# scientific reports

Check for updates

OPEN

Whole-genome sequencing reveals de-novo mutations associated with nonsyndromic cleft lip/palate

Waheed Awotoye[1,2✉], Peter A. Mossey[3], Jacqueline B. Hetmanski[4], Lord J. J. Gowans[5], Mekonen A. Eshete[6], Wasiu L. Adeyemo[7], Azeez Alade[2,8], Erliang Zeng[9], Olawale Adamson[7], Thirona Naicker[10], Deepti Anand[11], Chinyere Adeleke[2], Tamara Busch[2], Mary Li[2], Aline Petrin[1,12], Babatunde S. Aregbesola[13], Ramat O. Braimah[13], Fadekemi O. Oginni[13], Ayodeji O. Oladele[13], Abimbola Oladayo[2], Sami Kayali[2], Joy Olotu[14], Mohaned Hassan[2], John Pape[2], Peter Donkor[15], Fareed K. N. Arthur[5], Solomon Obiri-Yeboah[16], Daniel K. Sabbah[17], Pius Agbenorku[15], Gyikua Plange-Rhule[18], Alexander Acheampong Oti[16], Rose A. Gogal[19], Terri H. Beaty[4], Margaret Taub[4], Mary L. Marazita[20], Michael J. Schnieders[19], Salil A. Lachke[11,21], Adebowale A. Adeyemo[22], Jeffrey C. Murray[23] & Azeez Butali[1,2✉]

The majority (85%) of nonsyndromic cleft lip with or without cleft palate (nsCL/P) cases occur sporadically, suggesting a role for de novo mutations (DNMs) in the etiology of nsCL/P. To identify high impact protein-altering DNMs that contribute to the risk of nsCL/P, we conducted whole-genome sequencing (WGS) analyses in 130 African case-parent trios (affected probands and unaffected parents). We identified 162 high confidence protein-altering DNMs some of which are based on available evidence, contribute to the risk of nsCL/P. These include novel protein-truncating DNMs in the *ACTL6A*, *ARHGAP10*, *MINK1*, *TMEM5* and *TTN* genes; as well as missense variants in *ACAN*, *DHRS3*, *DLX6*, *EPHB2*, *FKBP10*, *KMT2D*, *RECQL4*, *SEMA3C*, *SEMA4D*, *SHH*, *TP63*, and *TULP4*. Many of these protein-altering DNMs were predicted to be pathogenic. Analysis using mouse transcriptomics data showed that some of these genes are expressed during the development of primary and secondary palate. Gene-set enrichment analysis of the protein-altering DNMs identified palatal development and neural crest migration among the few processes that were significantly enriched. These processes are directly involved in the etiopathogenesis of clefting. The analysis of the coding

[1]Iowa Institute for Oral Health Research, University of Iowa, Iowa City, IA, USA. [2]Department of Oral Pathology, Radiology and Medicine, College of Dentistry, University of Iowa, Iowa City, IA, USA. [3]Department of Orthodontics, University of Dundee, Dundee, UK. [4]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [5]Department of Biochemistry and Biotechnology, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [6]Surgical Department, School Medicine, Addis Ababa University, Addis Ababa, Ethiopia. [7]Department of Oral and Maxillofacial Surgery, University of Lagos, Lagos, Nigeria. [8]Department of Epidemiology, College of Public Health, University of Iowa, Iowa City, IA, USA. [9]Division of Biostatistics and Computational Biology, College of Dentistry, University of Iowa, Iowa City, IA, USA. [10]Department of Pediatrics, University of KwaZulu-Natal, Durban, South Africa. [11]Department of Biological Sciences, University of Delaware, Newark, USA. [12]Department of Orthodontics, University of Iowa, Iowa City, IA, USA. [13]Department of Oral and Maxillofacial Surgery, Obafemi Awolowo University, Ile-Ife, Osun A234, Nigeria. [14]Department of Anatomy, University of Port Harcourt, Choba, Nigeria. [15]Department of Surgery, School of Medicine and Dentistry, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [16]Department of Maxillofacial Sciences, School of Medicine and Dentistry, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [17]Department of Child Oral Health and Orthodontics, School of Medicine and Dentistry, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [18]Department of Child Health, School of Medicine and Dentistry, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [19]Center for Biocatalysis and Bioprocessing (CBB), University of Iowa, Iowa City, USA. [20]Center for Craniofacial and Dental Genetics, Department of Oral and Craniofacial Sciences, School of Dental Medicine, and Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [21]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, USA. [22]National Human Genomic Research Institute, Bethesda, MD, USA. [23]Department of Pediatrics, University of Iowa, Iowa City, IA, USA. ✉email: awotoye@uiowa.edu; Azeez-butali@uiowa.edu

sequence in the WGS data provides more evidence of the opportunity for novel findings in the African genome.

Abbreviations

BAM	Binary alignment map
CADD	Combined annotation dependent depletion
CNV	Copy number variant
DAVID	Database for annotation, visualization and integrated discovery
DNM	De novo mutation
GATK	GenomeAnalysisToolKit
GMKF	Gabriella–Miller kids first
GQ	Genotype quality
GSEA	Gene set enrichment analysis
GVCF	Genome variant call format
HapMap	Haplotype map
HPE	Holoprosencephaly
HWE	Hardy–Weinberg equilibrium
Indel	Insertions or deletions
IRB	Institution Review Board
MAF	Minor allele frequency
MGI	Mouse genome informatics
NAMD	NAnoscale molecular dynamics
nsCL/P	Nonsyndromic cleft lip with or without cleft palate
nsCLO	Nonsyndromic cleft lip only
nsCLP	Nonsyndromic cleft lip and palate
nsCPO	Nonsyndromic cleft palate only
PCR	Polymerase chain reaction
PolyPhen	Polymorphism phenotyping
QC	Quality control
RD	Read depth
SAM	Sequence alignment map
SIFT	Sorting Intolerant from tolerant
SNP	Single nucleotide polymorphism
SysFACE	Systems tool for craniofacial expression
VCF	Variant call format
WGS	Whole-genome sequencing
YRI	Yoruba in Ibadan

Nonsyndromic clefts of the lip with or without cleft palate (nsCL/P) represent one of the most common types of birth defect in humans and the most common of the craniofacial region[1]. These are developmental malformation resulting from the failure of the well-coordinated complete fusion of the facial prominences during embryogenesis[2]. This birth defect is one of the sub-types of nonsyndromic orofacial clefts (nsOFCs) which accounts for 70% of the orofacial clefts (OFCs). Other sub-types of nsOFCs include nonsyndromic cleft palate only (nsCPO). The combined global prevalence of OFCs is reported to be 1 in 700 livebirths[3].

Associated impairments due to these malformations include feeding problems, speech defects, malocclusion, and esthetics problems. Studies have reported a significant increase in overall mortality in people with nsCL/P[4]. Families of affected individuals are often stigmatized and have reported huge burdens on their financial, psychological and social well-being[5]. Effective management involves a team of specialists who conduct surgical repair of the defect and manage challenges in dental, speech and psychology[6,7]. Due to the need for multi-disciplinary expertise over the life course from birth to adulthood, the cost of management and the negative impact on oral health-related quality of life, nsCL/P poses a huge public health burden.

Genetic and environmental factors have been shown to contribute to the risk of nsCL/P at the population level. The role of genetics in the etiology of clefts have been well-reported with an estimated heritability rate between 50 and 80%[8]. Although all cleft cases tend to show familial tendency but only about 40% have been causally linked to genetic risk factors[9]. Familial aggregation of orofacial clefts indicates the contribution of genetics to the risk of these defects. A study found that first degree relatives of families with CL/P have 32 times higher risk of recurrence compared to those without positive family history[10]. This study also reported that the relative risk of CP recurrence among first degree relative is 56 times higher than that of the population-based control[10]. As the degree of segregation increases, the relative risk of clefts in affected families continues to reduce[10]. This is evident from the relative risks in third degree relatives not significantly different from population-based control[10]. Additionally, twin studies reported a concordance rate of clefts in monozygotic twins (40–60%) to be higher than that reported in dizygotic twins (3–5%)[11]. Other studies have reported that a nsCL/P affected parent has a 3.2% chance of having an affected offspring[10]. Following the birth of an affected child to such parent, the risk of the defect in another birth increases to 15.8%[10]. However, unaffected parents with an affected child have a 4.4% risk of having another child born with nsCL/P[10]. Putting these together, a strong genetic component in the etiopathogenesis of clefts becomes discernable.

