Building ceu1kgv

VJ Carey

April 18, 2015

1 Introduction

The collection of 1KG genotype calls and expression data for CEU HapMap lines is a complicated process. Channing has worked with various images of the CEU data, obtained before central distribution at ArrayExpress. Here we document clearly how to combine ArrayExpress data with VCF on 1KG.

2 Expression data

Briefly, the ArrayExpress E-MTAB-198.processed.1.zip file includes two matrices of normalized expression values, one with 60 samples, one with 109. The latter is converted to a matrix and the probe IDs are converted to nuIDs using lumiHumanIDMapping package.

It is regrettable that the ArrayExpress function fails (as of 11/15/13) to work with a request for this E-MTAB resource.

3 Genotype data

We use VCF representations of genotype calls as found in such 1000 genome resources as

```
ALL.chr22.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr22.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tbi
```

Code that imports the genotype information and converts to a *snpStats* `SnpMatrix` instance is

```
> dochrGr = function(chrn, ids, which, genome="hg19",
+  path2vcf = "/proj/rerefs/reref00/1000Genomes_Phase1_v3/ALL/") {
+ Require(VariantAnnotation)
+ Require(snpStats)
+  allvc = dir(path2vcf, patt="ALL.*gz$",
+   full=TRUE)
+  f_ind = grep(paste0(chrn, "\\."), allvc)
+  cpath = allvc[f_ind]
+  stopifnot(file.exists(cpath))
+  stopifnot(length(cpath)==1)
+  vp = ScanVcfParam( info=NA, geno="GT", fixed="ALT", samples=ids,
+         which=which )
+  tmp = readVcf( cpath, genome=genome, param=vp )
+  sz = prod(dim(tmp))
+  if (sz == 0) return(NULL)
+  genotypeToSnpMatrix(tmp)$genotypes
+ }
```

The chromosome-specific SnpMatrix indices are stored in inst/parts of the source image of this package.

4 Integrative data container

We use *GGBase* getSS to combine expression and genotype data in a compact format.

```
> library(GGBase)
> c22 = getSS("ceu1kgv", "chr22")
> c22

SnpMatrix-based genotype set:
number of samples:  79
number of chromosomes present:  1
annotation: lumiHumanAll.db
Expression data dims: 46713 x 79
Total number of SNP: 494328
Phenodata: An object of class 'AnnotatedDataFrame'
  rowNames: NA06984 NA06986 ... NA12890 (79 total)
  varLabels: IID FID ... V12 (18 total)
  varMetadata: labelDescription
```

5 Session information

```
> sessionInfo()

R version 3.2.0 (2015-04-16)
Platform: x86_64-unknown-linux-gnu (64-bit)
```

```
Running under: Ubuntu 14.04.2 LTS

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] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] GGBase_3.30.0   snpStats_1.18.0 Matrix_1.2-0    survival_2.38-1

loaded via a namespace (and not attached):
 [1] lattice_0.20-31      IRanges_2.2.0        XML_3.98-1.1
 [4] digest_0.6.8         GenomeInfoDb_1.4.0   grid_3.2.0
 [7] DBI_0.3.1            xtable_1.7-4         stats4_3.2.0
[10] RSQLite_1.0.0        zlibbioc_1.14.0      genefilter_1.50.0
[13] XVector_0.8.0        annotate_1.46.0      S4Vectors_0.6.0
[16] splines_3.2.0        tools_3.2.0          Biobase_2.28.0
[19] parallel_3.2.0       AnnotationDbi_1.30.0 BiocGenerics_0.14.0
[22] GenomicRanges_1.20.1
```