NORMENT
Norwegian Centre for
Mental Disorders Research

Survey of available GWAS and gene expression data + TSD overview

Oleksandr Frei

UiO : University of Oslo
Oslo University Hospital
UNIVERSITY OF BERGEN
Haukeland University Hospital
Norwegian Centre of Excellence
The Research Council of Norway

Overview

- What are typical columns in GWAS summary statistics?
- Where to download GWAS results?
- How to re-format GWAS summary statistics tables?

Practical demo – prereq. for tomorrow's tutorial on LDSR

- Download summary statistics
- Install python via Anaconda
- Harmonize column names using Python Pandas package
- Further details in https://etherpad.net/p/gwasOslo

Introduction to TSD service (tjeneste for sensitive data)

PGC

Psychiatric Genomics Consortium

CNCR CTGLAB

BD Genetics

NORMENT

GIANT CONSORTIUM

Social Science Genetic Association Consortium

biobank uk

project MinE

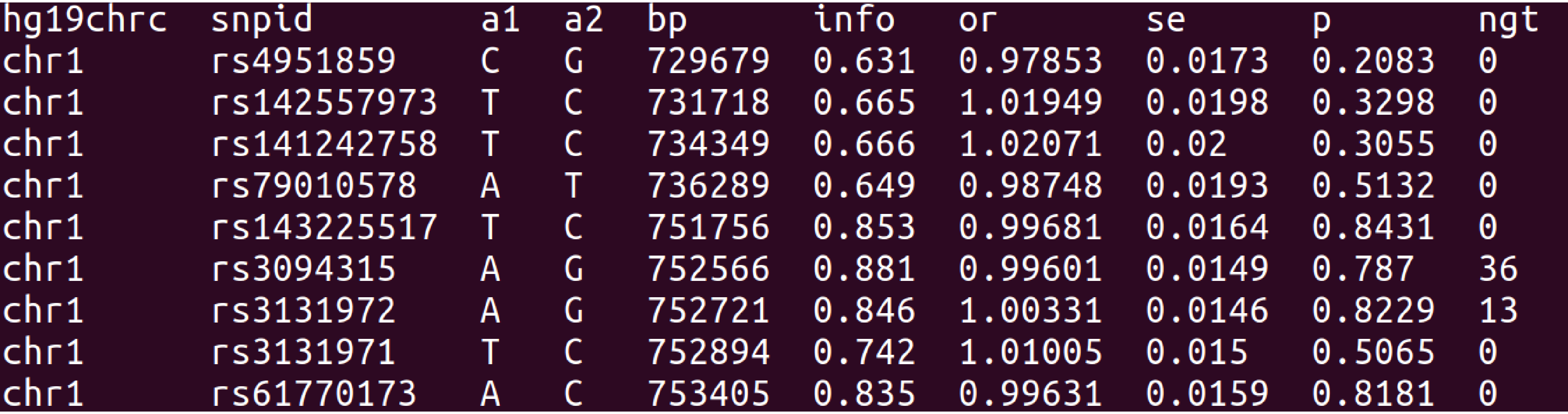
23andMe

ENIGMA

CHARGE CONSORTIUM
Cohorts for Heart & Aging Research in Genomic Epidemiology

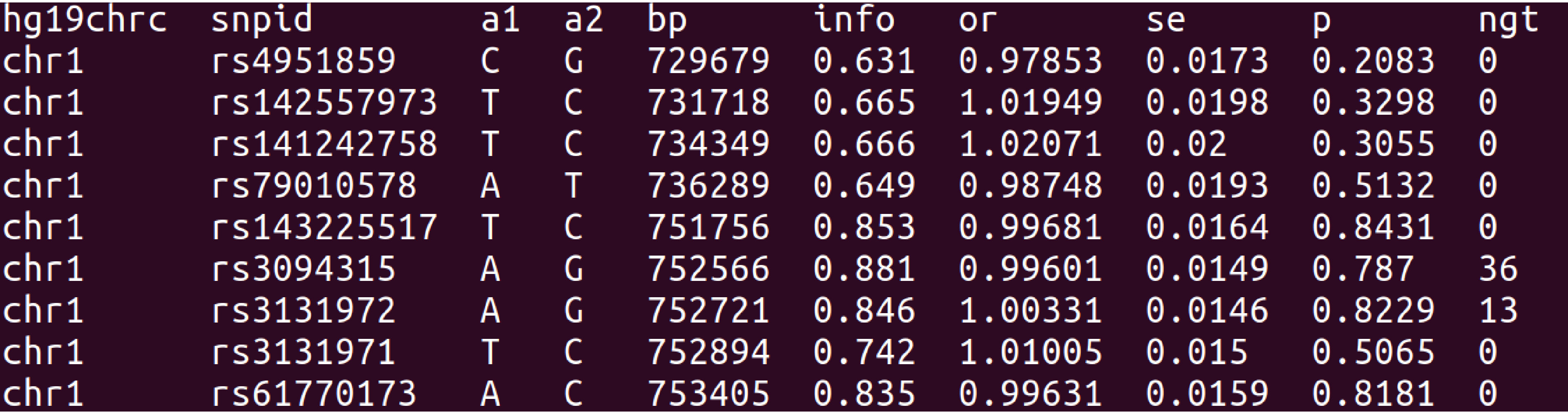
IGAP
INTERNATIONAL
GENOMICS OF ALZHEIMER'S PROJECT

- PGC Schizophrenia 2014 GWAS



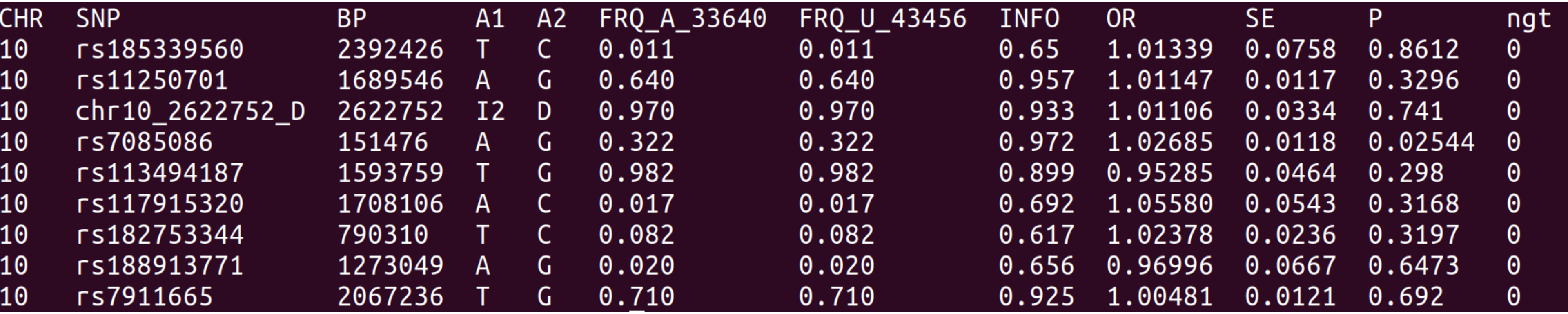
```
hg19chrc   snpid         a1   a2   bp        info    or        se       p        ngt
chr1       rs4951859     C    G    729679    0.631   0.97853   0.0173   0.2083   0
chr1       rs142557973   T    C    731718    0.665   1.01949   0.0198   0.3298   0
