Representation and parallelism in genomic QTL discovery

Vince Carey, PhD

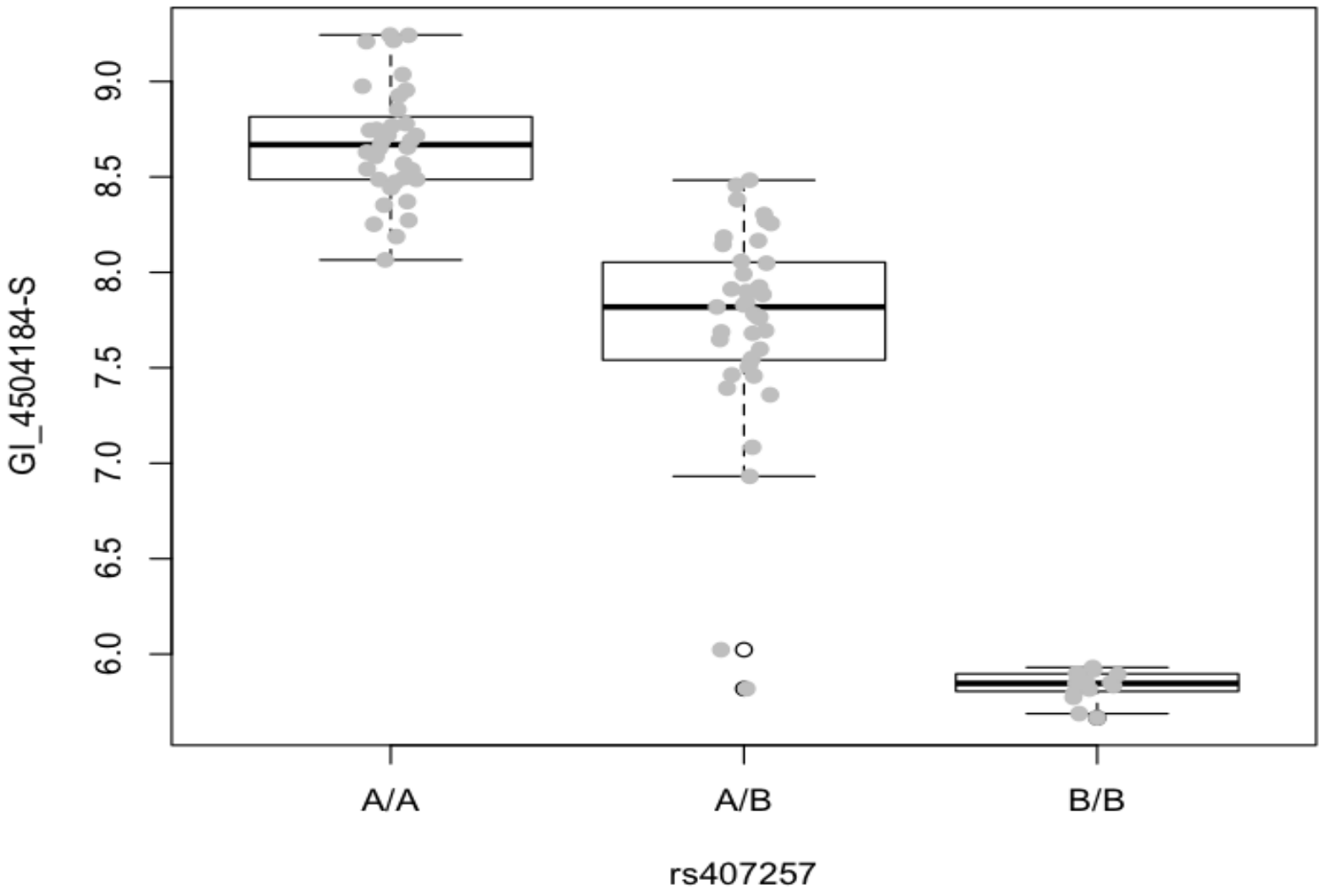
Channing Division of Network Medicine

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)



