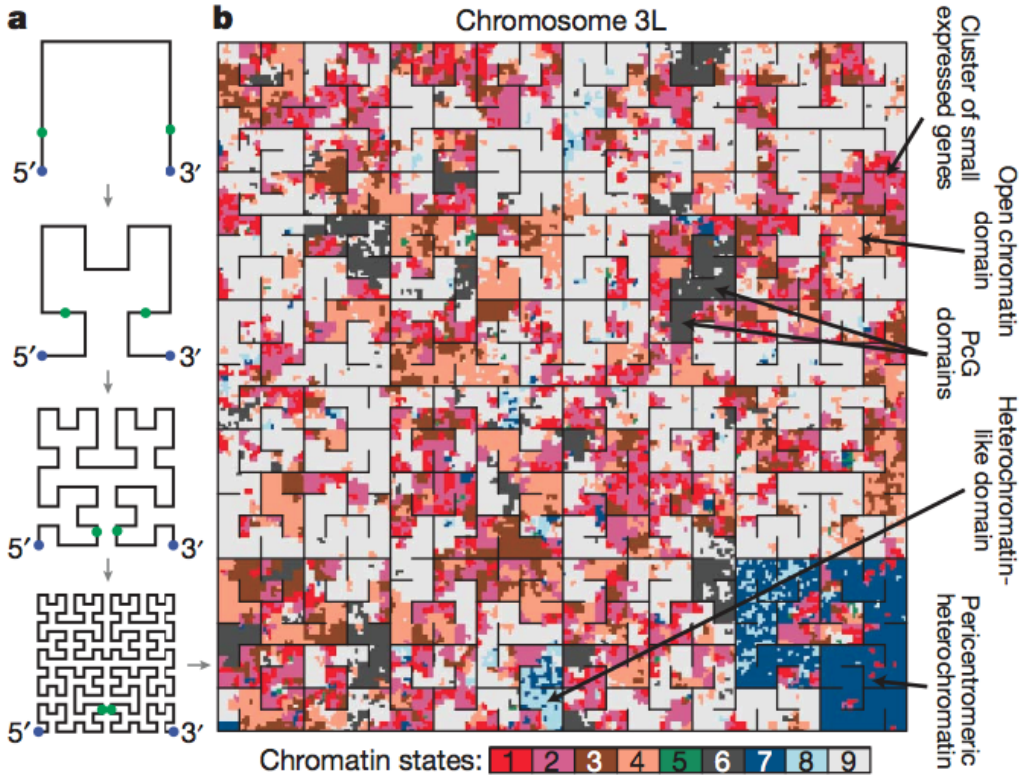
Data structures and methods for some integrative analyses

Vince Carey PhD

Harvard Medical School

The discreet charm of CSAMA listeners



So far four audience members have tapped me on the shoulder to help me understand fly chromosome structure … glad to have your support!

3 x 15’

- eQTL: sensitivity analysis, removal of extraneous variation
- dsQTL: genetics of chromatin accessibility – unraveling eQTL mechanism?
- CCLE: reproducing an application of elasticnet (combining lasso and ridge regression) to tumor chemosensitivity

