Phenotypes Prediction with Deep Learning

Sapienza

Buzatu Ionelia

April 2019

Acknowledgements

I want to thank...

Contents

List of Figures

Chapter 1

Introduction

UK Biobank is a national and international health resource with unparalleled research opportunities, open to all bona fide health researchers. UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. Scientists, please ensure you read the background materials before registering. [1]

Genotyping has been undertaken on all 500,000 participants. In addition to information collected during the baseline assessment(such as the eye measures and saliva samples), 100,000 UK Biobank participants have worn a 24-hour activity monitor for a week, and 20,000 have undertaken repeat measures. A programme of online questionnaires is being rolled out (diet, cognitive function, work history and digestive health) and UK Biobank has embarked on a major study to scan (image) 100,000 participants (brain, heart, abdomen, bones carotid artery). UK Biobank is linking to a wide range of electronic health records (cancer, death, hospital episodes, general practice).

The following chapters deonstrates the application of deep learning on phenotypic predictions exploting all the single nucleotide polymorphisms (SNPs) of each participant.

Chapter 2

The Data

2.1 Files Description

The `ukb8627.tab`[21GB] consists of the phenotype data of all the participants that have been taking part since the UK biobank started to collect their data. Over the years 502638 participants data has been collected but Y withdraw consent to participate to the project. Currently the total genomes are 486383. The 'f.eid' is the id of the patient and 10951 columns each rapresant a Field. data showcase, also described in the readme.

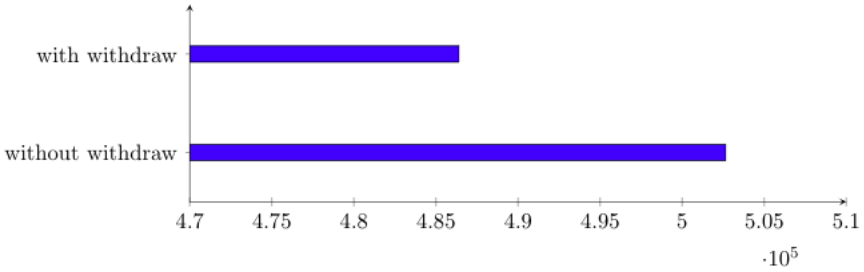

Figure 2.1: Participants

File names	
Calls BIM	`ukb_snp_chrN_v2.bim`
Calls FAM	`ukbA_cal_v2_sP.fam`
Relatedness	`ukbA_rel_sP.txt`
Imputation BGEN	`ukb_imp_chrN_v3.bgen`
Imputation BGI	`ukb_imp_chrN_v3.bgen.bgi`
Imputation MAF+info	`ukb_mfi_chrN_v3.txt`
Imputation sample chr1-22	`ukbA_imp_chrN_v3_sP.sample`
Imputation sample chrX	`ukbA_imp_chrX_v3_sP.sample`
Imputation sample chrXY	`ukbA_imp_chrXY_v3_sP.sample`
Haplotypes BGEN	`ukb_hap_chrN_v2.bgen`
Haplotypes BGI	`ukb_hbg_chrN_v2.bgi`
Intensity	`ukb_int_chrN_v2.bin`
Confidences	`ukb_con_chrN_v2.txt`

The `ukb_category.txt`[7.1K] describes categories ids with hierarchy.

The `ukb_field.txt`[234K] Description of fields ids with parent category. Field with biological similarities are grouped under a specific category. The below table shows all the fieald under the category CANCER

Field ID	**Description**
40021	Cancer record origin
40009	Reported occurrences of cancer
40005	Date of cancer diagnosis
40008	Age at cancer diagnosis
40006	Type of cancer: ICD10
40016	Type of cancer: ICD10 addendum
40013	Type of cancer: ICD9
40011	Histology of cancer tumour
40017	Type of cancer: ICD9 addendum
40012	Behaviour of cancer tumour
40019	Cancer report format

The `genetic data` consists of:
- log2ratios
- b-allele-frequencies
These files contain the B Allele Frequency baf and Log2Ratio log2r transformed intensitiy values for performing CNV calling. There is a separate file for baf and log2r per chromosome. These are plaintext files with space separated columns. The rows corresond to markers ordered as the calls BIM file and the columns correspond to samples ordered as the calls FAM file.

a) `Calls`
The genotype data calls are in binary PLINK format (.bed, .bim, .fam) [3] The BIM file determines the order of markers in the calls and all of the other genotype data sets. The SNP id is the rsid where it is available or the Affymetrix SNP id otherwise. The positions are in GRCh37 coordinates. The FAM file contains the id of the participants and determines the order of samples in the calls and all of the other genotype data sets. NOTE: that the fam is the same for all beds, so take as reference `ukb_cal_chr1_v2.fam`

b) indIDS[3.7M] : Ids of the individuals as found in the FAM file.

c) `ukb_snp_bim`
Bim files for the ‘calls‘ data above

d) `confidencies`
These files contain the Affymetrix ’confidence’ that a genotype belongs to the call cluster. This is a plaintext file with space spearated columns. Values are in the range 0-1 with 0 being most confident. Missing values are represented by -1. The order of markers and Samples are given by the BIM and FAM files. e) `haplotypes`

Phased haplotypes in BGEN format. The sample file lists the order of the samples in the .bgen files.

f) `The imputations`
contain imputed genotype of the individuals. The imputed genotype calls are in BGEN v1.2 format (.bgen, .sample, .bgi). The sample file lists the order of the samples in the .bgen files. The sample file includes the 'Sex' field for every sample (corresponding to 'Inferred.Gender' in the SampleQC file). The list of variants in the files can be found with bgenix [2]. The 1st column of the marker list file is alternate ids which is a unique identifier for each marker. For markers in the genotype data set alternate ids is the genotype marker id (rs id in SNPQC file). The second column is rsid or the reference panel marker id, it is not guaranteed to be unique. The alleles in the imputation are aligned with REF/ALT, first allele is the ref allele on the fwd strand.

g) `intensities`
contains A/B Intensity values measured by Affymetrix. Two intensity values A/B for each genotype (marker, indivudal) pair each represented as a 4-byte float. The set of A,B values for each marker are ordered consecutively by sample (analagous to a matrix with rows=SNPs and columns=Samples) e.g. SNP1SAMPLE1A SNP1SAMPLE1B SNP1SAMPLE2A SNP1SAMPLE2B ... Missing pairs of intensities are represented by -1 -1. The order of the markers and Samples are given by the BIM and FAM files with the calls. Affymetrix transform the A,B values into 'contrast' and 'strength' for their calling algorithm. The values are: contrast (X) = log2(A/B) strength (Y) = log2(AB)/2

h) `The relatedness`
file lists the pairs of individuals related up to the third degree in the data set. It is a plaintext file with space separated columns.

ID1	string	Sample id for individual 1 in related pair.
ID2	string	Sample id for individual 2 in related pair.
HetHet	numeric	Fraction in common of heterozygous genotype.
IBS0	numeric	Fraction sharing zero alleles (output from KING software).
Kinship	numeric	Estimate of the kinship coefficient output from KING software.

2.2 Data Processing

Chapter 3

Previous Work

Chapter 4

Good and Bad Questions

Chapter 5

Conclusion

Bibliography

[1] *About UK Biobank*. URL: https://www.ukbiobank.ac.uk/.

[2] *Bgenix - Variants*. URL: https://bitbucket.org/gavinband/bgen/wiki/bgenix.

[3] *Calls Formats and PLINK*. URL: https://www.cog-genomics.org/plink/1.9/formats.