, Nigeria

Introduction

The human face is probably the most commonly used descriptor of a person and has an extraordinary role in human evolution, social interactions, clinical applications as well as forensic investigations. The influence of genes on facial appearance can be seen in the striking resemblance of monozygotic twins as well as amongst first degree relatives, indicating high heritability [1,2].

Uncovering the genetic background for regulation of craniofacial morphology is not a trivial task. Human craniofacial development is a complex multistep process, involving numerous signalling cascades of factors that control neural crest development, followed by a number of epithelial-mesenchymal interactions that control outgrowth, patterning and skeletal differentiation [2]. The mechanisms involved in this process include various gene expression and protein translation patterns, which regulate cell migration, positioning and selective apoptosis, subsequently leading to development of specific facial prominences. These events are precisely timed and are under hormonal and metabolic control. Most facial features of the human embryo are recognizable from as early as 6 weeks post conception, developing rapidly *in utero* and continuing to develop during childhood and adolescence [3,4]. Development of the face and brain are interconnected and occur at the same time as limb formation. Facial malformations therefore, frequently occur with brain and limb abnormalities and vice versa. Consequently, the development of both craniofacial and limb segments is likely to be regulated by the same genes and epistatic interactions. Genetic regulation of craniofacial development involves several key morphogenic factors such as *HOX, WNT, BMP, FGF* as well as hundreds of other genes and intergenic regulatory regions, incorporating numerous polymorphisms [2].

The loci involved in the craniofacial syndromes may in fact influence the extraordinary variety of human facial appearances, in the same way that genes responsible for albinism have been shown to be involved in normal pigmentation phenotypes [5]. Additionally, non-genetic components such as nutrition, climate and socio-economic environment may also affect human facial morphology via epigenetic regulation of transcription, translation and other cellular mechanics. To date, both the genetic and even more so, the epigenetic regulation of the craniofacial morphology are poorly understood.

The genetic basis of craniofacial morphogenesis has been explored in numerous animal models with multiple loci shown to be involved [2]. The majority of human studies in this field have focused on the genetics of various craniofacial disorders such as craniosynostosis and cleft lip/palate [6,7], which may provide a link to regulation of normal variation of the craniofacial phenotype, as for example observed between cleft-affected offspring and the increase of facial width seen in non-affected parents [8]. These studies have identified several genes with numerous genetic variants that may contribute to normal variation of different facial features, such as cephalic index, bizygomatic distance and nasal area measurements [9-11]. Studies of other congenital disorders involving manifestation of craniofacial abnormalities such as Alagille syndrome (*JAG1* and *NOTCH2* gene mutations), Down syndrome (chromosome 21 trisomy - multiple genes), Floating-Harbor syndrome (*SRCAP* gene mutations) and Noonan syndrome (mutations in various genes such as *PTPN11* and *RAF1*) provide additional information on the candidate genes potentially involved in normal craniofacial development [12-17].

One of the main restrictions in performing genome - wide association studies in this field relies in requirement for a large number of DNA samples, accompanied with comprehensive facial phenotypes. In recent years, new digital technologies such as 3-Dimentional laser imaging have been used in numerous anthropometric studies. 3-D laser imaging allows accurate and rapid capture of facial morphology, providing a better alternative to traditional manual measurements of craniofacial distances [18-20]. The high-throughput genotyping technologies and digital methods for capturing facial morphology have been used in a number of recent studies that demonstrated a link between normal facial variation and specific genetic polymorphisms [21-23]. Despite these promising results, our current knowledge of craniofacial genetics is sparse.

This study aims to further define the polymorphisms associated with normal facial variation using a candidate gene approach. The advantage of a candidate gene approach over previous genome wide association studies (GWAS) is that it focuses on genes, which have previously been associated with craniofacial embryogenesis or inherited craniofacial syndromes, rather than screening hundreds of thousands of non-specific markers. This approach aims to increase the chances of finding significant associations between SNPs and visible traits and requires fewer samples for robust association analysis [24,25].

The current research aims to assist in uncovering the genetic basis of normal craniofacial morphology variation and enhance our understanding of craniofacial embryogenetics. These findings could be useful in building statistical models to predict facial appearance

from a forensic DNA sample where no suspect has been identified, thereby providing valuable investigative leads. It could also assist in identifying skeletal remains by allowing more accurate facial reconstructions.

Methods

Sample collection and ethics statement

A total of 623 unrelated individuals, mostly Bond University (Gold Coast, Australia) students, of Australian ancestry (mainly of European descent) were recruited. The participants provided their written informed consent to participate in this study, which was approved by the Bond University Ethics committee (RO-510). To minimize any age-related influences on facial morphology the samples were largely collected from volunteers aged between 18 and 40. The mean age of the volunteers was 26.6 (SD ± 8.9) years old. Samples from the individuals who had experienced severe facial injury and/or undergone facial surgery (e.g. nose or chin plastics) were ommited.

Each participant donated four buccal swabs (Isohelix, Cell Projects, Kent, UK). 3-Dimentional (3D) facial scans and three direct cranial measurements were obtained as described below. Samples with low DNA quantity or low quality facial scans were eliminated, leaving 587 DNA samples for subsequent DNA sequencing.

Additional information, such as: height, weight, age, sex, self-reported ancestry (based on the four grandparents' ancestry), eye lid (single or double), ear lobe (attached or detached), hair texture (straight, wavy, curly or very curly), freckling (none, light, medium or extensive), moles (none, few or many), as well as iris, skin, and hair pigmentation was collected by a single examiner in order to reduce potential variation. The pigmentation traits were arbitrary assigned according to colour charts [26-28].
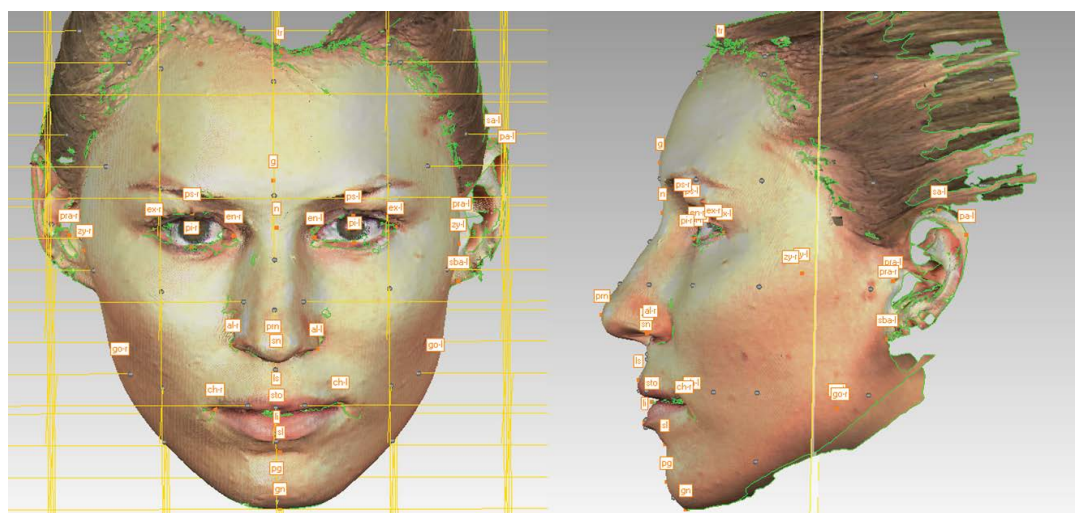
3D images collection and analysis

Craniofacial scans were obtained using the Vivid 910 3D digitiser (Konica Minolta, Australia) equipped with a medium range lens with a focal length of 14.5 mm. The scanner output images were of 640 × 480 pixels resolution for 3D and RGB data. Two daylight fluorescent sources (3400K/5400K colour temperature) were mounted at approximately 1.2 meters from the subject's head to produce ambient light conditions.

The scanner was mounted approximately one meter from the volunteer's head. Each volunteer remained in an upright seated position and was asked to keep neutral facial expression during the scan. Subjects with long hair pulled their hair behind the ears or were asked to wear a hair net. Glasses and facial jewelery were removed.