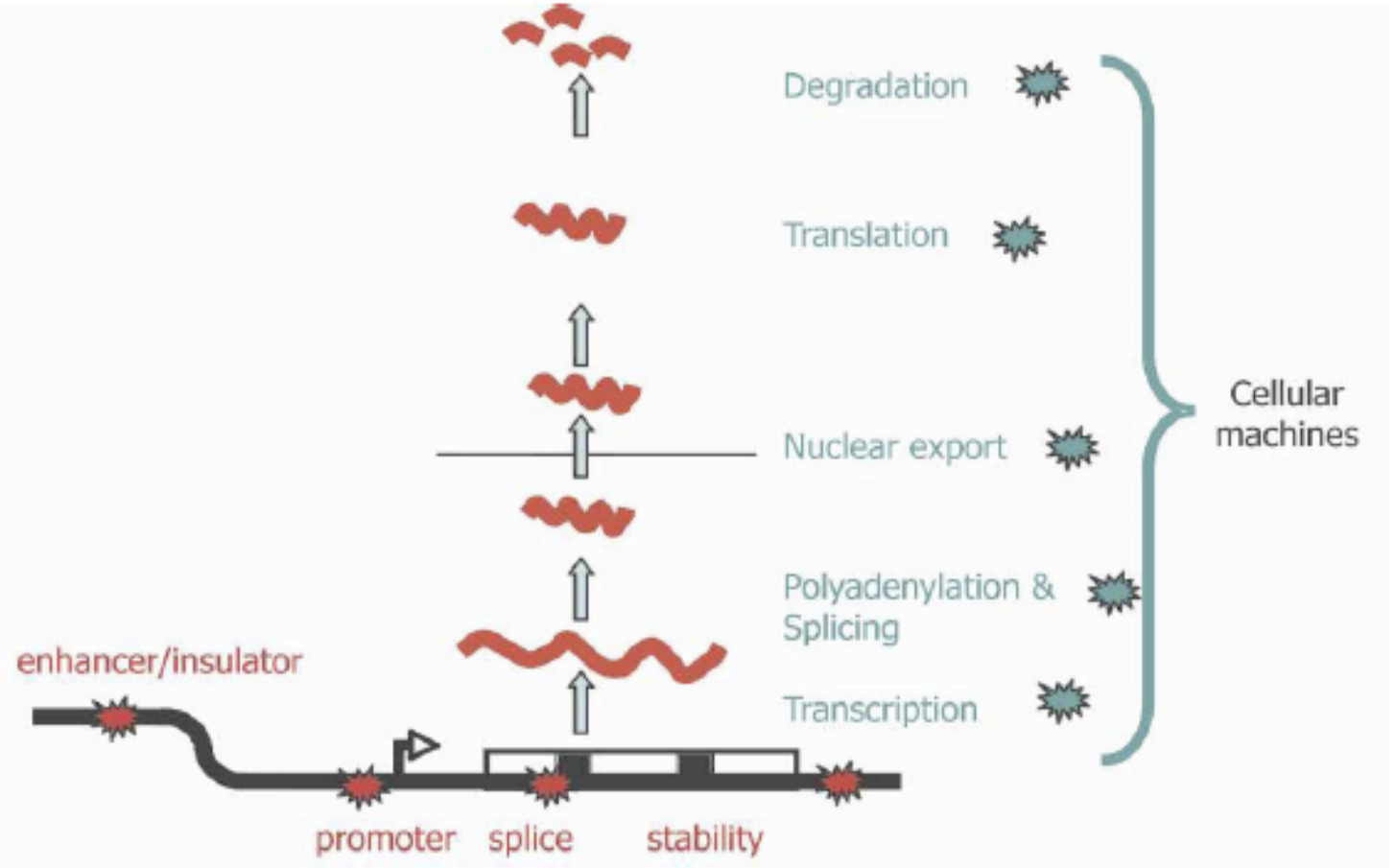
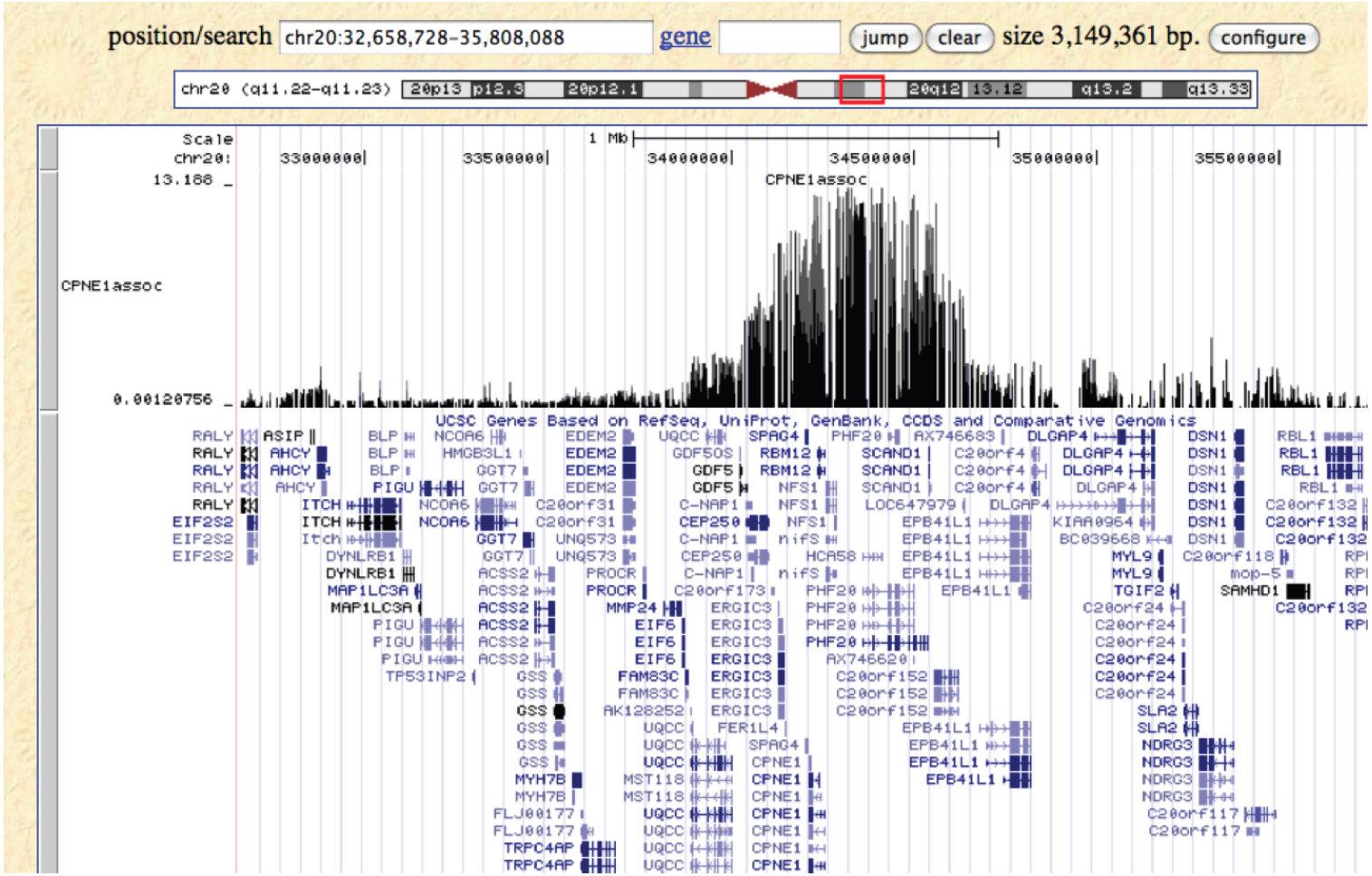


Figure 1. Plausible sites of action for genetic determinants of mRNA levels. Genetic variations influencing gene expression may reside within the regulatory sequences, promoters, enhancers, splice sites, and secondary structure motifs of the target gene and so be genetically in *cis* (red stars), or there may be variations in the molecular machinery that interact with *cis*-regulatory sequences and so act genetically in *trans* (blue stars).

Williams R et al. Genome Research 2007 vol. 17 (12) pp. 1707-1716

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

doi:10.1038/nature10808

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

Jacob F. Degner[1,2]*, Athma A. Pai[1]*, Roger Pique-Regi[1]*, Jean-Baptiste Veyrieras[1,3], Daniel J. Gaffney[1,4], Joseph K. Pickrell[1], Sherryl De Leon[4], Katelyn Michelini[4], Noah Lewellen[4], Gregory E. Crawford[5,6], Matthew Stephens[1,7], Yoav Gilad[1] & Jonathan K. Pritchard[1,4]

The mapping of expression quantitative trait loci (eQTLs) has emerged as an important tool for linking genetic variation to changes in gene regulation[1–5]. 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 DNase I sequencing to measure chromatin accessibility in 70 Yoruba lymphoblastoid cell lines, for which genome-wide genotypes and estimates of gene expression levels are also available[6–8]. 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 DNase I cleavage, often producing recognizable 'footprints' of decreased DNase I sensitivity[13,15–17].

