Project Request

Project #20973 : R package development for working with multi-omics human microbiome project data

Project name	**R package development for working with multi-omics human microbiome project data**
Project ID	20973
Approved user name	Mikhail Dozmorov
Institute affiliation	VIRGINIA COMMONWEALTH UNIVERSITY (Non-Profit)
Request date :	Renewal date :

Applicant Organization

Legal Name :	**VIRGINIA COMMONWEALTH UNIVERSITY**						
Department :	Biostatistics			Division :			
Street 1 :	830 East Main Street						
City :	Richmond	State :	VA	Zip :		Country :	USA

PI Contact Information

Name :	**Mikhail Dozmorov**			Position :	Principal Investigator		
Organization:	VIRGINIA COMMONWEALTH UNIVERSITY						
Street 1 :	830 East Main Street						
City :	Richmond	State :	VA	Zip :	23298	Country :	United States
Phone :	804-827-2055	:		Email :	mikhail.dozmorov@vcuhealth.org		

SO Contact Information

Name :	**Tina Cunningham**			Position :	Signing Official		
Organization:	VIRGINIA COMMONWEALTH UNIVERSITY						
Street 1 :	800 E. Leigh Street, Suite 3200						
City :	Richmond	State :	Virginia	Zip :	23298	Country :	USA
Phone :	804-828-6772	:		Email :	tlcunningham2@vcu.edu		

IT Director Contact Information

Name :	**Brian Bush**			Position :	Director of Information Technology		
Organization:	VIRGINIA COMMONWEALTH UNIVERSITY						
Department :	Biostatistics			Division :			
Street 1 :	830 East Main Street						
City :	Richmond	State :	VA	Zip :	23298	Country :	USA
Phone :	804-827-2172	:		Email :	brian.bush@vcuhealth.org		

Approved Research Use Statement

The main objective of the proposed research project is to provide an R software framework for processing, storing and analysis of the 1st and 2nd phase of the Human Microbiome Project (iHMP) data. These data are publicly available via the iHMP data portal (http://portal.hmpdacc.org/) and contains the collection of HMP Core Microbiome Sampling Protocol A (HMP-A, Study Accession: phs000228.v4.p1) (https://www.hmpdacc.org/ihmp/) and Multi-Omic Microbiome Study-Pregnancy Initiative (MOMS-PI) (Study Accession: phs001523.v1.p1). This is the largest collection of publicly available longitudinal multi-omics studies collected to understand the microbial biomarkers of preterm pregnancy, diabetes, and inflammatory bowel disease. While basic clinical annotation information, such as pregnancy or IBD outcome, is available through the data portal, the majority of participants data important for the analysis of disease outcomes, such as age, socioeconomic status, and medical history requires dbGAP access. The goal of this project is to securely integrate these data into the R software framework enabling authorized users the in-depth analysis of -omics data. This project extends our previous work on the integration of microbiome data with the controlled access dbGAP data (https://www.biorxiv.org/content/early/2018/08/29/299115). This study will advance: (1) the accessibility of iHMP data; (2) increase researchers' interest in iHMP data; (3) advance understanding of the connection between multi-omics measurements and physiological factors that cause spontaneous pre-term birth, the progression of type 2 diabetes and IBD. This is the collaborative project, which will include the following investigators: (1) Virginia Commonwealth University: Ekaterina Smirnova, Assistant Professor, Department of Biostatistics; Mikhail Dozmorov, Assistant Professor, Department of Biostatistics, and Jennifer Fettweis, Assistant Professor, Department of Microbiology and Immunology; (2) CUNY School of Public Health: Levi Waldron, Associate Professor, Department of Epidemiology and Biostatistics; (3) Johns Hopkins University: Ni Zhao, Department of Biostatistics.

Non-Technical Summary

The human microbiome contributes to human well-being, disease progression, pregnancy outcomes. Several consortia provide open-access microbiome data; however, software tools for the analysis of it remain undeveloped. This project will develop an R package enabling the integrative analysis of the microbiome data securely integrated with clinical annotations available for authorized users.