Each volunteer was scanned from a distance of approximately one meter from three different angles (front and two sides). The final merged 3D image was produced by semi-automatically aligning the three scans and manually cropping non-overlapping or superfluous data such as the neck area and hair using Polygone® software (Qubic, Australia). The complete coordinates of each merged 3D image were then saved in a 'vivid' file format (vvd) and exported to Geomagic® software (Qubic, Australia) for subsequent image processing.
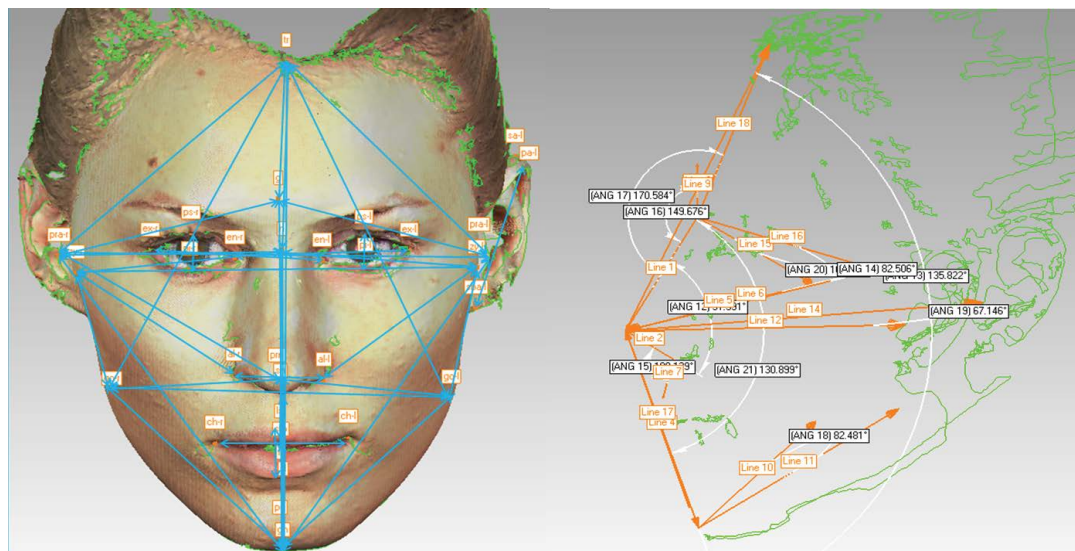
Based on the anthropometrical literature [29] 32 anthropometrical landmarks were manually identified on each 3D image using the Geomagic software (Figure

Figure 1: Anatomical position of the 32 manually annotated anthropometric landmarks used for calculation of linear and angular distances and ratios between the linear distances. Some landmarks are not clearly visible due to image orientation.

gn = Gnathion, pg= Pogonion, sl = Sublabiale, li = Labiale Inferius, sto = Stomion, ls = Labiale superius, ch-r = Right Chelion, ch-l = Left Chelion, go-r = Gonion Right, go-l = Left Gonion, sn = Subnasale, prn= Pronasale, al-r = Right Alare; al-l= Left Alare, n = Nasion (in fact this is Sellion), g= Glabella; tr = Tragion, en-l = Left Endocanthion, en-r = Right Endocanthion, ex-r = Right Exocanthion; ex-l = Left Exocanthion, ps-r = Right Palpebrale superius, ps-l = Left Palpebrale superius, pi-r = Right Palpebrale inferius, pi-l = Left Palpebrale inferius, zy-r = Right Zygion Right, zy-l = Left Zygion, pra-r = Right Tragion, pra-l = Left Tragion, sba-l = Left Subalare, sa-l = Left Superaurale, pa-l = Left Postaurale.

Figure 2: Illustration of linear and angular distances calculated from manually annotated landmark coordinates.

1 and Supplemental Table S1). Each landmark was represented by 'x', 'y' and 'z' coordinates as part of the Cartesian coordinate system. The Cartesian coordinates were used for subsequent calculation of 86 Euclidean distances, including 54 linear distances, 10 angular distances and 21 indices (ratios) between the linear distances (Figure 2 and Table 1).

Additionally, three direct cranial measurements: maximum cranial breadth (Euryon – Euryon), maximum cranial length (Gonion – Opisthocranium) and maximum cranial height (Vertex – Gnathion), were collected manually using a digital spreading calliper (Paleo-Tech Concepts, USA). Based on the craniofacial and body height measurements, three craniofacial ratios were calculated: Cephalic index: (eu-eu)/(g-op), Head width – Craniofacial height index: (eu-eu)/(v-gn) and Head – Body height index: (v-gn)/(body height), as summarised in Table 1. The linear and angular facial distances were used to calculate 10 principal components (PCs). Additionally, the details on the eyelid and earlobe morphology (single/double and attached/detached respectively) were recorded. In total, 104 craniofacial traits for each individual were recorded and used for genetic association analyses. All the facial measurements were recorded by a single analyst to reduce potential variability.

Candidate genes and SNPs selection

Two complementary strategies were used to generate a preliminary list of candidate genes and genetic markers. The first focused on searching the literature and web resources for candidate genes involved either in normal craniofacial variation or in craniofacial malformations in humans and model organisms (Supplemental Table S2).

The search for candidate genes focused not only on specifically defined craniofacial disorders, but also on genetic syndromes with various manifestations of craniofacial malformations, such as Down syndrome, Noonan Syndrome, Floating-Harbor Syndrome and others, as detailed in the Supplemental Table S2. The main resources for locating candidate genes in the animal models were Mouse Genome Informatics [30] and AmiGo databases [31] The main resources for identifying candidate genes in the human genome were OMIM [32] and GeneCards websites [33]. The online search was performed using the keywords such as: "craniofacial", "craniofacial mutants", "craniofacial anomalies", "craniofacial dimorphism" and "facial morphology" (as summarized in the Supplemental Appendix S1).

The second approach initially implemented a broad search for high Fst SNPs, such as ancestry informative markers (AIMs), with the rationale that many genes affecting craniofacial traits would express significantly different allele frequencies across populations, following non-neutral evolutionary pressure. AIMs were selected from a variety of published and online resources [34-43].

The relevant genes obtained by both approaches were subsequently analysed for potential involvement in the craniofacial embryogenesis, limb development and bilateral body symmetry. It should be noted however, that the final candidate gene list was not limited to craniofacial genes and included high Fst SNPs in genes with unknown function as well as markers located in intergenic regions, potentially possessing regulatory functions.

The resulting set of SNPs was further screened for high Fst SNPs (≥ 0.45) in three '1000 genomes' populations (CAU, ASW, CHB) using ENGINES browser [44] as well as potentially functional polymorphisms, such as non-synonymous SNPs [45], markers in transcription factor binding sites [46] and splicing sites [47] using various web resources, as detailed in the Supplemental Appendix S1 and reviewed elsewhere [48]. The candidate markers

Table 1: Craniofacial anthropometric measurements recorded in the study and used for genetic association analyses.

Manual (direct) craniofacial measurements
• V - Gn (Maximum Craniofacial height)
• Eu - Eu (Maximum Head Width)
• G - Op (Maximum Head Length)
• Cephalic index: (eu - eu)/(g - op)
• Head width – Craniofacial height index: (eu - eu)/(v - gn)
• Head – Body height index: (v - gn)/(body height)

3D (indirect) facial measurements
Linear facial distances
• Total face height: tr - gn
• Face width: zy - zy
• Morphological face height: n - gn
• Physiognomical face height: n - sto
• Lower profile height: prn - gn
• Lower face height: sn - gn
• Lower third face depth: t(l) - gn
• Middle face depth: t(l) - prn
• Middle face height (right): go(r) - zy(r)
• Middle face height (left): go(l) - zy(l)
• Middle face width 1: t(r) - t(l)
• Middle face width 2 (left): zy(l) - al(l)
• Middle face width 2 (right): zy(r) - al(r)
• Upper face depth: (left): t(l) - tr
• Upper face depth: (right): t(r) - tr
• Upper third face depth: t(l) - n
• Forehead height: g - tr
• Extended forehead height: tr - n
• Glabella – Gnathion distance: g - gn
• Supraorbital depth: t(l) - g
• Trichion – Zygion distance (left): tr - zy(l)
• Trichion – Zygion distance (right): tr - zy(r)
• Nasion – Zygion distance (left): n - zy(l)
• Nasion – Zygion distance (right): n - zy(r)
• Zygion – Gnathion distance (left): zy(l) - gn
• Zygion – Gnathion distance (right): zy(r) - gn
• Interendocanthal width: en - en
• Interexocanthal width: ex - ex
• Eye fissure width (left): en(l) - ex(l)
• Eye fissure width (right): en(r) - ex(r)
• Eye fissure height (left): ps(l) - pi(l)
• Eye fissure height (right): ps(r) - pi(r)
• Ear height (left): sa(l) - sba(l)
• Ear width (left): t(l) - pa(l)
• Nasal bridge length: n - prn
• Nose height: n - sn
• Nose width: al - al
• Nasal tip protrusion: sn - prn
• Ala length (left): prn - al(l)
• Ala length (right): prn - al(r)
• Gonion – Trichion distance (left): go(l) - tr
• Gonion – Trichion distance (right): go(r) - tr
• Gonion – Glabella distance: g - pg
• Pronasale – Gonion distance (left): prn - go(l)

- Pronasale – Gonion distance (right): prn - go(r)
- Chin height: sl - gn
- Mandibular region depth (right): t(r) - gn
- Mandible width: go - go
- Mandible height: sto - gn
- Lower jaw depth (left): gn - go(l)
- Lower jaw depth (right): gn - go(r)
- Mouth width: ch - ch
- Upper vermilion height: ls - sto
- Lower vermilion height: li - sto

