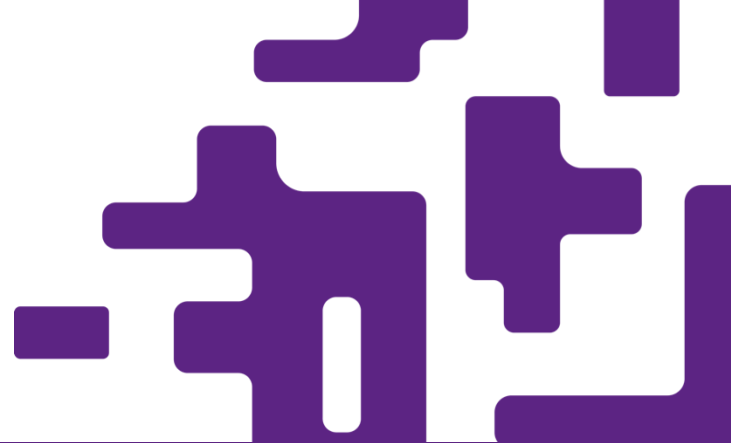




NORMENT

Norwegian Centre for
Mental Disorders Research



Documentation and version control procedures, inside and outside TSD

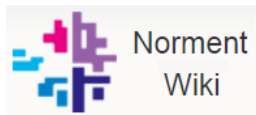
NORMENT & FHI GWAS seminar

Oleksandr Frei, Dec 11, 2018

Contents

- Documentation:

- <http://norment.awiki.org/dokuwiki/>



- Version control inside TSD:

- <http://p33-tl01-l:3000>



- Version control outside TSD:

- <https://github.com/precimed/>

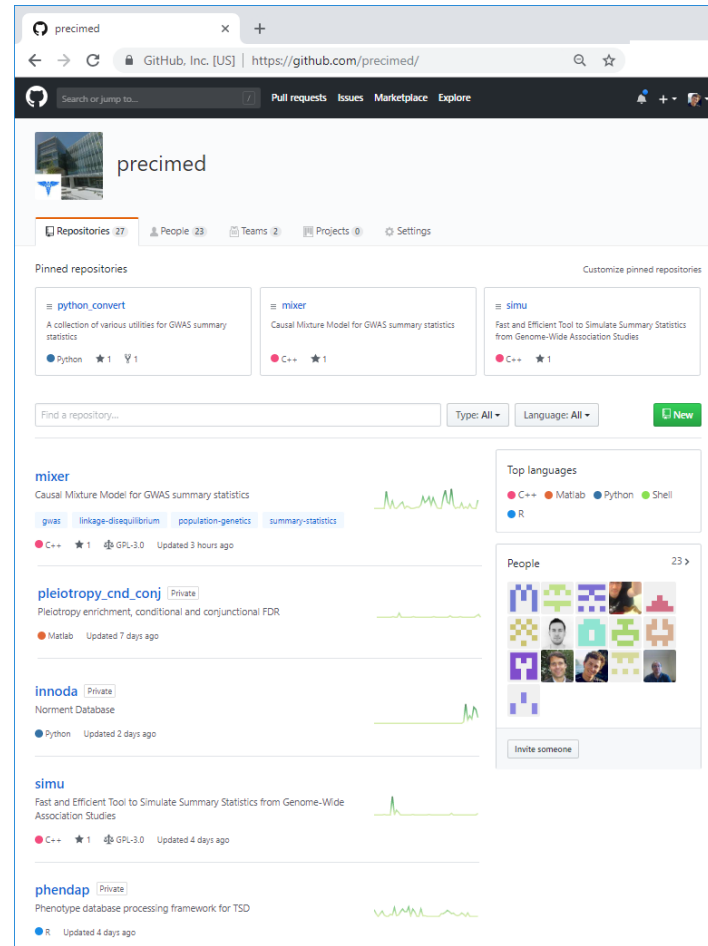


Contents

- Biostats code
 - Code version control and why this is important
 - Brief overview of the biostats tools
 - Practicalities (how to access, how to contribute, etc)
- Biostats inventory for GWAS Summary Statistics
 - Why this is important?
 - Overview of the data in the inventory
 - Practicalities (how to access, how to contribute, caveats, etc)

Code version control and why this is important?

- Everyone has immediate access to the latest version of the code
- Improves user experience:
 - README files (with markup syntax)
 - Releases (additional data, e.g. compiled binaries)
 - Issue tracker
- Improves developers experience:
 - Transparent history of the code
 - Promotes best practices (e.g. code review)
 - People are less afraid of making mistakes
 - Allows to seamlessly sync code across multiple machines
 - Access to free continuous integration, etc
- No more code on USB flash sticks, or on dropbox!



