PI: **Marazita, Mary L**	Title: Genetic analysis of quantitative facial variation	
Received: 09/19/2013	FOA: PAR11-210	Council: 01/2014
Competition ID: ADOBE-FORMS-B1	FOA Title: CENTER FOR INHERITED DISEASE RESEARCH (CIDR) HIGH THROUGHPUT GENOTYPING AND SEQUENCING RESOURCE ACCESS (X01)	
1 X01 HG007821-01	Dual:	Accession Number: 3621679
IPF: 2059802	Organization: UNIVERSITY OF PITTSBURGH AT PITTSBURGH	
Former Number:	Department: Oral Biology	
IRG/SRG: ZHG1 SRC (99)	AIDS: N	Expedited: N
Subtotal Direct Costs (excludes consortium F&A) Year 1: 0	Animals: N Humans: Y Clinical Trial: N Current HS Code: 20 HESC: N	New Investigator: N Early Stage Investigator: N
Senior/Key Personnel:	*Organization:*	*Role Category:*
Mary Marazita	University of Pittsburgh	PD/PI
Seth Weinberg Ph.D	University of Pittsburgh	MPI
Eleanor Feingold Ph.D	University of Pittsburgh	MPI
Richard Spritz M.D.	University of Colorado School of Medicine	Co-Investigator

Appendices

Appendi

SF 424 (R&R) APPLICATION FOR FEDERAL ASSISTANCE

15. ESTIMATED PROJECT FUNDING

a. Total Federal Funds Requested	0.00
b. Total Non-Federal Funds	0.00
c. Total Federal & Non-Federal Funds	0.00
d. Estimated Program Income	0.00

16. * IS APPLICATION SUBJECT TO REVIEW BY STATE EXECUTIVE ORDER 12372 PROCESS?

a. YES □ THIS PREAPPLICATION/APPLICATION WAS MADE AVAILABLE TO THE STATE EXECUTIVE ORDER 12372 PROCESS FOR REVIEW ON:

DATE:

b. NO □ PROGRAM IS NOT COVERED BY E.O. 12372; OR

☒ PROGRAM HAS NOT BEEN SELECTED BY STATE FOR REVIEW

17. By signing this application, I certify (1) to the statements contained in the list of certifications* and (2) that the statements herein are true, complete and accurate to the best of my knowledge. I also provide the required assurances * and agree to comply with any resulting terms if I accept an award. I am aware that any false, fictitious. or fraudulent statements or claims may subject me to criminal, civil, or administrative penalities. (U.S. Code, Title 18, Section 1001)

☒ *** I agree**

**** The list of certifications and assurances, or an Internet site where you may obtain this list, is contained in the announcement or agency specific instructions.***

18. SFLLL or other Explanatory Documentation

Add Attachment | Delete Attachment | View Attachment

19. Authorized Representative

Prefix: Mr. * First Name: Allen Middle Name:

* Last Name: DiPalma Suffix:

* Position/Title: Director

* Organization: University of Pittsburgh

Department: Office of Research Division:

* Street1: 123 University Place

Street2:

* City: Pittsburgh County / Parish:

* State: PA: Pennsylvania Province:

* Country: USA: UNITED STATES * ZIP / Postal Code: 15213-2303

* Phone Number: 412 624 7400 Fax Number: 412 624 7409

* Email: offres@offres.pitt.edu

*** Signature of Authorized Representative**	*** Date Signed**
Jacob Stempky	09/19/2013

20. Pre-application Add Attachment | Delete Attachment | View Attachment

424 R&R and PHS-398 Specific Table Of Contents

Page Numbers

SF 424 R&R Face Page 1

Table of Contents 3

Research & Related Other Project Information 4

Project Summary/Abstract (Description) 5

Public Health Relevance Statement (Narrative attachment) 6

Facilities & Other Resources 7

Equipment 14

Research & Related Senior/Key Person 17

Biographical Sketches for each listed Senior/Key Person 20

PHS 398 Specific Cover Page Supplement 34

PHS 398 Specific Research Plan 36

Specific Aims 37

Research Strategy 38

Inclusion Enrollment Report 44

Human Subjects Sections 45

Protection of Human Subjects 45

Women &Minorities 49

Planned Enrollment Table 50

Children 51

Multiple PI Leadership Plan 52

Bibliography & References Cited 53

Consortium/Contractual 55

Resource Sharing Plan 56

Appendix

Number of Attachments in Appendix: 1

RESEARCH & RELATED Other Project Information

OMB Number: 4040-0001
Expiration Date: 6/30/2016

1. Are Human Subjects Involved? ☒ Yes □ No

1.a. If YES to Human Subjects

Is the Project Exempt from Federal regulations? □ Yes ☒ No

If yes, check appropriate exemption number. □ 1 □ 2 □ 3 □ 4 □ 5 □ 6

If no, is the IRB review Pending? □ Yes ☒ No

IRB Approval Date: 08/14/2013

Human Subject Assurance Number: 00006790

2. Are Vertebrate Animals Used? □ Yes ☒ No

2.a. If YES to Vertebrate Animals

Is the IACUC review Pending? □ Yes □ No

IACUC Approval Date:

Animal Welfare Assurance Number:

3. Is proprietary/privileged information included in the application? □ Yes ☒ No

4.a. Does this Project Have an Actual or Potential Impact - positive or negative - on the environment? □ Yes ☒ No

4.b. If yes, please explain:

4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed? □ Yes □ No

4.d. If yes, please explain:

5. Is the research performance site designated, or eligible to be designated, as a historic place? □ Yes ☒ No

5.a. If yes, please explain:

6. Does this project involve activities outside of the United States or partnerships with international collaborators? □ Yes ☒ No

6.a. If yes, identify countries:

6.b. Optional Explanation:

7. Project Summary/Abstract Abstract.pdf Add Attachment Delete Attachment View Attachment

8. Project Narrative Narrative.pdf Add Attachment Delete Attachment View Attachment

9. Bibliography & References Cited Bibliography.pdf Add Attachment Delete Attachment View Attachment

10. Facilities & Other Resources Resources.pdf Add Attachment Delete Attachment View Attachment

11. Equipment Equipment.pdf Add Attachment Delete Attachment View Attachment

12. Other Attachments Add Attachments Delete Attachments View Attachments □

Genetic analysis of quantitative facial variation

PD/PIs: Mary L. Marazita, Ph.D., Seth Weinberg, Ph.D., Eleanor Feingold, PhD.

ABSTRACT

In this proposal we request genotyping to investigate genes underlying quantitative facial variation. The parent grant for this project is U01-DE020078 ("3D Analysis of Normal Facial Variation: Data Repository and Genetics") which is part of the FaceBase consortium (www.FaceBase.org, Hochheiser et al 2011). A core objective of the FaceBase consortium is to facilitate the rapid generation of new data resources that will enable researchers to investigate the genetic factors underlying both normal and abnormal facial development. Fully achieving this objective requires the construction of large minable databases comprised of detailed quantitative information on craniofacial structures as well as genotypic variation. Advances in non-invasive 3D imaging have made it feasible to capture high-resolution, geometrically accurate 3D surface scans of the face in large numbers of individuals. Moreover, with the advent of high-throughput genotyping, it is now possible to marry quantitative phenotypic data derived from 3D images to vast quantities of genetic information. Establishing such a phenotype-genotype database would represent a major step forward in the search for the genetic determinants of facial form and variation. Under the parent grant, the 3D facial image data is now available through FaceBase (known as the Three Dimensional Facial Norms dataset, i.e. TDFN) and the current proposal is requesting the genotyping to augment the image data resource.

The motivation for this project springs from the fact that very little is known about how variation in specific genes relates to the diversity of facial forms commonly observed in humans. Viable candidates for these morphogenes originate from a number of sources: tissue expression studies, animal models with targeted or spontaneous mutations, and genetic syndromes with craniofacial manifestations. Importantly, understanding the genetic basis for normal facial variation also has important implications for health-related research. For example, this work has the potential to shed light on the factors influencing liability to common craniofacial anomalies such as orofacial clefts. There is now ample evidence that certain facial features (e.g., increased midfacial retrusion) characterize individuals genetically at-risk for orofacial clefts (e.g., biological relatives of affected cases). While these predisposing facial features are statistically over-represented in at-risk groups, they are also common in the general population. Since many of the current candidate genes for clefting are thought to play a critical role in facial morphogenesis, variation in these genes may also underlie normal variation in these facial features. These candidate genes, however, probably represent only a small fraction of the total number of loci influencing facial variation. The GWAS approach proposed here has the potential to facilitate the discovery of new loci that may play an important role in both normal and abnormal facial development.

This CIDR proposal has three major goals: (1) to genotype 3,300 individuals from the FaceBase TDFN project; (2) to perform GWAS analyses of both midfacial linear measures and principal components derived from geometric morphometrics; (3) to perform a meta-analysis with data from the FaceBase Tanzanian facial image project.

Genetic analysis of quantitative facial variation
PD/PIs: Mary L. Marazita, Ph.D., Seth Weinberg, Ph.D., Eleanor Feingold, PhD.

NARRATIVE

The goal of this proposal is to investigate the genetics of quantitative variation in facial measures in order to understand genetic factors underlying facial development

Genetic analysis of quantitative facial variation
PD/PIs: Mary L. Marazita, Ph.D., Seth Weinberg, Ph.D., Eleanor Feingold, PhD.

Facilities and Resources

In the following sections we will describe the following specific resources available to the project:

(1) University of Pittsburgh – general resources
- **a. Center for Craniofacial and Dental Genetics (Dr. Marazita)**
- **b. Department of Human Genetics (Dr. Feingold)**

(1) University of Pittsburgh

All of the general resources of the University of Pittsburgjh, a prominent research university, will be available to this project (libraries, computer system, email/online services, technical training, etc).

The University of Pittsburgh Office of Research at the Health Sciences offers support services, including assistance in developing competitive grant applications, for investigators throughout the Graduate School of Public Health, the School of Dental Medicine, the School of Health and Rehabilitation Sciences, the School of Medicine, the School of Nursing, and the School of Pharmacy.

The University of Pittsburgh Office of Clinical Research at the Health Sciences facilitates the process of clinical research, promotes its value, and fosters communication among the various entities engaged in its conduct. The Office of Clinical Research promotes clinical research at the University of Pittsburgh and its affiliated institutions by providing research-related resources for volunteers, sponsors, investigators and research staff.

The Clinical and Translational Science Institute (CTSI) is an initiative that extends across all six schools of the health sciences—and beyond. Through the establishment of 10 core divisions, CTSI is channeling existing resources while also developing an infrastructure to develop and support a wide range of other institutional resources that are all designed to foster clinical and translational research.

The CTSI provides information on the vast resources and assistance available from CTSI to investigators engaged in biomedical research in order to enhance how efficiently and effectively advances might reach individual patients and the population as a whole.
CTSI, through its Research Facilities and Clinical Resources (CRRF) core, supports multiple research centers and networks devoted to advancing clinical and translational research. Researchers can utilize these Clinical and Translational Research Centers (CTRCs) and Research Networks for the provision of facilities, staff, equipment, laboratory testing, and other research resources in both inpatient and outpatient settings.