Collaborators

Internal

Ekaterina Smirnova
Assistant Professor
VIRGINIA COMMONWEALTH UNIVERSITY -- Department: Biostatistics
830 East Main Street
Richmond, VA 23298 USE
Phone: 804-827-0461 Email: ekaterina.smirnova@vcuhealth.org

Change Log

Date	Changed Details

Consent Group Information

phs000228.v4.p1 : HMP Core Microbiome Sampling Protocol A (HMP-A)

Consent Group : 1

Name : Human microbiome research
Abbreviation : HMP
Request Date :
Use Limitation : These data may be used only for studies related to the human microbiome.

Data Use Certification Agreement

NIH Human Microbiome Project - Core Microbiome Sampling Protocol A (HMP-A)

Introduction and Statement of Policy

The National Institutes of Health (NIH) has developed central data repositories to archive and distribute the results of studies provided by Contributing Investigators examining the relationship between genomic data (e.g., genotype, sequence, or epigenetic information) and phenotype. Such studies include genome-wide association studies, medical sequencing, and molecular diagnostic assays. Implicit in the establishment of the NIH data repositories, for example the database of Genotypes and Phenotypes (dbGaP), is the view that scientific progress in this area will be greatly enhanced if the data produced by these studies are readily available to all investigators in the research community.

Dataset access will be provided to research investigators who, along with their institutions, have certified their agreement with the expectations and terms of access detailed below. It is the intent of the NIH and NHGRI that Approved Users of NIH-provided datasets recognize any restrictions on data use delineated within the original informed consent agreements of contributing studies, as identified by the submitting institutions and stated on database websites.

Definitions of terminology used in this document are found in the Appendix.

The parties to this agreement include: the Principal Investigator (PI) requesting access to the genomic study dataset ("the Approved User"), his/her home institution as represented by the Institutional Signing Official designated through the eRA Commons system ("the Requester"), and the NHGRI, NIH. The effective date of this agreement shall be the Project Approval Date, as specified on the Data Access Committee approval notification.

Terms of Access

1. Research Use

The Requester agrees that if access is approved, the Principal Investigator named in the Data Access Request (DAR) submitted to the NIH, those named in the "Senior/Key Person Profile" portion of the DAR, which should include the Information Technology Director or his/her designee, and any trainee or employee working on the proposed research project under the direct supervision of these individuals, shall become Approved Users of the requested dataset(s). Research use will occur solely in connection with the research project described in the DAR, which includes a 1-2 paragraph description of the research objectives and design. New uses of these data outside those described in the DAR will require submission of a new DAR; modifications to the research project will require submission of an amendment to this application (e.g., the addition of new aims related to the approved project, adding or deleting collaborators from the same institution, and the potential addition of new NIH genomic datasets to an approved project). The Requester and all Approved Users may use the dataset(s) only in accordance with the parameters described on the NIH database Web site for the appropriate research use, and any limitations on such use, of the dataset(s) and as required by law.

Research access to the requested dataset(s) is granted for a period of one (1) year as defined below.

Contributing Investigators, or their direct collaborators, who provided the data or samples used to generate an NIH genomic dataset and who have appropriate IRB approval, if applicable, for broader use of the data are exempt from the limitation on the scope of the research use as defined in the DAR.

2. Institutional and Approved User Responsibilities

The Requester agrees through the submission of the Data Access Request (DAR) that the PI

named in the DAR has reviewed and understands the principles for responsible research use and data handling of the genomic datasets as defined in the NIH GWAS Data Sharing Policy and detailed in this Data Use Certification (DUC) agreement. The Requester and Approved Users further acknowledge that they are responsible for ensuring that all uses of the data are consistent with federal, state, and local laws and regulations and any relevant institutional policies. Through submission of the DAR, the Principal Investigator also agrees to submit annual data use reports to the appropriate NIH Data Access Committee (DAC) describing the research use of the Approved Users as described under *"Research Use Reporting"* below.

Approved Users who may have access to personal identifying information for research participants in the original study at their institution or through their collaborators, may be required to have IRB approval. By approving and submitting the attached Data Access Request, the Institutional Signing Official provides assurance that relevant institutional policies and applicable federal, state, or local laws and regulations (if any) have been followed, including IRB approval if required. The Institutional Signing Official also assures through the approval of the Data Access Request that other organizations within the institution with relevant authorities (e.g., the Office of Human Subjects Research, the Office of Information Technology, the Office of Technology Transfer, etc.) have reviewed the relevant sections of the NIH GWAS Data Sharing Policy and the associated procedures and are in agreement with the principles defined.

