Representation and parallelism in genomic QTL discovery

Vince Carey, PhD
Channing Division of Network Medicine
Harvard Medical School
Road map of talk

• Brief review of scientific objective: cis-gQTL detection
• Software pkg + data pkg can be effective for high volume data
• Special data representations have been important
• Thorough sensitivity analysis requires high performance
• Sensitivity to basic tuning parameter settings exists for cis-eQTL
eQTLs are SNP associated with expression variation (here for gene GSTT1, chr22)
Figure 1. Plausible sites of action for genetic determinants of mRNA levels. Genetic variations influencing gene expression may reside within the regulatory sequences, promoters, enhancers, splice sites, and secondary structure motifs of the target gene and so be genetically in cis (red stars), or there may be variations in the molecular machinery that interact with cis-regulatory sequences and so act genetically in trans (blue stars).

Localizing the specific determinant of variation is difficult.
A daunting problem of statistical inference?

• An eQTL search is essentially $O(10000)$ GWAS
  – The phenotypes are the components of the transcriptome
  – The predictors are SNP contents at as many as 37 million “1KG” SNP (imputed)
  – Single SNP tests are a reasonable but difficult starting place

• The special case of “cis” testing: SNP of interest are near the gene
  – Will focus on the gene-centric question: does gene $g$ have an eQTL nearby?
  – How far to go?
  – How to calibrate the tests?
LETTER

DNase I sensitivity QTLs are a major determinant of human expression variation

Jacob F. Degner1,2,*, Athma A. Pai1,*, Roger Pique-Regi1,*, Jean-Baptiste Veyrieras1,3, Daniel J. Gaffney1,4, Joseph K. Pickrell1, Sherryl De Leon2, Katelyn Michelini4, Noah Lewellen4, Gregory E. Crawford5,6, Matthew Stephens7,7, Yoav Gilad1 & Jonathan K. Pritchard1,4

The mapping of expression quantitative trait loci (eQTLs) has emerged as an important tool for linking genetic variation to changes in gene regulation1-5. However, it remains difficult to identify the causal variants underlying eQTLs, and little is known about the regulatory mechanisms by which they act. Here we show that genetic variants that modify chromatin accessibility and transcription factor binding are a major mechanism through which genetic variation leads to gene expression differences among humans. We used DNaseI sequencing to measure chromatin accessibility in 70 Yoruba lymphoblastoid cell lines, for which genome-wide genotypes and estimates of gene expression levels are also available6-8. We obtained a total of 2.7 billion uniquely and enhancer-associated histone marks. Furthermore, bound transcription factors protect the DNA sequence within a binding site from DNaseI cleavage, often producing recognizable ‘footprints’ of decreased DNaseI sensitivity13,15-17.

We collected DNase-seq data for 70 HapMap Yoruba lymphoblastoid cell lines for which gene expression data and genome-wide genotypes were already available6-8. We obtained an average of 39 million uniquely mapped DNase-seq reads per sample, providing individual maps of chromatin accessibility for each cell line (see Supplementary Information for all analysis details). Our data allowed us to characterize the distribution of DNaseI cuts within individual hypersensitive sites at extremely high resolution. As expected, the DHSs coincided to a great
Figure 3 | Relationship between dsQTLs and eQTLs. a, Example of a dsQTL SNP that is also an eQTL for the gene SLFN5. The SNP disrupts an interferon- genotype at the
Genotype-specific association of dsQTLs and a typical example. d, Box plot showing that rs4953223 (green) is associated with higher DNase I sensitivity (relative to genome-wide average) than rs4953223 (black) in the CC genotype. e, NF-κB motif GAAATTTGG.
Degner, Pai, Pique-Regi et al. (2012): \(-\log_{10}(P)\), LCLs, 76 Nigerian HAPMAP ids. DNase sensitivity QTLs (dsQTLs) by DN

Schadt et al. (2007): \(-\log_{10}(P)\), liver, 427 ids, European descent

Myers et al. (2007): \(-\log_{10}(P)\), cortex from control brain, 279 ids, European descent

Stranger et al. (2007): \(-\log_{10}(P)\), LCLs, 210 HAPMAP ids, 4 single populations.

Veyrieras et al. (2006): \(-\log_{10}(P)\), LCLs, 210 HAPMAP ids, multi-population.

Veyrieras et al. (2006): posterior probability, LCLs, 210 HAPMAP ids, multi-population.
Another problem: global expression measures exhibit significant technical variation.
Another problem: SNPs have varying minor allele frequencies

- Observed genotype at a SNP has possible values AA, AB, BB
- Additive genetic model uses the count of B alleles, for example, as a continuous predictor in linear regression
- When B alleles are rare, the slope estimator has higher variance
- Can set a lower bound on MAF for SNP to be tested, but this is unpleasant
Some underpinnings of an eQTL solution (Bioconductor GGtools)

• Data packages with decomposed genotype data manage the SNP volumes: GGdata, hmyriB36, dsQTL
• Clayton’s snpStats package: a byte-code for genotype probabilities
  – Compact representation of large SNP sets
  – Special code for GLMs to conduct GWAS
• Adler et al.’s ff package: flexible matrix-like interface to ‘flat files’ external to RAM
• R’s parallel package for concurrent computing on multicore hardware
• A decouple/recouple approach to computing genome-wide FDR
Representing (uncertain) SNP genotypes: David Clayton’s byte-sized encoding
Nota bene

- This representation can be used to handle pure genotype calls or Mach or Beagle imputation outputs as posterior genotype distributions
- Hand-coded GLM provided in snpStats to operate on this representation as ind/dep variable
- Question: templating for modeling algorithms? – see RcppEigen
ff to reduce memory consumption

How the creation of n values effects the run-time virtual memory address space:

ff object:

```r
> ffOb <- ff("foo", 8000000)
> aVal <- ffOb[1:2000000]
```

native R vector:

```r
> rObj <- numeric(8000000)
> aVal <- rObj[1:2000000]
```

- **Data storage** (reduced due to memory-mapped pages of flat files)
- **Indices storage** (reduced due to packing of index sequences)
- **Resulting vector storage**

The amount of memory required by an ff object.

