

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

VJ Carey

November 1, 2018

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	alleles_as_ambig
	<Rle>	<IRanges>	<Rle>	<character>	<character>
[1]	ch1	10327	*	112750067	Y
[2]	ch1	10440	*	112155239	M
[3]	ch1	10469	*	117577454	S
[4]	ch1	10492	*	55998931	Y
[5]	ch1	10519	*	62636508	S
...
[1849434]	ch1	249232732	*	80129254	R
[1849435]	ch1	249232742	*	28850958	S
[1849436]	ch1	249232749	*	77296965	R
[1849437]	ch1	249232757	*	28782254	Y
[1849438]	ch1	249232758	*	28837504	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("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 3.5.1 Patched (2018-07-12 r74967)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.5 LTS
```

```
Matrix products: default
BLAS: /home/biocbuild/bbs-3.8-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.8-bioc/R/lib/libRlapack.so
```

```
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_0.99.7
 [2] ceu1kg_0.20.0
 [3] GGtools_5.18.0
 [4] Homo.sapiens_1.3.1
 [5] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
 [6] org.Hs.eg.db_3.7.0
```

- [7] GO.db_3.7.0
- [8] OrganismDbi_1.24.0
- [9] GenomicFeatures_1.34.0
- [10] GenomicRanges_1.34.0
- [11] GenomeInfoDb_1.18.0
- [12] AnnotationDbi_1.44.0
- [13] IRanges_2.16.0
- [14] S4Vectors_0.20.0
- [15] Biobase_2.42.0
- [16] BiocGenerics_0.28.0
- [17] data.table_1.11.8
- [18] GGBase_3.44.0
- [19] snpStats_1.32.0
- [20] Matrix_1.2-14
- [21] survival_2.43-1

loaded via a namespace (and not attached):

- [1] colorspace_1.3-2 biovizBase_1.30.0
- [3] htmlTable_1.12 XVector_0.22.0
- [5] base64enc_0.1-3 dichromat_2.0-0
- [7] rstudioapi_0.8 hexbin_1.27.2
- [9] bit64_0.9-7 splines_3.5.1
- [11] knitr_1.20 Formula_1.2-3
- [13] Rsamtools_1.34.0 annotate_1.60.0
- [15] cluster_2.0.7-1 graph_1.60.0
- [17] BiocManager_1.30.3 compiler_3.5.1
- [19] httr_1.3.1 backports_1.1.2
- [21] assertthat_0.2.0 lazyeval_0.2.1
- [23] acepack_1.4.1 htmltools_0.3.6
- [25] prettyunits_1.0.2 tools_3.5.1
- [27] bindrcpp_0.2.2 gtable_0.2.0
- [29] glue_1.3.0 GenomeInfoDbData_1.2.0
- [31] reshape2_1.4.3 dplyr_0.7.7
- [33] Rcpp_0.12.19 biglm_0.9-1
- [35] Biostrings_2.50.0 gdata_2.18.0
- [37] rtracklayer_1.42.0 iterators_1.0.10
- [39] stringr_1.3.1 ensemblDb_2.6.0
- [41] gtools_3.8.1 XML_3.98-1.16
- [43] zlibbioc_1.28.0 scales_1.0.0
- [45] BSgenome_1.50.0 VariantAnnotation_1.28.0
- [47] hms_0.4.2 ProtGenerics_1.14.0
- [49] SummarizedExperiment_1.12.0 RBGL_1.58.0

[51] AnnotationFilter_1.6.0	RColorBrewer_1.1-2
[53] curl_3.2	memoise_1.1.0
[55] gridExtra_2.3	ggplot2_3.1.0
[57] biomaRt_2.38.0	rpart_4.1-13
[59] latticeExtra_0.6-28	stringi_1.2.4
[61] RSQLite_2.1.1	genefilter_1.64.0
[63] checkmate_1.8.5	caTools_1.17.1.1
[65] BiocParallel_1.16.0	rlang_0.3.0.1
[67] pkgconfig_2.0.2	matrixStats_0.54.0
[69] bitops_1.0-6	lattice_0.20-35
[71] ROCR_1.0-7	purrr_0.2.5
[73] bindr_0.1.1	GenomicAlignments_1.18.0
[75] htmlwidgets_1.3	bit_1.1-14
[77] tidyselect_0.2.5	plyr_1.8.4
[79] magrittr_1.5	R6_2.3.0
[81] gplots_3.0.1	Hmisc_4.1-1
[83] DelayedArray_0.8.0	DBI_1.0.0
[85] pillar_1.3.0	foreign_0.8-71
[87] RCurl_1.95-4.11	nnet_7.3-12
[89] tibble_1.4.2	crayon_1.3.4
[91] KernSmooth_2.23-15	progress_1.2.0
[93] grid_3.5.1	blob_1.1.1
[95] digest_0.6.18	xtable_1.8-3
[97] ff_2.2-14	munsell_0.5.0
[99] Gviz_1.26.0	