BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 11 2009, pages 1447–1448
doi:10.1093/bioinformatics/btp169

Genetics and population analysis

Data structures and algorithms for analysis of genetics of gene expression with Bioconductor: GGtools 3.x

Vincent J. Carey[1,*], Adam R. Davis[2], Michael F. Lawrence[3], Robert Gentleman[3] and Benjamin A. Raby[1]

[1]Channing Laboratory, Department of Medicine, [2]I2B2 National Center for Biocomputing, Brigham and Women's Hospital, 75 Francis St. Boston MA 02115 and [3]Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, WA 98109, USA.

Received on January 5, 2009; revised on February 27, 2009; accepted on March 19, 2009

Advance Access publication April 5, 2009

Associate Editor: Alex Bateman

ABSTRACT

Summary: Associations between DNA polymorphisms and mRNA abundance are a natural target of genetic investigations, and microarrays facilitate genome-wide and transcriptome-wide surveys of these associations. This work is motivated by emerging requirements for data architectures and algorithm interfaces to allow flexible exploration of public and private archives of genotyping and expression arrays. Using R/Bioconductor facilities, Phase II HapMap genotypes and Illumina 47K expression assay results archived on multiple populations may be interactively explored and analyzed using commodity hardware.

Availability and Implementation: Open Source. Bioconductor 2.3 packages GGtools, GGBase, GGdata, hmyriB36. Freely available on the web at http://www.bioconductor.org.

Contact: stvjc@channing.harvard.edu

1 INTRODUCTION

Numerous publications have addressed relationships between DNA polymorphisms and gene expression distributions (Cheung *et al*., 2005; Dixon *et al*., 2007; Göring *et al*., 2007; Schadt *et al*., 2008; Stranger *et al*., 2007). Interpretation of associations discovered in such analyses is complex and incomplete (Williams *et al*., 2007). Given the ease with which high-density genotypes and transcriptome-wide mRNA abundance measures can be obtained using microarrays, public and private archives of experiments in 'integrative genomics' will grow in number and scope. In this article, we describe a transparent and flexible approach to managing and interrogating data from integrative genomics experiments using R/Bioconductor.

2 METHODS

2.1 Data structures

The basic data structures are

(1) Expression assay results: a $G \times N$ array of numerical assay values, preprocessed so that N samples can be compared with respect to any of G genes.

(2) Genotyping results: an $S \times N$ array of SNP genotyping results, typically encoding homozygous rare or common allele, or heterozygous state.

(3) Sample-level data: an $N \times R$ collection of attributes assessed on each of the N samples, encoding demographic variables, disease-state indicators and relevant technical processing details.

(4) Versioned structural annotation: data on locations of SNPs are evolving as genome sequences are revised. Rapid resolution of bulk SNP location queries is required—queries may be specified in terms of chromosome, segment, proximity to gene or dbSNP identifiers.

(5) Structured containers for inferential results. let Γ denote some (possibly improper) subset of $1,\ldots,G$ defining indices of a gene set, let Σ denote some subset of $1,\ldots,S$, and let $|Q|$ denote the number of elements of any set Q. When $|\Gamma| \times |\Sigma|$ association tests have been performed, additional computations will often be desired to filter associations, conduct testing diagnostics and visualize results in conjunction with other experimental data. Package *GGBase* defines formal `snpScreenResult` containers that manage test results for downstream programmatic interrogation. The containers are readily transformed to custom UCSC genome browser tracks using facilities in Bioconductor packages *IRanges* and *rtracklayer*.

Bioconductor's *Biobase* package defines the `eSet` class for unified representation of genome-scale assay results, sample-level annotation and experiment-level metadata. The *GGtools* package extends the `eSet` class to `smlSet` in which expression array results are coupled to a highly efficient representation of high-density genotypes due to Clayton and Leung (2007), in the `snpMatrix` package.

2.2 Interface, algorithms

The `gwSnpTests` method is defined for various parameter sets to control execution of statistical tests over the transcriptome/SNP genome data contained in an `smlSet` instance. The most basic calls take the form `gwSnpTests(f, sms, cnum)` where `f` is an R linear modeling formula, `sms` is an `smlSet` and `cnum` is a chromosome label. The formula `f` has the form `y[~x1+...]` where `y` may be a gene symbol or name of a gene set as defined by the *GSEABase* package, and `[~x1+...]` is an (optional) linear predictor specification based on variables contained in the `phenoData` of `sms`. Such calls lead to estimation of generalized linear models relating expression in `y` to (implicitly) all SNP genotypes on chromosome `cnum`, optionally controlling for confounders in `x1+...` if present. The underlying computations use

*To whom correspondence should be addressed.