Code on Github

- `python_convert` – a collection of various utilities for GWAS summary statistics
- `pleiotropy_cnd_conj` – pleiotropy enrichment, conditional and conjunctive FDR
- `mixer` – univariate and bivariate causal mixture model
- `simu` – simulate synthetic GWAS summary statistics

The screenshot shows the GitHub profile page for the user 'precimed'. The page header includes the GitHub logo, the username 'precimed', and navigation links for 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. Below the header, there's a section for 'Pinned repositories' featuring four repositories: 'python_convert', 'mixer', 'simu', and 'pleiotropy_cnd_conj'. Each repository card displays its name, a brief description, and a star count. Below the pinned repositories, there's a search bar and a list of repositories. The list includes 'mixer', 'pleiotropy_cnd_conj', 'innoda', 'simu', and 'phendap', each with a description, language, and update time. On the right side, there's a 'Top languages' section showing C++, Matlab, Python, and Shell, and a 'People' section showing a grid of user avatars.

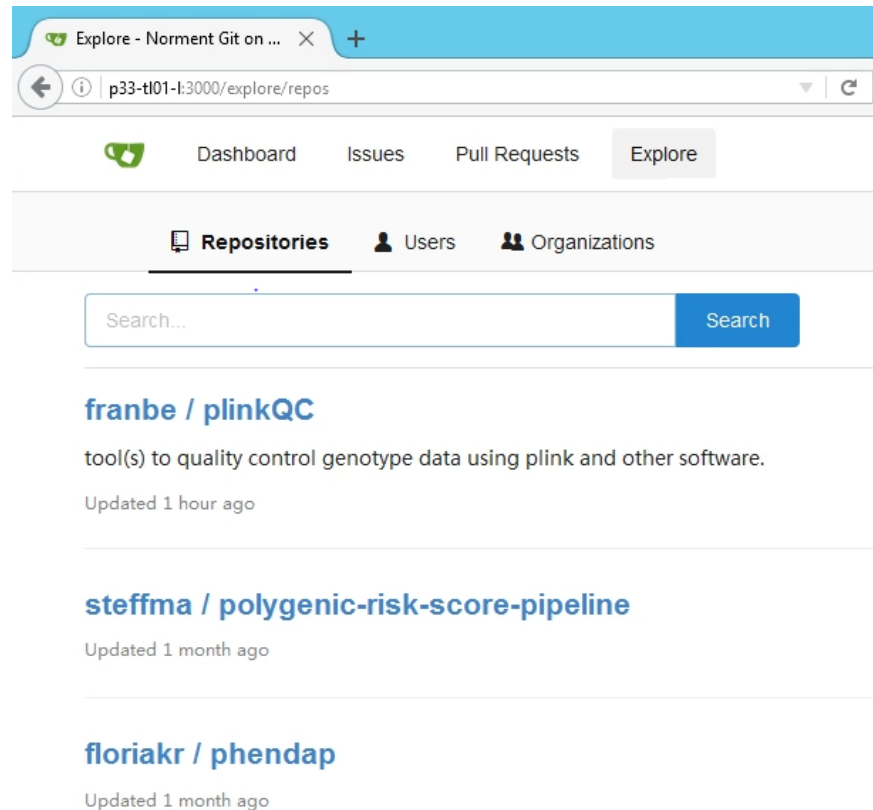
Code on TSD

- Gitea (free open-source alternative to Github)
- To access, login to TSD p33. Then open a browser and go to

<http://p33-t101-1:3000>

Contains code for

- NORMENT phenotype database
- Polygenic risk score pipeline
- Genotype QC and imputation pipeline
- TSD-specific tools



Practical stuff

- Use <https://github.com/precimed/> to access our publicly available tools
- Some repositories are private. To enable your access:
 - Create your free personal GitHub account : <https://github.com/join>
 - Send your GitHub login to oleksandr.frei@gmail.com and ask for access to precimed
 - You will receive an invitation e-mail from GitHub. Accept the invitation, and go to <https://github.com/precimed/>
 - Details: http://norment.awiki.org/dokuwiki/how_to_access_code_github.com_precimed
- You are always welcome to contribute!
 - Create issue ticket if something does not work / is not clear
 - Edit readme files directly on the GitHub (OK to commit directly to master branch)
 - Submit pull requests
 - Create your own repositories. Private repositories is OK, and this is free for precimed

