Data Preparations in Haplin

DATA PREPARATION

Three main ways to prepare data:

1. Native Haplin data format (Not much used)
2. Convert from ped to Haplin using `pedToHaplin` (not much used)
3. Convert from ped to Julia format, let Haplin read from ff data

The Julia format was developed (by Julia Romanowska, using the `ff` package), to replace the GenABEL data format.

More about data formats:

`https://folk.uib.no/gjessing/genetics/software/haplin/docu/data_format/`

DATA PREPARATION, .PED AND .MAP

Produce ped and map files, for instance from PLINK:

Input files: `pres.bed`, `pres.bim`, `pres.fam`

```
plink --bfile data/pres --alleleACGT --recode --out data/pres
```

pres.ped

```
18 1 3 2 1 1 G G T T A A
18 2 0 0 2 0 G G T A A A
18 3 0 0 1 0 G G T A A A
19 1 3 2 1 0 G G T A A A
19 2 0 0 2 0 G G T A A T
19 3 0 0 1 1 G G A A A A
```

pres.map

```
chrom    snp    pos
1        rs1    0
1        rs3    0
1        rs5    0
```

Read raw data into R:

```
tmp <- genDataRead(file.in = "data/pres.ped",
 file.out = "pres", dir.out = "data", format = "ped")
```

Quick info about data object:

```
tmp
```

Info about tmp

```
This is raw genetic data read in through genDataRead.
It contains the following parts:
   cov.data, gen.data, aux

with following dimensions:
  - covariate variables =  id.fam, id.c, id.f, id.m, sex, cc
      (total  6  covariate variables),
  - number of markers =  429 ,
  - number of data lines =  1659
```

Data preparation, Family and covariate information

```
showPheno(tmp)
```

Info about family/covars

```
     id.fam id.c id.f id.m sex cc
[1,] "1"    "1"  "3"  "2"  "2" "1"
[2,] "1"    "2"  "0"  "0"  "2" "0"
[3,] "1"    "3"  "0"  "0"  "1" "1"
[4,] "2"    "1"  "3"  "2"  "1" "0"
[5,] "2"    "2"  "0"  "0"  "2" "0"
```

```
nindiv(tmp)
```

```
[1] 1659
```

```
nfam(tmp)
```

```
[1] 550
```

Pre-process data into Julia format:

```
pres.data <- genDataPreprocess(tmp, map.file = "data/pres.map",
 dir.out = "data", ncpu = 3)
```

Quick info about data object:

```
pres.data
```

Info about pres.data

```
This is preprocessed data, ready for haplin analysis.

It contains the following parts:
   cov.data, gen.data, aux

with following dimensions:
- number of covariate variables =  10
- number of markers =  429
- number of individuals/families =  559
```

Data quality control

Haplin reports:

- The number of missing values at each SNP.
  (NOTE: this includes family members not genotyped.)
- Allele frequencies at each SNP.
- The number of families with Mendelian inconsistencies.
- A simple test for Hardy-Weinberg equilibrium.

However, there are no pre-cleaning tools in Haplin.

DATA QUALITY CONTROL

- Data should always be cleaned in advance.
  Use, for instance, PLINK.
- Nice hits should **always be checked "after the facts"**, since poor data quality can cause false positives.
- Haplin assumes Hardy-Weinberg equilibrium in its models.
  You can choose a multiplicative response model: `response = "mult"`
  - Less dependent on the HWE assumption
  - Faster
  - Easier to interpret
  - Good choice for initial screening

More on the Julia data format

```
vignette("B_Reading_data")
```