We collected DNase-seq data for 70 HapMap Yoruba lymphoblastoid cell lines for which gene expression data and genome-wide genotypes were already available[6–8]. 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 DNase I cuts within individual hypersensitive sites at extremely high resolution. As expected, the DHSs coincided to a great

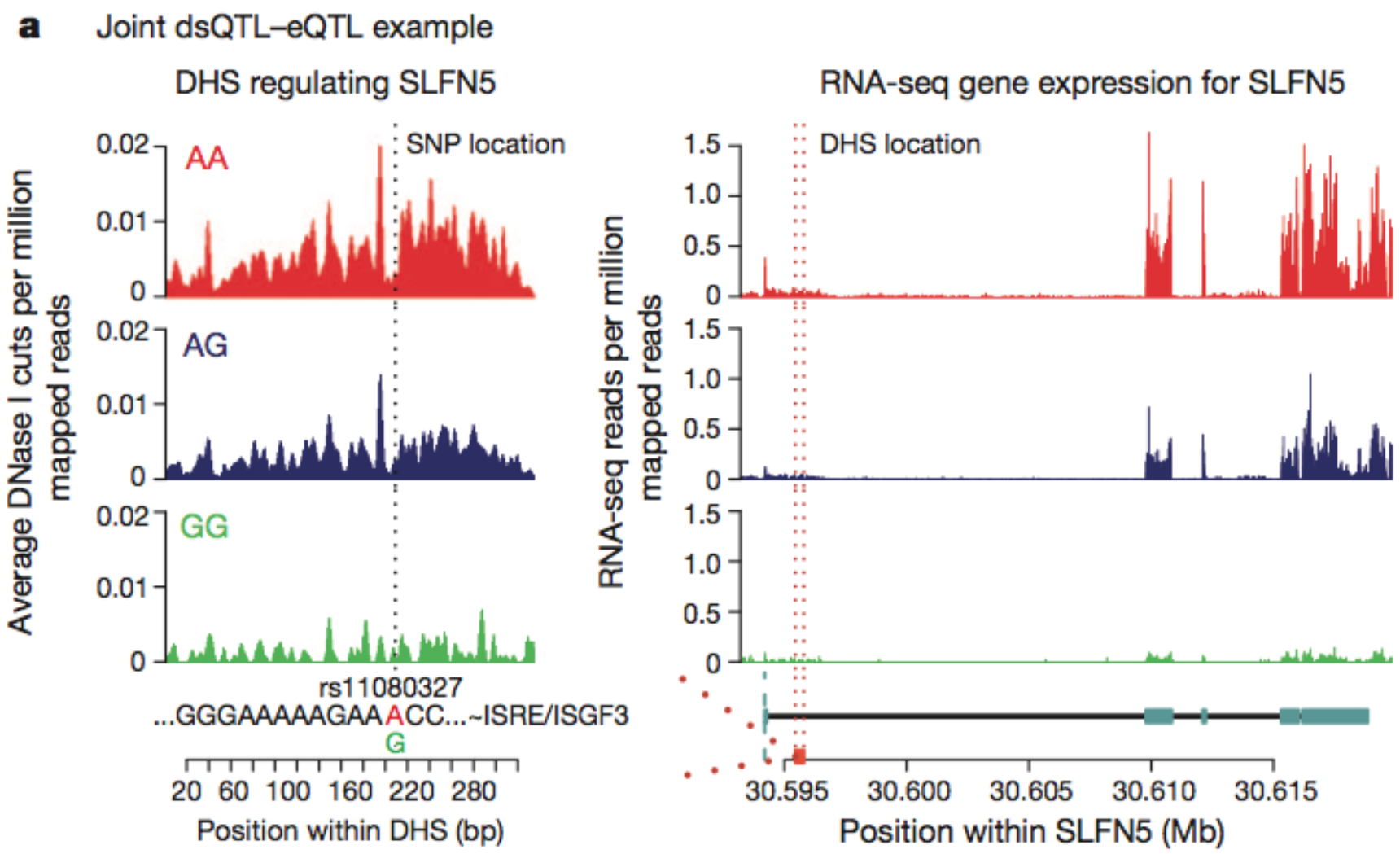
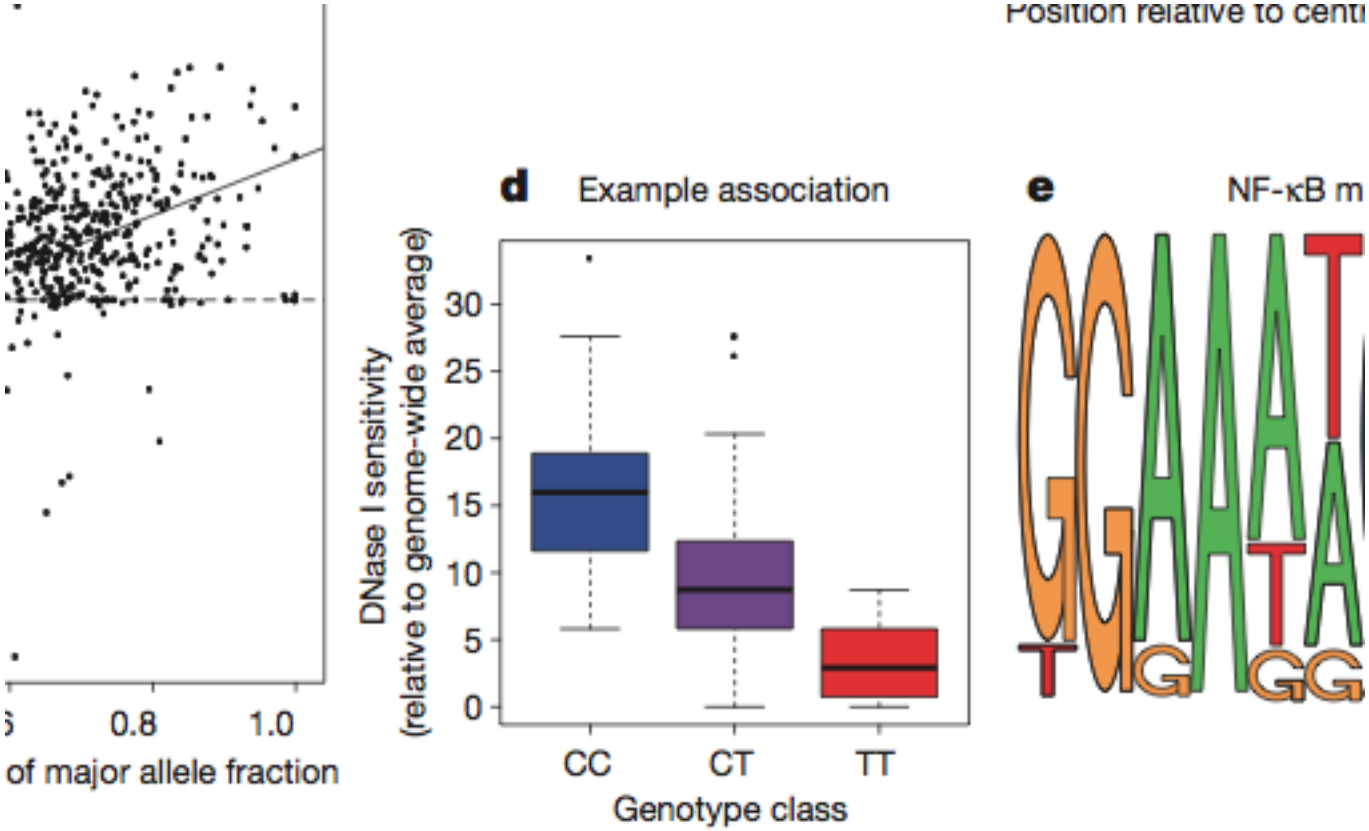


Figure 3 | Relationship between dsQTLs and eQTLs. a, Example of a dsQTL SNP that is also an eQTL for the gene *SLFN5*. The SNP disrupts an interferon- (right) measurer genotype at the p