chr1       rs141242758   T    C    734349    0.666   1.02071   0.02     0.3055   0
chr1       rs79010578    A    T    736289    0.649   0.98748   0.0193   0.5132   0
chr1       rs143225517   T    C    751756    0.853   0.99681   0.0164   0.8431   0
chr1       rs3094315     A    G    752566    0.881   0.99601   0.0149   0.787    36
chr1       rs3131972     A    G    752721    0.846   1.00331   0.0146   0.8229   13
chr1       rs3131971     T    C    752894    0.742   1.01005   0.015    0.5065   0
chr1       rs61770173    A    C    753405    0.835   0.99631   0.0159   0.8181   0
```

- PGC Schizophrenia 2014 GWAS



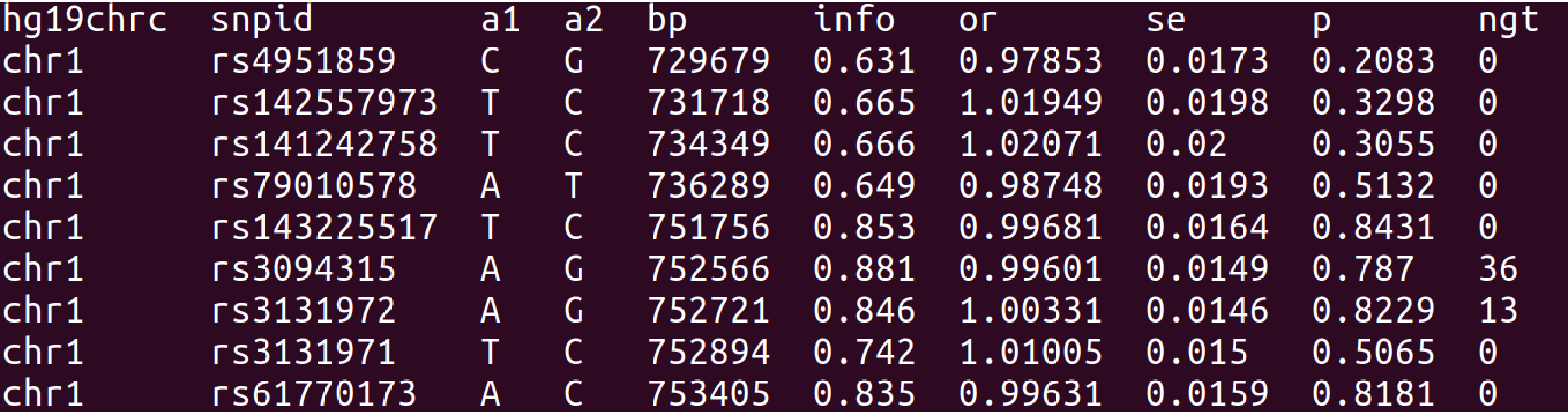
```
hg19chrc   snpid          a1   a2   bp       info    or        se       p        ngt
chr1       rs4951859      C    G    729679   0.631   0.97853   0.0173   0.2083   0
chr1       rs142557973    T    C    731718   0.665   1.01949   0.0198   0.3298   0
chr1       rs141242758    T    C    734349   0.666   1.02071   0.02     0.3055   0
chr1       rs79010578     A    T    736289   0.649   0.98748   0.0193   0.5132   0
chr1       rs143225517    T    C    751756   0.853   0.99681   0.0164   0.8431   0
chr1       rs3094315      A    G    752566   0.881   0.99601   0.0149   0.787    36
chr1       rs3131972      A    G    752721   0.846   1.00331   0.0146   0.8229   13
chr1       rs3131971      T    C    752894   0.742   1.01005   0.015    0.5065   0
chr1       rs61770173     A    C    753405   0.835   0.99631   0.0159   0.8181   0
```

