

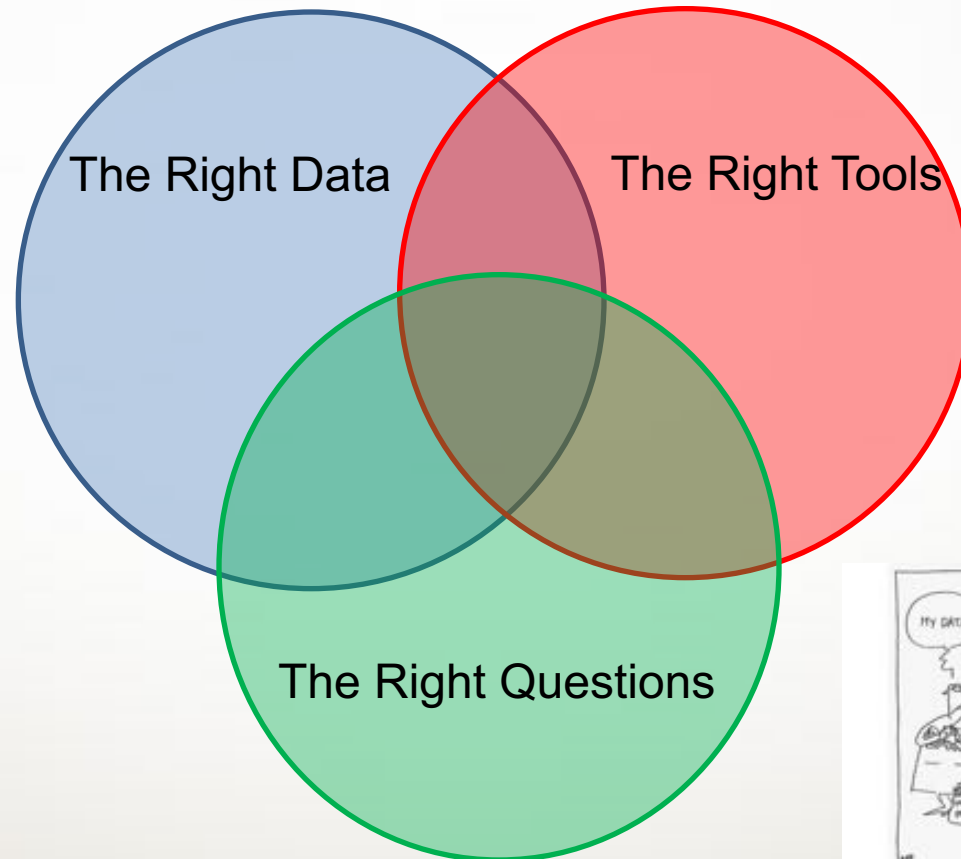
# The PGC Data Access Portal and Genomic Privacy

Lea Davis, PhD  
Assistant Professor  
Department of Medicine  
Vanderbilt University

# Outline of talk

- Context for the operations of the PGC Data Access Committee
- Describe structure of PGC DAC
- Walk through an example submission
- Goal is to help you get access to the data you need asap!

# Data sharing is a key element to scientific discovery in perpetuity



# A tension exists

Maintaining genomic  
privacy

Sharing data and  
democratizing science



# What are the concerns of research participants?

- The Center for Genetic Privacy and Identity in Community Settings
- Literature suggests participants fear that information will not be kept private
- Fears of genetic discrimination and putting family members at risk
- Greater concern among non-white minorities, patients, caregivers
- Potential for benefit outweighed concerns about privacy

Socio-Genomic Contract: We agree to maximize our research efforts as a community while at the same time protecting the data from misuse.

# Elements of the socio-genomic contract

- Maximize benefit by agreement to share data
  - Psychiatric Genomics Consortium
- Safeguards to protect data from misuse
  - Workgroup evaluation process
- Infrastructure to enable both
  - Data Access Committee

# PGC Data Access Committee

- Goal is to develop and maintain simple, efficient, and secure procedures for PGC investigators to access PGC data.
  - 1) liaise with the contributing party and ensure we are in compliance with international, federal, and local data protections
  - 2) provide a transparent infrastructure for requesting data access
  - 3) communicate with LISA administrators and ensure access is granted to the correct data sets



# PGC Data Access Committee

- Not an approval or permission granting body.
- Approvals remain with workgroup
- Permissions remain with data contributors

# Structure of PGC DAC

- DAC reps are your point of contact for questions or problems
- Ex-officio member, Dr. Patrick Sullivan, and administrative assistant, Ms. Tamara Biondi.
- Contact info on website
  - About the PGC -> People



