

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

VJ Carey

August 10, 2010

## 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.

Genotype calls are held in a list of `snp.matrix` instances:

```
> library(ceu1kg)
```

```
Loading package ff 2.1-2
```

```
-getOption("fftempdir")==="/var/folders/4D/4DI98FkjGzqOK2niUTEHSE+++TM/-Tmp-//RtmpTQNdf"
-getOption("ffextension")==="ff"
-getOption("ffdrop")==TRUE
-getOption("fffinonexit")==TRUE
-getOption("ffpagesize")==65536
-getOption("ffcaching")==="mmnoflush" -- consider "ffeachflush" if your system stalls
-getOption("ffbatchbytes")==16777216 -- consider a different value for tuning your sys
```

```
Attaching package ff
```

```
> if (!exists("ceu1KG.sml")) data(ceu1KG.sml)
```

```
> names(ceu1KG.sml)
```

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

```
> ceu1KG.sml[[1]]
```

```
A snp.matrix 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:

```
> if (!exists("ceulkgMeta_20")) data(ceulkgMeta_20)
> ceulkgMeta_20[1:3]
```

GRanges with 3 ranges and 5 elementMetadata values

	seqnames	ranges	strand	ref	alt	depth
	<Rle>	<IRanges>	<Rle>	<character>	<character>	<character>
rs6078030	chr20	[ 9098, 9098]	*	C	T	192
rs4814683	chr20	[ 9795, 9795]	*	G	T	332
rs34147676	chr20	[10731, 10731]	*	C	A	228
	ancest	alleleCnt				
	<character>	<character>				
rs6078030	.	24				
rs4814683	.	40				
rs34147676	.	14				

seqlengths

chr20

NA

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 41 of these 90 and the associated genotype plus expression data for these 41 are stored in an 'smlSet' container:

```
> data(ceulkg)
```

```
> ceulkg
```

snp.matrix-based genotype set:

number of samples: 41

number of chromosomes present: 22

annotation: illuminaHumanv1.db

Expression data dims: 47293 x 41

Phenodata: An object of class "AnnotatedDataFrame"

sampleNames: NA06985, NA06994, ..., NA12874 (41 total)

varLabels and varMetadata description:

famid: hapmap family id

persid: hapmap person id

...: ...

male: logical TRUE if male

(7 total)

## **2 Imputation of 1KG genotypes to 19 ‘phase II only’ CEU individuals**

have been included in the