

# Data Access Demystified: A complete guide to obtaining genotype data for PGC members

Lea K. Davis, PhD – On behalf of the PGC DAC  
Assistant Professor  
Division of Genetic Medicine  
Department of Medicine  
Vanderbilt University Medical Center

# Presentation Outline

- Glossary of terms
  - the difference between approval, permission, and access!
- Know your representatives
  - Introduction to PGC Data Access Committee
- Introduction to LISA file structure and the PGC “data packages”
- Writing secondary analysis proposals
  - Tips for writing your proposal
- How to acquire data access once your proposal is approved
- The data access request portal
- Behind the scenes of your data access request
- "I'm a PI but my analyst needs access"... and other FAQ
- Working with repositories (NIMH; PGC dbGAP bundle)
- Great, but does it work?

# Glossary of frequently used terms

- Approval = A workgroup approves of an analysis idea
- Permission = The data owner (or provider) allows data to be shared
- Access = Analysts get “hands on” the data
- Data set = An original collection of genotypes (i.e., case/control, family, case/pseudo-control)
- Data package = collection of data sets that can be accessed together
- DPS = Data Permission Sheet contains list of all data sets, summary information, and type of permission required
- SURFsara = Organization that develops and offers advanced computational infrastructure, services and expertise.
- LISA = Lisa system is a cluster supported by University of Amsterdam (UvA), The VU University Amsterdam (VU), and The SURF organization
- GCC = Genetic Cluster Computer (part of the Dutch LISA cluster) and home to PGC data

# PGC Data Access Committee

- Formed May of 2015
  - Thirteen members
  - Bi-weekly teleconference meetings
- The DAC is not an approval or permission granting body.
- Approvals remain with workgroup
- Permissions remain with data owners
- DAC goal is to develop and maintain simple and efficient procedures for PGC investigators to access PGC data.
- (Be kind to your DAC rep, this is a labor of love 😊)



# PGC Data Access Committee

- DAC serves as a clearinghouse to
  - 1) liaise with the workgroups and ensure PGC investigators are in compliance with international, federal, and local data access requirements
  - 2) provide a transparent infrastructure for requesting data access
  - 3) communicate with LISA administrators and ensure access is granted to the correct data sets

# Meet your DAC reps!

- DAC reps are your point of contact for questions or problems
- In addition to the central committee members and the disorder representative the group includes ex-officio member, Dr. Patrick Sullivan, and administrative assistant, Ms. Krista Latta.



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)



Cathryn Lewis  
(MDD)



Raymond Walters  
(SUD)



Richard Anney  
(ASD)

# How to contact the DAC Reps

- Info on PGC Website (“About the PGC -> People -> Data Access Committee”)

## Disorder Representatives

*Please note that all email addresses end with @gmail.com*

Phenotype Group	Disorder Representative	Email Address
ADHD	Marta Ribasés	<a href="mailto:pgc.dac.add@gmail.com">pgc.dac.add</a>
ANO	Karen Mitchell	<a href="mailto:pgc.dac.ano@gmail.com">pgc.dac.ano</a>
AUT	Ric Anney	<a href="mailto:pgc.dac.aut@gmail.com">pgc.dac.aut</a>
BIP	Eli Stahl	<a href="mailto:pgc.dac.bip@gmail.com">pgc.dac.bip</a>
MDD	Cathryn Lewis	<a href="mailto:pgc.dac.mdd@gmail.com">pgc.dac.mdd</a>
OCD/TS	Jeremiah Scharf	<a href="mailto:pgc.dac.toc@gmail.com">pgc.dac.toc</a>
PTSD	Laramie Duncan	<a href="mailto:pgc.dac.pts@gmail.com">pgc.dac.pts</a>
SCZ	Jo Knight	<a href="mailto:pgc.dac.scz@gmail.com">pgc.dac.scz</a>
SUD	Raymond Walters	<a href="mailto:pgc.dac.sud@gmail.com">pgc.dac.sud</a>

# Where is the PGC data?

## The Genetic Cluster Computer



Unique to the PGC is the centralization of data on a protected server at the Vrije University (Amsterdam, the Netherlands). Investigators wishing to access data must submit a proposal for approval and complete the PGC data access procedures.