The Research Conduct and Compliance Office oversees and facilitates the conduct of ethical and regulation-compliant human and animal subject research through an integrated system of research review, audit, and educational programs.

The Institutional Animal Care and Use Committee oversees the University's animal programs, facilities and procedures ensuring the appropriate care, use, and humane treatments of animals being used for research, testing, and education.

The University of Pittsburgh Institutional Review Board (FWA# 00006790) is an appropriately constituted administrative body established to protect the rights and welfare of human subjects (including patients) recruited to participate in research activities.

The Department of Environmental Health & Safety provides resources for workplace health and safety, biological safety, laboratory safety, chemical hygiene, other emergencies, and research support.

The University of Pittsburgh Office of Research serves as both a center of advocacy for research and related activities, and facilitator of the research environment. The functional areas supported by the Office of Research staff include information services, project and proposal development assistance, and grants and contracts administration for pre-award and selected post-award tasks.

The Office of International Services (http://www.ois.pitt.edu/) provides immigration related services to graduate students and postdoctoral scholars to ensure a smooth and accurate process. Additional resources and individual consultations are available to assist with issues such as obtaining driver's license and social security number .

(1a) CCDG Facilities & Other Resources

CCDG General Description

Since 2001, the ***Center for Craniofacial and Dental Genetics (CCDG)***, directed by Dr. Mary Marazita (grant co-PD/PI), has been an active research program within the Department of Oral Biology in University of Pittsburgh School of Dental Medicine (SDM). The major research focus of the CCDG is to combine molecular genetic and statistical approaches to identify genes contributing to complex human phenotypes, primarily those increasing risk for craniofacial and dental disorders. CCDG faculty members conduct molecular studies of biochemical pathways involved in the development and function of the orofacial complex. Major projects include: 1) investigating genetic contributions to orofacial clefting—with an emphasis on related sub-clinical phenotypes—in families from multiple national and international sites (OFC); 2) studying the genetic, microbial, and behavioral risk factors for early childhood caries and other oral health pathologies in Appalachia (COHRA); and 3) serving as the coordinating hub for the NIDCR's FaceBase consortium. The CCDG develops computational methods to improve gene identification efforts, and also provides secondary genetic data analysis to collaborators around the world. The CCDG is one of the major research strengths of the University of Pittsburgh School of Dental Medicine (SDM), responsible for about 80% of all research funding through the SDM. Center faculty are encouraged and actively supported in efforts to expand the Center's research program.

CCDG Physical Space

Most of the CCDG faculty are located at the Bridgeside Point #1 Building in Pittsburgh, which is less than a mile from the main Health Sciences Oakland campus of the University of Pittsburgh. With approximately 19,000 square feet of dedicated office and computing space on two floors, the Bridgeside Point location houses 36 private offices, 27 cubicles, a fully equipped conference room for 40 people, a 350 square foot wet lab, two kitchens and a lunch room, and several small common areas. There is a dedicated research area for private consultation and performance of study protocols with research participants. CCDG investigators at Bridgeside Point have direct access to two administrative specialists. Two CCDG faculty are housed in 2,800 square feet of wet-lab and office space at the School of Dental Medicine in Salk Hall, which is part of the Oakland campus of the University of Pittsburgh.

CCDG Computer Resources

1. **Datacenter Infrastructure** The CCDG servers exist in a VMware virtual datacenter infrastructure. The physical servers are connected to the University of Pittsburgh's network via a firewalled secure server zone that only allows our workstation firewall zone to access it. Accounts and permissions to file server shares are allowed through the University-maintained single sign-on resource based on Microsoft's Active Directory framework. No local accounts on machines exist that allow user access. Outside study collaborators have limited access to drop files in designated separate areas through the use of University-

provided resource accounts that have limited functionality. Permissions are reviewed to make sure that currently active staff and faculty have access to files for their individual studies and no others. Physical hosts for the center's servers are maintained off-site at the University's data center at RIDC Park in Blawnox, PA. The datacenter is maintained by the University of Pittsburgh Computer Services and Systems Development and our hosts and virtual servers are monitored 24x7 and fully backed-up.

The CCDG's hosted servers provide file sharing, application serving, and databases to center faculty and staff. The center runs Progeny Lab 8 database server (a proprietary genetics lab solution that tracks samples, families, and genomics), Teleforms Workgroup v10.1 (a proprietary paper forms/scanned OCR based data acquisition solution), and MS-SQL server for database applications.

The center's virtual workstation environment exists on a VMware VSphere 4.1 server platform at Bridgeside Point. The VMware server cluster consists of two Dell PowerEdge R710s, ESX 4.1 hosts, each with 96GB of memory, two Dell PowerEdge R610s each with 12GB of memory for VCenter4.1 server and Symantec Backup Exec server, a directly attached DELL PowerVault MD1000 with 8TB of storage space for direct disk to disk backup, Dell Power Vault 120T tape library with a 16 LTO2 tape cartridge capacity, EqualLogic 6000e iSCSI SAN with 8TB of usable space, and a Data Robotics Drobo Elite iSCSI SAN with 8TB of usable space. The CCDG supports over 40 Windows 7 virtual workstations for local and remote access connections through the above VSphere 4.1 infrastructure using VMware View 4.6. At the center we deploy 16 Wyse P-20 terminals with built-in VMware View client for virtual workstations and 6 Ubuntu kiosk terminals with VMware View client for virtual workstations. In addition we have a secure dropbox server for remote site data uploads and downloads that is virtualized on our VSphere 4.1 environment. All necessary firewall rule sets are in place so that only physical machines on the center's workstation network or authorized accounts via the secure VLAN for the department can access the View connection server and be granted account-level authorized access to a virtual desktop.

The CCDG's virtual development and test environment for the FaceBase project exists on a VMware VSphere 5.0 platform, also at Bridgeside Point. The VMware server cluster consists of two Dell PowerEdge R620s, ESXi 5.0 hosts, each with 192GB of memory and a directly attached Dell PowerVault MD3200 SAS storage array with 20TB of usable storage space. The environment includes a virtual VCenter 5.0 controller/server and the capacity for 20-30 virtual servers using CentOS, Ubuntu, and Windows server OSes. This allows developers and systems analysts a prototyping and testing environment separate from our production environment, which exists at the University's Network Operations Center (NOC) in Blawnox, PA. Maintenance, management, upgrades, and service is all performed locally by a dedicated staff of systems analysts and software developers specific to CCDG operations.

The center has a number of physical workstations for the various tasks of data collection, analysis and reporting. In addition, there are a small number of physical servers for specific local needs of data storage and replication. The CCDG maintains 15 Apple 27" iMac workstations for video editing, software development, data analysis and reporting, and 8 Mac Pro tower workstations for statistical analysis, numerical methods simulations, video editing (Final Cut Professional), and general computing. There are 2 Apple Mac Mini servers for software installation and general file sharing. Our center has 15 MacBooks available to staff and faculty for remote access and travel purposes. Our PC workstations consist of 17 Dell Precision workstations running either Red Hat Enterprise workstation or Windows XP and Windows 7 64-bit, and we also have 14 other Dell general-purpose workstations for OCR scanning data entry and general office tasks running windows operating system and utilizing the MS-Office application suite. All of our windows workstations (virtual or physical) are part of our active directory and group policy schemas to help us better manage security and desktop functionality.

2. **CCDG Mathematical Modeling and Computation (MMC) Lab Linux Cluster (INDY, Manika Govil, Director):** Indy, MMC lab's 21-node compute cluster, is designed to allow development, testing, and application of statistical genetic methods requiring the use of high speed computing. The specifications for this cluster were developed by Dr. Manika Govil, Director MMC lab, in consultation with collaborators at the Battelle Center for Mathematical Medicine, Columbus, Ohio. Indy is housed in a controlled-environment

facility at the University of Pittsburgh's Network Operations Center. The NOC also provides essential system monitoring and backup support for the cluster.

Indy is installed with CentOS v6.x operating system (an enterprise class Linux distribution) and is ready to compile and run parallel programs using the OpenMP, MPI and MPI-2 libraries. Sun GridEngine software is used for process allocation. The cluster has both the GCC and Intel suites of compilation tools, as well as a growing list of statistical genetic software available for use in analyses of genetic data.

The cluster is equipped with a hi-speed Infiniband network allowing for 7Gbps IP networking between all of the nodes. The cluster runs several performance monitoring packages including Ganglia and Nagios. Indy currently comprises 19 compute nodes, 1 head node and 1 web server node.

All 19 compute nodes on Indy are equipped with 80GB SATA 2.5" solid state drive to reduce I/O time during computation. Of these 19 nodes, 4 compute nodes each have 2 8-core Intel Xeon E5-2670 2.6GHz processors with 20MB cache, and 256GB of memory. The remaining 15 compute nodes have 2 6-Core Intel Xeon X5670 2.93GHz processors with 12MB cache. Memory of these 15 nodes ranges from 24GB to 144GB with the following distribution: 10 nodes with 24GB, 2 nodes with 48GB, 2 nodes with 96GB, and 1 compute node with 144GB memory.

The head node provides storage for and access to the cluster and is configured with 2 Intel 8-Core Intel Xeon E5-2650 2.0GHz processors with 20MB cache, a 120GB SATA 2.5" solid state drive for boot and operating system and 8 3TB SAS Hard Drives with 64MB cache, effectively providing approximately 12TB storage (with RAID 5). The head node is further equipped with a fiber-channel HBA which connects to the NOC's storage fabric, including its enterprise tape backup facilities.

The web server node is configured with 2 Intel Quad Core Xeon E5620 2.40GHz chips with 12MB cache, and 4 2TB SATAII Enterprise Hard Drives with 64MB cache, effectively providing approximately 5TB storage (with RAID 5). Similar to the head node, the web server node is also equipped with a fiber-channel HBA which connects to the NOC's storage fabric, including its enterprise tape backup facilities.

Indy is available as a resource to the faculty and compute staff at the Center for Craniofacial and Dental Genetics for their ongoing statistical genetic research requiring the use of high speed computing.

3. **CCDG 3D Facial Imaging and Morphometrics Lab (Seth Weinberg, Director)** The CCDG Imaging and Morphometrics Lab is located within the main Bridgeside Point research facility. The 250 square-foot lab is equipped with four identical Dell PC precision workstations with 27" ultra high-resolution monitors and two 27" Mac workstations. Each workstation is dedicated to biomedical image processing and analysis, with access to high-powered imaging software packages such as 3dMDvultus (for 3D facial surface analysis), Amira (for CT and microCT analysis), and a number of other morphometrics and visualization programs.

CCDG Laboratory Resources

(1) **Vieira Laboratory** Dr. Vieira's research group is located on the 6th floor of Salk Hall at the University of Pittsburgh School Of Dental Medicine. The Vieira laboratory occupies 1000 square feet. The lab has all of the routine basic laboratory equipment including microcentrifuges, scales, microscopes for immunehistochemistry and surface microhardness testing, -80OC freezers, -20OC freezers, refrigerators, bacterial shakers/incubators, a spectrophotometer, and PCR machines, as well as real time PCR, with additional shared laboratory space, including a dedicated tissue culture room with incubators and three tissue culture hoods. Shared dedicated rooms include a radioactivity room, darkroom, walk-in cold room, and a PCR room. The Oral Biology Department supplies dishwashing, Millipore water, and an autoclaving facility. There is ready access to a variety of core facilities, including a DNA sequencing core, oligonucleotide core, protein chemistry core and tissue culture core. The group has 15 personal computers, and full access to the large-scale computational power of the CCDG.