The amount of memory required by a native R vector object.
Flexible approach to concurrent, interruptible computing

- Multiple cores on one machine can simultaneously populate an ff archive.
- Archive for a chromosome is harvested for best SNP per gene when all genes are done.
- This applies to both the observed association scores and associations under permutation.
- When all chromosomes are done, the full permutation realization is assembled from the chromosome-specific realizations.
N.B. rhdf5 assessment

• I chose ff well before rhdf5 matured
• Recent comparisons show that for this application, the two approaches have reasonably similar performance
  – Multicore writes seem OK for this application
  – Chromosomes to nodes, genes to cores
• It would be nice to have an abstraction for “out of memory” computations so that alternate back-ends can easily be compared and swapped
Still needed for sensitivity analyses

• Managing one run is reasonably tractable
• Specifying and managing the results of a sensitivity search – still difficult
• Most clusters will have some kind of job submission/management system
  – Our group uses SGE/N1, with qmake ...
• These are often not well-matched to requirements of statistical investigations
The BatchJobs “vignette”, TU Dortmund

Computing on high performance clusters with R: Packages BatchJobs and BatchExperiments

Bernd Bischl, Michel Lang, Olaf Mersmann, Jörg Rahnenführer, Claus Weihs
Figure 2: Relationship of **BatchExperiment** functions. Grey rectangulars require user input. White boxes represent internal functions. A straight arrow stands for direct passing of the object or function, a squiggly line denotes passing of the evaluated result.
Upshots

• High level tools are emerging to smooth the path from statistical computing requirements to effective use of available hardware
  – CPUs/GPUs
  – Disk
  – Network
• Mastery will take work
• The environment is volatile
• Some example results:
> g3[1:5]

GRanges with 5 ranges and 11 elementMetadata cols:

<table>
<thead>
<tr>
<th>seqnames</th>
<th>ranges</th>
<th>strand</th>
<th>snpid</th>
<th>snploc</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr17</td>
<td>[73127217, 73127217]</td>
<td>*</td>
<td>rs3736075</td>
<td>73127217</td>
</tr>
<tr>
<td>chr16</td>
<td>[641445, 641445]</td>
<td>*</td>
<td>rs35585285</td>
<td>641445</td>
</tr>
<tr>
<td>chr2</td>
<td>[65605703, 65605703]</td>
<td>*</td>
<td>rs2217969</td>
<td>65605703</td>
</tr>
<tr>
<td>chr16</td>
<td>[24047271, 24047271]</td>
<td>*</td>
<td>chr16:24047271</td>
<td>24047271</td>
</tr>
<tr>
<td>chr1</td>
<td>[68737383, 68737383]</td>
<td>*</td>
<td>rs59129922</td>
<td>68737383</td>
</tr>
</tbody>
</table>

radiusUsed fdr probe excl maf

<table>
<thead>
<tr>
<th>radiusUsed</th>
<th>fdr</th>
<th>probe</th>
<th>excl</th>
<th>maf</th>
</tr>
</thead>
<tbody>
<tr>
<td>50000</td>
<td>0.0001943635</td>
<td>01YintXoV6ek6AxLqA</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>50000</td>
<td>0.0003236246</td>
<td>02KG50KDEI610L36kU</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>50000</td>
<td>0.0000000000</td>
<td>03tSCXVYVVCnRPBA</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>50000</td>
<td>0.0000000000</td>
<td>04.BQghp5olySHS17o</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>50000</td>
<td>0.0000000000</td>
<td>04jcQvhWQopUsop5I</td>
<td>0</td>
<td>0.05</td>
</tr>
</tbody>
</table>

nperm npc bestfdr sym

<table>
<thead>
<tr>
<th>nperm</th>
<th>npc</th>
<th>bestfdr</th>
<th>sym</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>(32,31,21,19) *1.50</td>
<td>0</td>
<td>NT5C</td>
</tr>
<tr>
<td>3</td>
<td>(32,31,21,19) *1.50</td>
<td>0</td>
<td>WFIKKN1</td>
</tr>
<tr>
<td>3</td>
<td>(32,31,21,19) *1.50</td>
<td>0</td>
<td>SPRED2</td>
</tr>
<tr>
<td>3</td>
<td>(32,31,21,19) *1.50</td>
<td>0</td>
<td>PRKCB</td>
</tr>
<tr>
<td>3</td>
<td>(32,31,21,19) *1.50</td>
<td>0</td>
<td>WLS</td>
</tr>
</tbody>
</table>

seqlengths:

chr1 chr10 chr11 chr12 chr13 chr14 ... chr4 chr5 chr6 chr7 chr8 chr9
NA NA NA NA NA NA ... NA NA NA NA NA NA NA

> sum(values(g3)$fdr <= 0.05)

[1] 3261
Conclusions

• Genomic annotation (gene, SNP names/locations) conveniently available through a given API in 2005, much has changed ... refactor?

• Basic R/bioc facilities facilitate thorough sensitivity analysis

• Sensitivity is apparently present, so criteria for choosing tuning parameters should be sought