Bioconductor tools for genetics of expression etc.

VJ Carey, Ph.D.
Channing Division of Network Medicine
Brigham and Women's Hospital
Harvard Medical School

Road map

- Some basic concepts and recent literature
- Exercises
  - Feature filtering and eQTL detection with SNP; comparison to GWAS catalog loci
  - Deep DNA sequencing in the vicinity of eQTL (Complete Genomics diversity panel)
  - Transcript variants and allelic imbalance with RNA-seq [NO – see ggtut tut11.pdf section 5]
  - dsQTL: variants associated with DNaseI hypersensitivity

Task 1: Upgrade your packages

source("http://bioconductor.org/scratch-repos/vince.R")

We'll use, among others

GGtools – structures and functions for genetics of expression

genetw12 – backbone with vignette underlying talk

cgdv17 – complete genomics diversity panel

dsQTL – genetic determinants of DNaseI hypersensitivity

Task 2: compute all the objects we'll want to talk about

```
Sweave(system.file("doc/genetw12.Rnw",
    package="genetw12"))
```

Will take 10 mins or so while we go through literature

LETTERS

edited by Jennifer Sills

Retraction

AFTER ONLINE PUBLICATION OF OUR REPORT "GENETIC SIGNATURES OF EXCEPTIONAL LONGEVity in humans" (*1*), we discovered that technical errors in the Illumina 610 array and an inadequate quality control protocol introduced false-positive single-nucleotide polymorphisms (SNPs) in our findings. An independent laboratory subsequently performed stringent quality control measures, ambiguous SNPs were then removed, and resultant genotype data were validated using an independent platform. We then reanalyzed the reduced data set using the same methodology as in the published paper. We feel the main scientific findings remain supported by the available data: (i) A model consisting of multiple specific SNPs accurately differentiates between centenarians and controls; (ii) genetic profiles cluster into specific signatures; and (iii) signatures are associated with ages of onset of specific age-related diseases and subjects with the oldest ages. However, the specific details of the new analysis change substantially from those originally published online to the point of becoming a new report. Therefore, we retract the original manuscript and will pursue alternative publication of the new findings.

PAOLA SEBASTIANI,[1]* NADIA SOLOVIEFF,[1] ANNIBALE PUCA,[2] STEPHEN W. HARTLEY,[1] EFTHYMIA MELISTA,[3] STACY ANDERSEN,[4] DANIEL A. DWORKIS,[3] JEMMA B. WILK,[5] RICHARD H. MYERS,[5] MARTIN H. STEINBERG,[6] MONTY MONTANO,[3] CLINTON T. BALDWIN,[6,7] THOMAS T. PERLS[4]*

[1]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA. [2]IRCCS Multimedica, Milano, Italy; Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate, 20122, Italy. [3]Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA. [4]Section of Geriatrics, Department of Medicine, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. [5]Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA. [6]Departments of Medicine and Pediatrics, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. [7]Center for Human Genetics, Boston University School of Medicine, Boston, MA 02118, USA.

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.T.P.)

Biostatistics (2012), **xx**, **xx**, *pp*. 1–10
doi:10.1093/biostatistics/kxr055

Learning from our GWAS mistakes: from experimental design to scientific method

CHRISTOPHE G. LAMBERT*
Golden Helix Inc., PO Box 10633, Bozeman, MT 59719, USA
lambert@goldenhelix.com

LAURA J. BLACK
College of Business, Montana State University, PO Box 173040, Bozeman, MT 59717-3040, USA and Greer Black Company, PO Box 3607, Bozeman, MT 59772-3607, USA

SUMMARY

Many public and private genome-wide association studies that we have analyzed include flaws in design, with avoidable confounding appearing as a norm rather than the exception. Rather than recognizing flawed research design and addressing that, a category of quality-control statistical methods has arisen to treat only the symptoms. Reflecting more deeply, we examine elements of current genomic research in light of the traditional scientific method and find that hypotheses are often detached from data collection, experimental design, and causal theories. Association studies independent of causal theories, along with

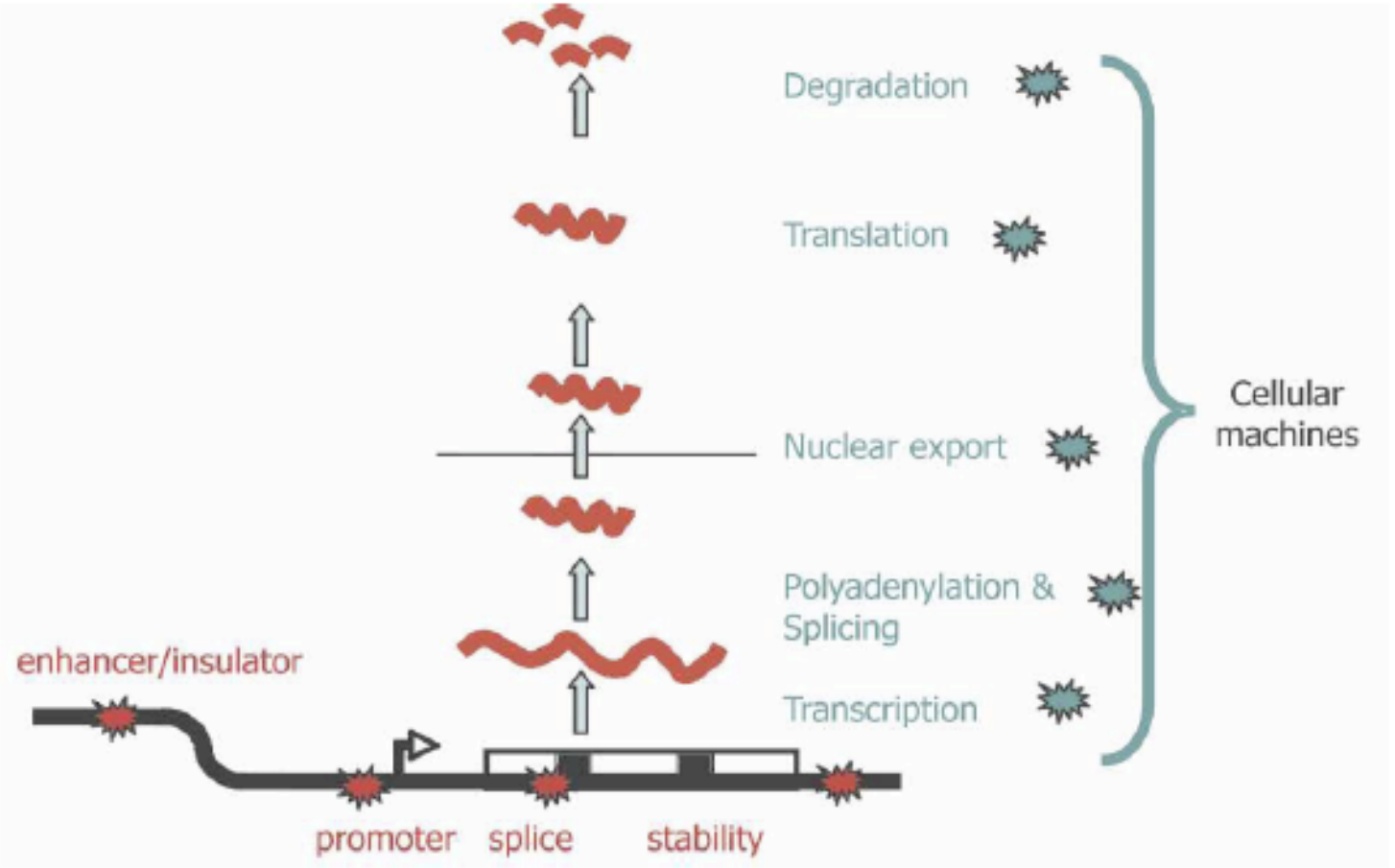
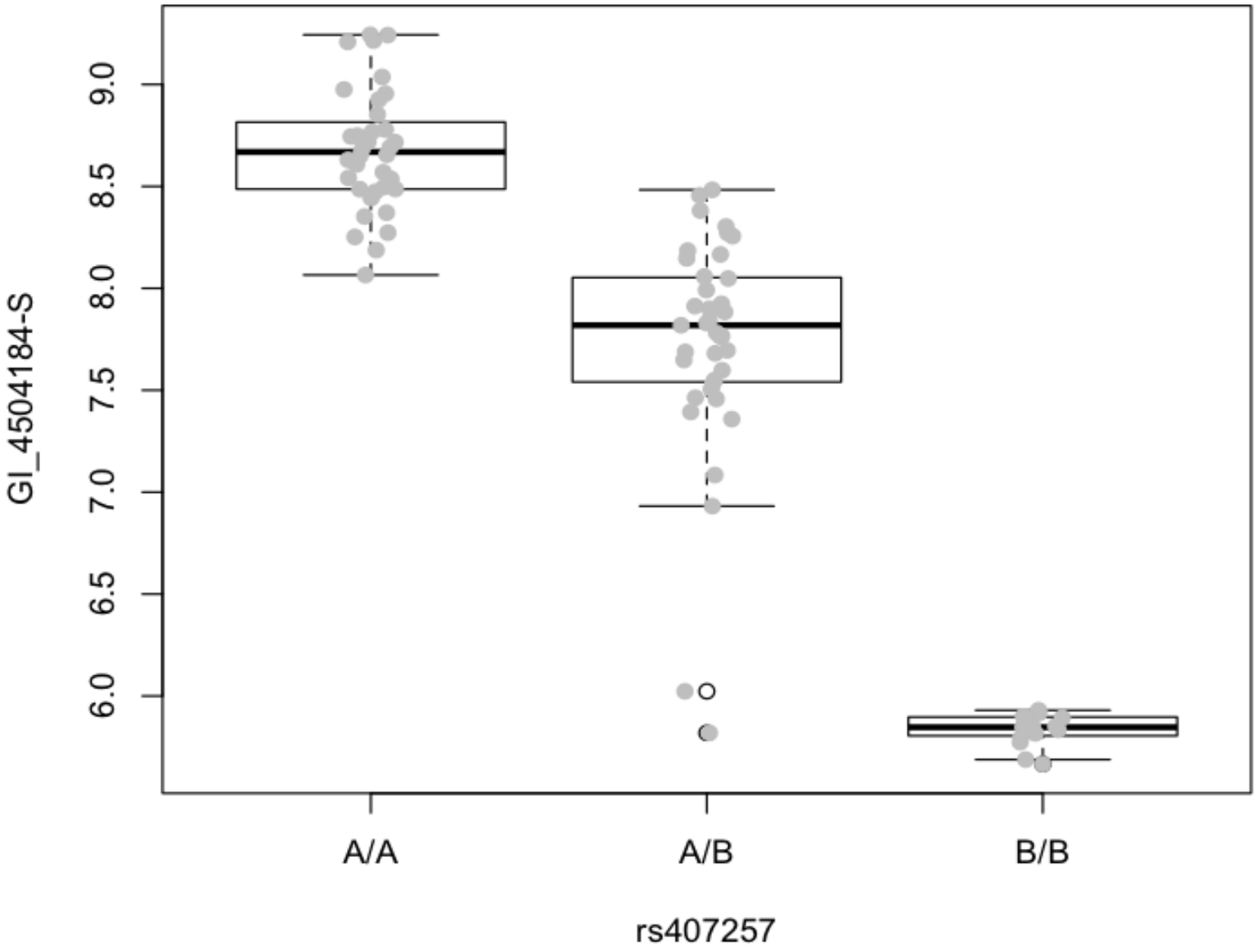


Figure 1. Plausible sites of action for genetic determinants of mRNA levels. Genetic variations influencing gene expression may reside within the regulatory sequences, promoters, enhancers, splice sites, and secondary structure motifs of the target gene and so be genetically in *cis* (red stars), or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

Average expression varies by genotype – why?