PGC Schizophrenia 2014 EUR-only GWAS



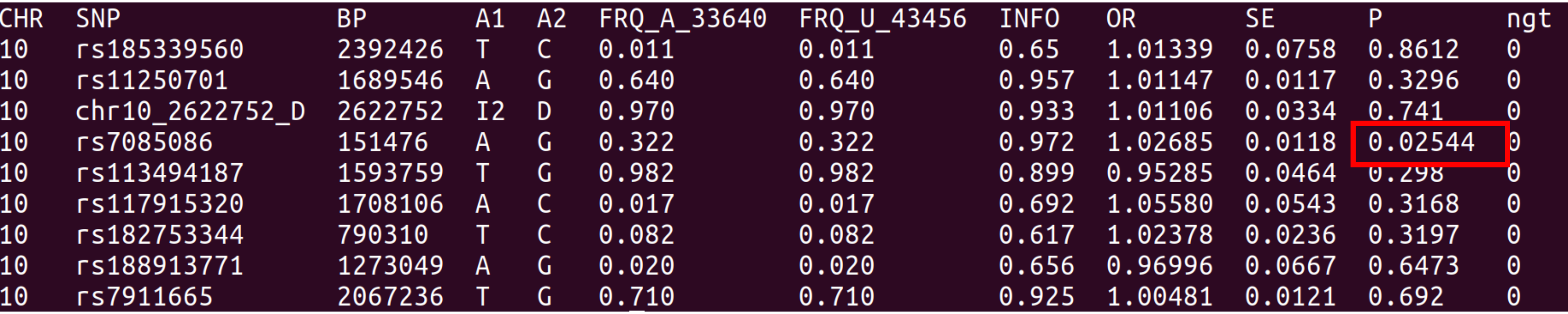
```
CHR  SNP                BP        A1   A2   FRQ_A_33640  FRQ_U_43456  INFO    OR        SE       P         ngt
10   rs185339560        2392426   T    C    0.011        0.011        0.65    1.01339   0.0758   0.8612    0
10   rs11250701         1689546   A    G    0.640        0.640        0.957   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
10   rs113494187        1593759   T    G    0.982        0.982        0.899   0.95285   0.0464   0.298     0
10   rs117915320        1708106   A    C    0.017        0.017        0.692   1.05580   0.0543   0.3168    0
10   rs182753344        790310    T    C    0.082        0.082        0.617   1.02378   0.0236   0.3197    0
10   rs188913771        1273049   A    G    0.020        0.020        0.656   0.96996   0.0667   0.6473    0
10   rs7911665          2067236   T    G    0.710        0.710        0.925   1.00481   0.0121   0.692     0
```

- PGC Schizophrenia 2014 GWAS



```
hg19chrc  snpid         a1  a2  bp      info   or       se      p       ngt
chr1      rs4951859     C   G   729679  0.631  0.97853  0.0173  0.2083  0
chr1      rs142557973   T   C   731718  0.665  1.01949  0.0198  0.3298  0
chr1      rs141242758   T   C   734349  0.666  1.02071  0.02    0.3055  0
chr1      rs79010578    A   T   736289  0.649  0.98748  0.0193  0.5132  0
chr1      rs143225517   T   C   751756  0.853  0.99681  0.0164  0.8431  0
chr1      rs3094315     A   G   752566  0.881  0.99601  0.0149  0.787   36
chr1      rs3131972     A   G   752721  0.846  1.00331  0.0146  0.8229  13
chr1      rs3131971     T   C   752894  0.742  1.01005  0.015   0.5065  0
chr1      rs61770173    A   C   753405  0.835  0.99631  0.0159  0.8181  0
```

PGC Schizophrenia 2014 EUR-only GWAS



```
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
10   rs113494187      1593759  T   G   0.982         0.982         0.899  0.95285  0.0464  0.298    0
10   rs117915320      1708106  A   C   0.017         0.017         0.692  1.05580  0.0543  0.3168   0
10   rs182753344      790310   T   C   0.082         0.082         0.617  1.02378  0.0236  0.3197   0
10   rs188913771      1273049  A   G   0.020         0.020         0.656  0.96996  0.0667  0.6473   0
10   rs7911665        2067236  T   G   0.710         0.710         0.925  1.00481  0.0121  0.692    0
```

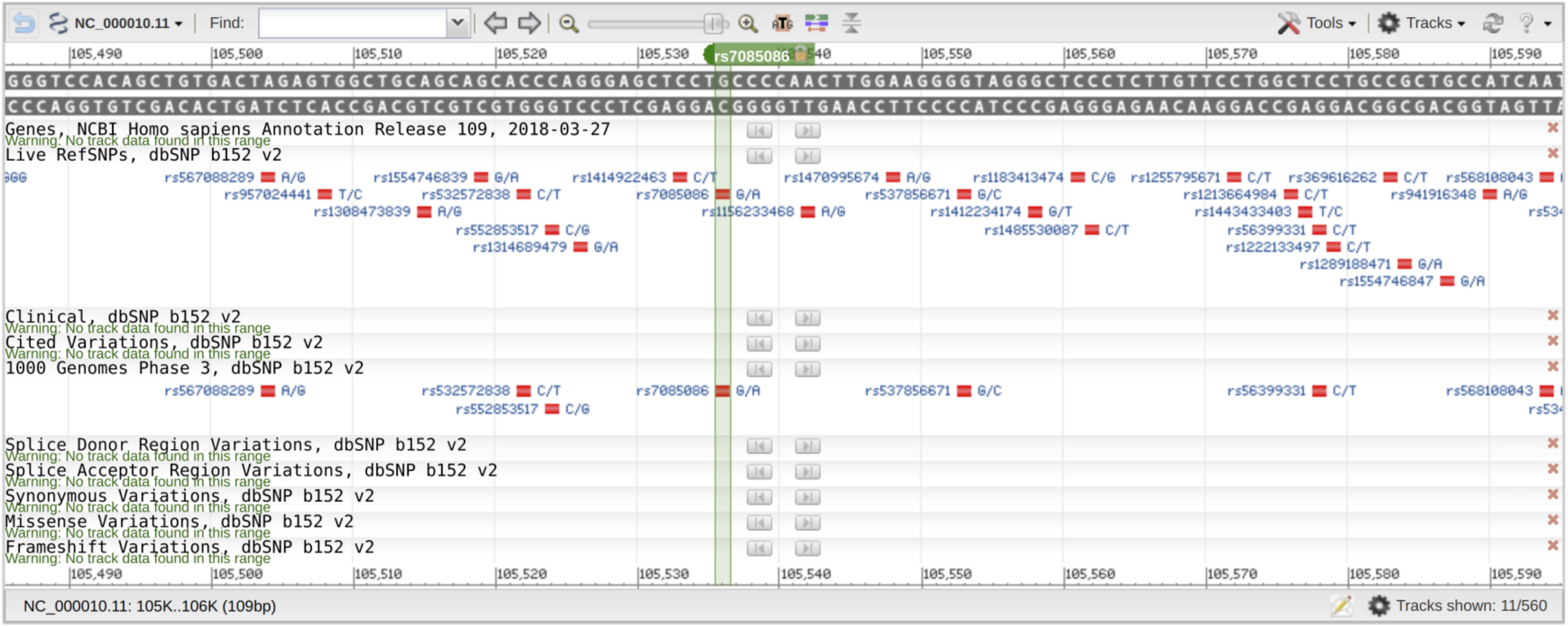
NORMENT
Norwegian Centre for
Mental Disorders Research

https://www.ncbi.nlm.nih.gov/snp/?term=rs7085086

1. □ rs7085086 *[Homo sapiens]*

ctgcagcagcacccagggagctcct[A/G]ccccaacttggaaggggtagggctc

Chromosome:	10:105536
Validated:	by 1000G,by 2hit 2allele,by cluster,by frequency
Global MAF:	A=0.2370/1187
HGVS:	CM000672.2:g.105536G>A, NC_000010.10:g.151476G>A, NW_003571043.1:g.95536G>A