Angular facial distances

- Nasal tip angle: (n - prn - sn)
- Nasal vertical prominence angle: (tr - prn - gn)
- Transverse nasal prominence angle 1: (zy(l) - prn - zy(r))
- Transverse nasal prominence angle 2: (t(l) - prn - t(r))
- Nasolabial angle: (prn - sn - ls)
- Nasofrontal angle: (g - n - prn)
- Nasion depth angle: (zy(l) - n - zy(r))
- Nasomental angle: (n - prn - pg)
- Forehead nasal angle: (tr - n - prn)
- Chin prominence angle: (go(l) - gn - go(r))

Ratios (indices)

- Forehead height ratio: (tr - n)/(go(r) - go(l))
- Upper face height ratio: (n - sn)/(go(r) - go(l))
- Lower face height ratio: (sn - gnx)/(go - go)
- Anterior face height 1 ratio: (n - gn)/(go - go)
- Anterior face height 2 ratio: (n - gn)/(zy - zy)
- Face height index: (n - gn)/(tr - gn)
- Upper – Lower face ratio: (tr - g)/(sn - gn)
- Upper face height ratio: (n - sn)/(sn - gn)
- Upper face width ratio: (n - sn)/(zy - zy)
- Total anterior face height ratio: (tr - gn)/(zy - zy)
- Mouth width ratio: (ch - ch) × 100/(en - en)
- Mandible – Face width ratio: (go - go)/(zy - zy)
- Mandible index: (sto - gn) × 100/(go - go)
- Mandible – Interexocanthion distance ratio (go - go)/(ex - ex)
- Interendocanthion distance ratio: (en - en)/(al - al)
- Intercanthal index: (en(r) - en(L))/(ex(r) - ex(l)
- Intercanthal – Intracanthal index: (ex(r) - en(r))/(en(l) - ex(l))
- Nasal index: (al - al) × 100/(n - sn)
- Nose - face height index: (n - sn) /(n - gn)
- Nose - face width index: (al - al)/(zy - zy)
- Nasal tip protrusion – nose width index: (sn - prn)/(al - al)
- Nasal tip protrusion –nose height index: (sn - prn)/(n - sn)

search resulted in identification of 1,319 SNPs, located in 177 genes and intergenic regions, as discussed in the Results section. In addition to the craniofacial candidate markers, 464 pigmentation and ancestry - informative polymorphisms were selected from the literature (Supplemental Appendix S1).

The chromosomal locations of final candidate markers (n = 1,783) were submitted to the custom Ampliseq primer design pipeline (Life Technologies), according to manufacturer recommendations. Following primer design failure for 881 markers, the marker list was redesigned to include alternative SNPs in high linkage disequilibrium with the original markers, resulting in 1,670 candidate polymorphisms. The final custom Ampliseq panel was manufactured as two separate pools of 849 and 847 primer pairs, with each amplicon length between 125 bp and 225 bp, covering 15.78 kb of the human genome. This panel comprised of 902 originally targeted candidate markers and 4,732 alternative tag markers, as a replacement for candidate SNPs that failed primer design pipeline. Subsequent genomic annotation of the total amplicon sequences for SNPs with

MAF ≤ 1% resulted in 8,518 markers as detailed in Supplemental Table S1 (Supplemental Appendix S2).

DNA extraction and quantification

DNA was purified from buccal swabs using the Isohelix DDK isolation kit (Cell Projects, Kent, UK) according to the manufacturer instructions. DNA samples were quantified using a Real Time quantitative PCR (q-PCR) method using a Bio-Rad CFX96 (Bio-Rad, Gladesville, Australia). This assay amplified a 63bp region of the OCA locus. The primer sequences were 5'-GCTGCAG-GAGTCAGAAGGTT-3' (forward primer) and 5'- CATTTG-GCGAGCAGAATCC-3' (reverse primer) at a final concentration of 200 mM. All DNA samples were additionally quantified using Qubit 2.0 fluorimeter (Invitrogen) prior to library construction as per manufacturer recommendations.

SNP genotyping and data analysis

Multiple DNA libraries were constructed from sets of 32 Ion Xpress™ (Life Technologies) barcoded samples using the Ion AmpliSeq™ library Kit 2.0 (Life Technologies) in conjunction with two custom primer mixes that were pooled according to manufacturer recommendations. Libraries were quantified using the Ion Library Quantitation kit (Life Technologies) and pooled in equal amounts for emulsion PCR, which was performed using the OneTouch™ 2 instrument (Life Technologies) according to manufacturer recommendations. All DNA samples (n = 587) were genotyped by massively parallel sequencing on the Personal Genome Machine (PGM) (Life Technologies) using the Sequencing 200 v2 kit and 316 Ion chips (Life Technologies).

Raw sequencing data were collected and processed on the Torrent Suite Server v3.6.2 using default settings. Alignment and variant calling were performed against the human genome reference (hg19) sequence at low stringency settings. Binary alignment map (BAM) files were generated and exported into Ion Reporter™ (IR) cloud-based software for SNP annotation along with the reference hotspot file. The IR analysis resulted in generation of the individual variant caller files (VCF) with genotype calls for each sample as well as sequencing performance statistics.

To reduce potential bias of the self-reported ancestry, ancestry inferences were obtained by 3,302 markers using STRUCTURE version 2.3.4 with default parameters as per software developer recommendations [49]. SNPs in long-range Linkage Disequilibrium (> 100,000 bp) were excluded from the STRUCTURE run. The ancestry was estimated based on four predefined population clusters: Europeans, East Asians, South Asians and Africans, according to software developer recommendations. Relative allele calls for four predefined HapMap population clusters (CEU, YRI, CHB and JPT) were used as reference populations [50]. The ancestry origin was estimated as a single (unmixed) source where the main ancestry cluster could be affiliated with at least 80% of the total mixed ancestry. The samples with mixed ancestry (> 20% admixture) were assigned to an 'Admixture' cluster.

Association analyses were performed using SNP & Variation Suite v7 (SVS) (Golden Helix, Inc., Bozeman, MT) and replicated using PLINK v1.07 software [51]. Statistical analyses in both software programs were performed using linear regression with quantitative phenotypes, and logistic regressions with binary phenotypes under the assumption of an additive genetic model, while each genotype was numerically encoded as 0, 1 or 2. Population stratification correction, incorporated by EIGENSTRAT program was implemented in the analyses [52,53]. In order to reduce any potential confounding effects, all the craniofacial traits association analyses were performed using sex, BMI and EIGENSTRAT ancestry clusters as covariates. In PLINK, p-values were adjusted using the '–adjust' option. The final reported association results are based on the PLINK statistical analyses with the EIGENSTRAT PCA clusters, BMI and gender as covariates.

Annotation analysis of the significantly associated genes was performed using GeneCards, ENTREZ and UniProtKB web portals [33,54]. The MalaCards web site was used to detect association between the genes and hereditary syndromes [55]. The GeneMania web site was used to identify a functional network among the genes and encoded proteins [56]. Gene ontology web resource was used to find orthologs of human genes in other organisms [31,57]. The MGI database was used to search for the phenotype in relevant craniofacial mouse gene mutants [30]. The dbSNP, 1000 genomes, SNPnexus and Alfred websites were used for SNP annotations [58-61].

The SNP Annotation and Proxy Search (SNAP) web portal was used to find SNPs in linkage disequilibrium (LD) and generate LD plots, based on the CEU population panel from the 1000 genomes data set, within a distance of up to 500 kb and an r^2 threshold of 0.8 [62].

The Regulome database and potentially functional database (PFS) searches were implemented to annotate SNPs with known and predicted regulatory elements in the intergenic regions of the *H. sapiens* genome [47,63].

Results

Candidate genes search and sequencing data filtering

The online database search for candidate SNPs revealed a list of 150 genes with 310 annotations. In parallel, a search for high Fst markers, resulted in identification of additional 938 targets, for a total of 1,088 genes and intergenic regions (a detailed list of online resources is summarized in the Supplemental Appendix S1).

Manual examination of the candidates list revealed no apparent link with normal craniofacial development

or malformations for 592 of these genes and these were therefore excluded. The remaining 496 regions were further screened for non-synonymous and potentially functional SNPs, as well as SNPs with high population differentiation, which resulted in a shortlist of 269 genes and intergenic regions.

Subsequent analysis of these 269 genes/regions for functional annotation, resulted in 177 candidate genes/intergenic regions, possessing 1,319 genetic polymorphisms potentially involved in various stages of human embryonic development, including: embryonic morphogenesis, sensory organ development, tissue development, pattern specification process, tissue morphogenesis, ear development, neural tube morphogenesis, epithelium development, chordate embryonic development and morphogenesis of an epithelium (Supplemental Appendix S1). Notably, the majority of these markers are located in introns and intergenic regions, including 5' and 3' UTRs. This outcome is quite expected, based on the initial assumption that most of the genetic variants influencing phenotypic traits lie in the noncoding part of the human genome, as demonstrated by numerous genome – wide association studies [64]. Such sequence polymorphisms are mostly rare (MAF < 0.1) and have tiny effect size, while thousands of them required to jointly affect complex traits, such as facial morphology or body height [65-68].

