IBD Genetics

International IBD Genetics Consortium (IIBDGC)

Version_28.January.2019

Memorandum of Understanding

This Memorandum of Understanding (MOU)describes the arrangement by which members of the International IBD Genetics Consortium (IIBDGC) will submit data for the purpose of group projects, and in return have the opportunity to participate in those projects and propose projects of their own. Such data will be referred to below as "IIBDGC-managed data." This MOU has been approved by the IIBDGC Management Committee (MC)—any changes will require approval of the MC.

1. Mission of the International IBD Genetics Consortium

The mission of the IIBDGC is to (1) identify the genetic risk factors for the inflammatory bowel diseases (IBD), Crohn's disease (CD), ulcerative colitis (UC), and related clinical phenotypes, (2) to determine how these risk factors interact with non-genetic risk factors impact on an individual's risk to develop IBD, CD, UC and related phenotypes, (3) perform these studies in a collaborative fashion and to broadly share the results with the international scientific community in a responsible fashion. While the original projects were limited to genomic DNA genotype data, the IIBDGC will incorporate all different data types generated on different sample types that are necessary to achieve its mission.

2. IIBDGC Management Committee (MC)

The Management Committee (MC) is responsible for (1) guiding the scientific direction of the IIBDGC, (2) providing oversight of IIBDGC projects, (3) communication with IIBGC members, (4) conflict resolution, (5) outreach to the broader IBD and general scientific communities.

3. General Requirements for IIBDGC Membership

Investigators wishing to participate in IIBDGC projects must comply with the following requirements:

- Submit data from a minimum of 500 research participants; exceptions to permit submitting fewer samples (at least 100) for cohorts of particular interest (e.g., rare phenotypes) will require approval by the MC and will be considered on a case-by-case basis.
- Submit corresponding core phenotype data as described in Cleynen et al. (PMID: 26490195), formatted according to the IIBDGC's Core Sub-Phenotype Submission Guidelines (***see Appendix-II***); specific projects may have additional phenotype requirements.
- Submit individual-level genotype, sequence, and/or other genomic data; if that is not possible, the MC may approve submission of summary data on a case-by-case basis. Genetic/genomic data should be submitted in accordance with the requirements of the specific project(s) for which they are intended. These project-specific requirements are outlined in Appendices to this document.
- Provide a letter stating that you have authority from your Institutional Review Board (IRB) or equivalent to share the data with the IIBDGC *and* in a public repository such as dbGaP or EMBL-EBI; if you do not have authority to share individual-level genotype data, the MC will consider permitting you to share summary data instead. Those who have already submitted their data to a public repository need only affirm this below and indicate how the data may be accessed from that location.
- Adherence to the principles outlined in the current document.

4. IIBDGC Projects & Working Groups: Proposals, Inclusiveness, and Transparency

Access and use of IIBDGC-managed data is limited to projects that have obtained MC approval. Any IIBDGC member (or group of members) wishing to initiate a project must complete a Project Proposal form. All approved Project Proposals are made available to the full IIBDGC membership. Project Leaders are required to (1) form a Working group that ensures the successful completion of the project, (2) make Working Group activities open to all IIBDGC members, (3) report progress to the MC on a regular basis as defined by the MC, (4) ensure that all communications of research results (abstracts, manuscripts, etc.) obtain prior MC approval.

5. Publication Policy

IIBDGC Working Groups are responsible for drafting communication of research results (abstracts, manuscripts, etc.). IIBDGC Working Groups are responsible for drafting author lists. Author lists will reflect the contributions of IIBDGC members and IIBDGC Working Groups. Author lists should include the IIBDGC banner author. Working Groups must obtain MC approval of documents (e.g. manuscripts) and author lists prior to submission.

6. Intellectual Property

The IIBDGC aims to make all of its results and findings publicly available in a timely and responsible manner. The IIBDGC will not pursue the protection/filings of any intellectual property that it generates. IIBDGC members are free to pursue intellectual property protection of their own data at any time, but such IP-related activities can't prevent or delay the public release or publication of IBDGC results and findings. Other than the members own data, use of IIBDGC-managed data or results in intellectual property filings is allowed only after public release.

7. Data Management Requirements

All IIBDGC-managed data (e.g. genotype, phenotype, demographic, etc.) will be securely stored and managed by the NIDDK IBD Genetics Consortium (IBDGC) Data Coordinating Center (DCC) on behalf of the IIBDGC members (see below for details regarding Data Commons). All data should be submitted with a local participant or sample ID containing no identifying information (identifying information such as names, medical record numbers, or other will not be accepted). Upon receipt of new data, the DCC will assign a unique ID that will be used for all subsequent distribution and analysis.

The DCC will submit data to dbGaP on behalf of the IIBDGC, under the direction of the IIBDGC MC, unless those data have already been deposited elsewhere (see section below on Data Sharing).

