UK Biobank data access and software for genome-wide association analysis

Hui Guo

Centre for Biostatistics

16 Nov 2018

How to access the data?

- Useful links from Research Infrastructure
  - UK Biobank data full release

http://ri.itservices.manchester.ac.uk/hosted-data-sets/ukbiobank/

  - UK Biobank Helper Modulefile

http://ri.itservices.manchester.ac.uk/csf-apps/software/applications/ukbiobank/

- To access the genotype data, you need to be
  - an approved applicant of UK Biobank
  - a user of the University's CSF/iCSF/DPSF
  - a member of the University's dataset-ukbiobank-full Unix group.

- Phenotype data are received directly from UK Biobank, according to different applications.

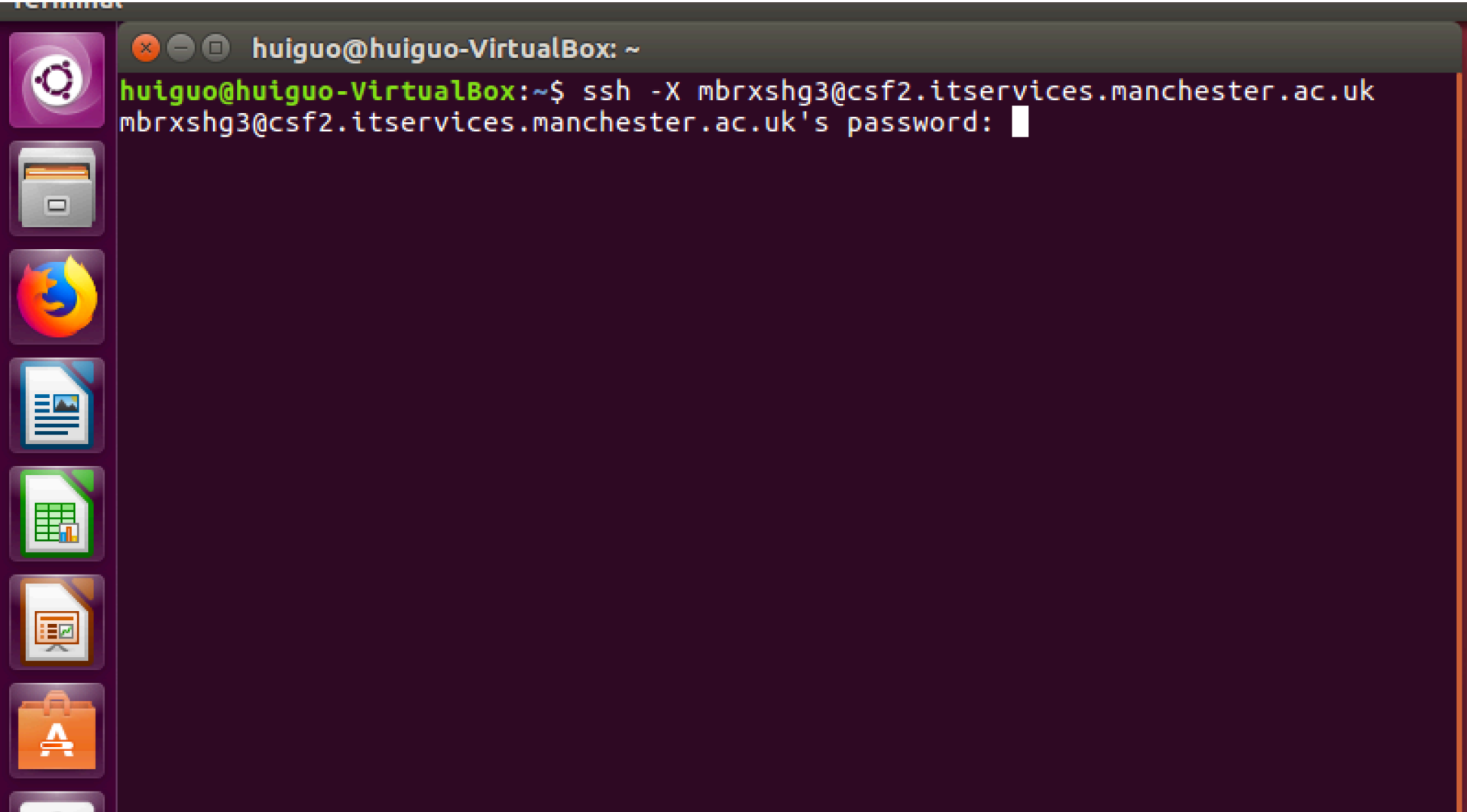
How big is genotype data ?