**Lea Davis**  
Vanderbilt University Medical Center



**Danielle Posthuma**  
Vrije University



**Stephan Ripke**  
Harvard University



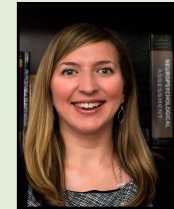
**Jo Knight**  
(SCZ)



**Jeremiah Scharf**  
(TSOCD)



**Laramie Duncan**  
(PTSD)



**Karen Mitchell**  
(AN/ED)



**Eli Stahl**  
(BIP)



**Mark Adams**  
(MDD)



**Raymond Walters**  
(SUD)

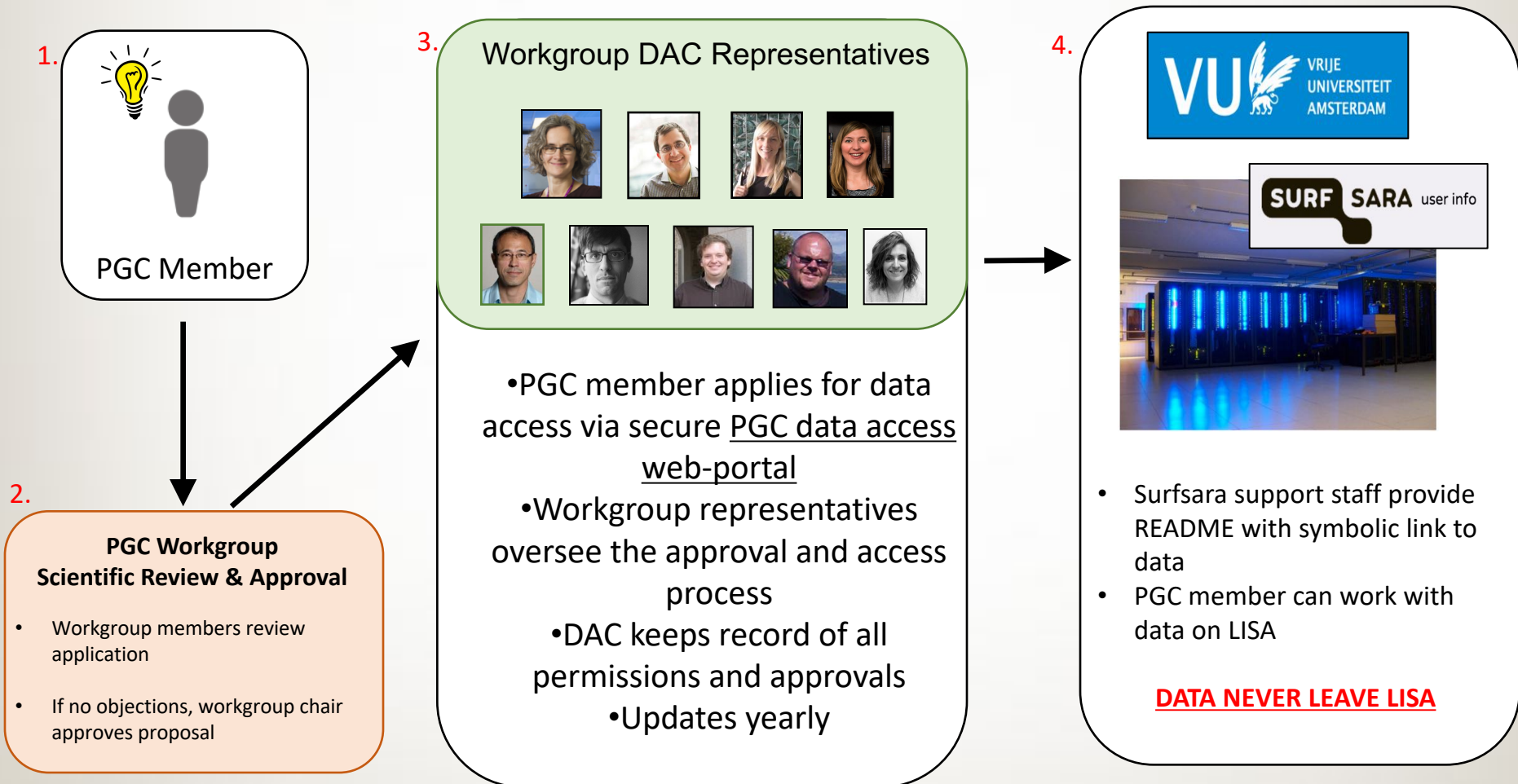


**Richard Anney**  
(ASD)



**Marta Ribases**  
(ADHD)

# Structure of Data Access Request



# Step 1: Writing and Submitting a Proposal

- Review the “Data Access -> How To” section of the PGC website
- Contact the workgroup chair and DAC rep
  - Is there already a proposal?
  - Is the data you need available?
- Review the publication policies found under “Workgroups”
- Use the current template found under “Data Access -> Documents for Data Access”
- Sign and submit the “PGC Analyst Memo”
  - “Data Access -> Documents for Data Access”

