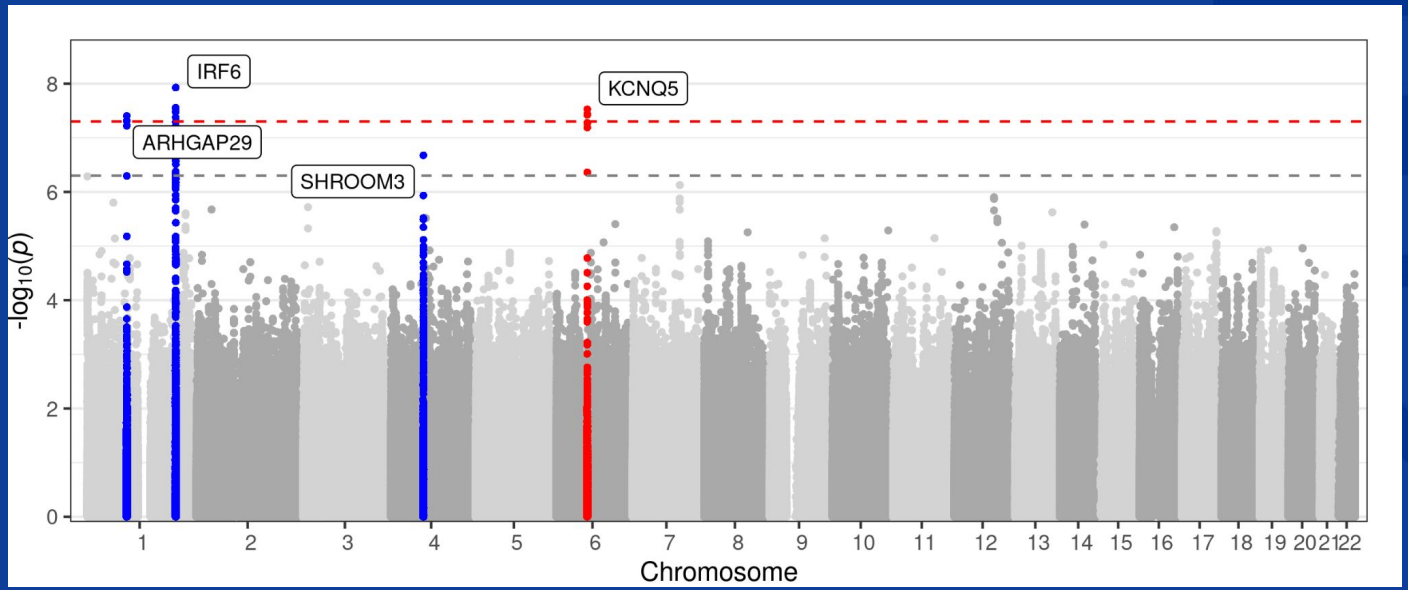
University of Pittsburgh | School of Public Health

Genome-wide Association Study of Cleft Lip with or without Palate in a Filipino Population



Dylan Maher
Masters Defense
July 30th, 2024

University of Pittsburgh | School of Public Health

Background

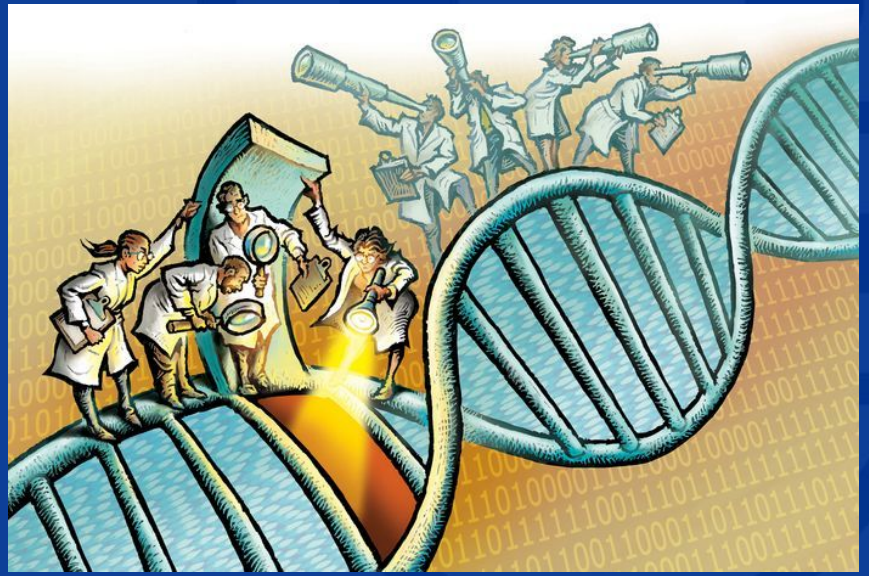
Cleft Lip with or without Palette (CL/P)

- One of the most common congenital anomalies worldwide
- Estimated prevalence
  - ~1/1000 (global)
  - ~1/500 (Filipino)
- Significant healthcare costs to affected & families
- A lot of “missing heritability”

University of Pittsburgh | School of Public Health

Heritability

- Heritability estimates high (~0.9) [family studies]
  - Primarily additive
  - Some contribution from non-shared environment
- Polygenic (uncertain to what degree)
- Common variants (~0.3) [Ludwig et al. (2017)]
- ~⅔ variance unexplained



University of Pittsburgh | School of Public Health

Outline

Methods

- ❏ Population Structure Analysis
- ❏ GWAS
  - ❏ fastGWA-GLMM
  - ❏ Covariate Justification
- ❏ TWAS
  - ❏ Note on Interpretation
- ❏ Gene-Based Test
  - ❏ m-BAT
  - ❏ “Masking effects”
- ❏ Fine-Mapping
- ❏ COJO
  - ❏ “Conditional”
  - ❏ “Joint”

Results

- ❏ GWAS
  - ❏ Significant SNPs found
  - ❏ Knowns/unknowns
- ❏ TWAS
  - ❏ Confirmatory Results
- ❏ Gene-Based Test
  - ❏ Chromosome 1 region
- ❏ Fine-Mapping
  - ❏ Chromosome 1 region
- ❏ COJO
  - ❏ Functional information

University of Pittsburgh | School of Public Health

Methods

- Genotyping performed by CIDR (4114 samples)
  - Infinium Global Diversity Array
- Imputation via TopMed Reference Panel
  - Filtered based on genotype probability (90%)
- Standard quality checks
  - Call & error rate
  - Sex check
  - Relatedness
  - Pop structure
  - Batch effects
  - ME
  - HWE

CIDR Center for Inherited Disease Research
Johns Hopkins University
Mapping a Genetic Path to Better Health

University of Pittsburgh | School of Public Health

Population Structure Analysis

PCA performed in PLINK

- 77,052 independent SNPs
- MAF > 0.01
- Pairwise LD (R^2 < 0.1)

PCA conducted on Philippines samples (n=2174)

- kinship estimation to categorize by relatedness
- PC/Biplots visually inspected (stratified by phenotype and DNA source)
- No systematic allele frequency differences detected

PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses

University of Pittsburgh | School of Public Health

Population Structure Analysis

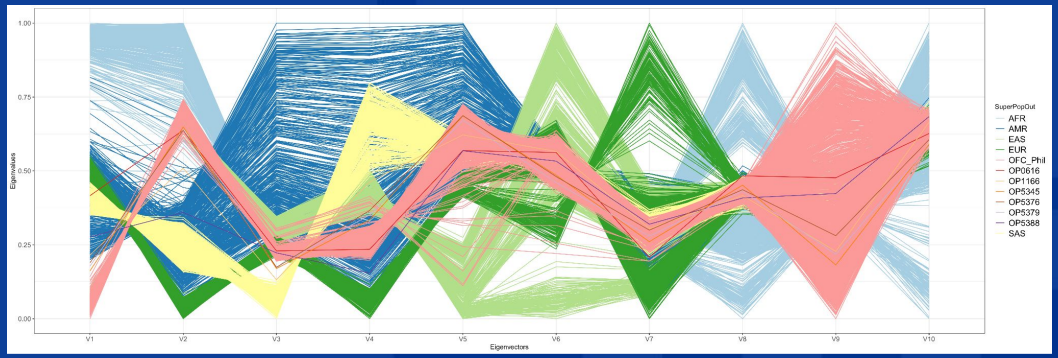
PCA performed on Philippines dataset combined with 1000 Genomes Project data (n=3202)
Kinship estimation as before