**GENOTYPE DATA NEVER LEAVE THE LISA SERVER.**

# LISA Specs

- very active helpdesk with people highly knowledgeable in statistical genetics software
- very good relationship with several LISA directors who often provide much more than we ask or pay for. PGC has used LISA since 2007, with hardly any costs involved for the users.
- DP has contributed 300k€, Dutch Brain foundation 70k€
- LISA primarily funded by VU

PGC has contributed ~170k over the past 7 years to buy a TB file server, extra computing power and a high mem node. Donations from: [Pat Sullivan](#), [Matt Keller](#), [Danielle Posthuma](#), [Lea Davis](#), [Ole Andreasson](#), [Pamela Sklar](#), [Jonathan Sebat](#), [Mark Daly](#), [Shaun Purcell](#), [Steve Faraone](#), [Doug Levinson](#), [Ken Kendler](#)

Total number of cores	7856
Total amount of memory	26 TB
Total peak performance	149 TFlop/sec
Disk space	100 TB for the home file systems

Additional for PGC:

66TB for home dir

1 HighMem node (1TB internal memory)



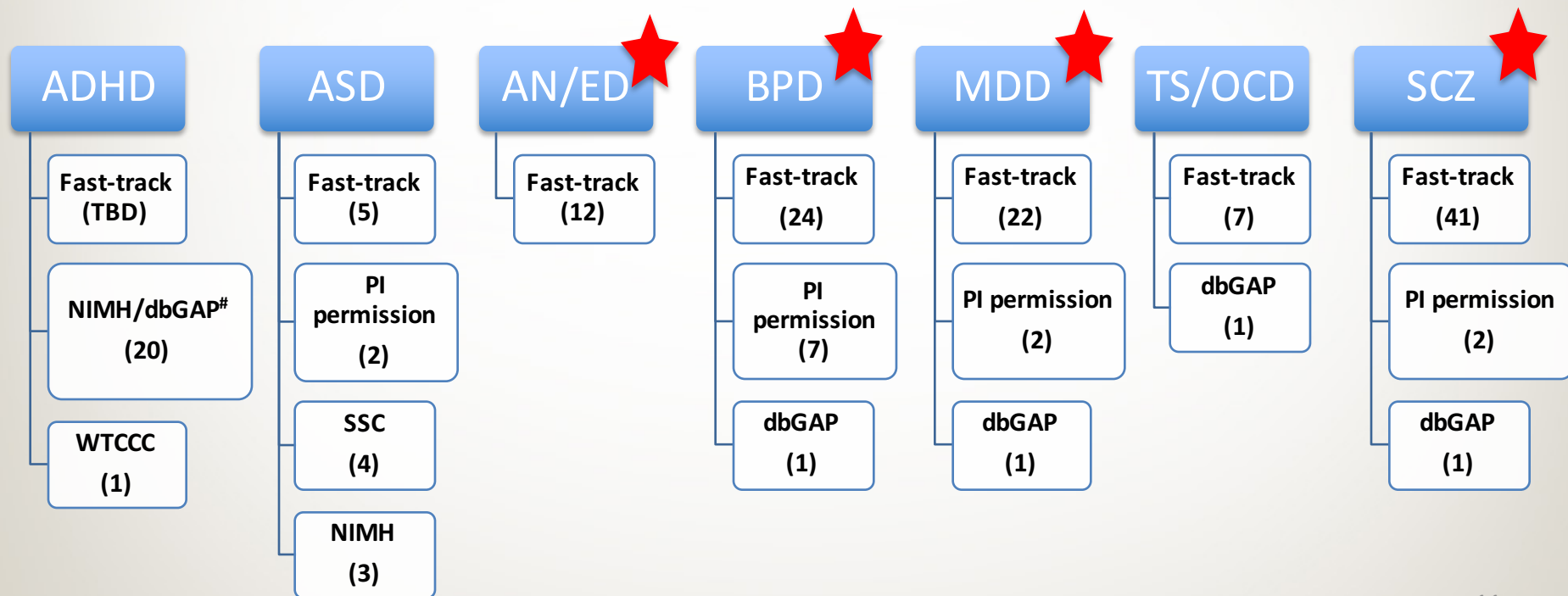
# PGC use of LISA

- 448 accounts
- 243 currently not active
- 205 currently active (including course accounts & accounts only used for data-upload), around 100-150 are running analyses occasionally or frequently
- Total home directory usage varies between 40-70 TB
- *main purpose*: primary and secondary analyses and DAC repository

# Organization of PGC Data

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.

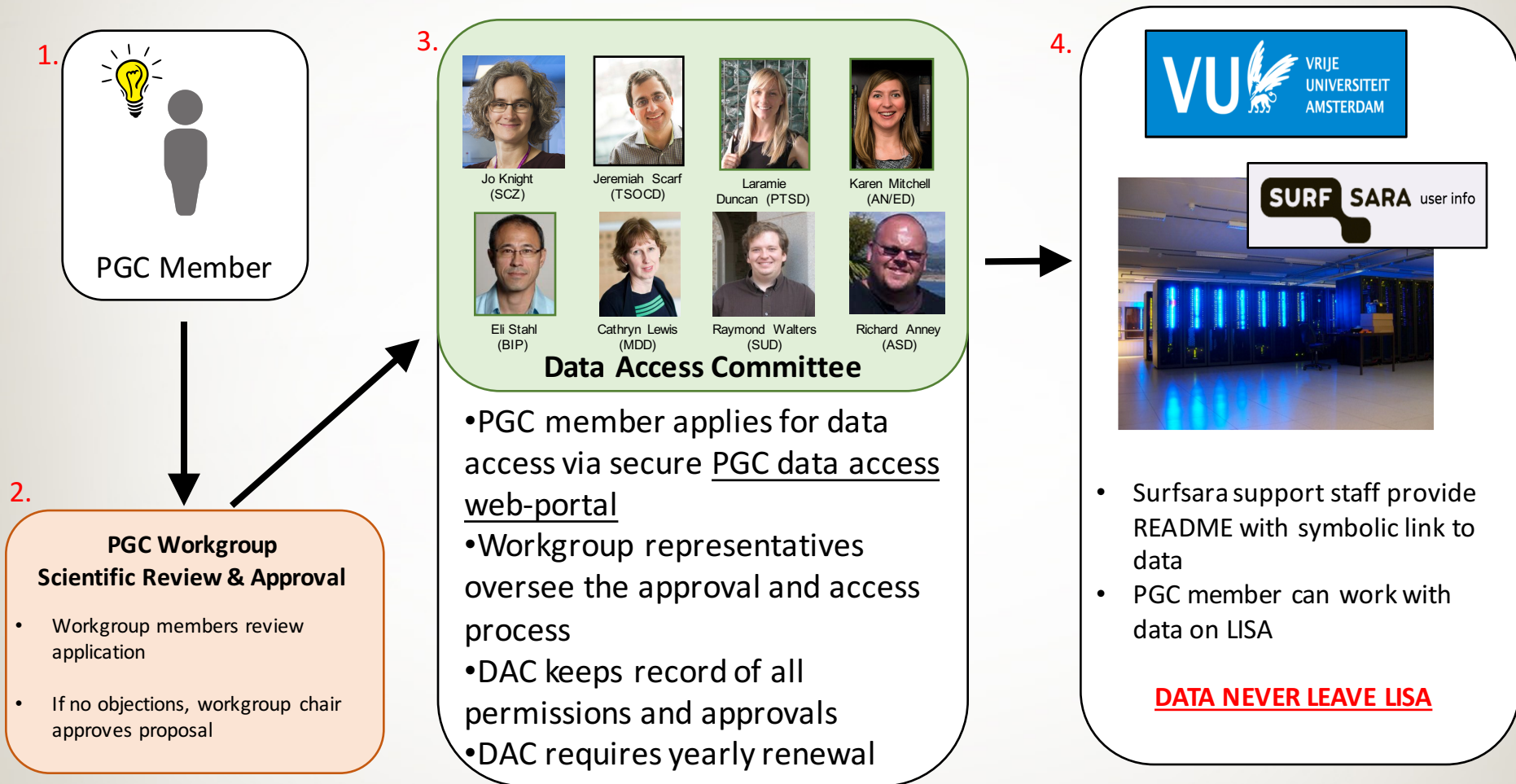
(PTSD and SUD will be added after the publication of primary workgroup manuscripts.)