Veyrieras et al 2008 PMID 18846210

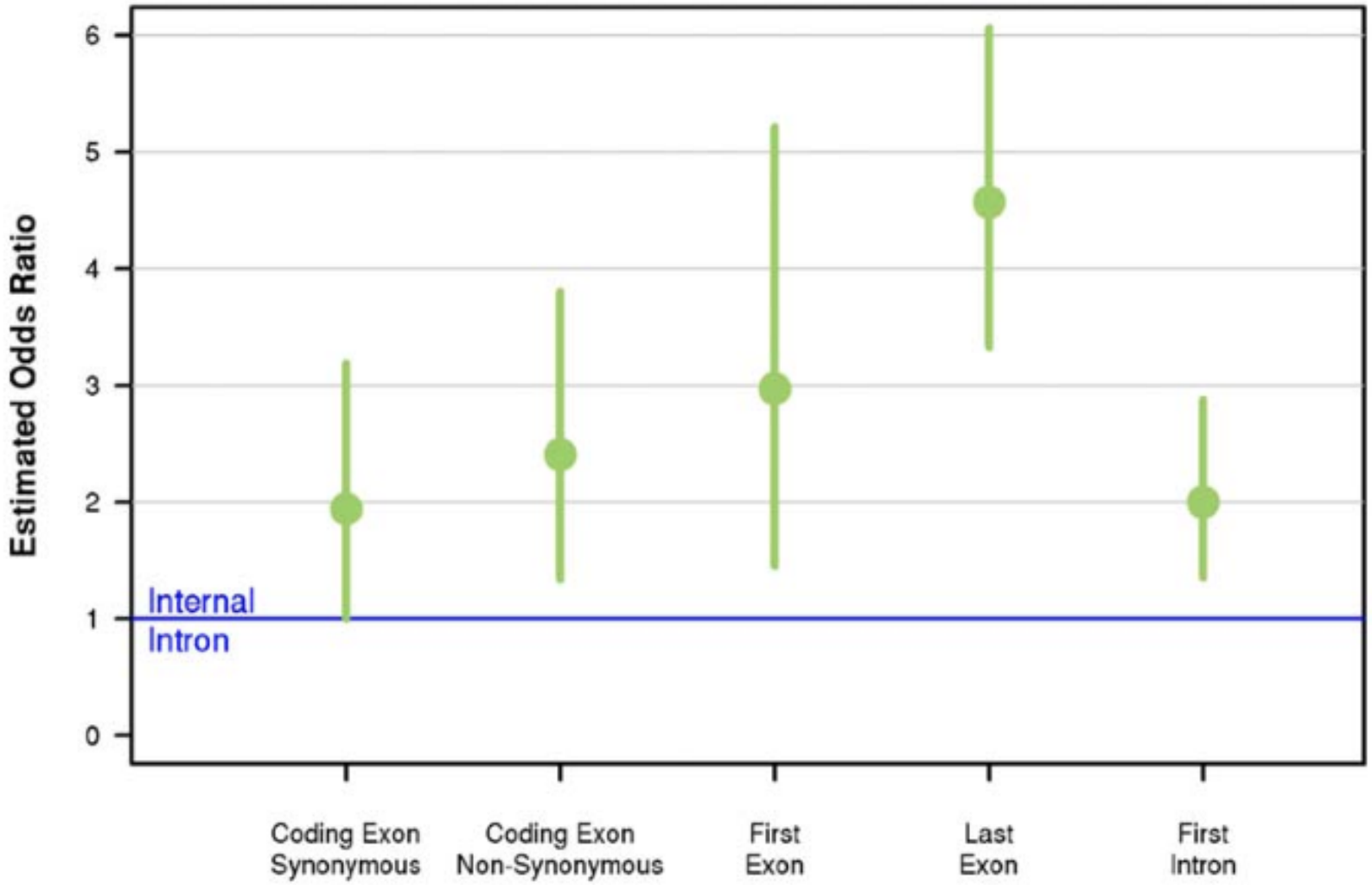


Figure 5. Expression-QTNs are under-represented in coding sequence introns, even after controlling for posit
shows the odds ratios for the probability that a SNP in a particular part of the gene (e.g., coding exon) is inferred to be an
probability for a SNP in an "internal" intron (i.e., an intron within the coding sequence). The odds ratios are estimated using t

Schemata for SNP-associated splicing events (Coulombe-Huntington 2009)

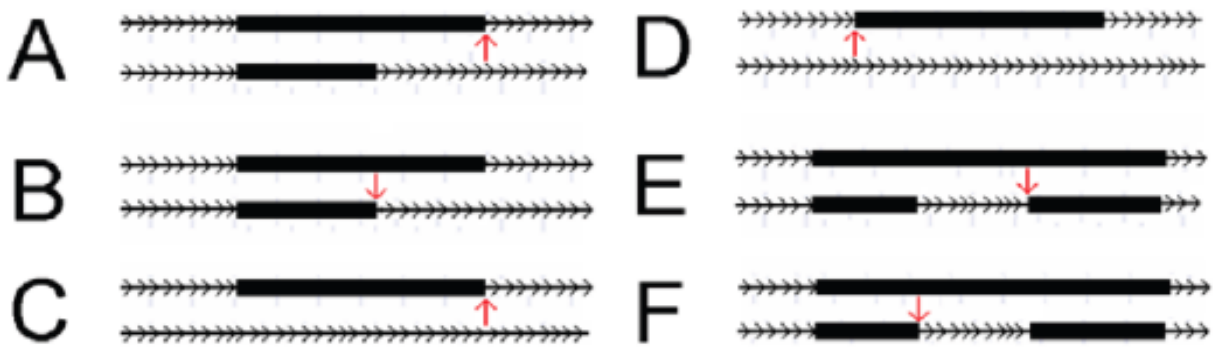


Figure 2. AS type and affected splice-site for SNPs identified in Table 2 and Table 3. The arrow indicates the splice-site affected by the polymorphism. The genes are read from left to right, as indicated by the intersecting arrow heads. The type of AS event and which splice-site is affected is essential to understanding the relation between the probeset expression change and the theoretical efficiency of splicing. In (A,C,D), the correlation should be positive since the use of the splice-site produces a longer transcript, while in (B,E,F), an inverse relation is expected since the use of the splice-site produces a shorter transcript. doi:10.1371/journal.pgen.1000766.g002

PMID 20011102

Gaffney et al. Dissecting the regulatory architecture of gene expression QTLs, Genome Biology 2012 (PMID 22293038)

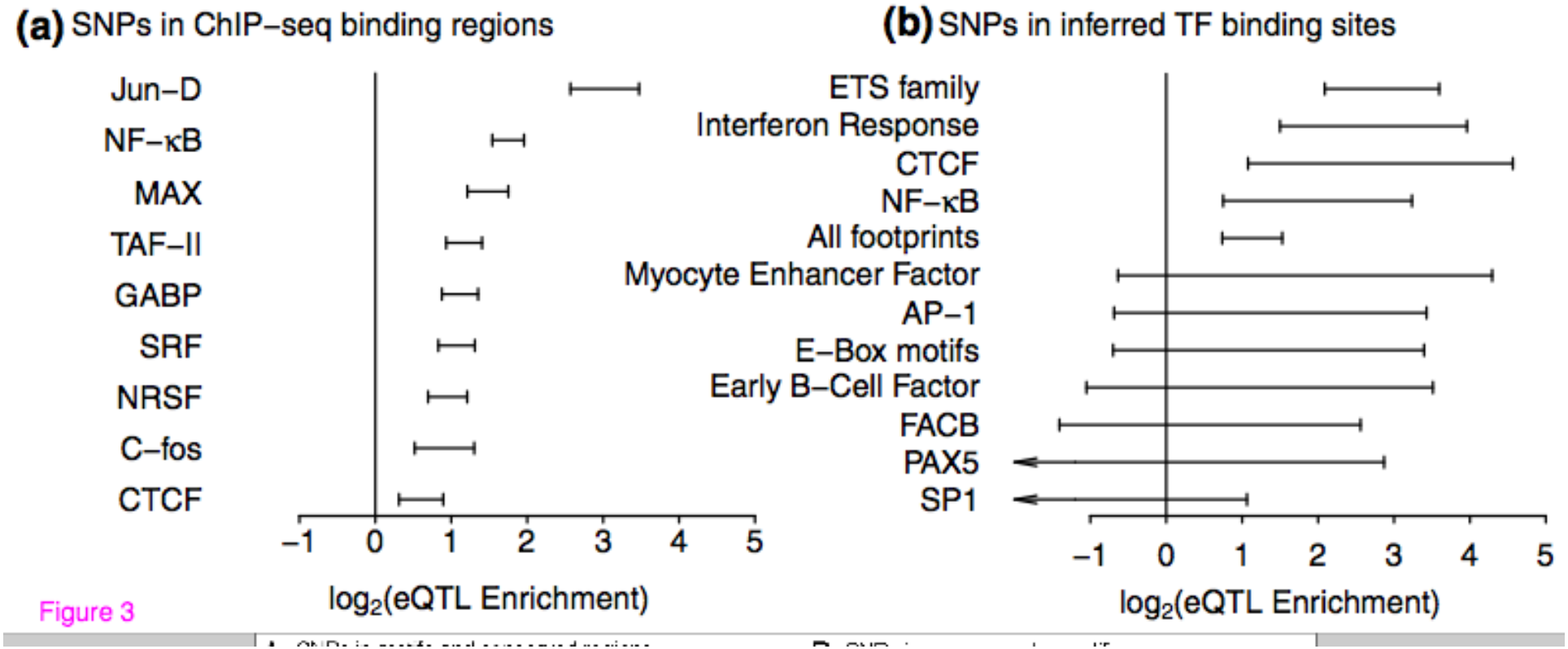
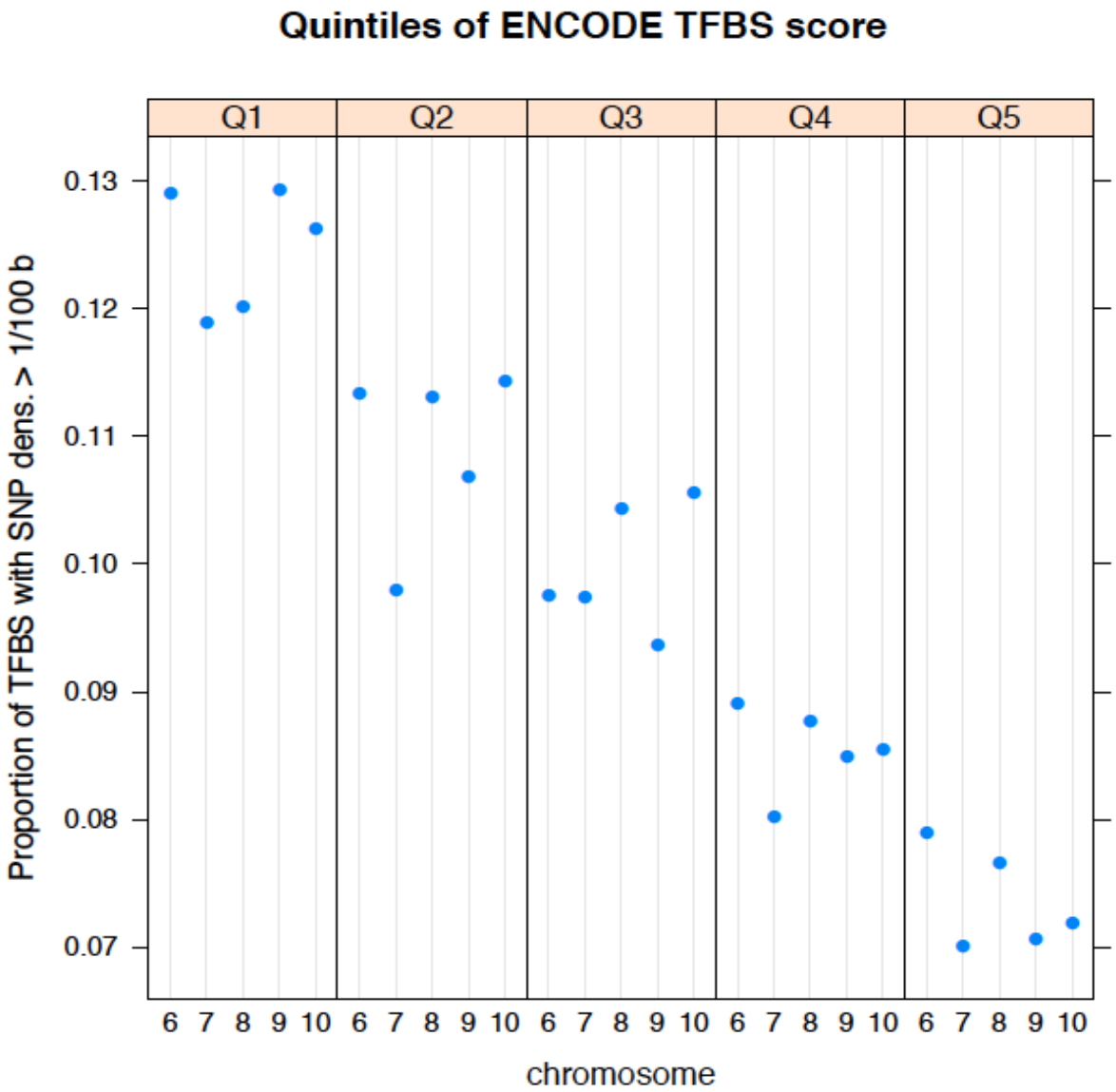


