PGC Data Access – Procedure for 2015 and beyond (Posthuma, Ripke, Daly)
dd. May 28, 2015

Background: Now that PGC has grown, we are in need of transparent and a more professional structure for granting data access to PGC genotypic data. Specifically, for many studies, the most complete analyses involve not only data contributed by PGC members (that are generally available for all secondary analyses) but also involve restricted access data (that requires each institution and investigator gain explicit approval for) as well as data shared only for specific primary analyses and not available for secondary or cross-disease uses.

Proposal: In order to clarify and facilitate PGC data use, we propose a data access committee (DAC) that will create clear disease-specific guidelines for data access, a transparent structure on LISA for file access, and will track data access approvals.

Data Access Committee (DAC)
We propose to set-up a DAC responsible for granting access to PGC-genotype data, given data access restrictions. The goal is to clarify responsibilities for data access approval, remove administrative load from primary analysts and to make the process as standardized as possible.

The DAC will be made up of the following individuals:

Disease Group Representatives

1. **Dataset Coordinator (DC)** - *the primary analyst for each disease*: The Dataset Coordinator will make sure, that a standardized google sheet, the Dataset Permission Sheets (DPS) will be created and kept updated with information about the access requirements and physical location for each dataset in the collection.
2. **Disease Representative (DR)** – *the responsible investigator from each disease group*: The Disease Representative is responsible for evaluating whether all necessary approvals are in place (e.g., dbGAP permissions, scientific approval from the disease group). He/she gets this information from the DPS (see above) and keeps track with filling Researcher Permission Sheets (RPS), With this information, the DR creates a checklist that is directly accessible by other PGC members.

Central Participants

3. **Permission Granter (PG):** a representative at LISA cluster (Danielle Posthuma or her designate) with the right to change permission groups of PGC members. She gets this information from the RPS and acts on requests from the DR only (see above). Also responsible for keeping permissions from dbgap and other sources for data deposit updated on Lisa.

4. **Administrative Helper (AH):** keeping lists and files up to date, organizing links, providing information to PGC members, also adding to wiki, archiving documents, keeping minutes of any periodic DAC conference calls.
5. **Chair**: Overseeing the PGC-DAC process, making sure all documents are up-to-date, and taking final responsibility to ensure the DAC process runs smoothly.

DPS and RPS are google sheets, accessible/editable by all DAC members (and readable by all PGC members).

Workflow a PGC member seeking genotype access for a disease:

If done correctly, then the main analyst (dataset coordinator) is not needed in this process

- Apply for account on GCC/Lisa
- Submit proposal (group.analysis.proposal.doc) to disease working group and obtain proof of approval
- Ask the Disease Representative for checklist of necessary documents.
- Collect necessary approvals as listed in checklist
- Send completed checklist + all necessary documents/ (including the scientific approval and analyst memo) to the AH, who checks for completeness and archives documents. If incomplete, AH informs PGC member
- AH sends complete application to DR for checks of single datasets.
- If approved: DR adds the name to the Researcher Permission Sheet and informs the Permission Granter. If not approved: DR informs PGC member
- The Permission Granter adds this member to the corresponding unix permission group.
- The DC informs the member about physical location of the datasets in question, also adds the member to the genotype access wiki (google groups) for information about how to work with the data
- This should not take longer than 2-3 weeks.

Proposals for persons in the above mentioned groups.

Dataset Coordiator (main analyst):

Ripke: scz, bip, mdd, aut, ocd
Duncan: pts, ano

Disease Representative

mdd: Cathryn Lewis

bip: Eli Stahl
scz: Jo Knight
aut: Ric Anney
add: someone to be nominated by ben/Steve
alc: Arpana Agrawal
pts: Karestan Koenen
ano: Cynthia Bulik
ocd: Carol Matthews
alz: ??

Permission Granter:
Danielle Posthuma

Administrative Helper:
Krista Latta

Chair: Stephan Ripke & Danielle Posthuma

Permissions are provided for one year. If a projects is not finished, and initial permissions are outdated (e.g. for dbGAP), a renewal must be sent to the DAC. All finished projects need to be closed out.

Physical data repository on LISA

We will create a new account on lisa, called pgc_dac which is used solely for sharing genotype data. There will be a separate folder for each disease, and within each disease a spate folder for each data freeze. Users that are granted permission by the DAC will be added to the relevant unix group, which will be disease specific. Folders can only be accessed by users that have been added to the correct unix groups. Unix groups will be specific for a particular data-freeze (the _fxx extension).

Unix groups:

dac_mdd_fxx
dac_bip_fxx
dac_scz_fxx
dac_aut_fxx
dac_add_fxx
dac_alc_fxx
dac_pts_fxx
dac_ano_fxx
dac_ocd_fxx

Work Plan to get this group started:

- agree on this proposal [all]
- dataset coordinators to move all datasets to main place
- dataset coordinators to set-up a central google sheet (Dataset Permission Sheets - DPS)
- create central Researcher Permission Sheet (RPS)
- ask the disease groups to nominate disease representatives (approach our proposed ones directly)
- disease representatives to create checklists for new applicants
- Find a way for already granted permissions (recontact, directly adding to the RPS)
- update the Analyst memo and Gatekeeper docs

- set up a google group or email account pgc-dac, wiki page with the checklists.

Details on work Plan:

The first task will be the creation of the Dataset Permission Sheets. This involves creating a standardized Google Sheet of all currently included cohorts, the contact person for that cohort and what is needed to gain access. This list needs to be updated whenever new cohorts are added. We probably need to keep distinct freezes for each disease.
Laramie prepared a PGC-dataset_record doc that lists in detail the permissions needed, and embargo. This can be used as addendum if more detail is needed, while the excel sheet provides an overview of which datasets need action. If there are any data restrictions this should also be included. Thus, for every disease group we need to have a file describing the following columns, the DAC will provide a template on google docs or similar:

Disease	One of mdd, bip, scz, aut, alc, pts, ano, ocd
Cohort	Name of cohort
PGC name	Short name in classic PGC format
PI	PI of cohort/main contact
Genotyping platform	A6.0, I550, etc
Sample Size, post QC, pre-dedup	cases, controls
PGC_DAC_req	Permission needed to conduct secondary analyses within PGC. For example: no requirement, dbGAP permission (including number), WTCCC permission
PGCrelease_version	Freeze of the PGC datafiles in which this cohort is included
Data_use_restrictions	E.g. Can be used solely for investigating dis X, can be used for cross-disorder, none, etc

We probably want to list also the datasets, that won't be found on LISA (e.g. the pharma dataset) for completeness

We need a similar list for the Researcher Permission Sheet.