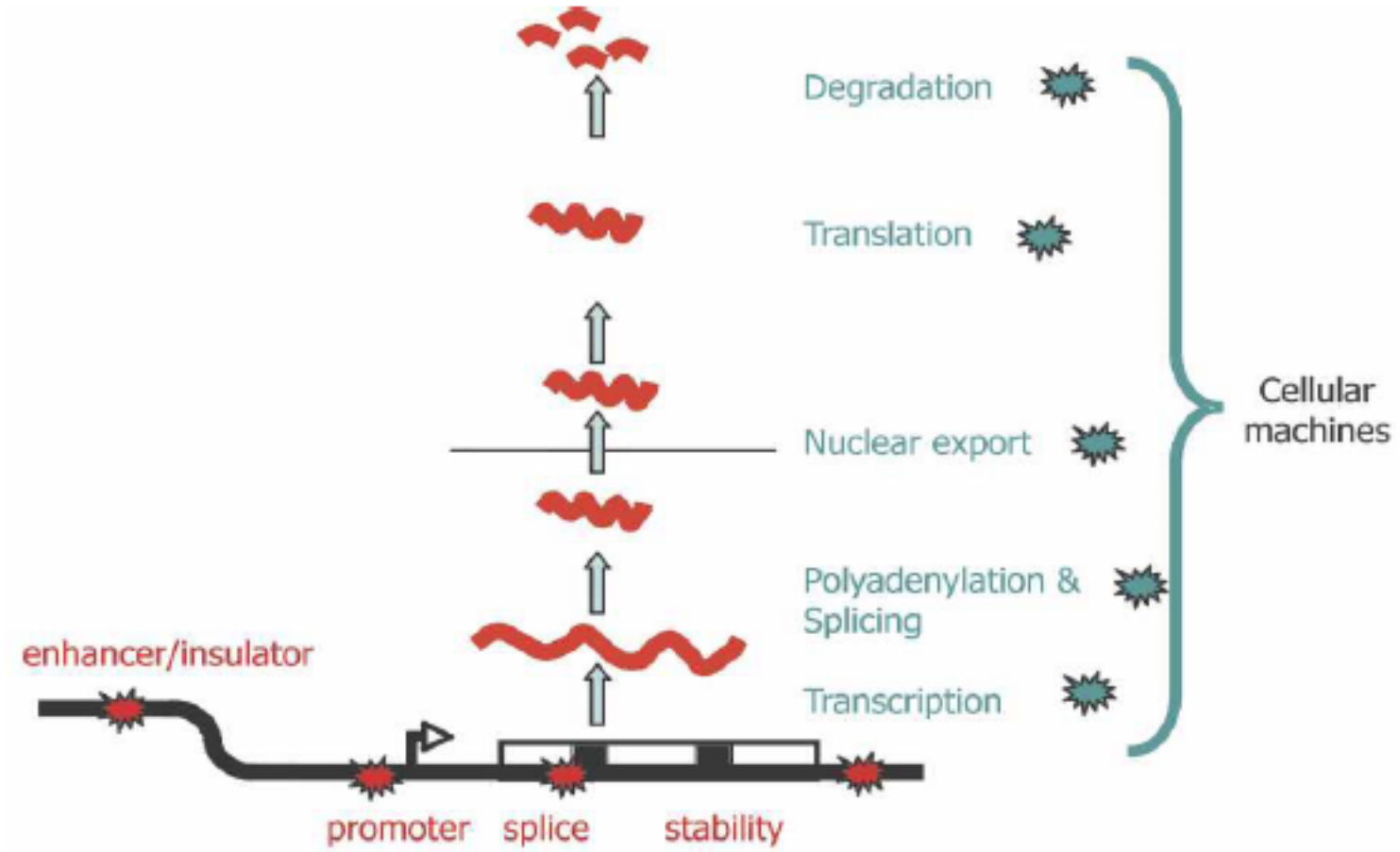
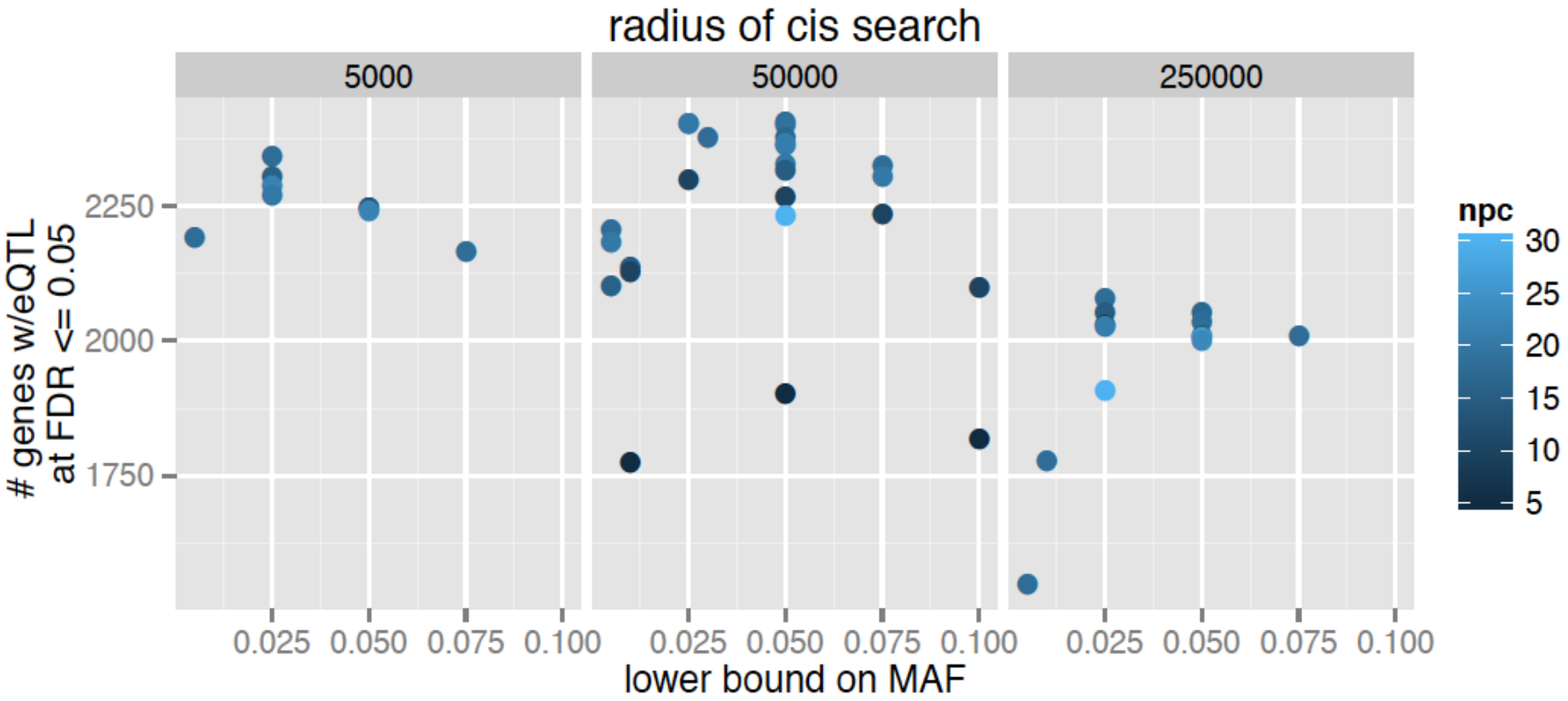


Figure 1. Plausible sites of action for genetic determinants of mRNA levels. Genetic variations influencing gene expression may reside within the regulatory sequences, promoters, enhancers, splice sites, and secondary structure motifs of the target gene and so be genetically in *cis* (red stars), or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

Opportunities for greedy tuning of cis-eQTL search



OPINION

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

Abstract | High-throughput technologies are widely used, for example to assay genetic variants, gene and protein expression, and epigenetic modifications. One often overlooked complication with such studies is batch effects, which occur because measurements are affected by laboratory conditions, reagent lots and personnel differences. This becomes a major problem when batch effects are correlated with an outcome of interest and lead to incorrect conclusions. Using both published studies and our own analyses, we argue that batch effects (as well as other technical and biological artefacts) are widespread and critical to address. We review experimental and computational approaches for doing so.

Many technologies used in biology — and hardware, along with highly trained personnel differences — affected by both biological and non-biological factors. Here we focus on batch effects, a common and powerful source of variation in high-throughput experiments.

Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study. For example, batch effects may occur if a subset of experiments was run on Monday and another set on Tuesday, if two technicians were responsible for different subsets of the experiments or if two different lots of reagents, chips or instruments were used. These effects are not exclusive to high-throughput biology and genomics research[1], and batch effects also affect low-dimensional molecular measurements, such as northern blots and quantitative PCR. Although batch effects are difficult or impossible to detect in low-dimensional assays, high-throughput technologies provide enough data to detect and even remove them. However, if not properly dealt with, these effects can have a particularly strong and pervasive impact. Specific examples have been documented in published studies[2,3] in which the biologi-