Despite extensive genome-wide association studies (GWAS) which have identified about 60 common risk loci, ~ 75% of the estimated heritability of liability to nsCL/P remains unexplained[12–22]. The contribution of rare coding variants has also been investigated to identify some of this missing heritability[23–25]. However, a significant knowledge gap still remains in our understanding of the genetics of nsCL/P. nsCL/P can either be sporadic or familial. Majority (80%) of the cases are sporadic, thus, suggesting a role for de novo mutations (DNMs)[26]. A few studies have examined the role of these DNMs in the etiology of nsCL/P through targeted sequencing analysis of candidate genes in affected families[27,28]. With the advent of next-generation sequencing, discovery of DNMs that may contribute to the risk of nsCL/P has yielded more positive results. The first large scale whole-genome sequencing (WGS) study reported an enrichment of DNMs in multi-ethnic nsCL/P case-parent trios of European, Colombian, or Taiwanese ancestry[29]. However, the role of these DNMs is yet to be studied on the African continent which has populations with the most genetic diversity and provides opportunity for novel findings[30].

Using the WGS data from nsCL/P African case-parent trios generated as part of the Gabriella Miller Kids First (GMKF) Pediatric Research Consortium, here we investigated the role of high impact DNMs and here we identify some that could increase the risk of nsCL/P.

Results

Samples and variants filtration.
Following our deep phenotyping and samples recruitment through the African Craniofacial Anomalies Network (AfriCRAN)[14], 150 case-parent trios (i.e., each trio consists of an affected child and unaffected parents as depicted in Fig. 1A) were selected for whole-genome sequencing at the Broad Institute. The ethnicity was determined at the point of recruitment and during the quality control. At the point of recruitment, case-parent trios were included provided their parent and grandparents were of African descent. Additionally, our quality control confirmed that these cohorts' principal components of ancestry (PCA) clustered with the HapMap sample PCs from west African countries. These samples were part of the cohorts used in the first African cleft GWAS which has been published[14].

The quality control (QC) process checked for completeness of the sequenced genomes, Mendelian errors and relatedness; this resulted in dropping 20 trios. The variants from the remaining 130 case-parent trios were filtered for high quality by ensuring that each had a genotype quality of at least 20 (GQ ≥ 20) and read depth of at least 10 (RD ≥ 10). These high-quality variants were further filtered to identify those predicted to have a high impact on the gene product and present in the case but not the parents, i.e., a de novo mutation (DNM) (MAF < 1%; protein-truncating and missense consequences). Details of the number of variants from each data filtering steps are shown in Fig. 1B. This resulted in the identification of 162 protein-altering DNMs which averages 1.2 DNMs within the coding region per case (Fig. 1C). All the protein-altering DNMs were sanger validated.

Novel protein-altering DNMs contribute to the risk of nsCL/P.
We found 162 protein-altering DNMs (Supplementary Table 1) 17 of them in genes recognized to play roles in craniofacial development and potentially contribute to the risk of nsCL/P (Table 1). Copy number variations (CNVs) involving these genes indicate that these genes could well be involved in craniofacial morphogenesis (Table 1). Mouse knockouts of *Acan,Dhrs3, Kmt2d, Recql4, Shh* and *Tp63* showed orofacial cleft phenotypes. The remaining genes: *ACTL6A, ARHGAP10, FKBP10, MINK1, TMEM5, TTN* and *TULP4* lack mouse knockout models to support their involvement in OFCs but other craniofacial dysmorphologies have been reported. Knockout of *Actl6a* in mice was reported to be embryonically lethal as the mice did not survive beyond developmental stage E6.5[31].

The DNM in *TTN* was found in the exon 49 which changed the codon for the amino acid Arginine to a stop codon at position 4738. The consequence of this premature stop codon is a truncation in the polypeptide chain and hence result in a loss of function of the gene product. Similar mutation consequences were found in the *ARHGAP10* and *MINK1* genes. However, the protein-truncating mutations in *ACTL6A* and *TMEM5* are due to an initiator codon and splice donor variants, respectively. The missense mutation in *DHRS3* changes a Serine amino acid to leucine at position 37. This position falls in the *DHRS3* catalytic domain which is critical for its enzymatic function in vitamin A metabolism. The mutation in *TP63* lies in the highly conserved sterile alpha motif (SAM) domain which is critical for protein–protein interaction of the molecule. Interestingly, the damaging protein-altering DNM discovered in *SHH* gene (p.Ser362Leu) in this study has been previously reported as the cause of a syndromic cleft(OMIM #142945)[32]. Following this discovery, a detailed review of medical records of the case carrying this DNM was done but we found no evidence of any other structural birth defect. This suggests that this mutation may contribute to the risk of syndromic as well as nonsyndromic clefts[33].

Bioinformatics analysis showed pathogenicity of identified DNMs.
The pathogenicity of these DNMs was predicted using in-silico tools to investigate the effect of the DNMs on gene expression, protein structures and functions. We used Combined Annotated Dependent Depletion (CADD) tool to ascertain the deleteriousness of the nucleotide changes that caused the DNMs. Additionally, we used Sorting Intolerant From Tolerant (SIFT) and Polymorphism Phenotyping (PolyPhen2) tools to predict how damaging the amino acid changes (resulting from the DNMs) are to the protein; and finally used HOPE to identify the effect of the amino acid changes on the protein structure and function. DNMs in the *ACTL6A, ARHGAP10, MINK1, TMEM5* and *TTN* were predicted to be among the top 0.1% most deleterious mutations in the human genome (see Table 1 for CADD scores). The DNMs in the *DHRS3, SHH, TP63* and *TULP4* genes were predicted to be among the top 1% most deleterious mutations in the human genome (see Table 1 for CADD scores) while the other genes (DNMs in *ACAN, FKBP10, KMT2D* and *RECQL4*) are among the top 10% deleterious mutations (Table 1). Among the missense DNMs, only variants in *ACAN* and *FKBP10* were predicted by both SIFT and PolyPhen2 to be well tolerated and are benign. Other missense DNMs were predicted by at least one of the two in-silico tools to be deleterious or damaging (Table 1).
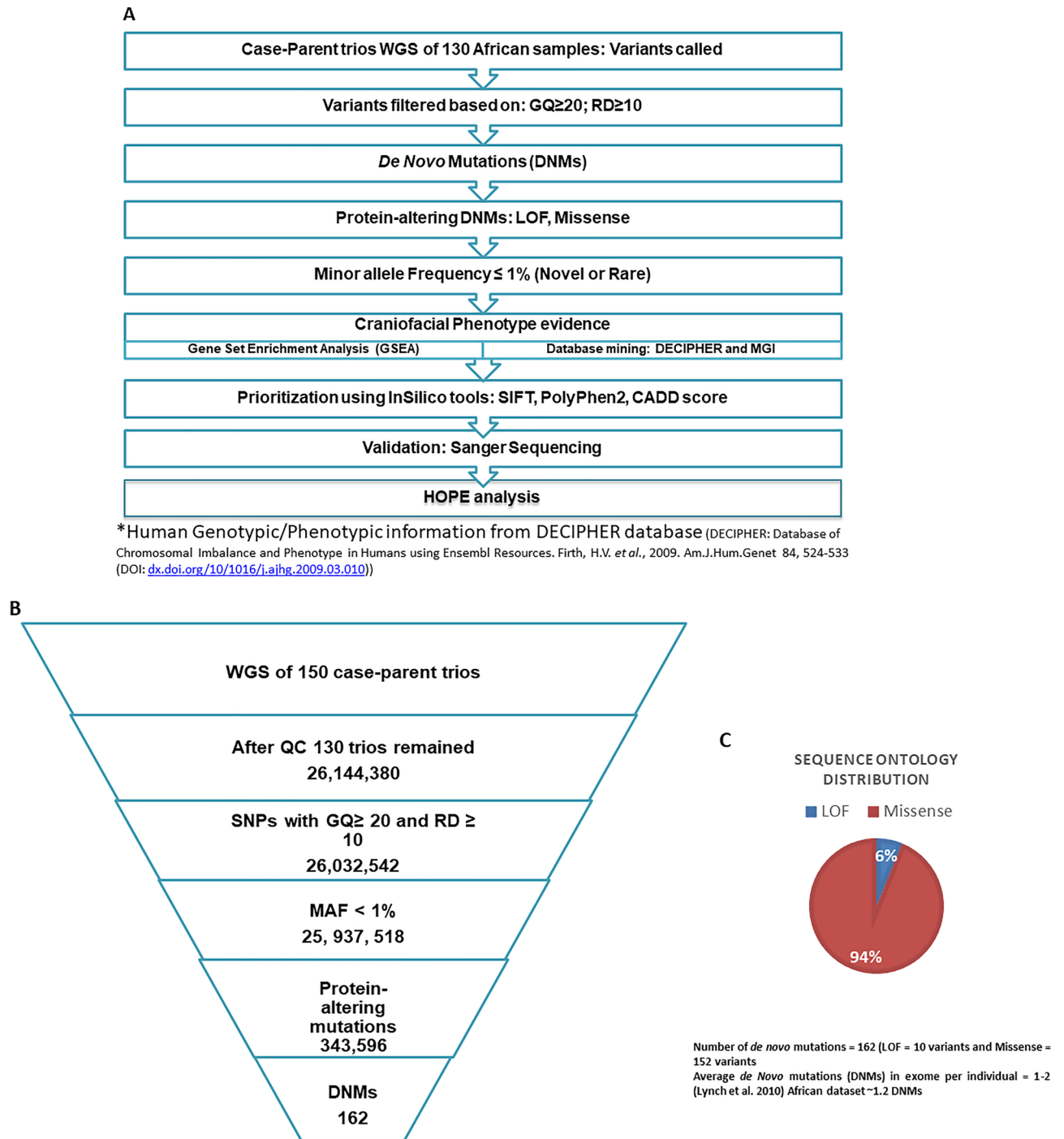
A

