I/O Suite, VCF (1000 Genome) and HapMap

Hin-Tak Leung

April 13, 2013

Contents

1 Introduction

snpMatrix version 1.19.0.18 (April 2013) introduces the functionality of reading VCF files; this document is the primary documentation as well as the test suite for correct operation using data from the 1000 Genome Project and the HapMap Project. It also serves as tutorial for data manipulation.

This vignette is explicitly checked to work on 32-bit windows against large files. The Chromosome 6 data file from 1000 Genome Project is 9GB in size. We extract the MHC region as follows:

```
> library(snpMatrix)
```

```
> in.1000g <- read.snps.vcf("1000g-release/20110521/ALL.chr6.phase1_release_v3.20101123.snps_i
+     from = 2.5e+07, to = 3.2e+07)
```

```
       Scanning data across chromosome 6
Converting 1092 samples at 106880 loci out of 1092 x 2424425
```

```
> proc.time()
```

```
    user   system  elapsed
1103.001   40.374 1389.035
```

```
> summary(in.1000g$snp.support)
```

```
   Loci.ID             Chrom              Position          AlleleA
 Length:106880      Length:106880      Min.   :25000000   Length:106880
 Class :character   Class :character   1st Qu.:27051192   Class :character
 Mode  :character   Mode  :character   Median :29148234   Mode  :character
                                       Mean   :28787088
                                       3rd Qu.:30591879
                                       Max.   :31997129

   AlleleB             Alleles
 Length:106880      G/A    :20846
 Class :character   C/T    :20542
 Mode  :character   A/G    :14248
                    T/C    :14132
                    C/A    : 4670
                    G/T    : 4616
                    (Other):27826
```

1.1 Ethnic Composition of 1000G vs HapMap

We have some supplementary tables for ethnic groups in 1000G and HapMap, as well as their (anonymized) sample names. The tables are somewhat larger than the currated and released genotypes, but provides some ideas about the overlaps of the samples used in each ethnic group.

```
> hapmap.info <- read.table("hapmap.ncbi.nlm.nih.gov/downloads/samples_individuals/relationshi
+     header = TRUE, sep = "\t")
> rownames(hapmap.info) <- hapmap.info$IID

> samples.1000g <- rownames(in.1000g$snp.data)
> length(hapmap.info[, "population"])
```

```
[1] 1301
```

```
> summary(hapmap.info[, "population"])
```

```
ASW CEU CHB CHD GIH JPT LWK MEX MKK TSI YRI
 90 180  90 100 100  91 100  90 180 100 180
```

```
> summary(hapmap.info[samples.1000g, "population"])
```

```
 ASW  CEU  CHB  CHD  GIH  JPT  LWK  MEX  MKK  TSI  YRI NA's
  53   82   85    0    0   80   97   56    0   98   88  453
```

```
> info.1000g <- read.table("1000g-release/20110521/supporting/phase1_samples_integrated_2010112
+     header = TRUE, sep = "\t", fill = TRUE)
> rownames(info.1000g) <- info.1000g$Individual.ID
> length(info.1000g[, "Population"])
```

```
[1] 1398
```

```
> summary(info.1000g[, "Population"])
```

```
ASW CEU CHB CHS CLM FIN GBR IBS JPT LWK MXL PUR TSI YRI
102 137  97 150  93  93  91  21  89  99  99  87  98 142
```

```
> summary(info.1000g[samples.1000g, "Population"])
```

```
ASW CEU CHB CHS CLM FIN GBR IBS JPT LWK MXL PUR TSI YRI
 61  85  97 100  60  93  89  14  89  97  66  55  98  88
```

The phase 1 1000 G data consists of 1092 samples (see first section), HapMap 2010-08 phaseII+III consists of 209 YRI samples (87 in 1KG), 174 CEU samples (81 in 1KG), 139 CHB samples (91 in 1KG), 116 JPT samples (85 in 1KG).

It would appear that "relationships_w_pops_121708.txt" from HapMap is significantly out-dated.