It is possible that for specific projects some data types (e.g. large-scale sequence data) may be stored on another data platform at the request of the IIBDGC MC. In such circumstances, equivalent Data Management Requirements will be respected and will be made transparent to IIBDGC members via Project description included as an Appendix to the current MOU.

8. Data Access

Data access will be provided through the NIDDK IBDGC Data Commons (https://ibdgc.datacommons.io/) that uses the resources of the University of Chicago's Data Commons (https://ctds.uchicago.edu/datacommons/). Specific projects may also request access to compute resources within the commons. Before accessing the commons, users are required to provide proof that they have completed the required Human Subjects training at their local institution (e.g., a copy of their certificate) as well as complete the NIH Secure Remote Computing training course (https://irtsectraining.nih.gov/publicuser.aspx). Certificates of completion should be sent to Sondra Birch (sbirch@uchicago.edu), together with your Gmail address (authentication to the commons is through Google's OAuth).

Investigators who have submitted data as well as other researchers and students working directly under their supervision may access IIBDGC-managed data for two purposes:

- Investigators may access and download their own data at any time without explicit permission from the MC. This MOU shall in no way limit an investigator's ability to use and publish his or her own data.
- Shared data may be accessed ONLY as part of an MC-approved project. Project proposals must be submitted to the MC for approval *before* data are accessed. In some cases, a project may also require completion of one or more additional Data Use Agreements (DUAs). Data accessed for a specific project must be used solely for completing the objectives of that project as described in the project proposal. The project leader(s) are responsible for ensuring that data are being used for this purpose, and that they are being handled securely (including that any additional special restrictions attaching to a particular dataset are being followed). No sharing or use of IIBDGC-managed data outside of MC-approved projects is permitted. No publications of results obtained from IIBDGC-managed data is allowed without MC approval of the manuscript.

Investigators who have not submitted data but wish to use IIBDGC-managed data may also submit a project proposal to the MC.

IMPORTANT: Absolutely no redistribution of IIBDGC-managed data is permitted with the exception of an investigator's own data. Any data downloaded for local analysis must be handled and stored securely.

In the event that other/additional data management/data access infrastructure is required in the future, such a modification would require IIBDGC MC approval. It is possible that for specific projects some data types (e.g. large-scale sequence data) may be stored on another data platform at the request of the IIBDGC MC. In such circumstances, equivalent Data Access procedures will be respected and will be made transparent to IIBDGC members via Project description included as an Appendix to the current MOU.

9. Data Sharing

Individual-level genetic and/or genomic data (or in cases where that is not possible, summary data) will be submitted to dbGaP by the DCC within one year of the date on which the data became available for analysis. Data that have already been deposited in a public repository will not be submitted to dbGaP, but instead instructions will be included on how the data may be accessed.

10. Use of the IIBDGC name

No Participant shall use the name or logo of the IIBDGC for any commercial purposes (e.g. promotional material or other public announcement or disclosure) without the prior written consent of the MC.

11. Conflict Resolution

If any unresolved conflicts arise, the MC must be informed. MC decisions will be final.

12. Electronic Execution

Each Member intends that an electronic copy of its signature stored in a PDF software application format shall be regarded as an original signature and agrees that this MOU may be executed in any number of counterparts, each of which shall be effective upon delivery and thereafter shall be deemed an original, and all of which shall be taken to be one and the same instrument.

Acknowledgement of Memo of Understanding

I understand and accept all of the requirements and responsibilities outlined above. I understand that failure to comply with this MOU may result in loss of access to IIBDGC-managed data as well as a notification being sent to my institutional representative.

Name and Signature

Institutional Affiliation

Institutional Address

Telephone

Email

Are you representing a group a local/regional/national group of researchers/clinicians (Y/N)?

Name of local/regional/national group that you are representing or indicate “not applicable” ***

Name and Signature ***

Date

*** Please complete and return the “MOU list of contributors and Key Personnel.xlsx”

APPENDIX – I

IIBDGC List of Contributors and Key Personnel

Please fill out and return the attached excel spreadsheet with the names and contact information for contributors to the cohort and the key personnel from your group that will require access to the IIBDGC Data Commons (maintained by the NIDDK IBDGC Data Coordinating Center on the University of Chicago's Data Commons, on behalf of the IIBDGC).

APPENDIX – II

International IBD Genetics Consortium Core Sub-Phenotype Submission Guidelines

Please see attached document

International IBD Genetics Consortium Core Sub-Phenotype Submission Guidelines

Contact: pschumm@uchicago.edu

Revision: 1.1
Date: 2019-01-15

Table of Contents

General Instructions

The objective of this effort is to collect data on a core set of sub-phenotype fields for all of the subjects included in the international GWAS and ImmunoChip cohorts. This request is being made of all those who submitted genotype data to the International Consortium. We are providing a list of the samples you submitted using your original IDs, and request that you submit data for these samples only, using the exact same IDs. If you wish to use another set of IDs, or if you find a problem with the IDs you submitted, please let us know.