(2) **Wendell Laboratory** Dr. Wendell's lab is located on the 6th floor of Salk Hall at the University of Pittsburgh School Of Dental Medicine. It consists of independent laboratory space totaling 400 sq. ft. and access to the

resources at the Vieira Laboratory, with additional shared laboratory space, including a dedicated tissue culture room with incubators and three tissue culture hoods. Major equipment available include an Eppendorf 96 well PCR machine and 2 dedicated heap filtered PCR workstations, high speed and ultra-centrifuges, and a Brady LabExpert label printer with associated barcode scanners. Shared dedicated rooms include a radioactivity room, darkroom, walk-in cold room, and a PCR room. The Oral Biology Department supplies dishwashing, Millipore water, and an autoclaving facility.

(3) **Bridgeside Point Laboratory** The CCDG maintains a 350 sq. ft. wet lab at Bridgeside Point. The lab has a refrigerator, -20^{o}C freezer, and two 26 cu. Ft. -80^{O}C freezers for long-term sample storage (uninterrupted power, emergency power backup, online alarm systems, CO_2 backup tanks). The lab has a bench-top autoclave for sterilization of dental supplies, and a computer to monitor the freezers and access the sample databases and logs.

CCDG Clinical Research Resources

1. **Bridgeside Point** The dedicated research area on the fourth floor of Bridgeside Point includes a waiting area, two cubicles with dental chairs and supplies, a cubicle with a clinic table and a portable ultrasound machine, and four private offices that serve multiple purposes—obtaining informed consents, performing research interviews, gathering 3D and video images, collecting saliva samples. The research area is appropriate for children, including babies as young as one-month in age. Changing facilities are available, and small tables, games, and toys are provided. A kitchen is next to the waiting area, and restrooms are down the hall.

2. **School of Dental Medicine, Salk Hall** The clinic of the School of Dental Medicine is equipped with 263 operatories for general and specialized dental treatment, which have been recently renovated into semi-private treatment rooms. The collection of biological specimens for the Dental Registry and DNA Repository occurs in this clinical setting. The Continuing Dental Education Suite on the second floor of Salk Hall includes a lecture hall, six semiprivate operatories, and a waiting room. It is available for training and calibration of dental research examiners..

CCDG Office Space and Resources

1. **Bridgeside Point** Office resources include centralized business centers on each floor, equipped with high-speed black & white printer/copiers, high resolution color printers, and scanning stations. The CCDG maintains a large format printer capable of 36” or 42” wide poster prints and 2 Xerox Color Phaser printers. There are 4 high output b & w laser printers available to center faculty, staff, and students. The fax machine is on the fifth floor. Additional printers are available to the administrative staff, and for specialized purposes, e.g., printing sample barcode labels.

2. **School of Dental Medicine, Salk Hall** CCDG faculty also have complete access to the common administrative, research, and clinical resources of the Department of Oral Biology in the SDM at Salk Hall. Dr. Vieira has five offices (800 square feet) for research and clerical support staff. Dr. Wendell has a fully equipped 200 sq. ft. office directly across from laboratory and secretarial support. He has access to 2 Dell desktop computers and an HP Officejet printer. Laserjet B/W and Color printers are networked in an adjacent room. All computers are equipped with the usual host of software, such as MS Office, Adobe Acrobat, Endnote, as well specialty software for SNP design and analysis work (ABI SNP browser, ABI SDS (Sequence Detection Systems), Sigma Plot, ChemBioOffice, Haploview, Vector NTI).

(1b) Department of Human Genetics (Dr. Feingold)

1. Additional computational resources available to the CCDG through the Department of Human Genetics include:

The Center for Computational Genetics at GSPH maintains a computational grid which is optimized for high-throughput genetic and genomic analytic projects. This resource is composed of 64 computational cores with 2GB of RAM allocated to each core. Process and resource allocation is managed using the Sun Grid Engine suite. Local (temporary) storage on the compute nodes is in excess of 300Gb. Users home directories are maintained on a 4Tb RAID5 NAS unit, which is backed up via private network to a 9Tb SAN disk array. All of these machines are secured behind a firewall, which allows for only encrypted communication from the external network (using SSH). All extraneous services on the firewall are turned off, including email, file sharing, printing, etc., so that the machine is as secure as possible. All user accounts are password-restricted - with strong password policies dictating the content of passwords and requiring password rotation - and all data on the grid is restricted using access-control lists so that only appropriate project members have access to a particular projects' data. Various software packages are available covering a wide range of genetic and genomic analytic needs, and all machines are additional enabled for interprocess communication using standard message passing frameworks (mpi).

Center for Simulation and Modeling: The University of Pittsburgh's Center for Simulation and Modeling (SAM) is the premier shared high-performance computating (HPC) facility in the University community, and represents investments of hardware and human capital from several University Schools/Departments. The Center also serves as a collaboration portal, having assembled a group of more than 50 collaborators from across the University who are engaged in computational research in Chemistry, Biology, Physics, Astronomy, Mathematics, Computer Science, Economics and several of the departments in the Swanson School of Engineering, as well as faculty from the Schools of Public Health, Medicine, and the Graduate School of Public and International Affairs. The Center employs full-time research faculty whose expertise cover a wide range of areas in HPC and academic research, including parallel programming for distributed, shared memory, and graphical processing units (GPUs), high-throughput data-intensive computing, and various areas of theoretical and computational science and engineering. The SAM faculty are responsible for preparing training and educational material, teaching, cluster user support and consulting, and focused software development and research support for various projects at Pitt. SAM provides user support, training, and project management services on a continual basis through web 2.0 based platforms (http://core.sam.pitt.edu and http://collab.sam.pitt.edu), as well as organizing year round workshops and training sessions on cluster usage, parallel programming, and various topics in HPC based research. The Center also acts as liaison for national computational resources, through partnerships with the Pittsburgh Supercomputing Center and the NSF/XSEDE Campus Champions program.

SAM provides in-house HPC resources allocated for shared usage free of charge for campus researchers. Computational resources consist of a heterogeneous grid/cluster comprised of 200 8-core Intel Westmere, 45 12-core Intel Nehalem, and 23 48-core AMD Magny-Cours compute nodes, adding up to a total of 3244 computation-only CPU cores, with a maximum of 128GB per node shared memory. Several Nehalem nodes have general purpose NVIDIA GPU accelerator cards, for a total of 16 GPU cards comprising 5504 GPU cores. Most nodes are connected via a fast Infiniband low latency network fabric. Process and resource allocation is managed using the PBS/Moab suite. Local (temporary) storage on the compute nodes is typically 1-3TB. Users home directories are maintained on a 40TB RAID5 NAS unit, with a redundant array providing online snapshots/backup. All of these machines are secured behind the University firewall, which allows for only encrypted communication from the external network (using SSH). All extraneous services are controlled at the firewall, including email, file sharing, printing, etc., so that the cluster is as secure as possible. All user accounts are password-restricted - with strong password policies dictating the content of passwords and requiring password rotation - and all data on the cluster is restricted using access-control lists so that only appropriate project members have access to a particular projects' data. Also available is PittGRID, a (Condor-based) grid computing platform which recycles unused CPU cycles across campus and makes them available for research.

SAM hardware resources are located in the University of Pittsburgh's Network Operations Center (NOC), which provides 9680 square feet of dedicated data center floor space. The NOC is operated as a secure environment with both external and internal video surveillance. The location of the facility is not made public, and physical access to the building is restricted by a secure access card system. There is a 10Gbit, high-bandwidth, fiber-optic dedicated network connection that ties the facilities to the main campus area (the NOC also maintains internet connectivity to the Pittsburgh Supercomputing Center (PSC) and Internet2). A specialized in-row, refrigerant-based cooling infrastructure is dedicated for support of HPC clusters. The facilities are protected by a preemptive fire detection and chemical-based suppression system. Power is supplied via two separate utility power feeds. There are redundant uninterruptable power supply (UPS) units and redundant diesel generators with 4,000 gallons of diesel fuel stored on site. Power is monitored in real time by systems that provide both visual and audio alarms. Switching between power feeds, UPSs, and the diesel generators is fully automated. All server racks are equipped with rack power distribution units, and redundant, monitored power feeds and switches. A power management utility provides power and cooling trend analysis. Network engineers staff the facility 24 hours a day, seven days a week.

Next-generation sequencing at SAM: Computational capacity for next-generation sequencing applications at the University of Pittsburgh is growing daily, due to the increased interest in the methods on campus. Primary processing of sequence information (the conversion of raw images and signal intensities into basecalls and quality scores) is typically done on dedicated computer clusters associated with each next-generation sequencer. Secondary analysis of the data (assembly/alignment, annotation, and variant calling) typically occurs at much larger HPC centers, such as at the SAM. Sequence assembly pipelines have been constructed to use the HPC resources housed in SAM, utilizing a number of different alignment tools (BWA, Bowtie, BFAST, barraCUDA, Novoalign, and the CLCbio tools) and post-processing tools (GATK, samtools, vcftools, bedtools, PICARD, VAAST, SiFT, etc.). Assembly of whole genome (human) data sets using a total of 48 cores (10 sequential processes each utilizing an entire Magny-Cours node, or 400 processes each utilizing 4 cores, with 12 processes running concurrently) can be completed in less than 15 hours. Exome sequencing data sets can be aligned to the reference human genome using a much more reasonable computing allocation (8 cores) in roughly 5 hours per individual. Post-processing and variant calling pipelines (which include local realignment and quality-score recalibration for better variant calling, and concordance analysis over multiple variant calling algorithms, as well as annotation of the resulting variant calls) typically occupy 1-8 cores (depending on the analysis) for an average of 12-15 hours per sample.

Genetic analysis of quantitative facial variation
PD/PIs: Mary L. Marazita, Ph.D., Seth Weinberg, Ph.D., Eleanor Feingold, PhD.