# Step 2: Obtain an account on LISA

1. Instructions on website “Data access -> How To”
2. Annual renewal of account: send Danielle Posthuma ([d.posthuma@vu.nl](mailto:d.posthuma@vu.nl)) an e-mail, with [hic@surfsara.nl](mailto:hic@surfsara.nl) in cc, stating you are still working on a PGC-approved project
3. Acknowledge GCC in all publications and presentations

**Genetic Cluster Computer**

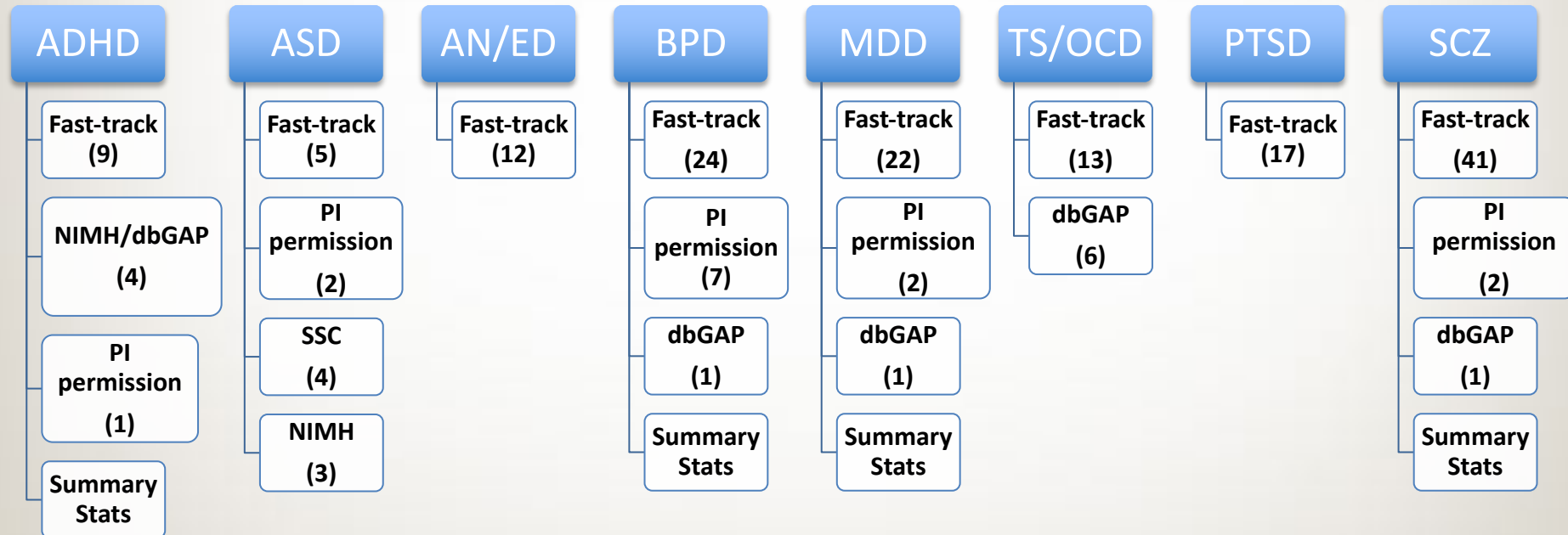
Welcome Technical Details **Obtaining Access** Tutorial Software GCC citations

Researchers seeking access to the genetics cluster can apply to the principal investigator (Posthuma), with a brief description of their intended analyses and research field. They will receive a user name from the administrators at SURFsara. There is no restriction on processor time usage, although some restrictions apply to the number of nodes that can be occupied per user and the number of jobs that can be submitted simultaneously. A batch queuing system will coordinate node usage, analogous to the current batch queuing systems used at the Lisa cluster at SURFsara. Please fill out the form below to obtain access:

Name\*  
E-mail address\*  
Affiliation\*  
Nationality\*  
Address\*  
Phone number\*  
Project Leader  
Estimated cpu-hours needed  
Brief description of your project (no more than 250 words)\*  
Send

# Step 3: Decide what you want to request

Individual data sets are organized into **data packages** based on the required permissions. All workgroups have a **fast-track** data package that can be accessed immediately with an approved proposal.



