

ceu1kg: resources for exploring the 1000 genomes data on individuals of central European ancestry in Bioconductor

VJ Carey

November 2, 2013

1 Introduction

Using results of next generation sequencing experiments, a consortium of geneticists produced calls for SNP at approximately 8 million loci of the genomes of individuals of central European ancestry.

Full genotype calls are held in a folder of SnpMatrix instances:

```
> library(ceu1kg)
> dir(system.file("parts", package="ceu1kg"))

[1] "chr1.rda" "chr10.rda" "chr11.rda" "chr12.rda" "chr13.rda" "chr14.rda"
[7] "chr15.rda" "chr16.rda" "chr17.rda" "chr18.rda" "chr19.rda" "chr2.rda"
[13] "chr20.rda" "chr21.rda" "chr22.rda" "chr3.rda" "chr4.rda" "chr5.rda"
[19] "chr6.rda" "chr7.rda" "chr8.rda" "chr9.rda"

> lk = load(dir(system.file("parts", package="ceu1kg"),full=TRUE)[1])
> c1gt = get(lk)
> c1gt

A SnpMatrix with 60 rows and 605756 columns
Row names: NA06985 ... NA12874
Col names: chr1:533 ... chr1:247196267
```

Metadata about the loci are provided in GRanges instances available from SNPlocs packages. Here we consider the 2010 November release.

```
> library(SNPlocs.Hsapiens.dbSNP.20101109)
> if (!exists("c1loc")) c1loc = getSNPlocs("ch1", as.GRanges=TRUE)
> c1loc
```

GRanges with 1849438 ranges and 2 metadata columns:

| | seqnames | ranges | strand | RefSNP_id |
|-----------|----------|------------------------|--------|-------------|
| | <Rle> | <IRanges> | <Rle> | <character> |
| [1] | ch1 | [10327, 10327] | * | 112750067 |
| [2] | ch1 | [10440, 10440] | * | 112155239 |
| [3] | ch1 | [10469, 10469] | * | 117577454 |
| [4] | ch1 | [10492, 10492] | * | 55998931 |
| [5] | ch1 | [10519, 10519] | * | 62636508 |
| ... | ... | ... | ... | ... |
| [1849434] | ch1 | [249232732, 249232732] | * | 80129254 |
| [1849435] | ch1 | [249232742, 249232742] | * | 28850958 |
| [1849436] | ch1 | [249232749, 249232749] | * | 77296965 |
| [1849437] | ch1 | [249232757, 249232757] | * | 28782254 |
| [1849438] | ch1 | [249232758, 249232758] | * | 28837504 |

alleles_as_ambig

<character>

| | |
|-----------|-----|
| [1] | Y |
| [2] | M |
| [3] | S |
| [4] | Y |
| [5] | S |
| ... | ... |
| [1849434] | R |
| [1849435] | S |
| [1849436] | R |
| [1849437] | Y |
| [1849438] | R |

seqlengths:

| ch1 | ch2 | ch3 | ch4 | ch5 | ch6 | ch7 | ch8 | ... | ch19 | ch20 | ch21 | ch22 | chX | chY | chMT |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|------|------|-----|-----|------|
| NA | NA | NA | NA | NA | NA | NA | NA | ... | NA | NA | NA | NA | NA | NA | NA |

```
> rsn1 = paste("rs", elementMetadata(c1loc)$RefSNP_id, sep="")
> length(intersect(rsn1, colnames(c1gt)))
```

```
[1] 401489
```

```
> ext1 = grep("chr", colnames(c1gt))
> ext1 = as.numeric(gsub("chr1:", "", colnames(c1gt)[ext1]))
> length(intersect(ext1, start(c1loc)))
```

```
[1] 1608
```

The last computation shows that most of the 1KG locations are not in dbSNP.

The Bioconductor *GGdata* package includes HapMap phase II genotypes on 90 CEU individuals in 30 trios, coupled with expression data as distributed at the Sanger GENEVAR project (<ftp://ftp.sanger.ac.uk/pub/genevar/>). The 1KG genotypes are available for 43 of these 90 and the associated genotype plus expression data for these 43 can be acquired using `getSS`, for any chromosome or set of chromosomes.

```
> c20 = getSS("ceu1kg", "chr20")
> c20
```

The above code throws warning because the genotype data are present for 60 individuals, but only 43 have expression values. To create the same structure without a warning:

```
> data(eset) # assume ceu1kg is first in line, yields ex in global
> c1m = c1gt[sampleNames(ex),]
> c1ss = make_smlSet( ex, list(chr1=c1m) )
> c1ss
```

```
SnpMatrix-based genotype set:
number of samples: 43
number of chromosomes present: 1
annotation: illuminaHumanv1.db
Expression data dims: 47293 x 43
Total number of SNP: 605756
Phenodata: An object of class 'AnnotatedDataFrame'
  sampleNames: NA06985 NA06994 ... NA12874 (43 total)
  varLabels: famid persid ... male (7 total)
  varMetadata: labelDescription
```

2 Session information

```
> sessionInfo()
```

```
R version 3.0.2 Patched (2013-10-30 r64123)
Platform: x86_64-apple-darwin10.8.0 (64-bit)
```

```
locale:
[1] C
```

```
attached base packages:
[1] parallel stats4 splines stats graphics grDevices utils
[8] datasets methods base
```

other attached packages:

```
[1] SNPlocs.Hsapiens.dbSNP.20101109_0.99.6
[2] ceu1kg_0.0.34
[3] Biobase_2.22.0
[4] GGtools_4.10.0
[5] Rsamtools_1.14.1
[6] Biostrings_2.30.0
[7] GenomicRanges_1.14.3
[8] XVector_0.2.0
[9] IRanges_1.20.4
[10] BiocGenerics_0.8.0
[11] GGBase_3.24.0
[12] snpStats_1.12.0
[13] Matrix_1.1-0
[14] survival_2.37-4
```

loaded via a namespace (and not attached):

```
[1] AnnotationDbi_1.24.0    BSgenome_1.30.0        DBI_0.2-7
[4] GenomicFeatures_1.14.0  RCurl_1.95-4.1         RSQLite_0.11.4
[7] VariantAnnotation_1.8.2 XML_3.98-1.1           annotate_1.40.0
[10] biomaRt_2.18.0          bit_1.1-10             bitops_1.0-6
[13] ff_2.2-12              genefilter_1.44.0       grid_3.0.2
[16] lattice_0.20-24         rtracklayer_1.22.0     tools_3.0.2
[19] xtable_1.7-1            zlibbioc_1.8.0
```