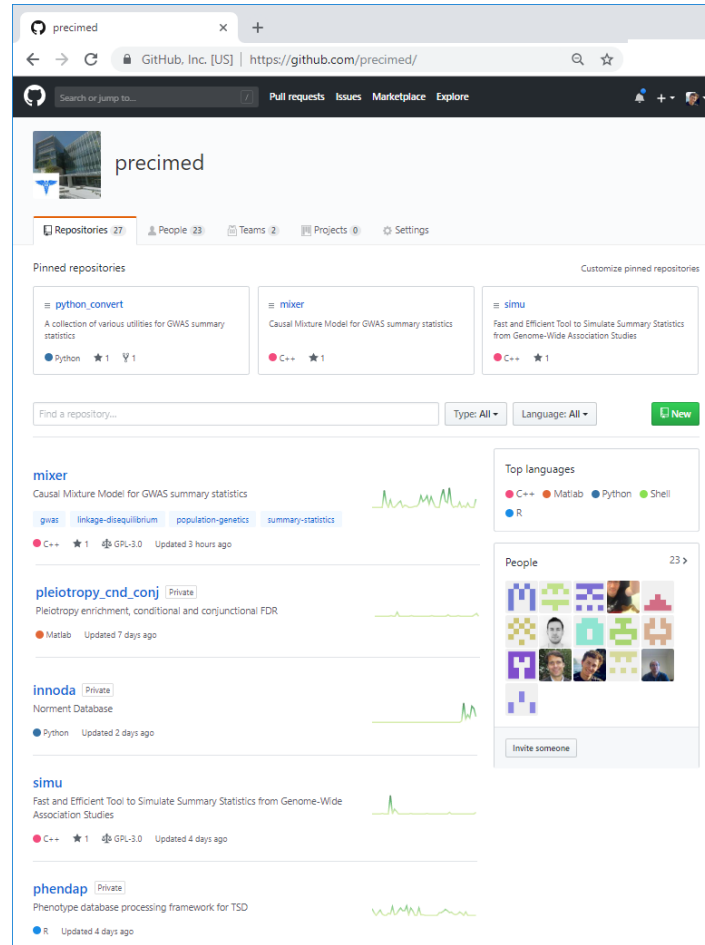
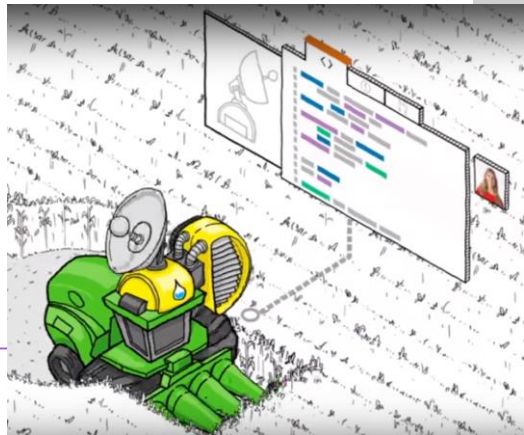
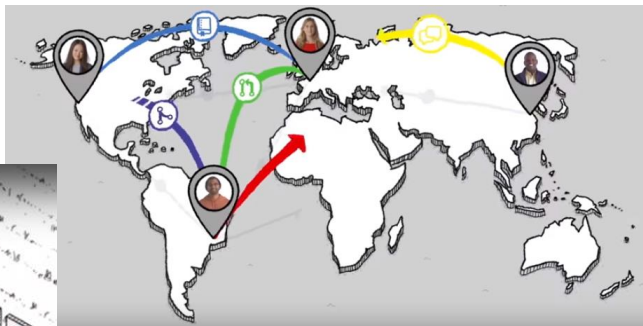
Code version control and why this is important?

Because end users and developers are building software together – which is difficult, but GitHub makes it fun.



What is GitHub?

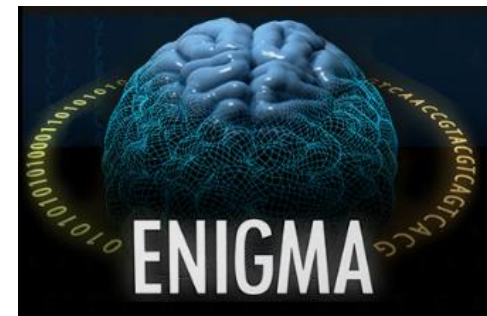
<https://www.youtube.com/watch?v=w3jLJU7DT5E>



The screenshot shows the GitHub profile for 'precimed'. The header includes the repository count (27), people (23), teams (2), projects (0), and settings. The pinned repositories section lists three projects: 'python-convert' (Python, 1 star, 1 fork), 'mixer' (C++, 1 star, 1 fork), and 'simu' (C++, 1 star, 1 fork). The main repository list shows 'mixer' (Causal Mixture Model for GWAS summary statistics, C++, 1 star, 1 fork, updated 3 hours ago), 'pleiotropy_cnd_conj' (Private, Pleiotropy enrichment, conditional and conjunctive FDR, Matlab, updated 7 days ago), 'innoda' (Private, Normant Database, Python, updated 2 days ago), 'simu' (Fast and Efficient Tool to Simulate Summary Statistics from Genome-Wide Association Studies, C++, 1 star, 1 fork, updated 4 days ago), and 'phendap' (Private, Phenotype database processing framework for TSD, R, updated 4 days ago). The right sidebar shows 'Top languages' (C++, Matlab, Python, Shell, R) and 'People' (23).



Psychiatric Genomics Consortium



GWAS summary statistics

- Important for various projects in NORMENT
 - Polygenic overlap between traits
 - Polygenic risk scores
 - New methods development
- Challenges
 - Not all summary statistics are publicly available
 - Formats is not standardized (particularly, column names differ a lot)
 - Some essential columns might be missing (chromosome, position, SNP rs#, etc)
 - Old genomic builds (hg18), old version of dbSNP (i.e. outdated rs#), duplicated SNPs, missing sample size, invalid effect direction, effect direction w.r.t. alternative allele, ...
- The solution is...

Transparent automated pipeline for harmonizing GWAS summary statistics

All steps are
documented
in the log files.