Table 1 | **Batch effects seen for a range of high-throughput technologies**

Study description*	Known variable used as a surrogate			Principal components used as a surrogate			Association with outcome	Refs
	Surrogate‡	Confounding (%)§	Susceptible features (%)‖	Principal components rank of surrogate (correlation)¶	Principal components rank of outcome (correlation)#	Susceptible features (%)**	Significant features (%)‡‡	
Data set 1: gene expression microarray, Affymetrix (N_p = 22,283)	Date	29.7	50.5	1 (0.570)	1 (0.649)	91.6	71.9	9
Data set 2: gene expression, Affymetrix (N_p = 4167)	Date	77.6	73.7	1 (0.922)	1 (0.668)	98.5	62.2	2
Data set 3: mass spectrometry (N_p = 15,154)	Processing group	100	51.7	2 (0.344)	2 (0.344)	99.7	51.7	3
Data set 4: copy number variation, Affymetrix (N_p = 945,806)	Date	29.2	99.5	2 (0.921)	3 (0.485)	99.8	98.8	16
Data set 5: copy number variation, Affymetrix (N_p = 945,806)	Date	12.2	83.8	1 (0.553)	1 (0.137)	99.8	74.1	17
Data set 6: gene expression, Affymetrix (N_p = 22,277)	Processing group	NA	83.8	5 (0.369)	NA	97.1	NA	18
Data set 7: gene expression, Agilent (N_p = 17,594)	Date	NA	62.8	2 (0.248)	NA	96.7	NA	18

nent will be highly correlated with cancer status. Principal components capture both biological and technical variability and, in some cases, principal components can be estimated after the biological variables have been accounted for[15]. In this case, the principal components primarily quantify the effects of artefacts on the high-throughput data. Principal components can be compared to known variables, such as processing group or time. If the principal components do not correlate with these known variables, there may be an alternative, unmeasured source of batch effects in the data.

involving the sTCC study, we examined the extent of batch effects for eight other published or publicly available data sets (TABLE 1) using the following approach. First, we identified a surrogate for batch effects (such as date or processing group) for each data set. We then used simple linear models to measure the level of confounding between this surrogate and the study outcome (for example, case or control) when available. Note that the more confounding there is, the more likely it is that batch variability can be confused with biological variability. We then summarized the

Upshots

- eQTLs are, in principle, identifiable by simple linear modeling of relationship between average expression and SNP genotype
- There is specificity to tissues, …
- Works of Stegle, Storey, Leek et al. indicate that removal of PCs and allied factors from expression array archives is important for improving sensitivity of eQTL detection

dsQTL identification

LETTER

doi:10.1038/nature10808

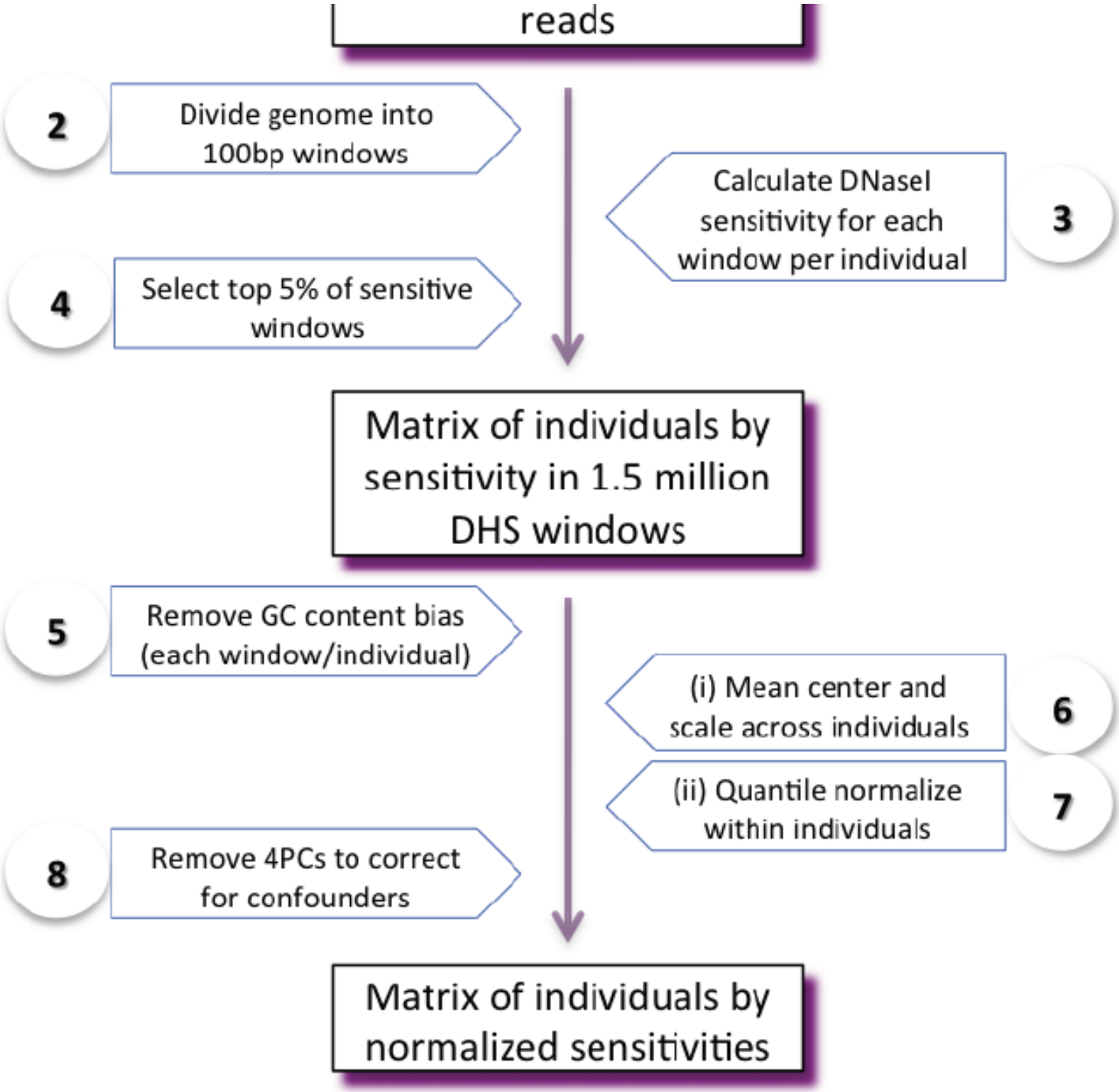
DNase I sensitivity QTLs are a major determinant of human expression variation