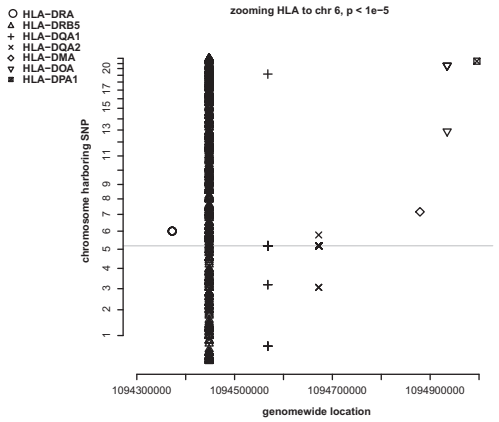

Fig. 1. HLA class 2 gene set eQTL candidates. The gray line is the line of identity (gene location = SNP location), where *cis* associations would be found. Glyphs are plotted where association test $P < 10^{-5}$.

byte-level genotype representations as provided in the `snp.rhs.tests` function of `snpMatrix`.

3 RESULTS AND CONCLUSION

Schadt *et al.* (2008) describe a set of genes in the HLA class II family with evidence of *cis*- or *trans*-associated SNP. We use the *GSEABase* package to define the set of genes represented on the Illumina WG-6 V1 expression array in an R variable `hla2set`. We filter the HapMap CEU cohort to $N = 60$ parents in an R variable `hmFou`. The *GGtools* method `gwSnpTests` computes associations between all HLA class II genes and all Phase II SNP with syntax:

```
hla2run <- gwSnpTests(hla2set~male,
                      hmFou, snpdepth(250))
```

Using a 64-bit Sun Blade with 8 GB RAM, running R 2.9.0, the $9 \times 4 \cdot 10^6$ linear model calculations specified above complete in under 8 min. On a 4 GB 2.6 GHz MacBook Pro the wall clock time is 4 min and 56 s, when the task is split up into collection in chunks of up to four genes at a time with intermittent garbage collection performed manually. Figure 1 is the output of the `masterSnps` function applied to `hla2run` to indicate locations for genes and SNPs with associations possessing sufficiently small *P*-values. *P*-values may be adjusted for multiplicities using Bioconductor's `multtest` package as desired.

Figure 2 employs the *rtracklayer* package to visualize a component of `hla2run` as computed above. The fine structure of the series of association tests is visible at user-selectable resolution

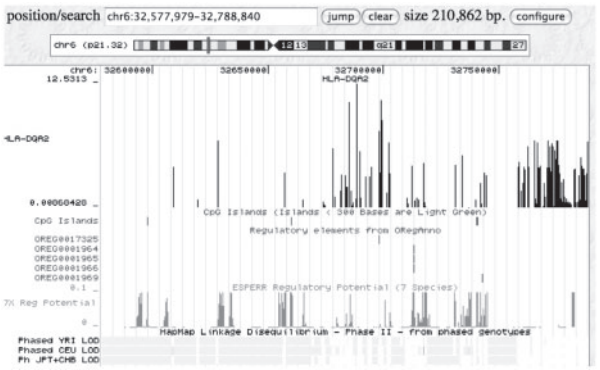

Fig. 2. A view of association results using UCSC browser.

because the association scores are automatically converted to a custom track in the UCSC browser. A high-scoring SNP is seen to be in the vicinity of OREGAnno element 13 725, a CTCF binding site.

In summary, the adoption of efficient container designs for genome-scale assays of mRNA abundance and SNP genotype facilitates interactive analysis of reasonably large integrative genomics studies ($N \approx 10^3, G \approx 10^5, S \approx 10^6$) on commodity hardware. Adoption of object designs in R for inferential outcomes facilitates downstream amalgamation of association patterns with diverse annotation resources, simplifying programmatic interpretation processes.

Funding: National Institutes of Health (P41 HG004059-01 and R01 HL086601-01).

Conflict of Interest: none declared.

REFERENCES

Cheung,V.G. *et al*. (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, 1365–1369, 1476–4687 (Electronic).

Clayton,D. and Leung,H.T. (2007) An R package for analysis of whole-genome association studies. *Hum. Hered.*, **64**, 45–51.

Dixon,A.L. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.*, **39**, 1202–1207.

Göring,H.H. *et al*. (2007) Discovery of expression qtls using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.*, **39**, 1208–1216.

Schadt,E. E. *et al*. (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, 1020–1032.

Stranger,B.E. *et al*. (2007) Population genomics of human gene expression. *Nat. Genet.,* **39**, 1217–1224.

Williams,R.B. *et al*. (2007) The influence of genetic variation on gene expression. *Genome Res.*, **17**, 1707–1716.