

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

VJ Carey

October 31, 2017

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("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 3.4.2 (2017-09-28)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.3 LTS
```

```
Matrix products: default
BLAS: /home/biocbuild/bbs-3.6-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.6-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_0.16.0
[3] GGtools_5.14.0
[4] Homo.sapiens_1.3.1
[5] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
[6] org.Hs.eg.db_3.4.2
[7] GO.db_3.4.2
[8] OrganismDbi_1.20.0
[9] GenomicFeatures_1.30.0
[10] GenomicRanges_1.30.0
[11] GenomeInfoDb_1.14.0
[12] AnnotationDbi_1.40.0
[13] IRanges_2.12.0
[14] S4Vectors_0.16.0
[15] Biobase_2.38.0
[16] BiocGenerics_0.24.0
[17] data.table_1.10.4-3
[18] GGBase_3.40.0
[19] snpStats_1.28.0
[20] Matrix_1.2-11
[21] survival_2.41-3
```

loaded via a namespace (and not attached):

```
[1] colorspace_1.3-2 biovizBase_1.26.0
[3] htmlTable_1.9 XVector_0.18.0
[5] base64enc_0.1-3 dichromat_2.0-0
[7] hexbin_1.27.1 bit64_0.9-7
[9] interactiveDisplayBase_1.16.0 splines_3.4.2
[11] knitr_1.17 Formula_1.2-2
[13] Rsamtools_1.30.0 annotate_1.56.0
[15] cluster_2.0.6 graph_1.56.0
[17] shiny_1.0.5 compiler_3.4.2
[19] httr_1.3.1 backports_1.1.1
[21] assertthat_0.2.0 lazyeval_0.2.1
```

[23]	acepack_1.4.1	htmltools_0.3.6
[25]	prettyunits_1.0.2	tools_3.4.2
[27]	gtable_0.2.0	GenomeInfoDbData_0.99.1
[29]	reshape2_1.4.2	Rcpp_0.12.13
[31]	biglm_0.9-1	Biostrings_2.46.0
[33]	gdata_2.18.0	rtracklayer_1.38.0
[35]	iterators_1.0.8	stringr_1.2.0
[37]	mime_0.5	ensemldb_2.2.0
[39]	gtools_3.5.0	XML_3.98-1.9
[41]	AnnotationHub_2.10.0	zlibbioc_1.24.0
[43]	scales_0.5.0	BSgenome_1.46.0
[45]	VariantAnnotation_1.24.0	BiocInstaller_1.28.0
[47]	ProtGenerics_1.10.0	SummarizedExperiment_1.8.0
[49]	RBGL_1.54.0	AnnotationFilter_1.2.0
[51]	RColorBrewer_1.1-2	yaml_2.1.14
[53]	curl_3.0	memoise_1.1.0
[55]	gridExtra_2.3	ggplot2_2.2.1
[57]	biomaRt_2.34.0	rpart_4.1-11
[59]	latticeExtra_0.6-28	stringi_1.1.5
[61]	RSQLite_2.0	genefilter_1.60.0
[63]	RMySQL_0.10.13	checkmate_1.8.5
[65]	caTools_1.17.1	BiocParallel_1.12.0
[67]	rlang_0.1.2	pkgconfig_2.0.1
[69]	matrixStats_0.52.2	bitops_1.0-6
[71]	lattice_0.20-35	ROCR_1.0-7
[73]	GenomicAlignments_1.14.0	htmlwidgets_0.9
[75]	bit_1.1-12	plyr_1.8.4
[77]	magrittr_1.5	R6_2.2.2
[79]	gplots_3.0.1	Hmisc_4.0-3
[81]	DelayedArray_0.4.0	DBI_0.7
[83]	foreign_0.8-69	RCurl_1.95-4.8
[85]	nnet_7.3-12	tibble_1.3.4
[87]	KernSmooth_2.23-15	progress_1.1.2
[89]	grid_3.4.2	blob_1.1.0
[91]	digest_0.6.12	xtable_1.8-2
[93]	ff_2.2-13	httpuv_1.3.5
[95]	munsell_0.4.3	Gviz_1.22.0