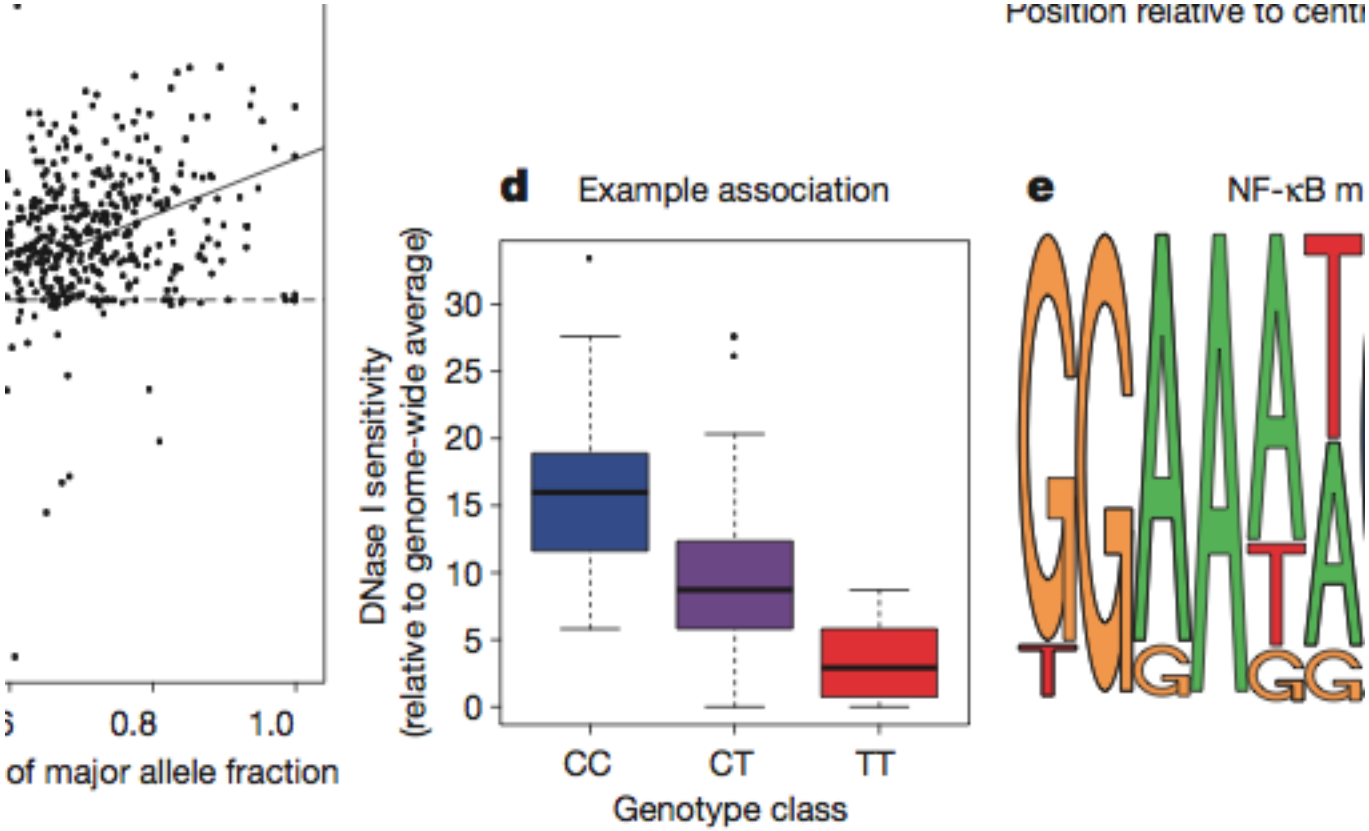
Jacob F. Degner[1,2]*, Athma A. Pai[1]*, Roger Pique-Regi[1]*, Jean-Baptiste Veyrieras[1,3], Daniel J. Gaffney[1,4], Joseph K. Pickrell[1], Sherryl De Leon[4], Katelyn Michelini[4], Noah Lewellen[4], Gregory E. Crawford[5,6], Matthew Stephens[1,7], Yoav Gilad[1] & Jonathan K. Pritchard[1,4]

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


reads
2
Divide genome into 100bp windows
3
Calculate DNaseI sensitivity for each window per individual
4
Select top 5% of sensitive windows
Matrix of individuals by sensitivity in 1.5 million DHS windows
5
Remove GC content bias (each window/individual)
6
(i) Mean center and scale across individuals
7
(ii) Quantile normalize within individuals
8
Remove 4PCs to correct for confounders
Matrix of individuals by normalized sensitivities



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

Different approaches to dsQTL data representation

- Chicago/GEO
  - Filtered and normalized DHS assay results in 70 bed files, indexed to hg18: 1.4GB gzipped on GEO, metadata not directly bound
  - Imputed genotype data harbored separately as name, loc, alleles, expected B allele count per indiv/SNP: chr1 = .3GB gzipped, 5 text bytes/SNP
- Bioconductor (dsQTLtools, not posted yet)
  - .8GB compressed `SummarizedExperiment` for DHS plus .08GB for 4 million imputed genotypes