Here is a description for each of the ethnic groups in the 1000 Genome Project as on 20111108, up to just over 2800 individuals, from the 1092 in 20110521:

Count	Group	Description
102	ACB	African Caribbean in Barbados
112	ASW	HapMap African ancestry individuals from SW US
100	CDX	Chinese Dai in Xishuangbanna, China
186	CEU	CEPH individuals
162	CHB	(CHB) Han Chinese in Beijing
129	CHD	Chinese in metropolitan Denver, CO
153	CHS	(CHB) Han Chinese South
111	CLM	Colombian in Medellin, Colombia
101	FIN	HapMap Finnish individuals from Finland
104	GBR	British individuals from England and Scotland (GBR)
117	GIH	HapMap Gujarati India individuals from Texas
150	IBS	Iberian populations in Spain
131	JPT	JPT Japanese individuals
122	KHV	Kinh in Ho Chi minh City, Vietnam
122	LWK	(LWK) Luhya individuals
232	MKK	HapMap Maasai individuals from Kenya
108	MXL	HapMap Mexican individuals from LA California
105	PEL	Peruvian in Lima, Peru
123	PUR	Puerto Rican in Puerto Rico
117	TSI	Toscan individuals
229	YRI	(YRI) Yoruba individuals

2 1000 Genome vs HapMap YRI (Africans)

The number of SNPs and samples in common are:

```
> in.hapmap <- read.HapMap.data("hapmap-genotypes/2010-08_phaseII+III/forward/genotypes_chr6_YR
```

```
Reading 209 samples
current line [0] : rs4097465 A/C chr6 3...
current line [20000] : rs9368446 A/G chr6 1...
current line [40000] : rs2480000 C/T chr6 2...
current line [60000] : rs6909790 A/G chr6 3...
current line [80000] : rs1680902 C/G chr6 4...
current line [100000] : rs6917528 A/C chr6 5...
current line [120000] : rs12199026 C/T chr6 ...
current line [140000] : rs9344532 A/G chr6 8...
current line [160000] : rs910423 A/G chr6 99...
current line [180000] : rs12211182 C/G chr6 ...
current line [200000] : rs1338841 A/G chr6 1...
current line [220000] : rs7746658 C/T chr6 1...
current line [240000] : rs9383986 A/G chr6 1...
current line [260000] : rs9459114 C/T chr6 1...
last line [270474] : rs4599694 A/G chr6 1...
```

```
EOF reached after 270475 snps
```

```
> snps.in.common <- rownames(in.hapmap$snp.support)[rownames(in.hapmap$snp.support) %in%
+     rownames(in.1000g$snp.support)]
> length(snps.in.common)
```

```
[1] 5336
```

```
> snp.support.1000g <- droplevels(in.1000g$snp.support[snps.in.common,
+     ])
> snp.support.hapmap <- in.hapmap$snp.support[snps.in.common, ]

> summary(snp.support.hapmap$Assignment)
```

```
 ./.  A/C  A/G  A/T  C/G  C/T  G/T
   5  441 1863  342  465 1785  435
```

```
> samples.in.common <- rownames(in.hapmap$snp.data)[rownames(in.hapmap$snp.data) %in%
+     rownames(in.1000g$snp.data)]
> length(samples.in.common)
```

```
[1] 87
```

```
> snp.data.1000g <- in.1000g$snp.data[samples.in.common, snps.in.common]
> snp.data.hapmap <- in.hapmap$snp.data[samples.in.common, snps.in.common]
```

We also note that a few "SNPs" in common are not assigned in the HapMap input. The reason is quite plainly obvious — from the 1KG end, we learn that those variants are not single nucleotide changes:

```
> snp.support.1000g[snp.support.hapmap$Assignment == "./.", ]
```

```
           Loci.ID Chrom Position AlleleA AlleleB Alleles
rs5875188 rs5875188     6 29248564       C      CT    C/CT
rs5875197 rs5875197     6 29574595       A      AC    A/AC
rs5875231 rs5875231     6 30070538       G      GT    G/GT
rs5875236 rs5875236     6 30118544    GAGA       G  GAGA/G
rs5875254 rs5875254     6 30386624       A      AG    A/AG
```

```
> subset(snp.support.hapmap, Assignment == "./.")
```

```
          dbSNPalleles Assignment Chromosome Position Strand
rs5875188          -/T        ./.       chr6 29356543      +
rs5875197          -/C        ./.       chr6 29682575      +
rs5875231          -/T        ./.       chr6 30178517      +
rs5875236        -/AGA        ./.       chr6 30226524      +
rs5875254          -/G        ./.       chr6 30494603      +
```