Those of you who have all of your phenotype data in an electronic database and who have sent us your database schema may, if you wish, simply dump or export *all* of your data for the subjects in question and submit them to us. As I've said before, the advantage of this is that if we decide we want additional fields in the future, you won't have to resubmit (assuming those fields are already present in your database). Also, you won't have to prepare a file as described below. If you'd like to do this, please email Phil Schumm (pschumm@uchicago.edu) to arrange for the transfer.

For those preparing their own upload file(s), please make sure to follow the specifications provided in the Field Definitions carefully. Files that do not conform to these specifications cannot be processed.

Note that data for different disease types (i.e., Crohn's disease, ulcerative colitis and indeterminate colitis) may be combined in the same file, as may data for both affected and unaffected subjects. To do this, simply leave blank the fields that do not apply to a given subject. If you prefer, you may upload separate files.

Required Fields

All fields marked as required must be included. The remaining fields should be submitted if you have those data in electronic form, or if you can enter them within two weeks. If this is not the case, we request that you submit your data without these items, so as not to delay our initial compilation of the data. You may then submit these fields by themselves at any time afterward.

Here is a list of the fields that are not initially required:

- Year started smoking
- Year stopped smoking
- Cigarettes per day
- Number of operations for abdominal disease
- Oral steroids
- IV steroids
- Anti-TNF or cyclosporine
- Indication for anti-TNF/cyclosporine
- Other immunomodulatory drugs

File Format

1) The preferred file format for submissions is tab-delimited ASCII (UTF-8 encoding), with the first row containing the column names. Alternatively, an Excel file may be submitted, also with column names in the first row.

2) If you do submit an Excel file, please note that any special formatting (e.g., the use of bold, color, highlighting, etc.) will be ignored, as only the data themselves will be read out of the file.

3) If you wish to leave a field blank because the item is inapplicable or the information is unavailable, simply leave the corresponding cell empty (do this for both string and numeric fields). *Do not* use zeros, periods, or other characters to indicate missing data. Note that this is different from the response "Unknown" which should be used as indicated in the field definitions below.

4) All year fields should be supplied as 4-digit integers (i.e., you must include the century).

5) Do not enclose string values in quotes.

File Submission

When you are ready to submit your file(s), please upload them to your home directory on the SFTP server (i.e., just as you did with the genotype data), and email Phil Schumm.

Field Definitions

All Subjects

The following information is requested for all subjects (both affected and unaffected).

Identifying Information

As noted above, we have provided a list of the samples you submitted and request that you use these same sample IDs when submitting your phenotype data. In addition, please indicate the center from which the sample was initially obtained, and, if the sample IDs provided by the supplying center differ from those you previously submitted with the genotype data, please provide those IDs as well. This is to facilitate identification of overlap between different submissions, and to facilitate future communication with the supplying center (if necessary).

Sample ID

Field name: `orig_sample_id`

Description: Sample ID supplied with genotype data

Type: String

Validation: Required; must be identical to corresponding value in the sample list you were provided.

Supplying center

Field name: `center`

Description: Name of center supplying sample

Type: String

Validation: Required

Note

Please use a consistent designation within your submission (i.e., if two samples come from the same center, their values for `center` should be identical).

Supplying center sample ID

Field name: `center_sample_id`

Description: Sample ID provided by the supplying center

Type: String

> **Note**
>
> Required if and only if different from `orig_sample_id`.

Affection Status, Disease Type and Basic Demographics

These fields were requested during submission of the ImmunoChip genotype data, and sex, affection status and disease type are also present in the GWAS data. However, because there was some missing data for these fields, and since these fields should be easily accessible to you, we are asking that you resubmit them. This may permit us to fill in some gaps, but perhaps more importantly, will help serve as a sanity check on the submission process. If we find inconsistencies between these fields and the data you submitted previously, we shall work with you to resolve them.

Sex

Field name: `sex`

Type: String

Values: "Male", "Female", "Unknown"

Validation: Required

Year of birth

Field name: `yob`

Type: Integer

Range: 1900 or greater

> **Note**
>
> This is a required fieldleaving it blank indicates that it was not collected and/or cannot be retrieved.

IBD affection status

Field name: `affection`

Type: String

Values: "Unaffected", "Affected", "Unknown"

Validation: Required

Disease type

Field name: `diag`

Type: String

Values: "Crohn’s Disease", "Ulcerative Colitis", "Indeterminate", "Unknown"

Validation: Required if `affection` equals "Affected"; otherwise must be left blank

Unrelated control

Field name: `control`

Description: Indicator for unrelated, healthy controls (1 for controls, 0 otherwise)

Type: Numeric

Values: 0, 1

Validation: Required

Note

Those designated as controls should have no family history of IBD.