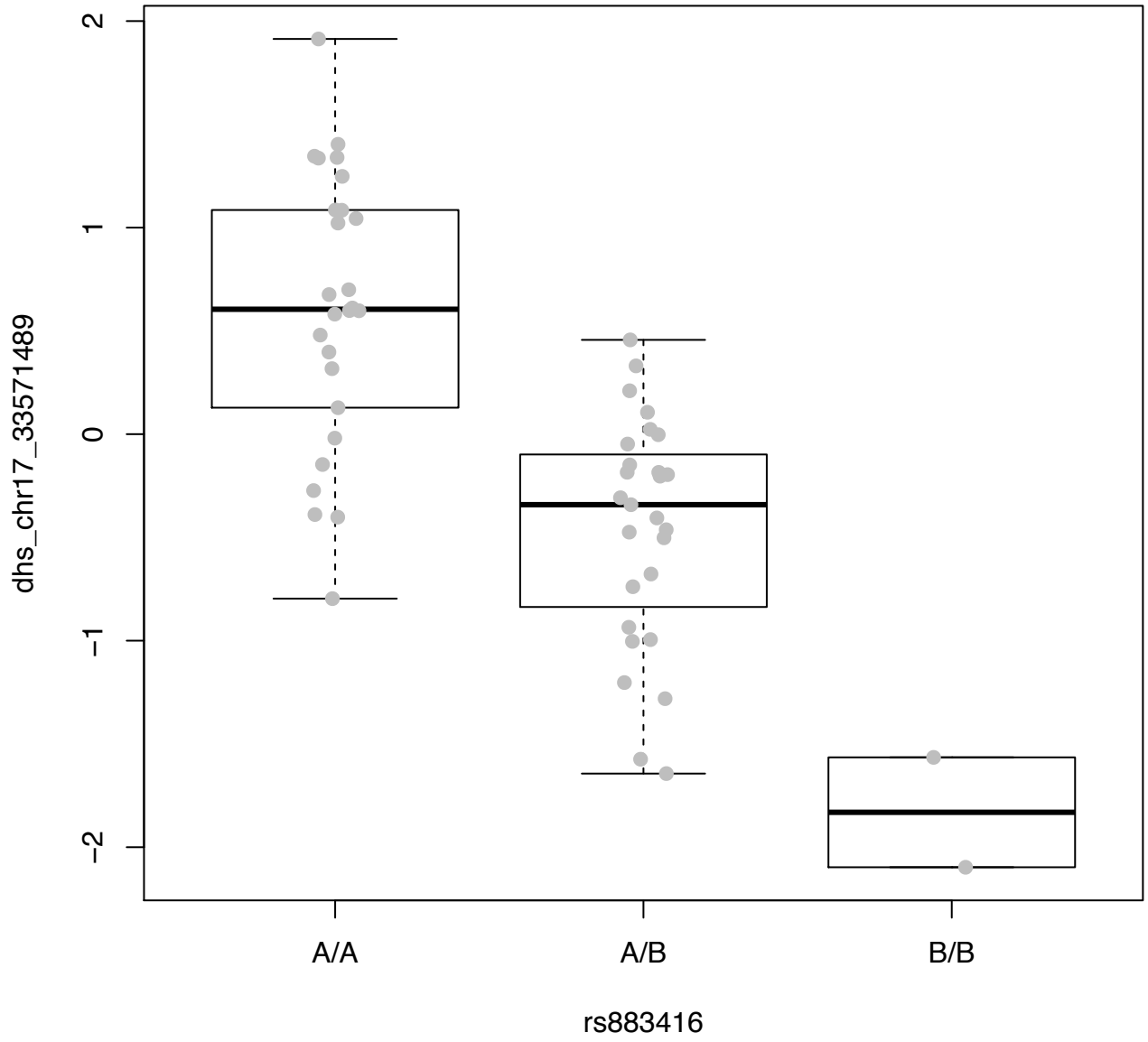
SummarizedExperiment instance

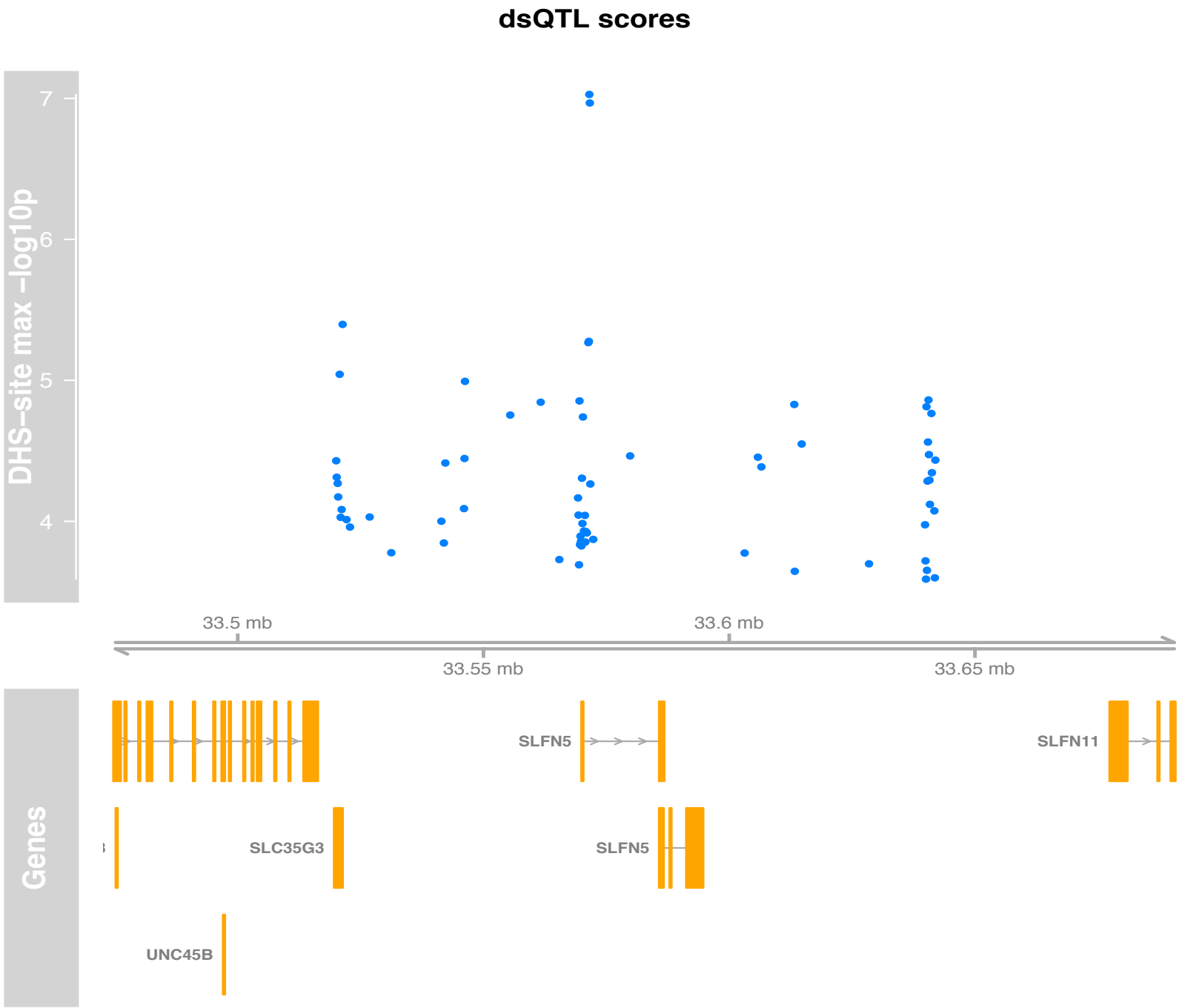
```
> DHStop5_hg19
class: SummarizedExperiment
dim: 1465442 70
exptData(2): MIAME annotation
assays(1): scores
rownames(1465442): dhs_chr1_10402 dhs_chr1_10502 ...
   dhs_chr22_51228236 dhs_chr22_51234736
rowData metadata column names(0):
colnames(70): NA18486 NA18498 ... NA19239 NA19257
colData names(9): naid one ... male isFounder

> assays(DHStop5_hg19)$scores[1:5,1:3]
                  NA18486    NA18498    NA18499
dhs_chr1_10402 -0.8932210 -0.3633581 -0.4540041
dhs_chr1_10502 -0.1523477 -0.1704101 -1.0598971
dhs_chr1_13239  0.4360728 -0.1159094  1.2505193
dhs_chr1_13939 -0.5259945 -0.8212344  0.1145535
dhs_chr1_16039 -0.9991160  0.2092481  0.3199874
```

```
> s1 = dsqNearGene("SLFN5")
> s1
dsqLook instance for SLFN5 w/ radius 1000.
best DHS site: dhs_chr17_33571489.
```