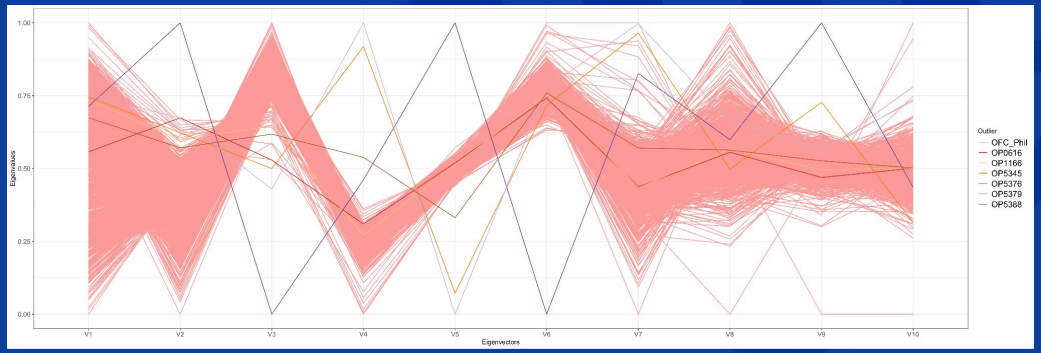
PCA performed on CL/P cases/controls

First five PCs deemed sufficient for inclusion as covariates

Final Sample

- 517 controls
- 882 cases
- 1399 total





University of Pittsburgh | School of Public Health

GWAS

GWAS performed with fastGWA-GLMM (part of GCTA software package)

Sparse GRM to capture relatedness, estimates variance components, performs score tests for each variant

Shown to be computationally efficient & produce well-calibrated test statistics for common & rare variants

TECHNICAL REPORT
https://doi.org/10.1038/s41588-021-00954-4
nature genetics
Check for updates

A generalized linear mixed model association tool for biobank-scale data

Longda Jiang[1,2,4], Zhili Zheng[1,4], Hailing Fang[2,3] and Jian Yang[1,2,3]

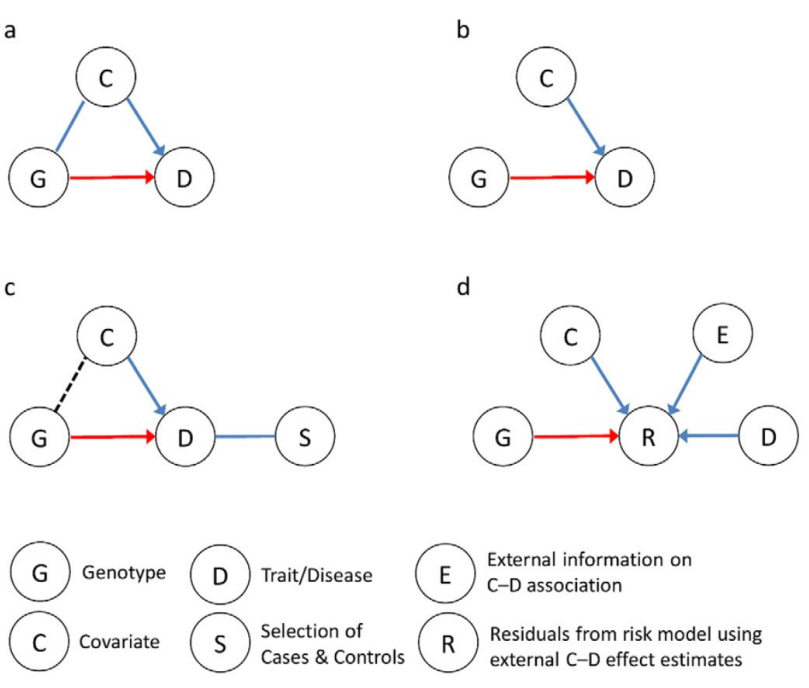
GCTA
a tool for Genome-wide Complex Trait Analysis

Evaluation of GENESIS, SAIGE, REGENIE and fastGWA-GLMM for genome-wide association studies of binary traits in correlated data

Anastasia Gurinovich[1]*, Mengze Li[2], Anastasia Leshchyk[2], Harold Bae[3], Zeyuan Song[4], Konstantin G. Arbeev[5], Marianne Nygaard[6], Mary F Feitosa[7], Thomas T Perls[8] and Paola Sebastiani[1]

University of Pittsburgh | School of Public Health

GWAS (covariates)



ANALYSIS

nature genetics

Including known covariates can reduce power to detect genetic effects in case-control studies

Matti Pirinen[1], Peter Donnelly[1,2] & Chris C A Spencer[1]

factors, including the prevalence of the disease studied. When the disease is common (prevalence of >20%), the inclusion of covariates typically increases power, whereas, for rarer diseases, it can often decrease power to detect new genetic associations.

OPEN ACCESS Freely available online

PLOS | GENETICS

Perspective

The Covariate's Dilemma

Joel Mefford[1,2], John S. Witte[1,2]*

1 Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America, **2** Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America

University of Pittsburgh | School of Public Health

GWAS (covariates)

where $p_i = P(Y_i = 1|a, b, \boldsymbol{c}, g_i, \boldsymbol{x}_i)$. If the genotype G is independent of the covariates $\boldsymbol{X}$ in the general population, then the arguments of Proposition 1 show that $\boldsymbol{X}$ is not a confounder of G-Y association, and thus the model $\mathcal{M}^*$ is valid for testing the genetic effect also in this case. The expected value of the element of the Fisher information matrix

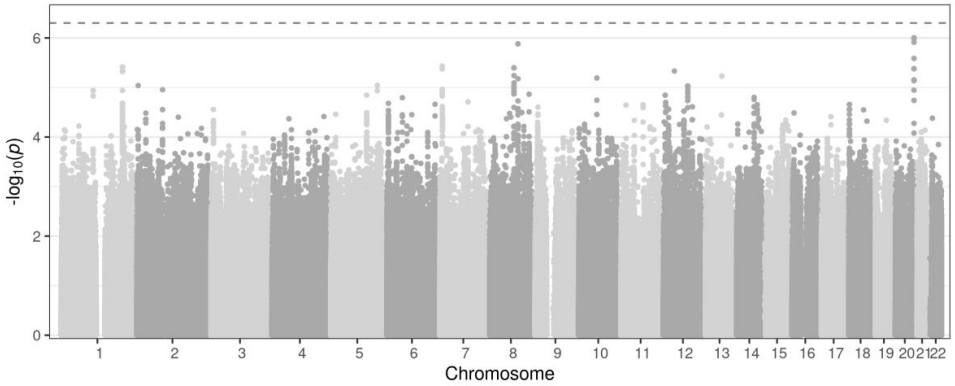


Figure 10: GWAS run using sex as the phenotype. No SNPs met suggestive threshold of $5e-07$ (indicated by the grey dashed line).

Including known covariates can reduce power to detect genetic effects in case-control studies.

Supplementary Information

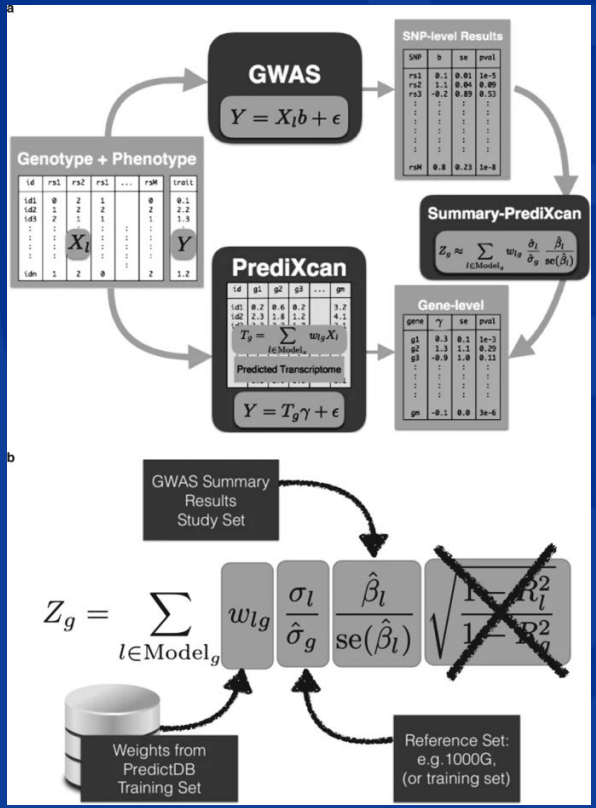
Matti Pirinen, Peter Donnelly and Chris Spencer