It is anticipated that, at least in some cases, these datasets will be updated with additional information. Unless otherwise indicated, all statements herein are presumed to be true and applicable to the access and use of all versions of these datasets.

3. Public Posting of Approved User's Research Use Statement

The Principal Investigator agrees that, if he or she becomes an Approved User, information about the PI and the approved research use may be posted on a public, US government web site that describes approved research projects. The information may include the Approved User's name and institution, project name, Research Use Statement, and a Non-technical Summary of the Research Use Statement. In addition, citations resulting from the use of NIH genomic datasets may be posted on NIH data repository websites.

4. Non-Identification

Approved Users agree not to use the requested datasets, either alone or in concert with any other information, to identify or contact individual participants from whom phenotype data and DNA samples were collected. This provision does not apply to research investigators operating with specific IRB approval, pursuant to 45 C.F.R. 46, to contact individuals within datasets or to obtain and use identifying information under an approved IRB research protocol. All investigators conducting "human subjects research" within the scope of 45 C.F.R. 46 must comply with the requirements contained therein.

5. Non-Transferability

The Requester and Approved Users agree to retain control over the data and further agree not to distribute data obtained through this Data Access Request to any entity or individual not covered in the submitted Data Access Request. If Approved Users are provided access to NIH genomic datasets for inter-institutional collaborative research described in the Research Use Statement of the Data Access Request, and all members of the collaboration are also Approved Users through their home institution(s), data obtained through this Data Access Request may be securely transmitted within the collaborative group. All data security practices and other terms of use defined in this agreement and the *dbGaP Security Best Practices* for the raw data are expected to be followed for the derived data, including any transmission of the data.

The Requester and Approved Users acknowledge responsibility for ensuring the review and agreement to the terms within this Data Use Certification and the appropriate research use of NIH genomic data by research staff associated with any approved project, subject to applicable laws and regulations. NIH genomic datasets obtained through this Data Access Request, in whole or in part, may not be sold to any individual at any point in time for any purpose.

Approved Users agree that if they change institutions during the access period, they will submit a new Data Access Request and Data Use Certification in which the new institution agrees to the NIH GWAS data use policy before data access resumes. Any versions of data stored at the prior institution for the approved use will be destroyed and documented through a final Data Use Report as described below. However, if advance written notice and approval by the NHGRI Data Access Committee is obtained to transfer responsibility for the approved research project to another Approved User within the same institution the data may not need to be destroyed.

6. Data Security and Data Release Reporting

The Requester and Approved Users, including the institutional Information Technology Director or his/her designee, acknowledge the intent of the NIH that they have reviewed and agree to handle the requested dataset(s) according to the current *dbGaP Security Best Practices*, including its detailed description of requirements for security and encryption. These include, but are not limited to:

- all Approved Users have completed all required computer security training required by their institution, for example, the http://irtsectraining.nih.gov/, or the equivalent;
- the data will always be physically secured (for example, through camera surveillance, locks on doors/computers, security guard);
- servers must not be accessible directly from the internet, (for example, they must be behind a firewall or not connected to a larger network) and unnecessary services should be disabled;
- use of portable media, e.g., on a CD, flash drive or laptop, is discouraged, but if necessary then they should be encrypted consistent with applicable law;
- use of updated anti-virus/anti-spyware software;
- security auditing/intrusion detection software, detection and regular scans of potential data intrusions;
- use of strong password policies for file access.
- all copies of the dataset should be destroyed, as permitted by law, whenever any of the following occurs:
  - the DUC expires and renewal is not sought;
  - access renewal is not granted;
  - the NHGRI requests destruction of the dataset;
  - the continued use of the data would no longer be consistent with the DUC.

In addition, the Requester and Approved Users agree to keep the data secure and confidential at all times and to adhere to information technology practices in all aspects of data management to assure that only authorized individuals can gain access to NIH genomic datasets. This agreement includes the maintenance of appropriate controls over any copies or derivatives of the data obtained through this Data Access Request.

