

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

VJ Carey

April 18, 2015

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 object 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

seqinfo: 25 sequences from an unspecified genome; no seqlengths

```
> 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("ceukg", "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 ceukg 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 3.2.0 (2015-04-16)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: Ubuntu 14.04.2 LTS
```

```
locale:
 [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8      LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8     LC_NAME=C
 [9] LC_ADDRESS=C             LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
```

```
[1] stats4      parallel  stats      graphics  grDevices  utils      datasets
[8] methods     base
```

other attached packages:

```
[1] SNPlocs.Hsapiens.dbSNP.20101109_0.99.6
[2] GenomicRanges_1.20.1
[3] GenomeInfoDb_1.4.0
[4] IRanges_2.2.0
[5] S4Vectors_0.6.0
[6] ceu1kg_0.6.0
[7] Biobase_2.28.0
[8] BiocGenerics_0.14.0
[9] GGtools_5.4.0
[10] data.table_1.9.4
[11] GGBase_3.30.0
[12] snpStats_1.18.0
[13] Matrix_1.2-0
[14] survival_2.38-1
```

loaded via a namespace (and not attached):

```
[1] Rcpp_0.11.5          biovizBase_1.16.0    lattice_0.20-31
[4] Rsamtools_1.20.0     Biostrings_2.36.0    gtools_3.4.2
[7] digest_0.6.8         plyr_1.8.1           chron_2.3-45
[10] futile.options_1.0.0 acepack_1.3-3.3      RSQLite_1.0.0
[13] ggplot2_1.0.1        gplots_2.16.0        zlibbioc_1.14.0
[16] GenomicFeatures_1.20.0 annotate_1.46.0       gdata_2.13.3
[19] hexbin_1.27.0        rpart_4.1-9          proto_0.3-10
[22] splines_3.2.0        BiocParallel_1.2.0  stringr_0.6.2
[25] foreign_0.8-63       RCurl_1.95-4.5       bit_1.1-12
[28] biomaRt_2.24.0       munsell_0.4.2        rtracklayer_1.28.0
[31] biglm_0.9-1          Gviz_1.12.0          nnet_7.3-9
[34] ff_2.2-13           Hmisc_3.15-0         matrixStats_0.14.0
[37] XML_3.98-1.1         GenomicAlignments_1.4.0 MASS_7.3-40
[40] bitops_1.0-6         grid_3.2.0           xtable_1.7-4
[43] gtable_0.1.2         DBI_0.3.1            scales_0.2.4
[46] KernSmooth_2.23-14  XVector_0.8.0        ROCR_1.0-7
[49] reshape2_1.4.1       genefilter_1.50.0    latticeExtra_0.6-26
[52] futile.logger_1.4    Formula_1.2-1        lambda.r_1.1.7
[55] RColorBrewer_1.1-2   iterators_1.0.7      tools_3.2.0
[58] dichromat_2.0-0     BSgenome_1.36.0     AnnotationDbi_1.30.0
[61] colorspace_1.2-6     cluster_2.0.1        caTools_1.17.1
[64] VariantAnnotation_1.14.0
```