## Step 4: Acquire any additional permissions

1. Fast-track data package requires no additional permissions (~85% of all available PGC data!)
2. Most workgroups have at least one repository held data package (e.g., dbGAP or NIMH)
  1. **\*\*dbGAP “PGC bundle” is available \*\***
3. Some workgroups have PI held data sets that require explicit written permission from the PI
  1. **The DAC rep will point you to the right people**



# PGC dbGAP Collection



## dbGaP Collection: Psychiatric Genomics Consortium (PGC) dbGaP Datasets

### Collection Description

The collection includes cases and controls used in the GWAS meta-analysis of neuropsychiatric phenotypes studied by investigators in the Psychiatric Genomics Consortium including major depression, bipolar disorder, attention deficit and hyperactivity disorder, schizophrenia, Tourette Syndrome, and obsessive-compulsive disorder. The consent type for data sets includes general research use as well as phenotype-specific restricted usage. This dbGaP collection was created to facilitate identification of datasets used in PGC analyses in order to expedite the application and ascertainment process for PGC members wishing to access data within the consortium. Non-PGC members may also apply for this dbGaP collection.

- Thank you to Anjie Addington, Nicki North, and Doug Levinson!
- 10 dbGAP data sets
- 1 application, 1 progress report, 1 renewal
- Reviewed by assigned DACs
- **Must be consistent with the data use limitations**
  - **Restricted use = can not combine with data on other phenotypes**




# Collect your documents

- LISA username
- Signed analyst memo
- Signed WTCCC memo
- Proposal
- Documentation of proposal approval
- Documentation of  
dbGAP/NIMH/SFARI/Investigator approval

# Documentation of proposal approval

9/20/2017      Gmail - Proposal for Review (Davis\_Petryshen) - 071817\_pgc.secondary.analysis.template.v2\_LKD\_SSPGC.pdf

      Lea Davis <lea.k.davis@gmail.com>

---

**Proposal for Review (Davis\_Petryshen) - 071817\_pgc.secondary.analysis.template.v2\_LKD\_SSPGC.pdf**

---

**Knight, Jo** <jo.knight@lancaster.ac.uk>      Mon, Aug 14, 2017 at 4:12 AM  
Cc: "TPETRYSHEN@mgh.harvard.edu" <TPETRYSHEN@mgh.harvard.edu>, "lea.k.davis@gmail.com" <lea.k.davis@gmail.com>

Hi,  
There have been no comments on this project so you have PI approval to go ahead.  
Cheers  
Jo

---

**From:** Knight, Jo  
**Sent:** 31 July 2017 08:02  
**To:** PGC3-SCZ  
**Cc:** TPETRYSHEN@mgh.harvard.edu; lea.k.davis@gmail.com  
**Subject:** Proposal for Review (Davis\_Petryshen) - 071817\_pgc.secondary.analysis.template.v2\_LKD\_SSPGC.pdf

Dear All,  
Please find a proposal attached for renewal.  
Please send any comments within two weeks.  
Cheers,  
Jo

**Date**

**Sent to PIs listed on proposal**

**Approval Status**

**Workgroup chair or representative**


# Documentation of dbGAP approval

dbGAP data set

Sent to PIs  
listed on  
proposal

Approval Status

9/20/2017 Gmail - APPROVAL of your request [#58319-2] for access phs000092/HR

 Gmail Lea Davis <lea.k.davis@gmail.com>

---

**APPROVAL of your request [#58319-2] for access phs000092/HR**

dbgap-reply@ncbi.nlm.nih.gov <dbgap-reply@ncbi.nlm.nih.gov> Wed, Aug 9, 2017 at 2:34 PM  
Reply-To: dbgap-help@ncbi.nlm.nih.gov  
To: lea.k.davis@gmail.com  
Cc: steve.munoz@vanderbilt.edu

This is an automated message from NCBI dbGaP (the Database of Genotypes and Phenotypes) Authorized Access system. Do not reply back to this message or send email to [dbgap-reply@ncbi.nlm.nih.gov](mailto:dbgap-reply@ncbi.nlm.nih.gov)

Dear Lea Davis,

NIH has **APPROVED** your request [#58319-2] for the dataset **Health Research in GENEVA Study of Addiction: Genetics and Environment (SAGE)** access as part of your project titled #15355: "The Genetic Basis for Sexual Dimorphism in Neuropsychiatric Disease".

The following comments were provided:  
*Thank you for applying to dbGaP. Please note all external collaborators must independently apply for dbGaP approved access. This dataset can only be used in research consistent with this data use limitation and cannot be combined with other datasets of other phenotypes. Uses inconsistent with the data use limitation are considered a violation of the Data Use Certification.*

- Before accessing the data, please **REVIEW** the terms of access of the [Data Use Agreement](#) that you and have signed.
- All external collaborators must submit an independent Project Request from their institution and be approved to access the datasets before any data may be shared or exchanged

Data Access Link(s):

Data Portal	Access Link(s)	Portal Technical Help Desk
dbGaP	<a href="#">Data Access Request</a>	<a href="#">dbGaP Help</a>

- If you have questions related to your access request for the 'GENEVA Study of Addiction: Genetics and Environment (SAGE)', please contact the 'National Human Genome Research Institute' Data Access Committee at [nhgridac@mail.nih.gov](mailto:nhgridac@mail.nih.gov).
- If you have any questions regarding Authorized Access Portal please contact the NCBI dbGaP [Help Desk](#).
- dbGaP FAQ: <https://www.ncbi.nlm.nih.gov/books/NBK5295/>

Please do not reply to this message.