Requesters and Approved Users agree to notify the NHGRI Data Access Committee of any unauthorized data sharing, breaches of data security, or inadvertent data releases that may compromise data confidentiality within 24 hours of when the incident is identified. As permitted by law, notifications should include the known information regarding the incident and a general description of the activities or process in place to fully define and remediate the situation. Within 3 business days of the NHGRI Data Access Committee notification, the Requester, through the Approved User and the Institutional Signing Official, agree to submit to the NHGRI Data Access Committee a more detailed written report including the date and nature of the event, actions taken or to be taken to remediate the issue(s), and plans or processes developed to prevent further problems, including specific information on timelines anticipated for action.

All notifications and written reports of data security incidents should be sent to:

NHGRI Data Access Committee at Urgent_NHGRIDAC@mail.nih.gov.

NOTE: Please email NHGRIDAC@mail.nih.gov to communicate with the NHGRI Data Access Committee on all other matters.

The NHGRI, the NIH, or another entity designated by the NIH may, as permitted by law, also investigate any data security incident. Approved Users and their associates agree to support such investigations and provide information, within the limits of applicable local, state and federal laws and regulations. In addition, Requesters and Approved Users agree to work with the NHGRI and the NIH to assure that plans and procedures developed to address identified problems are mutually acceptable consistent with applicable law.

7. Intellectual Property

By requesting access to genomic dataset(s), the Requester and Approved Users acknowledge the intent of the NIH that anyone authorized for research access through the attached Data Access Request follow the intellectual property principles within the NIH GWAS Policy for Data Sharing as summarized below:

> Achieving maximum public benefit is the ultimate goal of data distribution through the NIH genomic data repositories. The NIH believes that these data should be considered as pre-competitive, and urges Approved Users to avoid making IP claims derived directly from the genomic dataset(s). However, the NIH also recognizes the importance of the subsequent development of IP on downstream discoveries, especially in therapeutics, which will be necessary to support full investment in products to benefit the public.
>
> It is expected that these NIH-provided data, and conclusions derived there from, will remain freely available, without requirement for licensing. The NIH encourages broad use of genomic datasets coupled with a responsible approach to management of intellectual property derived from downstream discoveries in a manner consistent with the NIH's Best Practices for the Licensing of Genomic Inventions and the NIH Research Tools Policy.

8. Research Dissemination and Acknowledgement of NIH Genomic Datasets

It is the intent of the NIH to promote the dissemination of research findings from NIH genomic dataset(s) as widely as possible through scientific publication or other appropriate public dissemination mechanisms. Approved Users are strongly encouraged to publish their results in peer-reviewed journals and to present research findings at scientific meetings, etc.

In accord with the NIH GWAS Policy for Data Sharing, and as expressed through the submission of the DAR, Approved Users acknowledge the NIH's expectation **that they will not *submit* findings using the NIH Human Microbiome Project - Core Microbiome Sampling Protocol A (HMP-A), or updated versions thereof, for publication or presentation for a period of exclusivity for Contributing Investigators concluding with the Embargo Date identified on the dbGaP or other NIH genomic data repository homepage.**

Approved Users agree to acknowledge the NIH data repository, the Contributing Investigator(s) who contributed the phenotype data and DNA samples from his/her original study, and the primary funding organization that supported the contributing study in all oral and written presentations, disclosures, and publications resulting from any analyses of the data. Approved Users further agree that the acknowledgment shall include the dbGaP accession number to the specific version of the dataset(s) analyzed.

A sample statement for the acknowledgment of the NIH Human Microbiome Project - Core Microbiome Sampling Protocol A (HMP-A) dataset(s) follows:

Funding support for the development of NIH Human Microbiome Project - Core Microbiome Sampling Protocol A (HMP-A) was provided by the NIH Roadmap for Medical Research.

Clinical data from this study were jointly produced by the Baylor College of Medicine and the Washington University School of Medicine. Sequencing data were produced by the Baylor College of Medicine Human Genome Sequencing Center, The Broad Institute, the Genome Center at Washington University, and the J. Craig Ventor Institute. These data were submitted by the EMMES Corporation, which serves as the clinical data collection site for the HMP. Publications resulting from this dataset should acknowledge the above mentioned groups. Publications resulting from this dataset should also cite the original publication of the HMP

analyzing these data.