# reduces to 3 dbGAP data sets

\*WTCCC data set requires signed analyst assurance

# Process for PGC investigators to obtain individual level genotype data





# Tips for a writing a secondary analysis proposal

Psychiatric Genomics Consortium

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

» Data Access » How To

DATA ACCESS

Open Source Philosophy

How To

Documents for Data Access

Data Access Portal

## How To

### Find Available Datasets

1. GWAS summary statistics are available without restriction on the [Downloads](#) page.
2. Access to individual level data from each phenotype may be requested through the Data Access Portal. Consult the [PGC Workgroup DAC representative](#) for a list of available datasets.

### Request Individual Level Data

1. The first step is to consult your [workgroup DAC representative](#) to make sure that the data you are requesting is appropriate for your analysis plan. In addition, your DAC representative can let you know if there are any special permission requirements needed. You should allow at least a week for your DAC rep to respond.
2. You will need to agree to the terms of the [analyst memo](#) and send a signed copy to your workgroup chair.
3. You will need to obtain an account on the [LISA \(GCC\) cluster](#). This usually takes only a couple of days.
4. You will write a [secondary analysis proposal](#) and submit it to your [workgroup chair](#) (and cross-disorder chair, if necessary). Typically secondary analysis proposals are reviewed within three weeks.
5. All workgroups have a "fast-track" data package that is available once the workgroup approves the proposal. However, some additional data sets may require explicit permission from individual investigators, dbGAP, NIMH, or a sponsoring foundation. You will need to secure any special permissions for data sets not included in the "fast-track" data package. This is the most time consuming step and can take anywhere from 3-6 weeks. For most phenotypes, the vast majority of the data is available through the "fast-track" data package.
6. Once you have an approved proposal and documentation of any "special permissions" you may request data through the [Data Access Portal](#).

### Write a Secondary Analysis Proposal

The [secondary analysis proposal](#) includes a brief description of the rationale for the proposal, the analytic plans, the data being requested, the individuals who will be involved in the work, the timeline, and the plans for publication. When trying to decide how much detail to include, it is better to err on the side of too much rather than too little so that the workgroup has enough information on which to make a recommendation. The approval is awarded to the project, meaning that approval is given for the project that you describe. We realize that analytic strategies can sometimes change through the course of analysis. If you find that your plan has significantly shifted course, you should speak with your DAC representative to determine if the project has changed enough to be considered a new project.

### Gain Access to LISA (Genetic Cluster Computer (GCC))

Most PGC analyses are done on LISA, also known as the Genetic Cluster Computer (GCC), in the Netherlands. You or someone in your group will need to have an account on this cluster in order to access individual level genotype data.

You can apply for a LISA/GCC account [here](#).

The PGC is deeply grateful to our colleagues in the Netherlands. The [GCC](#) is supported by an [Netherlands Organisation for Scientific Research](#) Medium Investment grant (480-05-003) to Prof Danielle Posthuma, by the [VU University Amsterdam](#), and by the [Dutch Brain Foundation](#) to Prof Roel Ophoff. The GCC is hosted by the [Dutch National Computing and Networking Services](#).

