Pacific Symposium on Biocomputing 14:380-390 (2009)

INTERPRETING GENETICS OF GENE EXPRESSION: INTEGRATIVE ARCHITECTURE IN BIOCONDUCTOR *

V. J. CAREY

Channing Laboratory
Brigham and Women's Hospital
Harvard Medical School
181 Longwood Ave.
Boston MA 02115 USA
E-mail: stvjc@channing.harvard.edu

R. GENTLEMAN

Program in Computational Biology
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Ave. N, M2-B876
Seattle WA 98109 USA

Several influential studies of genotypic determinants of gene expression in humans have now been published based on various populations including HapMap cohorts. The magnitude of the analytic task (transcriptome vs. SNP-genome) is a hindrance to dissemination of efficient, thorough, and auditable inference methods for this project. We describe the structure and use of Bioconductor facilities for inference in genetics of gene expression, with simultaneous application to multiple HapMap cohorts. Tools distributed for this purpose are readily adapted for the structure and analysis of privately-generated data in expression genetics.

1. Introduction

Figure 1 depicts findings in a general population study of genetics of gene expression. On the left we plot distributions of expression of gene HLA-DRB1 (measured on an Illumina WG-1 platform, as distributed by Sanger Institute in the Genevar project[1]) against genotypes for SNP rs9271367, assayed in 60 individuals from the CEU (central european ancestry) CEPH

*This work was supported in part by NIH P41 HG004059-01 (R Gentleman, PI), NIH R01 HL086601-01 (B Raby, PI), NIH R01 HG003646-01 (R Lazarus, PI).

Pacific Symposium on Biocomputing 14:380-390 (2009)

HapMap cohort. On the right we plot a gender-adjusted measure of association of HLA-DRB1 expression with SNP rare allele counts for all phase II HapMap SNP (3.9 million loci). Interest in HLA-DRB1 stems from the report of Schadt et al.[2] who used a cohort of human liver samples to identify an eQTL at rs9272723. We display the third most-associated SNP (gender-adjusted nominal $p = 1.8 \times 10^{-9}$) in the CEU cohort as it decomposes the expression distribution among three genotype groups most effectively. This association was not reported in the multipopulation eQTL tables published by Stranger et al.[3], presumably for lack of formal genome-wide significance. It is noteworthy that the immortalized B-cell samples assayed by Stranger et al. carry signal similar in nature to that reported for the liver samples by Schadt et al.[2] Using tools described in this paper, it is straightforward to show that among 60 founders in the Yoruba CEPH cohort, the rare allele copy number for rs9271367 is also associated with HLA-DRB1 at a nominal p of 1.5×10^{-7}, but that the primary distinction in expression distributions falls between homozygous common and all others.

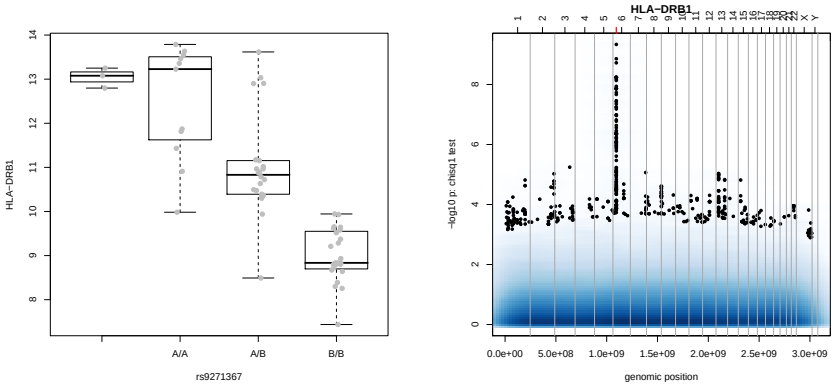

Figure 1. Left: y axis: Expression of HLA-DRB1 in 60 CEU HapMap individuals; x axis: genotype groups formed using SNP rs9271367. Right: y axis: measure of association between HLA-DRB1 expression and rare allele count ($-\log_{10} p$-value for Cochran-Armitage test for trend); x axis: locations of HapMap Phase II SNP according to build 36 release 23a.