- R Under development (unstable) (2013-01-08 r61589), `x86_64-apple-darwin10.8.0`
- Locale: `en_US.US-ASCII/en_US.US-ASCII/en_US.US-ASCII/C/en_US.US-ASCII/en_US.US-ASCII`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, splines, stats, stats4, tools, utils
- Other packages: AnnotationDbi 1.21.9, Biobase 2.19.2, BiocGenerics 0.5.6, BiocInstaller 1.9.6, Biostrings 2.27.8, codetools 0.2-8, DBI 0.2-5, digest 0.6.0, dsQTLtools 0.0.5, GenomicFeatures 1.11.6, GenomicRanges 1.11.21, GGBase 3.21.2, GGtools 4.7.17, GO.db 2.8.0, hmyriB36 0.99.16, Homo.sapiens 1.0.0, IRanges 1.17.24, lattice 0.20-13, Matrix 1.0-10, org.Hs.eg.db 2.8.0, OrganismDbi 1.1.9, Rsamtools 1.11.14, RSQLite 0.11.2, snpStats 1.9.2, survival 2.37-2, TxDb.Hsapiens.UCSC.hg19.knownGene 2.8.0, weaver 1.25.0
- Loaded via a namespace (and not attached): annotate 1.37.3, biomaRt 2.15.0, bit 1.1-9, bitops 1.0-5, BSgenome 1.27.1, ff 2.2-10, genefilter 1.41.1, graph 1.37.4, grid 3.0.0, RBGL 1.35.0, RCurl 1.95-3, rtracklayer 1.19.6, VariantAnnotation 1.5.28, XML 3.95-0.1, xtable 1.7-0, zlibbioc 1.5.0


best dsQTL near SLFN5 rad. 50000
–log10 assoc p. = 7.028
dhs_chr17_33571489
2
1
0
–1
–2
A/A
A/B
B/B
rs883416


dsQTL scores
DHS–site max –log10p
7
6
5
4
33.5 mb
33.6 mb
33.55 mb
33.65 mb
Genes
SLFN5
SLFN11
SLC35G3
SLFN5
UNC45B

AnnotationHub for DnaseI peaks

```
> library(AnnotationHub)
> ah = AnnotationHub()
> nah = names(ah)
> ds = grep("dnase", nah, ignore.case=TRUE, value=TRUE)
> length(ds)
[1] 686
> ds[1:6]
[1] "goldenpath.dm2.database.bdtnpDnase_0.0.1.RData"

[2] "goldenpath.hg16.database.nhgriDnaseHs_0.0.1.RData"

[3] "goldenpath.hg17.database.sangamoDnaseHs_0.0.1.RData"

[4] "goldenpath.hg18.database.wgEncodeUwDnaseSeq_0.0.1.RData"

[5] "goldenpath.hg19.encodeDCC.wgEncodeAwgDnaseUniform.wgEncodeAwgDnaseDuke8988t
UniPk.narrowPeak_0.0.1.RData"
[6] "goldenpath.hg19.encodeDCC.wgEncodeAwgDnaseUniform.wgEncodeAwgDnaseDukeAosmc
UniPk.narrowPeak_0.0.1.RData"
```

Current metadata

```
> dst1 = ah[[ds[5]]]
Retrieving 'goldenpath/hg19/encodeDCC/wgEncodeAwgDnaseUniform/wgEncodeAwgDnaseDu
ke8988tUniPk.narrowPeak_0.0.1.RData'
> args(ahinfo)
function (hub, path)
NULL
> ahinfo(ah, ds[5] )
 From: EncodeDCC
 Version: ENCODE Jan 2011 Freeze
 Description: wgEncodeAwgDnaseDuke8988tUniPk
 Genus and Species: Homo sapiens
 Genome: hg19
 BiocVersion: 2.12, 2.13
 Tags: 8988T, wgEncodeAwgDnaseUniPk, DnaseSeq, ENCODE Jan 2011 Freeze, wgEncodeE
H001103, Duke, 80fadeb7a14a72add38203910d937f50, wgEncode, 1700000, wgEncodeAwgD
naseDuke8988tUniPk, None, narrowPeak, Peaks, wgEncodeAwgDnaseUniform
```

Upshots

- Representation of DnaseI HS can take various forms at various scales, tracks are nice but archive is complicated
- DS-seq archive very substantial even as filtered, but a SummarizedExperiment container can manage it
- Searching for dsQTL with substantial parallelism and small RAM footprint: Martin and Val's Streamer, scanVCF

Some machine learning with CCLE

LETTER

doi:10.1038/nature11003

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina[1,2,3]†*, Giordano Caponigro[4]*, Nicolas Stransky[1]*, Kavitha Venkatesan[4]*, Adam A. Margolin[1]†*, Sungjoon Kim[5], Christopher J. Wilson[4], Joseph Lehár[4], Gregory V. Kryukov[1], Dmitriy Sonkin[4], Anupama Reddy[4], Manway Liu[4], Lauren Murray[1], Michael F. Berger[1]†, John E. Monahan[4], Paula Morais[1], Jodi Meltzer[4], Adam Korejwa[1], Judit Jané-Valbuena[1,2], Felipa A. Mapa[4], Joseph Thibault[5], Eva Bric-Furlong[4], Pichai Raman[4], Aaron Shipway[5], Ingo H. Engels[5], Jill Cheng[6], Guoying K. Yu[6], Jianjun Yu[6], Peter Aspesi Jr[4], Melanie de Silva[4], Kalpana Jagtap[4], Michael D. Jones[4], Li Wang[4], Charles Hatton[3], Emanuele Palescandolo[3], Supriya Gupta[1], Scott Mahan[1], Carrie Sougnez[1], Robert C. Onofrio[1], Ted Liefeld[1], Laura MacConaill[3], Wendy Winckler[1], Michael Reich[1], Nanxin Li[5], Jill P. Mesirov[1], Stacey B. Gabriel[1], Gad Getz[1], Kristin Ardlie[1], Vivien Chan[6], Vic E. Myer[4], Barbara L. Weber[4], Jeff Porter[4], Markus Warmuth[4], Peter Finan[4], Jennifer L. Harris[5], Matthew Meyerson[1,2,3], Todd R. Golub[1,3,7,8], Michael P. Morrissey[4]*, William R. Sellers[4]*, Robert Schlegel[4]* & Levi A. Garraway[1,2,3]*