In terms of molecular function, AmiGO showed that craniofacial candidate markers might be involved in a range of regulatory activities including: protein dimerization activity, chromatin binding, regulatory region DNA binding, sequence-specific DNA binding RNA polymerase II transcription factor activity, sequence-specific distal enhancer binding activity, heparin binding, RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription, BMP receptor binding and transmembrane receptor protein serine/threonine kinase binding (Supplemental Appendix S1). The final candidate marker list annotation revealed that almost 99% of the genomic regions (which may cover multiple markers) are associated with one or two genes with 62% of genomic regions located up to 500 kb downstream of a transcription start site (data not shown). This is again an anticipated outcome, given that non-coding variants are likely to reside within cis-regulatory genomic elements, such as enhancers [69,70].

Subsequent analysis of candidate SNPs for mouse phenotype associations confirmed that orthologous candidate markers were previously detected in mouse models displaying abnormal morphology of the skeleton, head, viscerocranium and facial area, as well as specific malformations of the eye, ear, jaw, palate, limbs, digits and tail (data not shown).

Targeted massively parallel sequencing of the 587

samples resulted in 8,518 genetic markers, with the majority of markers (> 5,000) represented by rare polymorphisms with MAF ≤ 1% (data not shown). This outcome was a result of identification of novel and rare markers in individual DNA samples. Three of the DNA samples did not produce high quality genotypes because of poor DNA quality or unsuccessful library and template preparation.

Data quality control steps were performed by removing markers of low genotype quality (GQ > 10) and sequencing depth (DP > 10X). Further filtering of markers using a 2% MAF cut-off resulted in 3,075 markers (Supplemental Appendix S2). Additional filtering based on the HWE threshold of p-value ≥ 0.01 resulted in 2,332 markers. The mean sequencing depth for significantly associated markers was 58 fold (± 48.9 SD).

Genetic association analysis

The association analyses were performed using a linear regression model, incorporating EIGENSTRAT-generated PCAs as well as sex and BMI as covariates. The use of covariates in the statistical analysis aims to reduce the risk of introducing confounding effects, which can result in false positive associations. While sexual dimorphism in the craniofacial morphology is well-known [71], BMI is likely affect certain craniofacial traits, since the soft facial tissue may change significantly with weight gain or loss. Notably, this potential confounding factor has to date been disregarded in association studies of normal craniofacial morphology. Age was not considered a significant covariate, given that average age of the subjects in this study was 27 (± 8.9 SD). Nevertheless, the potential effect of age as a cofactor was assessed on three craniofacial traits and found not significant (data not shown).

While the majority of current GWA studies rely on a p-value < 5.00E-08 significance threshold, some publications suggest this threshold may be too stringent, especially for complex traits that are regulated by a large number of small effect alleles [72,73]. In contrast to GWAS, candidate gene studies undertake a more focused genetic strategy, concentrating on a relatively limited number of putative markers. As this study analysed a significantly lower number of SNPs than usual GWA-studies, we could use a higher p-value cut-off, since the smaller sample size means the probability of false positive at extremely low p-values is itself lower. Nevertheless, we decided to keep the traditional GWAS p-value significance threshold (< 5.00E - 08) in order to reduce the possibility of detecting false positive results.

In addition, we subsequently applied a more stringent Bonferroni – corrected threshold in order to minimize the chance of detecting spurious associations. Following the association analysis of 104 craniofacial phenotypes with 2,332 genetic markers, the significance threshold

Barash et al. J Genet Genome Res 2018, 5:041

• Page 8 of 17 •

Table 2: Results of genetic association analyses between candidate SNPs and craniofacial traits, including all genomic markers reached the unadjusted p - value threshold of < 5.00E - 08.

gene/intergenic region	rs#	chromosomal location	observed alleles	MAF	genomic annotation	UNADJ	BONF	HOLM	BETA	SE
nasal width (al - al)										
15q26.1	**rs8035124**	**15:92105708**	**A/C**	**3.08E - 01**	**intergenic**	**1.52E - 10**	**1.74E - 07**	**1.74E - 07**	**1.37E + 00**	**2.09E - 01**
EYA2	rs58733120	20:45803852	C/G	2.29E - 02	intronic	5.37E - 10	6.15E - 07	6.14E - 07	4.98E + 00	7.86E - 01
11q15.4	rs1482795	11:7850345	C/T	1.71E - 01	intergenic	7.68E - 10	8.78E - 07	8.77E - 07	1.56E + 00	2.48E - 01
AGXT2	rs37369	5:35037115	C/T	1.77E - 01	missense	1.04E - 09	1.19E - 06	1.19E - 06	1.51E + 00	2.43E - 01
9q22.32 (downstream to PTCH1)	rs57585041	9:98205221	G/T	2.95E - 02	intergenic	6.05E - 09	6.92E - 06	6.90E - 06	3.63E + 00	6.13E - 01
EYA1	rs79867447	8:72127562	C/T	2.19E - 02	intronic	3.92E - 08	4.49E - 05	4.47E - 05	4.89E + 00	8.73E - 01
nasal tip protrusion (sn - prn)										
17q24.3	rs10512572	17:69512099	A/G	1.67E - 01	intergenic	2.22E - 08	2.54E - 05	2.54E - 05	- 9.43E - 01	1.66E - 01
cephalic index										
CACNB4	**rs16830498**	**2:152814028**	**C/T**	**9.06E - 02**	**intronic**	**7.57E - 11**	**8.67E - 08**	**8.67E - 08**	**3.01E + 00**	**4.52E - 01**
MYO5A	rs2290332	15:52611451	A/G	2.19E - 01	synonymous	5.56E - 10	6.37E - 07	6.37E - 07	1.99E + 00	3.15E - 01
ZEB1	rs59037879	10:31745993	A/T	2.49E - 02	intronic	6.27E - 10	7.18E - 07	7.17E - 07	6.24E + 00	9.85E - 01
COL11A1	rs4908280	1:103420759	G/T	3.14E - 01	intronic	1.66E - 09	1.91E - 06	1.90E - 06	- 1.70E + 00	2.77E - 01
EYA1	rs1481800	8:72131426	A/G	3.62E - 01	intronic	2.07E - 09	2.37E - 06	2.36E - 06	1.66E + 00	2.72E - 01
TEX41	rs10496971	2:145769943	G/T	1.87E - 01	intronic	5.32E - 09	6.09E - 06	6.06E - 06	1.99E + 00	3.36E - 01
PCDH15	rs10825273	10:55968685	C/T	2.82E - 01	intronic	9.93E - 09	1.14E - 05	1.13E - 05	1.71E + 00	2.94E - 01
COL11A1	rs11164649	1:103444679	G/T	3.15E - 01	intronic	1.70E - 08	1.95E - 05	1.94E - 05	- 1.63E + 00	2.85E - 01
5q14.3	rs373272	5:84818656	A/G	4.22E - 01	intergenic	2.40E - 08	2.75E - 05	2.73E - 05	1.53E + 00	2.69E - 01
nasal index (al - al/n - sn)										
AGXT2	**rs37369**	**5:35037115**	**C/T**	**1.77E - 01**	**missense**	**1.25E - 10**	**1.43E - 07**	**1.43E - 07**	**4.03E + 00**	**6.12E - 01**
EYA2	rs58733120	20:45803852	C/G	2.29E - 02	intronic	9.46E - 09	1.08E - 05	1.08E - 05	1.18E + 01	2.03E + 00
RP11 - 408B11.2	rs7311798	12:85808703	C/T	9.71E - 02	intergenic	1.77E - 08	2.02E - 05	2.02E - 05	5.00E + 00	8.73E - 01
11q15.4	rs1482795	11:7850345	C/T	1.71E - 01	intergenic	1.83E - 08	2.09E - 05	2.09E - 05	3.66E + 00	6.40E - 01
EYA1	rs79867447	8:72127562	C/T	2.19E - 02	intronic	3.53E - 08	4.03E - 05	4.02E - 05	1.26E + 01	2.24E + 00
nose - face width index (al - al/zy - zy)										
EYA1	rs79867447	8:72127562	C/T	2.19E - 02	intronic	1.10E - 09	1.26E - 06	1.26E - 06	3.75E + 00	6.02E - 01
EYA2	rs58733120	20:45803852	C/G	2.29E - 02	intronic	3.38E - 09	3.86E - 06	3.86E - 06	3.25E + 00	5.39E - 01
EGFR	rs17335905	7:55131384	C/T	3.30E - 02	intronic	4.74E - 08	5.42E - 05	5.41E - 05	2.26E + 00	4.07E - 01
nasolabial angle (prn - sn - ls)										
SMAD1	rs17020235	4:146418167	A/G	3.59E - 02	intronic	2.07E - 09	2.36E - 06	2.36E - 06	- 1.15E + 01	1.87E + 00
transverse nasal prominence angle (t - l) - prn - (t - r)										
ZEB1	**rs59037879**	**10:31745993**	**A/T**	**2.49E - 02**	**intronic**	**5.31E - 12**	**6.07E - 09**	**6.07E - 09**	**4.77E + 00**	**6.69E - 01**
17q24.3	**rs10512572**	**17:69512099**	**A/G**	**1.67E - 01**	**intergenic**	**1.38E - 11**	**1.57E - 08**	**1.57E - 08**	**1.51E + 00**	**2.17E - 01**

