nature portfolio

Corresponding author(s): Brian C. Schutte, Robert A. Cornell, David Rice

Last updated by author(s): Sep 17, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Imaging data of transgenic mouse embryos were collected with Adobe Photoshop Elements 11 (Adobe Inc., San Jose, CA, USA). Cell images were collected using the Leica TCS-SPE confocal microscope and were processed with Fiji software distribution of ImageJ (v2.9.0), https://imagej.net/software/fiji/downloads.
Data analysis	R (v3.6.1 and v4.3.2), https://cran.r-project.org; AxiomGT1 (Thermo Fisher Scientific, Santa Clara, CA, USA); REGENIE (v2.2.4), https://github.com/FINNGEN/regenie; GCTA (v1.93.2), https://cnsgenomics.com/software/gcta; PLINK (v2.00a2.3LM), https://www.cog-genomics.org/plink; Eagle (v2.3.5), https://alkesgroup.broadinstitute.org/Eagle; Beagle 4.1 (version 08Jun17.d8b) and Beagle 5.0 (v.28Sep18.793) http://faculty.washington.edu/browning/beagle/beagle.html; LocusZoom (v1.3), https://genome.sph.umich.edu/wiki/LocusZoom_Standalone; METAL (v2011-03-25), http://csg.sph.umich.edu/abecasis/Metal; HOMER http://homer.ucsd.edu/homer; BCFtools (v1.18), https://www.htslib.org/download/; GraphPad Prism 10, https://www.graphpad.com/.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The GWAS data generated in this study have been deposited in the GWAS Catalog under accession codes GCST90448958, GCST90448959 and GCST90448960. Individual-level genotypes and register data from FinnGen participants can be accessed by approved researchers via the Fingenious portal (https://site.fingenious.fi/en) hosted by the Finnish Biobank Cooperative FinBB (https://finbb.fi/en/), in accordance with appropriate permissions. The individual-level data from the Estonian Biobank are available under restricted access administered by the Estonian Genome Center of the University of Tartu in accordance with the regulations of the Estonian Human Genes Research Act; access can be obtained by application at www.biobank.ee. Other data underlying this article may be shared on a reasonable request to the corresponding author/s.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Findings do not apply to one sex only. Genotype-determined sex was included as a covariate in the GWAS. Subjects whose genetic sex did not match with sex provided from registries were excluded. Non-syndromic orofacial clefts are observed in both sexes, therefore data were not analyzed separately for males and females.
Reporting on race, ethnicity, or other socially relevant groupings	The FinnGen cohort contained individuals with the Finnish ancestry only. Non-Finnish individuals, as determined by principal component analysis (PCA), were excluded. The Estonian Biobank contained individual of Estonian ethnicity and non-Estonian individuals, as determined by genetic PCA, were excluded. We ascertained biological sex from genetic analyses and included as a covariate in all GWAS. We included genotyping batch and the first ten principal components (PCs) of genotypes as covariates to account for potential confounders of batch effect and population structure, respectively.
Population characteristics	There is a total of 520,105 participants in the FinnGen study with an average age of 62.1 ± 19 (mean ± standard deviation) years, 282,064 are female. As samples were mainly collected through legacy collections and hospital biobanks, the estimated FinnGen participant may be expected to have more diseases than Finnish population average. Phenotype data come from electronic health record data from Finnish health registries. The Finnish Discovery GWAS included 173,746 females and 135,408 males, with the following phenotype breakdown: CP only (157 females and 71 males), CLP (62 females and 59 males) and CL only (28 females and 26 males). Replication cohort from the FinnGen study contained 200,100 individuals (113,132 females and 86,968 males) with 65 CP cases (116 females and 49 males). The Estonian Biobank cohort consisted of Estonian individuals, also determined by PCA. Among the 71 CP cases in the Estonian cohort there were 54 females and 17 males. Since the phenotype under the study is observed at birth, age is not a factor that contributes to disease onset. Nevertheless, age was used as a continuous covariate in the GWAS.
Recruitment	The collected samples consist of two entities: 1) legacy samples, mainly collected by the THL (National Institute for Health and Welfare in Finland) and 2) prospective samples which were mainly collected by hospital biobanks. Almost all of the Finnish biobanks are part of the FinnGen study.
Ethics oversight	The FinnGen project has been approved by the Coordinating Ethics Committee of the Helsinki and Uusimaa Hospital District. Ethical approval for the Estonian Biobank was obtained from the Research Ethics Review Committee of the University of Tartu

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All available subjects with non-syndromic orofacial clefts and individuals with no orofacial clefts in FinnGen Data Freezes 7 through 12 and the Estonian Biobank, that passed genetic data QC, were included in the study as cases and controls. No sample size calculation was performed. However, we consider our GWAS of the cleft palate (CP) phenotype, the primary phenotype the paper is focused on, well-powered as we replicated a previously identified genome-wide significant hit at the GRHL3 gene.
Data exclusions	Genotyped samples that failed pre-established quality control criteria (standard cutoff values for sample and genotype missingness), samples with estimated non-Finnish ancestry (as determined by PCA in Kurki et al. Nature, 2023), samples mismatched between genetically imputed

sex and sex in their registered phenotype data, duplicated samples/twins, and samples with syndromic forms of cleft lip with or without palate and syndromic cleft palate were excluded from the study.

Replication

We replicated association of lead SNPs with non-syndromic cleft palate from 3 independent GWAS signals, that reached genome-wide statistical significance in the FinnGen discovery cohort (from Data Freeze 7), in an independent cohort from FinnGen (cases that became available in Data Freezes 8 through 12) and in a sample set from the Estonian Biobank.