GGGTCCACAGCTGTGACTAGAGTGGCTGCAGCAGCACCCAGGGAGCTCCTGCCCCAACTTGGAAGGGGTAGGGCTCCCTCTTGTTCCTGGCTCCTGCCGCTGCCATCAAT
CCCAGGTGTCGACACTGATCTCACCGACGTCGTCGTGGGTCCCTCGAGGACGGGGTTGAACCTTCCCCATCCCGAGGGAGAACAAGGACCGAGGACGGCGACGGTAGTTA

Genes, NCBI Homo sapiens Annotation Release 109, 2018-03-27
Warning: No track data found in this range

Live RefSNPs, dbSNP b152 v2

Clinical, dbSNP b152 v2
Warning: No track data found in this range

Cited Variations, dbSNP b152 v2
Warning: No track data found in this range

1000 Genomes Phase 3, dbSNP b152 v2

Splice Donor Region Variations, dbSNP b152 v2
Warning: No track data found in this range

Splice Acceptor Region Variations, dbSNP b152 v2
Warning: No track data found in this range

Synonymous Variations, dbSNP b152 v2
Warning: No track data found in this range

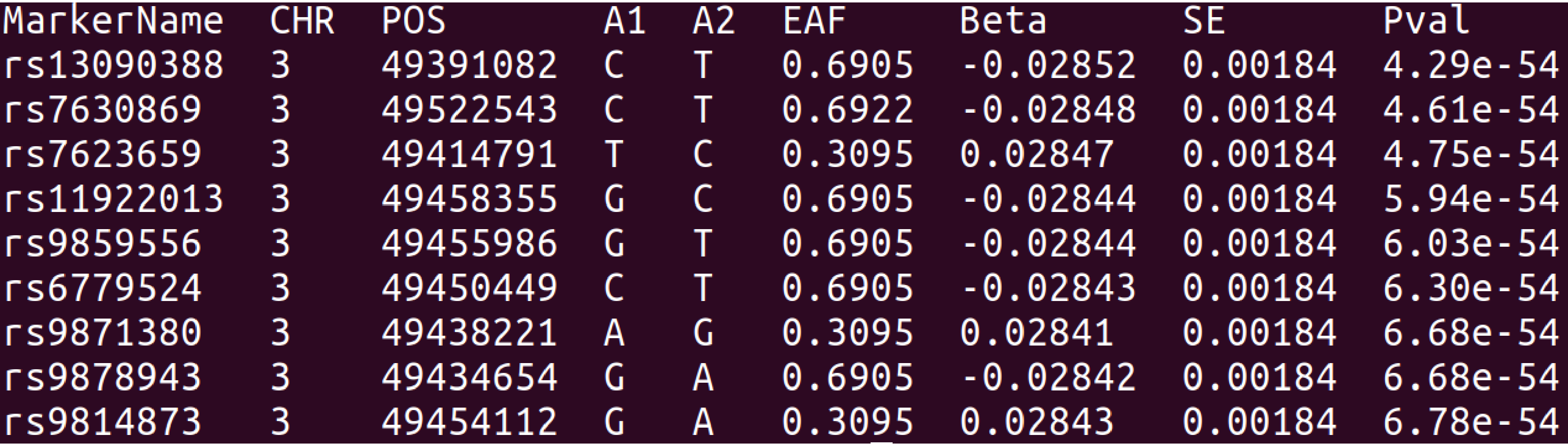
Missense Variations, dbSNP b152 v2
Warning: No track data found in this range

Frameshift Variations, dbSNP b152 v2
Warning: No track data found in this range

NC_000010.11: 105K..106K (109bp)

Tracks shown: 11/560

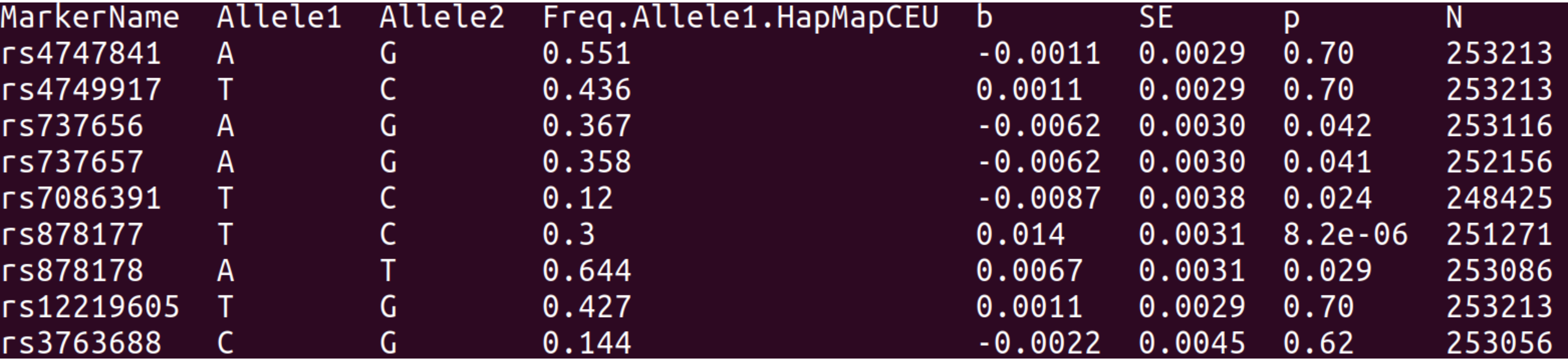
SSGAC Educational Attainment GWAS



```
MarkerName  CHR  POS       A1  A2  EAF     Beta      SE       Pval
rs13090388  3    49391082  C   T   0.6905  -0.02852  0.00184  4.29e-54
rs7630869   3    49522543  C   T   0.6922  -0.02848  0.00184  4.61e-54
rs7623659   3    49414791  T   C   0.3095  0.02847   0.00184  4.75e-54
rs11922013  3    49458355  G   C   0.6905  -0.02844  0.00184  5.94e-54
rs9859556   3    49455986  G   T   0.6905  -0.02844  0.00184  6.03e-54
rs6779524   3    49450449  C   T   0.6905  -0.02843  0.00184  6.30e-54
rs9871380   3    49438221  A   G   0.3095  0.02841   0.00184  6.68e-54
rs9878943   3    49434654  G   A   0.6905  -0.02842  0.00184  6.68e-54
rs9814873   3    49454112  G   A   0.3095  0.02843   0.00184  6.78e-54
```

