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.

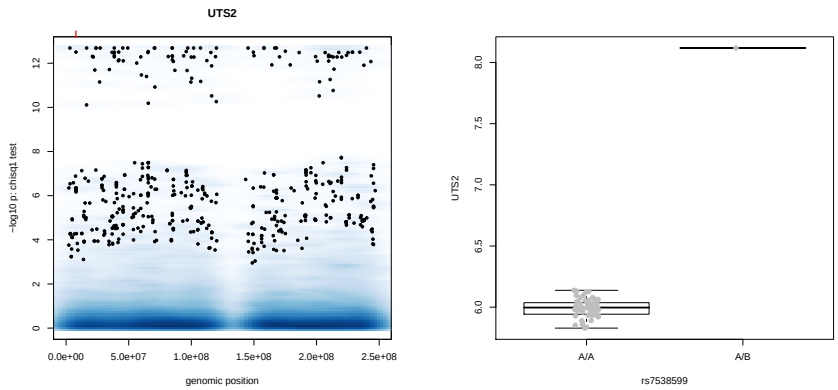

Figure 2. Left: whole-chromosome association analysis for gene UTS in the CEU founders. Right: the expression-genotype configuration that gives rise to the "many eQTL" appearance in the whole-chromosome analysis.

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.

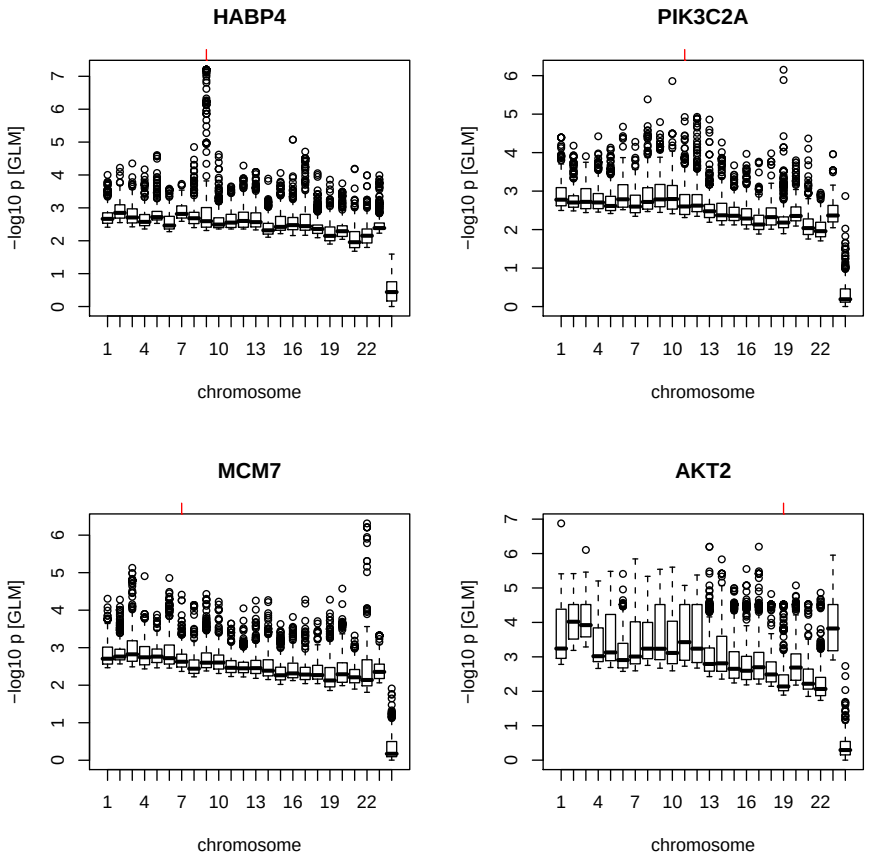

Figure 3. Left: whole-genome association analyses for four members of the FOXF2 motif-based gene set. Tick marks on upper bounding box are approximate location of coding region for each gene for which eQTL were assessed.

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.

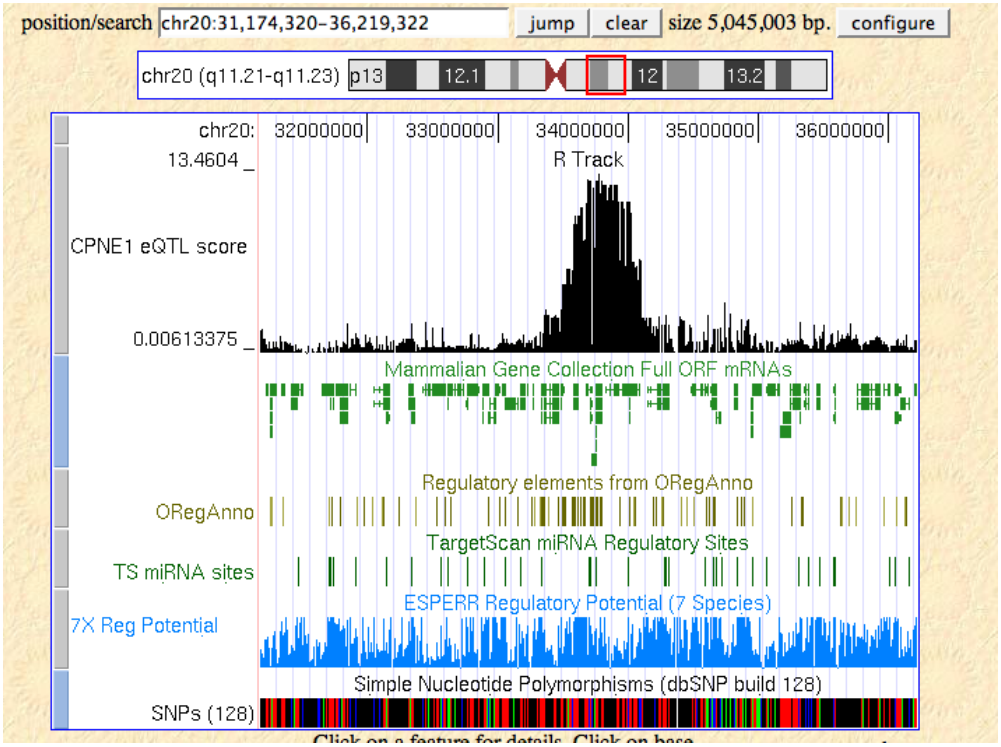

Figure 4. UCSC browser with custom track based on tests for eQTL for CPNE1. The score is $-\log_{10} p$ for the linear regression of log expression on copy number of SNP rare allele.

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).