In the work of Stranger and colleagues "a detailed association analysis identified at least 1,348 genes with association signals in *cis* and at least 180 in *trans*". The analyses involved filtering of expression probes on the

Pacific Symposium on Biocomputing 14:380-390 (2009)

basis of overall "variance and population differentiation", filtering of SNPs on the basis of minor allele frequency (confined to 5% and above in all CEPH cohorts) and considered both linear regression and rank correlation measures of association between allele counts and expression, applied only to unrelated individuals. The *trans* investigation was limited by confining attention to nonsynonymous SNP known to have *cis*-associtions, SNP thought to be involved in splicing (via Ensembl v41 annotation), and SNP found in sequences defining microRNA.

The combination of Genevar and HapMap resources provides potent tools for increasing our knowledge of genomic structures contributing to expression variation. A number of the steps taken in the pioneering analyses of Stranger *et al.* constitute concessions to computational, inferential and annotational barriers that will be lowered as research and computational prowess mature. We have undertaken in Bioconductor (www.bioconductor.org) to design and disseminate data structures, algorithms, and concrete software packages that simplify research into genetics of gene expression, both with public Genevar/HapMap data, and with privately generated data.

2. Data structures

2.1. *Abstract data types and methods*

An instance `x` of the `smlSet` class (`sml` denoting SNP-matrix list) defined in Bioconductor package *GGBase* satisfies the following basic constraints. If `x` holds information on N individuals assayed for expression through G features of a transcriptome-wide array, then `exprs(x)` returns a $G \times N$ matrix of expression measures, and `featureNames(x)` is a G-vector of strings encoding expression probe identifiers. If the individuals were genotyped on S_c SNP loci on chromosome c, $c = 1, \ldots, C$ (or, more generally, $c \in C$ a set of tokens enumerating chromosomes in an organism), then `snps(x, c)` returns an $S_c \times N$ matrix of SNP genotype assignments of the form A/A, A/B, B/B, NA, where the latter token denotes unavailable genotype. Identifiers for SNPs are retrieved using `snpNames(x, c)`. Genome-wide coordinates of SNP are retrieved using `getSnpLocs(x)`.

A primary method for conducting genome-wide association tests for phenotypes defined by gene expression (mRNA abundance) measures is `gwSnpScreen`, defined in the Bioconductor *GGtools* package. The call `res = gwSnpScreen(formula, smlSet [, chrnum])` returns an object `res` with S_c inference results if `chrnum` identifies chromosome `c`; if `chrnum` is omit-

Pacific Symposium on Biocomputing 14:380-390 (2009)

ted, then the object has $S = \sum_c S_c$ results. The type of analysis conducted depends upon the `formula` passed. In general this will have the form `gs ~ x1 [+ x2 + ...]` where `gs` is an instance of the `GeneSet` class defined by Bioconductor package *GSEABase*, and the right-hand side is a linear predictor expressed in Wilkinson-Rogers notation, using N-dimensional variables typically found in the `phenoData` component of the `smlSet`. These variables can be employed in addition to the rare allele counts for all SNP in $K \times S$ individual tests for association of SNP genotype with expression values for each of the K elements of the gene set `gs`. Finally, metadata about expression reporters can be housed in a `featureData` component, metadata about sample-level variable can be housed in a `varMetadata` component, and the MIAME schema for the experiment can reside in an `experimentData` component; designs for all these structures are defined in package *Biobase*.

The entity `res` returned by `gwSnpScreen` is an instance of a formal class that can be interrogated to identify most strongly associated SNP (adjusting for covariates in the linear predictor), and that can be easily visualized as in the left panel of Figure 1. Inferential summaries are those generated by the *snpMatrix* package genome-wide association test functions `single.snp.tests` (no covariate adjustment) or `snp.rhs.tests` (generalized linear model with covariates)[4]. The return object also includes `call` information to indicate how it was created.