### Contents

- [Find Available Datasets](#)
- [Request Individual Level Data](#)
- [Write a Secondary Analysis Proposal](#)
- [Gain Access to LISA \(Genetic Cluster Computer \(GCC\)\)](#)

## Tips for submitting a secondary analysis proposal

- Review the “Data Access -> How To” section of the new PGC website
- Contact the workgroup chair and DAC rep
  - Is there already a proposal?
  - Is the data you need available?
- Review the publication policies (“workgroups”)
- Be sure to use the template found under “Data Access -> Documents for Data Access”
- Sign and submit the “PGC Analyst Memo”
  - “Data Access -> Documents for Data Access”

# Proposal approved! Now what?

- Obtain an account on LISA
  - Instructions on website “Data access -> How To”
  - <http://geneticcluster.org>
  - 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
  - Please acknowledge GCC in all publications and presentations (instructions on website)

# Proposal approved! Now what?

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

# Proposal approved! Permission obtained! (Now what?)

- Have the following documents in hand
  - LISA username
  - Signed analyst memo
  - Signed WTCCC memo
  - Proposal
  - A copy of the email from the workgroup chair stating your proposal has been approved
  - Copies of any additional permissions required

# Using the PGC Data Access Portal

## Welcome to the PGC Data Access Portals!

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

- 1) your LISA 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)

go to [pgc.unc.edu](http://pgc.unc.edu)

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.

**BIP Data Access  
Portal**

**MDD Data Access  
Portal**

**SCZ Data Access  
Portal**

# User friendly portal interface

## PGC Data Access Portal

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

**Name\***

First Name

Last Name

**Institution\***

**Phone\***

**Institutional Email\***

**Address\***

City

State

Zip Code

Country

**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?

# Request data sets

Select a phenotype.  
Coming soon: For  
cross disorder  
proposals, select  
multiple  
phenotypes

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?

LISA username

ldavis

Providing your username here will speed up the process of providing access to data you are requesting.

Name of your proposal

Genetic analysis of MDD

PGC Phenotype\*

- ☐ Schizophrenia  
☒ Major Depressive Disorder  
☐ Obsessive Compulsive Disorder and Tourette Syndrome  
☐ Anorexia  
☐ Post Traumatic Stress Disorder  
☐ Bipolar Disorder  
☐ Autism Spectrum Disorder  
☐ Attention Deficit Hyperactivity Disorder  
☐ Check All

Please select the phenotype or phenotypes that you are interested in accessing. Later you will be able to specify data freezes and upload documentation.



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

Users upload any required documentation.

Significantly reduces reliance on email attachments!

Slide courtesy: Lea Davis

# 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?

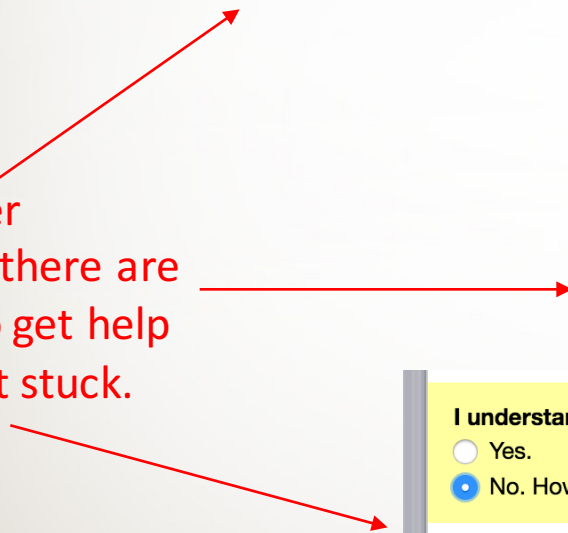
Please refer to the [PGC Website](#) or contact the Major Depression data set representative, [Dr. Cathryn Lewis](#), for information on how to submit an analysis proposal.