Figure 3

From *Ranges paper in progress; a demonstrative calculation – upshot is that there are covariates of TFBS:X relationships whose accommodation may be important



LETTER

doi:10.1038/nature10808

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

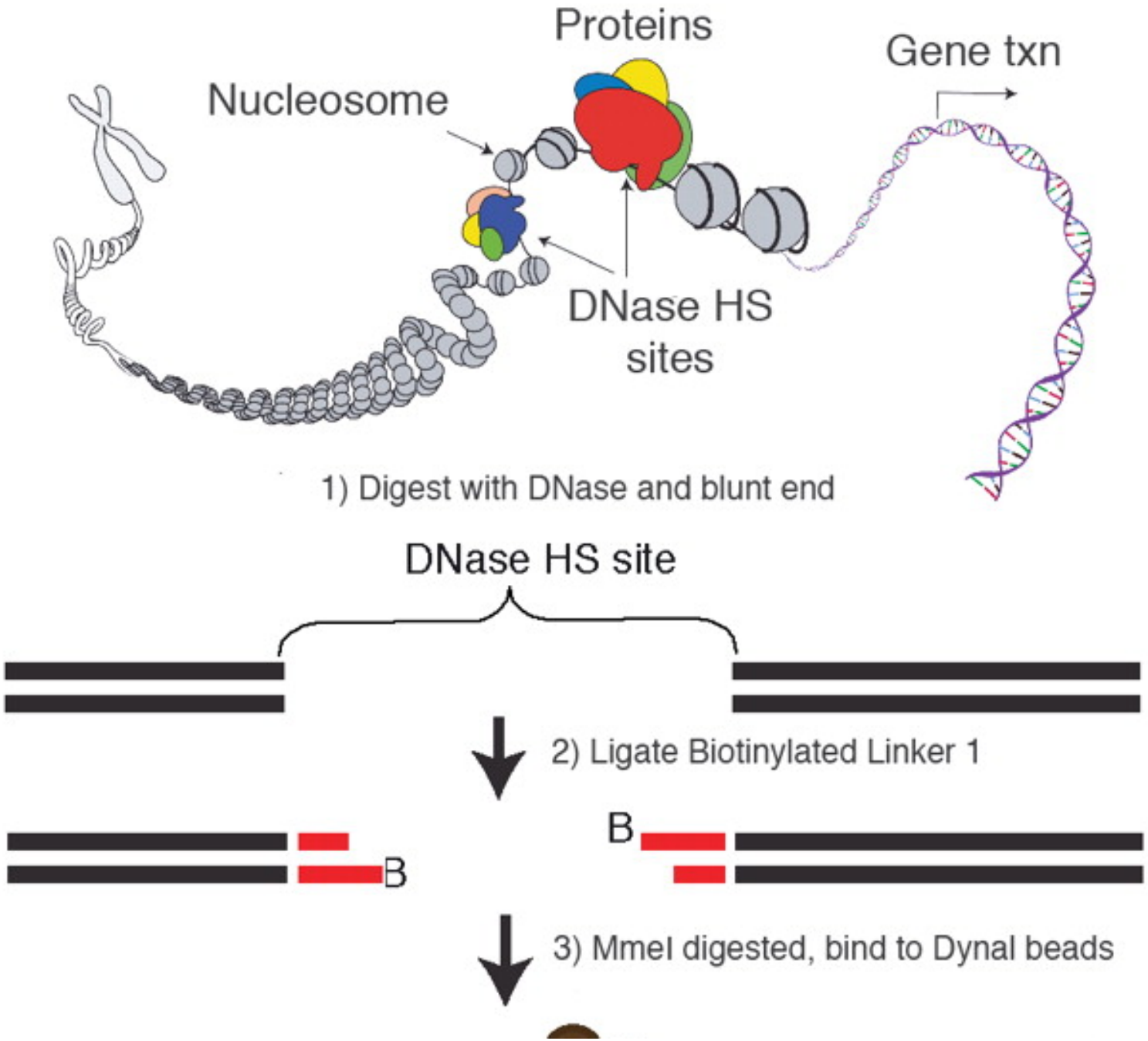
Jacob F. Degner[1,2]*, Athma A. Pai[1]*, Roger Pique-Regi[1]*, Jean-Baptiste Veyrieras[1,3], Daniel J. Gaffney[1,4], Joseph K. Pickrell[1], Sherryl De Leon[4], Katelyn Michelini[4], Noah Lewellen[4], Gregory E. Crawford[5,6], Matthew Stephens[1,7], Yoav Gilad[1] & Jonathan K. Pritchard[1,4]

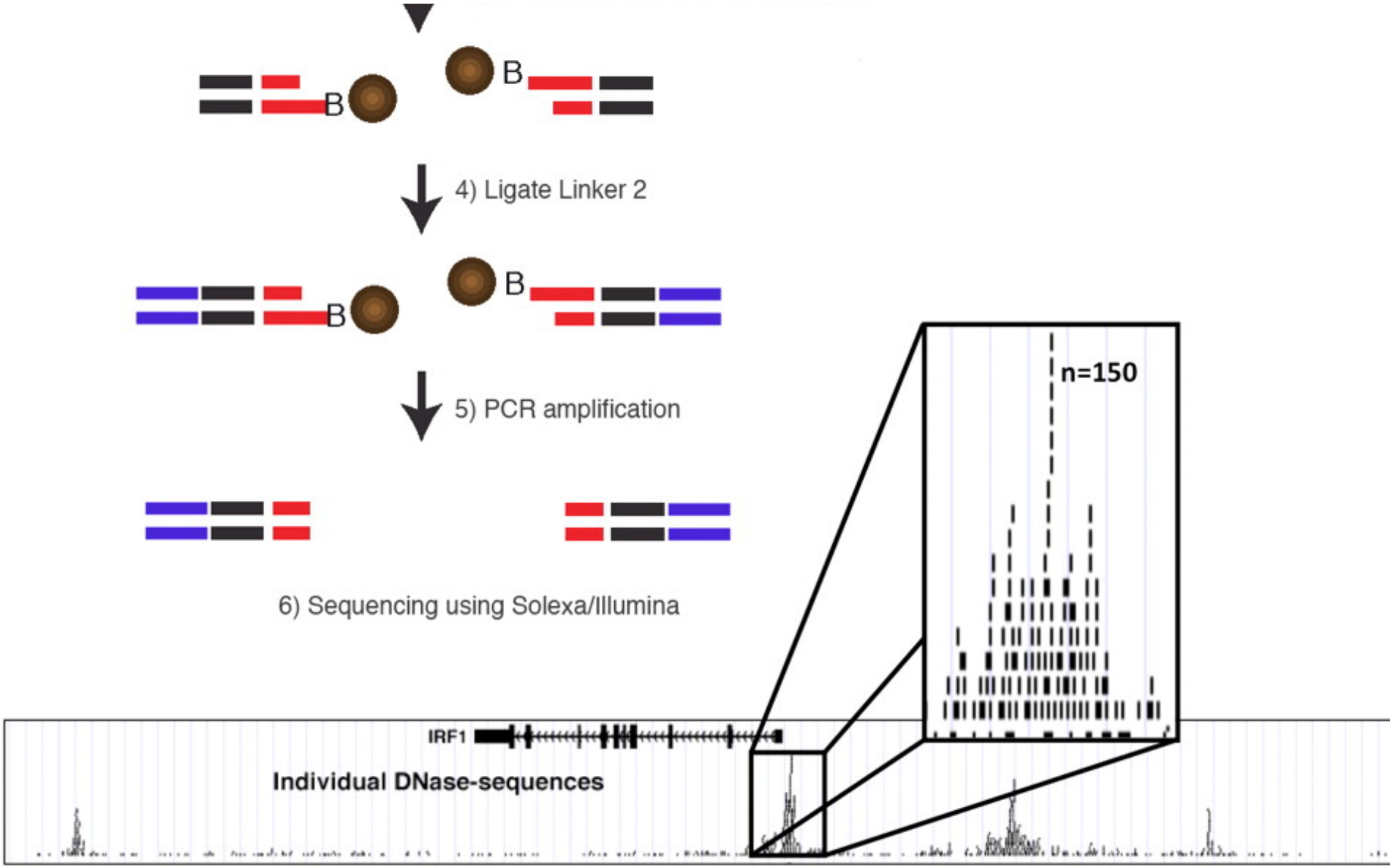
The mapping of expression quantitative trait loci (eQTLs) has emerged as an important tool for linking genetic variation to changes in gene regulation[1–5]. However, it remains difficult to identify the causal variants underlying eQTLs, and little is known about the regulatory mechanisms by which they 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

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

PMID 22307276


Proteins
Gene txn
Nucleosome
DNase HS
sites
1) Digest with DNase and blunt end
DNase HS site
2) Ligate Biotinylated Linker 1
B
B
3) MmeI digested, bind to Dynal beads


B
B
B
B
4) Ligate Linker 2
5) PCR amplification
6) Sequencing using Solexa/Illumina
n=150
IRF1
Individual DNase-sequences