2.2. *Statistical internals*

Of primary interest are collections of p-values summarizing the statistical strength of the genotype-expression associations. These are derived from percentiles of the Chi-squared distribution with one d.f. evaluated at the score statistic for the gene- and SNP-specific parameter β_{gs} in the linear model for log expression of gene g

$$Y_{gi} = \alpha_g + \beta_{gs} c_{si} + \gamma_{gs}^t Z_i + e_{gsi},$$

where c_{si} is the copy number of the rare allele for SNP s on subject i, Z_i is a q-vector of confounders of the association betwen genotype and expression, and e_{gsi} is a Gaussian disturbance with zero mean and constant (gene- and SNP-specific) variance over all subjects. These p-values are provided in nominal form; of them may be transformed to false discovery rates, or to other corrected versions accounting for multiple comparisons, using the `multtest` package of Bioconductor.

Pacific Symposium on Biocomputing 14:380-390 (2009)

2.3. *Concrete illustrations*

The *GGtools* package provides an exemplar `smlSet` instance called `hmceuB36.2021`:

```
> library(GGtools); data(hmceuB36.2021)

> hmceuB36.2021

snp.matrix-based genotype set:
number of samples:  90
number of snp.matrix:  2
annotation:
 exprs: illuminaHumanv1.db
 snps: snp locs package: GGBase ; SQLite ref: hmceuAmbB36_23a_dbconn
Expression data: 47293 x 90
Phenodata: An object of class "AnnotatedDataFrame"
  sampleNames: NA06985, NA06991, ..., NA12892  (90 total)
  varLabels and varMetadata description:
    famid: hapmap family id
    persid: hapmap person id
    ...: ...
    isAdad: logical TRUE if person is a father
    (9 total)
```

This structure is easy to work with interactively as it maintains information on only two chromosomes, 20 and 21. The genome-wide data are available separately in the *GGdata* package as `hmceuB36`.

Decoding and translation of expression reporter nomenclature is carried out using standard platform-specific Bioconductor SQLite annotation data packages. Maintenance of metadata on millions of SNP reporters is more challenging. We investigated several approaches to storing and responding to queries about SNP identifiers, locations, and allele assignments, including web services and netCDF. At present we use SQLite tables; when an `smlSet` instance is brought into scope, a SQLite connection is created and linked to the instance for query resolution. The primary use cases of interest concern visualization and computation of SNP-gene distances, and so only retrieval of large blocks of locations (genome-wide or chromosome-wide) are provided at present. Fine-grained queries on SNP metadata can be carried out using special packages devoted to dbSNP data serialization or the Bioconductor *biomaRt* package.

Pacific Symposium on Biocomputing 14:380-390 (2009)

To conduct chromosome-wide inference on a small gene set, we create `hmFou`, a restriction of the samples in `hmceuB36.2021` to the cohort founders (parents), and then:

```
> library(GSEABase)
> s1 = GeneSet(c("CPNE1", "RPS26"),
     geneIdType = SymbolIdentifier())

> f1 = gwSnpScreen(s1 ~ male, hmFou, chrnum(20))

> f1

multi genome-wide snp screen result:
gene set used as response:
setName: NA
geneIds: CPNE1, RPS26 (total: 2)
geneIdType: Symbol
collectionType: Null
details: use 'details(object)'
there are 2 results.
the call was:
gwSnpScreen(sym = s1 ~ male, sms = hmFou, cnum = chrnum(20))
```

The formula dictates that each gene will be analyzed correcting for gender. The resulting object is an instance of class multiGwSnpScreenResult, a list of elements of class cwSnpScreenResult. Each of these elements can be plotted, or interrogated for significant SNP.

Genome-wide inference on a gene set of even modest size can be mechanically challenging. In the phase II HapMap context, each gene yields 4 million inference measures, many of which are of no importance. The `gwSnpScreen` method for gene sets can receive a `snpdepth` parameter. If this parameter has value D, results on all SNPs but those yielding the D smallest p-values, per chromosome, are discarded at the earliest possible moment.