MAJOR EQUIPMENT

In the following sections we will describe the major equipment available to the project:

Equipment	**#**
COMPUTATIONAL	
VMware VSphere 4.1 server platform (CCDG)	
Dell PowerEdge R710, ESX 4.1 host, 96GB memory	2
Dell PowerEdge R610, 12GB memory	2
Dell PowerVault MD1000, 8TB storage space	1
Dell Power Vault 120T tape library, 16 LTO2 tape cartridge capacity	1
EqualLogic 6000e iSCSI SAN, 8TB usable space	1
Data Robotics Drobo Elite iSCSI SAN, 8TB usable space	1
Workstations supported by VMware VSphere 4.1 server platform	
Windows 7 Virtual Workstation	>40
Wyse P-20 Terminal	16
Ubuntu Kiosk Terminal	6
Virtual Servers provided by VMware VSphere 4.0 infrastructure (CCDG)	
Windows Server 2003, 4GB memory, 6 TB SAN Storage	1
Windows Server 2003, MS-SQL 2003 server 2GB memory, 500 GB SAN Storage	1
Windows Server 2003 Teleforms applications 2GB memory, 500 GB SAN Storage	1
Linux Server CentOS 5 statistical applications 2GB memory, 500 GB SAN Storage	1
VMware VSphere 5.0 platform (FaceBase)	
Dell PowerEdge R620, ESXi 5.0 host, 192GB memory	2
Dell PowerVault MD3200 SAS storage array, 20TB usable storage space	1
Workstations supported by Vmware VSphere 5.0 platform	
CentOS, Ubuntu, and Windows server Oses Virtual Server	20-30 capacity
Equipment	**#**

VMware VSphere 4.0 platform (FaceBase)	
Dell PowerEdge R710, ESX host, 96GB memory	2
Dell PowerEdge R610, VCenter server, 12 GB memory	1
SAN storage available to Vsphere cluster: 20 TB	
Workstations supported by 4.0 platform	
CentOS 5 Virtual servers (for FaceBase.org)	20
MMC Lab Linux Cluster (Indy)	
Head Node [2 8-Core Intel Xeon E5-2650 2.0GHz processors (each 20MB cache), a 120GB SATA 2.5” solid state drive, 8 3TB SAS hard drives (each 64MB cache)]	1
Web Server Node [2 Intel Quad Core Xeon E5620 2.40GHz chips (each 12MB cache), 4 2TB SATAII Enterprise hard drives (each 64MB cache)]	1
Compute Node [2 8-core Intel Xeon E5-2670 2.6GHz processors (each 20MB cache, 256GB memory)]	4
Compute Node [2 6-Core Intel Xeon X5670 2.93GHz processors (each 12MB cache, memory varies)]	15
3D Facial Imaging and Morphometrics Lab	
Dell PC Precision Workstation, 27” Ultra-High-Resolution Monitor	4
Mac Workstation, 27"	2
Other	
Apple 27” iMac Workstation	15
Mac Pro Tower Workstation	8
Apple Mac Mini-Server	2
MacBook	15
Dell Precision Workstation	17
Dell General-Purpose Workstation	14
Lab Personal Computer	12
Dell Optiplex GX620 Computer	1
HHP Dolphin 9500 Mobile Computer	1
Phenotyping Equipment	**#**
Cameras	
3D Camera, 3DMD Face System	3
3D Imaging System, Genex	1
Digital Video Camera, Canon 7D	3
Video Camera, Sony DCR-TR V70	1
Extraoral Camera, Canon Rebel (with macro lens and ring flash)	4
Ultrasound Systems	

Portable Ultrasound Machine, Sonosite M Turbo	7
Ultrasound Transducer (7.5L45), Siemens	1
Multisystem Digital Converter 2 (for converting PALS ultrasound footage to NTSB), Markertek Video supply	1
Ultrasound Probe, SP10-16 Linear, GE Medical	1
Other	
MinXray (220V) and Tripod Stand, Aseptico	1
Nasometer, Kay Pentax	2
Spreading Calipers & Sliding Calipers, Seritex	2
Spreading Calipers, Seritex	2
Head Light, Dental, Light-Tech, Inc	2

PHS 398 Cover Page Supplement

OMB Number: 0925-0001

1. Project Director / Principal Investigator (PD/PI)

Prefix:	Dr.	* First Name:	Mary
Middle Name:			
* Last Name:	Marazita		
Suffix:			

2. Human Subjects

Clinical Trial? ☒ No □ Yes

* Agency-Defined Phase III Clinical Trial? □ No □ Yes

3. Applicant Organization Contact

Person to be contacted on matters involving this application

Prefix:	Mr.	* First Name:	Allen
Middle Name:			
* Last Name:	DiPalma		
Suffix:			
* Phone Number:	412 624 7400	Fax Number:	412 624 7409
Email:	offres@offres.pitt.edu		

* Title:	Director Office of Research		
* Street1:	123 University Place		
Street2:			
* City:	Pittsburgh		
County/Parish:			
* State:	PA: Pennsylvania		
Province:			
* Country:	USA: UNITED STATES	* Zip / Postal Code:	15213-2303

PHS 398 Cover Page Supplement

4. Human Embryonic Stem Cells

* Does the proposed project involve human embryonic stem cells? ☒ No □ Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/. Or, if a specific stem cell line cannot be referenced at this time, please check the box indicating that one from the registry will be used:

Cell Line(s): □ Specific stem cell line cannot be referenced at this time. One from the registry will be used.

OMB Number: 0925-0001

PHS 398 Research Plan

1. Application Type:

From SF 424 (R&R) Cover Page. The response provided on that page, regarding the type of application being submitted, is repeated for your reference, as you attach the appropriate sections of the Research Plan.

*Type of Application:

☒ New ☐ Resubmission ☐ Renewal ☐ Continuation ☐ Revision

2. Research Plan Attachments:

Please attach applicable sections of the research plan, below.

1. Introduction to Application (for RESUBMISSION or REVISION only)		Add Attachment	Delete Attachment	View Attachment
2. Specific Aims	Specific Aims.pdf	Add Attachment	Delete Attachment	View Attachment
3. *Research Strategy	Research Strategy2.pdf	Add Attachment	Delete Attachment	View Attachment
4. Inclusion Enrollment Report	Inclusion.pdf	Add Attachment	Delete Attachment	View Attachment
5. Progress Report Publication List		Add Attachment	Delete Attachment	View Attachment

Human Subjects Sections

6. Protection of Human Subjects	HumanSubj.pdf	Add Attachment	Delete Attachment	View Attachment
7. Inclusion of Women and Minorities	WomenMinor.pdf	Add Attachment	Delete Attachment	View Attachment
8. Targeted/Planned Enrollment Table	Inclusion.pdf	Add Attachment	Delete Attachment	View Attachment
9. Inclusion of Children	Children.pdf	Add Attachment	Delete Attachment	View Attachment

Other Research Plan Sections

10. Vertebrate Animals		Add Attachment	Delete Attachment	View Attachment
11. Select Agent Research		Add Attachment	Delete Attachment	View Attachment
12. Multiple PD/PI Leadership Plan	MultiPDPI.pdf	Add Attachment	Delete Attachment	View Attachment
13. Consortium/Contractual Arrangements	col consort.PDF	Add Attachment	Delete Attachment	View Attachment
14. Letters of Support		Add Attachment	Delete Attachment	View Attachment
15. Resource Sharing Plan(s)	Sharing.pdf	Add Attachment	Delete Attachment	View Attachment

16. Appendix Add Attachments Remove Attachments View Attachments

SPECIFIC AIMS

In this proposal we request genotyping to investigate genes underlying quantitative facial variation. The parent grant for this project is U01-DE020078 ("3D Analysis of Normal Facial Variation: Data Repository and Genetics") which is part of the FaceBase consortium (www.FaceBase.org, Hochheiser et al 2011). A core objective of the FaceBase consortium is to facilitate the rapid generation of new data resources that will enable researchers to investigate the genetic factors underlying both normal and abnormal facial development. Fully achieving this objective requires the construction of large minable databases comprised of detailed quantitative information on craniofacial structures as well as genotypic variation. Advances in non-invasive 3D imaging have made it feasible to capture high-resolution, geometrically accurate 3D surface scans of the face in large numbers of individuals. Moreover, with the advent of high-throughput genotyping, it is now possible to marry quantitative phenotypic data derived from 3D images to vast quantities of genetic information. Establishing such a phenotype-genotype database would represent a major step forward in the search for the genetic determinants of facial form and variation. Under the parent grant, the 3D facial image data is now available through FaceBase (known as the Three Dimensional Facial Norms dataset, i.e. TDFN) and the current proposal is requesting the genotyping to augment the image data resource.

The motivation for this project springs from the fact that very little is known about how variation in specific genes relates to the diversity of facial forms commonly observed in humans. Viable candidates for these morphogenes originate from a number of sources: tissue expression studies, animal models with targeted or spontaneous mutations, and genetic syndromes with craniofacial manifestations. Importantly, understanding the genetic basis for normal facial variation also has important implications for health-related research. For example, this work has the potential to shed light on the factors influencing liability to common craniofacial anomalies such as orofacial clefts. There is now ample evidence that certain facial features (e.g., increased midfacial retrusion) characterize individuals genetically at-risk for orofacial clefts (e.g., biological relatives of affected cases). While these predisposing facial features are statistically over-represented in at-risk groups, they are also common in the general population. Since many of the current candidate genes for clefting are thought to play a critical role in facial morphogenesis, variation in these genes may also underlie normal variation in these facial features. These candidate genes, however, probably represent only a small fraction of the total number of loci influencing facial variation. The GWAS approach proposed here has the potential to facilitate the discovery of new loci that may play an important role in both normal and abnormal facial development.

This CIDR proposal has three major goals:

Goal 1: To genotype subjects recruited and imaged for the FaceBase TDFN dataset (approximately 3,300 individuals). Requested is the Illumina HumanOmniExpress plus Exome and custom content (to cover genomic regions with previous evidence of association with quantitative facial variation).

Goal 2: To analyze the TDFN image and genotype data to identify genetic loci that influence normal facial variation.

2a: Midfacial variation Given the rich surface capture available from 3D facial imaging, there are an essentially limitless number of potential measurements to investigate. In keeping with the current aims of FaceBase (ie a focus on Midfacial development), we will initially analyze a subset of standard midfacial linear distances.

2b: Geometric morphometrics To seek genetic factors underlying general facial shape, we will also analyze principal components of linear measures derived from geometric morphometric approaches.

Goal 3: To perform a meta-analysis with data collected in an additional FaceBase project that collected facial images and genotypes in 3,700Tanzanians (co-I Spritz is the PI of the Tanzanian project).

RESEARCH STRATEGY

Significance and Evidence of a Genetic Component

In this proposal we request genotyping to investigate genes underlying quantitative facial variation. The parent grant for this project is U01-DE020078 ("3D Analysis of Normal Facial Variation: Data Repository and Genetics") which is part of the FaceBase consortium (www.FaceBase.org, Hochheiser et al 2011). A core objective of the FaceBase consortium is to facilitate the rapid generation of new data resources that will enable researchers to investigate the genetic factors underlying both normal and abnormal facial development.

The Identification of Genes Underlying Normal Facial Shape in Humans: Progress to Date

Numerous lines of converging evidence indicate that variation in facial morphology has a strong genetic basis; these include human heritability studies using twin and parent-offspring designs (Byard et al., 1985; Carels et al., 2001; Ermakov et al., 2006; Martinez-Abadias et al., 2009), craniofacial syndromes resulting from known gene mutations (Gorlin et al., 2001), transgenic animal models with distinctive craniofacial phenotypes (Morriss-Kay et al., 1996; Hallgrimsson et al., 2006; Perlyn et al., 2006), and studies mapping QTLs for craniofacial shape in several mammalian models (Cheverud et al., 1997; Haworth et al., 2001; Klingenberg et al., 2001; Sherwood et al., 2008). However, we still have a very poor understanding of how variation in specific genes relates to the diversity of facial forms commonly observed in humans. To date only a handful of peer-reviewed studies have explored this topic explicitly. Coussens and van Daal (2005) reported a significant association between a single SNP within the tyrosine kinase domain of the FGFR1 gene and cranial vault shape; mutations in this gene result in craniosynostosis, which is characterized by extreme distortions in vault shape due to premature suture fusion. In a very limited study, Boehringer et al. (2011) investigated genotype-phenotype associations in a small number of SNPs in candidate genes relevant to orofacial clefting, identifying a few possible hits for measures of nose and facial width; both of these traits have been shown to be differ in samples at risk for orofacial clefting (i.e. cleft lip and cleft palate; Weinberg et al., 2008). In 2012, two independent GWA studies were carried out on large Caucasian samples of healthy subjects using 3D facial imaging and a combination of traditional and more advanced morphometric methods to derive phenotypes (Paternoster et al., 2012; Liu et al., 2012). Between these two studies, a handful of intriguing genome-wide significant results were obtained, but only a single gene has been partially replicated – PAX3. In both studies variation in PAX3 SNPs was associated with anatomical changes in the interorbital region; an intriguing finding, given that mutations in PAX3 cause Waardenburg Type 1 syndrome which is characterized by hypertelorism among other morphological abnormalities.