Race

Field name: `race`

Type: String

Values: "American Indian/Alaskan Native", "East-Asian", "South-Asian", "Asian (Unspecified)", "Black/African American", "Middle-Eastern", "Native Hawaiian/Pacific Islander", "Other", "Unknown", "White"

Validation: Required

Other race (specify)

Field name: `race_other`

Type: String

Max length: 80

Validation: Required if `race` equals "Other"; otherwise must be blank

Hispanic

Field name: `hispanic`

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Jewish ethnicity

Field name: `jewish`

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

> **Note**
>
> Should ideally be determined based on status of grandparents, with those who have two or more Jewish grandparents being marked "Yes"; may be based on self-report if information on grandparents is unavailable.

Smoking History

Smoking status

Field name: `smoking`

Description: Smoking status at diagnosis/ascertainment (i.e., within the three months immediately prior)

Type: String

Values: Affected subjects: "Yes", "Ex-smoker", "No", "Unknown"; Unaffected subjects: "Current smoker","Ex-smoker","Non-smoker","Unknown"

Validation: Required

> **Note**
>
> For affected subjects this field represents smoking status at diagnosis of IBD, whereas for unaffected subjects it represents smoking status at time of ascertainment. Note that the possible values differ in these two cases.

Year started smoking

Field name: `smoking_start`

Type: Integer

Range: 1900 or greater

Validation: Allowed only if `smoking` equals "Yes", "Current smoker" or "Ex-smoker". Must be greater than or equal to `yob`. For affected subjects, must be less than or equal to `diag_yr`.

Year stopped smoking

Field name: `smoking_stop`

Type: Integer

Range: 1900 or greater

Validation: Allowed only if `smoking` equals "Ex-smoker". Must be greater than or equal to `smoking_start`. Must be greater than or equal to `yob`. For affected subjects, must be less than or equal to `diag_yr`.

Cigarettes per day

Field name: `nocigar_day`

Description: Average number of cigarettes smoked per day

Type: Integer

Range: 0 or greater

Validation: Allowed only if `smoking` equals "Yes", "Current smoker" or "Ex-smoker".

Affected Subjects Only

Diagnostic Information

Year of diagnosis

Field name: `diag_yr`

Type: Integer

Range: 1900 or greater

Validation: Must be greater than or equal to `yob` and less than or equal to `review_yr`

> **Note**
>
> This is a required fieldleaving it blank indicates that it was not collected and/or cannot be retrieved.

Age at diagnosis

Field name: `diag_age`

Type: Integer

Range: [0, 99]

> **Note**
>
> This is an alternative field for use in cases where the year of diagnosis is unavailable; otherwise, it should be left blank.

Year of last review

Field name: `review_yr`

Description: Year in which the subject's phenotype data were last updated

Type: Integer

Range: 1900 or greater

Validation: Must be greater than or equal to both `yob` and `diag_yr`

> **Note**
>
> This is a required fieldleaving it blank indicates that it was not collected and/or cannot be retrieved.

Family history of IBD

Field name: `family_history`

Description: Family history of IBD in first *or* second degree relatives

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Family type

Field name: `family_type`

Type: String

Values: "CD", "UC", "Mixed", "Unknown"

Validation: Required when `family_history` is equal to "Yes"; otherwise must be left blank

Extra-Intestinal Manifestations

Primary sclerosing cholangitis

Field name: `liver_psc`

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Ankylosing spondylitis

Field name: `joint_as`

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Disease Treatment

Oral steroids

Field name: `oralster_diag`

Description: Ever received oral steroids (for treatment of IBD)?

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

IV steroids

Field name: `ivster_diag`

Description: Ever received IV steroids (for treatment of IBD)?

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Anti-TNF or cyclosporine

Field name: `atnfcyclo_diag`

Description: Ever received anti-TNF or cyclosporine (for treatment of IBD)?

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Indication for anti-TNF/cyclosporine

Field name: `atnfcyclo_ind`

Description: If received anti-TNF or cyclosporine, was this for treatment of an acute flare, chronic refractory IBD, or both?

Type: String

Values: "Acute flare", "Chronic refractory IBD", "Both", "Unknown"

Validation: Required when `atnfcyclo_diag` is equal to "Yes"; otherwise must be left blank

Other immunomodulatory drugs

Field name: `immuno_diag`

Description: Ever received other immunomodulatory drugs (for treatment of IBD)?