Case-Parent trios WGS of 130 African samples: Variants called

Variants filtered based on: GQ≥20; RD≥10

De Novo Mutations (DNMs)

Protein-altering DNMs: LOF, Missense

Minor allele Frequency ≤ 1% (Novel or Rare)

Craniofacial Phenotype evidence

| Gene Set Enrichment Analysis (GSEA) | Database mining: DECIPHER and MGI |

Prioritization using InSilico tools: SIFT, PolyPhen2, CADD score

Validation: Sanger Sequencing

HOPE analysis

*Human Genotypic/Phenotypic information from DECIPHER database (DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources. Firth, H.V. *et al.*, 2009. Am.J.Hum.Genet 84, 524-533 (DOI: dx.doi.org/10/1016/j.ajhg.2009.03.010))

B

WGS of 150 case-parent trios

After QC 130 trios remained
26,144,380

SNPs with GQ≥ 20 and RD ≥ 10
26,032,542

MAF < 1%
25, 937, 518

Protein-altering mutations
343,596

DNMs
162

C

SEQUENCE ONTOLOGY DISTRIBUTION

■ LOF ■ Missense

6%

94%

Number of *de novo* mutations = 162 (LOF = 10 variants and Missense = 152 variants
Average *de Novo* mutations (DNMs) in exome per individual = 1-2 (Lynch et al. 2010) African dataset ~1.2 DNMs

Figure 1. Case-parent trios' definition, cleft sub-types, and Data filtration pipeline. (**A**) Data filtration pipeline used to identify the high confidence de novo mutations (DNMs) that contribute to the risk of nsCL/P. (**B**) Details of the number of variants from each data filtering steps which resulted in 162 DNMs in protein-coding genes. (**C**) Pie chart showing the distribution of the effects of the de novo Variants. Majority of the DNMs (94%) cause amino acid changes which alters the protein structures and functions while 6% cause loss of function mutation in the protein-coding genes.

Our analysis of the effect of the amino acid changes due to these missense DNMs showed that the resulting alterations in the protein structures could impact the functions of the molecules encoded by the mutated genes. Notably, the secondary structures and the protein interactions are affected (Supplementary Fig. 1). The amino acid changes identified in DHRS3, DLX6, SEMA3C, SEMA4D, SHH and TP63 occur at highly conserved region and/or critical domains (Supplementary Fig. 1).

The SHH mutation p.Ser362Leu we discovered lies in a highly conserved region of the protein; the serine residue is the only amino acid at this location in available vertebrate orthologs going back to Zebrafish. This

Chromosome Location Hg 38	Genes	HGVS c.	HGVS p.	Effect	Mouse Craniofacial phenotype (MGI)	CADD Score	SIFT Score	PolyPhen2 Score
15:88856906	*ACAN*	c.4321A>G	p.Thr1441Ala	missense	Yes	0.089	0.515	0.138
179563094	*ACTL6A*	c.2T>G	p.Met1Arg	protein-truncating	Nil	30	N/A	N/A
4:147875102	*ARHGAP10*	c.784C>T	p.Arg262Ter	protein-truncating	Nil	37	N/A	N/A
1:12617239	*DHRS3*	c.110C>T	p.Ser37Leu	missense	Yes	25	0.013	0.614
7:97007691	*DLX6*	c.490G>C	p.Gly164Arg	missense	Yes	29.5	0.021	0.746
1:22863094	*EPHB2*	c.869C>A	p.Pro290His	missense	Nil	25.6	0	1
17:41821726	*FKBP10*	c.1472C>A	p.Thr491Lys	missense	Yes	16.44	0.092	0.009
12:49050509	*KMT2D*	c.3079C>T	p.Leu1027Phe	missense	Yes	16.15	0	0.001
17:4887701	*MINK1*	c.1141C>T	p.Arg381Ter	protein-truncating	Yes	37	N/A	N/A
8:144512000	*RECQL4*	c.3304G>T	p.Asp1102Tyr	missense	Yes	14.77	0.029	0.94
7:80828605	*SEMA3C*	c.244A>G	p.Ile82Val	Missense	Nil	26.3	0.193	0.351
9:89381234	*SEMA4D*	c.1559G>A	p.Cys520Tyr	missense	Nil	24.4	0	1
7:155803204	*SHH*	c.1085C>T	p.Ser362Leu	missense	Yes	26.4	0.004	1
12:63802407	*TMEM5 (RXYLT1)*	c.743+2T>A		protein-truncating	Nil	32	N/A	N/A
3:189894226	*TP63*	c.1767C>G	p.Ile589Met	missense	Yes	22.5	0.001	0.998
2:178738241	*TTN*	c.14212G>A	p.Arg4738Ter	protein-truncating	Yes	37	N/A	N/A
6:158493629	*TULP4*	c.1688C>T	p.Ser563Phe	missense	Nil	24.5	0.011	0.951

SIFT and Polyphen2 SCORE Interpretation: Tolerated   Deleterious

Table 1. List of novel variants which have evidence suggestive of involvement in craniofacial development and role in development of nsCL/P. These evidences are based on the phenotype in humans with CNVs involving these genes as well as SNVs which are reported in the DECIPHER database (https://www.deciphergenomics.org/). *Hg38: Human genome build 38 (GRCh38). *Nil: No craniofacial phenotype in mouse studies; N/A: Insilico scores not available for protein-truncating mutations.

DNM occurs within the Hedgehog (Hh) domain located at the carboxyl terminal of the polypeptide. This region is important for protein auto processing, thus modifying of the N terminal of the protein which is critical for the protein interactions which mediates Hh signaling[34,35].

Using the updated neural network-based model called AlphaFold2, we computationally predicted the 3-dimensional structure of the SHH variants[36,37]. We then estimated the folding free energy (ΔG) change through thermodynamic analysis. First, this calculated the free-energy of the unfolded SHH variant then that of the folded variants. We then estimated the difference between these free-energies and used the value to predict the protein stability[36,37]. This analysis estimated the value of folding free energy change ($\Delta\Delta G$) associated with this p.Ser362Leu variant to be 4.984 kcal/mol with a standard deviation of 0.221. The folding free energy change measures the protein stability resulting from the amino acid change due to this mutation. The mutation causes a positive folding free energy change ($\Delta\Delta G$) which indicates that it is destabilizing the protein structure[37] (Fig. 2).

Palate development and neural crest migration biological processes are significantly disrupted by DNMs. Our gene set enrichment analysis (GSEA) identified significantly enriched processes (p < 0.05) involved in normal development of the lip and palate (Fig. 3). Among these biological processes showing at least nominal significant enrichment, we identified *palate development and neural crest migration* (p values
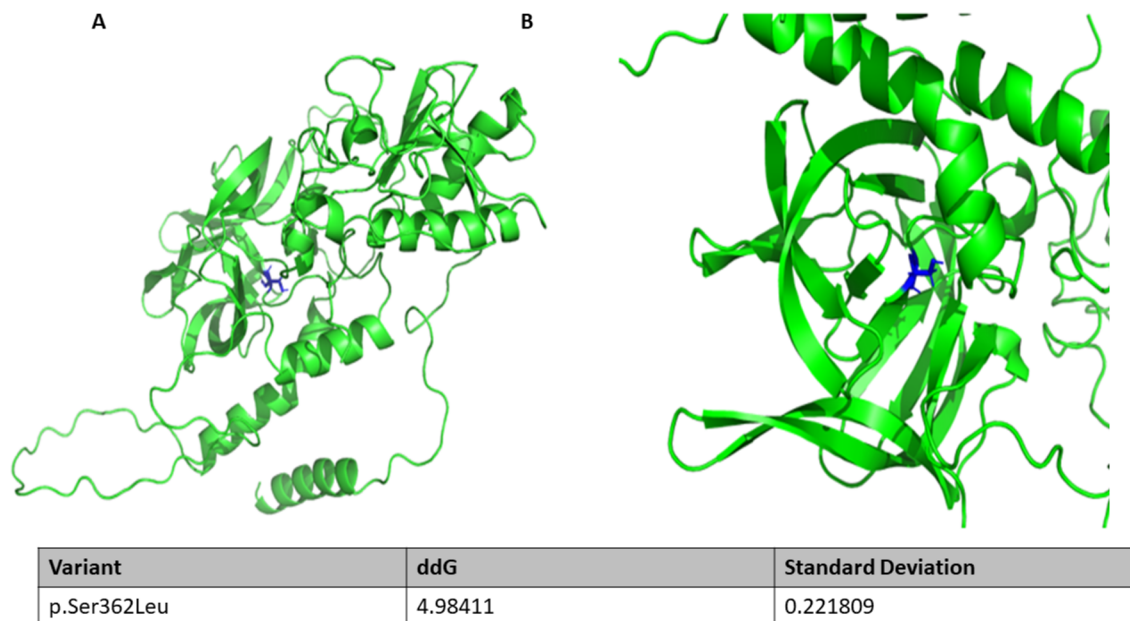
Variant	ddG	Standard Deviation
p.Ser362Leu	4.98411	0.221809