In summary, the few genetic studies completed to date have shown hits in biologically plausible genes that underlie predictable facial variations. It is quite clear, however, that these early studies are just scratching the surface and that the potential for additional discovery is great. Part of the problem with prior studies involves the methods used to capture the facial images from which the core phenotypes are derived. Both of the prior GWA studies (Paternoster et al., 2012; Liu et al., 2012), for example, relied on imaging methods that are susceptible to serious motion artifacts, potentially adding considerable noise to the data. This can be a major problem when attempting to detect loci with small effects on the phenotype, as is to be expected with normal quantitative facial traits. Thus, the current project represents a significant advance over previous efforts because the 3D facial images for our project were obtained using imaging technology that is not susceptible to these types of artifacts, capturing human faces in full 3D in less than 2 milliseconds. As a result, we anticipate that our facial phenotypes will be cleaner and therefore provide greater resolution for detecting putative loci.

Broader Public Health Relevance: From Normal to Abnormal Variation

The success of investigations into the genetic and environmental causes of craniofacial malformations depends on the acquisition of objective, reliable, and carefully collected data on the craniofacial phenotype. This need was formally recognized in 2004 when the National Institute of Dental and Craniofacial Research (NIDCR) posted a request for applications "to refine clinical characterization, to identify diagnostic biomarkers, to establish reliable subphenotypes, and to develop standardized and comprehensive phenotypic definitions" (NIDCR: DE04-004). Many individuals with genetic syndromes that affect the head and face present with very subtle morphological disturbances. Implicit in any description of facial dysmorphology is the notion that the phenomenon under consideration represents a deviation from some normal or baseline state. Thus, all descriptions of dysmorphology are comparative by nature. As a consequence, an understanding of what constitutes the range of normal variation for craniofacial features is essential for the necessary comparative studies. Although attempts to quantify the human face date back to antiquity, standardized methods for measuring the human face were only developed in the early 20th century (Hrdlicka, 1952; Kolar and Salter, 1997). In response to the need by the clinical community for population-based norms, large datasets comprised of standardized facial anthropometric or cephalometric measures were eventually constructed (Feingold and Bossert, 1974; Saksena et al., 1987; Farkas and Munro, 1987; Farkas, 1994). In order to fully capture the variation present in the population while simultaneously providing age- and sex-specific normative data, large numbers of healthy individuals were typically recruited for these projects. Today, these normative datasets are routinely used by clinicians and researchers to determine where a particular patient or subject sits relative to the population distribution (e.g., by constructing Z-scores).

Understanding the genetic basis for normal facial variation has important implications for human health, primarily because the boundary dividing normal from abnormal facial morphogenesis is often only a matter of degree. Many genetic syndromes affecting the face are characterized by subtle morphological changes, often involving quantitative traits with continuous distributions. The distribution for a given trait will often display substantial overlap between affected cases and healthy individuals. Thus, understanding the genetic factors that contribute to normal facial trait variation may provide valuable insights into the causes of craniofacial dysmorphology including the most common craniofacial birth defect, orofacial clefting (i.e. cleft lip and cleft palate). For example, there is now ample evidence that certain facial features (e.g., increased midfacial retrusion and greater upper facial width) characterize ostensibly unaffected individuals genetically at-risk for orofacial clefts (e.g., the immediate biological relatives of affected cleft cases) (Weinberg et al, 2008; 2009). While these predisposing facial features are statistically over-represented in at-risk groups, they are also common in the general population. Since many of the current candidate genes for clefting are also believed to play a critical role in facial morphogenesis, variation in these genes may also underlie normal variation in these facial features (Boehringer et al., 2011).

In the aggregate, craniofacial anomalies are among the most common birth defects, plus a very high proportion of human genetic syndromes involve malformations of the craniofacial complex, including midfacial structures (Gorlin et al., 2001; Jones, 2006). Orofacial clefting alone is associated with over 350 distinct syndromes (Leslie and Marazita, in press). Further, a large number of syndromes are associated with subtle facial deviations, such as hypoplastic facial features and/or altered facial proportions (Jones, 2006). Finally, because most syndromes are monogenetic and display Mendelian patterns of inheritance, classical mapping approaches have been highly successful in identifying causative genes. Although relatively rare, such syndromes can serve as a kind of natural experiment, providing important clues as to the genetic basis of facial form (see Leslie and Marazita, in press, for a recent review of such identified genes).

In summary, understanding the genetic factors that contribute to normal facial trait variation may provide valuable insight into the causes of craniofacial dysmorphology and common structural birth defects. Specifically, this work has the potential to shed light on the factors influencing liability to common craniofacial anomalies such as orofacial clefts.

Building a Resource for the Community

Note that all subjects are consented for full data-sharing, including the primary facial scans and genetic data. As part of NIDCR's FaceBase Consortium (Hochheiser et al., 2011), all the phenotype data will be available for data-sharing with any qualified investigators and the clinical community via the NIDCR-funded FaceBase web portal (www.FaceBase.org), following approval by the NIDCR FaceBase Data Access Committee. This includes not only derived craniofacial measurements and demographic descriptors, but the original 3D facial source images (with proper permissions). It is worth emphasizing that no other existing data repository contains facial phenotypic data like this; genetic data is currently the missing part that will make this a truly unique and valuable resource. If CIDR access is granted, the resulting genotypes will reside in dbGaP for individual-level data downloads.

Investigators given access to the full data will be able to self-define an almost infinite number of orofacial morphometric phenotypes and then study the genetics underlying those phenotypes in humans or animal models. However, the facial scans can be reconstituted as photographs deemed potentially identifiable by one of the participating IRBs (ie the University of Colorado IRB—co-I Spritz), whereas dbGaP policy is to accept only de-identified data. Therefore, NIDCR has determined that the full scans will be shared via the FaceBase Hub, and only a set of 15 commonly-used orofacial morphometric measurements derived from the scans, as well as all human genetic data, will be submitted to dbGaP. See the Appendix and the Research Strategy for a list of the morphometric measures to be submitted to dbGaP.

By making both the original phenotypic data and genotypes available to the broader scientific community, our hope is to create an unparalleled resource that will stimulate other investigators to explore questions related to human craniofacial genetics. We anticipate that the genotypic and phenotypic data we provide will facilitate additional genetic studies on human facial shape, from initial GWA discovery to meta-analysis to replication/extension studies.

Innovation

Combining Detailed Quantitative Phenotypes with Genotypes

Our project involves generating a variety of rich and detailed phenotypes for use in genetic analysis. Our proposed genetic analysis will focus mainly on multivariate shape phenotypes derived from principal components analysis. In addition, however, we will generate a number of standard anthropometric measurements capturing various dimensions of the human face and head. Furthermore, the raw 3D facial surfaces from which the measures are derived contain a huge amount of quantitative information (a typical facial surface is comprised of over 20,000 points). These surfaces are of great interest to those in the computer science community who possess the skills necessary to deal with this quantity of information. All of this information will be wedded to high density SNP data.

Unrivaled Public Access

Because our project is part of the FaceBase Consortium, all of the phenotypic data generated (including the original source 3D facial images for each subject in the dataset) will be available to investigators who apply for access. The genotype data generated from this CIDR proposal will be available through dbGaP. This level of data access will allow investigators to not only use the phenotypes we derive, but also derive their own novel phenotypes from the original source data, for use in their own customized analyses.

Approach

Sample Description

Data and samples from a total of **3,132** subjects is already in hand (see inclusion enrollment report)—note that enrollment and assessment is on-going so that by the time CIDR access is granted we anticipate a total of **3,300** subjects will be available for analysis. 3D facial images (see Figure 1 below) and basic demographics are collected (see Appendix for data dictionary), and DNA is extracted from saliva samples collected in Oragene kits. The phenotypes and biological samples available for this proposal are derived from three sources/projects: **3D Facial Norms (TDFN)**: This is a spoke project of the FaceBase Consortium (PIs: Weinberg and Marazita) and currently contains data from 2117 subjects; **Genetic Determinants of Orofacial Shape**: This is also a spoke project for FaceBase (PI: Spritz). Dr. Spritz (co-I in this application) will provide data and samples from 797 juvenile and adolescent subjects; **Extending the Phenotype of Nonsyndromic Orofacial Clefts**: Data and samples from 218 healthy controls recruited as part of this project (PI: Marazita) will also be made available.

All of the subjects are healthy unrelated Caucasians of recent European Ancestry as far back as their maternal and paternal grandparents. They are between the age of 3 and 40 years, with very close to uniform distribution across that age range. Exclusion criteria include (1) a personal or family history of any syndrome or other congenital condition with a craniofacial manifestation; (2) a personal history of facial plastic, reconstructive, or orthognathic surgery; and (3) a personal history of major facial trauma. Males with excessive facial hair were also excluded. Note that the FaceBase initiative funded two similar projects to collect 3D facial images and genetic data: the Weinberg/Marazita project with all Caucasians and the Spritz project with mostly Tanzanians and a subset of Caucasians. This current CIDR grant proposal is focusing on the European Caucasians; co-I Dr. Spritz is in the midst of a CIDR genotyping project of just the Tanzanians from his FaceBase project.

Types of Phenotypic Data Available for the Resource

From each contributing study, phenotypic data will include 3D facial surfaces (obj format), 24 3D facial landmarks, 29 anthropometric facial measurements and basic demographic descriptors (sex, age, etc.). Data from the three contributing projects used similar technology to obtain facial images and collected the same facial landmarks and demographic information. This will allow for seamless data integration. Refer to the Appendix for the data dictionary.