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Crohn's Disease Only

Macroscopic Disease Location

Upper GI

Field name: `dis_loc_gi`

Description: Disease location: Upper GI

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Jejunum

Field name: `dis_loc_jejunal`

Description: Disease location: Jejunum

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Ileum

Field name: dis_loc_ileal

Description: Disease location: Ileum

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Colorectal

Field name: dis_loc_colorectal

Description: Disease location: Colorectal

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Perianal/Perineal

Field name: dis_loc_perianal

Description: Disease location: Perianal/Perineal

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

CD disease behavior

Field name: dis_behavior

Type: String

Values: "B1", "B2", "B3", "Unknown"

Validation: Required

> **Note**
> Following Montreal classification, presence of B3 always trumps B2.

Surgery

Abdominal surgery for CD

Field name: surgery

Description: Abdominal surgery for complication or treatment of CD

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Year of first operation

Field name: `surgery_yr`

Type: Integer

Range: 1900 or greater

Validation: Permitted only if `surgery` equals "Yes"; otherwise must be blank. Must be greater than or equal to both `yob` and `diag_yr`, and must be less than or equal to `review_yr`.

> **Note**
>
> This is a required fieldleaving it blank indicates that it was not collected and/or cannot be retrieved.

Number of operations for abdominal disease

Field name: `operation_ad`

Description: Number of operations for abdominal disease

Type: Numeric

Range: [0, 21]

Validation: Permitted if `surgery` equals "Yes"; otherwise must be blank

UC/Indeterminate Only

Macroscopic Disease Location

Proctitis

Field name: `dis_loc_proctitis`

Description: Disease location: Proctitis

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Left-sided

Field name: `dis_loc_left`

Description: Disease location: Left-sided (to splenic flexure)

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Extensive

Field name: `dis_loc_extensive`

Description: Disease location: Extensive (beyond splenic flexure)

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Surgery

Colectomy

Field name: `surgery`

Description: Colectomy for complication or treatment of UC/IC

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required

Year of surgery (colectomy)

Field name: `surgery_yr`

Type: Integer

Range: 1900 or greater

Validation: Permitted only if `surgery` equals "Yes"; otherwise must be blank. Must be greater than or equal to both `yob` and `diag_yr`, and must be less than or equal to `review_yr`.

> **Note**
>
> This is a required fieldleaving it blank indicates that it was not collected and/or cannot be retrieved.

Surgery for dysplasia/cancer

Field name: `surgery_dysplasia`

Description: Indication for colectomy: Dysplasia/cancer

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required if `surgery` equals "Yes"; otherwise must be blank

Surgery for chronic continuous disease

Field name: `surgery_chronic`

Description: Indication for colectomy: Chronic continuous disease

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required if `surgery` equals "Yes"; otherwise must be blank

Surgery for acute fulminant disease

Field name: `surgery_acute`

Description: Indication for colectomy: Acute fulminant disease

Type: String

Values: "Yes", "No", "Unknown"

Validation: Required if `surgery` equals "Yes"; otherwise must be blank

APPENDIX - III

Genome-Wide Association Study 2019 - Requirements & Data Use

Primary objective of this IIBDGC Project

- Identification of genomic regions, genes, genetic variants, haplotypes associated with CD, UC, IBD and associated clinical phenotypes

Details of Genome-Wide Association Study

This collaborative project aims to assemble the largest number of IBD cases (and controls) from the broadest set of worldwide populations to advance our understanding of the genetic basis of IBD and associated phenotypes.

In order to participate in the above-mentioned study, I agree to the following:

- *Submit genotype data according to the "IIBDGC Instructions for GWAS Data Submission"*
- *Submit phenotype data according to the "International IBD Genetics Consortium Core Sub-Phenotype Submission Guidelines"*
- *Genotype data should be submitted no later than 2019-02-28; complete phenotype data are due by 2019-05-10*
- Certify that I am not aware of any other group that is also submitting this data in whole or part

Data Use - Please check one of the following:

□ I am including a letter authorizing me to share individual-level data in a public repository such as dbGaP, and have indicated which (if any) restrictions must be placed their use

□ I have already deposited my data with a public repository such as dbGaP or EMBL-EBI, and am including instructions on how to access them

□ I am not permitted to share individual-level data beyond the IIBDGC, but understand that my summary data will be deposited in a public repository

□ I have permission from the MC to submit only summary data to the IIBGC, and understand that these summary data will be deposited in a public repository

IIBDGC Instructions for GWAS Data Submission

These instructions are for submission of individual-level genotype and phenotype data as part of Phase I of the International IBD Genetics Consortium's (IIBDGC) new GWAS project. Phase II will include collection of summary statistics for those collaborators unable to share individual-level data.

Before submitting data, investigators must provide a signed copy of the IIBDGC Memo of Understanding to the Data Coordinating Center (DCC) of the NIDDK IBD Genetics Consortium. Upon receipt of the MOU, you (and your research staff) will be given access to the IBD Data Commons.