Consent Group

# Step 5: Submit Request

The screenshot shows the 'Data Access Portal' page of the Psychiatric Genomics Consortium. The header includes the consortium's name and a navigation menu with links like Home, About the PGC, PGC Workgroups, Data Access, Downloads, Tools, Training & Jobs, FAQ, Stat Gen, and Worldwide. A sidebar on the left lists 'DATA ACCESS' links: Open Source Philosophy, How To, Documents for Data Access, and Data Access Portal (highlighted). The main content area is titled 'Data Access Portal' and contains a welcome message: 'Welcome to the PGC Data Access Portals!'. Below this, it states: 'You will need the following information on hand to submit your request:'. A list of requirements follows: 1) your USA username, 2) your signed analyst memo, 3) your signed WTCCC memo, 4) your approved proposal, 5) a copy of the email from the workgroup chair stating your proposal has been approved, and 6) copies of any additional permissions required (consult your DAC representative). A note specifies: 'All documentation must be in PDF or Word format to upload on the Data Access Portal.' Below this, a prompt says: 'Please click on the button that corresponds to your data request.' There are six blue buttons arranged in two rows: 'ADHD Data Access Portal', 'BIP Data Access Portal', 'ED Data Access Portal' in the top row, and 'MDD Data Access Portal', 'PTSD Data Access Portal', 'SCZ Data Access Portal' in the bottom row. At the bottom, a green box contains the text: 'Portals for all workgroups will be brought online as soon as possible. The DAC will make announcements as soon as each new portal is ready!'.

Psychiatric Genomics Consortium

Home About the PGC PGC Workgroups Data Access Downloads Tools Training & Jobs FAQ Stat Gen Worldwide

» - Data Access - Data Access Portal

**DATA ACCESS**

- Open Source Philosophy
- How To
- Documents for Data Access
- Data Access Portal**

## Data Access Portal

**Welcome to the PGC Data Access Portals!**

You will need the following information on hand to submit your request:

- 1) your USA username
- 2) your signed analyst memo
- 3) your signed WTCCC memo
- 4) your approved proposal
- 5) a copy of the email from the workgroup chair stating your proposal has been approved
- 6) copies of any additional permissions required (consult your DAC representative)

All documentation must be in PDF or Word format to upload on the Data Access Portal.

Please click on the button that corresponds to your data request.

**ADHD Data Access Portal** **BIP Data Access Portal** **ED Data Access Portal**

**MDD Data Access Portal** **PTSD Data Access Portal** **SCZ Data Access Portal**

Portals for all workgroups will be brought online as soon as possible. The DAC will make announcements as soon as each new portal is ready!

# Step 5: Submit Request

## MDD PGC Data Access Portal

This form will walk you through the documentation you need to provide in order to gain access to MDD PGC data.

**Name\***

First Name

Last Name

**Institution\***

**Phone\***

**Institutional Email\***

Upon selection of a phenotype, the phenotype-specific section of the form appears.

Users upload required documentation.

Significantly reduces reliance on email attachments!

## Major Depressive Disorder

Here you will be able to specify your requested data freeze and upload required documentations.

Your request requires an analysis proposal that has been approved by the Major Depression PGC Workgroup.\*

- ☒ I am the PI of an approved proposal.
- ☐ I am a named analyst on an approved proposal.
- ☐ I do not have an approved proposal. How do I submit a proposal?

Please upload your approved proposal. Accepted formats include pdf and doc.\*

No file chosen

Please upload a copy of your MDD workgroup approval email. Accepted formats include pdf and doc.\*

No file chosen

Select data package\*

- ☒ Fast track (no permissions needed beyond PGC MDD group approval), 22 datasets
- ☒ BOMA, 1 dataset
- ☒ SHIP, 2 datasets
- ☒ GenRED (MGS GAIN dbGaP controls), 1 dataset
- ☒ Check All

Data generously contributed by the BoMa consortium requires consortium approval. Please upload your approval here (PDF or doc copy of approval email is sufficient).\*

No file chosen

Data generously contributed by the SHIP consortium requires consortium approval. Please upload your approval here (PDF or doc copy of approval email is sufficient).\*

No file chosen

MGS GAIN samples used in MDD analysis were obtained from dbGAP and therefore require dbGAP approvals to access on LISA. If you are an independent investigator, you must (a) submit your own dbGAP approval or (b) submit dbGAP approval in which you are named as a collaborator at your institution for data set phs000021.v3.p2. If you are a named analyst (and not an independent investigator) you may submit the dbGAP approval obtained by your supervisor at your institution for dbGAP data set phs000021.v3.p2.\*

- ☐ I can provide my own dbGAP approval.
- ☐ I am a named collaborator on the requisite dbGAP application.
- ☐ I can provide my supervisor's dbGAP approval for the requisite data sets.
- ☐ How do I gain dbGAP access?