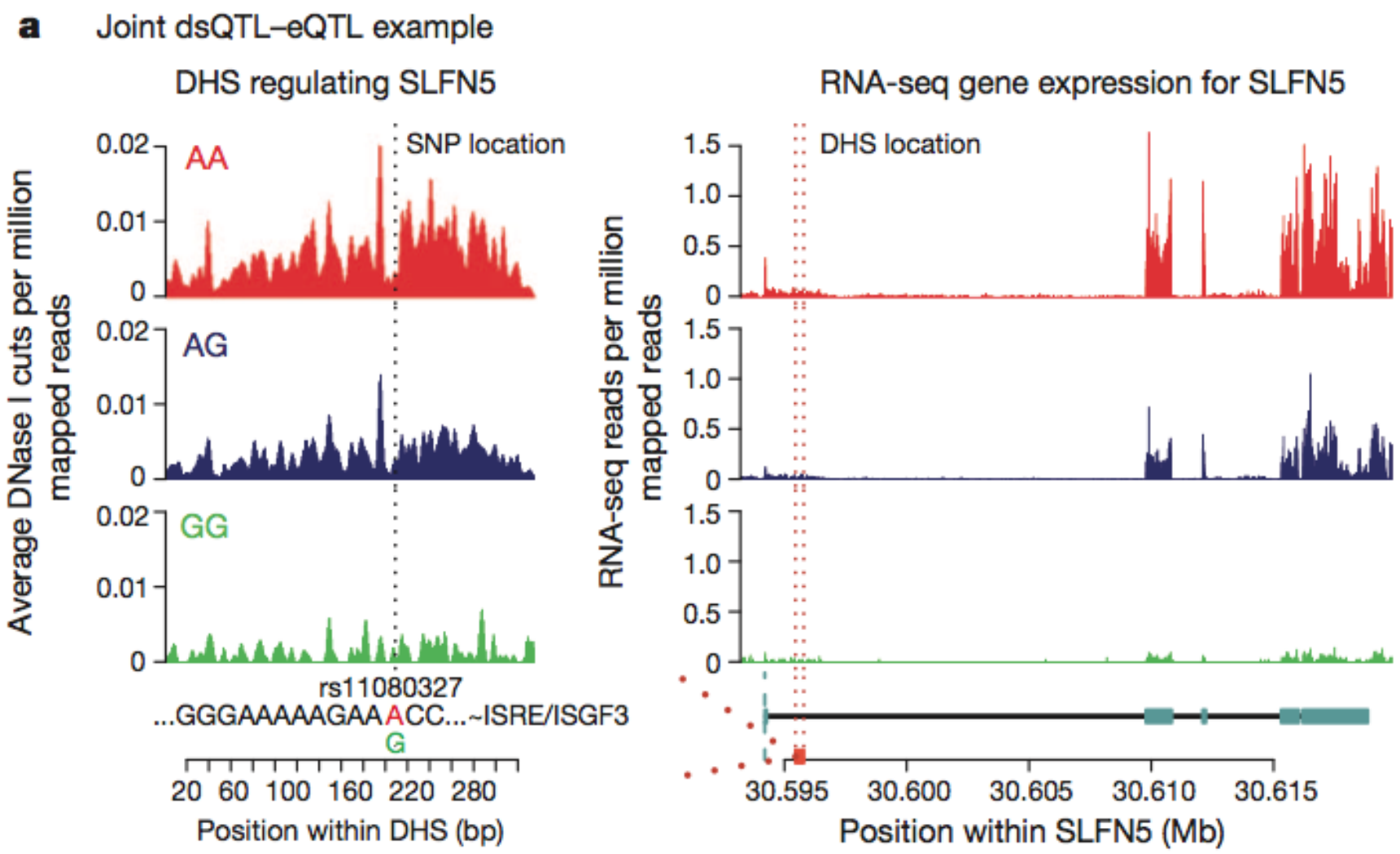
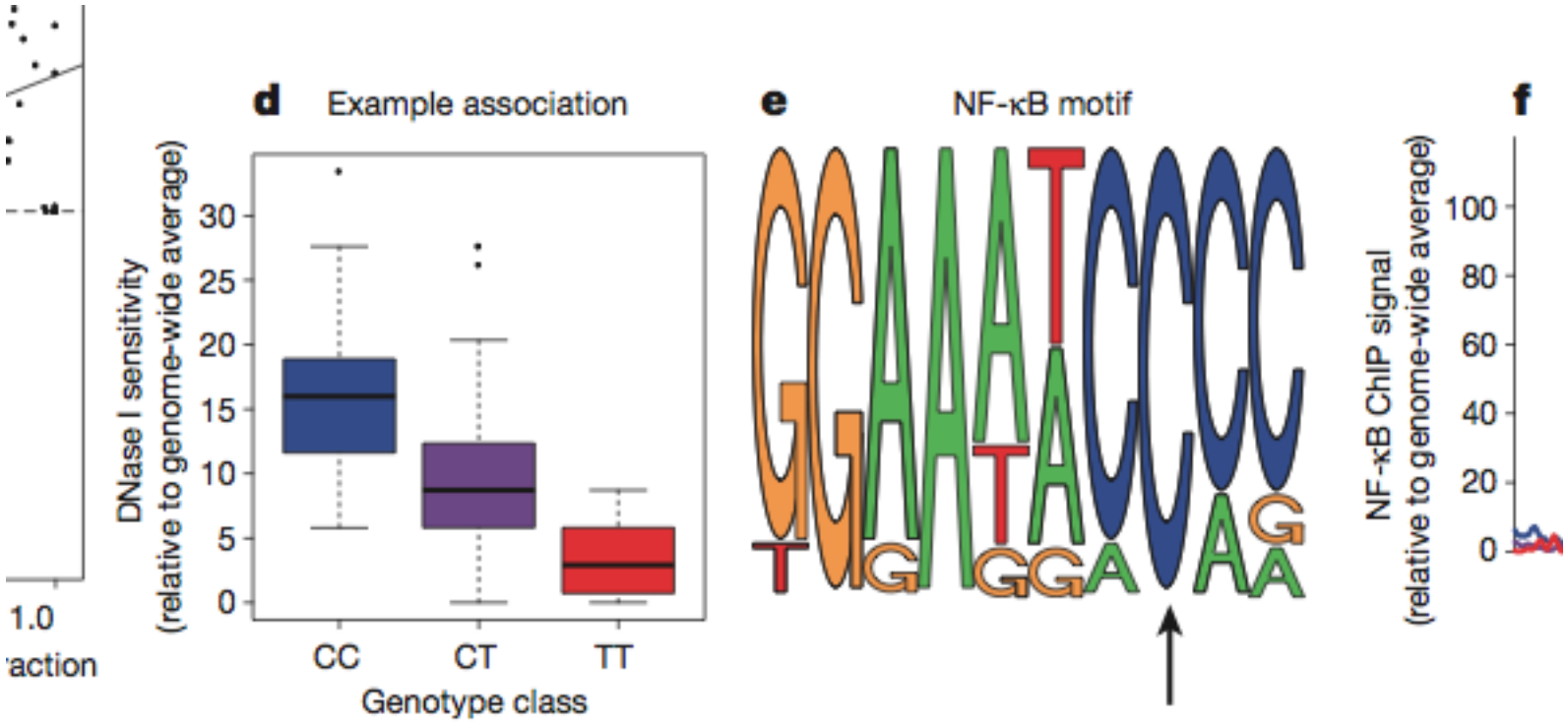


Figure 3 | Relationship between dsQTLs and eQTLs. a, Example of a dsQTL SNP that is also an eQTL for the gene *SLFN5*. The SNP disrupts an interferon- (right) measurer genotype at the p



TLs and a typical example. DNase I cut rates in 100-bp 40-kb (black) regions centred dsQTL (rs4953223). The black line indicates the positio **d**, Box plot showing that rs4953223 is strongly associat accessibility ($P = 3 \times 10^{-13}$). **e**, The T allele, which is a

Upshots

- Basic theory of structural impacts of DNA variation (in *cis*) on expression variation (Williams cartoon) has some observational confirmation
  - Expression-associated variants more common in exons (with some trend in location) than internal introns
  - Impacts of DNA variants on splicing events have been observed
  - Enrichment of eQTL among SNP located in insulators, enhancers; effects on chromatin accessibility
- What about phenotypic impacts?

OPEN ACCESS Freely available online

PLoS GENETICS

Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations

Alexandra C. Nica[1,2], Stephen B. Montgomery[1,2], Antigone S. Dimas[1,2], Barbara E. Stranger[1,3], Claude Beazley[1], Inês Barroso[1], Emmanouil T. Dermitzakis[1,2]*

1 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **2** Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland, **3** Harvard Medical School/Brigham and Women's Hospital, Boston, Massachusetts, United States of America

Abstract

The recent success of genome-wide association studies (GWAS) is now followed by the challenge to determine how the reported susceptibility variants mediate complex traits and diseases. Expression quantitative trait loci (eQTLs) have been implicated in disease associations through overlaps between eQTLs and GWAS signals. However, the abundance of eQTLs and the strong correlation structure (LD) in the genome make it likely that some of these overlaps are coincidental and not driven by the same functional variants. In the present study, we propose an empirical methodology, which we call Regulatory Trait Concordance (RTC) that accounts for local LD structure and integrates eQTLs and GWAS results in order to reveal the subset of association signals that are due to *cis* eQTLs. We simulate genomic regions of various LD patterns with both a single or two causal variants and show that our score outperforms SNP correlation metrics, be they statistical (r^2) or historical (D'). Following the observation of a significant abundance of regulatory signals among currently published GWAS

PMID 20369022

Table 1. Candidate *cis* results.

GWAS SNP	Complex Trait	Gene	RTC	Chr
rs2064689	Crohn's disease	WDR78	1	1
rs3129934	Multiple sclerosis	HLA-DRB1	1	6
rs2188962	Crohn's disease	SLC22A5	1	5
rs1015362	Burning and freckling	TRPC4AP	1	20
rs2735839	Prostate cancer	C19orf48	1	19
rs6830062	Height	LCORL	1	4
rs2242330	Parkinsons disease	TMPRSS11A	1	4
rs6441961	Celiac disease	LIMD1	0.92	3
rs660895	Rheumatoid arthritis	PSMB9	0.91	6
rs9652490	Essential tremor	ILMN_111363	0.91	15
rs1397048	Hemostatic factors	OR8H2	0.91	11
rs3825932	Type 1 diabetes	CTSH	0.91	15
rs2395185	Ulcerative colitis	ILMN_29412	0.9	6

Candidate genes (RTC Score >= 0.9) for *cis* regulatory mediated GWAS effects. The higher the score, the more likely it is that the GWAS SNP and the eQTL for the gene shown are tagging the same functional variant.

doi:10.1371/journal.pgen.1000895.t001

Scoring scheme for determining causal regulatory effects

We assess the likelihood of a shared functional effect between a GWAS SNP and an eQTL by quantifying the change in the statistical significance of the eQTL after correcting for the genetic effect of the GWAS SNP. We redo the SRC association of the eQTL genotype with the residuals from the standard LR of the "corrected-for" SNP against normalized expression values. We account for the LD structure in each hotspot interval separately by ranking ($Rank_{GWAS\ SNP}$) the impact on the eQTL (quantified by the adjusted association P-value after correction) of the GWAS SNP correction to that of correcting for all other SNPs in the same interval. By taking into account the total number of SNPs in the interval (N_{SNPs}), we can compare this ranking across different genes and intervals. For this purpose we define the regulatory trait concordance (RTC) Score ranked below ranging from 0 to 1, with values closer to 1 indicating causal regulatory effects.

$$RTC = \frac{N_{SNPs} - Rank_{GWAS\,SNP}}{N_{SNPs}}$$

Summary

- Genome-wide studies of impacts of DNA variation are exciting (acceptance of longevity signatures) and tricky (retraction of longevity signatures)
- Good experimental design is essential, but workflows are elaborate; real-time aspects may induce loss of design control
- Many decisions on data filtering and choice of analysis have no *a priori* justification, so sensitivities of findings to assumptions and optional choices should be assessed

A series of exercises, informally

- Represent and provide interfaces to the expression + genotype data on human cohorts so that effective eQTL searches can be conducted and statistically calibrated
- Relate published results on GWAS to eQTL that you identify
- Investigate arbitrary variants obtained through deep DNA- [and RNA-sequencing for information on individual contexts of eQTL, and on allelic imbalance in transcription]
- Connect normalized DNase-seq results with SNP genotyping to identify dsQTL