Barash et al. J Genet Genome Res 2018, 5:041

• Page 9 of 17 •

Gene/locus	rs#	Chromosomal location	Observed alleles	MAF	Annotation	UNADJ	BONF	HOLM	BETA	SE
AGXT2	rs37369	5:35037115	C/T	1.77E-01	missense	1.46E-09	1.66E-06	1.66E-06	1.31E+00	2.12E-01
LMNA	rs12076700	1:156055099	C/G	2.28E-01	intronic	1.54E-09	1.75E-06	1.75E-06	1.18E+00	1.92E-01
FAM49A	rs6741412	2:16815759	C/G	3.99E-01	intronic	2.75E-09	3.14E-06	3.13E-06	9.96E-01	1.64E-01
TEX41	rs10496971	2:145769943	G/T	1.87E-01	intronic	5.52E-09	6.30E-06	6.27E-06	1.27E+00	2.14E-01
RTTN	rs74884233	18:67813813	A/G	2.59E-02	intronic	1.20E-08	1.37E-05	1.37E-05	3.07E+00	5.30E-01
AC073218.1	rs892458	2:34667749	C/T	4.97E-01	intergenic	1.73E-08	1.98E-05	1.97E-05	9.61E-01	1.68E-01
PAX3	rs2289266	2:223089431	G/T	1.23E-01	intronic	1.95E-08	2.23E-05	2.21E-05	1.58E+00	2.76E-01
LHX8	rs12041465	1:75609049	A/C	2.37E-01	intronic	2.30E-08	2.62E-05	2.60E-05	1.21E+00	2.12E-01
AC073218.1	rs892457	2:34667721	G/A	4.98E-01	intergenic	3.43E-08	3.92E-05	3.89E-05	9.38E-01	1.67E-01
14q22.1 (upstream to BMP4)	rs2357442	14:52607967	A/C	2.03E-01	intergenic	4.40E-08	5.03E-05	4.98E-05	1.11E+00	1.99E-01
PC1 (EV = 1391.99)										
AGXT2	**rs37369**	**5:35037115**	**A/G**	**1.77E-01**	**missense**	**2.49E-11**	**2.85E-08**	**2.85E-08**	**-2.10E-02**	**3.08E-03**
FOXN3	**rs390345**	**14:89976534**	**A/G**	**2.47E-01**	**intronic**	**7.46E-11**	**8.55E-08**	**8.54E-08**	**-1.84E-02**	**2.77E-03**
FAM49A	rs6741412	2:16815759	G/A	3.99E-01	intronic	4.67E-10	5.35E-07	5.34E-07	-1.52E-02	2.40E-03
17q24.3	rs10512572	17:69512099	G/A	1.67E-01	intergenic	4.99E-10	5.71E-07	5.70E-07	-2.01E-02	3.17E-03
EYA1	rs79867447	8:72127562	C/T	2.19E-02	intronic	7.46E-10	8.54E-07	8.51E-07	-6.45E-02	1.03E-02
LMNA	rs12076700	1:156055099	C/G	2.28E-01	intronic	7.87E-10	9.01E-07	8.97E-07	-1.73E-02	2.76E-03
PCDH15	rs10825273	10:55968685	C/T	2.82E-01	intronic	1.04E-09	1.19E-06	1.19E-06	-1.68E-02	2.70E-03
14q22.1 (upstream to BMP4)	rs942316	14:54440983	A/C	1.21E-01	intergenic	2.66E-09	3.04E-06	3.02E-06	-2.36E-02	3.89E-03
TEX41	rs10496971	2:145769943	T/G	1.87E-01	intronic	3.71E-09	4.25E-06	4.22E-06	-1.84E-02	3.06E-03
RP11-785H20.1	rs7844723	8:122908503	C/T	4.37E-01	intergenic	8.11E-09	9.29E-06	9.21E-06	1.41E-02	2.41E-03
EYA2	rs58733120	20:45803852	G/C	2.29E-02	intronic	2.19E-08	2.51E-05	2.49E-05	-5.65E-02	9.94E-03
FAM49A	rs11096686	2:16815892	T/C	2.50E-01	intronic	2.25E-08	2.57E-05	2.55E-05	-1.65E-02	2.89E-03
EYA1	rs73684719	8:72131359	G/A	2.59E-02	intronic	2.62E-08	3.00E-05	2.96E-05	-4.94E-02	8.75E-03
XXYLT1	rs950257	3:194847650	T/A	3.87E-01	intronic	3.94E-08	4.51E-05	4.46E-05	-1.38E-02	2.48E-03

Highlighted in **bold**: genomic markers that reached Bonferroni corrected threshold (1.6E-07); Highlighted with green colour: angular distances; Highlighted with blue: linear distances; Highlighted with red: craniofacial indices; Highlighted with violet: principal component; gene/intergenic region: gene name/locus; rs#: reference SNP ID number; chromosomal location of the marker based on the GRCh37/hg19; observed alleles: common alleles observed in human genome based on dbSNP build 147; MAF: minor allele frequency; genomic annotation: genomic location of the marker; UNADJ: Unadjusted p-values. BONF: Bonferroni single-step adjusted. HOLM: Holm (1979) step-down adjusted; BETA: minor allele effect size; SE: standard error.

Barash et al. J Genet Genome Res 2018, 5:041

• Page 10 of 17 •

based on the Bonferroni correction with a desired α of 0.05 would be 2.06E - 07 (= 0.05/(2,332*104)).

However, it should be emphasized that the Bonferroni correction is widely considered over-conservative, especially in the case of complex phenotypic traits with small individual effects of each allele. Considering that our results confirm the previously published findings, we believe the GWAS p-value threshold is conservative enough to avoid or at least significantly reduce potentially spurious associations. Following this rationale, we report all the variants, which met the unadjusted genome-wide association p-value threshold, as suggestive. We believe these findings are useful for the future studies focusing on genetics of normal craniofacial morphology.

The results of the association analyses of the craniofacial traits are summarized in Table 2 and Supplemental Figures S1-S19. In general, following the application of a stringent Bonferroni-corrected GWAS threshold (adjusted p-value < 1.6E - 07), we observed five craniofacial traits being associated with six genomic markers. Specifically, nasal width with rs8035124 (p-value 1.74E - 07, Beta = 1.366, SE = 0.209), cephalic index with rs16830498 (p-value 8.67E - 08, Beta = 3.005, SE = 0.4518), nasal index with rs37369 (p-value 1.43E - 07, Beta = 4.025, SE = 0.6124), transverse nasal prominence angle with rs59037879 (p-value 6.07E - 09, Beta = 4.765, SE = 0.6685) and rs10512572 (p-value 1.57E - 08, Beta = 1.505, SE = 0.2171), and principal component (EV = 1391.99) with rs37369 (p-value 2.85E - 08, Beta = -0.021, SE = 0.003079) and rs390345 (p-value 8.55E - 08, Beta = -0.0184, SE = 0.002768). The polymorphisms: rs16830498, rs59037879 and rs390345 are intronic variants in *CACNB4*, *ZEB1* and *FOXN3* respectively; rs37369 is a missense mutation in the *AGXT2* gene and rs8035124 and rs10512572 are intergenic variants in 15q12.2 and 17q21.33 chromosomal locations respectively.

However, given the over-conservative nature of the Bonferroni correction and also assuming that polygenic traits are likely to be dominated by numerous alleles with small causal effect we also report suggestive associations reaching the unadjusted 5.00E - 08 p-value threshold. In summary, two linear distances (nasal width and nasal tip protrusion), two angular distances (nasolabial angle and transverse nasal prominence angle), three indices (cephalic index, nasal index and nose-face width index) and one principal component revealed significant associations with 33 SNPs in 28 genes and intergenic regions (Table 2).

These factors can be arbitrarily divided into three main categories based on their cellular function: 1) Genes with known roles in the craniofacial morphogenesis and/or mutated in various hereditary syndromes displaying craniofacial abnormalities; 2) Genes or pseudo-genes without known function in the craniofacial

morphology regulation or previously uncharacterized genes; and 3) Non-protein coding genes, such as lncRNA class genes. There are also a number of significant variants that are located in the intergenic regions, with or without proximity to open reading frames (ORFs).