Figure 2. AlhpaFold predicted protein structures of the SHH (**A,B**). Highlighted in blue is the side-chain where the mutation occurred ((**A**) closed-up view in (**B**)). This is located within the Hedgehog domain that is critical for the hedgehog signaling. Table shows thermodynamic prediction of the effect of the p.Ser363Leu SHH DNM on the protein stability. The amino acid change resulted in a change in the folding free energy by 4.984 kcal/mol (± 0.221). This change is predicted to be disease causing.
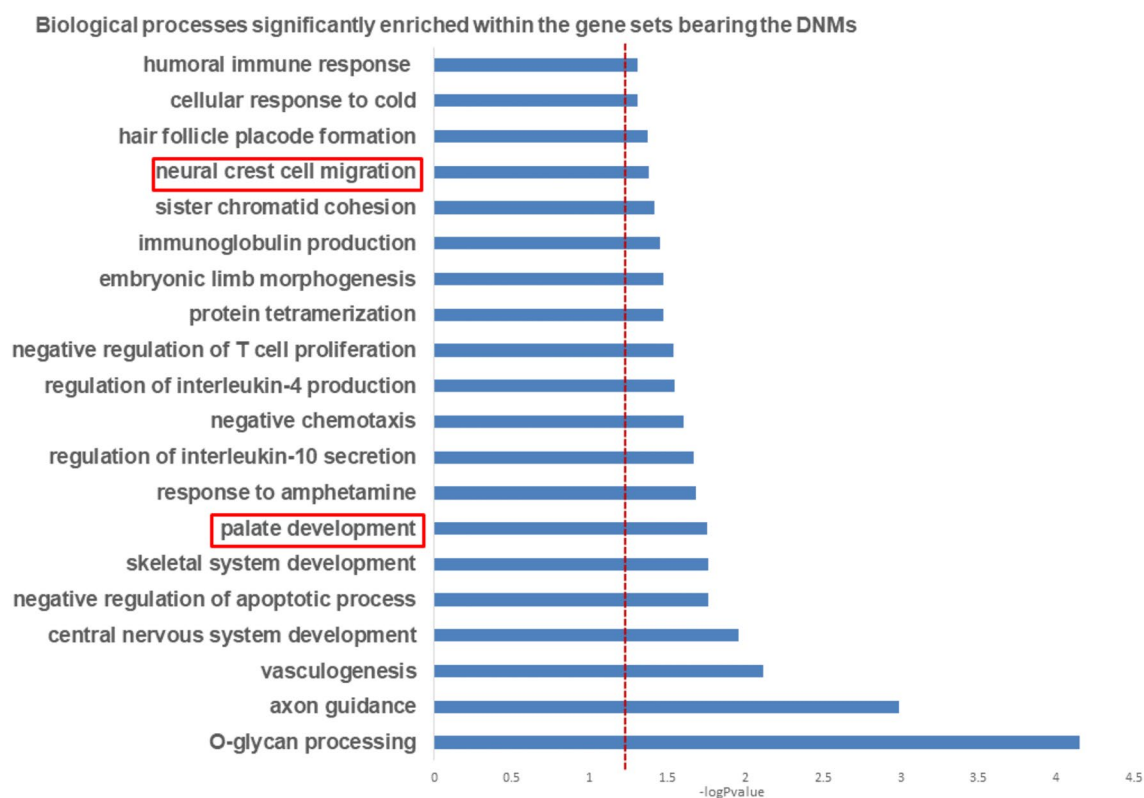
Figure 3. Graph showing significantly enriched BP from GSEA. The palate development and neural crest migration are among the processes significantly enriched ($p < 0.05$).

are 0.02 and 0.04; respectively). These biological processes have been causally linked to the etiopathogenesis of OFCs as disruption could manifest as craniofacial dysmorphology such as OFC.

Among our list of prioritized genes, the genes known to play significant roles in the development of the palate are *DHRS3*, *DLX6*, *EPHB2* and *SHH,* while those contributing to neural crest migration include *SEMA3C, SEMA4D* and *SHH*. This GSEA analysis identified *DLX6, EPHB2, SEMA3C* and *SEMA4D* as potential cleft candidate genes.

SysFACE analysis informs on gene expression in facial tissue development.   Using Systems tool for craniofacial expression-based gene discovery (SysFACE), we found that several genes with DNMs exhibit expression in mouse facial tissue development and are likely to contribute to the development of lip and palate (Fig. 4, Supplementary Fig. 2). The expression profiles showed that except for that of *TULP4* (whose ortholog was not detected in mouse), all the genes in Table 1 were found to be highly expressed in several facial tissues (Fig. 4A). Furthermore, majority of these candidates exhibit elevated expression in the E10.5 maxillary columnar epithelium compared to the E10.5 mandibular columnar epithelium (Fig. 4B). Similarly, majority of these candidates showed elevated expression in E10.5 maxillary arch compared to E10.5 mandibular arch (Fig. 4C). Several genes showed highest expression in the palate tissue (Fig. 4A). Moreover, SysFACE identified several other candidate genes that were expressed in mouse facial development (Supplementary Fig. 2).

Discussion

To our knowledge, this case-parent trios' analysis of the coding sequence in the genome to identify high impact protein-altering DNMs that could contribute to the risk of nsCL/P is a first of such analysis in an African population. Our approach utilized a trio-based study design and analysis to identify potentially pathogenic variants that can explain sporadic cases of nsCL/P. Our analysis also incorporated data from publicly available databases to identify those genes that could play a role in craniofacial development in animal models. We identified several high impact protein-truncating and missense DNMs that appear to contribute to the risk of nsCL/P. Additionally, we used thermodynamic analysis to investigate the effect of the amino acid change on the protein stability.

The protein-truncating DNMs were found in *ACTL6A*, *ARHGAP10*, *MINK1*, *TMEM5* and *TTN* genes, all of which are loci with substantial annotation, functional and/or animal model data supporting their roles in orofacial clefts. *ACTL6A* encodes an actin-related protein involved in chromatin remodeling and knockout mice do not survive beyond E6.5[31]. *ARHGAP10* is a member of the Rho GTPase family that is important in cell adhesion, migration and proliferation. This rho GTPase regulates WNT signaling by reducing the expression of *CTNNB1*[38]. The dysregulation of WNT signaling has been well reported in the etiology of cleft[39]. *MINK1* encodes misshapen-like kinase 1 which functions in cell–cell adhesion and migration. Mutant mice showed an abnormal tooth morphology indicating this gene plays some role in craniofacial development[31]. *TMEM5* encodes a transmembrane protein and mutations have been implicated in neural tube defects with some affected individuals presenting with clefts[40], further GWAS found association between common SNPs in non-coding regions near this gene family and CL/P[12]. Titin protein (encoded by *TTN*) is the largest protein molecule that plays a role in the development of the striated muscles. Genetic mutations in this gene cause congenital titinopathy: a birth defect characterized by myopathies (with cardiomyopathy)[41]. Cleft palate has also been reported in some individuals with this birth defect[42]. Mutant mice showed an increased apoptosis of cells in the frontonasal process, an important tissue that contribute to the development of the lip and palate and could easily contribute to development of OFCs[43]. Our expression analysis using the SysFACE tool provides additional evidence suggesting some role for the DNMs in *ACTL6A, ARHGAP10, MINK1* and *TMEM5* in the etiopathogenesis of nsCL/P. However, experimental evidence in model animals would confirm the role of these genes in the development of the lip and palate; and may shed light on these DNMs in the etiology of cleft.

Our analysis also found several missense DNMs in genes recognized as contributing to the risk of nsCL/P. We found a damaging mutation in *DHRS3,* which encodes an enzyme important in the metabolism of retinol. The DNM we identified is in the catalytic domain which is critical for enzymatic function. Mouse knockout models for this gene resulted in the cleft palate phenotype seen at E14.5[44]. Here, we report a damaging missense DNM in *TULP4* which encodes tubby-related protein 4 which functions in post-translation modification. This gene has only been reported in association studies where it was found to be associated with orofacial cleft in Filipinos and in Africans (Ethiopia, Ghana and Nigeria)[45,46]. We also identified damaging missense DNMs in cleft candidate genes *SHH* and *TP63*. The damaging mutation in *TP63* lies within a highly conserved sterile alpha motif (SAM) domain. Damaging mutations within this SAM domain have been reported to cause clefts[47,48] and genome-wide approaches found significant association between nsCL/P and common SNPs within *TP63*[13,49,50].

