Generate mock phenotype data for 1000 Genomes Phase 3 samples

By Bruno Ambrozio (https://about.me/bambrozio)

> Notebook based on the script "make_mock_phenotypes_script (https://app.terra.bio/#workspaces/broad-t2d-dev/Running_GWAS_in_Terra)", maintained by Terra (https://app.terra.bio/). Kuddos to them!

In [1]:

```
library(data.table)
set.seed(5)
```

In [2]:

```
# load the data from 1kg ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
ped <- fread("../integrated_call_samples_v3.20130502.ALL.panel",
             data.table = F,
             stringsAsFactors = F,
             fill = T)
```

In [3]:

```
head(ped[-1,])
```

	V1	V2	V3	V4	V5	V6
2	HG00096	GBR	EUR	male	NA	NA
3	HG00097	GBR	EUR	female	NA	NA
4	HG00099	GBR	EUR	female	NA	NA
5	HG00100	GBR	EUR	female	NA	NA
6	HG00101	GBR	EUR	male	NA	NA
7	HG00102	GBR	EUR	female	NA	NA

In [4]:

```
# fix the names
drops <- c("V5", "V6")
ped = ped[ , !(names(ped) %in% drops)]
names(ped) <- c("sample", "subpopulation", "superpopulation", "sex")
ped <- ped[-1,]
row.names(ped) <- 1:nrow(ped)
head(ped)
```

sample	subpopulation	superpopulation	sex
HG00096	GBR	EUR	male
HG00097	GBR	EUR	female
HG00099	GBR	EUR	female
HG00100	GBR	EUR	female
HG00101	GBR	EUR	male
HG00102	GBR	EUR	female

In [5]:

```
# What ancestries do we have?
unique(ped$superpopulation)
```

'EUR' 'EAS' 'AMR' 'SAS' 'AFR'

In [6]:

```
# how many people?
n <- nrow(ped)
n
```

2504

In [7]:

```
# add a age mock column
ped$age <- floor(runif(n, min = 30, max = 99))
head(ped)
```

sample	subpopulation	superpopulation	sex	age
HG00096	GBR	EUR	male	43
HG00097	GBR	EUR	female	77
HG00099	GBR	EUR	female	93
HG00100	GBR	EUR	female	49
HG00101	GBR	EUR	male	37
HG00102	GBR	EUR	female	78

In [8]:

```
# add bmi (quantitative phenotype) and t2d (case/control phenotype) mock columns
ped$bmi <- ped$t2d <- 0
head(ped)
```

sample	subpopulation	superpopulation	sex	age	t2d	bmi
HG00096	GBR	EUR	male	43	0	0
HG00097	GBR	EUR	female	77	0	0
HG00099	GBR	EUR	female	93	0	0
HG00100	GBR	EUR	female	49	0	0
HG00101	GBR	EUR	male	37	0	0
HG00102	GBR	EUR	female	78	0	0

For each ancestry, add t2d columns based on population prevalence. Grounded on:

Source: Spanakis, E., & Golden, S. (2013). Race/Ethnic Difference in Diabetes and Diabetic Complications. Current Diabetes Reports, 13(6), 814-823. doi: 10.1007/s11892-013-0421-9 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3830901/#S3title)

Table 1

Age-adjusted prevalence of diagnosed diabetes mellitus in the United States by race/ethnicity in adults ≥20 years of age [5]

Race/ethnic group	Age-adjusted prevalence (%)
Non Hispanic Whites	7.1
Asian- Americans	8.4
Hispanic-Americans overall	11.8
Non Hispanic Blacks	12.6
Alaska Natives	5.5
Native Americans	33

Table 1 is reproduced from Centers for Disease Control and Prevention National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States. Atlanta: U.S Department of Health and Human Services. 2011. Available from: http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf.