University of Pittsburgh | School of Public Health

TWAS

TWAS performed using S-PrediXcan
Uses pre-trained prediction models from GTEx

Caution warranted when interpreting results (de Leeuw, 2023)



ARTICLE

DOI: 10.1038/s41467-018-03621-1 OPEN

Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics

Alvaro N. Barbeira[1], Scott P. Dickinson[1], Rodrigo Bonazzola[1], Jiamao Zheng[1], Heather E. Wheeler[2,3], Jason M. Torres[4], Eric S. Torstenson[5], Kaanan P. Shah[1], Tzintzuni Garcia[6], Todd L. Edwards[7], Eli A. Stahl[8,9], Laura M. Huckins[8,9], GTEx Consortium, Dan L. Nicolae[1], Nancy J. Cox[5] & Hae Kyung Im[1]

University of Pittsburgh | School of Public Health

Gene-Based Test

Gene-based tests performed with mBAT-combo

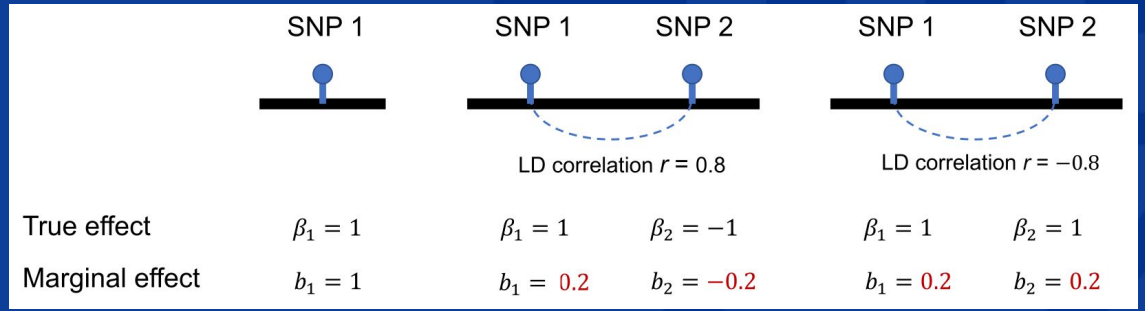
Recently developed (2023)

Shown to be well-powered to detect effects presence of “masking effects”

ARTICLE

mBAT-combo: A more powerful test to detect gene-trait associations from GWAS data

Ang Li,[1] Shouye Liu,[1] Andrew Bakshi,[2] Longda Jiang,[3] Wenhan Chen,[4] Zhili Zheng,[1] Patrick F. Sullivan,[5,6] Peter M. Visscher,[1] Naomi R. Wray,[1,7] Jian Yang,[8,9] and Jian Zeng[1,*]



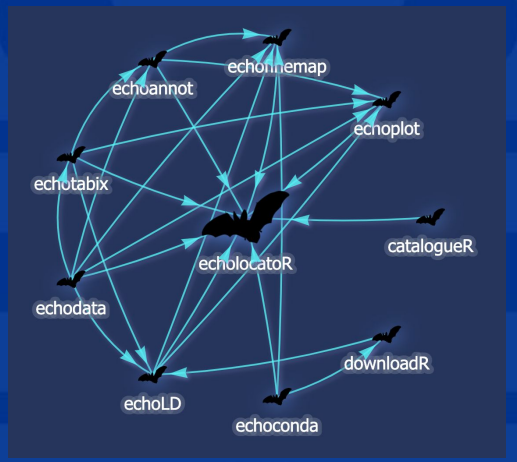
University of Pittsburgh | School of Public Health

Fine-mapping

Fine-mapping performed with echolocatoR package (part of echoverse suite)

Integrates multiple fine-mapping methods (ABF, FINEMAP, SuSiE, PAINTOR)

EAS subset of 1000G (phase 3) used to approximate LD



University of Pittsburgh | School of Public Health

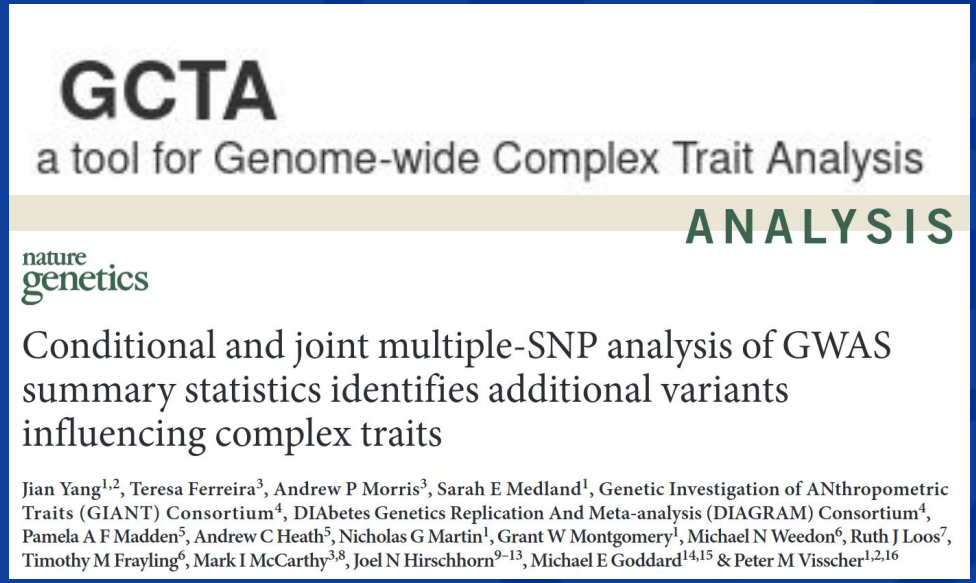
COJO Analysis

Approximates multiple regression in GWAS setting by leveraging local LD

Able to identify secondary signals not discernible through standard (marginal) analysis alone

“Conditional”: effect size of focal SNP adjusted for other SNPs in model

“Joint”: effect size of focal SNP estimated simultaneously with other SNPs in model


GCTA

a tool for Genome-wide Complex Trait Analysis

ANALYSIS

nature genetics

Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits

Jian Yang[1,2], Teresa Ferreira[3], Andrew P Morris[3], Sarah E Medland[1], Genetic Investigation of ANthropometric Traits (GIANT) Consortium[4], DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium[4], Pamela A F Madden[5], Andrew C Heath[5], Nicholas G Martin[1], Grant W Montgomery[1], Michael N Weedon[6], Ruth J Loos[7], Timothy M Frayling[6], Mark I McCarthy[3,8], Joel N Hirschhorn[9–13], Michael E Goddard[14,15] & Peter M Visscher[1,2,16]

University of Pittsburgh | School of Public Health

Outline

Methods

- ✓ Population Structure Analysis
- ✓ GWAS
  - ✓ fastGWA-GLMM
  - ✓ Covariate Justification
- ✓ TWAS
  - ✓ Note on Interpretation
- ✓ Gene-Based Test
  - ✓ m-BAT
  - ✓ “Masking effects”
- ✓ Fine-Mapping
- ✓ COJO
  - ✓ “Conditional”
  - ✓ “Joint”

Results

- ❑ GWAS
  - ❑ Significant SNPs found
  - ❑ Knowns/unknowns
- ❑ TWAS
  - ❑ Confirmatory Results
- ❑ Gene-Based Test
  - ❑ Chromosome 1 region
- ❑ Fine-Mapping
  - ❑ Chromosome 1 region
- ❑ COJO
  - ❑ Functional information

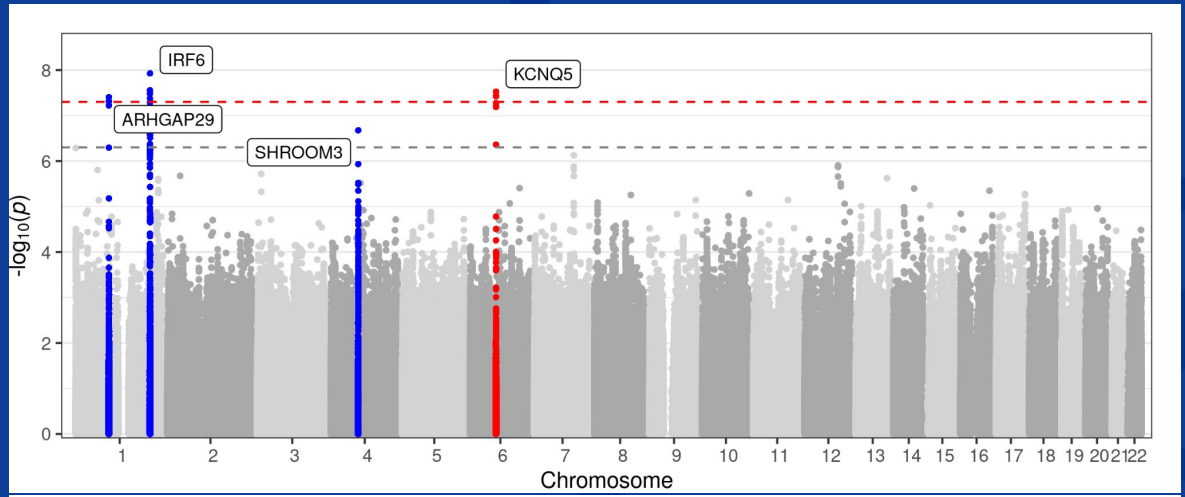
University of Pittsburgh | School of Public Health

GWAS

ARGHAP29 first associated in Beaty et al. (2010). Signal originally thought to be due to *ABCA4*.