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 rounded 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.**

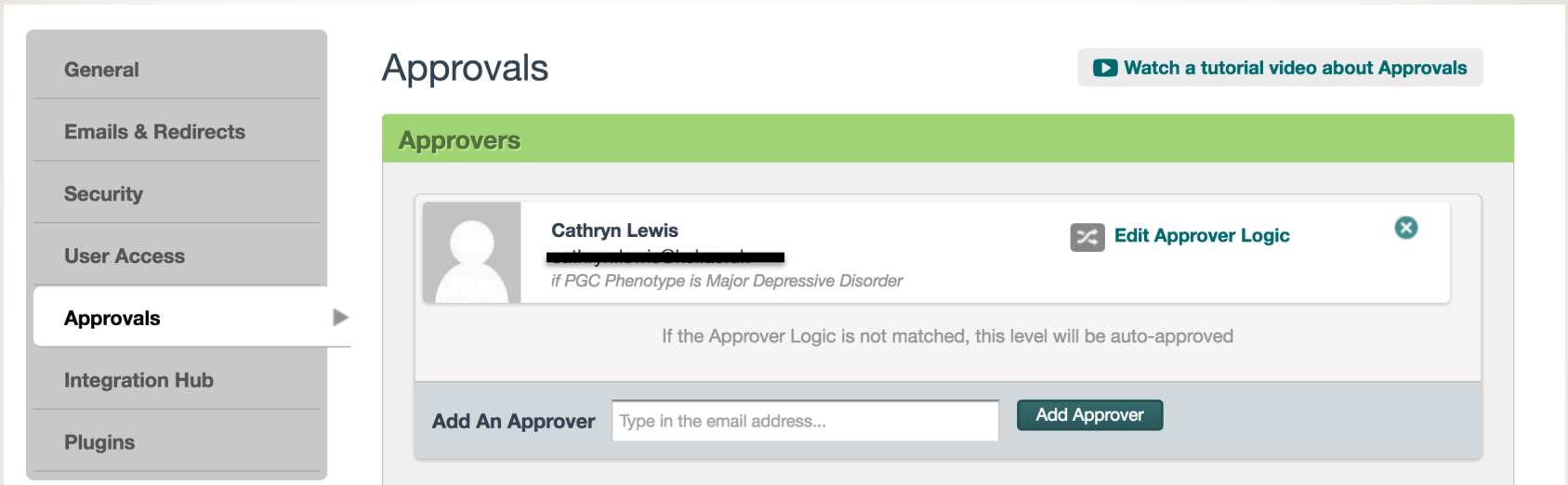
**You will be notified of the outcome within 1-2 weeks. If you experience delays or have questions or concerns, please contact Krista Latta ([krilatta@email.unc.edu](mailto:krilatta@email.unc.edu)).**

**~ the PGC Data Access team**

Encouragement

Contact info

# Behind the scenes of your access request



General

Emails & Redirects

Security

User Access

**Approvals**


Integration Hub

Plugins

## Approvals

[Watch a tutorial video about Approvals](#)

### Approvers

 **Cathryn Lewis**  
[Redacted Email Address]  
*if PGC Phenotype is Major Depressive Disorder*

[Edit Approver Logic](#)

If the Approver Logic is not matched, this level will be auto-approved

**Add An Approver**  **Add Approver**

- You receive a confirmation email when your request is submitted to the system
- Your request gets routed to the appropriate DAC representative for review
- You receive an approval email when your request is approved
- Upon approval, an email is automatically sent to the LISA help desk with your credentials and the requested data packages
- You will receive an email when you have been added to the appropriate user group

# Automated request sent to LISA helpdesk

PGC Data Access Request Portal



Inbox x



**danielle.posthuma@vu.nl** <noreply+0c02ffa8018d830c@formstack.com>

12:46 AM (14 hours ago) ☆



to sripke, danielle.posth., hic, krilatta, jeroen.engelbe., me ▾

this is an automated notification from Formstack:

Please add user [REDACTED] to the below Unix group(s); and make a symbolic link in his/her home directory to the corresponding README file for this in the READMEs directory in /home/pgcdac

mddw2v01  
mdd00001  
mdd00002  
mdd00003

please send confirmation about completion to [REDACTED] and;  
[REDACTED]

the PGC Data Access Team

# Access!

```
#####  
## README for PGC SCZ wave2 data access, accessible with permission group sczw2v01  
#####  
  
## first of all, check if you are member of permission group sczw2v01 (typing "groups" at the command line)  
## in fact otherwise you should not find this README in your homedirectory  
  
## link to dataset collection  
## https://www.broadinstitute.org/ricopili/  
## https://www.broadinstitute.org/ricopili/  
  
## for datastructure consult this wiki page:  
https://sites.google.com/a/broadinstitute.org/ricopili/  
  
## please find these slides there (starting with slide 20):  
ricopili_imputation_wcpog_oct15_website.pptx  
  
## to use data with ricopili pipeline:  
  
## create a symbolic link to your home directory  
## ln -s /  
  
## basic association test:  
postimp_navi_[version] --mds prune.bfile.cobg.PGC_SCZ49.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 --refiex refiex.sczw2v01 for a quick test (1 genomic chunk) to compare with primary outcome  
  
## it is recommend to test with the above mentioned --refiex 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.
  - Approval: You must be listed on the proposal
  - Permission: You must obtain permission from repositories (i.e., dbGAP, NIMH, SSC, etc.) and any individual data owners
  - Access: ANY member of your lab who wishes to access data must get their own LISA account (including you!)
  - PLEASE NEVER SHARE LOGIN CREDENTIALS

# 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

# Working with repositories

- WTCCC – requires only signed analyst memo
- dbGAP
  - 3 ADHD; 1 MDD/SCZ; 1 TS/OCD; 1 BPD will be combined into a “PGC dbGAP bundle” so all 6 sets can be requested at 1 time, in 1 application, with 1 progress report, and 1 renewal
  - dbGAP DACs
    - Joint Addiction, Aging, and Mental Health DAC
    - National Human Genome Research Institute DAC
  - Estimate ~4-6 weeks (usually shorter, sometimes longer if the government shuts down)

# Additional Repositories

- NIMH repository:  
[https://www.nimhgenetics.org/access\\_data\\_biomaterial.php](https://www.nimhgenetics.org/access_data_biomaterial.php)
- SSC repository:  
<https://sfari.org/resources/sfari-base>
- Need help? Visit the PGC website
  - FAQ
  - Data Access -> Documents for Data Access

# How are we doing so far?

Phenotype	Total number of requests	Total number of approvals	Total number of resubmissions	Approximate average response time
MDD	18	18	5	1.5 days
SCZ	24	24	7	1.3 days
BPD	10	10	4	3 days
AN/ED	6	6	1	1 day
<b>Total</b>	<b>67</b>	<b>67</b>	<b>19 (28%)</b>	<b>1.7 days</b>

Main reasons for resubmissions:

- 1) Wrong documentation uploaded
- 2) Out of date documentation uploaded
- 3) Analyst agreements missing signatures

**Your feedback on the portal is important!**

# Need help?

- The NEW 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)

# Join your colleagues who are accessing data!

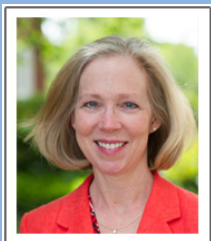




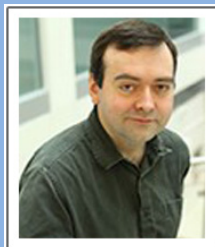


Huge thank you to LISA  
helpdesk!  
Jeroen Engleberts  
Zheng Meyer-Zhao

### Website Committee



Cindy  
Bulik  
(AN/ED)

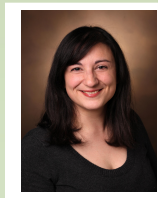


Gerome  
Breen  
(AN/ED)

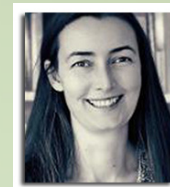
We wish to thank all data owners who  
have graciously shared their data with  
the PGC and worked with the DAC to  
make data available.

We wish to thank all 900,000+ people  
who have entrusted us with their  
genomes and their deepest struggles.

### 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)



Cathryn  
Lewis  
(MDD)



Raymond  
Walters  
(SUD)



Richard  
Anney  
(ASD)



Krista Latta  
(Admin  
Support  
Extraordinaire)



Patrick  
Sullivan  
(Ex-officio  
Member)