Other genes which have damaging DNMs identified in our study include *KMT2D* and *RECQL4. KMT2D* encodes a methyl transferase which functions in transcription activation. It has been reported to be associated with Kabuki syndrome in both humans and mice[51]. Protein-truncating *Kmt2d* in the neural crest cells result in a fully penetrant cleft palate phenotype[52]. *RECQL4* encodes a helicase which plays a role during DNA replication. Mutations in this gene have been associated with autosomal recessive Rapadilino syndrome (OMIM #266280) which in addition to limbs, joints and knee anomalies, and affected individuals may present with cleft palate[53,54]. Mutant mice recapitulated most of these phenotypes (including cleft palate)[55]. Although the DNMs reported in *FKBP10* and *ACAN* that encode a binding protein and an extracellular matrix protein (aggrecan) are predicted to be benign, studies have suggested they play roles in craniofacial development[56]. In-vitro studies and knockout experiments in mice provide evidence of the role of aggrecan in the etiology of cleft[57,58].

Gene-set enrichment analysis identified other genes on our list whose DNMs may also contribute to the risk of nsCL/P. These genes are involved in biological processes; palate development and neural crest migration which have been directly linked with the etiopathogenesis of orofacial clefts[59]. Although the distal-less homeobox 6
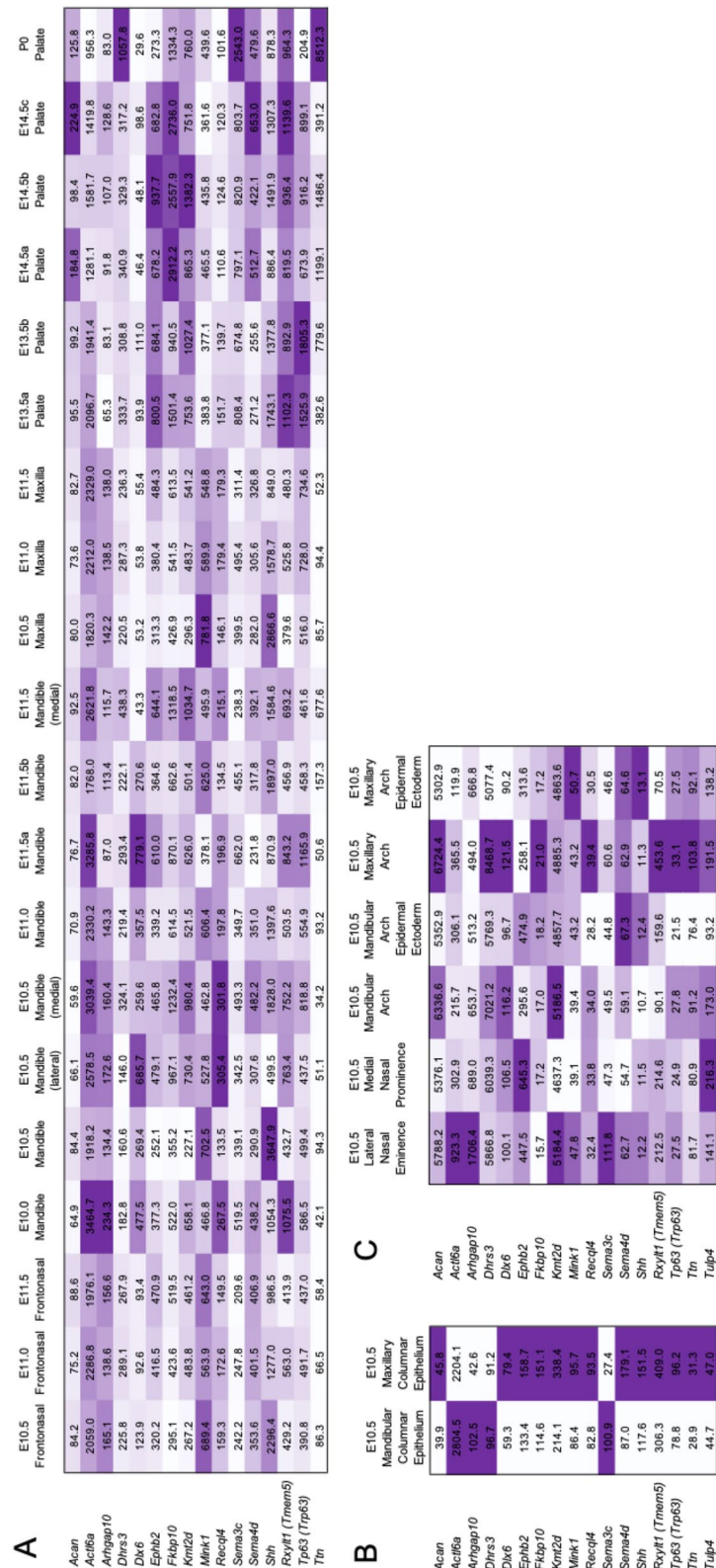
Figure 4. SysFACE-based expression analysis of candidate genes in mouse facial development. Expression of candidate genes in mouse facial development. Expression of candidate genes based on analysis using (**A**) GSE7759 microarray data generated on the Affymetrix Mouse Genome 430 2.0 Array platform, (**B**) FaceBase microarray data generated on the Affymetrix Mouse Gene 1.0 ST Array platform, and (**C**) GSE55965 microarray data generated on the Affymetrix Mouse Gene 1.0 ST Array platform. Heat-map denotes row-wise comparative expression of individual genes in different tissues at Embryonic (E) and/or postnatal (P) stages. Intensity of the color in the heat-map is representative of candidate gene expression and the average fluorescence signal intensity is shown.

(*Dlx6*) mouse knockouts show a number of craniofacial defects, cleft phenotypes have not been reported. The ephrin type B receptor 2 (*Ephb2*) is a member of the ephrin cell membrane receptors which bind to each other and initiate the Eph and ephrin signaling. The forward signaling of the bidirectional (forward-Eph and reverse-ephrin)signaling pathways is critical for normal palatogenesis[60,61]. The truncation of Ephb2 in an Ephb3 null (EphB2$^{lacZ/lacZ}$/EphB3$^{-/-}$) mice inhibited the proliferation of the palatal mesenchyme which resulted in cleft palate[60]. The sema domain (semaphorin) 3C and 4D (*SEMA3C* and *SEMA4D*) mouse knockout do not display craniofacial defects but are expressed in the 1st branchial arches which contribute indirectly to the development of the lip and palate[31]. Additionally, homozygous *Sema3c* knockout is perinatally lethal. These genes are highly expressed in the embryonic tissues critical to normal development of the lip and palate. With the aid of this SysFACE tool, we also identified novel genes contributing to the development of the lip and palate. We deduced DNMs in these genes may contribute to the risk of cleft in humans. Among other genes identified with the SysFACE tool, *MMP9* has been reported to be a key extra cellular matrix remodeling protein that could play a role in lip and palate development[62]. *Junb, Jup* and *Dnajc3* are among the genes highly expressed during the 3-way fusion that forms the lamdoidal junction in the developing face[63]. This 3-way fusion comprises fusion of the medial nasal process (MNP), lateral nasal process (LNP) and the maxillary process (MxP): MNP and LNP; MNP and MxP; LNP and MxP and is critical in the development of the lip and palate[63].

In conclusion, our analysis of the WGS in African nsCL/P case-parent trios led to the discovery of novel pathogenic genetic mutations likely to contribute to the risk of OFC. The findings of nsCL/P-risk protein-altering DNMs in some of these genes for the first time expands our knowledge of the genetic architecture of sporadic nsCL/P and provides further evidence to support the role of de novo mutations in the risk of the most common craniofacial birth defect.

Materials and methods

Study samples.   All the individuals recruited were of African ancestry from the 2 participating countries (Ghana and Nigeria). As per the established AfriCRAN protocol developed by Butali and colleagues, recruitments of infants with OFC and their parents were done during the evaluation of the affected child for surgical repair of their clefts. We recruited children affected with nsCL/P and the unaffected parents (father and mother) i.e., case-parent trios for this genetic study. In some situations, we recruited just mother and affected child i.e., dyads and in rare occasions, we recruited other family members like siblings and grand-parents. For the current study, only case-parent trios were included. Before a trio was recruited, the parents and grandparents must be ascertained to be of African ancestry and have reported no family history of any major birth defect. Following the study design, ethical approvals were obtained at the local institution review boards (IRBs) at the participating sites: Lagos University Teaching Hospital (ADM/DCST/HREC/VOL.XV/321), Obafemi Awolowo University Teaching Hospital (ERC/2011/12/01), Kwame Nkrumah University of Science and Technology (CHRPE/RC/018/130) and University of Iowa (IRB ID #: 201101720). The methods used in the recruitment at the different centers were carried out in accordance with statutory guidelines and regulations. Informed consent was obtained from all subjects included in this study.