The systematic translation of cancer genomic data into knowledge of tumour biology and therapeutic possibilities remains challenging. Such efforts should be greatly aided by robust preclinical model systems that reflect the genomic diversity of human cancers and for which detailed genetic and pharmacological annotation is available[1].

known cancer genes were assessed by mass spectrometric genotyping[13] (Supplementary Table 2 and Supplementary Fig. 1). DNA copy number was measured using high-density single nucleotide polymorphism arrays (Affymetrix SNP 6.0; Supplementary Methods). Finally, messenger RNA expression levels were obtained for each of the lines using Affymetrix

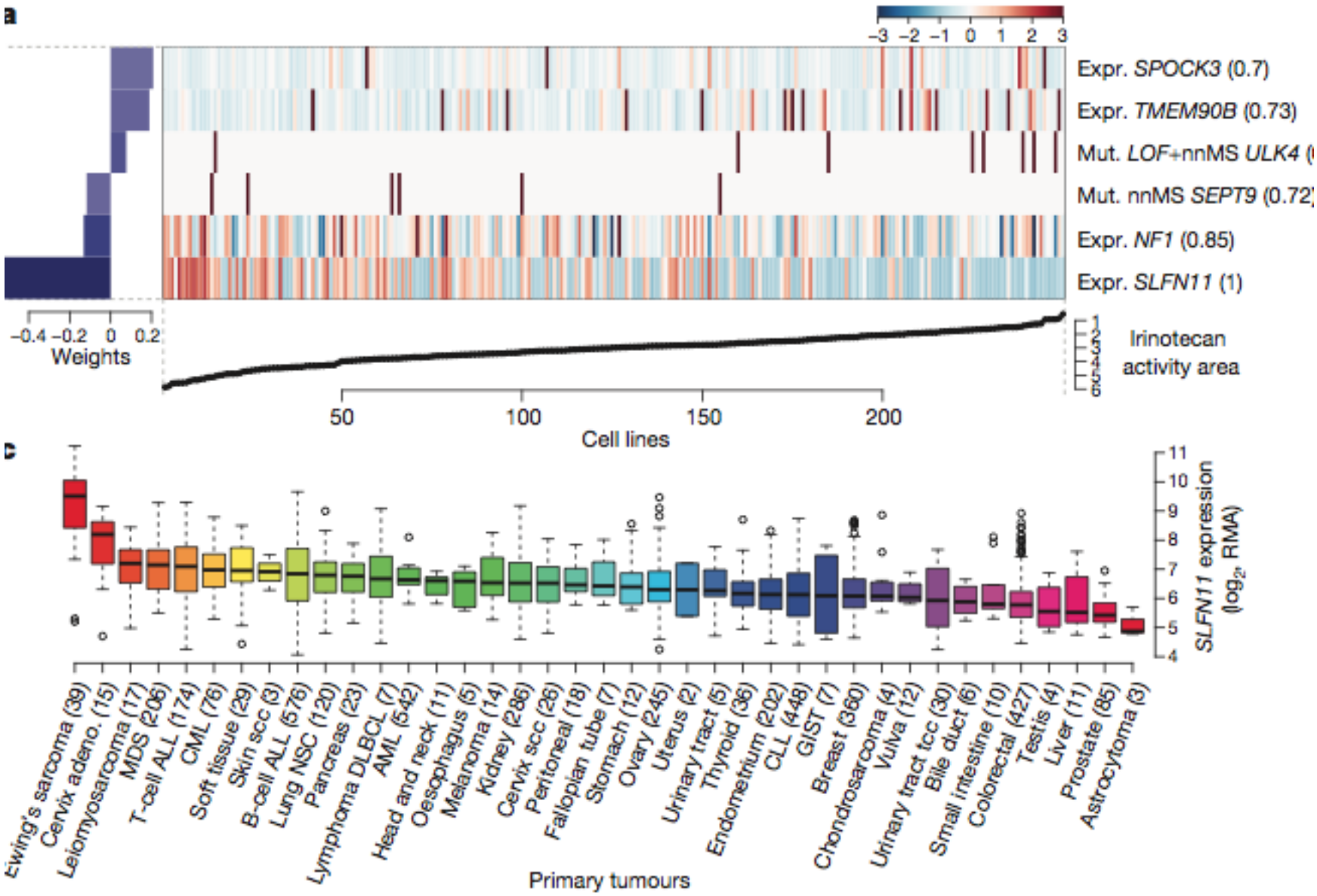


Figure 4 | Predicting sensitivity to topoisomerase I inhibitors. **a,** Elastic net regression analysis of genomic correlates of irinotecan sensitivity is shown for 250 cell lines. **b,** Dose–response curves for three Ewing's sarcoma cell lines

the mean growth inhibition (*n*
tumours. Box-and-whisker plo
each subtype, ordered by the n

Specification [edit]