The majority of associated markers (n = 21) are located in 17 protein-coding genes and pseudo-genes such as *AGXT2, CACNB4, COL11A1, EGFR, EYA1, EYA2, FAM49A, FOXN3, LHX8, LMNA, MYO5A, PAX3, PCDH15, RTTN, SMAD1, XXYLT1 and ZEB1*. Five variants are present in RNA-coding (lncRNA) genes, which include *AC073218.1, RP11-494M8.4, RP11-408B11.2, RP11-785H20.1* and *TEX41*. The rest of the markers (n = 6) are found in the intergenic regions, near the following genes and pseudogenes: *BMP4, HAS2-AS1, LOC124685, LOC100131241, MRPS36P3, PTCH1, SLC25A5P2, SV2B* and *TRNAY16P*.

Functional annotation of the significantly – associated markers revealed that one variant represent a missense mutation (rs37369), one SNP is a synonymous transversion (rs2290332), while 20 markers are found in intronic sequences and 11 markers located in intergenic regions (Table 1). As stated above, the noncoding genomic location for most of the markers makes sense as they expected to have mild affect on a normal craniofacial phenotype than exonic polymorphisms. This conclusion however, could be biased as *a priori*, most candidate markers were selected in the non-coding regions (Supplemental appendix S1). As expected, the majority of significantly associated SNPs (n = 27) are found in the regulatory elements of the genome, such as in transcription factor (TF) binding sites, and represent potentially functional SNPs (pfSNPs). These variants may be involved in regulation of gene expression by "fine tuning" of the normal craniofacial phenotype as part of the enhancer/silencer mechanisms, as has been recently suggested [69].

The nasal area measurements, using either "n", "prn", "sn" or "al" landmarks, produced the majority of the total number of significant associations (6 out of 8). These measurements include nasal width (al-al), nasal tip protrusion (sn-prn), nasolabial angle (prn-sn-ls), transverse nasal prominence angle (t_l-prn-t_r), nasal index (al-al/n-sn), and nose face width index (al-al/zy-zy). The apparent overrepresentation of associations with the nasal area is likely a result of the easier allocation and consequent superior reproducibility of the nasal area landmark measurements on 3D images. It may also be a result of a focused selection of candidate genes from the JAX mice database resource, which concentrated on mutants with nasal abnormalities.

The analysis of the direct cranial measurements and their relative indices revealed significant associations for cephalic index (CI) with 9 SNPs (Table 2). The association analysis of the principal components (PC) representing all the craniofacial measurements,

revealed one principal component (explaining 73.3% of all the craniofacial phenotypes) that was associated with 14 genetic markers (Table 2).

In contrast to most other craniofacial association studies that focused on a specific homogeneous population group (mostly Europeans), this study included samples from several population groups, which enabled investigation of the genetic factors influencing normal craniofacial morphology in different ethnicities [74]. Self-reported ancestry however, cannot be considered fully reliable, as demonstrated previously [75,76]. In order to address this issue we assessed the self-reported ancestry using STRUCTURE with 186 SNPs removed due to long-range disequilibrium [49]. Following the rationale that the best ancestry estimates are obtained using a large number of random markers [77], we used all the available markers (after MAF ≤ 2% filtering) in STRUCTURE analysis. The STRUCTURE analysis resulted in clusters of 367 Europeans, 51 East Asians, 43 South Asians and 16 Africans, with 107 samples designated as admixed ancestry (Figure 3). Of the samples tested with STRUCTURE, 459 (89%) were assigned the same ancestry cluster (sole or mixed origin) as the self-reported information. Of the remaining 57 individuals, 39 were estimated as 'admixture' (based on up to 20% admixture threshold) and 18 were assigned a single ancestry, different to the self-reported ancestry (Figure 3).

The risk of detecting false positive results due to population stratification was carefully assessed and further reduced by applying an EIGENSTRAT correction. Specifically, EIGENSTRAT's smartpca.perl script was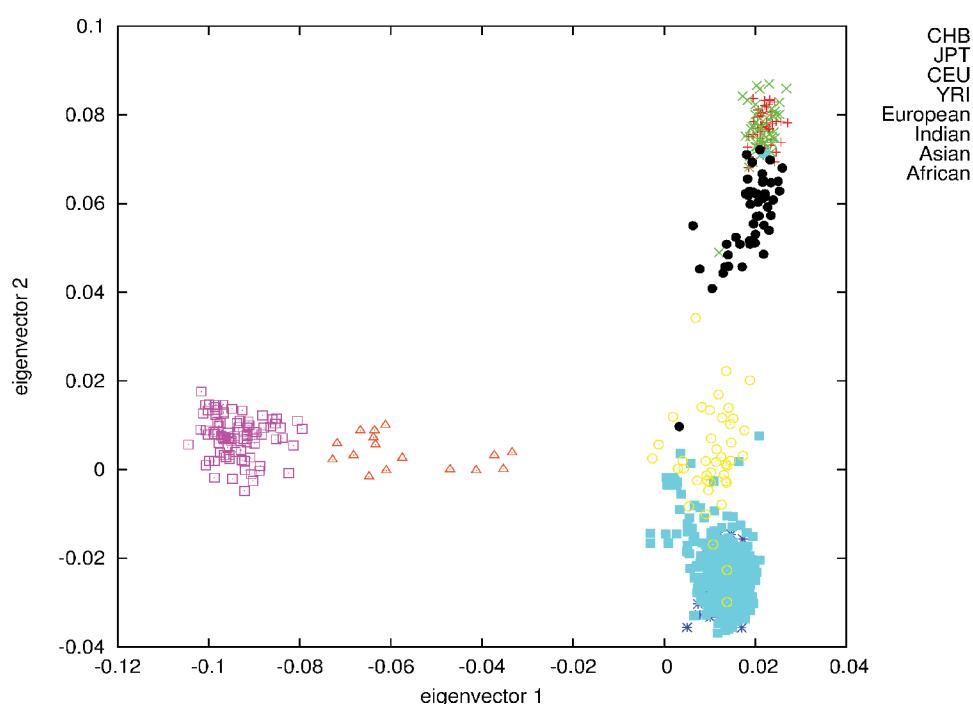 used to perform PCA-clustering in comparison to reference populations from HapMap reference clusters. The Q-Q plots of the associated traits confirmed the expected distribution of data after applied correction (Supplemental Figures S1-S8).

Due to a limited number of genotyped markers typed, genomic allele imputations were not performed.

Association analyses of the non-craniofacial traits

In our attempt to identify genetic markers influencing normal variation in craniofacial traits, we initially incorporated 464 markers previously associated with human pigmentation and ancestry – informative markers, while many of these overlap. These markers were included to validate the statistical methods used for the craniofacial traits association analysis.

The application of the Hardy-Weinberg equilibrium (HWE) threshold resulted in filtering 25% of the total number of SNPs. These markers included almost all the "key" SNPs, previously associated with pigmentation traits, such as rs12913832, rs1129038, rs8039195 and rs16891982. This is not surprising, since population-specific markers are likely not being in HWE 'a priori'. Another explanation for this observation is potential bias from partially uncorrected heterogeneous ancestry, since the ancestry correction algorithm can only minimize, rather than completely remove spurious associations [52]. In fact, the association analyses of the HWE non-filtered genotyping data with pigmentation traits (eye, skin and hair colour), demonstrated highly significant associations, fully concordant with the previously published sources (Supplemental Table S3). It should be noted however, that these results may not necessarily confirm the craniofacial markers associations.

Figure 3: Population structure as represented by plotting genomic PCs 1 and 2, using 270 HapMap individuals as anchor clusters.

Barash et al. J Genet Genome Res 2018, 5:041

• Page 12 of 17 •

Craniofacial genes and SNPs annotations

Due to space limitations, the functional annotation of the significantly associated SNPs and their potential role in regulating normal craniofacial phenotype has been provided in a Supplemental Appendix S3.

Discussion

This study focused on the identification of genetic markers in a set of candidate genes potentilly involved in normal craniofacial development. We report significant associations (Bonferroni adjusted p-value threshold of < 1.6E - 07) between 5 craniofacial traits (nasal width, cephalic index, nasal index, transverse nasal prominence angle and principal component) and 6 SNPs (rs8035124, rs16830498, rs37369, rs59037879, rs10512572 and rs390345) located in 6 genes and intergenic regions (15q26.1, 17q24.3, *CACNB4*, *AGXT2*, *ZEB1* and *FOXN3* respectively). We also report all the significant associations that met the less stringent GWAS threshold (unadjusted p-value < 5.00E - 08) between 8 craniofacial traits and 33 polymorphisms in 28 genes and intergenic regions, as Bonferroni correction is generally considered over-conservative, especially when analysing complex traits such as craniofacial morphology, which is likely to be influenced by a large number of alleles with relatively small individual effect, similar to body height [65-67,78-80]. In fact, most of the significant SNPs (31 out of 33) are noncoding variants, located in the regulatory regions such as: promoters, enhancers and splicing sites, as discussed below. These associations are suggestive and would require additional studies for potential confirmation.