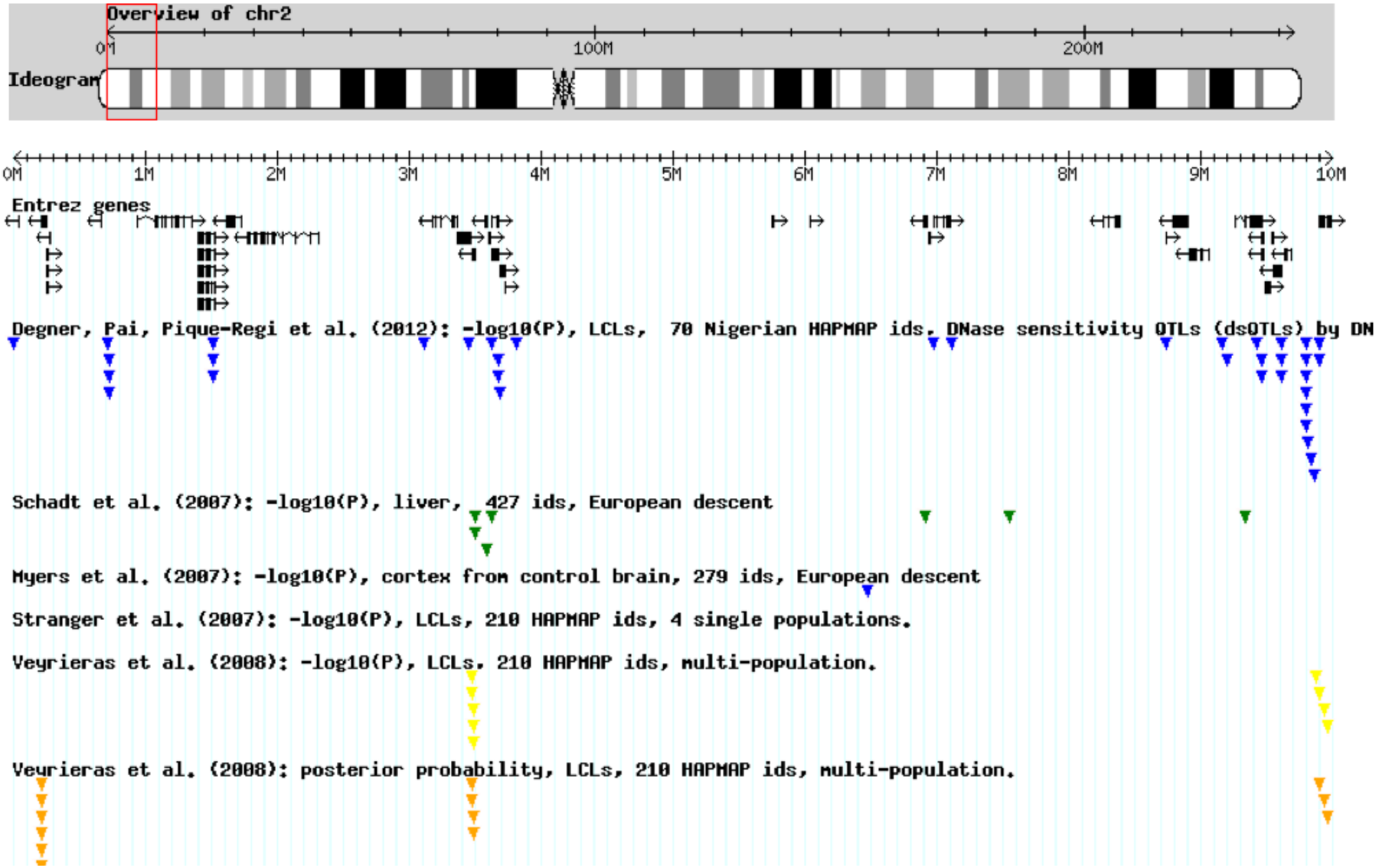


ation of dsQTLs and a typical example.
tion between DNase I cut rates in 100-bp
(green) and 40-kb (black) regions centred
lele-specific analysis of dsQTLs in

dsQTL (rs4953223). The bla
d, Box plot showing that rs4
accessibility ($P = 3 \times 10^{-13}$
DNase I sensitivity, disrupts

view

ils


Overview of chr2
0M
100M
200M
Ideogram
0M
1M
2M
3M
4M
5M
6M
7M
8M
9M
10M
Entrez genes
Degner, Pai, Pique-Regi et al. (2012): -log10(P), LCLs, 70 Nigerian HAPMAP ids, DNase sensitivity QTLs (dsQTLs) by DN
Schadt et al. (2007): -log10(P), liver, 427 ids, European descent
Myers et al. (2007): -log10(P), cortex from control brain, 279 ids, European descent
Stranger et al. (2007): -log10(P), LCLs, 210 HAPMAP ids, 4 single populations.
Veyrieras et al. (2008): -log10(P), LCLs, 210 HAPMAP ids, multi-population.
Veyrieras et al. (2008): posterior probability, LCLs, 210 HAPMAP ids, multi-population.

Another problem: global expression measures exhibit significant technical variation

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies

Oliver Stegle[1,2*9], Leopold Parts[3*9], Richard Durbin[3], John Winn[4*]