Genotype data

Illumina: Genotype data generated on the Illumina platform should be submitted in the form of a tab-delimited Final Report file, as produced by the Genome Studio software. When preparing the final report, please export all available fields, including the following:

> SNP Name, Sample ID, Allele1 – Top, Allele2 – Top, GC Score, Sample Name, Sample Group, Sample Index, SNP Index, SNP Aux, Allele1 – Forward, Allele2 – Forward, Allele1 – Design, Allele2 – Design, Allele1 – AB, Allele2 – AB, Chr, Position, GT Score, Cluster Sep, SNP, ILMN Strand, Customer Strand, Top Genomic Sequence, Theta, R, X, Y, X Raw, Y Raw, B Allele Freq, Log R Ratio, CNV Value, CNV Confidence

Data for all samples with at least a 90% call rate should be submitted (samples with lower call rates may also be included but are not required).

Affymetrix: TBD

Sample information

In a separate, tab-delimited file, please provide the information listed in the IIBDGC Core Sub-Phenotype Submission Guidelines under the sections Identifying Information and Affection Status, Disease Type and Basic Demographics. Data should be coded and formatted exactly as described in the guidelines.

In addition, please include a column indicating whether Immunochip data were submitted for this individual as part of the IIBDGC's Immunochip project. Submitting GSA data using the same sample IDs that were used for the Immunochip project will permit us to link the samples across projects.

Phenotype Data

The remaining phenotype data described in the Sub-Phenotype Submission Guidelines should be submitted by May 10th 2019 (the week before DDW 2019).

Public Data Sharing

As described in the MOU, data will be deposited with dbGaP (or EMBL-EBI, if appropriate) by the Data Coordinating Center for the NIDDK IBD Genetics Consortium (IBDGC DCC). To ensure that data are shared in accordance with the consent provided by this, please add a column to the Sample Information file indicating for each sample one of the following four consent groups:

1. General Research Use
2. Health/Medical/Biomedical
3. IBD research only
4. Summary statistics only

Access to all individual-level genotype data (1–3 above) *will require IRB approval*. Per new NIH guidelines, summary statistics for all samples will be made available without IRB approval.

Finally, please include a signed letter indicating that you have authority to share the data through dbGaP. The letter should be directed to:

Dr. Ronald A. Thisted
Scientific Director, NIDDK IBDGC DCC
Department of Public Health Sciences
University of Chicago

Data Submission

Data should be uploaded to the SFTP server maintained by the IBDGC DCC. To gain access to the server, please send an RSA key (at least 2048 bits long) to Phil Schumm (pschumm@uchicago.edu). If you need help in generating a key, please see the instructions (included in a separate document).

APPENDIX - IV

Whole Exome Sequencing Study 2019 - Requirements & Data Use

Primary objective of this IIBDGC Project

- Identification of genomic regions, genes, genetic variants, haplotypes associated with CD, UC, IBD and associated clinical phenotypes

Description of Whole Exome Sequencing Study

Please refer to APPENDIX V (entitled "*Overview of Broad IBD Exome Sequencing Program)* for an overview of the project that has been led by Mark Daly. Additional data (WES and/or WES data extracted from WGS) is welcome. VCF files generated as part of this project will be mirrored on the IIBDGC Data Commons described in the MOU.

In order to participate in the above-mentioned study, I agree to the following:

- *Submit sequence data according to the "IIBDGC Instructions for Sequence Data Submission"*
- *Submit phenotype data according to the "International IBD Genetics Consortium Core Sub-Phenotype Submission Guidelines"*
- *Genotype data should be submitted no later than 2019-02-28; complete phenotype data are due by 2019-05-10*
- Certify that I am not aware of any other group that is also submitting this data in whole or part

Data Use - Please check one of the following:

□ I am including a letter authorizing me to share individual-level data in a public repository such as dbGaP, and have indicated which (if any) restrictions must be placed their use

□ I have already deposited my data with a public repository such as dbGaP or EMBL-EBI, and am including instructions on how to access them

□ I am not permitted to share individual-level data beyond the IIBDGC, but understand that my summary data will be deposited in a public repository

□ I have permission from the MC to submit only summary data to the IIBGC, and understand that these summary data will be deposited in a public repository

APPENDIX - V

Overview of Broad IBD Exome Sequencing Program

With support from the Helmsley Charitable Trust and NHGRI, we are launching a transformative program of exome sequencing in IBD for which we are currently engaging collaborative partners. The overarching aim of the program is to define the full allelic spectrum of protein-altering variation in genes associated to IBD, assess their role in both CD and UC, and determine whether loss-of-function variants confer risk or protection in each gene in order to articulate the most opportune therapeutic targets. Recent technical innovations in DNA sequencing and analysis enable exome sequencing to take place at an unprecedented low cost and high accuracy and have facilitated the launch of this program, which aims to evaluate at least 50,000 exomes over the course of the next two years.