A case-parent trio was recruited following deep phenotyping of the type of cleft and ruling out other congenital anomalies. This ensured the case (affected child) had a nonsyndromic cleft phenotype while the parents were unaffected. A standardized phenotyping protocol was used by the surgeons during the physical examination, taking clinical photographs and detailing the cleft phenotypes in a clinical database as reported in our previously published works[14,64]. Echocardiography was used to rule out congenital cardiac defects. For each trio, the cleft status describing the type of cleft was recorded. Table 2 shows the number of trios, site of recruitment and their cleft status. The distribution of the cases based on cleft types is as shown in Table 2.

Saliva samples were collected from the parents and the affected child using the Oragene saliva tool kits. Each case-parent trio was assigned a unique identifier number and their epidemiological and clinical information were remotely uploaded into a secure REDCap database. Following de-identification, the saliva samples were shipped to the Butali laboratory at the University of Iowa for processing.

DNA extraction and XY genotyping.   Saliva samples received from the recruitment centers were labeled with their unique identifier (UNID) number. The DNA was isolated from the saliva samples using the Oragene DNA extraction protocol. Extracted DNA samples were quantified using Qubit (http://www.invitrogen.com/site/us/en/home/brands/Product-Brand/Qubit.html; ThermoFisher Scientific, Grand Island, NY). Stocks and working aliquots of each DNA samples were made for future testing.

We confirmed the reported sex using the TaqMan XY genotyping. Confirmation of the sex is an inhouse quality control (QC) used in the Butali laboratory. Working aliquots (25 µl) passing QC with DNA concentration ≥ 250 ng were shipped to the Broad Institute for whole genome sequencing supported by the Gabriella Miller Kids First program.

Whole-genome sequencing and variant calling.   Our nsCL/P case-parent trios' DNA samples were part of the cohorts sequenced under the Gabriella-Miller Kids First (GMKF) Pediatric Research Consortium (https://kidsfirstdrc.org/). This consortium was established and funded to address the knowledge gaps in the understanding role of the genetics in the etiology of structural birth defects and pediatric cancers. The WGS was conducted at the Broad Institute with entire genome sequenced an average of 30 times (30 × WGS). The binary alignment map (BAM) and sequence alignment map (SAM) files were obtained after the sequence data were aligned to the Human genome assembly GRCh38 (hg38). Alternate alleles (i.e., variants from the reference genome), were called when present using the GenomeAnalysisToolKit (GATK) pipelines at the Broad Institute (https://software.broadinstitute.org/gatk/best-practices/workflow). Briefly, these variants include single nucleo-

Country	Cleft status		Total
	Cleft lip	Cleft lip and palateCleft lip and palate	
Ghana	40	64	104
Nigeria	11	15	26
	51	79	130

Table 2. Distribution of the case-parent trios based on the country of origin and cleft status. Majority (79) of the trio have CLP. Each trio consisted of a nsCL/P affected child and unaffected parents. SIFT and Polyphen2 SCORE Interpretation: Tolerated Deleterious.

tide variants (SNVs) and Insertions or deletions (Indels), were called using the HaplotypeCaller in GVCF mode and GenotypeGVCFs for single-sample variant calling and the multiple-sample joint variant calling respectively. Variants were stored in a variant call format (VCF) file which was used for further analyses.

Quality control. The quality control of the case-parent nsCL/P African trios WGS data was done using PLINK v.1.9. Each individual in a case-parent trio were evaluated on variety of quality metrics. Individuals with missingness > 10%, inconsistency between the sex reported and the average homozygosity of X-chromosome or Hardy–Weinberg Equilibrium (HWE) < 1E−06 were excluded. Also, trios showing deviation from the expected degree of relatedness between the case (offspring) and parents, or case-parent trios with Mendelian errors outside three standard deviations from the mean were excluded. Additionally, individuals with variant calls beyond 4 standard deviations from mean heterozygote/homozygote ratio were excluded.

A case-parent trio was retained only when the offspring and parent samples meet these quality control thresholds. If at least one sample in a trio does not meet these thresholds, the entire case-parent trio was excluded. After the quality control, 130 out of 150 case-parent trios were retained for downstream analyses.

Analyses for de novo mutations (DNMs) contributing to risk of nsCL/P. Following the variant calling, we filtered for high confidence protein-altering DNMs using the data filtration pipeline in Fig. 1C. Variants were first filtered based on genotype quality (GQ) ≥ 20 and a read depth (DP) ≥ 10. The high-quality variants were then filtered for mutations present in the affected case but absent in the unaffected parents (DNMs). Potential DNMs passing these filtering steps were then examined for high impact/ protein-altering mutations. Such mutations are within the coding region of the genes and the selected consequences are protein-truncating and missense mutations creating altered gene products.

Following the identification of these coding DNMs, we filtered for those variants with minor allele frequency (MAF) ≤ 1% (0.01). We did this by comparing the identified DNMs to variants reported in the 1000 Genome database (https://www.internationalgenome.org/), Exome Variant Server database (https://evs.gs.washington.edu/EVS/) and Genome Aggregation Database (https://gnomad.broadinstitute.org/). Allelic frequencies in these public databases contains whole genome sequencing data from over 7000 African and African-American controls including individuals from Ghana and Nigeria.

We then identified those genes with DNMs with some evidence of involvement in human craniofacial development. This was achieved by mining the DECIPHER database (https://www.deciphergenomics.org/) to identify those genes with copy number variants (CNVs), indels and SNVs reported in individuals with craniofacial anomalies. We prioritized those genes recognized as associated with lip and palate anomalies or anomalies in other craniofacial structures. In a bid to identify the contribution of these DNMs to the risk of nsCL/P, we also mined the Mouse Genome Informatics (MGI) database. We focused on genes among our list with cleft phenotype in mouse knockouts.

Next, we predicted the pathogenicity of these protein-altering DNMs using the bioinformatic tools such as Sorting Intolerant From Tolerant, SIFT (http://sift.jcvi.org/)[65]. Polymorphism Phenotyping, PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/)[66] and Combined Annotation Dependent Depletion, CADD (https://cadd.gs.washington.edu/)[67]. We identified protein-altering DNMs predicted to be deleterious, damaging or among the topmost deleterious mutations in the human genome.

Furthermore, we investigated the effect of missense DNMs on the protein structure and function. We used the bioinformatic tool, Help you Protein Explained: HOPE (https://www3.cmbi.umcn.nl/hope/)[68] to predict effects of the amino acid change on the protein structure and function. Additional computational methods were used to predict the structural effects of the DNM on one of the most reported cleft candidate gene products. Starting from the predicted protein structures generated by AlphaFold2, we locally optimized the structure to relax its backbone torsions and performed sidechain optimization (i.e., *sidechain repacking*) to find the most favorable position for each sidechain and improve MolProbity scores[69]. Both optimizations were done with the AMOEBA polarizable force field[70,71]. We then used the optimized protein structure to calculate the protein stability change due to the amino acid changes. Folding free energy changes (i.e., *protein stability changes*) were measured using NAnoscale Molecular Dynamics (NAMD) by calculating the free energy change due to mutation for a folded and unfolded state of the wildtype and mutant protein[72]. The protein stability change ($\Delta\Delta G$) is defined as $\Delta\Delta G = \Delta G_{folded} - \Delta G_{unfolded}$. Generally, protein stability changes that are greater than 1 kcal/mol are more likely to cause disease. This analysis offered insight into the effect of the DNMs on protein functionality.

Sanger-sequencing validation. To eliminate false positive DNMs, we conducted Sanger-sequencing validations of selected high impact DNMs discovered through WGS in our case-parent trios. Briefly, we designed primers around the DNMs by including 500 base pairs upstream and downstream of the mutation locus. The primers were designed using primer3 (https://primer3.ut.ee/) and optimized for the application of the regions containing the DNMs. The optimized primers were used to amplify these loci in DNA samples using a DNA concentration of 4 ng/µl in a 10 µl polymerase chain reaction. A YRI HapMap sample was added to the plate as a negative control. Details of the primers and annealing temperatures are available from the Butali laboratory upon request. The PCR products were sent to Functional Biosciences, Inc., Madison, WI (https://functionalbio.com/) for sequencing. Sequence data were investigated to confirm the DNMs.

Gene set enrichment analysis (GSEA) and SysFACE based craniofacial gene expression analysis. We did a gene set enrichment analysis to identify those processes significantly enriched within our DNMs gene set. We used the Database for Annotation, Visualization and Integrated Discovery (DAVID) to identify the biological processes significantly enriched within gene sets with the DNMs. The list of the genes with DNMs were entered in as a query in the DAVID Bioinformatics Resources 6.8 (https://david.ncifcrf.gov/) and the analysis was run. The biological processes with at least a nominal p-value ($p < 0.05$) were selected, among others. Among those with suggestive significance value ($p < 0.05$), we identified those processes involved in lip and palatal development where genes on our list were involved.