# Getting help through the form

In order to access data on the LISA server, you must have a valid account.\*

- ☐ Yes, I have a LISA account.
- ☒ No, I don't have a LISA account. How do I get one?

To request an account on LISA please follow the instructions on the [Surfsara Help Page](#). **PLEASE DO NOT SUBMIT YOUR REQUEST UNTIL YOU HAVE OBTAINED AN ACCOUNT ON LISA!**

Your request requires an analysis proposal that has been approved by the Major Depression PGC Workgroup.\*

- ☐ I am the PI of an approved proposal.
- ☐ I am a named analyst on an approved proposal.
- ☒ I do not have an approved proposal. How do I submit a proposal?

Please refer to the [PGC Website](#) or contact the Major Depression data set representative, [Dr. Cathryn Lewis](#), for information on how to submit an analysis proposal.

I understand how to run jobs efficiently on the LISA cluster.\*

- ☐ Yes.
- ☒ No. How can I learn about submitting jobs efficiently on LISA?

[SurfSara](#) offers users substantial support to [LISA](#) users ranging from [submitting jobs efficiently](#) to available [software and libraries](#). Please review materials on their website and [contact](#) SurfSara help desk for further assistance. **PLEASE DO NOT SUBMIT YOUR REQUEST UNTIL YOU HAVE COMPLETED THE LISA USER TUTORIALS!**

Wherever possible there are places to get help if you get stuck.

# Authorize request with signature

Consistent with regulatory requirements

I understand that I am requesting access to de-identified genetic data. I promise not to use this data to attempt to re-identify research participants. I further promise to use data requested only for the purposes described in the attached proposal. I understand that no individual genotype data is to leave the LISA server. All supporting documentation is, to the best of my knowledge, accurate and current.\*

A handwritten signature in black ink, appearing to read 'T. H. [unclear]', is written over a horizontal line within a light gray rectangular box.

[\[clear\]](#)

Use your mouse or finger to draw your signature above





## **Congratulations!**

**Your data access request has been successfully submitted to the Data Access Committee and LISA administrators. Your request will be reviewed by the appropriate disease representatives and LISA administrators.**

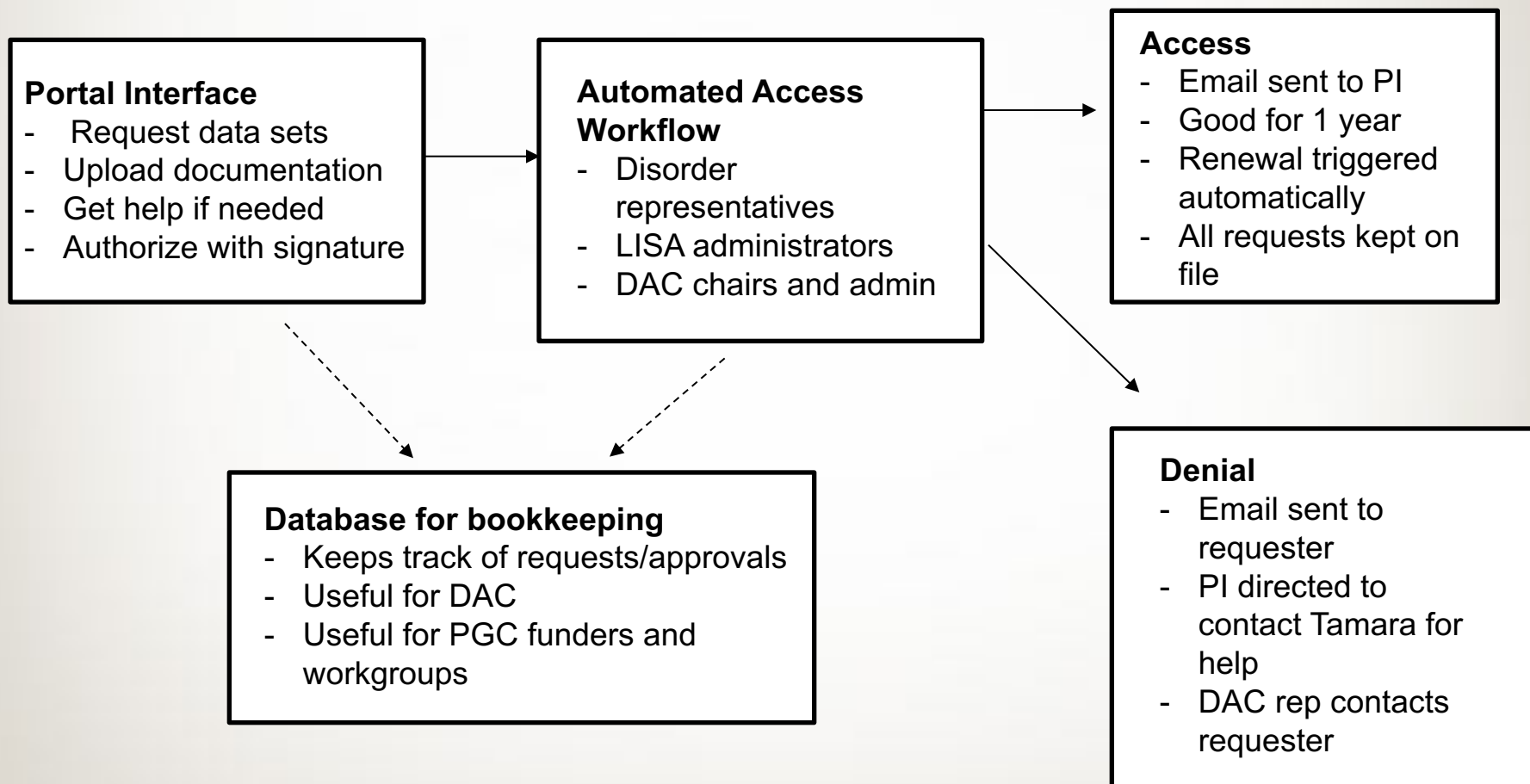
**You will be notified of the outcome within 1-2 weeks. If you experience delays or have questions or concerns, please contact Krista Latta ([krilatta@email.unc.edu](mailto:krilatta@email.unc.edu)).**