- ~ 0.5 million individuals
- ~ 92.7 million SNPs (genotyped + imputed)
- Size of compressed files:

  genotyped: 4TB

  imputed: 2TB

Where are the data ?



How to access the data ?

`cd /mnt/data-sets/ukbiobank/full-release`

- Imputed data:

  EGAD10001474/ukb_imp_chrN_v3.bgen

- Genotyped data (PLINK format):

  EGAD10001497/ukb_snp_chrN_v2.bim

  EGAD10001497/ukb_cal_chrN_v2.bed

- List of all files and description:

  Filelist.2018.txt

  Ukb_genetic_file_description.txt

- Useful document from UK Biobank:

http://www.ukbiobank.ac.uk/wp-content/uploads/2018/03/UKB-Genotyping-and-Imputation-Data-Release-FAQ-v3-2.pdf

Data	UKB File Naming	N	GB
Genotyping Data			
Genotyped SNP index *	ukb_snp_chrN_v2.bim	1-26	0.03
Calls	ukb_cal_chrN_v2.bed	1-26	92
Confidences	ukb_con_chrN_v2.txt	1-26	2900
Intensities	ukb_int_chrN_v2.bin	1-26	2900
CNV B-allele-freq	ukb_baf_chrN_v2.txt	1-26	1500
CNV log ratio	ukb_l2r_chrN_v2.txt	1-26	2300
Sample QC	ukb_sqc_v2.txt	n/a	0.3
HLA	ukb_hla_v2.txt	n/a	0.4
Imputation Data			
Imputation SNP index *	ukb_imp_chrN_v3.bgen.bgi	1-24	4
Imputed data	ukb_imp_chrN_v3.bgen	1-24	2100
Minor-allele freq + info scores *	ukb_mfi_chrN_v3.txt	1-24	4

Genome-wide association analysis (GWAS)

- SNPTEST v2.5.4-beta3

https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#introduction

- phenotype: categorical, single or multiple quantitative
- tests: Bayesian, frequentist
- relatively slow

Genome-wide association analysis (GWAS)

- PLINK 2

https://www.cog-genomics.org/plink/2.0/

- phenotype: single or multiple, binary or quantitative
- tests: frequentist
- very fast

Genome-wide association analysis (GWAS)

- BOLT-LMM v2.3.2

https://data.broadinstitute.org/alkesgroup/BOLT-LMM/

- phenotype: single quantitative or (reasonably) balanced binary
- sample size: > 5000
- accounting for relatedness
- tests: Bayesian

  BOLT-LMM: uses Gaussian mixture-model for both infinitesimal and non-infinitesimal genetic architectures, higher statistical power than traditional infinitesimal model BOLT-LMM-inf
- very fast

Example: GWAS on BOLT-LMM

```
Module load tools/env/ukbiobank-full-release
./apps/BOLT-LMM_v2.3.2/bolt \
--bed=$UKBB_GENOTYPED_DIR/ukb_cal_chr{1:22}_v2.bed \
--bim=$UKBB_GENOTYPED_DIR/ukb_snp_chr{1:22}_v2.bim \
--fam=name of family file \ # 6 columns
--remove=individuals_to_be_removed.txt \
--phenoFile=name of phenotype file \
--phenoCol=name of the column/phenotype in phenoFile \
--covarFile=name of covariate file \ # could be the same as phenoFile
--covarCol=name of binary covariate \
--qCovarCol=name of quantitative covariate \
--LDscoresFile=LDSCORE.1000G_EUR.tab.gz \ # provided in BOLT-LMM
--geneticMapFile=genetic_map_hg19_withX.txt.gz \ # provided in BOLT-LMM
--bgenFile=$UKBB_IMPUTATION_DIR/ukb_imp_chr{1:22}_v3.bgen \
--bgenMinMAF=threshold of minor allele frequency \
--bgenMinINFO=threshold of imputation INFO score \
--sampleFile=sample file generated from imputed data \
--statsFile=file name of output for all SNPs \
--numThreads=number of threads \
```

Genome-wide association analysis (GWAS)

- SAIGE

https://github.com/weizhouUMICH/SAIGE/

- phenotype: binary
- accounting for both case-control imbalance and relatedness
- slower than BOLT-LMM
- assuming infinitesimal model