To gain biological insights on the candidate genes among the DNM gene lists, we examined the expression in relevant craniofacial tissues using SysFACE (Systems tool for craniofacial expression-based gene discovery), as previously published[25,73]. Mouse craniofacial transcriptomics microarray data for maxilla, frontonasal and palate at embryonic (E) and post-natal (P) stages was used for examining gene expression. Transcriptomics data from public databases such as FaceBase (https://www.facebase.org) and National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (https://www-ncbi-nlm-nih-gov.udel.idm.oclc.org/geo/) was meta-analyzed as previously described[73]. The following FaceBase datasets (FB00000352, FB00000353, FB00000107, FB00000254, FB00000264, FB00000468.01, FB00000474.01, FB00000477.01, FB00000905) and NCBI Gene Expression Omnibus (GEO) datasets (GSE7759, GSE55965, GSE22989, GSE31004, GSE11400) were considered in the analysis. Craniofacial tissue expression, presented in fluorescence intensity units, were used to generate heatmap representation.

Data availability
Data available through dbGaP Accession Number: phs001997.

References

1. Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J. & Shaw, W. C. Cleft lip and palate. *Lancet* **374**, 1773–1785. https://doi.org/10.1016/s0140-6736(09)60695-4 (2009).
2. Smarius, B. *et al.* Accurate diagnosis of prenatal cleft lip/palate by understanding the embryology. *World J. Methodol.* **7**, 93–100. https://doi.org/10.5662/wjm.v7.i3.93 (2017).
3. Rahimov, F., Jugessur, A. & Murray, J. C. Genetics of nonsyndromic orofacial clefts. *Cleft Palate Craniofac. J.* **49**, 73–91. https://doi.org/10.1597/10-178 (2012).
4. Christensen, K., Juel, K., Herskind, A. M. & Murray, J. C. Long term follow up study of survival associated with cleft lip and palate at birth. *BMJ* **328**, 1405. https://doi.org/10.1136/bmj.38106.559120.7C (2004).
5. Hunt, O., Burden, D., Hepper, P. & Johnston, C. The psychosocial effects of cleft lip and palate: A systematic review. *Eur. J. Orthod.* **27**, 274–285. https://doi.org/10.1093/ejo/cji004 (2005).
6. Wehby, G. L. & Cassell, C. H. The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral Dis.* **16**, 3–10. https://doi.org/10.1111/j.1601-0825.2009.01588.x (2010).
7. Berk, N. W. & Marazita, M. L. Costs of cleft lip and palate: Personal and societal implications. *Cleft Lip Palate From Origin Treat.* **36**, 458–469 (2002).
8. Grosen, D. *et al.* Risk of oral clefts in twins. *Epidemiology* **22**, 313–319. https://doi.org/10.1097/EDE.0b013e3182125f9c (2011).
9. Vyas, T. *et al.* Cleft of lip and palate: A review. *J. Family Med. Prim. Care* **9**, 2621–2625. https://doi.org/10.4103/jfmpc.jfmpc_472_20 (2020).
10. Sivertsen, A. *et al.* Familial risk of oral clefts by morphological type and severity: Population based cohort study of first degree relatives. *BMJ* **336**, 432–434. https://doi.org/10.1136/bmj.39458.563611.AE (2008).
11. Little, J. & Bryan, E. Congenital anomalies in twins. *Semin. Perinatol.* **10**, 50–64 (1986).
12. Yu, Y. *et al.* Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nat. Commun.* **8**, 14364. https://doi.org/10.1038/ncomms14364 (2017).
13. Leslie, E. J. *et al.* Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. *Hum. Genet.* **136**, 275–286. https://doi.org/10.1007/s00439-016-1754-7 (2017).
14. Butali, A. *et al.* Genomic analyses in African populations identify novel risk loci for cleft palate. *Hum. Mol. Genet.* **28**, 1038–1051. https://doi.org/10.1093/hmg/ddy402 (2019).
15. van Rooij, I. A. *et al.* Non-syndromic cleft lip with or without cleft palate: Genome-wide association study in Europeans identifies a suggestive risk locus at 16p121 and supports SH3PXD2A as a clefting susceptibility gene. *Genes (Basel).* https://doi.org/10.3390/genes10121023 (2019).
16. Ludwig, K. U. *et al.* Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat. Genet.* **44**, 968–971. https://doi.org/10.1038/ng.2360 (2012).
17. Mangold, E. *et al.* Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat. Genet.* **42**, 24–26 (2010).
18. Birnbaum, S. *et al.* Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat. Genet.* **41**, 473–477 (2009).

19. Grant, S. F. *et al.* A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *J. Pediatr.* **155**, 909–913 (2009).

20. Beaty, T. H. *et al.* A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat. Genet.* **42**, 525–529. https://doi.org/10.1038/ng.580 (2010).

21. Sun, Y. *et al.* Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. *Nat. Commun.* **6**, 6414. https://doi.org/10.1038/ncomms7414 (2015).

22. Leslie, E. J. *et al.* A genome-wide association study of nonsyndromic cleft palate identifies an etiologic missense variant in GRHL3. *Am. J. Hum. Genet.* **98**, 744–754. https://doi.org/10.1016/j.ajhg.2016.02.014 (2016).

23. Al Mahdi, H. B. *et al.* Identification of causative variants contributing to nonsyndromic orofacial clefts using whole-exome sequencing in a Saudi family. *Genet. Test. Mol. Biomark.* **24**, 723–731. https://doi.org/10.1089/gtmb.2019.0233 (2020).

24. Aylward, A. *et al.* Using whole exome sequencing to identify candidate genes with rare variants in nonsyndromic cleft lip and palate. *Genet. Epidemiol.* **40**, 432–441. https://doi.org/10.1002/gepi.21972 (2016).

25. Liu, H. *et al.* Exome sequencing provides additional evidence for the involvement of ARHGAP29 in Mendelian orofacial clefting and extends the phenotypic spectrum to isolated cleft palate. *Birth Defects Res.* **109**, 27–37. https://doi.org/10.1002/bdra.23596 (2017).

26. Mossey, P. A. & Modell, B. Epidemiology of oral clefts 2012: An international perspective. *Front. Oral Biol.* **16**, 1–18. https://doi.org/10.1159/000337464 (2012).

27. Riley, B. M. *et al.* Impaired FGF signaling contributes to cleft lip and palate. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 4512–4517. https://doi.org/10.1073/pnas.0607956104 (2007).

28. Leoyklang, P., Siriwan, P. & Shotelersuk, V. A mutation of the p63 gene in non-syndromic cleft lip. *J. Med. Genet.* **43**, e28. https://doi.org/10.1136/jmg.2005.036442 (2006).

29. Bishop, M. R. *et al.* Genome-wide enrichment of de novo coding mutations in orofacial cleft trios. *Am. J. Hum. Genet.* **107**, 124–136. https://doi.org/10.1016/j.ajhg.2020.05.018 (2020).

30. Campbell, M. C. & Tishkoff, S. A. African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.* **9**, 403–433. https://doi.org/10.1146/annurev.genom.9.081307.164258 (2008).

31. Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A. & Richardson, J. E. Mouse genome database (MGD) 2019. *Nucleic Acids Res.* **47**, D801-d806. https://doi.org/10.1093/nar/gky1056 (2019).

32. Roessler, E. *et al.* The mutational spectrum of holoprosencephaly-associated changes within the SHH gene in humans predicts loss-of-function through either key structural alterations of the ligand or its altered synthesis. *Hum. Mutat.* **30**, E921–E935. https://doi.org/10.1002/humu.21090 (2009).

33. Basha, M. *et al.* Whole exome sequencing identifies mutations in 10% of patients with familial non-syndromic cleft lip and/or palate in genes mutated in well-known syndromes. *J. Med. Genet.* **55**, 449–458. https://doi.org/10.1136/jmedgenet-2017-105110 (2018).

34. Perler, F. B. Protein splicing of inteins and hedgehog autoproteolysis: Structure, function, and evolution. *Cell* **92**, 1–4. https://doi.org/10.1016/s0092-8674(00)80892-2 (1998).

35. Sasai, N., Toriyama, M. & Kondo, T. Hedgehog signal and genetic disorders. *Front. Genet.* https://doi.org/10.3389/fgene.2019.01103 (2019).

36. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589. https://doi.org/10.1038/s41586-021-03819-2 (2021).

37. Duan, J., Lupyan, D. & Wang, L. Improving the accuracy of protein thermostability predictions for single point mutations. *Biophys. J.* **119**, 115–127. https://doi.org/10.1016/j.bpj.2020.05.020 (2020).

38. Teng, J. P. *et al.* The roles of ARHGAP10 in the proliferation, migration and invasion of lung cancer cells. *Oncol. Lett.* **14**, 4613–4618. https://doi.org/10.3892/ol.2017.6729 (2017).