3. Knowledge-driven applications

3.1. *In silico appraisal of putative eQTL*

In Stranger's published list of multi-population *cis*-eQTL, gene UTS2 on chromosome 1 is distinguished as having SNP determinants of expression only in the Asian populations CHB-JPT. Figure 2 depicts the findings,

Pacific Symposium on Biocomputing 14:380-390 (2009)

chromosome-wide, in the CEU founders. The right-hand panel shows that the tests statistics in customary use are sensitive to outlying values for expression in small subsets of the data.

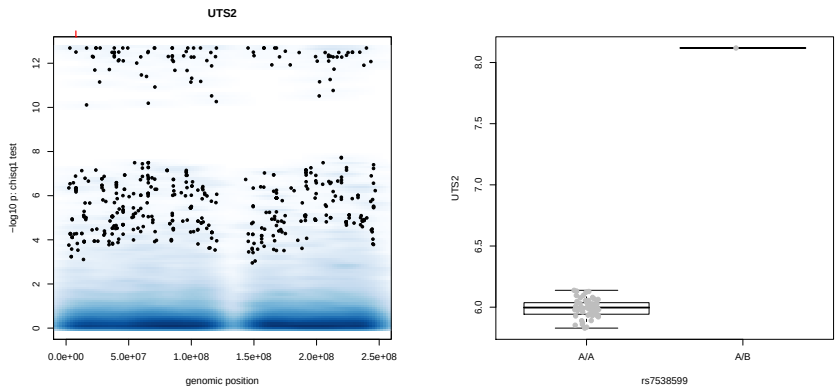

Figure 2. Left: whole-chromosome association analysis for gene UTS in the CEU founders. Right: the expression-genotype configuration that gives rise to the "many eQTL" appearance in the whole-chromosome analysis.

3.2. *Surveying a gene set for eQTL*

The *GSEABase* package provides convenient facilities for defining and translating gene sets between diverse nomenclatures. There are also facilities for importing reference gene set collections such as the Broad Institute's msigDB. We chose to study the motif-based set `V$FREAC2_01`, containing genes with promoter regions including a motif related to FOXF2 (forkhead box F2), because this set includes a gene (CPNE1) with a well-documented eQTL, and because FOXF2 is involved in activation of lung-specific proteins. Probes on the Illumina WG-1 expression array for the CEU founders were filtered to satisfy 1) membership in this gene set, 2) existence of unique Entrez identifier, and 3) in the case of multiple probes sharing an Entrez identifier, the probe with greatest IQR over all samples was retained. This yielded 201 probes; at time of writing, 140 have been analyzed as described here. Using genome-wide testing with snpdepth (as described above) set to 500 per chromosome, each gene is analyzed for eQTL in about three minutes on a Sun Blade with 8GB RAM.

Pacific Symposium on Biocomputing 14:380-390 (2009)

Figure 3 gives lightweight visualizations of per-chromosome distributions of assocation statistics for four genes from the FOXF2 gene set. Some association statistic had to satisfy $-\log_{10} p > 6$ to be included; HABP4 seems to possess a straight *cis*-eQTL; PIK3C2A has a complex appearance; MCM7 appears to have a *trans*-eQTL; AKT2 may possess several.