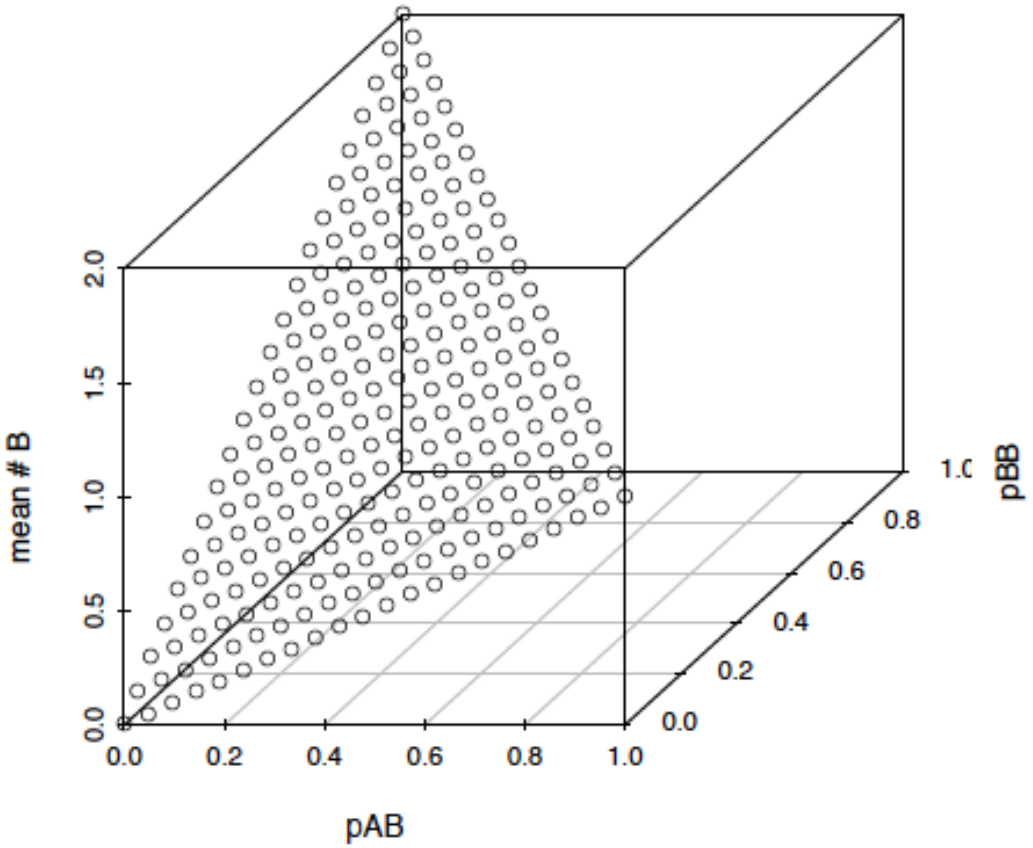
Representation with a package: `help(package="GGdata")`

```
 Information on package 'GGdata'

Description:

Package:       GGdata
Title:         all 90 hapmap CEU samples, 47K expression, 4mm SNP
Description:   data exemplars dealing with hapmap SNP reports, GWAS,
               etc.
Version:       1.0.17
Author:        VJ Carey <stvjc@channing.harvard.edu>
Maintainer:    VJ Carey <stvjc@channing.harvard.edu>
biocViews:     ExperimentData, HapMap
Depends:       R (>= 2.12.0), methods, Biobase (>= 2.5.5), GGBase,
               snpStats, illuminaHumanv1.db, AnnotationDbi
Enhances:      GGtools
LazyLoad:      yes
License:       LGPL
Built:         R 2.15.0; ; 2011-11-17 00:19:10 UTC; unix

Index:

hmceuB36                representations of HapMap snp data + expression
                        data
```

After `suppressPackageStartupMessages(library(Ggtools))`

```
> g22 = getSS("GGdata", "22")
> g22
SnpMatrix-based genotype set:
number of samples:  90
number of chromosomes present:  1
annotation: illuminaHumanv1.db
Expression data dims: 47293 x 90
Total number of SNP: 54786
Phenodata: An object of class "AnnotatedDataFrame"
  sampleNames: NA06985 NA06991 ... NA12892 (90
  total)
  varLabels: famid persid ... male (7 total)
  varMetadata: labelDescription
```

D. Clayton's snpStats bytecode for (potentially) uncertain calls



Expression and genotype data on the CEPH CEU HapMap cell lines

```
> g22 = getSS("GGdata", "22")
> exprs(g22)[1:5,1:5]
                NA06985   NA06991   NA06993   NA06994   NA07000
GI_10047089-S  5.983962  5.939529  5.912270  5.891347  5.906675
GI_10047091-S  6.544493  6.286516  6.244446  6.277397  6.330893
GI_10047093-S  9.905235 10.353804 10.380972  9.889223 10.155686
GI_10047099-S  7.993935  7.593970  8.261215  6.598430  6.728085
GI_10047103-S 11.882265 12.204753 12.249708 11.798415 12.015252
> as(smList(g22)[[1]][1:5, 1:5], "character")
        rs11089130 rs738829 rs7510853 rs10154488 rs915674
NA06985 "B/B"      "B/B"    "B/B"     "A/A"      "A/B"
NA06991 "A/B"      "B/B"    "B/B"     "A/A"      "B/B"
NA06993 "NA"       "B/B"    "B/B"     "A/A"      "B/B"
NA06994 "A/B"      "B/B"    "B/B"     "A/A"      "A/B"
NA07000 "B/B"      "B/B"    "B/B"     "A/A"      "B/B"
```

Supporting searches for genes possessing *cis* eQTL

```
> args(best.cis.eQTLs)
function (smpack = "GGdata", rhs = ~1, folderstem = "cisScratch",
    radius = 50000, shortfac = 100, chrnames = as.character(1:22),
    smchrpref = "", gchrpref = "", schrpref = "ch",
    geneApply = lapply,
    geneannopk = "illuminaHumanv1.db",
    snpannopk = "SNPlocs.Hsapiens.dbSNP.20100427",
    smFilter = function(x) nsFilter(MAFfilter(x, lower = 0.05),
        var.cutoff = 0.97), nperm = 2)
NULL
```

By default some very sharp filtering is performed.

A new filter and an initial search

```
fil.75_1 = function(x) regressOut(x, ~male)
fil.75_2 = function(x) clipPCs(x, 1:10)
fil.75_3 = function(x) MAFfilter(x, lower=0.05)
fil.75 = function(x) nsFilter(
     fil.75_1( fil.75_2 (fil.75_3(x))),
     var.cutoff=.75)
library(parallel)
options(mc.cores=parallel:::detectCores())
set.seed(1234)
b.75a <- best.cis.eQTLs(smpack = "GGdata", rhs = ~1,
   chrnames = "22",
      geneApply = mclapply, smFilter = fil.75)
```

Took 190 seconds on the student machine Monday morning.

What happened

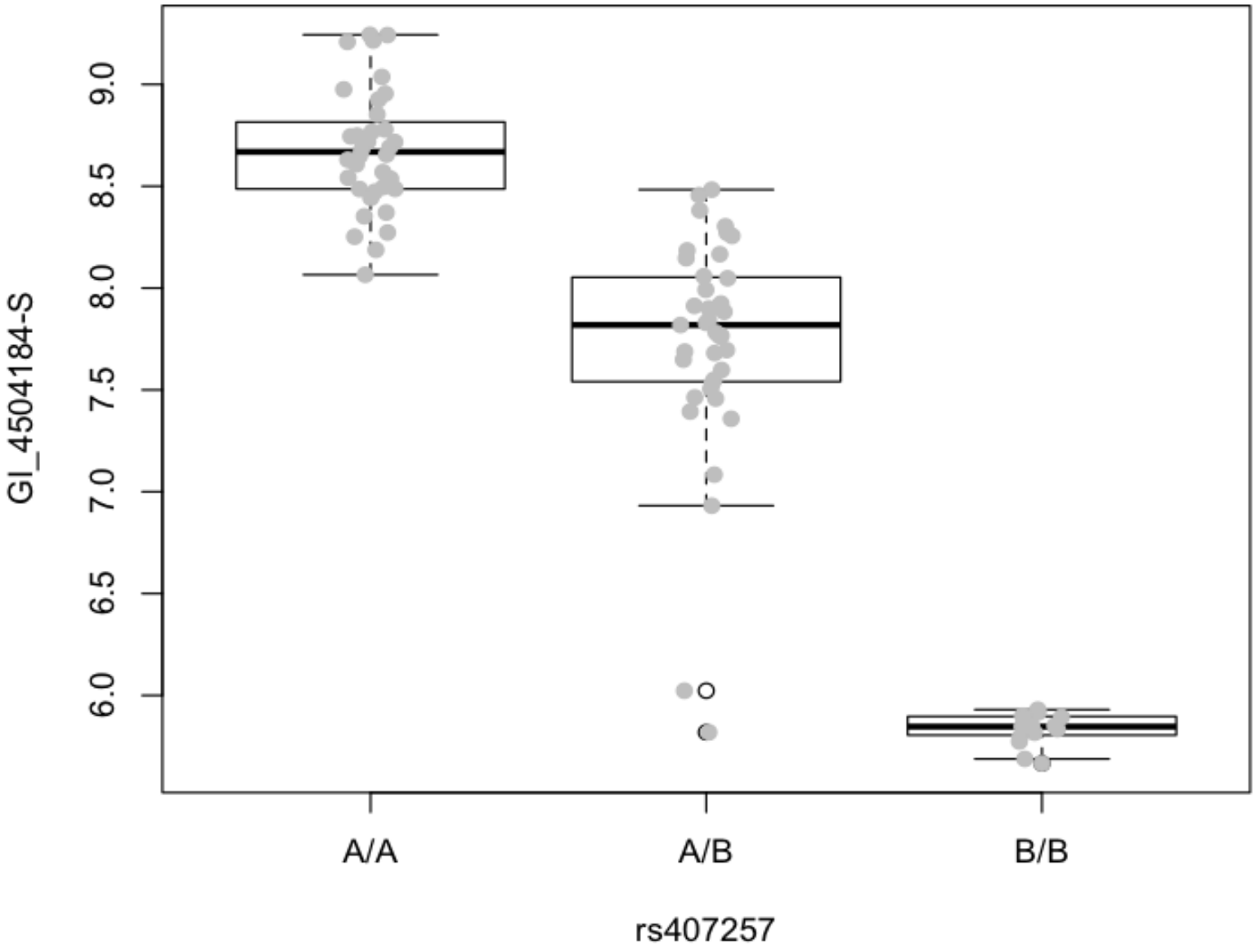
- chr22 genotype data on all 90 cell lines was extracted; SNP with MAF < 0.05 removed
- Expression data were filtered nonspecifically to probes with IQR in top quartile of all probes, then restricted to chr22
- All SNP x expression association tests were carried out, retaining score statistics, with gender covariate
- The best *cis* association (default radius 50kb) **per gene** was extracted
- Expression values permuted against genotypes twice, and plug-in estimates of FDR for the per-gene hypotheses "gene g has a *cis* eQTL" are obtained – these FDR are for the one-chromosome search; the procedure can produce whole-genome inferences, but these take more time