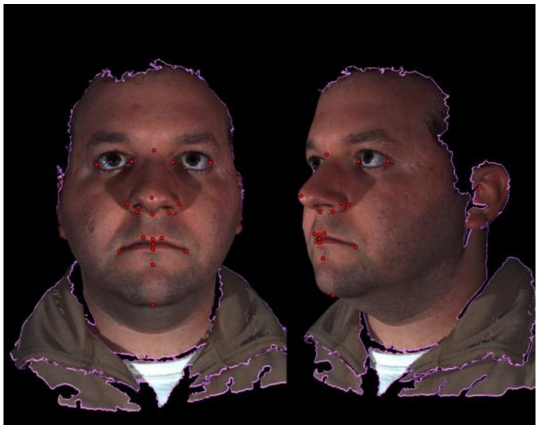
Phenotypic Data QC

In order to deal with a large quantity of 3D facial image data, our Pittsburgh team has developed a number of protocols and processes to facilitate uniform data collection and ensure quality. First, all individuals involved in data collection from 3D images go through a multi-phase training program, in which they learn how to manage, clean, and landmark 3D facial images. We have developed a number of custom automated "housekeeping" programs designed to locate and verify 3D image files and their associated measurement files. With the help of a landmark import utility program, 3D coordinates collected from facial images are automatically detected and imported into a relational database. A landmark checker program then automatically reads the coordinates from this database and detects placement errors. Common statistical methods such as outlier detection are then used to identify any remaining issues in the data. More than 6000 3D facial images from various projects have been processed by Drs. Weinberg and Marazita using this pipeline.

<u>Extracting Quantitative Facial Phenotypes</u>

A geometric morphometric (GM) analysis pipeline will be used to extract quantitative facial phenotypes from the landmark coordinate data (Dryden and Mardia, 1998; Zelditch et al., 2004). This type of analysis starts by placing the 3,123 landmark configurations (see Figure 1 for sample landmarks in the midface marked as red dots) derived from each subject's 3D facial surface image into a common coordinated system; this is accomplished via a generalized Procrustes superimposition (Rohlf, 1999). This step centers, rotates and scales the configurations in an iterative fashion until a solution minimizing the sum-of-squares is obtained. The resulting transformed coordinates reflect variation in shape and are referred to as Procrustes coordinates. The coordinate data can be further adjusted for traits such as age and body size using regression. The resulting coordinates are then treated as new variables and subjected to multivariate data reduction, typically principal components analysis (PCA). In the context of GM, PCA is designed to capture the major modes of unique shape variation in a dataset (Klingenberg, 2010). A major advantage of GM is in its visualization capacity. The intrinsic geometric properties of the data can be leveraged in order to reveal the patterns of shape variation as deformable morphs. In PCA, each axis of variation (each PC) can be modeled by deforming the mean shape according to the eigenvector weightings. The scores on the extracted PCs will be used as quantitative variables in our genetic analysis.

Figure 1. A 3D facial Surface showing the location of 24 standard anatomical landmarks



<u>Phenotypes that will be Initially Studied by our Group</u>

A number of standard anthropometric variables (simple linear distances) will also be calculated from the landmark coordinate data described in the previous section. A subset of these variables will be provided to dbGaP along with the genotype data (see Appendix for data dictionary) and all of the landmark coordinates for each subject used in the proposed analysis will be available to investigators via the FaceBase web portal. While all phenotypes will be made available to the research community, our own group will focus first on those that have been shown to be most relevant to orofacial clefting, specifically those related to Midfacial development (see table below for the traits along with descriptive statistics from the FaceBase TDFN dataset).

Linear Distance	**Landmarks**	**FaceBase Variable Name**	**Min**	**Max**	**Mean**	**SD**
Middle facial depth (Right)	sn - t_r	MidFaceDepth_R	97.00	146.96	124.6803	8.66952
Middle facial depth (Left)	sn - t_l	MidFaceDepth_L	94.47	146.19	124.1995	8.69851
Lower facial depth (Right)	gn - t_r	LowFaceDepth_R	101.23	177.15	140.2917	11.98008
Lower facial depth (Left)	gn - t_l	LowFaceDepth_L	99.89	175.69	139.7832	11.97712
Upper facial height	n - sto	UpFaceHeight	46.65	93.48	74.0649	6.67782
Lower facial height	sn -gn	LowFaceHeight	45.03	88.48	66.9002	6.77030

Intercanthal width	en_r - en_l	InCanthWidth	21.58	43.86	31.9973	3.03505
Nasal width	al_r - al_l	NasalWidth	23.25	44.80	33.1620	3.33149
Maximum facial width	zy-zy	MaxFaceWidth	85.00	168.00	130.0399	9.07496

Justification of Services Requested

Requested is the Illumina HumanOmniExpressExome panel (>700,000 GWAS SNPs, plus ~250,000 exome SNPs), plus the addition of up to 5000 custom beadtypes to cover genes/regions already implicated in the genetics of facial shape and form. The OmniExpressExome panel allows detection of common and rare variants (to MAF 1%). Further it has recently been demonstrated (Nelson et al, in press) that after imputation, the OmniExpress panel achieves up to 86.6% coverage of the genome in Caucasians (dependent on assumptions). The inclusion of the Exome SNP panel and custom content adds very little to cost, and allows the consideration of potentially functional SNPs within exomes across the genome. In total, this panel is the best combination of coverage, flexibility, and informativeness for Caucasians, while costing considerably less than a sequencing approach.

Data storage and management

We rely on SQL, ACCESS and ProgenyLab databases for management of clinical and genotype data. ProgenyLab is a pedigree and genetic lab management software program with an integrated, customizable relational database for data management and pedigree analysis. Sufficient hardware and software is available for automating the large number of statistical analysis steps necessary for this large-scale study (see Resources and Equipment files for description of computer systems). Drs. Marazita and Feingold are very experienced in managing large datasets, through multiple large cohort studies, multiple GWAS and sequencing studies, and through Dr. Marazita’s involvement as one of the PD/PIs for the FaceBase online data resource for craniofacial researchers.

Data Analyses

Data analysis team and support. The statistical genetics analysis team for our project consists of PD/PIs Marazita and Feingold, plus other faculty and staff who will assist in the strategy and in directing analyses by graduate students and junior faculty. All of these individuals have extensive experience with GWAS (and other) data analyses, having participated, for example, in multiple GENEVA projects and our other ongoing projects. Drs. Marazita and Feingold have collaborated closely since 2008.

Data QC and cleaning. The initial quality control and data cleaning for this project would be provided by CIDR through their collaboration with the University of Washington (Dr. Bruce Weir’s team). We have extensive experience interacting with this QC group since they were also the Data Coordinating Center for the GENEVA projects. Data cleaning will include all standard checks, including using various diagnostics to identify problematic samples and SNPs, identifying chromosomal abnormalities, checking for family relationships, testing and exploring any ethnic stratification, and checking genotype calling scatterplots as needed.

General GWAS methods We will conduct basic QTL GWAS analyses on the quantitative traits discussed above using PLINK (Purcell et al, 2007) with additive models. We will control for potential ethnic stratification using principal components. The X chromosome will be analyzed using Clayton’s statistic, as we have recently shown (not yet published) that this is far more robust than the commonly-used regression with (0, 1) coding for male genotypes. We will also conduct gene-level and pathway-level analyses, and are currently developing a new gene-based test. In addition, we will call CNVs using PennCNV and perform association tests SNP by SNP of copy number vs. each outcome. All analyses will also include imputed SNPs – imputation is done to 1000 genomes by the University of Washington team.

Covariates The two major covariates that will need to be handled in these analyses are age and sex. Sex will be included in all models as a simple covariate. Handling of age will depend to some extent on the phenotype under consideration, but in each case the first step will be thorough modeling of the relationship between the phenotype and age. We anticipate that for some analyses we will restrict ourselves to particular age subsets, but for the most part we will carefully model an age-adjusted phenotype and use all subjects.

Rare variants. Rare variants will be studied using burden-type tests that first group variants into regions (genes, pathways, functional prediction groups) of interest and then compare the composite variable to the phenotypic outcome. Our preferred test given the currently-available options is SKAT, though we are always considering other options as the literature evolves. Population stratification in rare variant tests is a very controversial issue, and we will consider adjusting using principal components as well as using local-ancestry approaches.

Multiple comparisons issues. Control of type I error is implemented by inspection of qq plots. We consider GWAS to be a hypothesis-generating procedure, not a hypothesis testing procedure, so we do not draw absolute limits on the p-values that we consider “interesting.” In general we do careful bioinformatic follow-up on any SNP that achieves a p-value of 10-5 or less.

Power. In part the justification of the power of our sample size is due to the success of other published GWAS of anthropometric traits using similar sample sizes. At present, we do not have any basis for estimating the likely sizes of individual gene contributions to the heritability of these traits. Therefore, we estimated the power for our sample size (n=3,300) at a range of effect sizes. The summary table below shows that power is quite high even for QTLs that explain a very modest proportion of trait variance. (Note that power does not depend on marker frequency but that low-frequency markers are unlikely to explain a high proportion of variance).

Power for QTL association with a sample size of 3,300

Percent of variance explained by the marker	Power for alpha = $.10^{-5}$	Power for alpha = 10^{-8}
.05%	.36	.05
.06%	.52	.10
.07%	.65	.18
.08%	.77	.28
.09%	.85	.39
1%	> .99	.51
2%	> .99	> .99
3%	> .99	> .99

Replication and other plans to follow-up the requested GWAS:

Meta-analysis with Tanzanian data. Co-I Dr. Spritz plans a meta-analysis with his phenotype and genotype data from Tanzania and this new study. He plans to first carry out genome-wide association analysis of the new Caucasian data imputed to 1KGP, using PLINK, followed by genome-wide meta-analysis between the Caucasian and Tanzanian Bantu datasets, using both random and fixed effects models, to identify segments of potential ancestrally shared association, evidenced by significant or near-significant association of shared effect alleles. To assess the likelihood that these candidate signals are shared ancestrally between Caucasian and Tanzanian Bantu, and to provide refined localization of ancestrally shared associations, we would then carry out trans-ethnic meta-analysis of these candidate regions using MANTRA (Morris, 2011), which unfortunately cannot be carried out on a genomewide basis due to computational limitations. We have previously used this approach successfully to refine localization of genetic association signals for vitiligo (Birlea et al., 2013).

Sequencing under the peaks of significance. Deep resequencing of candidate genes and regions is likely to be one of our most important follow-up tools, with the goal of detecting causal variants in the regions of GWAS peaks. In order to do this, we will pursue several complementary filtering and testing strategies (as we are currently doing in other sequencing studies of orofacial clefting): variant calling filtered against databases, filter against parents (to detect de novo variants), confirm identified variants in the extensive control population, and subject identified variants to functional prediction models. Mutations that are predicted to be deleterious and/or are seen most frequently in the dataset will undergo further testing. We also perform a number of secondary analyses, including looking at variants that are transmitted (not de novo) and comparing frequencies of variants and functionally-defined groups of variants between subgroups (e.g. defined by ethnicity).

Followup functional studies It is difficult to predict what our potential functional analyses will be following up this GWAS (and/or sequencing studies). As one avenue, note that the parent project is part of FaceBase, and several of the ten other FaceBase consortium projects are specifically aimed at testing involvement of selected genes in craniofacial development and in the pathogenesis of craniofacial anomalies. Within the collaborative consortium structure of FaceBase, any positive results from the current study will directly influence planning of loci to be studied by other consortium members working in animal models.

Inclusion Enrollment Report

This report format should NOT be used for data collection from study participants.