**~ the PGC Data Access team**

Encouragement

Contact info

# Structure of your request

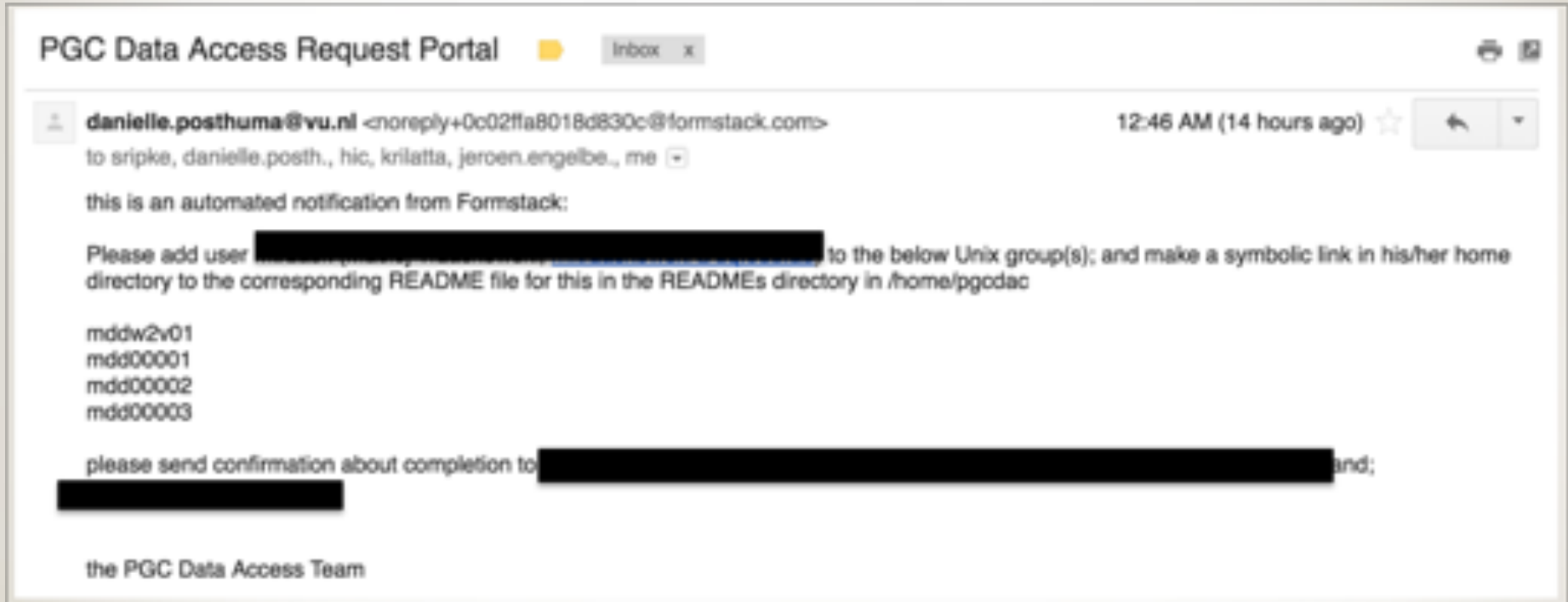


# Behind the scenes of your access request



1. You receive a confirmation email
2. DAC rep reviews and approves
3. You receive an approval email

# Automated request sent to LISA helpdesk



1. LISA help desk is notified
2. You will receive an email when you have been added to the appropriate user group

# Access!

```
#####  
## README for PGC SCZ wave2 data access, accessible with permission group scz2v01  
#####  
  
## first of all, check if you are member of permission group scz2v01 (typing "groups" at the command line)  
## in fact otherwise you should not find this README in your homedirectory  
  
## link to dataset collection  
[REDACTED]  
  
## for datastructure consult this wiki page:  
https://sites.google.com/a/broadinstitute.org/ricopili/  
  
## please find these slides there (starting with slide 28):  
ricopili_imputation_wcpq_oct15_website.pptx  
  
## to use data with ricopili pipeline:  
  
## create a symbolic link to your home directory  
## ln -s / [REDACTED]  
  
## basic association test:  
postimp_navi_[version] --mds prune.bfile.cobg_PGC_SC249.sh2.menv.mds_cov --coco 1,2,3,4,5,6,7,9,15,18 --out OUTNAME --addout OUTADDITION --triset triset_loc_scz49  
  
## --out and --addout specify naming of output-files (will be combined)  
## use --reflex reflex_scz2v01 for a quick test (1 genomic chunk) to compare with primary outcome  
  
## it is recommend to test with the above mentioned --reflex command to check if you get the same results as the original freeze.  
  
## please become a member of the google groups and ask your questions there:  
https://groups.google.com/a/broadinstitute.org/forum/#!forum/rp-users
```

# Frequently Asked Questions

## Requirements for principal investigators

- As the PI, you must assume responsibility for proper use of data in your lab.
  - You must be listed on the proposal