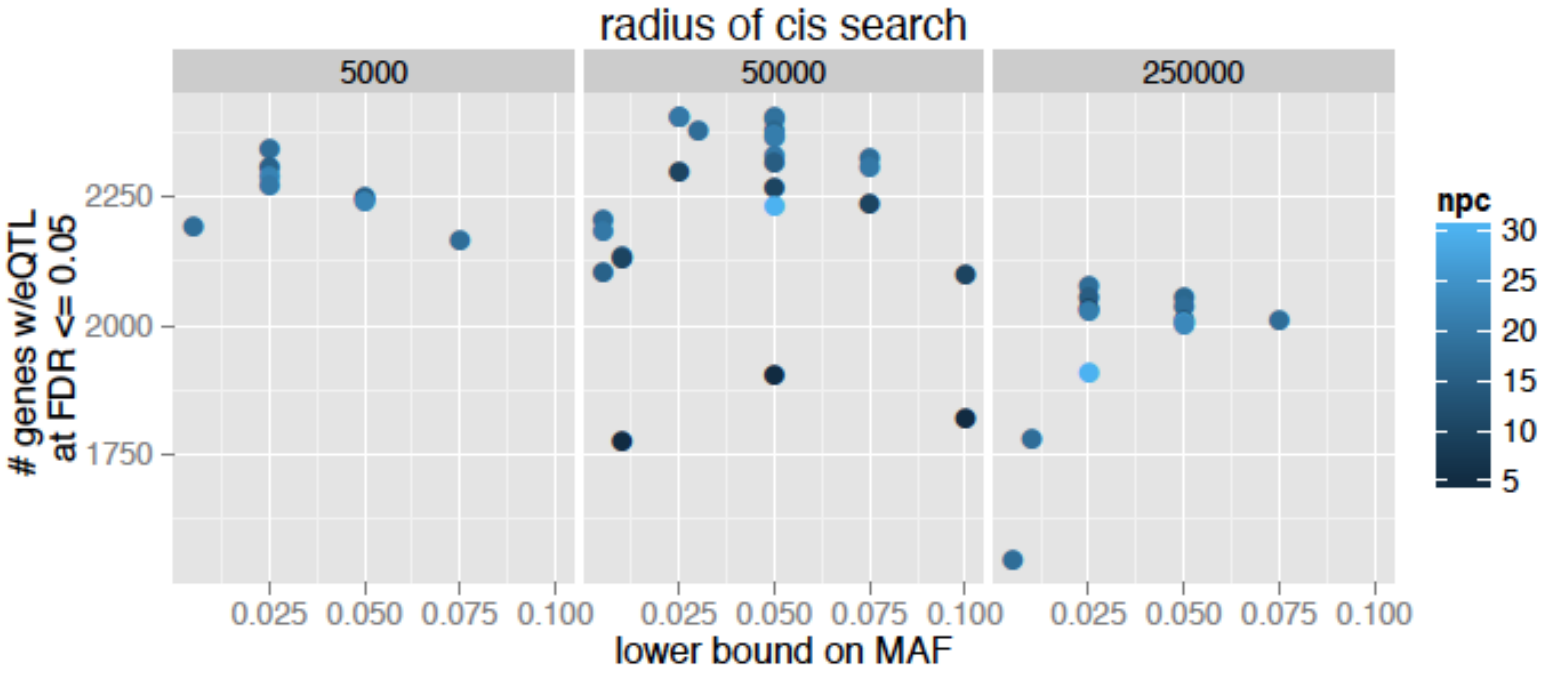
1 Max Planck Institutes Tübingen, Tübingen, Germany, 2 University of Cambridge, Cambridge, United Kingdom, 3 Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom, 4 Microsoft Research, Cambridge, United Kingdom

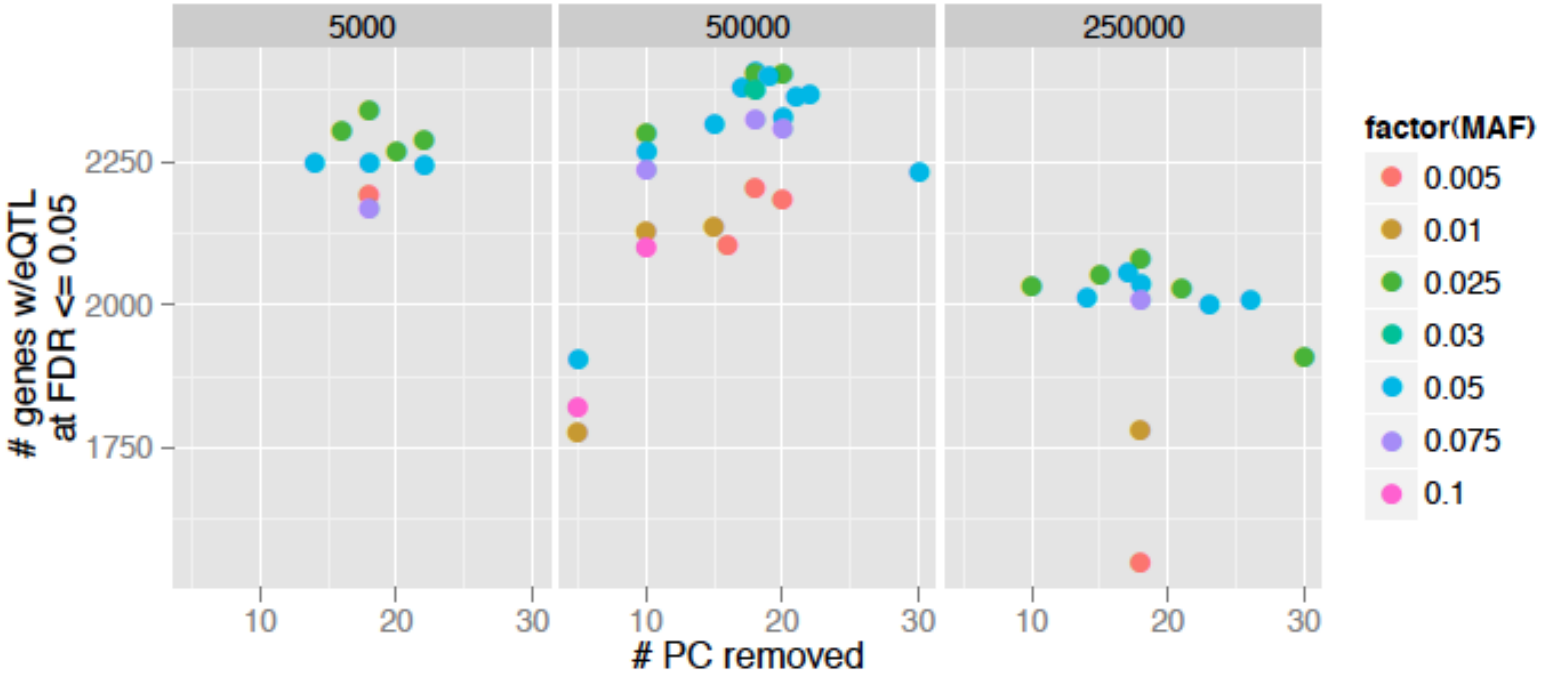
Abstract

Gene expression measurements are influenced by a wide range of factors, such as the state of the cell, experimental conditions and variants in the sequence of regulatory regions. To understand the effect of a variable of interest, such as the genotype of a locus, it is important to account for variation that is due to confounding causes. Here, we present VBQTL, a probabilistic approach for mapping expression quantitative trait loci (eQTLs) that jointly models contributions from genotype as well as known and hidden confounding factors. VBQTL is implemented within an efficient and flexible inference framework, making it fast and tractable on large-scale problems. We compare the performance of VBQTL with alternative methods for dealing with confounding variability on eQTL mapping datasets from simulations, yeast, mouse, and human. Employing Bayesian complexity control and joint modelling is shown to result in more precise estimates of the contribution of different confounding factors resulting in additional associations to measured transcript levels compared to alternative approaches. We present a threefold larger collection of *cis* eQTLs than previously found in a whole-genome eQTL scan of an outbred human population. Altogether, 27% of the tested probes show a significant genetic association in *cis*,