Raw summary stats as downloaded from
various consortia

- NORSTORE/GWAS_SUMSTAT/RAW
- MMIL/GWAS_Original_Summary_Stats

MISC

- supplementary data
- Liftover chain files
 - SNPdb tables
 - 2.5M, 9M reference
 - etc

Standardized summary stats file (STD)

<CONSORTIA>_<TRAIT>_<YEAR>[_qc][_lift][_comment][_noMHC].csv.gz

- CONSORTIA = name of the group that made summary stats (PGC, UKB, ...)
- TRAIT = phenotype abbreviation (SCZ, BIP, ...)
- YEAR = year of the GWAS publication
- qc = optional suffix added if some SNPs are excluded due to QC procedure
- lift = optional suffix added if summary stats are converted from old genomic build, enriched with CHR:POS information or enriched with SNP rs# information
- noMHC = optional suffix added if MHC regions is excluded from summary stats

Summary statistics converted to MATLAB format

- MAT_2M – summary statistics aligned to old 2.5M template
- MAT_9M – summary statistics aligned to new 9M template
- .mat files contains logpvec and zvec information
- .diff files which report difference to previously generated .MAT files

LDSR (LD score regression format)

- LDSR_Data - sumstats.gz files
- MATLAB_Data - converted to matlab
- LDSR_Results (h2, rg) – heritability and genetic correlation estimates

https://github.com/precimed/GWAS_SUMSTAT

https://github.com/precimed/python_convert/

Log file for PGC_SCZ_2014_EUR

Call:

```
./sumstats.py csv \  
  --force \  
  --out /space/syn03/1/data/GWAS/SUMSTAT/STD/INTERMEDIATE/PGC_SCZ_2014_EUR.csv \  
  --ncase-val 33640.0 \  
  --head 5 \  
  --auto \  
  --ncontrol-val 43456.0 \  
  --sumstats /space/syn03/1/data/GWAS/SUMSTAT/RAW/PGC_SCZ_2014_EUR/daner_PGC_SCZ49.sh2_mds10_1000G-frq_2.gz  
Beginning analysis at Thu Nov 9 07:00:31 2017 by oleksandr, host pip31.ucsd.edu  
Reading summary statistics file  
/space/syn03/1/data/GWAS/SUMSTAT/RAW/PGC_SCZ_2014_EUR/daner_PGC_SCZ49.sh2_mds10_1000G-frq_2.gz...  
File header:
```

CHR	SNP	BP	A1	A2	FRQ_A_33640	FRQ_U_43456	INFO	OR	SE	P	ngt	
10	rs185339560		2392426	T	C	0.011	0.011	0.65	1.01339	0.0758	0.8612	0
10	rs11250701		1689546	A	G	0.640	0.640	0.957	1.01147	0.0117	0.3296	0
10	chr10_2622752_D		2622752	I2	D	0.970	0.970	0.933	1.01106	0.0334	0.741	0
10	rs7085086		151476	A	G	0.322	0.322	0.972	1.02685	0.0118	0.02544	0

Interpret column names as follows:

CHR : CHR (Chromosome number)

SNP : SNP (Variant ID (e.g., rs number))

BP : BP (Base-pair position)

A1 : A1 (Allele 1, interpreted as ref allele for signed sumstat.)

A2 : A2 (Allele 2, interpreted as non-ref allele for signed sumstat.)

FRQ_A_33640 : None (Will be deleted)

FRQ_U_43456 : None (Will be deleted)

INFO : INFO (INFO score (imputation quality; higher --> better imputation))

OR : OR (Odds ratio (1 --> no effect; above 1 --> A1 is risk increasing))

SE : SE (standard error of the effect size)

P : PVAL (p-Value)

ngt : None (Will be deleted)

Done. 15358497 SNPs saved to /space/syn03/1/data/GWAS/SUMSTAT/STD/INTERMEDIATE/PGC_SCZ_2014_EUR.csv

Sample size: N=nan NCASE=33640.0 NCONTROL=43456.0



PGC_SCZ_2014_EUR

```
>zcat PGC_SCZ_2014_EUR_gc.csv.gz | head
```

A1	A2	BP	CHR	INFO	NCASE	NCONTROL	OR	PVAL	SE	SNP
I2	D	2622752	10	0.933	33640	43456	1.01106	0.741	0.0334	chr10_2622752_D
A	G	151476	10	0.972	33640	43456	1.02685	0.02544	0.0118	rs7085086
T	G	1593759	10	0.899	33640	43456	0.95285	0.298	0.0464	rs113494187
A	C	1708106	10	0.692	33640	43456	1.0558	0.3168	0.0543	rs117915320
T	C	790310	10	0.617	33640	43456	1.02378	0.3197	0.0236	rs182753344
A	G	1273049	10	0.656	33640	43456	0.96996	0.6473	0.0667	rs188913771
T	G	2067236	10	0.925	33640	43456	1.00481	0.692	0.0121	rs7911665
D	I3	587412	10	0.578	33640	43456	0.9071	0.9322	1.147	chr10_587412_I
I2	D	1734885	10	0.716	33640	43456	0.92997	0.1837	0.0546	chr10_1734885_D