39. Kurosaka, H., Iulianella, A., Williams, T. & Trainor, P. A. Disrupting hedgehog and WNT signaling interactions promotes cleft lip pathogenesis. *J. Clin. Investig.* **124**, 1660–1671. https://doi.org/10.1172/jci72688 (2014).

40. Vuillaumier-Barrot, S. *et al.* Identification of mutations in TMEM5 and ISPD as a cause of severe cobblestone lissencephaly. *Am. J. Hum. Genet.* **91**, 1135–1143. https://doi.org/10.1016/j.ajhg.2012.10.009 (2012).

41. Oates, E. C. *et al.* Congenital titinopathy: Comprehensive characterization and pathogenic insights. *Ann. Neurol.* **83**, 1105–1124. https://doi.org/10.1002/ana.25241 (2018).

42. Chauveau, C. *et al.* Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. *Hum. Mol. Genet.* **23**, 980–991. https://doi.org/10.1093/hmg/ddt494 (2014).

43. May, S. R., Stewart, N. J., Chang, W. & Peterson, A. S. A Titin mutation defines roles for circulation in endothelial morphogenesis. *Dev. Biol.* **270**, 31–46. https://doi.org/10.1016/j.ydbio.2004.02.006 (2004).

44. Billings, S. E. *et al.* The retinaldehyde reductase DHRS3 is essential for preventing the formation of excess retinoic acid during embryonic development. *FASEB J.* **27**, 4877–4889. https://doi.org/10.1096/fj.13-227967 (2013).

45. Vieira, A. R. *et al.* Fine mapping of 6q23.1 identifies TULP4 as contributing to clefts. *Cleft Palate Craniofac. J.* **52**, 128–134. https://doi.org/10.1597/13-023 (2015).

46. Gowans, L. J. *et al.* Association studies and direct DNA sequencing implicate genetic susceptibility loci in the etiology of non-syndromic orofacial clefts in Sub-Saharan African populations. *J. Dent. Res.* **95**, 1245–1256. https://doi.org/10.1177/0022034516657003 (2016).

47. Tsutsui, K. *et al.* A novel p63 sterile alpha motif (SAM) domain mutation in a Japanese patient with ankyloblepharon, ectodermal defects and cleft lip and palate (AEC) syndrome without ankyloblepharon. *Br. J. Dermatol.* **149**, 395–399. https://doi.org/10.1046/j.1365-2133.2003.05423.x (2003).

48. Zheng, J. *et al.* Tooth defects of EEC and AEC syndrome caused by heterozygous TP63 mutations in three Chinese families and genotype-phenotype correlation analyses of TP63-related disorders. *Mol. Genet. Genomic Med.* **7**, e704. https://doi.org/10.1002/mgg3.704 (2019).

49. Marazita, M. L. *et al.* Genome-scan for loci involved in cleft lip with or without cleft palate in consanguineous families from Turkey. *Am. J. Med. Genet. A* **126**, 111–122. https://doi.org/10.1002/ajmg.a.20564 (2004).

50. Marazita, M. L. *et al.* Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. *Hum. Hered.* **68**, 151–170. https://doi.org/10.1159/000224636 (2009).

51. Bjornsson, H. T. *et al.* Histone deacetylase inhibition rescues structural and functional brain deficits in a mouse model of Kabuki syndrome. *Sci. Transl. Med.* **6**, 135. https://doi.org/10.1126/scitranslmed.3009278 (2014).

52. Shpargel, K. B., Mangini, C. L., Xie, G., Ge, K. & Magnuson, T. The KMT2D Kabuki syndrome histone methylase controls neural crest cell differentiation and facial morphology. *Development* https://doi.org/10.1242/dev.187997 (2020).

53. Wang, L. L. *et al.* Association between osteosarcoma and deleterious mutations in the RECQL4 gene in Rothmund-Thomson syndrome. *J. Natl. Cancer Inst.* **95**, 669–674. https://doi.org/10.1093/jnci/95.9.669 (2003).

54. Maciaszek, J. L. *et al.* Enrichment of heterozygous germline RECQL4 loss-of-function variants in pediatric osteosarcoma. *Cold Spring Harb. Mol. Case Stud.* https://doi.org/10.1101/mcs.a004218 (2019).

55. Mann, M. B. *et al.* Defective sister-chromatid cohesion, aneuploidy and cancer predisposition in a mouse model of type II Rothmund-Thomson syndrome. *Hum. Mol. Genet.* **14**, 813–825. https://doi.org/10.1093/hmg/ddi075 (2005).

56. Lietman, C. D. *et al.* Connective tissue alterations in Fkbp10-/- mice. *Hum. Mol. Genet.* **23**, 4822–4831. https://doi.org/10.1093/hmg/ddu197 (2014).

57. Rittenhouse, E. *et al.* Cartilage matrix deficiency (cmd): A new autosomal recessive lethal mutation in the mouse. *J. Embryol. Exp. Morphol.* **43**, 71–84 (1978).

58. Bueno, D. F. *et al.* Human stem cell cultures from cleft lip/palate patients show enrichment of transcripts involved in extracellular matrix modeling by comparison to controls. *Stem Cell Rev. Rep.* **7**, 446–457. https://doi.org/10.1007/s12015-010-9197-3 (2011).

59. Deshpande, A. S. & Goudy, S. L. Cellular and molecular mechanisms of cleft palate development. *Laryngosc. Investig. Otolaryngol.* **4**, 160–164. https://doi.org/10.1002/lio2.214 (2019).

60. Risley, M., Garrod, D., Henkemeyer, M. & McLean, W. EphB2 and EphB3 forward signalling are required for palate development. *Mech. Dev.* **126**, 230–239. https://doi.org/10.1016/j.mod.2008.10.009 (2009).

61. Benson, M. D. & Serrano, M. J. Ephrin regulation of palate development. *Front. Physiol.* **3**, 376. https://doi.org/10.3389/fphys.2012.00376 (2012).

62. Smane-Filipova, L., Pilmane, M. & Akota, I. MMPs and TIMPs expression in facial tissue of children with cleft lip and palate. *Biomed. Pap. Med. Fac. Univ. Palacky Olomouc Czech Repub.* **160**, 538–542. https://doi.org/10.5507/bp.2016.055 (2016).

63. Li, H., Jones, K. L., Hooper, J. E. & Williams, T. The molecular anatomy of mammalian upper lip and primary palate fusion at single cell resolution. *Development* https://doi.org/10.1242/dev.174888 (2019).

64. Awotoye, W. *et al.* Genome-wide gene-by-sex interaction studies identify novel nonsyndromic orofacial clefts risk locus. *J. Dent. Res.* https://doi.org/10.1177/00220345211046614 (2021).

65. Sim, N. L. *et al.* SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457. https://doi.org/10.1093/nar/gks539 (2012).

66. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* https://doi.org/10.1002/0471142905.hg0720s76 (2013).

67. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894. https://doi.org/10.1093/nar/gky1016 (2019).

68. Venselaar, H., Te Beek, T. A., Kuipers, R. K., Hekkelman, M. L. & Vriend, G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinform.* **11**, 548. https://doi.org/10.1186/1471-2105-11-548 (2010).

69. Tollefson, M. R. *et al.* Structural insights into hearing loss genetics from polarizable protein repacking. *Biophys. J.* **117**, 602–612. https://doi.org/10.1016/j.bpj.2019.06.030 (2019).

70. Ponder, J. W. *et al.* Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **114**, 2549–2564. https://doi.org/10.1021/jp910674d (2010).

71. Shi, Y. *et al.* Polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.* **9**, 4046–4063. https://doi.org/10.1021/ct4003702 (2013).

72. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130. https://doi.org/10.1063/5.0014475 (2020).

73. Cox, L. L. *et al.* Mutations in the epithelial cadherin-p120-catenin complex cause Mendelian non-syndromic cleft lip with or without cleft palate. *Am. J. Hum. Genet.* **102**, 1143–1157. https://doi.org/10.1016/j.ajhg.2018.04.009 (2018).

Acknowledgements

Author contributions

W.A. and A.B. contributed to the conception and design, data acquisition, analysis, and interpretation, drafted and critically revised the manuscript; P.M., J.B.H., L.J.J.G., M.A.E., W.L.A., A.A., E.Z., O.A., T.N., D.A, C.A., T.B., M.L., A.P., B.S.A., R.O.B., F.O.O., A.O.O., A.O., S.K., J.O., M.H., J.P., P.D., F.K.N.A., S.O.-Y., D.K.S., P.A., G.P.-R., A.A.O., R.A.G., T.H.B., M.T., M.L.M., M.J.S., S.A.L., A.A. and J.C.M. contributed to the conception, data acquisition, analysis and interpretation and critically revised the manuscript. All authors gave final approval and agreed to be accountable for all aspects of the work.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-15885-1.

Correspondence and requests for materials should be addressed to W.A. or A.B.

Reprints and permissions information is available at www.nature.com/reprints.