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

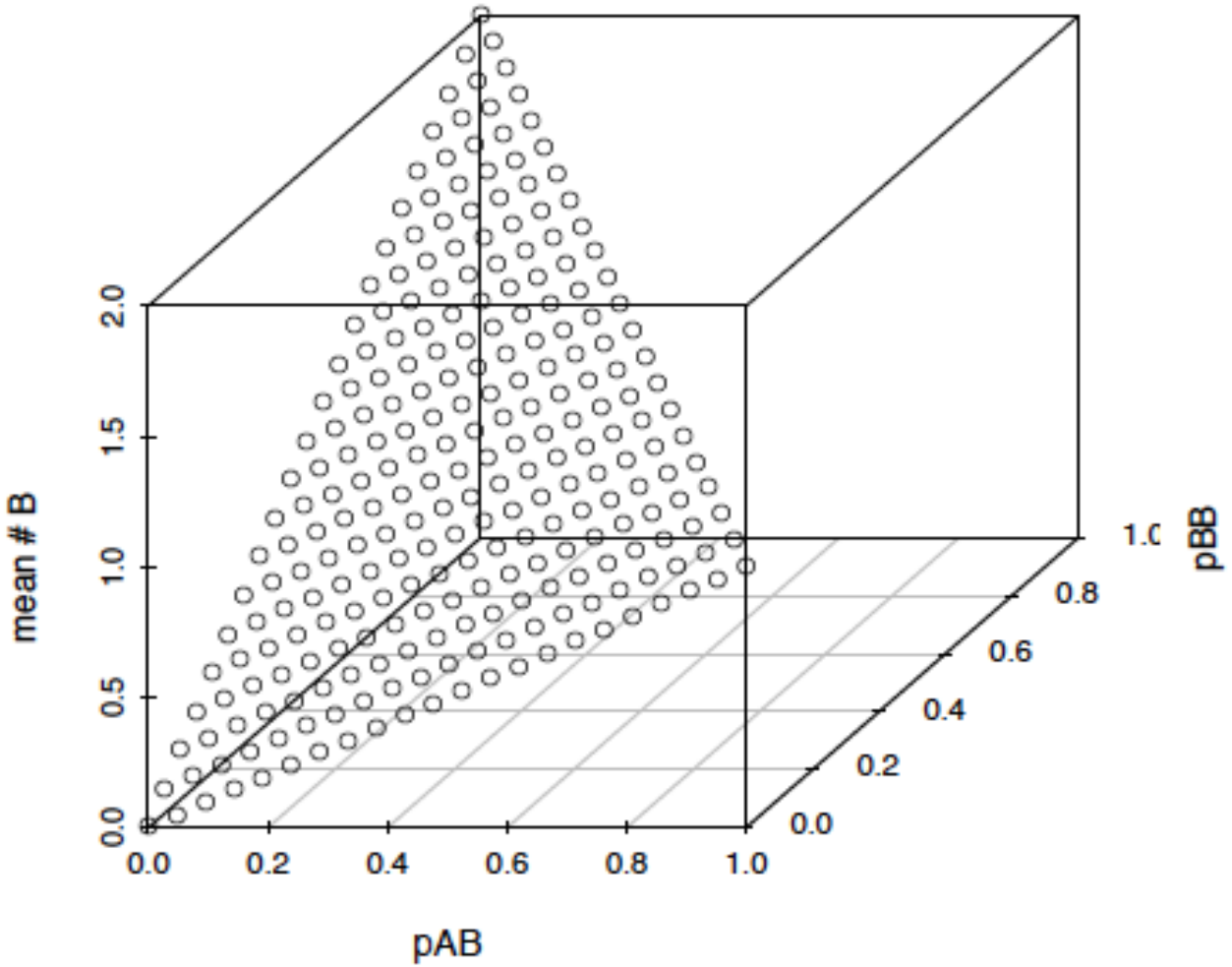

radius of cis search
5000
50000
250000
genes w/eQTL
at FDR <= 0.05
2250
2000
1750
0.025
0.050
0.075
0.100
lower bound on MAF
npc
30
25
20
15
10
5


5000
50000
250000
genes w/eQTL
at FDR <= 0.05
2250
2000
1750
10
20
30
# PC removed
factor(MAF)
0.005
0.01
0.025
0.03
0.05
0.075
0.1

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

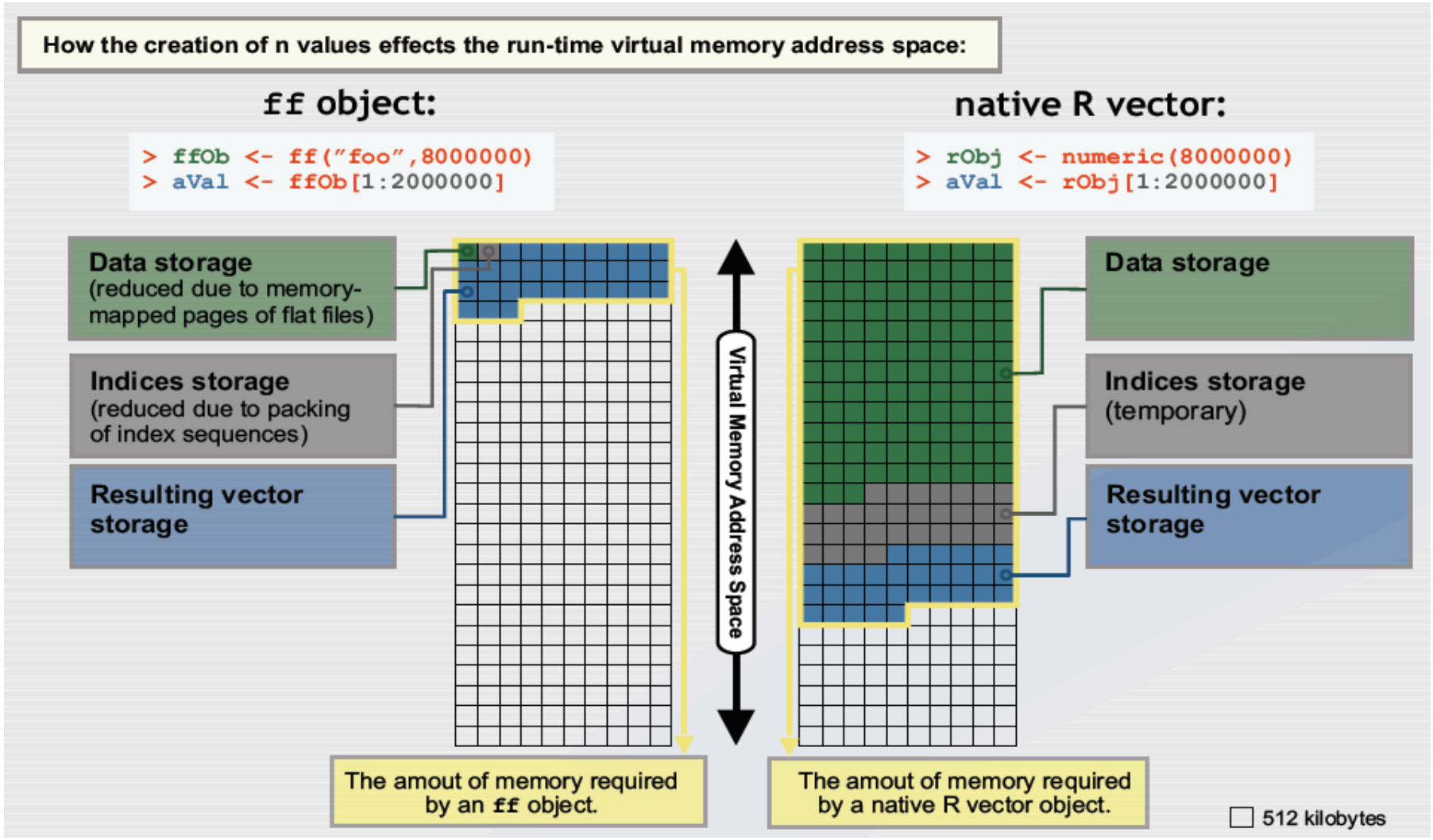
Representing (uncertain) SNP genotypes: David Clayton's byte-sized encoding