9. Research Use Reporting

To assure that NIH policies and procedures for genomic data use are adhered to, Approved Users agree to provide to the NHGRI Data Access Committee annual feedback on how these data have been used and any results that have been generated as a result of access to the data, including patents and publications. This information will be used by the NHGRI Data Access Committee staff for program evaluation activities, and may be considered by the NIH GWAS Governance committees as part of the NIH effort to provide ongoing oversight and management of all NIH genomic data sharing activities.

Approved Users thus agree to provide a brief Annual Data Use Report on the research specified within the DAR submitted with this DUC. Approved Users who are seeking renewal agree to provide specific information in a renewal DAR. Those not seeking renewal agree to provide specific information to the Data Access Committee via the contact information below. Annual Data Use Reports will provide information regarding potentially significant findings and publications or presentations that resulted from the use of the requested dataset(s), a summary of any plans for future research use, any violations of the terms of access described within this Data Use Certification and the implemented remediation, and information on any downstream intellectual property generated as a result of the data. Approved Users also may include general comments regarding topics such as the effectiveness of the NIH genomic data access process (e.g., ease of access and use), appropriateness of data format, challenges in following the policies, and suggestions for improving data access or the program in general if desired.

Approved Users agree to send the Annual Data Use Report prior to the anniversary of the Approved Access Date assigned by the DAC and specified within the manifest file provided to Approved Users by the NIH Data Repository at the time that data access is provided. It is agreed that the Annual Data Use Report will be shared with the NIH within the context of a renewal Data Access Request, or via a letter signed by the Institutional Signing Official and the Approved User.

Annual Data Use Reports should be submitted to:

The NHGRI Data Access Committee by e-mail at NHGRIDAC@mail.nih.gov, unless otherwise indicated in automated reminder messages from NCBI/dbGaP. Requests for continued data access should be made through dbGaP.

Note that any inadvertent or inappropriate data release incidents should be reported to the NHGRI Data Access Committee according to the agreements and instructions under Term 6.

10. Non-Endorsement, Indemnification

The Requester and Approved Users acknowledge that although all reasonable efforts have been taken to ensure the accuracy and reliability of NIH genomic data, the NIH, the NHGRI Data Access Committee, and Contributing Investigators do not and cannot warrant the results that may be obtained by using any data included therein. The NIH, the NHGRI Data Access Committee, and all contributors to these datasets disclaim all warranties as to performance or fitness of the data for any particular purpose.

No indemnification for any loss, claim, damage or liability is intended or provided by any party under this agreement. Each party shall be liable for any loss, claim, damage, or liability that said party incurs as a result of its activities under this agreement, except that the NIH, as an agency of the United States, may be liable only to the extent provided under the Federal Tort Claims Act, 28 U.S.C. 2671 et seq.

11. Termination and Violations

This Data Use Certification will be in effect for a period of one (1) year from the date the dataset(s) are made accessible to the Approved User ("Approved Access Date"). At the end of the access period, Approved Users agree to destroy all copies of the requested dataset(s), except as required by publication practices or law to retain them.

Consideration will be given to a renewal of this agreement upon submission of a new DAR. Copies of NIH genomic dataset(s) may not need to be destroyed if, with advance notice and approval by the NHGRI Data Access Committee, the project has been transferred to another Approved User. In this case, documentation must be provided that other Approved Users are using the dataset(s) under an active DAC approved research project at the same institution.

The Requester and Approved User acknowledge that the NIH or the NHGRI may terminate this agreement and immediately revoke access to all NIH genomic datasets at any time if the Requester is found to be no longer in agreement with the policies, principles and procedures of the NIH and the NHGRI.

****

By submission of the attached Data Access Request, the Requester through the Institutional Signing Official attests to the Approved Users' qualifications for access to and use of NIH genomic dataset(s) and certifies their agreement to the NIH principles, policies and procedures for the use of the requested datasets as articulated in this document, including the potential termination of access should a violation of any of these agreement terms be identified.