Here we learn something interesting: HapMap (dbSNPalleles) records those "SNPs" as insertions of 1 to 3 bases, while 1000 Genome's VCF record shows equivalent changes as an alternation between one base and a longer segment of 2-4 bases, of which the single base forms the first member. The latter is somewhat arbitrary as one can as well think of the changes as that between one (a different) base and a longe segment of whih the single base forms the *last* member. Also dbSNPalleles shows the alternatives alphabatically while 1000 Genome distiguishes the two alternatives as reference (possibly more common) vs alternate.

```
> summary(snp.support.1000g$Alleles)
```

```
  A/AC   A/AG    A/C    A/G    A/T    C/A   C/CT    C/G    C/T    G/A    G/C
     1      1    209    882    168    234      1    225    946    988    239
  G/GT    G/T GAGA/G    T/A    T/C    T/G
     1    212      1    174    833    221
```

To properly compare the two sets of data, we must cope with diferences in the two data sources in how we assigned "Allele A" (or the first allele) and "Allele B" (or the second allele).

We observe that HapMap always assigns alphabetically (see above), so we can determine whether to switch the alleles looking at whether the 1000 Genome assignment is in alphabetical order or not:

```
> to.flip <- (snp.support.1000g$AlleleA > snp.support.1000g$AlleleB)
> summary(to.flip)
```

```
   Mode   FALSE    TRUE    NA's
logical    2646    2690       0
```

```
> snp.data.1000g.flipped <- switch.alleles(snp.data.1000g, to.flip)
```

There are other ways of checking. For convenience, we use one of the non-public convenience routines:

```
> differences.of.two.platforms <- snpMatrix:::.snp.diff(snp.data.1000g.flipped,
+     snp.data.hapmap)
> summary(differences.of.two.platforms)
```

```
     common       common.called    diff.called       called.in.1     called.in.2
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.0000   Min.   : 2.00   Min.   :0
 1st Qu.:37.00   1st Qu.:37.00   1st Qu.: 0.0000   1st Qu.: 2.00   1st Qu.:0
 Median :38.00   Median :38.00   Median : 0.0000   Median :49.00   Median :0
 Mean   :58.72   Mean   :58.72   Mean   : 0.7894   Mean   :27.49   Mean   :0
 3rd Qu.:85.00   3rd Qu.:85.00   3rd Qu.: 0.0000   3rd Qu.:49.00   3rd Qu.:0
 Max.   :85.00   Max.   :85.00   Max.   :85.0000   Max.   :87.00   Max.   :0
```

```
> summary(as.factor(differences.of.two.platforms$diff.called))
```

```
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
4362  553  155   76   29   31   16   13    9    6    3    4    2    5    6    1
  16   18   19   20   21   22   23   24   25   26   29   30   31   32   33   35
   4    3    6    2    5    3    2    3    1    2    4    1    2    2    2    2
  37   38   50   53   59   66   69   72   73   74   76   82   83   84   85
   1    7    1    1    1    1    1    1    1    1    1    1    1    1    1
```

The two data set do not agree exactly; However, 82% (4362/5336) of SNPs do make the same calls on all the samples, This increases to 92% for at most one difference among the samples in common, to 95% for at most two differences in calls for samples. 5% of $\sim$ 87 samples is about 4. This observation is similar to and somewhat better than what we observe with SNPs called in Affymetrix vs Illumina for the 1958 Birth Cohort, which is 93% of SNPs shows call differences in fewer than 5% of samples[1].