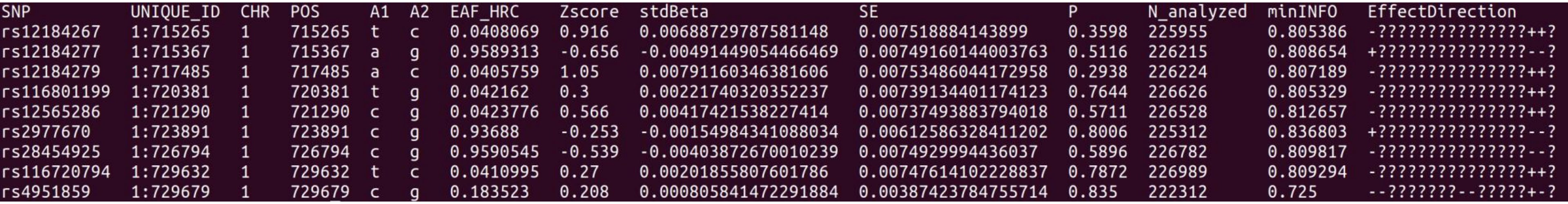
README file:
MarkerName: SNP rs number.
CHR: chromosome number.
POS: base pair position.
A1: effect allele.
A2: other allele.
EAF: A1 frequency in 1000 Genomes Phase 3 sample (CEU, GBR and TSI individuals).
Beta_*: Standardized regression coefficient, i.e. per-allele effect size on the phenotype that has been standardized to have unit variance.
SE_*: standard error of Beta
Pval_*: Nominal p-value of the null hypothesis that the coefficient is equal to zero.

- GIANT Height 2014



```
MarkerName  Allele1  Allele2  Freq.Allele1.HapMapCEU  b        SE      p        N
rs4747841   A        G        0.551                   -0.0011  0.0029  0.70     253213
rs4749917   T        C        0.436                   0.0011   0.0029  0.70     253213
rs737656    A        G        0.367                   -0.0062  0.0030  0.042    253116
rs737657    A        G        0.358                   -0.0062  0.0030  0.041    252156
rs7086391   T        C        0.12                    -0.0087  0.0038  0.024    248425
rs878177    T        C        0.3                     0.014    0.0031  8.2e-06  251271
rs878178    A        T        0.644                   0.0067   0.0031  0.029    253086
rs12219605  T        G        0.427                   0.0011   0.0029  0.70     253213
rs3763688   C        G        0.144                   -0.0022  0.0045  0.62     253056
```

- CTG Cognition 2018



```
SNP          UNIQUE_ID  CHR  POS     A1  A2  EAF_HRC    Zscore  stdBeta                SE                   P       N_analyzed  minINFO   EffectDirection
rs12184267   1:715265   1    715265  t   c   0.0408069  0.916   0.00688729787581148    0.007518884143899    0.3598  225955      0.805386  -???????????????++?
rs12184277   1:715367   1    715367  a   g   0.9589313  -0.656  -0.00491449054466469   0.00749160144003763  0.5116  226215      0.808654  +???????????????--?
rs12184279   1:717485   1    717485  a   c   0.0405759  1.05    0.00791160346381606    0.00753486044172958  0.2938  226224      0.807189  -???????????????++?
rs116801199  1:720381   1    720381  t   g   0.042162   0.3     0.00221740320352237    0.00739134401174123  0.7644  226626      0.805329  -???????????????++?
rs12565286   1:721290   1    721290  c   g   0.0423776  0.566   0.00417421538227414    0.00737493883794018  0.5711  226528      0.812657  -???????????????++?
rs2977670    1:723891   1    723891  c   g   0.93688    -0.253  -0.00154984341088034   0.00612586328411202  0.8006  225312      0.836803  +???????????????--?
rs28454925   1:726794   1    726794  c   g   0.9590545  -0.539  -0.00403872670010239   0.0074929994436037   0.5896  226782      0.809817  -???????????????--?
rs116720794  1:729632   1    729632  t   c   0.0410995  0.27    0.00201855807601786    0.00747614102228837  0.7872  226989      0.809294  -???????????????++?
rs4951859    1:729679   1    729679  c   g   0.183523   0.208   0.000805841472291884   0.00387423784755714  0.835   222312      0.725     --???????--?????+-?
```

Navigate through data sources:

Specific consortia:

- PGC - https://www.med.unc.edu/pgc/results-and-downloads
- SSGAC - https://www.thessgac.org/data
- CTG - https://ctg.cncr.nl/software/summary_statistics
- IBD Genetics - https://www.ibdgenetics.org/downloads.html

General catalogs:

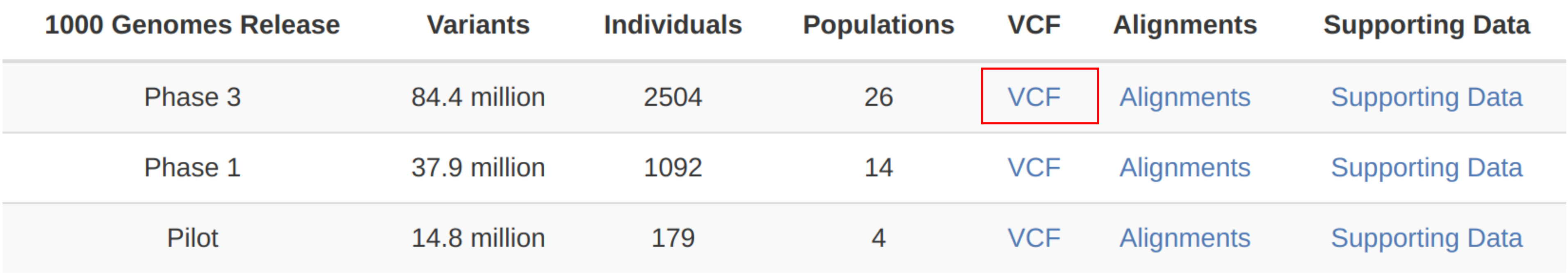
- GWAS catalog: https://www.ebi.ac.uk/gwas/
- LDhub: http://ldsc.broadinstitute.org/ldhub/
- UK Biobank: https://biobank.ctsu.ox.ac.uk/showcase/

Reference data:

- 1000 Genomes
- HRC
- dbSNP

1000 Genomes and HRC references

- http://www.internationalgenome.org/data/



1000 Genomes Release	Variants	Individuals	Populations	VCF	Alignments	Supporting Data
Phase 3	84.4 million	2504	26	VCF	Alignments	Supporting Data
Phase 1	37.9 million	1092	14	VCF	Alignments	Supporting Data
Pilot	14.8 million	179	4	VCF	Alignments	Supporting Data

https://vcftools.github.io/index.html <- software to work with VCF format

- http://www.haplotype-reference-consortium.org/

Tables manipulation exercise:

Input:

```
  hg19chrc        snpid a1 a2     bp  info      or     se      p  ngt
0     chr1    rs4951859  C  G 729679 0.631 0.97853 0.0173 0.2083    0
1     chr1  rs142557973  T  C 731718 0.665 1.01949 0.0198 0.3298    0
2     chr1  rs141242758  T  C 734349 0.666 1.02071 0.0200 0.3055    0
3     chr1   rs79010578  A  T 736289 0.649 0.98748 0.0193 0.5132    0
4     chr1  rs143225517  T  C 751756 0.853 0.99681 0.0164 0.8431    0
```

Intermediate result:

```
   CHR          SNP A1 A2     BP  INFO      OR     SE  PVALUE
0  chr1    rs4951859  C  G 729679 0.631 0.97853 0.0173 0.2083
1  chr1  rs142557973  T  C 731718 0.665 1.01949 0.0198 0.3298
2  chr1  rs141242758  T  C 734349 0.666 1.02071 0.0200 0.3055
3  chr1   rs79010578  A  T 736289 0.649 0.98748 0.0193 0.5132
4 chr1  rs143225517  T  C 751756 0.853 0.99681 0.0164 0.8431
```

Final result:

```
   CHR          SNP A1 A2     bp  INFO      OR     SE  PVALUE     N         Z
0    1    rs4951859  C  G 729679 0.631 0.97853 0.0173 0.2083 82315 -1.258254
1    1  rs142557973  T  C 731718 0.665 1.01949 0.0198 0.3298 82315  0.974517
2    1  rs141242758  T  C 734349 0.666 1.02071 0.0200 0.3055 82315  1.024710
3    1   rs79010578  A  T 736289 0.649 0.98748 0.0193 0.5132 82315 -0.653863
4    1  rs143225517  T  C 751756 0.853 0.99681 0.0164 0.8431 82315 -0.197930
```

```
>>> import pandas as pd
>>> df=pd.read_table('ckqny.scz2snpres.gz', delim_whitespace=True)
>>> df.head()
  hg19chrc        snpid a1 a2      bp   info       or      se      p  ngt
0     chr1    rs4951859  C  G  729679  0.631  0.97853  0.0173  0.2083    0
1     chr1  rs142557973  T  C  731718  0.665  1.01949  0.0198  0.3298    0
2     chr1  rs141242758  T  C  734349  0.666  1.02071  0.0200  0.3055    0
3     chr1   rs79010578  A  T  736289  0.649  0.98748  0.0193  0.5132    0
4     chr1  rs143225517  T  C  751756  0.853  0.99681  0.0164  0.8431    0

>>> df.rename(columns={'hg19chrc':'CHR', 'snpid':'SNP', 'a1':'A1', 'a2':'A2',
'or':'OR', 'se':'SE', 'p':'PVALUE', 'bp':'BP', 'info':'INFO'}, inplace=True)
>>> df.drop(columns=['ngt'], inplace=True)
>>> df.head()
    CHR          SNP A1  A2      BP   INFO       OR      SE  PVALUE
0  chr1    rs4951859  C   G  729679  0.631  0.97853  0.0173  0.2083
1  chr1  rs142557973  T   C  731718  0.665  1.01949  0.0198  0.3298
2  chr1  rs141242758  T   C  734349  0.666  1.02071  0.0200  0.3055
3  chr1   rs79010578  A   T  736289  0.649  0.98748  0.0193  0.5132
4  chr1  rs143225517  T   C  751756  0.853  0.99681  0.0164  0.8431

>>> df['N'] = 35476+46839
>>> import numpy as np
>>> from scipy import stats
>>> df['Z'] = -stats.norm.ppf(df['PVALUE'].values*0.5)*np.sign(np.log(df['OR'].values))
>>> df['CHR'] = pd.to_numeric(df['CHR'].str.replace('chr', ''), errors='coerce')
>>> df.head()
    CHR          SNP A1  A2      bp   INFO       OR      SE  PVALUE      N         Z
0     1    rs4951859  C   G  729679  0.631  0.97853  0.0173  0.2083  82315 -1.258254
1     1  rs142557973  T   C  731718  0.665  1.01949  0.0198  0.3298  82315  0.974517
2     1  rs141242758  T   C  734349  0.666  1.02071  0.0200  0.3055  82315  1.024710
3     1   rs79010578  A   T  736289  0.649  0.98748  0.0193  0.5132  82315 -0.653863
4     1  rs143225517  T   C  751756  0.853  0.99681  0.0164  0.8431  82315 -0.197930
>>> df = df[~df['CHR'].isnull()].copy()
>>> df['CHR']=df['CHR'].astype(int)
>>> df.to_csv('PGC_SCZ_2014.csv', index=False, sep='\t')
```