Randomization

As a retrospective, registry-based GWAS, randomization of samples was unattainable as sample assignment to cases and controls was based on clinical outcomes derived from electronic health records. Genotyping batches were mostly random (processed in batches of 5,000 in the order samples accrued). Genotyping batches were used as covariates to correct for batch effects. In addition, participant age and genomic principal components (1-10) were used as fixed-effect covariates in GWAS.

Blinding

Authors and analysts were blinded from the diagnosis of participants prior to recruitment. Newly recruited individuals were randomly asked for consent among individuals receiving cares in hospitals or when donating blood to the red cross blood bank. Personal identities were pseudonymyzed for all data analysis. In transgenic mouse reporter assays, independent observers who were asked to assess the enhancer activity based on the X-gal staining intensity of embryos, were blinded to the genotype of the transgene.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

1. The polyclonal rabbit anti-Irf6 antibody was obtained from Walid Fakhouri, School of Dentistry, University of Texas Health Science Center at Houston, Houston, TX 77054, USA. Generation of the antibody was described previously in Bailey et al., 2005. (PMID: 16049006).
2. Anti-H3K27Ac (Anti-acetyl-Histone H3 [Lys27] Antibody) - Supplier: Millipore-Sigma; Catalogue number: 07-360; Lot number: 3321098.
3. Normal Rabbit IgG - Supplier: Millipore-Sigma; Catalogue number: 12-370; Lot number: 3202364.
4. Pitx2 Antibody (H-1) - Supplier: Santa Cruz; Catalogue number: sc-390457; Lot number: H0918.
5. E-Cadherin Rabbit mAb - Supplier: ABclonal; Catalogue number: A20798; Lot number: 3501111120.
6. Secondary antibody (Mouse IgG Alexa Flour 488) - Supplier: Thermo Fisher; Catalog number: A-11001.
7. Secondary antibody (Rabbit IgG Alexa Flour 568) - Supplier: Thermo Fisher; Catalog number: A-11036.

Validation

1. The polyclonal rabbit anti-Irf6 antibody was validated in Western blot, immunofluorescence and co-immunoprecipitation as per Bailey et al., 2005 (PMID: 16049006).
2. Anti-H3K27Ac is known to work well for Western blotting and ChIP-Seq/ChIP-qPCR as mentioned on the manufacturer's website (https://www.emdmillipore.com/US/en/product/Anti-acetyl-Histone-H3-Lys27-Antibody,MM_NF-07-360) which also cites references that has used this antibody (https://www.emdmillipore.com/US/en/product/Anti-acetyl-Histone-H3-Lys27-Antibody,MM_NF-07-360#anchor_REF).
3. Rabbit IgG validation: https://www.sigmaaldrich.com/US/en/product/mm/12370 which also provides references that has used the Rabbit IgG https://www.sigmaaldrich.com/US/en/search/12-370?focus=papers&page=1&perpage=30&sort=relevance&term=12-370&type=citation_search
4. Anti-PITX2 is known to work well for immunostaining and Western blotting as mentioned on the manufacturer's website (https://www.scbt.com/p/pitx2-antibody-h-1?gad_source=1&gclid=CjwKCAiAzc2tBhA6EiwArv-i6WtN5d_JjE0MfSRcDSUAT6czjXzxBYK5hxDdKhxL9OTyYhZxu4bCdRoCh3QQAvD_BwE) which also cites references that has used this antibody.
5. Anti-ECAD is known to work well for immunostaining and Western blotting as mentioned on the manufacturer's website (https://abclonal.com/catalog-antibodies/ECadherinRabbitmAb/A20798) which also cites the references (in the publication page from the same link) that has used this antibody.
6. Mouse IgG Alexa Flour 488 is known to work well for Immunofluorescence and immunohistochemistry as mentioned on the manufacturer's website (https://www.thermofisher.com/antibody/product/Goat-anti-Mouse-IgG-H-L-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-11001) which also cites the references that has used this antibody.
7. Mouse IgG Alexa Flour 488 is known to work well for Immunofluorescence and immunohistochemistry as mentioned on the manufacturer's website (https://www.thermofisher.com/antibody/product/Goat-anti-Rabbit-IgG-H-L-Highly-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-11036) which also cites the references that has used this antibody.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Primary Epidermal Keratinocytes: Normal, Human, Neonatal Foreskin (HEKn) were purchased from American Type Culture Collection (ATCC PCS-200-010). Sex: male The human induced pluripotent stem cells UCSFi001-A [WTC11] were obtained from Coriell Institute for Medical Research under the identifier GM25256. Sex: male.
Authentication	G-banding karyotyping was performed through Diagnostic Cytogenetics; STR analysis was performed through WiCell.
Mycoplasma contamination	Testing for Mycoplasma contamination was performed using the MycoScope PCR kit (Genlantis, Fisher).
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in the study.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	All animals used in this study were of Mus musculus species and the FVB/NJ strain (7-8 weeks old egg donor females and 2-6 months old stud males) from the Jackson Laboratory, and the BDF-1 strain vasectomized males to generate pseudopregnancies (2 months to 2 years old) and the CD-1 strain recipient females (7-11 weeks old) from Charles River. Transgenic embryos were assayed at embryonic day 13.5.
Wild animals	The study did not involve wild animals.
Reporting on sex	Sex was not considered in the study design as the phenotype of interest do not apply to only one sex.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	All animal work done in this study was reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*