PGC_SCZ_2014_EUR

```
>zcat PGC_SCZ_2014_EUR_gc.csv.gz | head
```

A1	A2	BP	CHR	INFO	NCASE	NCONTROL	OR	PVAL	SE	SNP
I2	D	2622752	10	0.933	33640	43456	1.01106	0.741	0.0334	chr10_2622752_D
A	G	151476	10	0.972	33640	43456	1.02685	0.02544	0.0118	rs7085086
T	G	1593759	10	0.891						
A	C	1708106	10	0.691						
T	C	790310	10	0.611						
A	G	1273049	10	0.651						
T	G	2067236	10	0.921						
D	I3	587412	10	0.571						
I2	D	1734885	10	0.711						

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)



<https://xkcd.com/927/>

Age at menarche
Age at menopause
Agreeableness
Aggression
Alcohol consumption
Alzheimer disorder
Amyotrophic Lateral Sclerosis
Antisocial Behavior
Anxiety
ADHD
Autism Spectrum Disorder
Bipolar Disorder
Blood pressure
Body Mass Index
Borderline personality disorder
Breast Cancer
C Reactive Protein
Cannabis lifetime use
Chronotype
Cognitive performance
College completion
Colorectal cancer
Conscientiousness
Coronary artery disease
Crohn's disease

Depressive Symptoms
Educational attainment
Ever smoked
Extraversion
Former smoker
General cognitive ability
Gestational age
Height
High-density lipids
Hippocampal volume
Inflammatory Bowel Disease
Insomnia
Intelligence
Intracranial volume
Loneliness / social isolation
Low-density lipids
Lung cancer
Major Depressive Disorder
Major depressive disorder
Memory
Migraine
Multiple System Atrophy
Neuroticism
Obsessive-compulsive disorder
Openness

Ovarian cancer
Pallidum volume
Parkinson's disease
Posttraumatic Stress Disorder
Preterm birth
Prostate cancer
Putamen volume
Reaction Time
Rheumatoid arthritis
Schizophrenia
Sleep Duration
Smoke onset
Stroke
Subjective Well Being
Total Cholesterol
Triglycerides
Type2 Diabetes
Ulcerative Colitis
Various blood cell phenotypes
Verbal Numeric Reasoning
Vitamin B12
Vitamin D level
Voice break
Waist Hip Ratio
Years of Educational Attainment

SUMSTAT inventory



MMIL-Oslo GWAS Inventory v2



File Edit View Insert Format Data Tools Form Add-ons Help Last edit was on November 23



SHARE

0

Consortia										
	B	C	D	E	F	G	H	I	J	K
1	Consortia	Phenotype	Phenotype abbreviation	Quantitative or qualitative	CONSORTIA_TRAIT_YE	Total sample size	Number of cases (affected)	Number of controls (unaffected)	Year of the publication	Publication URL
2	CARDIOGRAM	Coronary artery disease	CAD	Case/control trait	CARDIOGRAM_CAD_2015		60801	123504	2015	https://www.nature.com/articles/nrg3286
3	CHARGE	Cognition	COG	Quantitative trait	CHARGE_COG_2015	53949			2015	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4588881/
4	COGENT	Cognition	COG	Quantitative trait	COGENT_COG_2017	35298			2017	http://www.nature.com/articles/nrg5186
5	COGENT	Cognition	COG	Quantitative trait	COGENT_COG_2017_nc	27888			2017	http://www.nature.com/articles/nrg5186
6	COGS	Prostate cancer	PROSTATE	Case/control trait	COGS_PROSTATE_2013		25074	24272	2013	https://www.nature.com/articles/nrg3286
7	CTG	Intelligence	INTELLIGENCE	Quantitative trait	CTG_INTELLIGENCE_20	78308			2017	http://www.nature.com/articles/nrg5186
8	CTG	Insomnia	INSOMNIA	Case/control trait	CTG_INSOMNIA_2017		32384	80622	2017	http://www.nature.com/articles/nrg5186
9	CTG	Insomnia	INSOMNIA	Case/control trait	CTG_INSOMNIA_2017_males		12863	40776	2017	http://www.nature.com/articles/nrg5186
10	CTG	Insomnia	INSOMNIA	Case/control trait	CTG_INSOMNIA_2017_females		19521	39846	2017	http://www.nature.com/articles/nrg5186
11	DIAGRAM	Type2 Diabetes	T2D	Case/control trait	DIAGRAM_T2D_2012		34840	114981	2012	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488881/
12	DIAGRAM	Type2 Diabetes	T2D	Case/control trait	DIAGRAM_T2D_2016				2016	https://www.nature.com/articles/nrg5186
13	EAGLE	Attention-Deficit/Hyperact	ADHD	Quantitative trait	EAGLE_ADHD_2016	17666			2016	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4588881/
14	GAMEON	Breast Cancer	BREAST	Case/control trait	GAMEON_BREAST_2013_BCAC		15863	40022	2013	http://www.nature.com/articles/nrg3286
15	GAMEON	Colorectal cancer	COLON	Case/control trait	GAMEON_COLON_2015_CORECT		5100	7529	2015	https://www.nature.com/articles/nrg5186
16	GAMEON	Lung cancer	LUNG	Case/control trait	GAMEON_LUNG_2014_TRICL_6studyw		12160	16838	2012	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488881/
17	GAMEON	Ovarian cancer	OVARIAN	Case/control trait	GAMEON_OVARIAN_2013_FOCI		3995	3277	2013	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488881/
18	GIANT	Body Mass Index	BMI	Quantitative trait	GIANT_BMI_2015_EUR	339224			2015	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4588881/
19	GIANT	Height	HEIGHT	Quantitative trait	GIANT_HEIGHT_2014	253288			2014	http://www.nature.com/articles/nrg5186
20	GIANT	Waist Hip Ratio	WHR	Quantitative trait	GIANT_WHR_2015_EUR	224459			2015	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4588881/
21	IGAP	Alzheimer disorder	AD	Case/control trait	IGAP_AD_2013		17008	37154	2013	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488881/
22	IIBDGC	Crohn's disease	CD	Case/control trait	IIBDGC_CD_2015_EUR		5956	14927	2015	http://www.nature.com/articles/nrg5186
23	IIBDGC	Inflammatory Bowel Dise	IBD	Case/control trait	IIBDGC_IBD_2015_EUR		12882	21770	2015	http://www.nature.com/articles/nrg5186