The association of the *PAX3* gene and the *COL11A1* gene with the transverse nasal prominence angle and cephalic index respectively, confirms previous findings [11,22,23,80]. In fact, an intronic SNP rs11164649 that was associated with cephalic index in the current study, was recently associated with normal-range effects in various craniofacial traits and used for their prediction [80], while the other variants in *COL11A1* (rs4908280) and in *PAX3* (rs2289266) have not been reported previously. The rest of the significantly - associated markers are also novel and include 20 polymorphisms in protein-coding genes and pseudo-genes, such as *AGXT2*, *CACNB4*, *EGFR*, *EYA1*, *EYA2*, *FAM49A*, *FOXN3*, *LHX8*, *LMNA*, *MYO5A*, *PCDH15*, *RTTN*, *SMAD1*, *TEX41*, *XXYLT1* and *ZEB1*. Interestigly, at least 8 of these genes (EYA1, EYA2, FAM49A, FOXN3, LHX8, SMAD1, TEX41, ZEB1) represent transcription factors, affecting numerous signaling cascades during craniofacial embryonic development (discussed below and in more detail in the Supplemental Appendix S3). Additional 7 significantly–associated SNPs are found in intergenic regions adjacent to several loci, such as *BMP4*, *LOC124685*, *LOC100131241* and *PTCH1*. Some of these genes and chromosomal segments were previously linked to craniofacial embryogenesis, while others represent novel associations.

Six genetic variants were found in lncRNA genes, which have not been previously linked to craniofacial morphogenesis. These findings may suggest an yet unexplored level of epigenetic regulation affecting craniofacial morphology. lncRNAs represent a recently discovered class of regulatory factors, whose expression is thought to be important for regulation of gene expression through direct competition with transcription factors in promoter regions, as well as by indirectly affecting gene expression via interaction with miRNAs, histone modifying complexes and other cellular factors [81,82]. The role of these ncRNA molecules in the craniofacial development clearly requires additional research.

Another interesting example of potential epigenetic regulation is the gene EYA1, which encodes a histone phosphatase protein and transcriptional coactivator for SIX gene family. This factor regulates transcription during organogenesis of various craniofacial domains, such as branchial arches, eye and ear and known being mutated in Otofaciocervical Syndrome 1 and Branchiootorenal Syndrome 1 (discussed in detail in the Supplemental Appendix S3). Based on ENCODE and Regulome analysis, at least two of the significantly associated SNPs in the intronic region of this gene (rs1481800 and rs73684719) are potentially functional variants, located distal to EYA1 transcription start site. These polymorphisms may essentially cis-regulate its expression, as well as trans-regulate expression of other craniofacial genes via chromatin modification mechanisms. The comprehensive role of epigenetic gene regulation in general, and in the craniofacial embryonic development in particular, is poorly understood. Our results support the importance of epigenetic regulation behind normal craniofacial morphology and require further research.

Notably, there is a limited number of recent studies showing existence of thousands of enhancer sequences, potentially active in the developing craniofacial complex in mice and in humans [69,83]. Given the extremely high complexity of the craniofacial morphology, its developmental regulation and consequent phenotypic variety is likely influenced by both cis and trans epistatic interactions between genetic and epigenetic regulatory factors, which require further investigation.

Even though a relatively large number of phenotypes were studied (92 linear and angular measurements and indices), this may still represent an oversimplification of the complexity of the human face. Despite the importance of the association between specific 3D measurements and SNPs demonstrated in this study, the association of facial shapes, represented by the principal components might better represent facial morphology complexity. However, only one of the 10 principle components analysed in this study, showed significant associations at the GWAS threshold level. While the explanation of this observation is unclear,

Barash et al. J Genet Genome Res 2018, 5:041

• Page 13 of 17 •

it is consistent with other similar studies [22,23]. The specific anthropometric measurements on the other hand, produced numerous significant associations, identifying many genes and intergenic regions that appear to play important roles in the development of normal human facial appearance.

We also note the relatively small sample size used in this study, which led to a major limitation: lack of replication of these results in a separate set of samples. However, the confirmation of the two previously associated genes (*PAX3* and *COL11A1*) supports the validity of our findings.

Conclusions

The high complexity of facial anatomy, and the composite nature of genetic regulation that affects its development, requires novel comprehensive approaches of capturing facial morphology. A number of such methods has recently revealed additional genes with a few polymorphisms associated with the normal craniofacial variation range in the Caucasian population [82,80,84]. Further studies may involve the use of similar or even more advanced methods to capture the majority of craniofacial variation and benefit from larger and more diverse population samples.

A recent attempt to predict facial appearance was performed using only 24 SNPs [85]. This approach has promise, although it is largely based on reconstruction of a 'facial composite image' via prediction of ancestry, sex, pigmentation and human perception of faces. While absolutely reasonable, this approach highlights the need for additional association studies looking at specific craniofacial traits. Genetic association studies of a large scope of anthropometric measurements and other facial phenotypes are essential to uncover the dense genetic and epigenetic landscape, affecting normal facial apperance and may subsequently being useful in its prediction.

Additional association studies utilizing large sample sizes and incorporating dense SNPs and DNA methylation panels or even whole genome sequencing approaches, in conjunction with morphologically adequate representation of the craniofacial characteristics would be a valuable adjunct to the promising results obtained in this study. These studies would likely require extensive collaboration between research groups and greatly improve our understanding of the dynamic gene expression influencing craniofacial morphology, which in turn enable better prediction of the visual appearance of a person from a DNA sequence.

Acknowledgments

Ethics and Consent to Participate

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

MB designed the study, carried out the molecular genetic studies, carried out the data analysis, participated in the statistical analysis and drafted the manuscript. PB performed the statistical analysis and drafted the manuscript. AvD participated in the design of the study and drafted the manuscript. All authors read and approved the final manuscript.

Availability of Data and Materials

The data supporting the conclusions of this article are included within the article and its additional files.

Funding

References

1. Kohn LAP (1991) The Role of Genetics in Craniofacial Morphology and Growth. Annual Review of Anthropology 20: 261-278.

2. Sperber GH, Sperber SM, Guttmann GD (2010) Craniofacial embryogenetics and development. (2nd edn). Shelton, CT: People's Medical Pub. House USA.

3. Richtsmeier JT, Cheverud JM (1986) Finite element scaling analysis of human craniofacial growth. J Craniofac Genet Dev Biol 6: 289-323.

4. Neubauer S, Gunz P, Hublin JJ (2009) The pattern of endocranial ontogenetic shape changes in humans. J Anat 215: 240-255.

5. Sturm RA (2009) Molecular genetics of human pigmentation diversity. Hum Mol Genet 18: R9-R17.

6. Shkoukani MA, Chen M, Vong A (2013) Cleft Lip - A Comprehensive Review. Front Pediatr 1: 53.

7. Kimonis V, Gold JA, Hoffman TL, Panchal J, Boyadjiev SA (2007) Genetics of Craniosynostosis. Semin Pediatr Neurol 14: 150-161.

8. Weinberg SM, Naidoo SD, Bardi KM, Brandon CA, Neiswanger K, et al. (2009) Face shape of unaffected parents with cleft affected offspring: combining three-dimensional surface imaging and geometric morphometrics. Orthod Craniofac Res 12: 271-281.

9. Coussens AK, Wilkinson CR, Hughes IP, Morris CP, van Daal A, et al. (2007) Unravelling the molecular control of calvarial suture fusion in children with craniosynostosis. BMC Genomics 8: 458.

Barash et al. J Genet Genome Res 2018, 5:041

• Page 14 of 17 •

10. Coussens AK, Hughes IP, Wilkinson CR, Morris CP, Anderson PJ, et al. (2008) Identification of genes differentially expressed by prematurely fused human sutures using a novel in vivo -in vitro approach. Differentiation 76: 531-545.

11. Boehringer S, van der Lijn F, Liu F, Gunther M, Sinigerova S, et al. (2011) Genetic determination of human facial morphology: links between cleft-lips and normal variation. Eur J Hum Genet 19: 1192-1197.

12. Nakamura A, Hattori M, Sakaki Y (1997) A novel gene isolated from human placenta located in Down syndrome critical region on chromosome 21. DNA Res 4: 321-324.

13. El Ghouzzi V, Legeai-Mallet L, Benoist-Lasselin C, Lajeunie E, Renier D, et al. (2001) Mutations in the basic domain and the loop-helix II junction of TWIST abolish DNA binding in Saethre-Chotzen syndrome. FEBS Lett 492: 112-118.