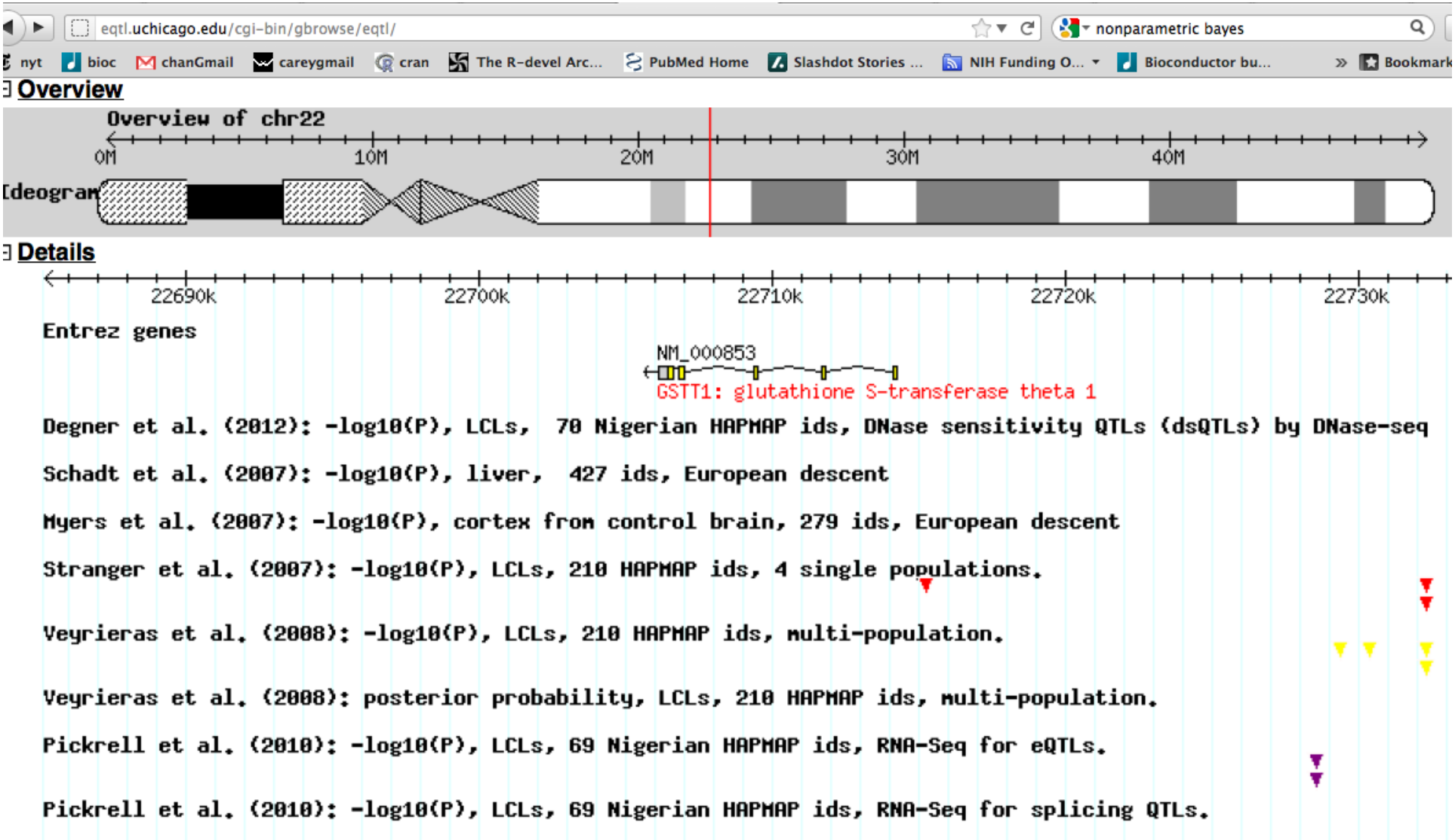
```
> b.75a
GGtools mcwBestCis instance.  The call was:
best.cis.eQTLs(smpack = "GGdata", rhs = ~1, chrnames = "22",
    geneApply = mclapply, smFilter = fil.75)
Best loci for 123 are recorded.
Top  4 probe:SNP combinations:
GRanges with 4 ranges and 5 elementMetadata cols:
               seqnames              ranges strand |     score       snpid
                  <Rle>           <IRanges>  <Rle> | <numeric> <character>
  GI_4504184-S       22 [24326141, 24434284]      * |     65.96    rs407257
  GI_8923587-S       22 [45655081, 45787834]      * |     54.66    rs738177
  GI_7262293-S       22 [51013450, 51116607]      * |     52.34   rs6151429
  GI_6005825-S       22 [43215772, 43461184]      * |     49.02   rs2038058
                  snploc radiusUsed       fdr
               <integer>  <numeric> <numeric>
  GI_4504184-S  24346550      50000         0
  GI_8923587-S  45731539      50000         0
  GI_7262293-S  51063477      50000         0
  GI_6005825-S  43334295      50000         0
  ---
  seqlengths:
         22
   51116607
====
use chromsUsed(), fullreport(), etc. for additional information.
```

Use sum(fdr(b.75a) <= 0.05) to count the number of genes with *cis* eQTL at FDR 0.05.

```
> fullreport(b.75a)[1:10]
GRanges with 10 ranges and 5 elementMetadata cols:
                seqnames              ranges strand |     score       snpid
                   <Rle>           <IRanges>  <Rle> | <numeric> <character>
   GI_4504184-S       22 [24326141, 24434284]      * |     65.96    rs407257
   GI_8923587-S       22 [45655081, 45787834]      * |     54.66    rs738177
   GI_7262293-S       22 [51013450, 51116607]      * |     52.34   rs6151429
   GI_6005825-S       22 [43215772, 43461184]      * |     49.02   rs2038058
  GI_25092724-S       22 [42854343, 42965829]      * |     43.75  rs16986101
  GI_22035699-A       22 [39695954, 39824393]      * |     41.90    rs909685
  GI_24497446-A       22 [45509726, 45633888]      * |     34.69    rs132863
  GI_38157977-A       22 [21871957, 22028323]      * |     30.24   rs5754100
  GI_34486096-S       22 [41713392, 41845328]      * |     24.51   rs4822025
  GI_42662524-S       22 [50939542, 51051328]      * |     22.80    rs131777
                  snploc radiusUsed       fdr
                <integer>  <numeric> <numeric>
   GI_4504184-S  24346550      50000         0
   GI_8923587-S  45731539      50000         0
   GI_7262293-S  51063477      50000         0
   GI_6005825-S  43334295      50000         0
  GI_25092724-S  42924632      50000         0
  GI_22035699-A  39747671      50000         0
  GI_24497446-A  45564427      50000         0
  GI_38157977-A  21916166      50000         0
  GI_34486096-S  41776646      50000         0
  GI_42662524-S  50991033      50000         0
  ---
  seqlengths:
         22
   51116607
```

Use `plot_EvG(probeId("GI_4504184-S"), rsid("rs407257"), g22)` to visualize top hit (or apply the filter to g22 to see the transformed relationship); the gene is GSTT1.




eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/
nonparametric bayes
nyt
bioc
chanGmail
careygmail
cran
The R-devel Arc...
PubMed Home
Slashdot Stories ...
NIH Funding O...
Bioconductor bu...
Bookmark
Overview
Overview of chr22
0M
10M
20M
30M
40M
Ideogram
Details
22690k
22700k
22710k
22720k
22730k
Entrez genes
NM_000853
GSTT1: glutathione S-transferase theta 1
Degner et al. (2012): -log10(P), LCLs, 70 Nigerian HAPMAP ids, DNase sensitivity QTLs (dsQTLs) by DNase-seq
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.
Pickrell et al. (2010): -log10(P), LCLs, 69 Nigerian HAPMAP ids, RNA-Seq for eQTLs.
Pickrell et al. (2010): -log10(P), LCLs, 69 Nigerian HAPMAP ids, RNA-Seq for splicing QTLs.

Enhancements, if time permits

- Non-biological expression heterogeneity is a major concern with such studies (see references to Stegle, Leek), and alternatives to clipPCs may be of interest – define an `SVAfilter` or `PEERfilter`
- Is the trio structure a concern for inference?
- How sensitive are the findings to the number of permutations used?
- How would you sharpen a p-value for a borderline finding?
- The vignette for *genetw12* includes a section on multi-SNP testing for genes, applied to genes without simple eQTL, using SKAT.

Summary

- *snpStats* representation and testing facilities allow rapid surveys of SNP-phenotype associations under various genetic models
- *GGtools* `best.cis.eQTLs` supports rapid and concise identification of eQTL in a gene-centric framework
- Filtering and modeling details affect performance and interpretability
- Additional facilities are available for *trans* and multipopulation applications

NHGRI GWAS catalog

```
> library(gwascat)
'gwcat' data frame now available, provides NHGRI GWAS cat records of 02/02/2012.
building 'gwrngs', GRanges for studies with located variants...done.
> gwc22 = subsetByChromosome(gwrngs, "chr22")
> gwc22
gwasloc instance with 110 records and 34 attributes per record.
Excerpt:
GRanges with 5 ranges and 3 elementMetadata values:
      seqnames               ranges strand |          Disease.Trait        SNPs   p.Value
         <Rle>            <IRanges>  <Rle> |            <character> <character> <numeric>
  [1]    chr22 [37258503, 37258503]      * |      Atopic dermatitis   rs4821544     6e-06
  [2]    chr22 [30423460, 30423460]      * |        IgA nephropathy     rs12537     1e-11
  [3]    chr22 [17057138, 17057138]      * |   HIV-1 viral setpoint   rs5746647     2e-06
  [4]    chr22 [48929569, 48929569]      * |      Pancreatic cancer   rs5768709     1e-10
  [5]    chr22 [37310046, 37310046]      * | Ankylosing spondylitis   rs2075726     9e-06
```

Find the GWAS loci closest to our best *cis* eQTL

```
> nearest(ranges(fullreport(b.75a)[1:8]), ranges(gwc22))
[1]   9 103 105  90  90  27 103  79
> gwc22[unique(.Last.value)]
gwasloc instance with 6 records and 34 attributes per record.
Excerpt:
GRanges with 5 ranges and 3 elementMetadata cols:
      seqnames               ranges strand |
         <Rle>            <IRanges>  <Rle> |
  [1]    chr22 [24295286, 24295286]      * |
  [2]    chr22 [44332570, 44332570]      * |
  [3]    chr22 [51017353, 51017353]      * |
  [4]    chr22 [43500212, 43500212]      * |
  [5]    chr22 [39687484, 39687484]      * |
                                                    Disease.Trait        SNPs   p.Value
                                                      <character> <character> <numeric>
  [1] Plasma levels of liver enzymes (gamma-glutamyl transferase)    rs2739330     2e-09
  [2]                                Plasma levels of liver enzymes   rs2281135     8e-16
  [3]                                                   Narcolepsy   rs5770917     6e-08
  [4]                                              Prostate cancer   rs5759167     6e-29
  [5]                                        Sudden cardiac arrest     rs54211     8e-07
  ---
```

Focus on asthma: approximate the regulatory trait concordance of Nica et al. (2010); *a priori* focus on chr17; find eQTL nearest the risk loci

```
> asgw = subsetByTraits(gwrngs, "Asthma")
> asgw17 = subsetByChromosome(asgw, "chr17")
> elementMetadata(asgw17)[,c(2,8,15,21,28,31)]
DataFrame with 5 rows and 6 columns
     PUBMEDID Disease.Trait    Mapped_gene Strongest.SNP.Risk.Allele   p.Value OR.or.beta
  <character>   <character>    <character>               <character> <numeric>  <numeric>
1    21804549        Asthma          GSDMB              rs11078927-?     2e-16         NA
2    21150878        Asthma ORMDL3 - GSDMA               rs6503525-C     5e-07       1.33
3    20860503        Asthma          GSDMB               rs2305480-G     1e-07       1.18
4    20860503        Asthma          GSDMA               rs3894194-A     5e-09       1.17
5    17611496        Asthma          GSDMB               rs7216389-T     9e-11       1.45

> library(parallel)
> set.seed(1234)
> lk17.6 = best.cis.eQTLs("GGdata", ~male, chrnames="17",
                smFilter=function(x) MAFfilter( nsFilter(x, var.cutoff=.6), lower=0.05),
                geneApply=mclapply)
> nsig = sum(fdr(lk17.6) <= 0.05)
> nsig
[1] 65
> library(illuminaHumanv1.db)
> nrst = nearest( ranges(asgw17), ranges(fullreport(lk17.6)[1:nsig]) )
> ind = unique(nrst)
> ind
[1] 14
> get(names(fullreport(lk17.6))[ind], illuminaHumanv1SYMBOL)
[1] "ORMDL3"
```

Computing the RTC

Tasks for computing the approximate RTC:

Obtain the genotypes for the GWAS SNP – cited as rs7216389

Obtain residuals for prediction of ORMDL3 expression by rs7216389 (GWAS SNP) genotype

Compute association statistic for eQTL against this pseudo phenotype, and obtain its **rank** in the collection of statistics obtained against the pseudo phenotypes generated by obtaining residuals against all other proximal SNP

RTC = (Nprox – **rank**)/Nprox

Code not yet available; but for this example, it is not necessary:

```
> g17 = getSS("GGdata", "17")
> g17c = as(smList(g17)[[1]], "character")
> table( eqtl=g17c[, "rs12950743"], gwas=g17c[,
"rs7216389"] )
     gwas
eqtl  A/A A/B B/B
  A/A  24   0   0
  A/B   0  49   0
  B/B   0   0  16
  NA    0   1   0
```

Summary

- *gwascat* package provides location information and metadata on major SNP-phenotype associations in replicated GWAS as curated by NHGRI
- Pairing of eQTL and GWAS findings is simplified with `nearest()`
- RTC algorithm simple to implement in R

Working with deeply sequenced DNA from Complete Genomics Diversity panel

```
library(cgdv17)
> data(popvec)
> popvec[1:5]
NA19700 NA19020 NA19701 NA19025 NA19703
   "ASW"   "LWK"   "ASW"   "LWK"   "ASW"
> table(popvec)
popvec
ASW CEU CHB GIH JPT LWK MKK MXL TSI YRI
  5   5   4   4   4   4   4   5   4   7
```