GWAS summary statistics

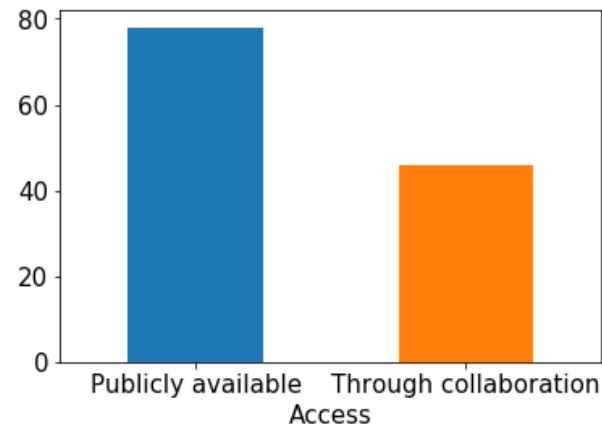
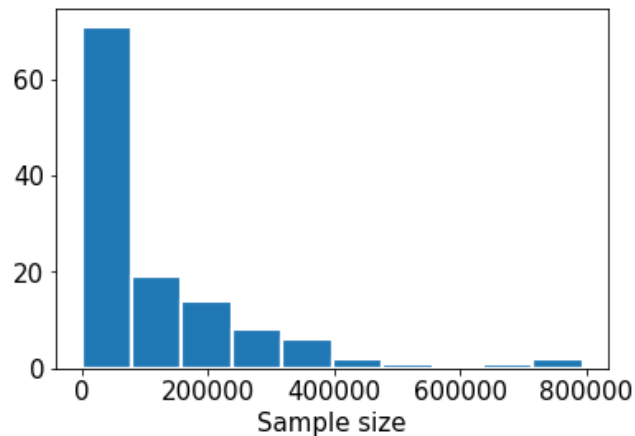
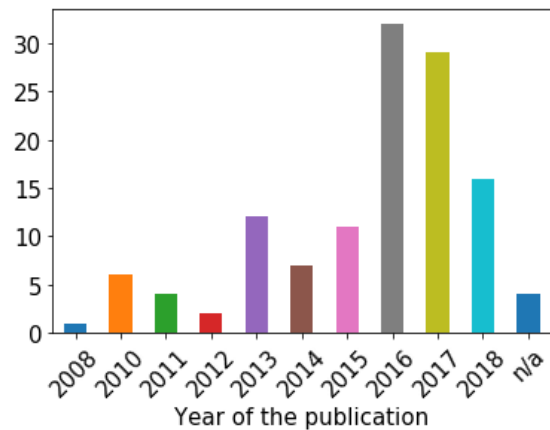


