

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

VJ Carey

October 18, 2014

## 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 3.1.1 Patched (2014-09-24 r66678)
Platform: i386-w64-mingw32/i386 (32-bit)
```

```
locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252
```

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

other attached packages:

- [1] SNPlocs.Hsapiens.dbSNP.20101109\_0.99.6
- [2] GenomicRanges\_1.18.1
- [3] GenomeInfoDb\_1.2.0
- [4] IRanges\_2.0.0
- [5] S4Vectors\_0.4.0
- [6] ceu1kg\_0.3.1
- [7] Biobase\_2.26.0
- [8] BiocGenerics\_0.12.0
- [9] GGtools\_5.2.0
- [10] data.table\_1.9.4
- [11] GGBase\_3.28.0
- [12] snpStats\_1.16.0
- [13] Matrix\_1.1-4
- [14] survival\_2.37-7

loaded via a namespace (and not attached):

[1] AnnotationDbi_1.28.0	BBmisc_1.7	BSgenome_1.34.0
[4] BatchJobs_1.4	BiocParallel_1.0.0	Biostrings_2.34.0
[7] DBI_0.3.1	Formula_1.1-2	GenomicAlignments_1.2.0
[10] GenomicFeatures_1.18.0	Gviz_1.10.0	Hmisc_3.14-5
[13] KernSmooth_2.23-13	MASS_7.3-35	R.methodsS3_1.6.1
[16] RColorBrewer_1.0-5	RCurl_1.95-4.3	ROCR_1.0-5
[19] RSQLite_0.11.4	Rcpp_0.11.3	Rsamtools_1.18.0
[22] VariantAnnotation_1.12.0	XML_3.98-1.1	XVector_0.6.0
[25] acepack_1.3-3.3	annotate_1.44.0	base64enc_0.1-2
[28] biglm_0.9-1	biomaRt_2.22.0	biovizBase_1.14.0
[31] bit_1.1-12	bitops_1.0-6	brew_1.0-6
[34] caTools_1.17.1	checkmate_1.4	chron_2.3-45
[37] cluster_1.15.3	codetools_0.2-9	colorspace_1.2-4
[40] dichromat_2.0-0	digest_0.6.4	fail_1.2
[43] ff_2.2-13	foreach_1.4.2	foreign_0.8-61
[46] gdata_2.13.3	genefilter_1.48.1	ggplot2_1.0.0
[49] gplots_2.14.2	grid_3.1.1	gtable_0.1.2
[52] gtools_3.4.1	hexbin_1.27.0	iterators_1.0.7
[55] lattice_0.20-29	latticeExtra_0.6-26	matrixStats_0.10.3
[58] munsell_0.4.2	nnet_7.3-8	plyr_1.8.1
[61] proto_0.3-10	reshape2_1.4	rpart_4.1-8
[64] rtracklayer_1.26.1	scales_0.2.4	sendmailR_1.2-1
[67] stringr_0.6.2	tools_3.1.1	xtable_1.7-4
[70] zlibbioc_1.12.0		