

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

VJ Carey

October 29, 2020

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 4.0.3 (2020-10-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 18.04.5 LTS
```

```
Matrix products: default
BLAS: /home/biocbuild/bbs-3.12-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.12-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_0.28.0
[3] GGtools_5.26.0
[4] Homo.sapiens_1.3.1
[5] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
[6] org.Hs.eg.db_3.12.0
```

[7] GO.db_3.12.0
[8] OrganismDbi_1.32.0
[9] GenomicFeatures_1.42.0
[10] GenomicRanges_1.42.0
[11] GenomeInfoDb_1.26.0
[12] AnnotationDbi_1.52.0
[13] IRanges_2.24.0
[14] S4Vectors_0.28.0
[15] Biobase_2.50.0
[16] BiocGenerics_0.36.0
[17] data.table_1.13.2
[18] GGBase_3.52.0
[19] snpStats_1.40.0
[20] Matrix_1.2-18
[21] survival_3.2-7

loaded via a namespace (and not attached):

[1] colorspace_1.4-1	ellipsis_0.3.1
[3] biovizBase_1.38.0	htmlTable_2.1.0
[5] XVector_0.30.0	base64enc_0.1-3
[7] dichromat_2.0-0	rstudioapi_0.11
[9] hexbin_1.28.1	bit64_4.0.5
[11] xml2_1.3.2	splines_4.0.3
[13] knitr_1.30	Formula_1.2-4
[15] Rsamtools_2.6.0	annotate_1.68.0
[17] cluster_2.1.0	dbplyr_1.4.4
[19] png_0.1-7	graph_1.68.0
[21] BiocManager_1.30.10	compiler_4.0.3
[23] httr_1.4.2	backports_1.1.10
[25] assertthat_0.2.1	lazyeval_0.2.2
[27] htmltools_0.5.0	prettyunits_1.1.1
[29] tools_4.0.3	gtable_0.3.0
[31] glue_1.4.2	GenomeInfoDbData_1.2.4
[33] reshape2_1.4.4	dplyr_1.0.2
[35] rappdirs_0.3.1	Rcpp_1.0.5
[37] biglm_0.9-2	vctrs_0.3.4
[39] Biostrings_2.58.0	rtracklayer_1.50.0
[41] iterators_1.0.13	xfun_0.18
[43] stringr_1.4.0	lifecycle_0.2.0
[45] ensemblDb_2.14.0	XML_3.99-0.5
[47] zlibbioc_1.36.0	scales_1.1.1
[49] BSgenome_1.58.0	VariantAnnotation_1.36.0

[51]	hms_0.5.3	MatrixGenerics_1.2.0
[53]	ProtGenerics_1.22.0	SummarizedExperiment_1.20.0
[55]	RBGL_1.66.0	AnnotationFilter_1.14.0
[57]	RColorBrewer_1.1-2	curl_4.3
[59]	memoise_1.1.0	gridExtra_2.3
[61]	ggplot2_3.3.2	biomaRt_2.46.0
[63]	rpart_4.1-15	latticeExtra_0.6-29
[65]	stringi_1.5.3	RSQLite_2.2.1
[67]	genefilter_1.72.0	checkmate_2.0.0
[69]	BiocParallel_1.24.0	rlang_0.4.8
[71]	pkgconfig_2.0.3	matrixStats_0.57.0
[73]	bitops_1.0-6	lattice_0.20-41
[75]	ROCR_1.0-11	purrr_0.3.4
[77]	GenomicAlignments_1.26.0	htmlwidgets_1.5.2
[79]	bit_4.0.4	tidyselect_1.1.0
[81]	plyr_1.8.6	magrittr_1.5
[83]	R6_2.5.0	generics_0.0.2
[85]	Hmisc_4.4-1	DelayedArray_0.16.0
[87]	DBI_1.1.0	pillar_1.4.6
[89]	foreign_0.8-80	RCurl_1.98-1.2
[91]	nnet_7.3-14	tibble_3.0.4
[93]	crayon_1.3.4	BiocFileCache_1.14.0
[95]	jpeg_0.1-8.1	progress_1.2.2
[97]	grid_4.0.3	blob_1.2.1
[99]	digest_0.6.27	xtable_1.8-4
[101]	ff_4.0.4	openssl_1.4.3
[103]	munsell_0.5.0	Gviz_1.34.0
[105]	askpass_1.1	