Subsequently shown to be *ARGAP29* by Leslie et al. (2012), possible pathway with *IRF6*.

Replicated several times since original identification



Note: Blue represents known loci, red represents novel locus

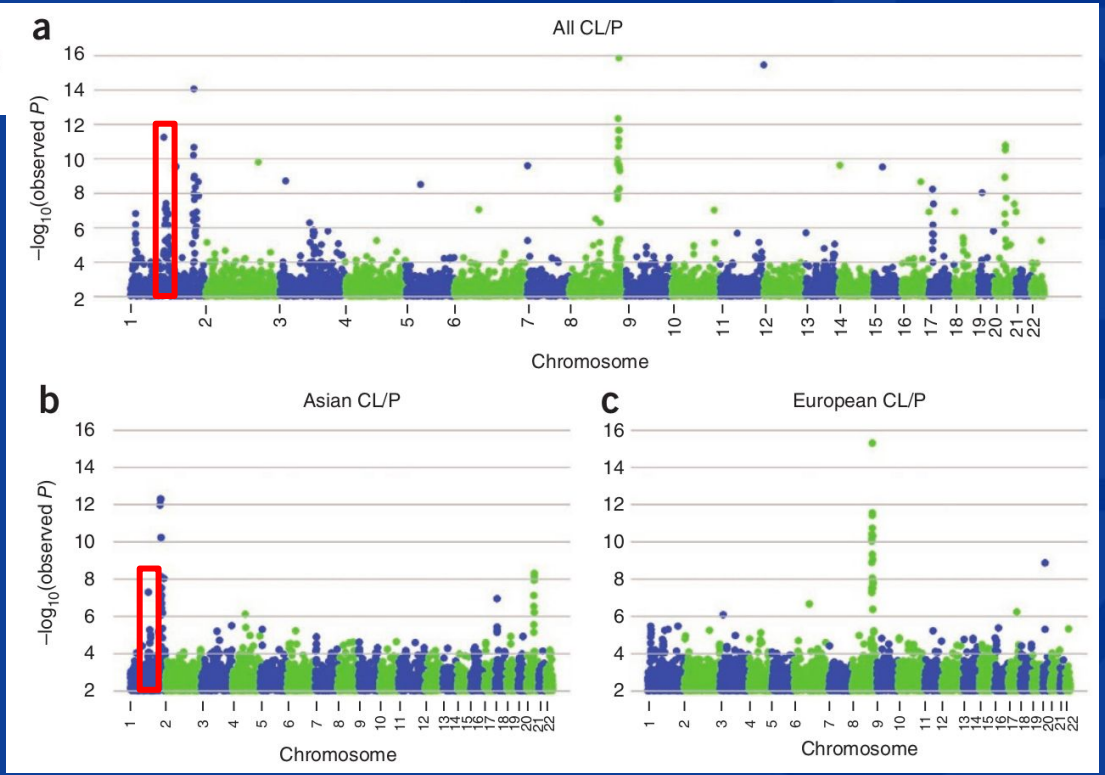
University of Pittsburgh | School of Public Health

ABCA4 (?)

A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near *MAFB* and *ABCA4*

ARGHAP29 first associated in Beaty et al. (2010). Signal originally thought to be due to *ABCA4*.

Subsequently shown to be *ARGAP29* by Leslie et al. (2012), possible pathway with *IRF6*.

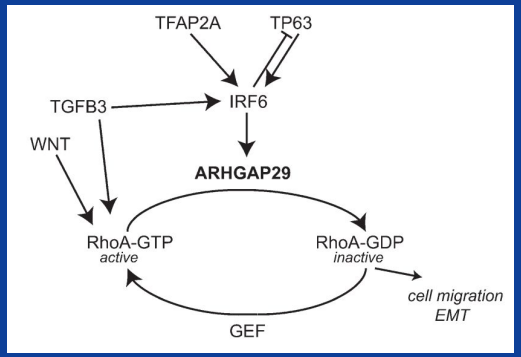


University of Pittsburgh | School of Public Health

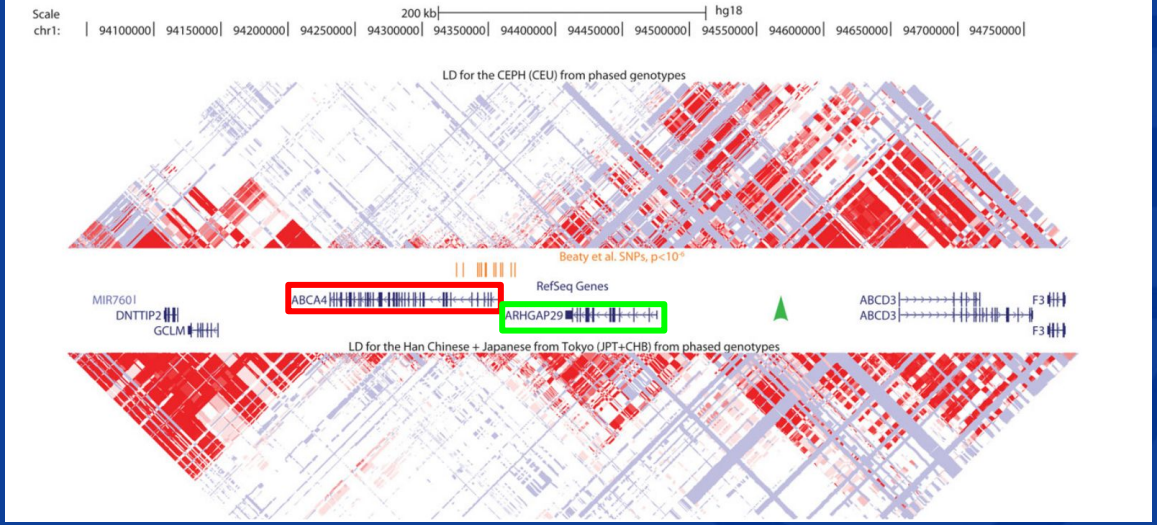
ARGHAP29 (!)

ARGHAP29 first associated in Beaty et al. (2010). Signal originally thought to be due to *ABCA4*.

Subsequently shown to be *ARGAP29* by Leslie et al. (2012), possible pathway with *IRF6*.



Expression and Mutation Analyses Implicate ARHGAP29 as the Etiologic Gene for the Cleft Lip with or without Cleft Palate Locus Identified by Genome-Wide Association on Chromosome 1p22