We should have gotten rid of those 5 non-SNPs just to be proper, so let's repeat the procedure; the conclusion is the same and not affected by those.

```
> strange.snps <- rownames(snp.support.hapmap)[snp.support.hapmap$Assignment ==
+     "./."]
> strange.snps
```

```
[1] "rs5875188" "rs5875197" "rs5875231" "rs5875236" "rs5875254"
```

```
> snps.in.common.5fewer <- snps.in.common[!(snps.in.common %in%
+     strange.snps)]
> snp.data.1000g.flipped.5fewer <- snp.data.1000g.flipped[, snps.in.common.5fewer]
> snp.data.hapmap.5fewer <- snp.data.hapmap[, snps.in.common.5fewer]
> differences.of.two.platforms.5fewer <- snpMatrix:::.snp.diff(snp.data.1000g.flipped.5fewer,
+     snp.data.hapmap.5fewer)
> summary(differences.of.two.platforms.5fewer)
```

```
     common       common.called    diff.called       called.in.1     called.in.2
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.0000   Min.   : 2.00   Min.   :0
```

[1] See the "MHC Subset Preparation" vignette

```
 1st Qu.:37.00   1st Qu.:37.00   1st Qu.: 0.0000   1st Qu.: 2.00   1st Qu.:0
 Median :38.00   Median :38.00   Median : 0.0000   Median :49.00   Median :0
 Mean   :58.78   Mean   :58.78   Mean   : 0.7901   Mean   :27.43   Mean   :0
 3rd Qu.:85.00   3rd Qu.:85.00   3rd Qu.: 0.0000   3rd Qu.:49.00   3rd Qu.:0
 Max.   :85.00   Max.   :85.00   Max.   :85.0000   Max.   :59.00   Max.   :0
```

> summary(as.factor(differences.of.two.platforms.5fewer$diff.called))

```
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
4357  553  155   76   29   31   16   13    9    6    3    4    2    5    6    1
  16   18   19   20   21   22   23   24   25   26   29   30   31   32   33   35
   4    3    6    2    5    3    2    3    1    2    4    1    2    2    2    2
  37   38   50   53   59   66   69   72   73   74   76   82   83   84   85
   1    7    1    1    1    1    1    1    1    1    1    1    1    1    1
```

In the above we saw that there were many called in 1000G but not in Hapmap, but none called in Hapmap which are not called in 1000G. This is more apparent if we look at the call rates, and the difference in the two platforms is largely due to the much lower call rate for HapMap in this region:

> summary(row.summary(snp.data.1000g.flipped.5fewer))

```
  Call.rate Heterozygosity
 Min.   :1   Min.   :0.1664
 1st Qu.:1   1st Qu.:0.2348
 Median :1   Median :0.2583
 Mean   :1   Mean   :0.2593
 3rd Qu.:1   3rd Qu.:0.2908
 Max.   :1   Max.   :0.3301
```

> summary(row.summary(snp.data.hapmap.5fewer))

```
  Call.rate      Heterozygosity
 Min.   :0.4646   Min.   :0.1794
 1st Qu.:0.4716   1st Qu.:0.2446
 Median :0.4721   Median :0.2695
 Mean   :0.6847   Mean   :0.2741
 3rd Qu.:0.9862   3rd Qu.:0.3009
 Max.   :0.9940   Max.   :0.3774
```

2.1 Further questions

```
> hapmap.25.to.32 <- subset(in.hapmap$snp.support, Position > 2.5e+07 &
+     Position < 3.2e+07)
> length(rownames(hapmap.25.to.32))
```

```
[1] 15136
```

The SNPs overlapping between 1000G and HapMap is only about 1/3 of HapMap's and only about 5% (of 106880) of 1000 Genome's in the region. What happens to the other 2/3 and the other 95%?

3 Repeating the Analysis with CEU (Caucasians), CHB (Northern Chinese), JPT (Japanese)

The commands are not shown below; the boxes shows corresponding results as the YRI panels.