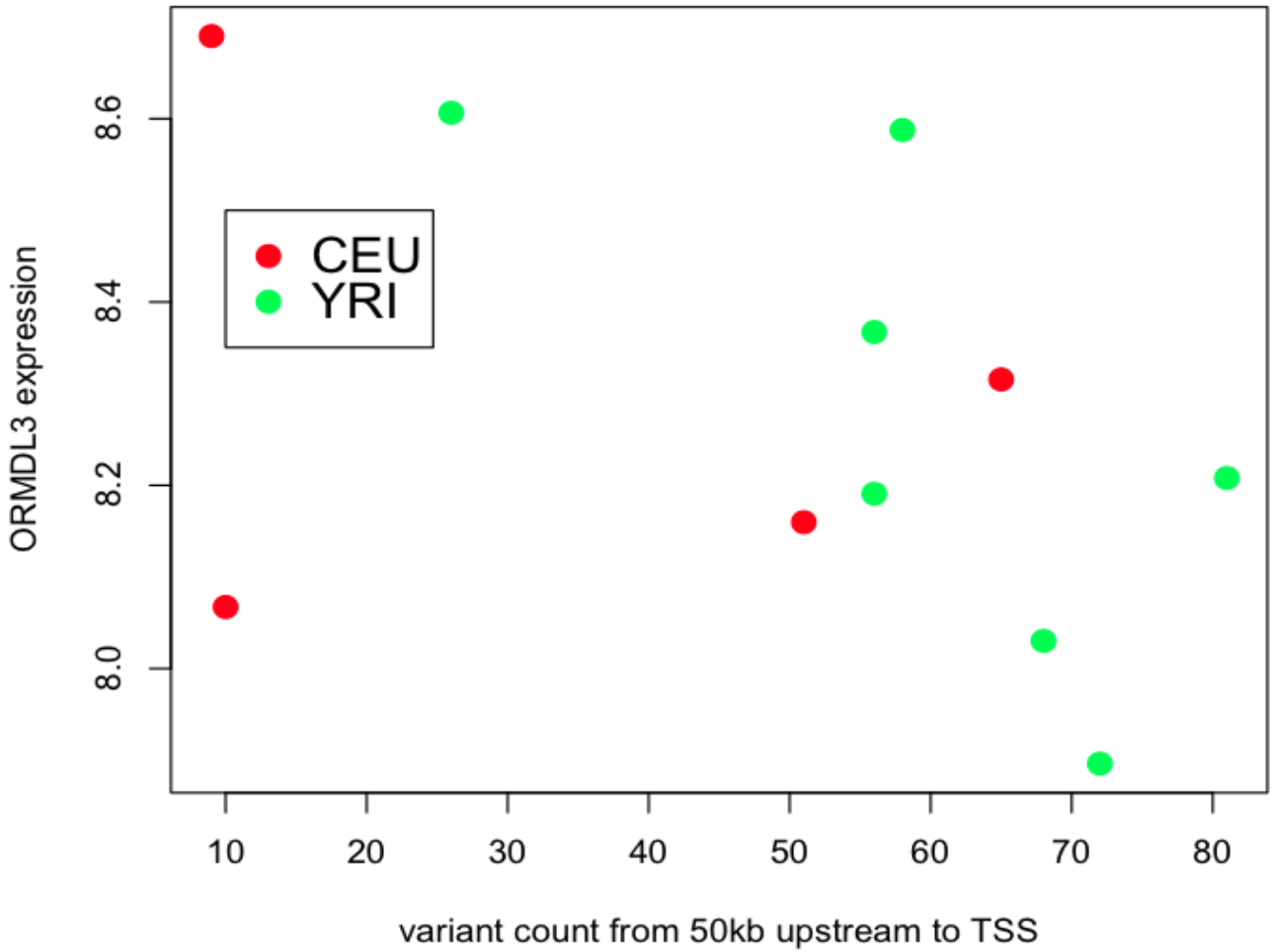
Different individuals present different sets of variants

```
> rv = getRVS("cgdv17")
> rv
raggedVariantSet instance with 46 elements.
some sampleNames:  NA06985 NA06994 ... NA21737 NA21767
> R85 = getrd(rv, "NA06985")
> length(R85)
[1] 174744
> R85[1:2]
GRanges with 2 ranges and 5 elementMetadata cols:
             seqnames     ranges strand |            REF
                <Rle>  <IRanges>  <Rle> | <DNAStringSet>
    chr17:1       17 [  1,  13]      * |  AAGCTTCTCACCC
  rs35998167      17 [302, 302]      * |               T
                                  ALT      QUAL         geno     depth
             <CompressedCharacterList> <numeric> <character> <integer>
    chr17:1                         .         0          ./.      <NA>
  rs35998167                       TA       139          1/0        12
```

Filtering variants on quality

```
> R85 = getrd(rv, "NA06985")
> length(R85)
[1] 174744
> summary(elementMetadata(R85)$QUAL)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       0      98     117     166    1714
> kp = which(elementMetadata(R85)$QUAL >= 166)
> R85hiq = R85[kp]
```

Vignette shows how to create: interpret



Summary

- Complete Genomics deep sequencing resources useful for methodologic development, complementary to 1000 genomes and other sequencing datasets
- TSV files transformed to VCF, one per individual
- Managing external VCF archives – work in progress
- Pad ragged variants to SnpMatrix – in vignette, simplifies association analysis

We'll skip RNA-seq variants

- ggtut and cheung2010 have relevant resources; the ggtut vignette addresses identifying allelic imbalance in transcription

DNase-seq and dsQTL

- Very new publication from Gilad/Pritchard lab (U Chicago)
- Data are distributed as bed files for normalized DNaseI hypersensitivity measures
  - Original assay tiled at 100bp
  - Filtered by authors to windows exhibiting DNaseI hypersensitivity (DHS) in top 5% of its distribution
  - Imputation to 1000 genomes genotypes, “mean GT”
- Search for SNPs or indels associated with variation in DHS across samples using 20kb radius (and also 1kb)

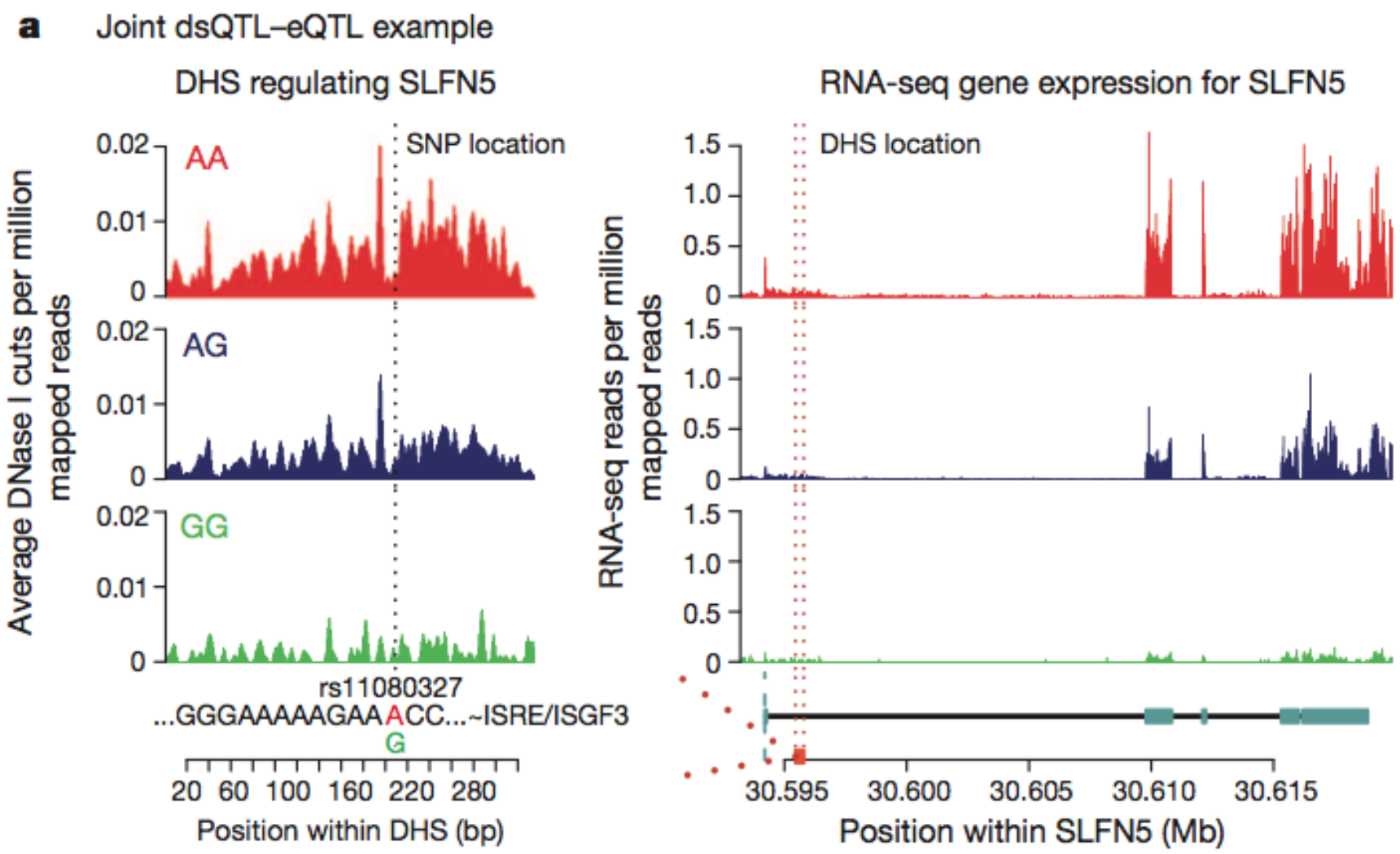


Figure 3 | Relationship between dsQTLs and eQTLs. a, Example of a dsQTL SNP that is also an eQTL for the gene *SLFN5*. The SNP disrupts an interferon- (right) measurer genotype at the p

Principles of managing and analyzing the dsQTL experiment

- Versioned R package for distribution and maintenance
- Formal coordination of sample assay data, metadata, and genotype data
  - How to connect high-dimensional assay (tiled genome) with genotype? smlSet is a reasonable low-cost approach for now
- Systematic extraction of location metadata from versioned packages and environments: *CHRLOC, *CHRLOCEND, getSNPlocs – not available

The package and some metadata

```
> library(dsQTL)
> data(package="dsQTL")
> data(DSQ_17)
> DSQ_17
class: SummarizedExperiment
dim: 105960 70
exptData(1): MIAME
assays(1): normDHS
rownames: NULL
rowData values names(0):
colnames(70): NA18486 NA18498 ... NA19239 NA19257
colData names(0):
> exptData(DSQ_17)[[1]]
Experiment data
  Experimenter name: Degner JF
  Laboratory: Department of Human Genetics, University of Chicago, Chicago,
    Illinois 60637, USA.
  Contact information:
  Title: DNaseâI sensitivity QTLs are a major determinant of human expression
    variation.
  URL:
  PMIDs: 22307276

  Abstract: A 252 word abstract is available. Use 'abstract' method.
```