University of Pittsburgh | School of Public Health

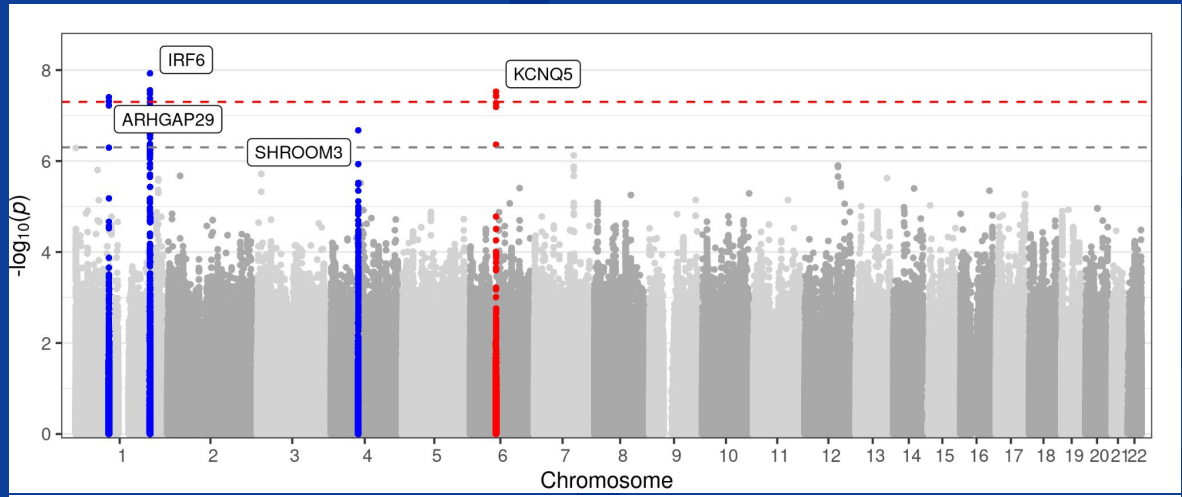
GWAS

IRF6: one of the most robustly-confirmed clefting loci in literature

First identified in 2004

Further solidified via follow-up functional studies

Recent (10Jul2024) preprint on medArxiv



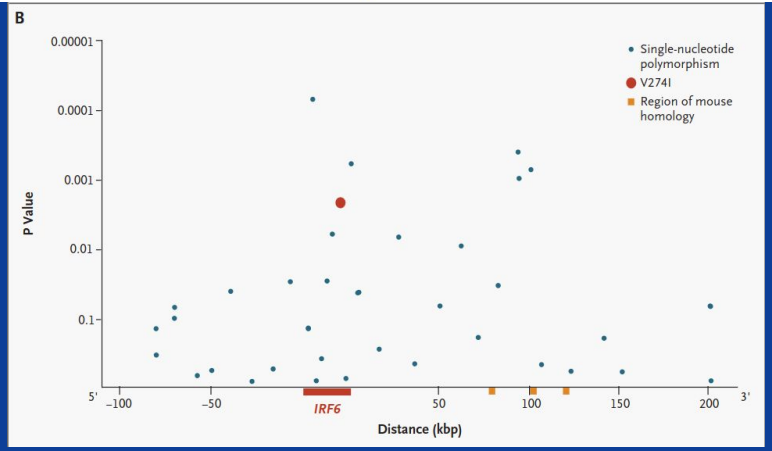
Note: Blue represents known loci, red represents novel locus

University of Pittsburgh | School of Public Health

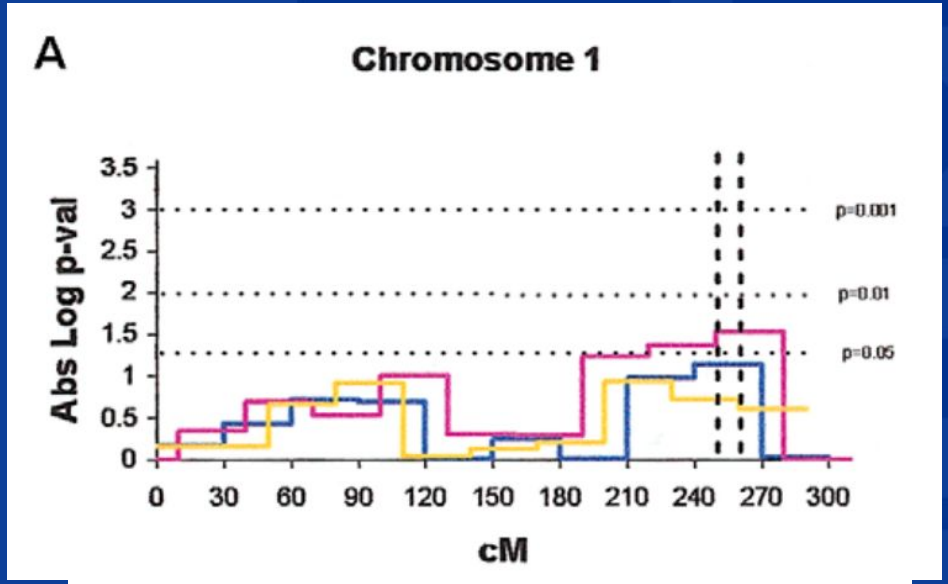
GWAS

ORIGINAL ARTICLE

Interferon Regulatory Factor 6 (*IRF6*) Gene Variants and the Risk of Isolated Cleft Lip or Palate



Zucchero et al. (2004)



Chromosome 1

The 1q32 region is the location for interferon regulatory factor–6 (*IRF6*) that was identified recently as the locus involved in van der Woude syndrome (VDWS

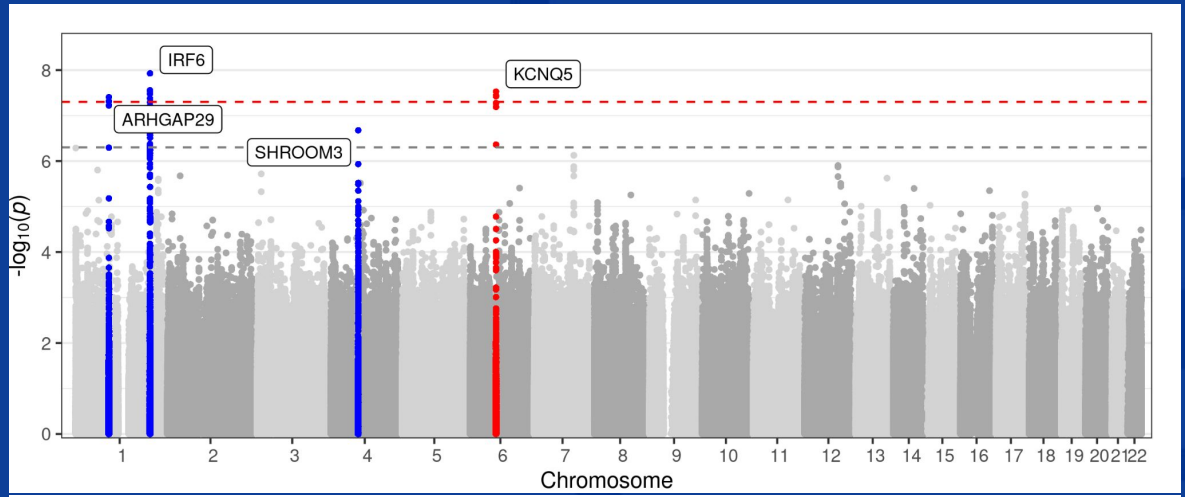
Marazita et al. (2004)

University of Pittsburgh | School of Public Health

GWAS

SHROOM3 nominated as potential associated locus in Leslie et al., 2017

Confirmatory evidence has come to light since then (Diaz et al., 2023 & Deshwar, 2023)



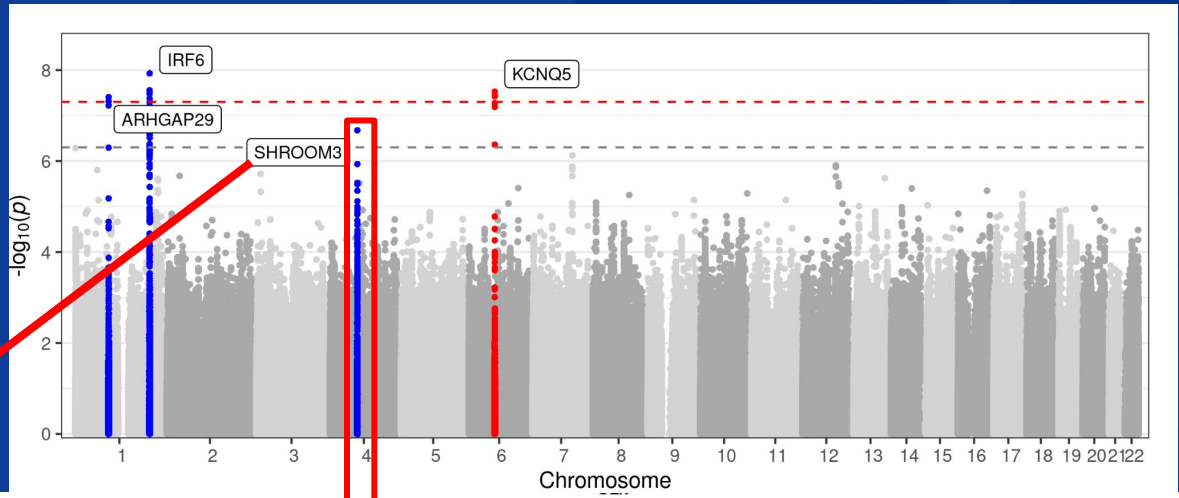
Note: Blue represents known loci, red represents novel locus