Demo

- Download GWAS data for PGC Schizophrenia GWAS, 2014
  - https://www.med.unc.edu/pgc/results-and-downloads,

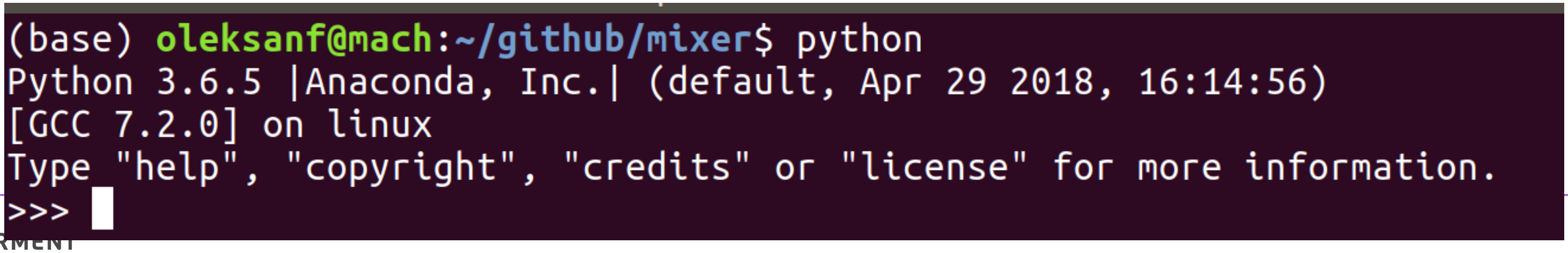
  Search for SCZ2, then “Download full SNP results”

- Download and install Anaconda (Python distribution)
  - https://www.anaconda.com/distribution/

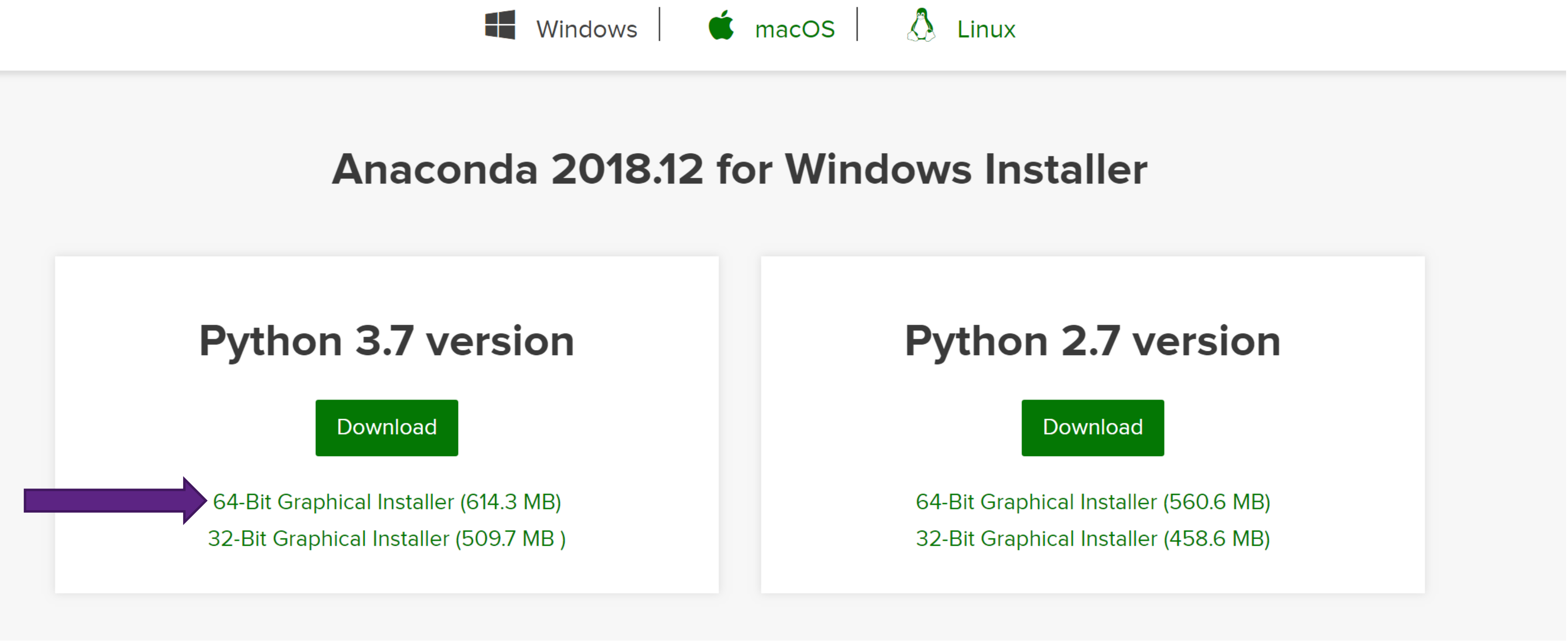
  (!) Ba careful about Mac | Windows | Linux – choose your system

  (!) Choose 64 bit. DO NOT choose 32 bit.