Souce: Ma, R., & Chan, J. (2013). Type 2 diabetes in East Asians: similarities and differences with populations in Europe and the United States. Annals Of The New York Academy Of Sciences, 1281(1), 64-91. doi: 10.1111/nyas.12098 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3708105/#__sec2title)

Table 1

Comparison of prevalence of diabetes, age distribution, and proportion undiagnosed in East Asian countries compared to the United States and Europe

Country	Diabetes prevalence in 2011 (%)	Estimated number of people affected in 2011	Proportion of DM subjects aged 20–39 in 2011 (%)	Estimated proportion of DM cases undiagnosed in 2011 (%)	IGT prevalence in 2011 (%)	Mean diabetes-related expenditure per person with DM (USD)
China	9.29	90,045,980	15.1	56.9	2.41[a]	194
Hong Kong	9.38	525,390	7.4	46.7	14.87	2,059
Macau	7.49	32,710	10.0	46.7	5.69	480
Taiwan	9.59	1,664,540	9.59	46.7	11.61	1,314
Mongolia	6.74	117,460	43.6	56.9	7.86	107
Japan	11.2	10,674,320	5.9	46.7	14.3	3,266
Dem. People's Republic of Korea	9.08	1,507,500	13.1	63.0	10.83	17
Republic of Korea	8.8	3,186,390	10.0	46.7	13.45	1,615
USA	10.94	23,721,760	13.9	27.7	11.97	8,468
Canada	10.80	2,716,140	6.0	27.7	12.2	5,106
United Kingdom	6.84	3,063,910	6.95	36.6	9.19	4,267
Australia	8.10	1,292,090	8.46	46.7	9.94	4,878

NOTE: the data source is based on projections from epidemiological surveys. Data source: Diabetes Atlas, Fifth edition, 2011. International Diabetes Federation.[1]

[a]IGT prevalence figures reported in the above reference is markedly different to that reported in a recent nationwide study, which reported age-standardized prevalence of diabetes of 9.7%, and 15.5% for prediabetes, including 11.9% with IGT.

In [9]:

```
t2d.prev <- data.frame(
  anc = c("EUR", "AMR", "AFR", "EAS", "SAS"),
  prev = c(7.1, 33, 12.6, 8.95, 17),
  case_bmi = c(20, 35, 30, 15, 10),
  control_bmi = c(25, 30, 30, 17, 8),
  stringsAsFactors = F
)
t2d.prev
```

anc	prev	case_bmi	control_bmi
EUR	7.10	20	25
AMR	33.00	35	30
AFR	12.60	30	30
EAS	8.95	15	17
SAS	17.00	10	8

In [10]:

```
new.ped <- list()
for (anc in unique(ped$superpopulation)){
  cur.ped <- ped[ped$superpopulation == anc,]
  nn <- nrow(cur.ped)
  ncase <- floor(nn*t2d.prev[t2d.prev$anc == anc, "prev"]/100)
  cur.ped[sample(1:nrow(cur.ped), ncase), "t2d"] <- 1
  cur.ped[cur.ped$t2d == 1, "bmi"] <- rnorm(nrow(cur.ped[cur.ped$t2d == 1,]),
                                            mean = t2d.prev[t2d.prev$anc == anc, "ca
                                            sd = 2)
  cur.ped[cur.ped$t2d == 0, "bmi"] <- rnorm(nrow(cur.ped[cur.ped$t2d == 0,]),
                                            mean = t2d.prev[t2d.prev$anc == anc, "co
                                            sd = 2)
  new.ped[[anc]] <- cur.ped
}

ped <- do.call(rbind, new.ped)
ped$sex <- ifelse(ped$sex == "male", "M", "F")

head(ped)
```

	sample	subpopulation	superpopulation	sex	age	t2d	bmi
EUR.1	HG00096	GBR	EUR	M	43	0	25.42418
EUR.2	HG00097	GBR	EUR	F	77	0	22.79394
EUR.3	HG00099	GBR	EUR	F	93	1	24.48213
EUR.4	HG00100	GBR	EUR	F	49	0	25.46661
EUR.5	HG00101	GBR	EUR	M	37	0	26.27279
EUR.6	HG00102	GBR	EUR	F	78	0	24.76994

In [23]:

```
fwrite(ped,
       file = paste0('1kg_phenotype_mock_', as.Date(Sys.time(), "%d/%m/%Y"), ".tsv")
       sep = "\t", row.names = F)
```