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



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

ort

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

Bernd Bischl, Michel Lang,
Olaf Mersmann,
Jörg Rahnenführer, Claus Weihs

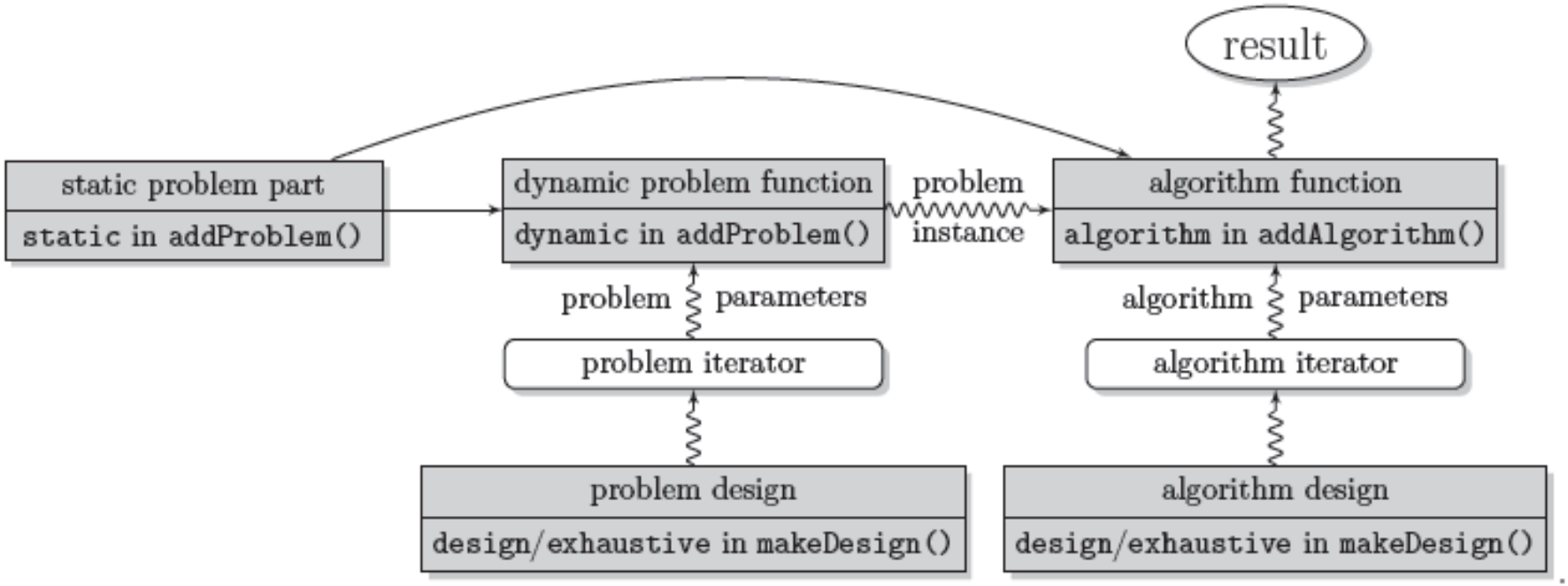
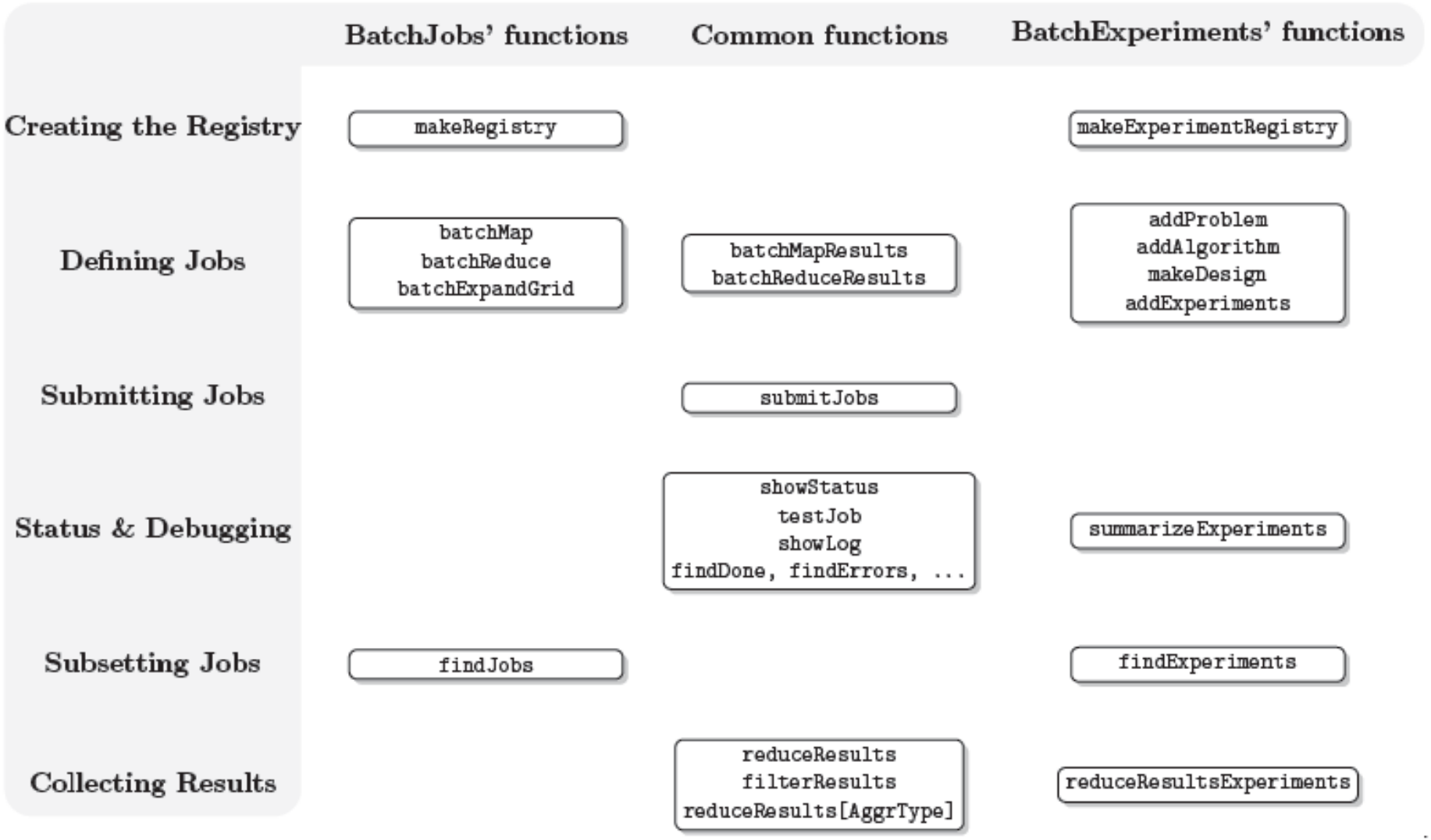