Study Title:	Genetic analysis of quantitative facial variation		
Total Enrollment:	3,132	**Protocol Number:**	
Grant Number:			

PART A. TOTAL ENROLLMENT REPORT: Number of Subjects Enrolled to Date (Cumulative) by Ethnicity and Race

	Sex/Gender			
Ethnic Category	**Females**	**Males**	**Unknown or Not Reported**	**Total**
Hispanic or Latino	54	84	0	138 **
Not Hispanic or Latino	1745	1249	0	2994
Unknown (individuals not reporting ethnicity)	0	0	0	0
Ethnic Category: Total of All Subjects*	1799	1333	0	3132 *
Racial Categories				
American Indian/Alaska Native	0	0	0	0
Asian	0	0	0	
Native Hawaiian or Other Pacific Islander	0	0	0	0
Black or African American	0	0	0	0
White	1799	1333	0	3132
More Than One Race	0	0	0	0
Unknown or Not Reported	0	0	0	0
Racial Categories: Total of All Subjects*	1799	1333	0	3132 *

PART B. HISPANIC ENROLLMENT REPORT: Number of Hispanics or Latinos Enrolled to Date (Cumulative)

Racial Categories	**Females**	**Males**	**Unknown or Not Reported**	**Total**
American Indian or Alaska Native				
Asian				
Native Hawaiian or Other Pacific Islander				
Black or African American				
White	54	84	0	138
More Than One Race				
Unknown or Not Reported				
Racial Categories: Total of Hispanics or Latinos**	54	84	0	138 **

* These totals must agree.
** These totals must agree.

Genetic analysis of quantitative facial variation
PD/PIs: Mary L. Marazita, Ph.D., Seth Weinberg, Ph.D., Eleanor Feingold, PhD.

PROTECTION OF HUMAN SUBJECTS

The samples and data for this project come from two projects, (1) one led by Drs. Seth Weinberg and Mary Marazita at the University of Pittsburgh, and (2) one led by Dr. Rich Spritz at the University of Colorado. The total number of study participants in hand is 3,125 (see inclusion enrollment report), 797 from the Colorado study and the remainder from Pittsburgh. Recruitment is on-going, thus at the time the samples are to be sent to CIDR we anticipate that there will be approximately 3,300 total samples.

(1) University of Pittsburgh (Weinberg and Marazita)

The primary goals of this study are to: (1) construct a repository comprised of normative anthropometric measures derived from 3D facial surface images of Caucasians, plus genome-wide SNP markers for each participant; and (2) identify SNPs associated with normal variation in midfacial morphology. Based on these goals, the parent study focused on recruiting a large sample of healthy males and females between the ages of three and 40. Recruiting efforts were carried out at three sites: the University of Pittsburgh, the Seattle Children's Research Institute and the University of Texas Health Science Center at Houston. IRB approval for this study is pending at the three sites. In general, all three sites will implement identical research protocols. However, when appropriate, site-specific details will be highlighted.

Risks to Human Subjects

<u>Human subjects involvement and characteristics</u> Enrollment was limited to individuals of European Caucasian ancestry (white, non-Hispanic) in an effort to ensure adequate sample size and to minimize potential population stratification in the genetic analysis component (see INCLUSION OF WOMEN AND MINORITIES section). Further, we are limiting our age range to individuals between three and 40 in order to ensure that the number of participants at each age is sufficient to capture the range of variation present in the population (50 subjects per age in each sex). Because this study is concerned with normative variation, healthy individuals culled from the general population will be recruited to participate. Criteria for excluding individuals primarily relate to issues that could affect craniofacial structure: a personal or family history of any syndrome or other congenital condition with a craniofacial manifestation; a personal history of facial plastic, reconstructive or orthognathic surgery or use of major orthodontic appliances; a personal history of major facial trauma. We will also exclude individuals with extensive facial hair, since this obscures relevant facial anatomy during 3D facial imaging.

<u>Sources of materials</u> Individuals enrolled in this study will have one or more images taken of their face, using a surface-based 3D photogrammetry system. Data in the form of 3D landmark coordinates and linear distance measurements will be collected from the 3D images. Basic demographic information will also be collected. All study participants will also be asked to contribute a saliva sample using Oragene collection kits for extracting DNA. Individuals may elect to participate partially in the study by, for example, agreeing to have their 3D image taken, but not agreeing to provide a DNA sample. All resulting data, saliva samples and 3D images will be obtained specifically for research purposes. Only designated individuals will have access to sensitive health information.

<u>Potential risks</u> There are very few foreseeable risks for individuals participating in this study. The 3D imaging method is entirely non-invasive and is similar in many respects to having a standard photograph taken. DNA will be collected through saliva samples, obviating the need for venipuncture. Some participants may feel uncomfortable sharing certain questions about personal or family medical history with the research team. Participants may also object to having their facial image and/or genotype data included as part of a repository, accessible by the broader scientific community. The greatest potential risk in this study is the inappropriate release of confidential and/or personally identifying information by members of the research staff.

Adequacy of Protection Against Risks

Recruitment and informed consent The same parameters for subject eligibility was used at the Pittsburgh, Seattle and Houston sites.

At least one full-time staff person at each site is dedicated to the recruitment/consenting of participants. Upon arrival of the potential subject at the site, all participants receive a general introduction to the research (purpose, procedures, etc.) and any questions that they have are answered. If an individual chooses to participate, a consent form is provided to read silently, after which designated site staff review the document in detail with them to be certain that they understand any potential risks. The consent forms are then signed and witnessed.

All aspects of the research are explained individually to each potential participant over age 10 in detail and any questions that they have are answered. Children under 10 years of age remain with their parents during the entire informed consent procedure. If a potential participant over age 10 chooses to consider participating, he/she is given a consent form to be read silently, which is then reviewed with them. The consent forms will then be signed and witnessed. One of the parents is required to co-sign the consent form for all children under 18 years of age.

Minimal exclusion criteria are employed regarding factors that raise doubts about the capacity to provide informed consent. These include chronic medical disease, life threatening illness, neurological trauma requiring hospitalization, neurodevelopmental disability, uncorrectable sensory handicap, and psychosis.

Protections against risk As described above, each participant was fully informed about the research protocol. Because this information was communicated on several occasions and also contained in a written detailed description prior to evaluation, there is ample opportunity for participants to withdraw if they perceive any risk. Since a history of significant acute or chronic medical illness is a disqualifying criterion, the first level of minimizing risk will involve adherence to the recruitment procedure described previously. As part of the need to reduce any lasting concerns or anxieties, subjects undergo a debriefing following completion of the protocol. Any concerns that arose during the course of the investigation will be addressed and resolved at this juncture.

Participants can exercise the right to refuse answering any questions that they perceive as inappropriate or overly invasive. As part of the 3D normative database component, we plan to make facial images, in addition to derived measurements, available for viewing and analysis. To protect participant privacy, only the geometric content of the images will be housed within the normative database (i.e., the identifying surface features like skin and hair will be stripped from the image, resulting in a monochromatic surface, prior to upload into the database). This process effectively renders the images non-identifiable and will be explained with the aid of visual examples to all participants. Thus, no information permitting personal identification of individuals will be made public. Participants still concerned about their image data being part of a publicly available database will have the option to elect to have all or some of their data withheld from this portion of the study. No clinical descriptions that might permit personal identifications will be published, and all clinical information will be stored in secure, encrypted computer files. All participants will receive a random alphanumeric ID upon enrolment, and this number will be used on all data forms and electronic files. Any and all paper data forms will be stored in lockable filing cabinets. We will work closely with the FaceBase hub to develop controlled access procedures for potentially sensitive data (e.g., GWAS markers).

Potential Benefits of the Proposed Research Participants enrolled in this study do not receive any direct benefits from this research other than reimbursement for travel expenses and participation. Likewise, there is no direct benefit to any other individuals.

Importance of the Knowledge to be Gained The proposed normative 3D facial database will serve as an important resource for investigators. As 3D facial imaging becomes more common in clinical and research environments, demand for high quality 3D normative data will increase. Currently, because no accessible repository containing 3D normative facial data is available, investigators must either rely on unreliable data sources or collect their own control samples at great expense and effort. The proposed normative database will contain both 3D surface images and detailed quantitative measures of facial morphology on a large number of healthy children and adults, available through FaceBase in a web-based minable format. Thus, it will serve as a rich source of high-quality control data for investigators interested in carrying out comparative morphological studies.

Because this repository will also include genome-wide SNP data, it has the potential to facilitate a variety of novel research endeavors. This database, we argue, will enable discoveries that improve our understanding of the genetic factors influencing normal variation in facial features. Consequently, an improved understanding of the nature of normative facial variation may, in turn, enhance our knowledge of the factors that underlie common forms of craniofacial dysmorphology.

(2) University of Colorado (Dr. Rich Spritz)

IRB approvals: Subjects were recruited under IRB approvals from the University of Colorado-Denver (Spritz) and University of California San Francisco (Klein). The University of Colorado-Denver IRB (COMIRB) was the IRB of record for the overall sample collection. Nevertheless, each site/research group underwent project review and approval from their respective individual governing IRBs before enrolling any subjects.

Subject recruitment: Subjects were recruited and enrolled at Children's Hospital Colorado and clinics and University of California San Francisco hospital and clinics.

Required education in the protection of human research participants: All study personnel completed human subjects research training curricula and certification as required by their individual parent institutions.

Subject inclusion and exclusion criteria: Only self-described white subjects (Non-Hispanic and Hispanic) were included. All subjects were children, ages 3-18.

Consent: Consent was obtained at the time of enrollment. As appropriate, the purpose, method, and voluntary nature of this study was explained to all participants/their parent(s)/decision-maker, and informed consent (and if applicable, subjects' assent) was obtained and kept on file by the research group enrolling the subject and collecting the data.

All subjects were explicitly consented for full sharing of 3D scans (which are deemed potentially recognizable images), genomewide genetic data, and standard parameters such as height, weight, head circumference, and non-identifying demographic data via the FaceBase Hub. Data access is limited to qualified applicant investigators per approval by a duly constituted NIDCR Data Access Committee.

Source of research material:

1) 3D orofacial morphometric scanning: 3D facial images were obtained using a Creaform rapid white light 3D photography system; the experience was essentially that of taking a facial photograph.
2) Height, weight, and head circumference were measured as covariates.
3) Saliva was obtained for preparation of DNA for genetic studies.
4) Information was provided by subjects or their parent(s)/decision-making guardian as to subjects' age, sex, self-described ethnicity, and other non-identifying demographic data.

Potential risks: There are no known physical risks to 3D orofacial morphometric imaging, provision of a saliva sample, genome-wide genetic analysis, and provision of information as to subjects' age, sex, self-described ethnicity. There is no possibility of unexpected diagnosis.

Protection against risks: There are no known physical risks associated with the procedures in this study. To preserve subject confidentiality, each study participant was assigned a unique subject identification number on enrollment in the study. This ID number is used to identify all subject information, facial scans, and other study data. ID numbers will be assigned by convention to prevent enrollment duplication and cross-site identifier clashes. Study investigators and other study staff will only have access to these codes. Only the local Study Coordinators who are in direct contact with the study subjects and the local Co-Investigator will have routine access to identifying information and codes applicable to subjects enrolled under auspices of that site. However, the study Investigators and the CU-Denver Study Coordinator will have access to all study information as may be necessary to ensure scientific integrity of the study, to ensure compliance with NIH, OHRP, HIPAA rules and regulations, and to ensure accurate compliance reporting.