The data on chromosome 2

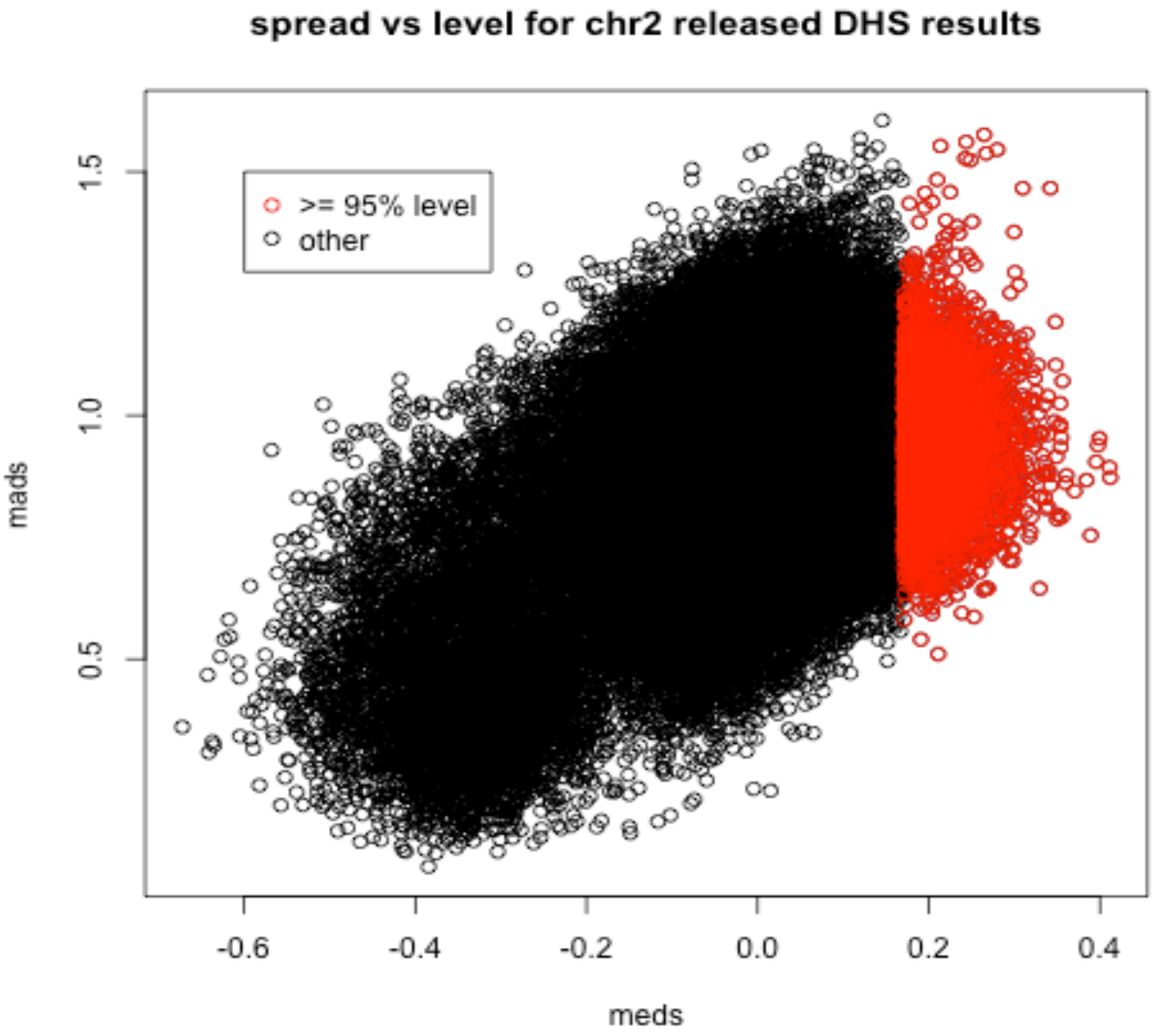
```
> data(DSQ_2)
> DSQ_2
class: SummarizedExperiment
dim: 96024 70
exptData(0):
assays(1): normedDHS
rownames: NULL
rowData values names(0):
colnames(70): NA18486 NA18498 ... NA19239 NA19257
colData names(0):
> assays(DSQ_2)[[1]][1:5,1:5]
        NA18486     NA18498    NA18499    NA18501    NA18502
[1,] -0.2684343 -0.78076674 -0.4840237  2.3894003 -1.0813642
[2,] -1.4445813  0.92170439  0.5812017  0.8627376  0.5186581
[3,]  0.7624075 -0.12340745 -1.1821308  1.4253179  0.3125592
> rowData(DSQ_2)[1:5]
GRanges with 5 ranges and 0 elementMetadata cols:
      seqnames       ranges strand
         <Rle>    <IRanges>  <Rle>
  [1]     chr2 [1202, 1301]      *
  [2]     chr2 [1602, 1701]      *
  [3]     chr2 [2002, 2101]      *
  [4]     chr2 [7502, 7601]      *
  [5]     chr2 [8802, 8901]      *
  ---
```

Borrowing the eQTL infrastructure

```
> d2 = getSS("dsQTL", "roundGT_2")
> d2
SnpMatrix-based genotype set:
number of samples:  70
number of chromosomes present:  1
annotation:
Expression data dims: 96024 x 70
Total number of SNP: 1336471
Phenodata: An object of class "AnnotatedDataFrame":
  none
> smList(d2)[[1]]
A SnpMatrix with  70 rows and  1336471 columns
Row names:  NA18486 ... NA19257
Col names:  chr2.140 ... chr2.242750984
```

Quiz

- The authors present/analyze data on the DHS sites achieving values at the 95th percentile or above over the entire experiment
  - What feature filtering principle is violated?
  - How, with complete assay results, could we explore sensitivity of findings to this choice? How could we (probably) enhance power of the study?


spread vs level for chr2 released DHS results
>= 95% level
other
mads
meds
0.5
1.0
1.5
-0.6
-0.4
-0.2
0.0
0.2
0.4

Anything strange?

```
> data(package="dsQTL")
Data sets in package 'dsQTL':

DSQ_17
DSQ_2
ch2locs
dsQTLCHR
dsQTLCHRLOC
dsQTLCHRLOCEND
ex (eset)
meanGT_chr2
```

Improvised compliant infrastructure

- Need to be able to create “cis maps”, lists of SNP proximal to feature of interest
- Standard approach: use *CHRLOC to determine gene location, getSNPlocs to determine SNP location
- The relevant resources don’t exist as centralized packages, but the standard APIs can be satisfied with stuff in arbitrary packages

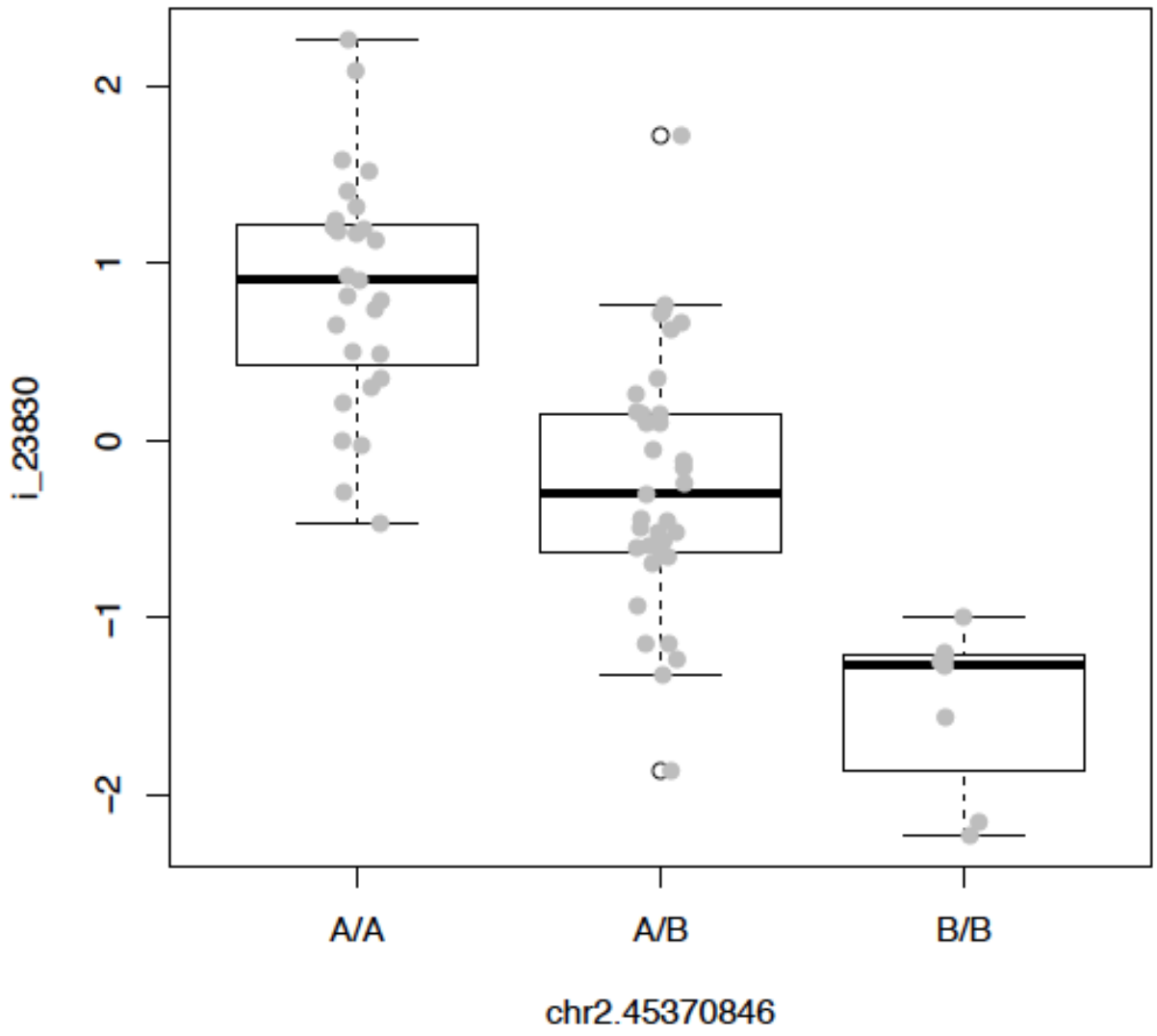
Allows reuse of available infrastructure

```
getSNPlocs = dsQTL::getSNPlocs  # force
n1 = best.cis.eQTLs(smpack="dsQTL", radius=2000,
   geneannopk="dsQTL",
   snpannopk="dsQTL", chrnames="2",
   smchrpref="roundGT_",
   smFilter =
          function(x) GTFfilter(x, lower=0.05)
   [23810:23830,], geneApply=mclapply)
```

These DHS features are selected deliberately

```
> n1
GGtools mcwBestCis instance.  The call was:
best.cis.eQTLs(smpack = "dsQTL", radius = 2000, chrnames = "2",
    smchrpref = "roundGT_", geneApply = mclapply, geneannopk = "dsQTL",
    snpannopk = "dsQTL", smFilter = function(x) GTFfilter(x,
        lower = 0.05)[23810:23830, ])
Best loci for 21 are recorded.
Top  4 probe:SNP combinations:
GRanges with 4 ranges and 5 elementMetadata cols:
          seqnames              ranges strand |     score          snpid
             <Rle>           <IRanges>  <Rle> | <numeric>   <character>
  i_23830        2 [45368802, 45373801]      * |     38.64 chr2.45370846
  i_23829        2 [45368702, 45373701]      * |     29.11 chr2.45370846
  i_23828        2 [45367802, 45372801]      * |     19.14 chr2.45370846
  i_23813        2 [45303002, 45308001]      * |      6.43 chr2.45307016
             snploc radiusUsed       fdr
          <integer>  <numeric> <numeric>
  i_23830  45370846       2000 0.0000000
  i_23829  45370846       2000 0.0000000
  i_23828  45370846       2000 0.0000000
  i_23813  45307016       2000 0.1666667
```

```
> plot_EvG(probeId("i_23830"), rsid("chr2.45370846"), getSS("dsQTL",
+     "roundGT_2"))
```



Upshots

- We can use the distributed bed files and genotypes to verify key assertions of the paper
- **`best.cis.eQTLs`** can be hijacked to establish QTL for regions exhibiting variability in DNaseI hypersensitivity
- Problem: the high resolution tiling takes us far beyond the cardinality of genes x SNP addressed by **`best.cis.eQTLs`**

Conclusions

- R/bioconductor principles can be deployed against integrative analysis tasks
  - General eQTL, GWAS catalog, rare variants, dsQTL
- Divide and conquer strategies are important
  - Iterate over arbitrary decomposition and combine as needed, perhaps much later
  - Design to make use of simple parallel execution
- Stretching R: SnpMatrix byte code, specially coded GLM score tests, out of memory (ff) archives of compressed test results
- Stretching Bioc: “faking” the structures for needed but unavailable annotation