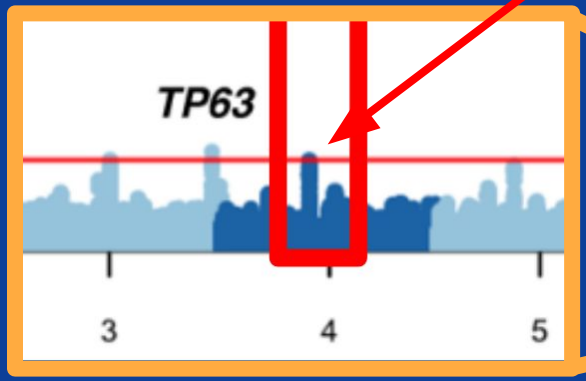
University of Pittsburgh | School of Public Health

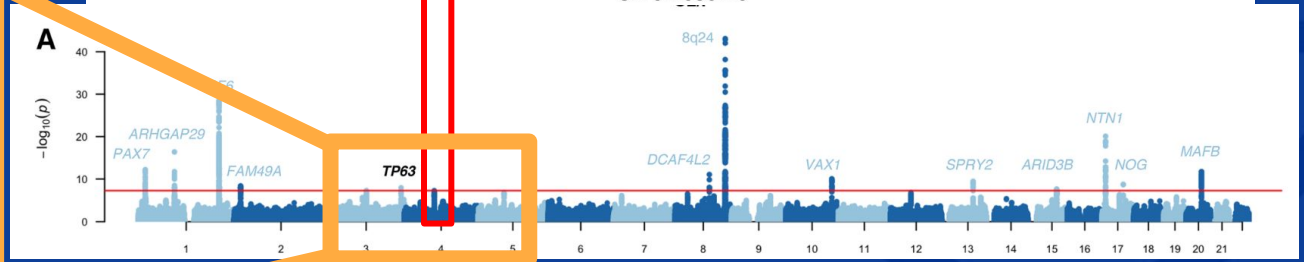
GWAS

SHROOM3 nominated as potential associated locus in Leslie et al., 2017

Confirmatory evidence has come to light since then (Diaz et al., 2023 & Deshwar, 2023)







Leslie et al. (2017)

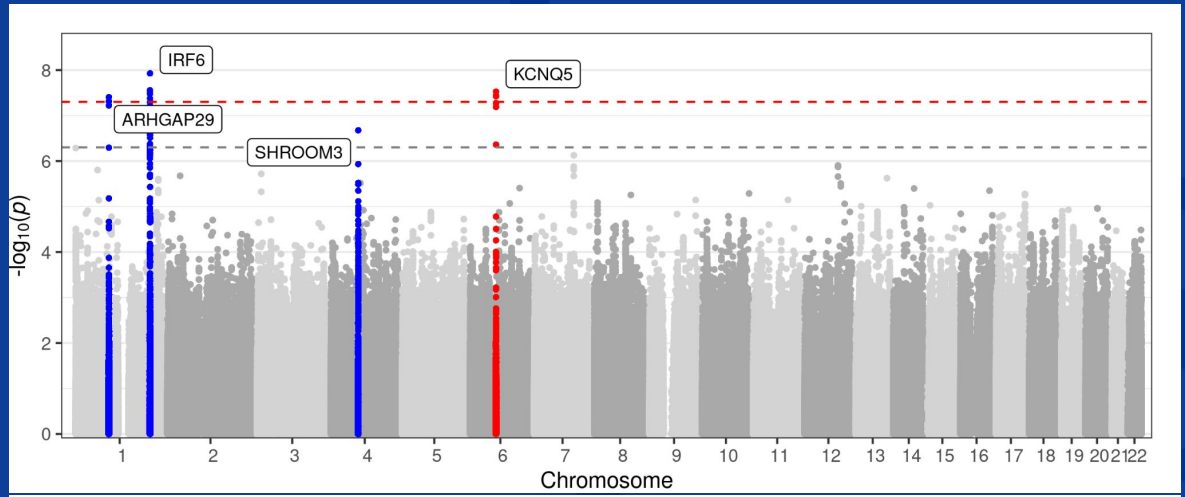
University of Pittsburgh | School of Public Health

GWAS

KNCQ5: putatively novel locus

Codes for potassium voltage-gated channel (subfamily Q member 5)

GWAS Catalog lists 136 associations spanning 95 studies (BMI, height, lung function, SCZ, ocular disorders)



Note: Blue represents known loci, red represents novel locus

University of Pittsburgh | School of Public Health

GWAS

KNCQ5: putatively novel locus

Codes for potassium voltage-gated channel (subfamily Q member 5)

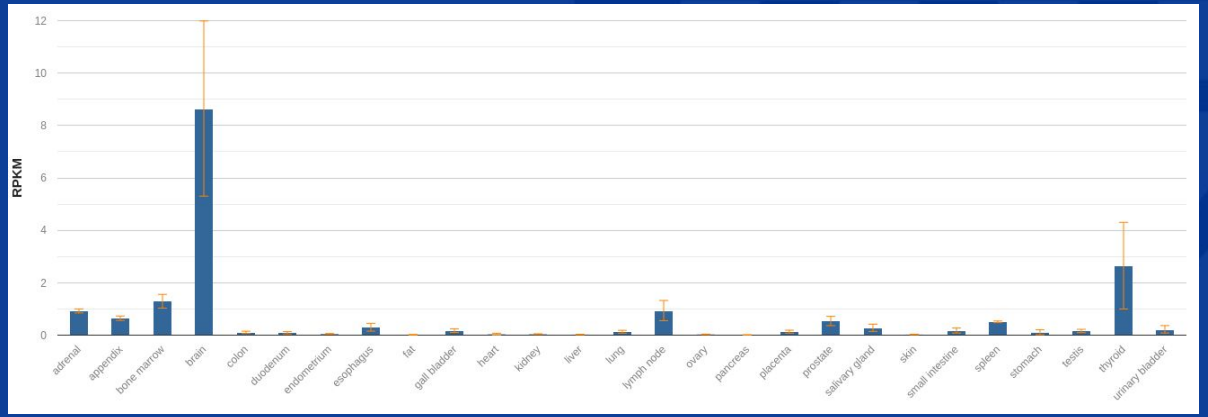
GWAS Catalog lists 136 associations spanning 95 studies (BMI, height, lung function, SCZ, ocular disorders)

KCNQ5, a Novel Potassium Channel Broadly Expressed in Brain, Mediates M-type Currents*

Received for publication, April 17, 2000, and in revised form, May 16, 2000
Published, JBC Papers in Press, May 17, 2000, DOI 10.1074/jbc.M003245200

Björn C. Schroeder, Mirko Hechenberger, Frank Weinreich, Christian Kubisch‡, and Thomas J. Jentsch§

From the Zentrum für Molekulare Neurobiologie Hamburg, Hamburg University, Martinistrasse 85, D-20246 Hamburg, Germany