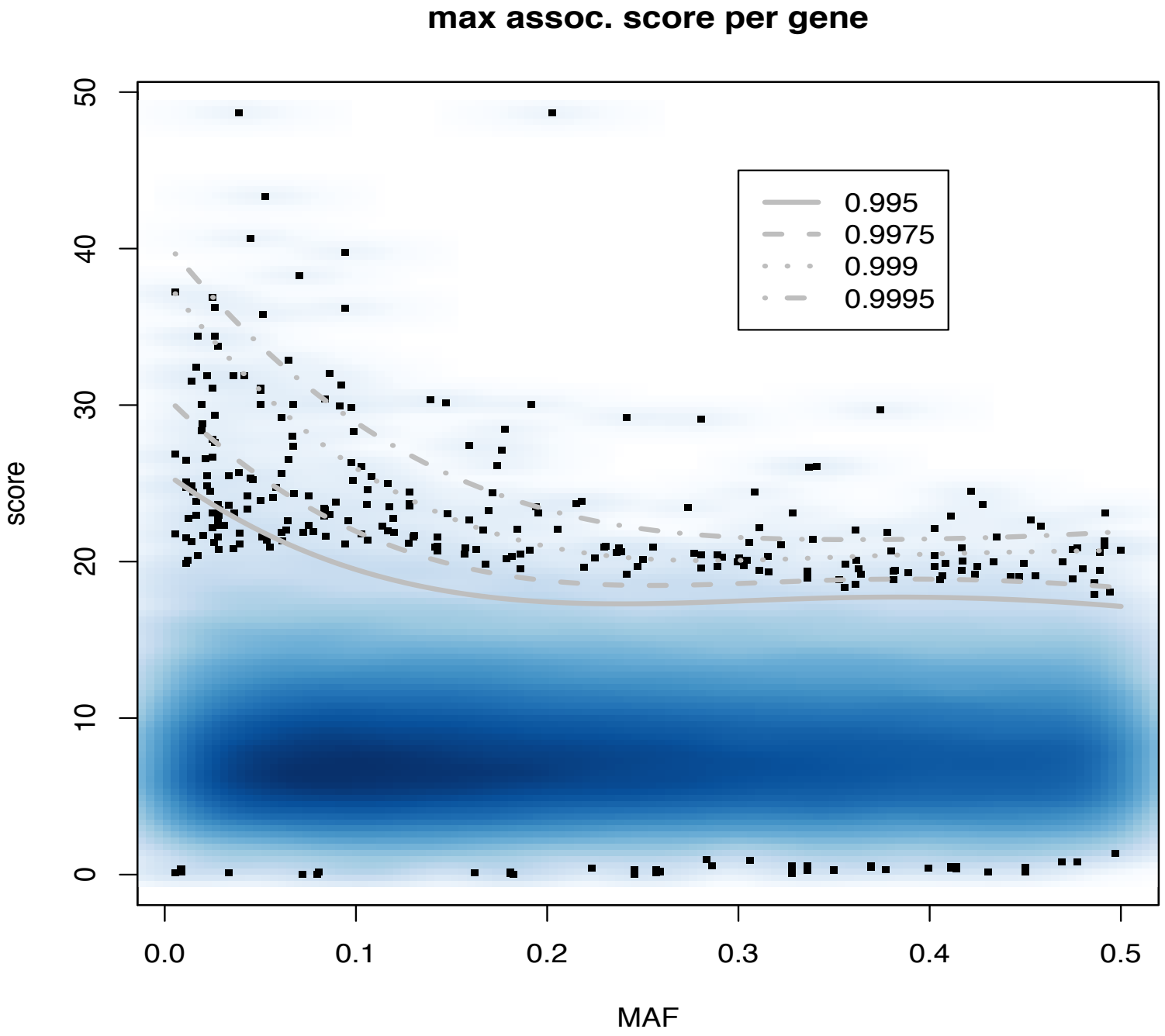
Figure 2: Relationship of **BatchExperiment** functions. Grey rectangulars require user input. White boxes represent internal functions. A straight arrow stands for direct passing of the object or function, a squiggly line denotes passing of the evaluated result.


BatchJobs' functions
Common functions
BatchExperiments' functions
Creating the Registry
makeRegistry
makeExperimentRegistry
Defining Jobs
batchMap
batchReduce
batchExpandGrid
batchMapResults
batchReduceResults
addProblem
addAlgorithm
makeDesign
addExperiments
Submitting Jobs
submitJobs
Status & Debugging
showStatus
testJob
showLog
findDone, findErrors, ...
summarizeExperiments
Subsetting Jobs
findJobs
findExperiments
Collecting Results
reduceResults
filterResults
reduceResults[AggrType]
reduceResultsExperiments

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:
      seqnames              ranges strand |           snpid    snploc
         <Rle>           <IRanges>  <Rle> |     <character> <integer>
  [1]    chr17 [73127217, 73127217]      * |      rs3736075  73127217
  [2]    chr16 [  641445,   641445]      * |     rs35585285    641445
  [3]     chr2 [65605703, 65605703]      * |      rs2217969  65605703
  [4]    chr16 [24047271, 24047271]      * | chr16:24047271  24047271
  [5]     chr1 [68737383, 68737383]      * |     rs59129922  68737383
      radiusUsed          fdr               probe        excl         maf
       <numeric>    <numeric>         <character> <character> <character>
  [1]      50000 0.0001943635 01YintXoV6ek6AxLqA           0        0.05
  [2]      50000 0.0003236246 02KG50KOEI610L36kU           0        0.05
  [3]      50000 0.0000000000 03tSCXVYVVCc.nRPBA           0        0.05
  [4]      50000 0.0000000000 04.BQghp5olySHSl7o           0        0.05
  [5]      50000 0.0000000000 04jcQvhWQopUosop5I           0        0.05
           nperm                 npc   bestfdr         sym
      <character>         <character> <numeric> <character>
  [1]           3 (32,31,21,19)*1.50         0        NT5C
  [2]           3 (32,31,21,19)*1.50         0     WFIKKN1
  [3]           3 (32,31,21,19)*1.50         0      SPRED2
  [4]           3 (32,31,21,19)*1.50         0       PRKCB
  [5]           3 (32,31,21,19)*1.50         0         WLS
  ---
  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
> sum(values(g3)$fdr <= 0.05)
[1] 3261
```


max assoc. score per gene
score
0
10
20
30
40
50
0.0
0.1
0.2
0.3
0.4
0.5
MAF
0.995
0.9975
0.999
0.9995

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