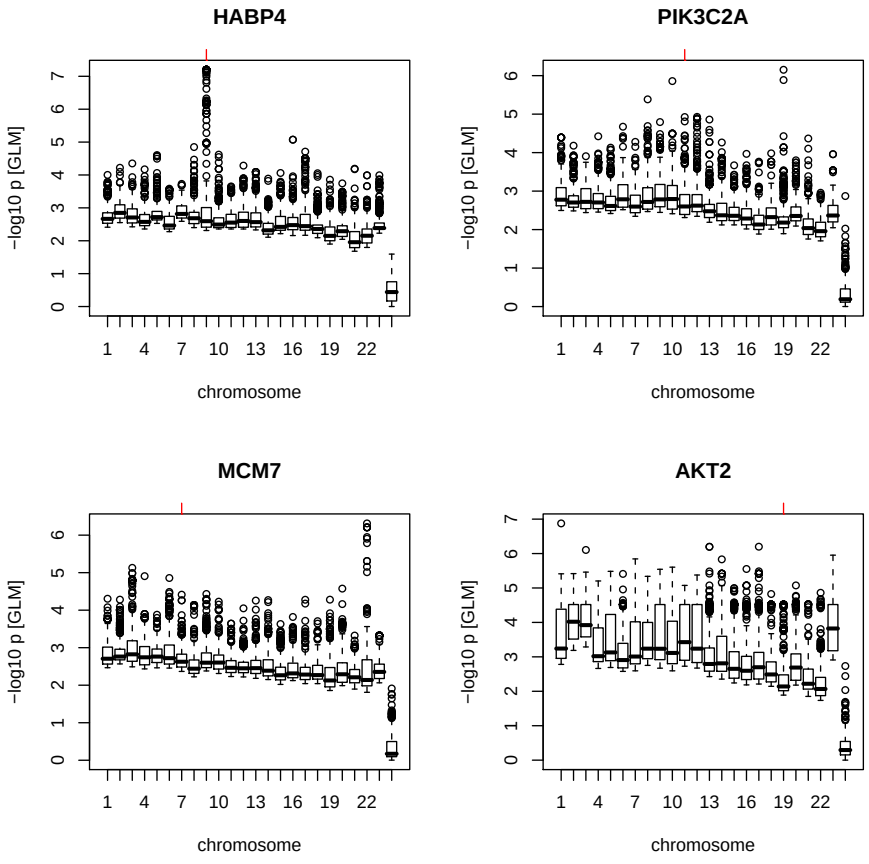

Figure 3. Left: whole-genome association analyses for four members of the FOXF2 motif-based gene set. Tick marks on upper bounding box are approximate location of coding region for each gene for which eQTL were assessed.

Pacific Symposium on Biocomputing 14:380-390 (2009)

3.3. *Combining SNP-expression association scores with reference information on regulatory elements*

Results of `gwSnpScreen` can be transformed to UCSC browser track inputs (WIG format) using the `toTrackSet` method in conjunction with the *rtracklayer* package. Figure 4 shows a fairly coarse view of SNP and putative regulatory regions in the vicinity of CPNE1. There are many non-synonymous coding SNP lying under the CNPE1-associated hump, along with various locations where there is evidence of regulatory elements. Much more information on functional impacts and correlates of polymorphic DNA must be brought to bear to further our understanding of diversity in gene expression.