University of Pittsburgh | School of Public Health

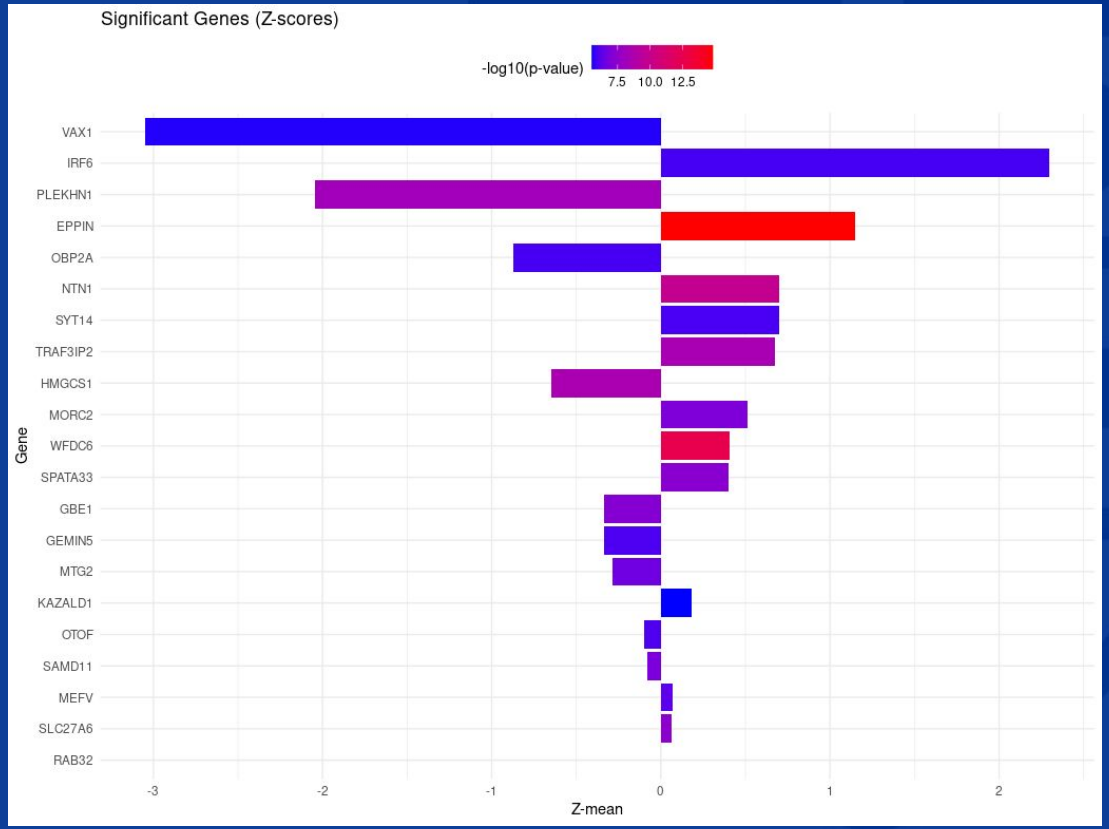
TWAS

Significance threshold of 2.47e-06 chosen to adjust for 20,215 genes tested

21 significant genes

- *VAX1*
- *NTN1*
- *IRF6*
- *SYT14*
- *PLEKHN1*

These five genes among top seven when ranked by absolute value mean Z-score

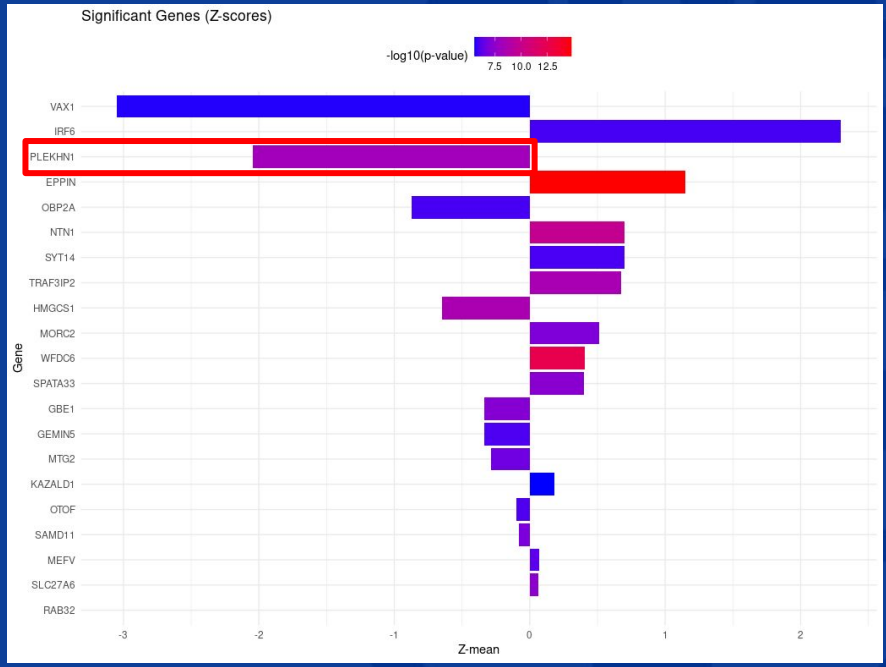


University of Pittsburgh | School of Public Health

TWAS

21 significant genes

- *VAX1*
- *NTN1*
- *IRF6*
- *SYT14*
- ***PLEKHN1***



DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Investigating the Roles of IRF6 in Epithelial Maturation, Craniofacial Development, and Orofacial Cleft Pathogenesis

Li (2018) [unpbulished dissertation]

spatiotemporal gene expression filters applied (Figure 28C), six putative IRF6 transcriptional target genes with putative deleterious *de novo* coding mutations were identified: *WNT11*, *ETV4*, *KEAP1*, *METRN*, *PLEKHN1* and *RAP1GAP*. Whether these coding mutations actually negatively affect the

University of Pittsburgh | School of Public Health

Gene-Based Test

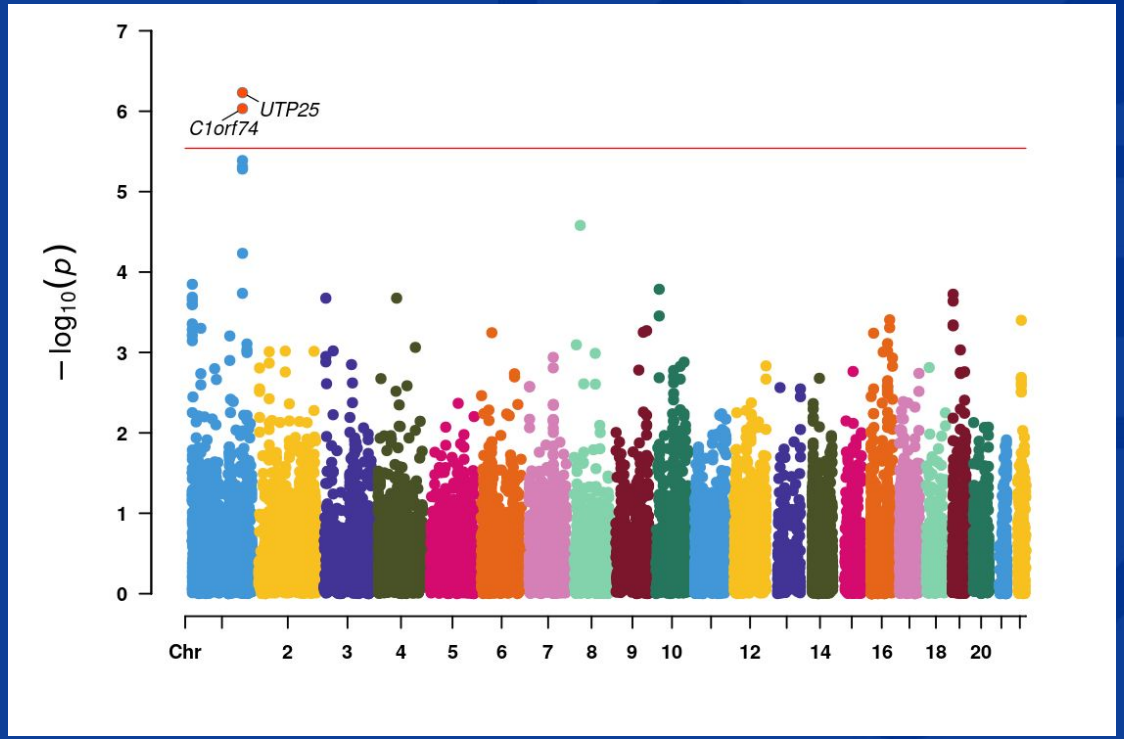
Significance threshold of 2.9e-06 chosen to adjust for 17,271 genes

2 significant genes

- C1orf74
- UTP25

4 potential “masking”

Gene	p-mBAT	p-fastBAT	p-mBATcombo
UTP25	**2.56e-06**	3.32e-07	**5.88e-07**
C1orf74	**4.65e-07**	1.08e-04	**9.27e-07**
TRAF3IP3	**2.07e-06**	1.99e-04	**4.10e-06**
IRF6	**2.53e-06**	8.40e-05	**4.92e-06**