Electronic data is stored in a secure server behind University firewalls, connected to the internal secure University networks, with access restricted only to members of the respective workgroups who require the data to perform their jobs. Server maintenance is performed by the University. Data access will be granted on an individual user level, with each user only accessing the data needed by that user. Subject identifiers and links to subject codes are stored on a freestanding non-networked computer. Data transmission to the FaceBase Hub will be via SFTP. All emails containing PHI or genetic information are encrypted.

Physical security of paper records is maintained by storing such records, including all forms and other identifiable information, in locked file cabinets in a locked office by the enrolling investigative team; identifiable paper records will be accessible only to appropriate authorized study personnel and to appropriate regulatory authorities. Physical records no longer needed will be shredded.

No publications will contain any personal or HIPAA identifiers or present identifiable information on individuals or specific families. However, consent will be obtained for potential publication of recognizable facial images, as may be necessary and appropriate.

Potential benefits of proposed research to subjects & others: There are no anticipated direct benefits for the study subjects. The proposed study may yield important information and methods for the analysis and diagnosis of genetic syndromes and defects of craniofacial development.

Importance of the knowledge to be gained: The knowledge to be gained from this research has potentially great importance, in that it may yield important information about structural abnormalities of individual genetic abnormalities, may lead to development of new tools to analyze such disorders, and may lead to development of a clinical tool to assist in differential diagnosis of genetic abnormalities in a clinical setting.

Collaborating sites: Each parent institution operates under an appropriate FWA. Individual IRB approval is required for the project at each site that is contributing data/information.

IRB policy on data sharing: Data sharing will be carried out at the levels required under FaceBase Consortium rules. All consent forms will include common language describing planned data-sharing. All data submitted to the FaceBase Hub will be de-identified; however, 3D facial images may be intrinsically recognizable. Data access via the FaceBase Hub will be limited to qualified applicant investigators approved by a duly constituted NIDCR Data Access Committee.

Genetic analysis of quantitative facial variation
PD/PIs: Mary L. Marazita, Ph.D., Seth Weinberg, Ph.D., Eleanor Feingold, PhD.

INCLUSION OF WOMEN AND MINORITIES

Individuals will be included in the study regardless of gender; our goal is enroll roughly equal numbers of males and females. Enrollment will be limited for this proposal to White individuals. This was deemed necessary because, to be of any practical use, craniofacial norms must be population-specific. Thus, race- and ethnicity-specific samples must be ascertained. This fact presents a major impediment to any study aiming to collect comprehensive normative craniofacial data, since no single proposal is likely to have sufficient resources to ensure that adequate sample sizes will be obtained in all possible demographic groups. The reality is that this goal will have to be accomplished in stages. In light of this, limiting our enrollment to a single racial/ethnic group will allow for adequate sample sizes to be obtained over a five-year period. A second reason for our recruitment strategy relates to the fact that allele frequencies can vary greatly across human populations, which, in the context of a genetic analysis, can introduce bias into the statistical methods used to assess associations between genotypes and measured traits.

We fully recognize that in order to be a useful resource the 3D normative database must eventually be expanded to include data on racial groups other than Caucasians, particularly given the demographic shifts currently taking place within the United States. As a first step, however, this project will facilitate the development of the basic infrastructure for this database. This infrastructure will be populated initially with subjects relevant to our genetic analysis of face shape, but can later be expanded to include additional groups. The scalable nature of this database is a core feature, and one that makes it wholly different from anything else in existence today. Thus, while we are focusing on Caucasian individuals in the current grant application, we are committed to pursuing additional funding to populate the database with individuals of other racial/ethnic backgrounds in the near future.

Inclusion Enrollment Report

This report format should NOT be used for data collection from study participants.

Study Title: Genetic analysis of quantitative facial variation

Total Enrollment: 3,132 **Protocol Number:**

Grant Number:

PART A. TOTAL ENROLLMENT REPORT: Number of Subjects Enrolled to Date (Cumulative) by Ethnicity and Race

	Sex/Gender			
Ethnic Category	**Females**	**Males**	**Unknown or Not Reported**	**Total**
Hispanic or Latino	54	84	0	138 **
Not Hispanic or Latino	1745	1249	0	2994
Unknown (individuals not reporting ethnicity)	0	0	0	0
Ethnic Category: Total of All Subjects*	1799	1333	0	3132 *
Racial Categories				
American Indian/Alaska Native	0	0	0	0
Asian	0	0	0	
Native Hawaiian or Other Pacific Islander	0	0	0	0
Black or African American	0	0	0	0
White	1799	1333	0	3132
More Than One Race	0	0	0	0
Unknown or Not Reported	0	0	0	0
Racial Categories: Total of All Subjects*	1799	1333	0	3132 *

PART B. HISPANIC ENROLLMENT REPORT: Number of Hispanics or Latinos Enrolled to Date (Cumulative)

Racial Categories	**Females**	**Males**	**Unknown or Not Reported**	**Total**
American Indian or Alaska Native				
Asian				
Native Hawaiian or Other Pacific Islander				
Black or African American				
White	54	84	0	138
More Than One Race				
Unknown or Not Reported				
Racial Categories: Total of Hispanics or Latinos**	54	84	0	138 **

* These totals must agree.
** These totals must agree.

Genetic analysis of quantitative facial variation
PD/PIs: Mary L. Marazita, Ph.D., Seth Weinberg, Ph.D., Eleanor Feingold, PhD.

INCLUSION OF CHILDREN

Children (according to NIH definition) between the age of three and 21 will be included in this study. Children under the age of three will be excluded because this age group will require a highly specialized recruitment effort, involving ascertainment at very fine age intervals. For instance, within the first year of life, samples (N = 50) of both boys and girls should ideally be recruited at monthly intervals. Such an intensive recruiting effort will require resources well beyond the scope of this proposal. Thus, to ensure our ability to obtain adequate sample sizes at each age interval, the current parent proposal focused on children three and older.

Genetic analysis of quantitative facial variation
PD/PIs: Mary L. Marazita, Ph.D., Seth Weinberg, Ph.D., Eleanor Feingold, PhD.

Multiple PD/PI Leadership Plan

Rationale: This project will have a multiple PD/PI structure with three PD/PIs sharing the oversight, planning, analysis planning, and results interpretation,

Structure: The three PD/PIs are Dr. Mary L Marazita, Dr. Seth Weinberg, and Dr. Eleanor Feingold, all at the University of Pittsburgh. They will jointly be responsible and accountable for the conduct of the proposed research.

Dr. Marazita will serve as the Project Director, as “contact PI”, and will coordinate all research efforts. She.. As contact PI, she will be responsible for communication between the PD/PIs and the NIH. In addition, she has the been the Principal Investigator on several other currently funded grants as well as several other grants now completed, so she has extensive experience in directing and administering complex research programs.

Dr. Weinberg is coPD/PI (along with Dr. Marazita) and director/contact PI of the parent grant that generated the study data/samples for this CIDR access request. He has particular expertise in 3D imaging and morphometrics. His primary responsibilities as a PD/PI in the current proposal will be to direct and coordinate all aspects of the facial phenotype data analysis and preparation, interpretation of results, and interacting with the FaceBase team for uploading data and creating FaceBase portal-based tools for summarizing the resulting data and making it available to the research community.

Dr. Feingold will serve as leader of the analysis team, which consists of several junior faculty, postdocs, programmers, and students. She have an extensive publication record in both applied and methodological statistical genetics, and has managed analysis teams for several large GWAS and sequencing projects over the last few years, as well as being co-PI of her own projects on the genetics of meiotic recombination and Down syndrome. Dr. Feingold has worked closely with Dr. Marazita and with most of the individuals on our analysis team for about five years. She led the analysis group for Dr. Marazita’s GENEVA dental caries GWAS and is now deeply involved in the follow-up work from that study. In addition, she has worked closely with CIDR personnel on many previous projects.

Process for Decisions on Scientific Direction: Drs. Marazita, Weinberg, and Feingold will meet on a regular basis to discuss progress of the project with CIDR, NIDCR, and the data analysis team, and will adhere to additional oversight structures required by the funding agency.

Communication/Meeting Plans: The PD/PIs will have regularly scheduled meetings (both in person and via teleconference).

Allocation of Resources: N/A, no budget.

Reporting Requirements: The PD/PIs will jointly be responsible for fulfilling all NIH reporting requirements, with coordination by the Project Director/contact PI (Dr. Marazita).

Resolution of Disputes: If any disputes arise between the PD/PIs, staff within the Senior Vice Chancellor’s office of the University of Pittsburgh will serve as mediators.

Genetic analysis of quantitative facial variation
PD/PIs: Mary L. Marazita, Ph.D., Seth Weinberg, Ph.D., Eleanor Feingold, PhD.

Resource Sharing

All of the investigators and institutions involved in the proposed project are committed to the concept that these results will be an important resource to the research community. The PD/PIs, Drs. Marazita, Weinberg, and Feingold, have been involved in many similar projects over the years, starting with the original Human Gene Mapping Workshops back in the 1980's. Both Drs. Marazita and Feingold have also been involved on an advisory level with one of the NIH genotyping service, the Center for Inherited Disease Research, each having served on the CIDR access committee for multiple years.

With respect to the project outlined in this application, any data generated by the project (subject to final IRB approvals) plus the facial phenotype data described in this application will be available for contribution to the controlled access databases of de-identified data that NIDCR, NCBI and NIH have developed (such as dbGAP and FaceBase). We are totally committed to the discovery of the etiologic factors leading to facial development and agree that such data sharing is an important component to the fastest progress in the greater research community. As another example of our commitment to shared resources, Dr. Marazita is a co-PD/PI of the NIDCR-funded FaceBase portal and database (FaceBase, www.FaceBase.org) that is collating genetic and phenotypic data (related to development of the face) from both human and animal models, in an effort to make new connections and provide an invaluable service to the scientific community.

Plans for ensuring the security and integrity of the data, and maintaining privacy of study participants.
Dr. Marazita and colleagues have a great deal of experience in maintaining confidential databases. For example, Dr. Marazita developed and directed the birth defects registry for the State of Virginia over a period of 6 years, and has been involved in numerous large-scale family studies (nationally and internationally), and manages the FaceBase data repository (see above). When appropriate, we will work closely with NIH program staff to ensure that our data is properly de-identified. We have the necessary resources to ensure that our study can follow those protocols.

Specific data sharing plans for the current project.
All subjects are consented for full data-sharing, including the primary facial scans and genetic data. All data developed under FaceBase will be available for data-sharing with any qualified investigators via the NIDCR FaceBase Hub (www.facebase.org/), following approval by the NIDCR FaceBase DAC. With respect to this project, outside investigators given access to the full data will be able to self-define an almost infinite number of orofacial morphometric phenotypes and then study the genetics underlying those phenotypes in humans and/or mice. However, the facial scans can be reconstituted as photographs deemed potentially identifiable by one of the participating IRBs (ie the University of Colorado IRB), whereas dbGaP policy is to accept only de-identified data. Therefore, NIDCR has determined that the full scans will be shared via the FaceBase Hub, and only a set of 15 commonly-used orofacial morphometric measurements derived from the scans, as well as all human genetic data, will be submitted to dbGaP. See the Appendix and the Research Strategy for a list of the morphometric measures to be submitted to dbGaP.