  Recommended download: Python 3.7 (but Python 2.7 is also OK)

- 

```
(base) oleksanf@mach:~/github/mixer$ python
Python 3.6.5 |Anaconda, Inc.| (default, Apr 29 2018, 16:14:56)
[GCC 7.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> 
```



Anaconda 2018.12 for Windows Installer

Python 3.7 version

64-Bit Graphical Installer (614.3 MB)

32-Bit Graphical Installer (509.7 MB)

Python 2.7 version

64-Bit Graphical Installer (560.6 MB)

32-Bit Graphical Installer (458.6 MB)

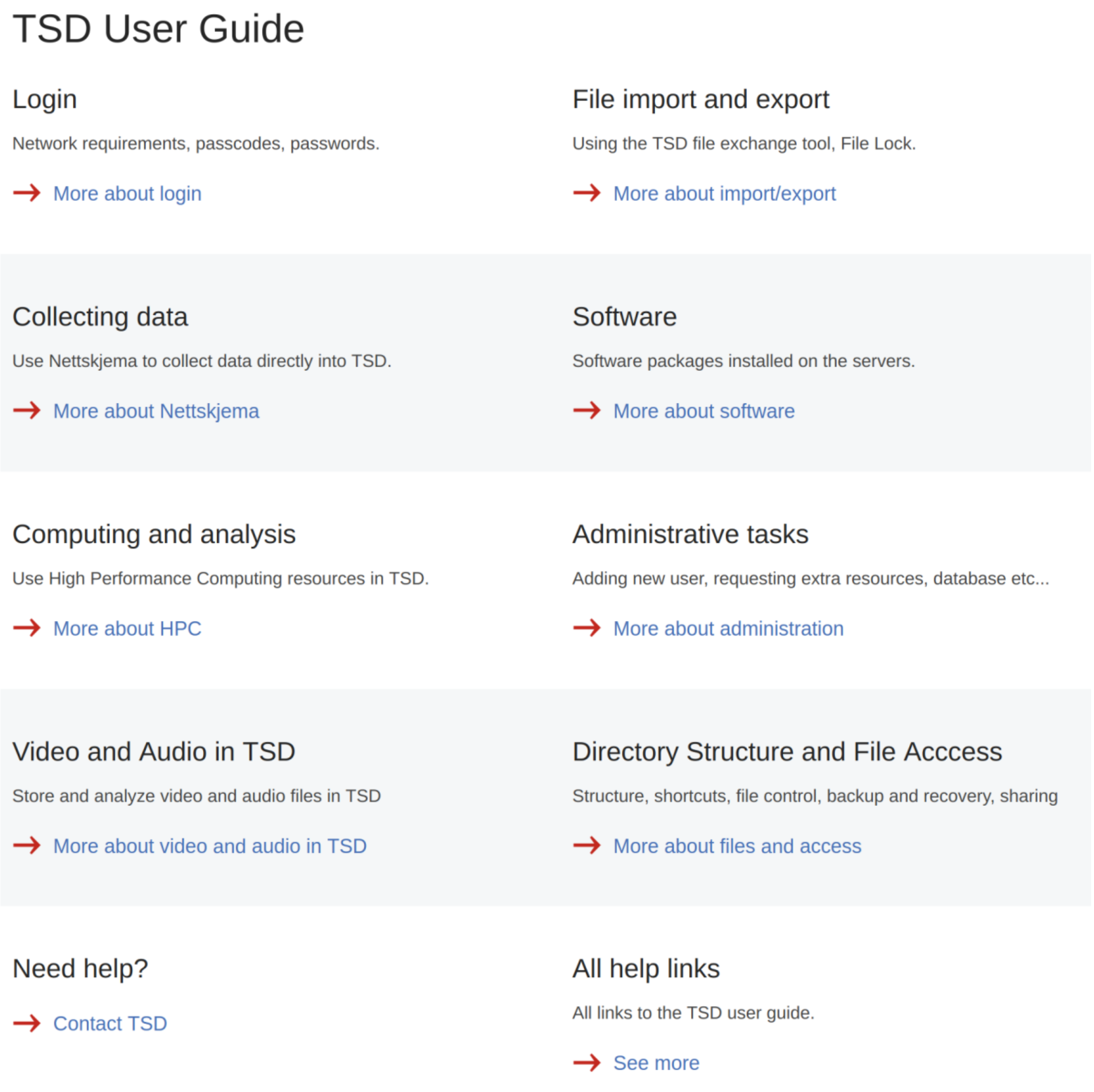
Links

- Consortia-specific GWAS locations
  - PGC - https://www.med.unc.edu/pgc/results-and-downloads
  - SSGAC - https://www.thessgac.org/data
  - CTG - https://ctg.cncr.nl/software/summary_statistics
  - IBD Genetics - https://www.ibdgenetics.org/downloads.html
  - B. Neale Lab - http://www.nealelab.is/uk-biobank
  - GIANT: Genetic Investigation of ANthropometric Traits - https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

- GWAS catalog: https://www.ebi.ac.uk/gwas/
- LDhub: http://ldsc.broadinstitute.org/ldhub/
- UK Biobank: https://biobank.ctsu.ox.ac.uk/showcase/

TSD introduction

- https://www.uio.no/english/services/it/research/sensitive-data/use-tsd/
- Good documentation!



TSD User Guide

Login

Network requirements, passcodes, passwords.

→ More about login

File import and export

Using the TSD file exchange tool, File Lock.

→ More about import/export

Collecting data

Use Nettskjema to collect data directly into TSD.

→ More about Nettskjema

Software

Software packages installed on the servers.

→ More about software

Computing and analysis

Use High Performance Computing resources in TSD.

→ More about HPC

Administrative tasks

Adding new user, requesting extra resources, database etc...

→ More about administration

Video and Audio in TSD

Store and analyze video and audio files in TSD

→ More about video and audio in TSD

Directory Structure and File Acccess

Structure, shortcuts, file control, backup and recovery, sharing

→ More about files and access

Need help?

→ Contact TSD

All help links

All links to the TSD user guide.

→ See more

TSD Demo

- (3) How to login to Linux machine
- (4) How to login to Windows machine
- (5) How to import data to TSD using file uploader

- (1) How to apply for TSD access for an existing project
- (2) How to reset TSD password (https://selfservice.tsd.usit.no/)

Advanced topics:

- How to extend your TSD project (request disk space, CPU hours, larger login nodes for your project)
- How to import large volumes of data to TSD

P.S. Gene expression data

- https://www.ebi.ac.uk/gxa/home
- http://portal.brain-map.org/
- https://www.proteinatlas.org/humanproteome/tissue/brain
- https://gtexportal.org/home/
- http://hbatlas.org/

NORMENT
Norwegian Centre for
Mental Disorders Research

NORMENT
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