Explore



NORMENT
Norwegian Centre for
Mental Disorders Research

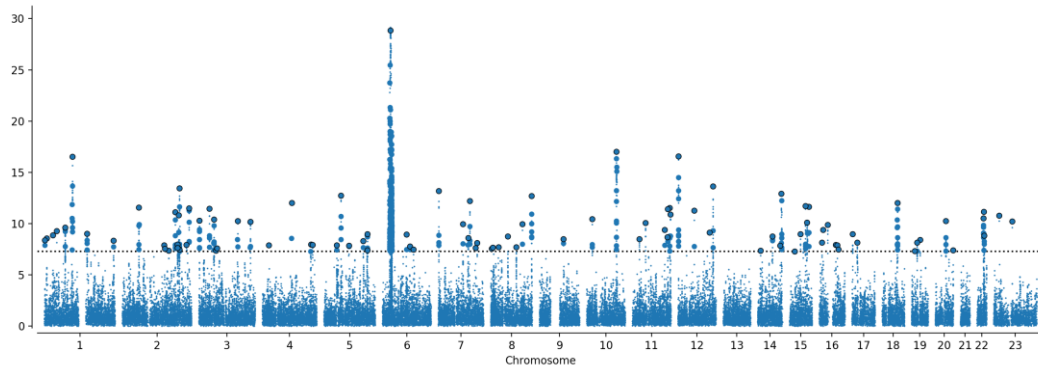
SUMSTAT inventory



Outcome of SUMSTAT pipeline

- All traits named as CONSORTIA_PHENOTYPE_YEAR_<optional tags>
- Format is standardized:
 - Column names are standardized (SNP, CHR, POS, A1, A2, PVAL, BETA, ...)
 - Sample size is provided (either as total N, or as number of cases / number of controls)
 - Chromosome and position always refer to hg19 / grch37 genomic build
- For each trait we generate
 - Manhattan plot
 - Table of genomic loci passing $5e-8$ significance, according to PGC definition
 - LD score regression (inflation, genetic correlation, partitioned heritability)
 - Matlab files compatible with conditional/conjunctive FDR rate
 - Polygenic risk scores in NORMENT TOP sample

PGC_SCZ_2014_EUR

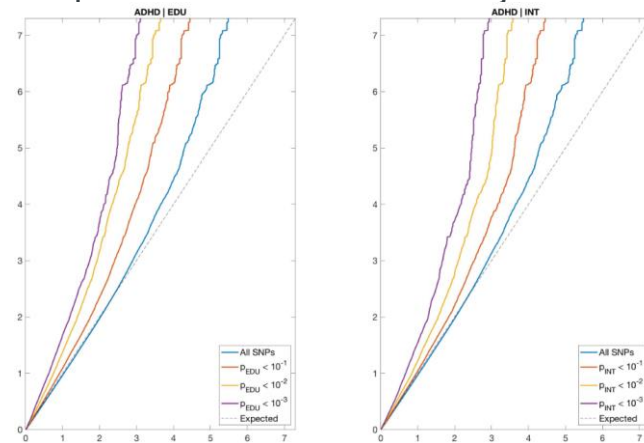


Loci analysis:

- 96 loci
- 141 lead SNPs
- 493 independent significant SNPs
- 53046 candidate SNPs

locusnum	CHR	LEAD_SNP	LEAD_BP	MinBP	MaxBP	PVAL	
38	6	rs13217619	28306671	25038442	33791998	1,44E-29	
57	10	rs11191419	104612335	104229588	105165184	9,24E-18	
65	12	rs2007044	2344960	2285731	2523772	2,63E-17	
6	1	rs1702294	98501984	98033259	98562260	2,79E-17	

Input .MAT files for cond/conj FDR



Results from LD score regression including pairwise genetic correlations across all traits in the inventory