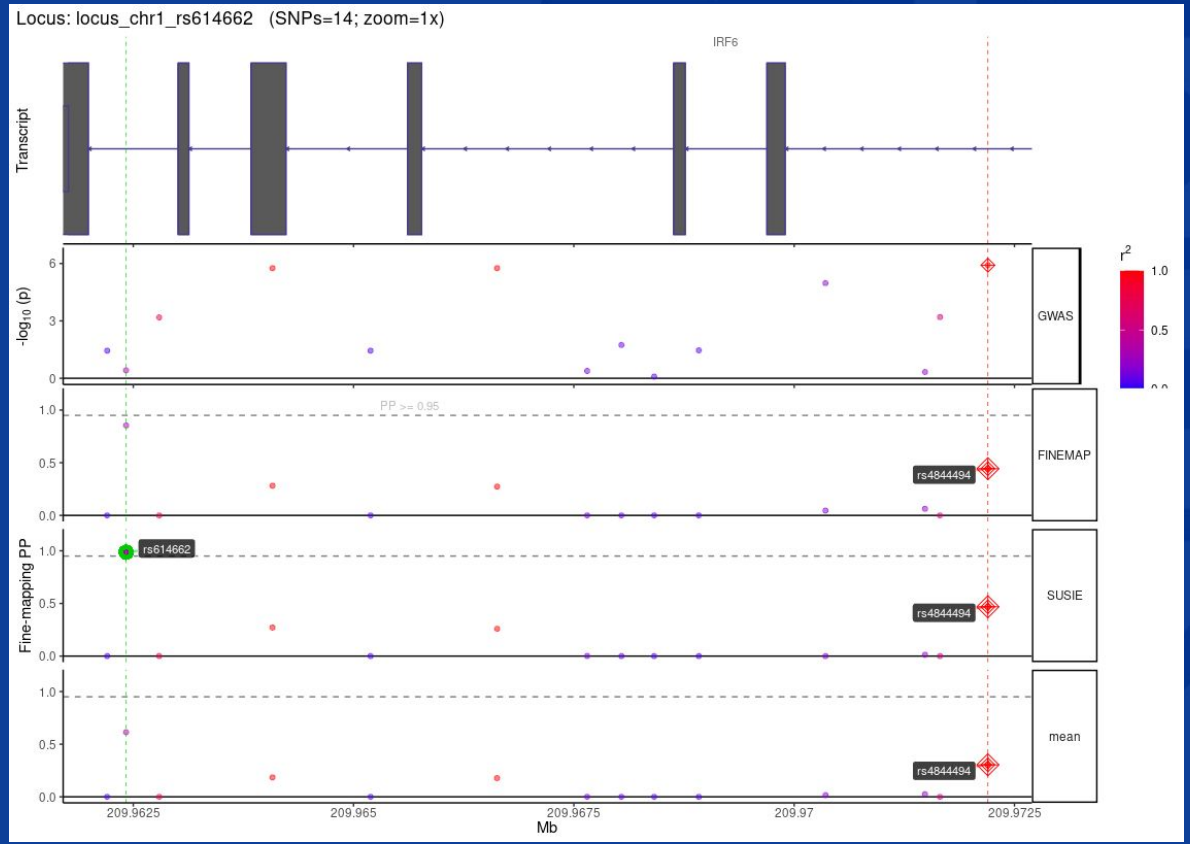


University of Pittsburgh | School of Public Health

Fine-Mapping

Initial analysis revealed interesting signal from rs614662, ranked low in GWAS but PP=1.0

This despite presence of other significant SNPs identified in surrounding region



University of Pittsburgh | School of Public Health

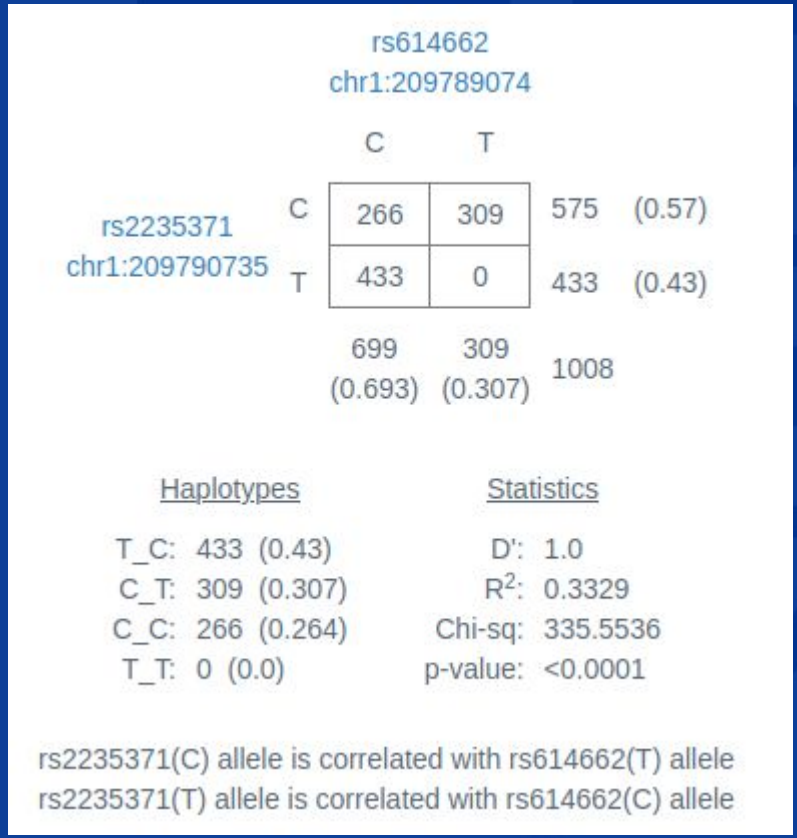
Fine-Mapping

Ultimately deemed artifact of sparse LD matrix construction

Found to be in perfect LD with rs2235371

rs2235371:

- gnomAD CADD: 23.5 (top 0.5% most deleterious substitutions in genome)
- REVEL score: 0.385 (moderate likelihood of pathogenicity)
- phyloP score: 8.90 (highly conserved across species)
- PolyPhen (max) score: 0.758 (possibly damaging to protein function)



University of Pittsburgh | School of Public Health

COJO

"Conditional on single SNP" analysis:

- 500Kb region centered on rs614662 extracted
- Independent analyses run conditioning on
  - rs614662
    - 27 significant SNPs
    - p < 5e-08
  - rs2235371
    - 0 significant SNPs
    - P_min = 0.522

Results consistent with previous interrogation

GCTA

a tool for Genome-wide Complex Trait Analysis

ANALYSIS

nature genetics

Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits

Jian Yang[1,2], Teresa Ferreira[3], Andrew P Morris[3], Sarah E Medland[1], Genetic Investigation of ANthropometric Traits (GIANT) Consortium[4], DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium[4], Pamela A F Madden[5], Andrew C Heath[5], Nicholas G Martin[1], Grant W Montgomery[1], Michael N Weedon[6], Ruth J Loos[7], Timothy M Frayling[6], Mark I McCarthy[3,8], Joel N Hirschhorn[9–13], Michael E Goddard[14,15] & Peter M Visscher[1,2,16]

University of Pittsburgh | School of Public Health

COJO

Conditional and Joint Analysis

- 1MB region centered on midpoint between two SNPs extracted
- COJO parameters:
  - 10Mb window (assumption: SNPs > 10Mb apart in complete LE)
  - Collinearity threshold lowered from 0.9 to 0.7, allowing more SNPs to enter model simultaneously

Joint: only rs4329516 significant (p=1.18e-08)
Conditional: no significant SNPs

Takeaway: rs4329516 possibly relevant, but functional evidence less persuasive than rs2235371

GCTA

a tool for Genome-wide Complex Trait Analysis

ANALYSIS

nature genetics

Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits

Jian Yang[1,2], Teresa Ferreira[3], Andrew P Morris[3], Sarah E Medland[1], Genetic Investigation of ANthropometric Traits (GIANT) Consortium[4], DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium[4], Pamela A F Madden[5], Andrew C Heath[5], Nicholas G Martin[1], Grant W Montgomery[1], Michael N Weedon[6], Ruth J Loos[7], Timothy M Frayling[6], Mark I McCarthy[3,8], Joel N Hirschhorn[9–13], Michael E Goddard[14,15] & Peter M Visscher[1,2,16]

University of Pittsburgh | School of Public Health

Discussion

GWAS

- one novel locus, *KCNQ5*
- Replicated three known loci
  - *ARHGAP29*
  - *IRF6*
  - *SHROOM3*

Gene-Based Test

- Potential masking effects
  - *C1orf74*
  - *TRAF31P3*
  - *IRF6*
  - *UTP25*

TWAS

- Identified several significant genes
- Some previously identified
  - *VAX1*
  - *NTN1*
  - *IRF6*
  - *PLEKHN1*

Fine-Mapping/COJO

- Evidence (tentatively) points to specific SNPs
  - rs4329516
  - rs2235371

University of Pittsburgh | School of Public Health

Further Work

(for Dr. Shaffer's next student)

Unanswered Questions

- What is the functional role of *KCNQ5*?
- What accounts for the remaining unexplained variation?
  - Rare variants
    - Population-specific
    - Family-specific
  - Environment
  - Interaction effects
- What are the causal mechanisms?
- How to make gathered genetic insights actionable?

University of Pittsburgh® | School of Public Health

FIN.