This program is part of Centers for Common Disease Genomes (CCDG) initiative at the Broad Institute, which includes parallel exome sequencing efforts in psychiatric and cardiometabolic disease. **The National Human Genome Research Institute (NHGRI)** has funded a collaborative large-scale sequencing effort to comprehensively identify rare risk and protective variants contributing to multiple common disease phenotypes. This initiative explores a range of diseases with the goal of:

- Undertaking variant discovery for enough different examples of disease architectures and study designs to better understand the general principles of genomic architecture underlying common, complex inherited diseases.
- Understand how best to design rare variant studies for common disease.
- Develop resources, informatics tools, and innovative approaches and technologies for multiple disease research communities and the wider biomedical research community.

Data sharing with these other programs will enable maximum power and efficiency by using a 2:1 (or higher) case:control ratio for some elements of this IBD program, and power will be further advanced with technical and allele frequency data from nearly 100,000 exomes already sequenced,.

The long term aim of this program is a complete assessment of the role of rare coding variation to IBD risk and protection. This will include ***articulation of the full allelic series at the nearly 200 genes identified by GWAS*** – characterizing the role of each gene in disease and in some cases allowing the immediate identification of actionable therapeutic hypotheses – as well as ***genes with high-impact variation not yet flagged by GWAS*** studies. In order to achieve these ultimate goals, we propose to set up a collaborative infrastructure for exome sequencing studies based on the successful model of the International IBD Genetics Consortium and other genetics consortia and to provide sequencing and analysis support for many such studies.

Initial Projects

As with GWAS, achieving the long term objective of the complete articulation of the contribution of rare variation will ultimately require the analysis of tens of thousands of samples and a worldwide collaborative effort. We propose to begin with a series of projects that take advantage of established strategies to increase the power of genetic studies to discover high-impact variants, or which address specific questions of clinical importance over and above IBD risk. We specifically aim to launch projects in the following areas:

- Deeper exploration of infantile and early-onset cases likely to harbor a heavier genetic load of IBD risk variants at all loci, including severe protein-coding mutational spectrum less frequently seen in adult patients
- Case-control samples from isolated/founder populations (e.g., Finland, Ashkenazi Jews, Quebec) as the allele frequency spectrum and relationship of effect size and frequency is significantly more favorable in populations that have passed through a recent bottleneck

- Exploration of non-European populations/ethnicities not as widely surveyed to date in GWAS or sequencing studies (Latino, East Asian and African American) – providing a much more complete assessment of variation in each gene
- IBD samples with compelling clinical data such as extremes of response to commonly used therapies, severe adverse response to common IBD drugs, particularly severe phenotypic manifestations (eg: presenting with pan-colitis, IBD patients with multiple extra-intestinal manifestations) – as such studies can assess important IBD-related clinical questions and simultaneously be used in the discovery of IBD risk variation. Other potential collections that might be prioritized include:
    - IBD in the elderly. Along with the infantile IBD, it could be captured under the spectrum "IBD onset in the extremes of age".
    - IBD onset associated with migration and abrupt change of environment – could be integrated with studies of non-European populations described above
    - Sample collections with detailed environmental information that may inform on underlying pathogenic mechanisms or particularly long and detailed longitudinal studies

In order to maximize the impact of the program, and considering the extensive effort behind the curation and collection of individual sample sets by collaborative partners, we would propose that individual studies in the categories above, focused on a specific population or clinical hypothesis, would work with our team to collaboratively write up and publish findings surrounding those specific samples. Simultaneously, data from all such projects will assembled for world-wide analyses of all samples later in the program.

General principles of collaboration

A “Collaborative Study Group” (CSG) is defined as an investigator or team of investigators who have collected an IBD sample that both the CSG and our group are enthusiastic to perform an exome sequencing study on. To initiate a project, a representative of the CSG will be asked to provide an example consent form to the Broad Institute compliance office (to insure we understand what can be done and what data release may be permitted) and will be asked to sign a Memorandum of Understanding regarding expectations about exome sequencing results (attached). After these preliminary administrative steps (and any additional required by the CSG), DNA transfer and exome sequencing could be initiated. The CSG will be free to use and publish on their data, with an expectation that an initial publication describing the CSG-specific experiment will be developed collaboratively by the CSG and Broad/Daly teams.

Data Storage and Access

Exome sequencing data, both raw BAMs (if desired), and processed VCF generated by the current production version the state-of the-art Picard-GATK pipeline will be provided to the CSG from the Broad Institute following the application of standard QC procedures to insure accuracy and sample fidelity according to a data freeze schedule worked out for each project.

With the exception of an extremely valuable sample with consents that forbid it, samples accepted into this study should be those for which deposit in a restricted access database (e.g., dbGAP, Wellcome Trust, EGA/EBI, etc.) is permitted. Data deposition will be performed by the Broad Institute to the appropriate repository at an agreed upon time in order to maximize the eventual use and value of the data to the broader research community.