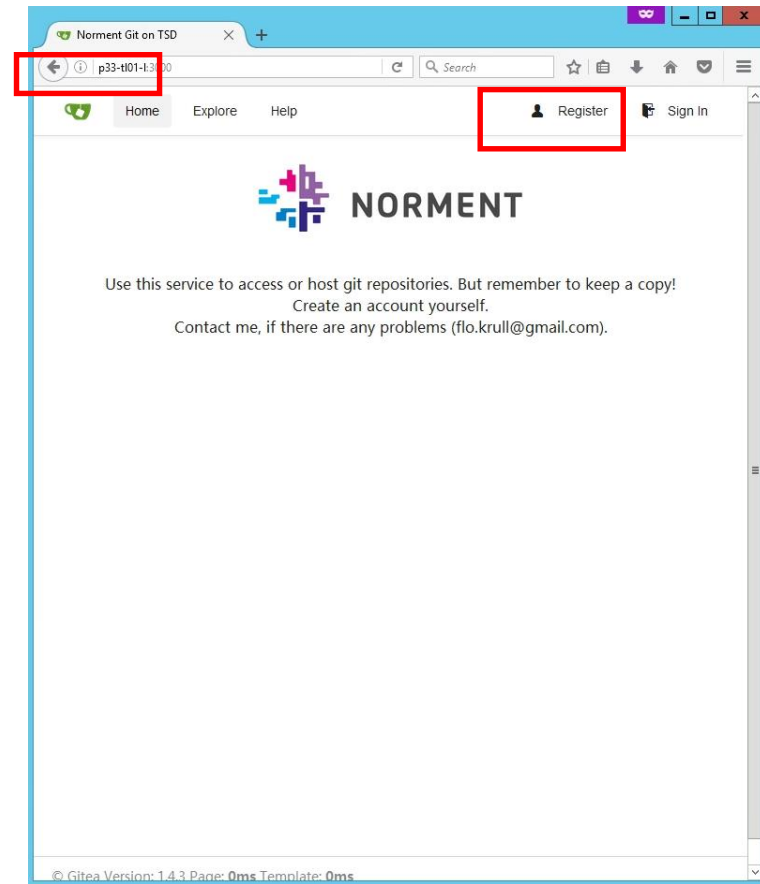
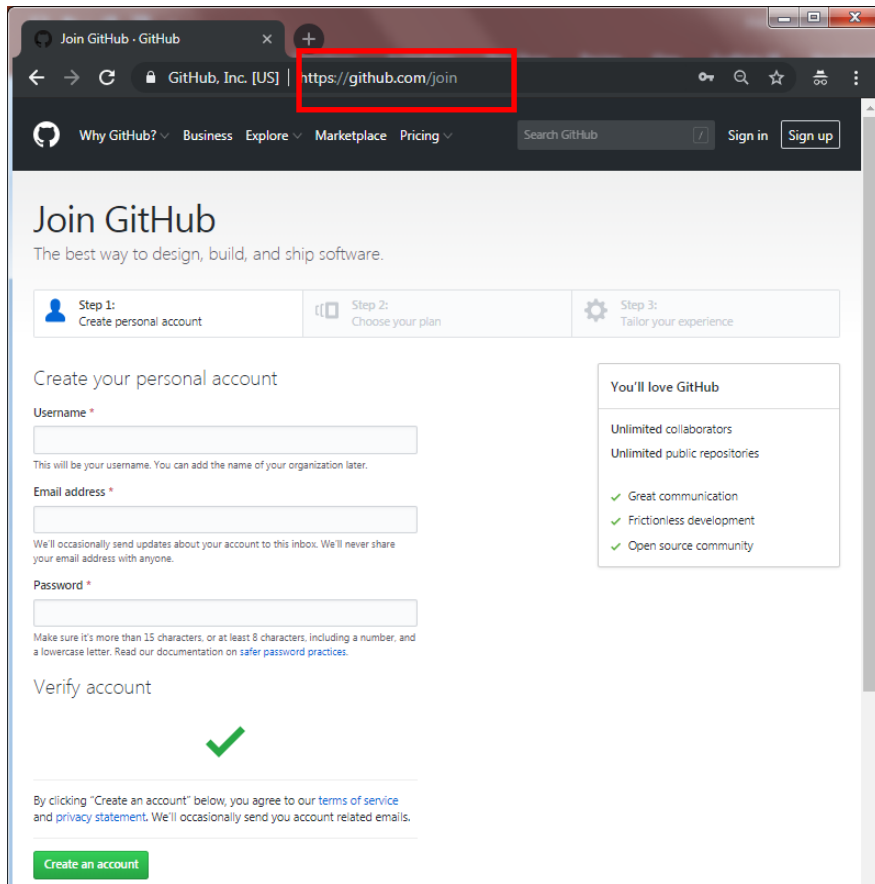
Total Observed scale h²: 0.4545 (0.0184)
 Lambda GC: 1.5883
 Mean Chi²: 1.7678
 Intercept: 1.0588 (0.0119)

Practicalities

- **Access inventory spreadsheet here:**
<https://docs.google.com/spreadsheets/d/19cURugXQQyLgfLU-gwCReuWK99DcpODOpSyDkaR-bow/edit#gid=1515118726>
- **Always check whether the data is “Public” or “Through Collaboration”**
- **Access SUMSTAT inventory on TSD:**
 - `/tsd/p33/data/durable/s3-api/mmil/SUMSTAT/RAW` (original, raw data)
 - `/tsd/p33/data/durable/s3-api/mmil/SUMSTAT/STD` (standardized files)
 - `/tsd/p33/data/durable/s3-api/mmil/SUMSTAT/ANALYSIS` (Manhattan & Loci)
 - `/tsd/p33/data/durable/s3-api/mmil/SUMSTAT/LDSR/LDSR_Results` (h2, rg)
- **Request to add more summary statistics to the inventory:**
<https://docs.google.com/forms/d/e/1FAIpQLSd2wNKmrZlYmlc4j8ByM9m3d1XCtyAI-U7h5uBwMJ3azvjKxA/viewform>
- **More information is in the wiki:**
http://norment.awiki.org/dokuwiki/where_to_find_summary_stats

Caveats

- Multiple versions of data for a given paper:
 - Excluding specific cohorts (“_no23andMe”, “_noUKB”, “_noTOP”)
 - Different populations (“_EUR”)
- Prolific consortia produce multiple GWAS studies per year on a given phenotype:
 - PGC_MDD_2018 - N. Wray et al.
 - PGC_MDD_2018_Howard - D. Howard et al.
- Data accessed through collaboration
 - $\text{PGC_MDD_2018_no23andMe} + \text{23andMe_MDD_2016} = \text{PGC_MDD_2018_with23andMe}$
- Effect direction quite often has issues
 - Good idea to check genetic correlations between your phenotype and related traits in the inventory: for example with other GWAS on the same phenotype
- Investigate sample overlap between studies (especially UK Biobank)



<https://norment.awiki.org/dokuwiki/>



NORMENT

Norwegian Centre for
Mental Disorders Research

Oslo University Hospital HF
Division of Mental Health and Addiction
Psychosis Research Unit/TOP
Ullevål Hospital, building 49
P.O. Box 4956 Nydalen
N-0424 Oslo
Norway