3.1 CEU

```
Reading 174 samples
current line [0] : rs412135 A/G chr6 53...
current line [20000] : rs2179343 C/G chr6 1...
current line [40000] : rs2876639 A/G chr6 2...
current line [60000] : rs9391734 A/G chr6 3...
current line [80000] : rs9395019 C/T chr6 4...
current line [100000] : rs1359069 C/T chr6 5...
current line [120000] : rs1753505 C/T chr6 7...
current line [140000] : rs9450163 A/G chr6 8...
current line [160000] : rs2202733 A/C chr6 9...
current line [180000] : rs2032568 A/G chr6 1...
current line [200000] : rs13201806 C/G chr6 ...
current line [220000] : rs9402956 A/G chr6 1...
current line [240000] : rs9479040 C/T chr6 1...
current line [260000] : rs7755325 G/T chr6 1...
last line [272383] : rs4304215 A/G chr6 1...
EOF reached after 272384 snps
```

```
[1] 5379
```

```
[1] 81
```

```
 ./.  A/C  A/G  A/T  C/G  C/T  G/T
   6  443 1872  346  469 1807  436
```

```
  A/AC   A/AG    A/C    A/G    A/T    C/A   C/CT    C/G    C/T    G/A    G/C
     2      1    212    887    169    232      1    226    958    991    241
  G/GT    G/T GAGA/G    T/A    T/C    T/G
     1    213      1    177    845    222
```

```
           Loci.ID Chrom Position AlleleA AlleleB Alleles
rs5875188 rs5875188     6 29248564       C      CT    C/CT
rs5875197 rs5875197     6 29574595       A      AC    A/AC
rs5875231 rs5875231     6 30070538       G      GT    G/GT
rs5875236 rs5875236     6 30118544    GAGA       G  GAGA/G
rs5875254 rs5875254     6 30386624       A      AG    A/AG
rs4645838 rs4645838     6 31543404       A      AC    A/AC
```

```
   Mode   FALSE    TRUE    NA's
logical    2670    2709       0
```

```
     common       common.called     diff.called        called.in.1      called.in.2
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.0000   Min.   : 4.00   Min.   :0
 1st Qu.:35.00   1st Qu.:35.00   1st Qu.: 0.0000   1st Qu.: 4.00   1st Qu.:0
 Median :35.00   Median :35.00   Median : 0.0000   Median :46.00   Median :0
 Mean   :53.05   Mean   :53.05   Mean   : 0.6377   Mean   :27.31   Mean   :0
 3rd Qu.:77.00   3rd Qu.:77.00   3rd Qu.: 0.0000   3rd Qu.:46.00   3rd Qu.:0
 Max.   :77.00   Max.   :77.00   Max.   :76.0000   Max.   :81.00   Max.   :0
```

```
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
4725  367   89   38   20   17   15    8    7   10    8    2    4    3    2    2
  16   17   18   19   20   21   22   23   25   26   27   28   29   30   31   32
   4    2    1    6    1    5    2    2    2    4    1    1    3    2    6    1
  34   35   51   57   59   61   62   63   75   76
   4    3    1    1    1    1    1    1    2    4
```

The agreements are better at 88% (4725/5379) 95% and 96% respectively.

```
  Call.rate Heterozygosity
 Min.   :1   Min.   :0.007994
 1st Qu.:1   1st Qu.:0.173080
 Median :1   Median :0.231456
 Mean   :1   Mean   :0.222587
 3rd Qu.:1   3rd Qu.:0.270125
 Max.   :1   Max.   :0.348020
```

```
  Call.rate      Heterozygosity
 Min.   :0.4501   Min.   :0.01426
 1st Qu.:0.4566   1st Qu.:0.18935
 Median :0.4575   Median :0.24299
 Mean   :0.6628   Mean   :0.24041
 3rd Qu.:0.9840   3rd Qu.:0.29465
 Max.   :0.9948   Max.   :0.40374
```