Requesters and the Principal Investigator further acknowledge that they have shared this document and the NIH GWAS data sharing policies and procedures for access and use of genomic datasets with any Approved Users, appropriate research staff, and all other Key Personnel identified in the DAR.

Institutional Signing Officials acknowledge that they have considered the relevant NIH GWAS policies and procedures, that they have shared this document and the relevant policies and procedures with appropriate institutional organizations, and have assured compliance with local institutional policies related to technology transfer, information technology, privacy, and human subjects research.

Appendix

Definitions of Terminology

Annual Data Use Report: A report submitted to the DAC on the anniversary of access approval summarizing the analysis of NIH genomic datasets obtained through the Data Access Request and any significant findings derived from the work.

Approved User: Post-DAC approval will include the PI, collaborators at the home institution who are named in the "Senior/Key Person Profile" portion of the DAR, the IT Director or designee named in the "Senior/Key Person Profile" portion of the DAR, and trainees or staff to these investigators.

Contributing Investigator: The researcher who submitted the genomic dataset to dbGaP.

Data Access Request: SF 424 (R&R) cover pages and requested attachments, if any.

Data Derivative: any data including individual-level data or aggregate genomic data that stems from the original dataset obtained through dbGaP. Excepted from this term is summary information that is expected to be shared through community publication practices.

Final Data Use Report: A final report submitted to the DAC at the conclusion of the approved access period when no additional access is sought, or when leaving an institution. This report should summarize the analysis of genomic study datasets obtained through the Data Access Request and any significant findings derived from the work.

Information Technology Director: Someone with the authority to vouch for the IT capacities at an institution, or higher-level division of an institution (e.g., the School of Medicine).

Institutional Signing Official: Someone with the authority to sign on behalf of the Requester and credentialed through the eRA system as such.

Requester: The home institution/organization for the Primary Investigator (PI) that will use the requested data.

Senior/Key Persons: Collaborators at the home institution, and the IT Director or designee.

Addendum to the Data Use Certification Agreement
Modification of Data Security Terms and Best Practices

Effective for all dbGaP Data Access Requests submitted on or after March 23, 2015,
Section 6 of the Data Use Certification Agreement is replaced in its entirety by the following:

6. Data Security and Data Release Reporting

The Requester and Approved Users, including the institutional IT Director, acknowledge NIH's expectation that they have reviewed and agree to manage the requested dataset(s) according to the current NIH Security Best Practices for Controlled-Access Data Subject to the GDS Policy and the institutional IT security requirements and policies, and that the institution's IT security requirements and policies are sufficient to protect the confidentiality and integrity of the NIH controlled-access data entrusted to the Requester.

If approved by NIH to use cloud computing for the proposed research project, as outlined in the Research and Cloud Computing Use Statements of the Data Access Request, the Requester acknowledges that the IT Director has reviewed and understands the cloud computing guidelines in the NIH Security Best Practices for Controlled-Access Data Subject to the GDS Policy.

Requesters and PIs agree to notify the NHGRI DAC of any unauthorized data sharing, breaches of data security, or inadvertent data releases that may compromise data confidentiality within 24 hours of when the incident is identified. As permitted by law, notifications should include any known information regarding the incident and a general description of the activities or process in place to define and remediate the situation fully. Within 3 business days of the NHGRI DAC notification, the Requester, through the PI and the Institutional Signing Official, agree to submit to the NHGRI Data Access Committee a detailed written report including the date and nature of the event, actions taken or to be taken to remediate the issue(s), and plans or processes developed to prevent further problems, including specific information on timelines anticipated for action.

All notifications and written reports of data security incidents should be sent to:
NHGRI Data Access Committee URGENT: nhgridac@mail.nih.gov
GDS mailbox: gds@mail.nih.gov

NIH, or another entity designated by NIH may, as permitted by law, also investigate any data security incident. Approved Users and their associates agree to support such investigations and provide information, within the limits of applicable local, state, and federal laws and regulations. In addition, Requesters and Approved Users agree to work with the NHGRI and NIH to assure that plans and procedures that are developed to address identified problems are mutually acceptable and consistent with applicable law.