To further insure the greatest community use of these data in functional studies and therapeutic development programs, results from each study (e.g., sites discovered, allele frequencies, and variant and gene level association statistics) will be made publicly available on a cloud-based storage system, FireCloud, upon completion of analyses at or before the time of initial manuscript submission.

Broader collaborative activities

We expect a number of collaborative studies to be launched as described above. The commitment of each CSG would then be not only to its own project, but also to participate in a collaborative mega-analysis of data starting in the second year of this program. A steering committee consisting of one member from each CSG and chaired by Dr. Daly will meet periodically by teleconference once several projects are underway in order to coordinate data sharing and global analyses, which we would aim to model after the successful existing genetics consortium activities in IBD. While the program website will keep all participants, and the world, informed about what projects are underway or planned, data production timetables and ultimately results, it would be expected that the collaborative network of CSGs would engage more freely in data sharing, would be kept up to date on progress of other studies, and work together to define the ultimate worldwide data analyses that should reach the required scale to fully articulate the role of rare variation. Thus a commitment to secure data sharing and full integration of each CSGs data in global meta/mega analyses is a requirement.

Unless explicitly precluded by consents, it is expected that control data will be able to be shared with other exome sequencing studies in the CCDG program – though this will never be done without explicit notification to the steering committee. Reciprocally, the ultimate IBD-focused mega-analyses will enlist controls from these same parallel CCDG efforts.

Frequently Asked Questions

Can one participate if one has sequencing performed elsewhere? As with prior successful consortia, we would hope that this serves as a catalyst for a data sharing, processing and analysis activity that goes far beyond that supported by this study. We will expect to invest in importing and reprocessing data from other projects external to the Broad Institute (providing those data back to contributors and keeping it for mega-analyses described above) and also for potentially developing ways of integrating site and allele count information from processed studies that cannot engage in data sharing at the same level.

Is microbiome/metagenome sequencing available? While this current funding opportunity does not support microbiome sequencing, this is available at the Broad Institute and we would enthusiastically support projects that propose both exome and microbiome studies from the same individual should support for the microbiome component be available elsewhere. We encourage contacting Ramnik Xavier for information about microbiome sequencing available at the Broad Institute. Current studies focus on microbiome analysis of diagnostic gut biopsy samples, and longitudinal fecal samples collected across a period of disease perturbation (treatment, diet, disease flares, etc). The Broad institute is committed developing and defining best practices for research and interpretation of studies involving the human microbiome.

Can additional sequencing at the same low exome price be available to partners with their own funding support? CSGs and other IBD partners with their own sources of support for expanding their studies or performing related studies not directly supported by current funding will be able to develop further projects using the same exome sequencing capabilities at the low price point currently available (the current subsidized price at Broad is at or below $200/exome available for non-NIH funding sources). These data (raw and identically processed) would be provided back directly to the group, though it would be our fervent hope that such data, if relevant or even to extend control numbers, would be available for the joint mega-analyses in IBD.

Please feel free to contact **Mark Daly** with any questions or comments.

NIH Genomic Data Sharing (GDS) Policy

To promote robust sharing of human and non-human data from a wide range of genomic research, and to provide appropriate protections for research involving human data, the National

Institutes of Health (NIH) issued the NIH Genomic Data Sharing Policy (GDS Policy). The policy became effective on January 25, 2015, and applies to NIH-funded research (e.g., grants, contracts, and extramural research) that generates large-scale human or non-human genomic data, regardless of the funding level, as well as the use of these data for subsequent research. The policy includes standards for sharing human and non-human genomic data; mechanisms for accessing large-scale genomic data; and expectations for institutional certification, IRB review, and broad, unspecified participant consent. The GDS website provides vast resources including guidance, FAQs, updates, templates, etc. for researchers, institutions, and Institutional Review Boards (IRBs). Investigators using federal funds to generate human sequence data are required to deposit that data into dbGaP. Additionally, many top tier journals are requiring deposition of data - regardless of the funding source - in order to publish. Additionally, the GDS policy requires that consents used to collect samples after 1/25/2015 reference future use, broad sharing, and that data will be deposited into a repository. Samples collected before 1/25/2015 must not be inconsistent with these ends.

Below is sample language regarding future use, broad sharing and data deposition.

> **1. Sharing and Future use -** Your samples, genomic data and health information will be stored and shared with other researchers. The samples and information will be available for any research question, such as research to understand what causes certain diseases (for example heart disease, cancer, or psychiatric disorders), development of new scientific methods, or the study of where different groups of people may have come from.
>
> **2. Repository language -** Your individual genomic data and health information will be put in a controlled-access database. This means that only researchers who apply for and get permission to use the information for a specific research project will be able to access the information. Your genomic data and health information will not be labeled with your name or other information that could be used to identify you. Researchers approved to access information in the database will agree not to attempt to identify you.