3.2 CHB

```
Reading 139 samples
current line [0] : rs4097465 A/C chr6 3...
current line [20000] : rs1318748 C/T chr6 1...
current line [40000] : rs9466378 A/G chr6 2...
current line [60000] : rs2507961 A/G chr6 3...
current line [80000] : rs9462896 C/T chr6 4...
current line [100000] : rs9382543 C/T chr6 5...
current line [120000] : rs2273566 C/T chr6 7...
current line [140000] : rs7749272 A/G chr6 8...
current line [160000] : rs4262195 C/T chr6 9...
current line [180000] : rs11753218 A/C chr6 ...
current line [200000] : rs9490889 C/T chr6 1...
current line [220000] : rs6900658 C/G chr6 1...
current line [240000] : rs13200107 A/T chr6 ...
current line [260000] : rs508144 A/G chr6 16...
last line [276250] : rs4599694 A/G chr6 1...
EOF reached after 276251 snps
```

```
[1] 5449
```

```
[1] 91
```

```
 ./.  A/C  A/G  A/T  C/G  C/T  G/T
   6  447 1897  348  481 1830  440
```

```
  A/AC   A/AG    A/C    A/G    A/T    C/A   C/CT    C/G    C/T    G/A    G/C
     2      1    215    907    171    233      1    233    974    995    247
  G/GT    G/T GAGA/G    T/A    T/C    T/G
     1    218      1    177    852    221
```

```
           Loci.ID Chrom Position AlleleA AlleleB Alleles
rs5875188 rs5875188     6 29248564       C      CT    C/CT
rs5875197 rs5875197     6 29574595       A      AC    A/AC
rs5875231 rs5875231     6 30070538       G      GT    G/GT
rs5875236 rs5875236     6 30118544    GAGA       G  GAGA/G
rs5875254 rs5875254     6 30386624       A      AG    A/AG
rs4645838 rs4645838     6 31543404       A      AC    A/AC
```

```
   Mode   FALSE    TRUE    NA's
logical    2723    2726       0
```

```
     common       common.called    diff.called       called.in.1     called.in.2
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.0000   Min.   : 1.00   Min.   :0
 1st Qu.:41.00   1st Qu.:41.00   1st Qu.: 0.0000   1st Qu.: 1.00   1st Qu.:0
 Median :41.00   Median :41.00   Median : 0.0000   Median :50.00   Median :0
 Mean   :61.45   Mean   :61.45   Mean   : 0.7783   Mean   :28.77   Mean   :0
 3rd Qu.:90.00   3rd Qu.:90.00   3rd Qu.: 0.0000   3rd Qu.:50.00   3rd Qu.:0
 Max.   :90.00   Max.   :90.00   Max.   :89.0000   Max.   :91.00   Max.   :0
```

```
   0    1    2    3    4    5    6    7    8    9   10   12   13   14   15   16
4664  429  108   49   23   29   19   21    7    9    5    6    4    6    4    3
  17   18   19   20   21   23   24   25   26   27   28   30   31   32   33   34
   1    2    3    3    1    2    3    2    4    2    2    2    1    1    1    2
  35   36   37   38   39   40   41   48   68   69   71   72   82   83   85   89
   1    6    2    1    3    1    5    1    1    1    1    1    1    1    3    2
```

86% (4664/5449), 93%, 95%.

```
   Call.rate Heterozygosity
 Min.   :1   Min.   :0.07121
 1st Qu.:1   1st Qu.:0.19747
 Median :1   Median :0.24059
 Mean   :1   Mean   :0.23781
 3rd Qu.:1   3rd Qu.:0.27776
 Max.   :1   Max.   :0.35236
```

```
   Call.rate      Heterozygosity
 Min.   :0.4318   Min.   :0.08164
 1st Qu.:0.4435   1st Qu.:0.21463
 Median :0.4441   Median :0.26107
 Mean   :0.6838   Mean   :0.25948
 3rd Qu.:0.9884   3rd Qu.:0.30847
 Max.   :0.9971   Max.   :0.42379
```