14. Kamath BM, Stolle C, Bason L, Colliton RP, Piccoli DA, et al. (2002) Craniosynostosis in Alagille syndrome. Am J Med Genet 112: 176-180.

15. Hood RL, Lines MA, Nikkel SM, Schwartzentruber J, Beaulieu C, et al. (2012) Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. Am J Hum Genet 90: 308-313.

16. Roper RJ, Baxter LL, Saran NG, Klinedinst DK, Beachy PA, et al. (2006) Defective cerebellar response to mitogenic Hedgehog signaling in Down's syndrome mice. Proc Natl Acad Sci U S A 103: 1452-1456.

17. Croonen EA, van der Burgt I, Kapusta L, Draaisma JM (2008) Electrocardiography in Noonan syndrome PTPN11 gene mutation--phenotype characterization. Am J Med Genet A 146: 350-353.

18. Fourie Z, Damstra J, Gerrits PO, Ren Y (2011) Evaluation of anthropometric accuracy and reliability using different three-dimensional scanning systems. Forensic Sci Int 207: 127-134.

19. Toma AM, Zhurov A, Playle R, Ong E, Richmond S (2009) Reproducibility of facial soft tissue landmarks on 3D laser-scanned facial images. Orthod Craniofac Res 12: 33-42.

20. Kovacs L, Zimmermann A, Brockmann G, Baurecht H, Schwenzer-Zimmerer K, et al. (2006) Accuracy and Precision of the Three-Dimensional Assessment of the Facial Surface Using a 3-D Laser Scanner. IEEE Trans Med Imaging 25: 742-754.

21. Coussens AK, van Daal A (2005) Linkage disequilibrium analysis identifies an FGFR1 haplotype-tag SNP associated with normal variation in craniofacial shape. Genomics 85: 563-573.

22. Paternoster L, Zhurov AI, Toma AM, Kemp JP, St Pourcain B, et al. (2012) Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. Am J Hum Genet 90: 478-485.

23. Liu F, Van der Lijn F, Schurmann C, Zhu G, Chakravarty MM, et al. (2012) A genome-wide association study identifies five loci influencing facial morphology in Europeans. PLoS Genet 8: e1002932.

24. Michel S, Liang L, Depner M, Klopp N, Ruether A, et al. (2010) Unifying candidate gene and GWAS Approaches in Asthma. PLoS One 5: e13894.

25. Sun J, Jia P, Fanous AH, Webb BT, van den Oord EJ, et al. (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. Bioinformatics 25: 2595-6602.

26. Hrdy DB (1978) Analysis of hair samples of mummies from Semma South (Sudanese Nubia). Am J Phys Anthropol 49: 277-282.

27. Fitzpatrick TB (1988) The validity and practicality of sun-reactive skin types I through VI. Archives of Dermatology 124: 869.

28. Sturm RA, Larsson M (2009) Genetics of human iris colour and patterns. Pigment Cell Melanoma Res 22: 544-562.

29. Farkas LG (1994) Anthropometry of the head and face. (2nd edn). New York: Raven Press.

30. Website MMRW. The Jackson Laboratory, Bar Harbor, Maine.

31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

32. Online Mendelian Inheritance in Man O. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD).

33. Rebhan M, ChalifaCaspi V, Prilusky J, Lancet D (1997) GeneCards: Integrating information about genes, proteins and diseases. Trends in Genetics 13: 163.

34. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

35. Zhou N, Wang L (2007) Effective selection of informative SNPs and classification on the HapMap genotype data. BMC Bioinformatics 8: 484.

36. Kosoy R, Nassir R, Tian C, White PA, Butler LM, et al. (2009) Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 30: 69-78.

37. Paschou P, Lewis J, Javed A, Drineas P (2010) Ancestry informative markers for fine-scale individual assignment to worldwide populations. J Med Genet 47: 835-847.

38. Londin ER, Keller MA, Maista C, Smith G, Mamounas LA, et al. (2010) CoAIMs: a cost-effective panel of ancestry informative markers for determining continental origins. PLoS One 5: e13443.

39. Bulbul O, Filoglu G, Altuncul H, Aradas AF, Ruiz Y, et al. (2011) A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type. Forensic Sci Inter: Genet Suppl 3: e500-e501.

40. Tandon A, Patterson N, Reich D (2011) Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays. Genet Epidemiol 35: 80-83.

41. Phillips C, Freire Aradas A, Kriegel AK, Fondevila M, Bulbul O, et al. (2013) Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries. Forensic Sci Int Genet 7: 359-366.

42. Gettings KB, Lai R, Johnson JL, Peck MA, Hart JA, et al. (2014) A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. Forensic Sci Int Genet 8: 101-108.

43. Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, et al. (2014) Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet 10: 23-32.

44. Amigo J, Salas A, Phillips C (2011) ENGINES: exploring single nucleotide variation in entire human genomes. BMC Bioinformatics 12: 105.

45. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.

46. Marinescu V, Kohane I, Riva A (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC Bioinformatics 6: 79.

47. Wang J, Ronaghi M, Chong SS, Lee CGL (2011) pfSNP: An integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses. Human Mutation 32: 19-24.

48. Coassin S, Brandstätter A, Kronenberg F (2010) Lost in the space of bioinformatic tools: A constantly updated survival guide for genetic epidemiology. The GenEpi Toolbox. Atherosclerosis 209: 321-335.

49. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

50. Thorisson G, Smith A, Krishnan L, Stein L (2005) The International HapMap Project Web site. Genome Res 15: 1592.

51. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

52. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

53. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190.

54. UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 42: 191-198.

55. Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, et al. (2013) MalaCards: an integrated compendium for diseases and their annotation. Database (Oxford) 18.

56. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38: 214-220.

57. Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput Biol 5: e1000431.

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

59. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

60. Chelala C, Khan A, Lemoine NR (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics 25: 655-661.

61. Rajeevan H, Cheung KH, Gadagkar R, Stein S, Soundararajan U, et al. (2007) ALFRED: an allele frequency database for microevolutionary studies. Evol Bioinform Online 1: 1-10.

62. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938-2939.

63. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22: 1790-1797.

64. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, et al. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45: 896-901.

65. Visscher PM (2008) Sizing up human height variation. Nature Genetics 40: 489-490.

66. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241.

67. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46: 1173-1186.

68. Hallgrimsson B, Mio W, Marcucio RS, Spritz R (2014) Let's Face It - Complex Traits Are Just Not That Simple. PLoS Genet 10: e1004724.

69. Attanasio C, Nord AS, Zhu Y, Blow MJ, Li Z, et al. (2013) Fine tuning of craniofacial morphology by distant-acting enhancers. Science 342: 1241006.

70. Tak YG, Farnham PJ (2015) Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin 8: 57.

71. Weston EM, Friday AE, Lio P (2007) Biometric evidence that sexual selection has shaped the hominin face. PLoS One 2: e710.

72. Panagiotou OA, Ioannidis JP, Genome-Wide Significance P (2012) What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int J Epidemiol 41: 273-286.

73. Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genome wide association scans. Genet Epidemiol 32: 227-234.

74. Cooper RS, Tayo B, Zhu X (2008) Genome-wide association studies: implications for multiethnic samples. Hum Mol Genet 17: 151-155.

75. Barnholtz-Sloan JS, Chakraborty R, Sellers TA, Schwartz AG (2005) Examining population stratification via individual ancestry estimates versus self-reported race. Cancer Epidemiol Biomarkers Prev 14: 1545-1551.

76. Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, et al. (2005) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 76: 268-275.

77. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459-463.

78. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109: 1193-1198.

79. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, et al. (2011) Beyond Missing Heritability: Prediction of Complex Traits. PLOS Genetics 7: e1002051.

80. Claes P, Liberton DK, Daniels K, Rosana KM, Quillen EE,

et al. (2014) Modeling 3D facial shape from DNA. PLoS Genet 10: e1004224.

81. Lau E (2014) Non-coding RNA: Zooming in on lncRNA functions. Nat Rev Genet 15: 574-575.

82. Peschansky VJ, Wahlestedt C (2014) Non-coding RNAs as direct and indirect modulators of epigenetic regulation. Epigenetics 9: 3-12.

83. Wu H, Nord AS, Akiyama JA, Shoukry M, Afzal V, et al. (2014) Tissue-Specific RNA Expression Marks Distant-

Acting Developmental Enhancers. PLoS Genet 10: e1004610.

84. Peng S, Tan J, Hu S, Zhou H, Guo J, et al. (2013) Detecting Genetic Association of Common Human Facial Morphological Variation Using High Density 3D Image Registration. PLoS Comput Biol 9: e1003375.

85. Claes P, Hill H, Shriver MD (2014) Toward DNA-based facial composites: preliminary results and validation. Forensic Sci Int Genet 13: 208-216.

Barash et al. J Genet Genome Res 2018, 5:041

• Page 17 of 17 •