  - You must obtain permission from repositories (i.e., dbGAP, NIMH, SSC, etc.) and any individual data owners
  - ANY member of your lab who wishes to access data must get their own LISA account (including you!)
  - NEVER SHARE LOGIN CREDENTIALS
  - INDIVIDUAL GENOTYPES NEVER LEAVE LISA

# Requirements for trainees and staff

- If you are not the PI of your lab, you must
  - Sign and submit the PGC analyst memo
  - Sign and submit the WTCCC analyst memo
  - Obtain your own LISA username
  - Submit documentation with your PI and institution listed
  - NEVER SHARE LOGIN CREDENTIALS
  - INDIVIDUAL GENOTYPES NEVER LEAVE LISA



# Working with repositories

- WTCCC – requires only signed analyst memo
  - Thanks to Cathryn Lewis
- dbGAP
  - PGC dbGAP collection so all 10 sets can be requested at 1 time, in 1 application, with 1 progress report, and 1 renewal
  - dbGAP DACs
  - Estimate ~4-6 weeks (usually shorter, sometimes longer if the government shuts down)

# Need help?

- The PGC website!
  - <https://www.med.unc.edu/pgc>
  - Data Access
  - Workgroup descriptions
  - Meet the workgroup chairs and DAC contacts
  - FAQ
  - Join the stat gen calls
- Submitting & running jobs, queueing  
LISA website or helpdesk: [hic@surfsara.nl](mailto:hic@surfsara.nl)



Huge thank you to LISA  
helpdesk!  
Jeroen Engleberts  
Zheng Meyer-Zhao

We wish to thank all data owners who  
have graciously shared their data with  
the PGC and worked with the DAC to  
make data available.

Heartfelt thanks to all 900,000+ people  
who have entrusted us with their  
genomes and their greatest hopes for  
recovery.



Tamara  
Biondi  
(Admin  
Support)



Cathryn Lewis

## Data Access Committee



Lea Davis  
Vanderbilt University



Danielle Posthuma  
Vrije University



Stephan Ripke  
Harvard University



Jo  
Knight  
(SCZ)



Jeremiah  
Scarf  
(TSOCD)



Laramie  
Duncan  
(PTSD)



Karen  
Mitchell  
(AN/ED)



Eli Stahl  
(BIP)



Mark  
Adams  
(MDD)



Raymond  
Walters  
(SUD)



Richard  
Anney  
(ASD)



Marta  
Ribases  
(ADHD)



Patrick  
Sullivan  
(Ex-officio  
Member)