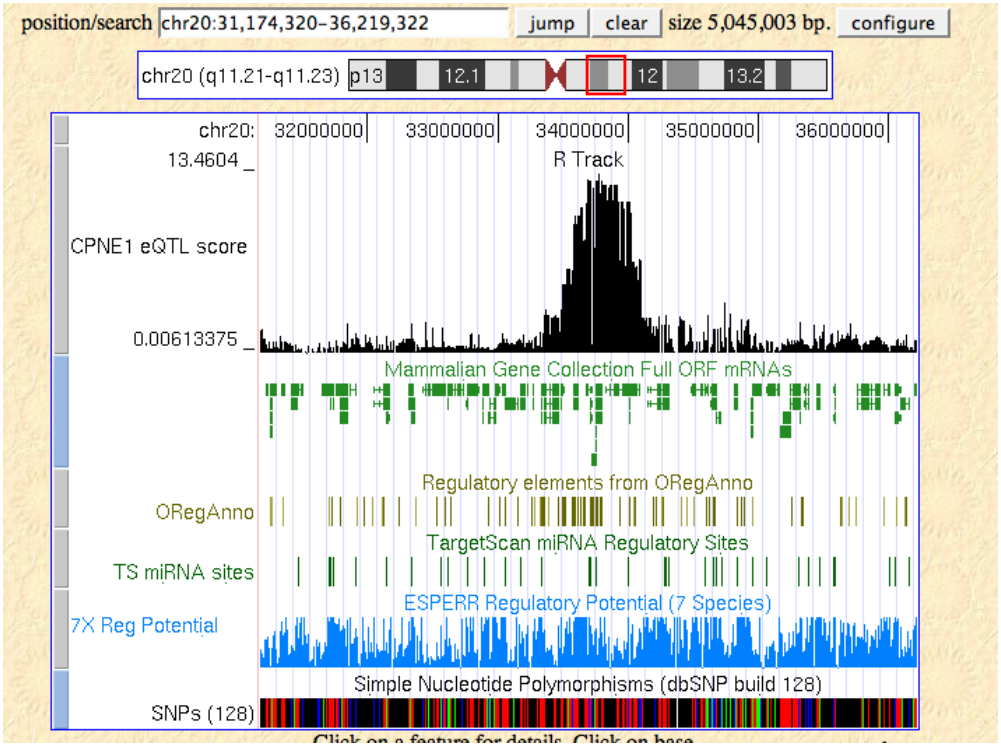

Figure 4. UCSC browser with custom track based on tests for eQTL for CPNE1. The score is $-\log_{10} p$ for the linear regression of log expression on copy number of SNP rare allele.

Pacific Symposium on Biocomputing 14:380-390 (2009)

4. Discussion

In a recent survey of expression genetics, Williams and colleagues[5] suggest that the search for genetic explanation of expression variation is "somewhat simplistic", citing the many non-genetic determinants along with the wide variety of mechanisms by which genetic variation could affect gene expression. These authors also identify a number of technical problems of interpretation of eQTL statistics, including the effects of polymorphisms in hybridization probes, expression array batch effects, and effects due to expression array normalization. They complain that "a more disappointing general observation is that the ability to combine independent studies, even those carried out upon the same organism, is severely compromised by the multiplicity of mapping panels, genetic markers, statistical methodology, genes on arrays, and array platforms".

The approach described in this paper to investigating the relationships between expression variation and genotypic variation represents a step towards facilitating broader integration of multiple experiments and multiple forms of biologic metadata in studies of expression genetics.

- First, multiassay surveys of cohorts are represented in unified and coordinated objects with relatively simple but rich query resolution support. These objects can contain hundreds of samples with millions of SNPs and be manipulated interactively on commodity hardware.
- Second, genome-wide statistical analyses of expression-genotype associations are conducted using high-level facilities (including general covariate adjustments, and formulas involving gene sets as dependent variables) with good performance thanks to detailed programming with byte-level representations of SNP genotypes due to D. Clayton (package *snpMatrix*). These analyses also occupy coordinated computational objects that may be programatically transformed, queried, visualized as needed to identify biologically important interpretations.
- Third, a specific mechanism for integrating expression-genotype analysis results with biologic metadata available in the UCSC genome browser has been created on the basis of the *rtracklayer* package. The importation and visualization shown in this paper are complemented by the bidirectional aspect of the browser interface. Information on regulatory structures can be imported back into R for numerical and statistical analysis, to permit detailed

Pacific Symposium on Biocomputing 14:380-390 (2009)

interpretation of observed *cis*- and *trans*- relationships.

It is well-acknowledged that much work remains to be done to create knowledge from the results of expression genetics experiments. Transparent and extensible computational architectures for representing and interpreting these experiments will play a fundamental role in these efforts.

References

1. URL: `www.sanger.ac.uk/humgen/genevar/`
2. E. Schadt, C. Molony, E. Chudin, et al., *PLoS Biology* 6(5):e107 (2008).
3. B. Stranger, A. Nica, M. Forrest, et al., *Nat. Genet.*, 39(10):1217 (2007).
4. D. Clayton, H. Leung, *Hum. Hered.*, 64:41 (2007).
5. R. Williams, E. Chan, M. Cowley et al., *Genome Res.*, 17:1707 (2007).