3.3 JPT

```
Reading 116 samples
current line [0] : rs4097465 A/C chr6 3...
current line [20000] : rs562777 C/T chr6 10...
current line [40000] : rs2487980 A/G chr6 2...
current line [60000] : rs2269475 C/T chr6 3...
current line [80000] : rs833048 C/G chr6 43...
current line [100000] : rs3823039 A/G chr6 5...
```

```
current line [120000] : rs7773186 C/T chr6 7...
current line [140000] : rs4593335 C/G chr6 8...
current line [160000] : rs9320243 A/G chr6 9...
current line [180000] : rs9400378 C/G chr6 1...
current line [200000] : rs9385310 C/T chr6 1...
current line [220000] : rs2206056 C/T chr6 1...
current line [240000] : rs6901988 A/C chr6 1...
current line [260000] : rs2281403 C/T chr6 1...
last line [276025] : rs4599694 A/G chr6 1...
EOF reached after 276026 snps
```

```
[1] 5429
```

```
[1] 85
```

```
 ./.  A/C  A/G  A/T  C/G  C/T  G/T
   5  449 1889  348  481 1820  437
```

```
  A/AC   A/AG    A/C    A/G    A/T    C/A   C/CT    C/G    C/T    G/A    G/C
     1      1    216    904    171    234      1    233    969    991    247
  G/GT    G/T GAGA/G    T/A    T/C    T/G
     1    215      1    177    846    221
```

```
           Loci.ID Chrom Position AlleleA AlleleB Alleles
rs5875188 rs5875188     6 29248564       C      CT    C/CT
rs5875197 rs5875197     6 29574595       A      AC    A/AC
rs5875231 rs5875231     6 30070538       G      GT    G/GT
rs5875236 rs5875236     6 30118544    GAGA       G  GAGA/G
rs5875254 rs5875254     6 30386624       A      AG    A/AG
```

```
   Mode   FALSE    TRUE    NA's
logical    2712    2717       0
```

```
     common       common.called    diff.called     called.in.1     called.in.2
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 3.00   Min.   :0
 1st Qu.:38.00   1st Qu.:38.00   1st Qu.: 0.00   1st Qu.: 3.00   1st Qu.:0
 Median :39.00   Median :39.00   Median : 0.00   Median :46.00   Median :0
 Mean   :55.04   Mean   :55.04   Mean   : 0.69   Mean   :29.27   Mean   :0
 3rd Qu.:82.00   3rd Qu.:82.00   3rd Qu.: 0.00   3rd Qu.:47.00   3rd Qu.:0
 Max.   :82.00   Max.   :82.00   Max.   :81.00   Max.   :85.00   Max.   :0
```

```
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   16
4738  345   96   50   27   31   21   17    9    3    7    5    7    5    7    3
  18   19   20   21   22   23   24   25   27   30   31   32   33   34   35   36
   3    2    2    3    6    2    2    1    2    2    2    3    4    3    1    1
  37   38   50   52   53   61   63   76   79   80   81
   4    5    1    1    1    1    1    2    1    1    1
```

87% (4738/5429), 94%, 95%.

```
   Call.rate Heterozygosity
 Min.   :1   Min.   :0.002579
 1st Qu.:1   1st Qu.:0.177012
 Median :1   Median :0.212010
 Mean   :1   Mean   :0.211869
 3rd Qu.:1   3rd Qu.:0.253638
 Max.   :1   Max.   :0.368024
```

```
   Call.rate      Heterozygosity
 Min.   :0.3982   Min.   :0.0008617
 1st Qu.:0.4056   1st Qu.:0.1980918
 Median :0.4060   Median :0.2464093
 Mean   :0.6557   Mean   :0.2450945
 3rd Qu.:0.9875   3rd Qu.:0.2980944
 Max.   :0.9952   Max.   :0.4563239
```

4 Uninteresting details

> proc.time()

```
    user   system  elapsed
1184.411   41.013 1471.968
```

> sessionInfo()

```
R version 2.15.3 (2013-03-01)
Platform: i686-redhat-linux-gnu (32-bit)

locale:
 [1] LC_CTYPE=en_GB.utf8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.utf8        LC_COLLATE=en_GB.utf8
 [5] LC_MONETARY=en_GB.utf8    LC_MESSAGES=en_GB.utf8
 [7] LC_PAPER=C                LC_NAME=C
```

```
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.utf8 LC_IDENTIFICATION=C

attached base packages:
[1] grDevices datasets  splines   graphics  utils     stats     methods
[8] base

other attached packages:
[1] snpMatrix_1.19.0.18 Matrix_1.0-11       lattice_0.20-13
[4] survival_2.37-2

loaded via a namespace (and not attached):
[1] grid_2.15.3  tools_2.15.3
```