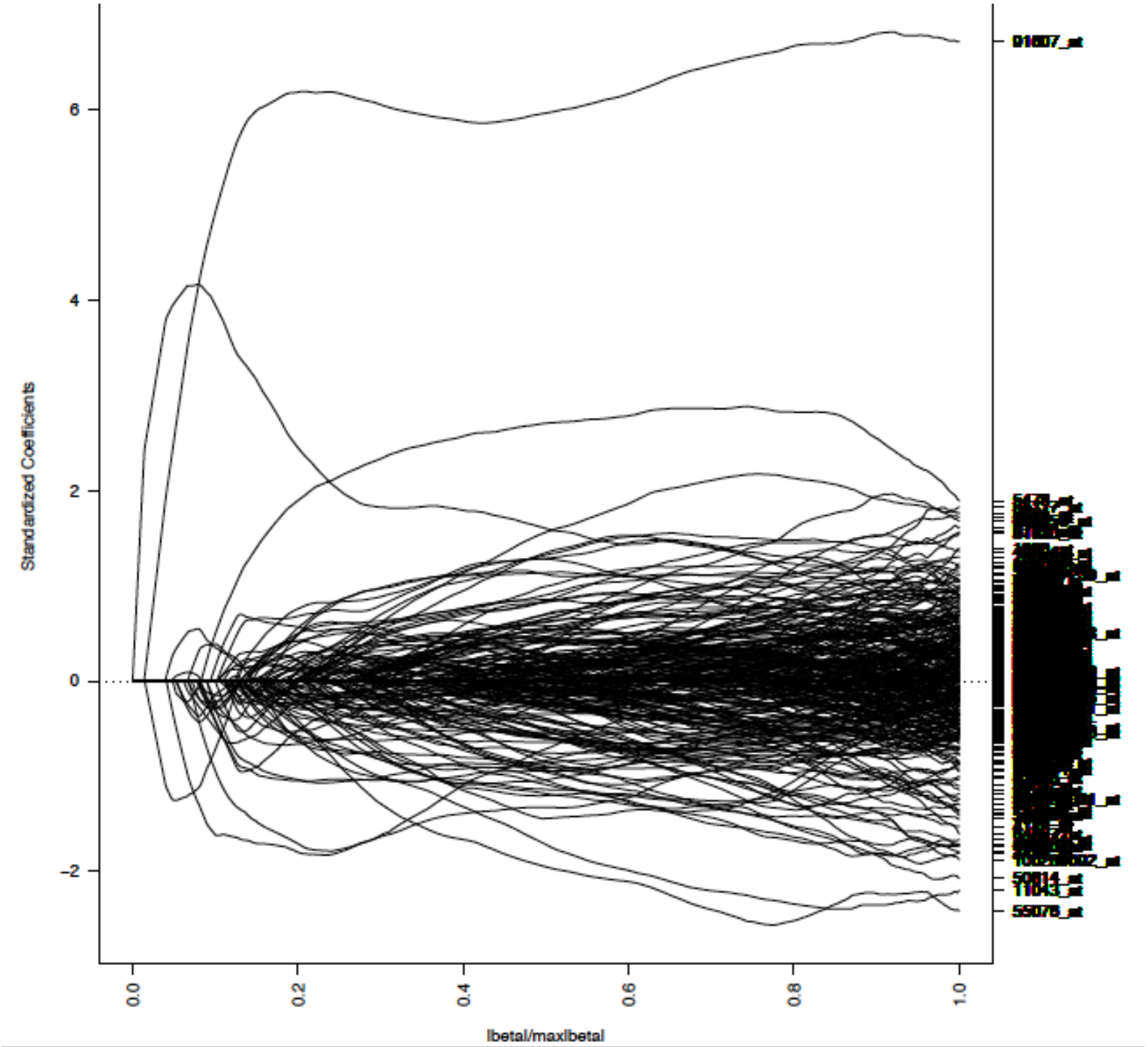
The elastic net method which overcomes the limitations of the LASSO (least absolute shrinkage and selection operator) method which uses a penalty function based on

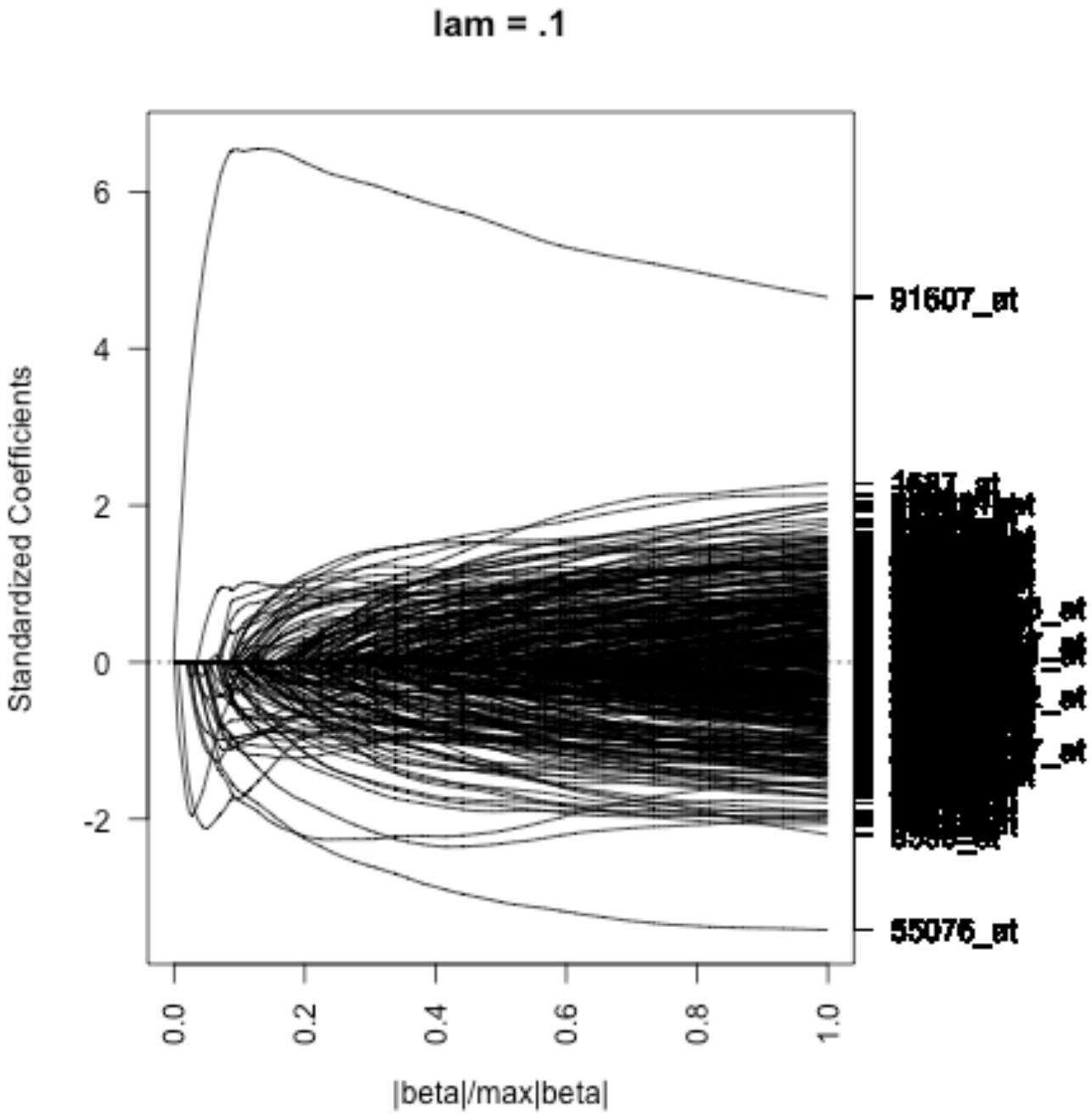
$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

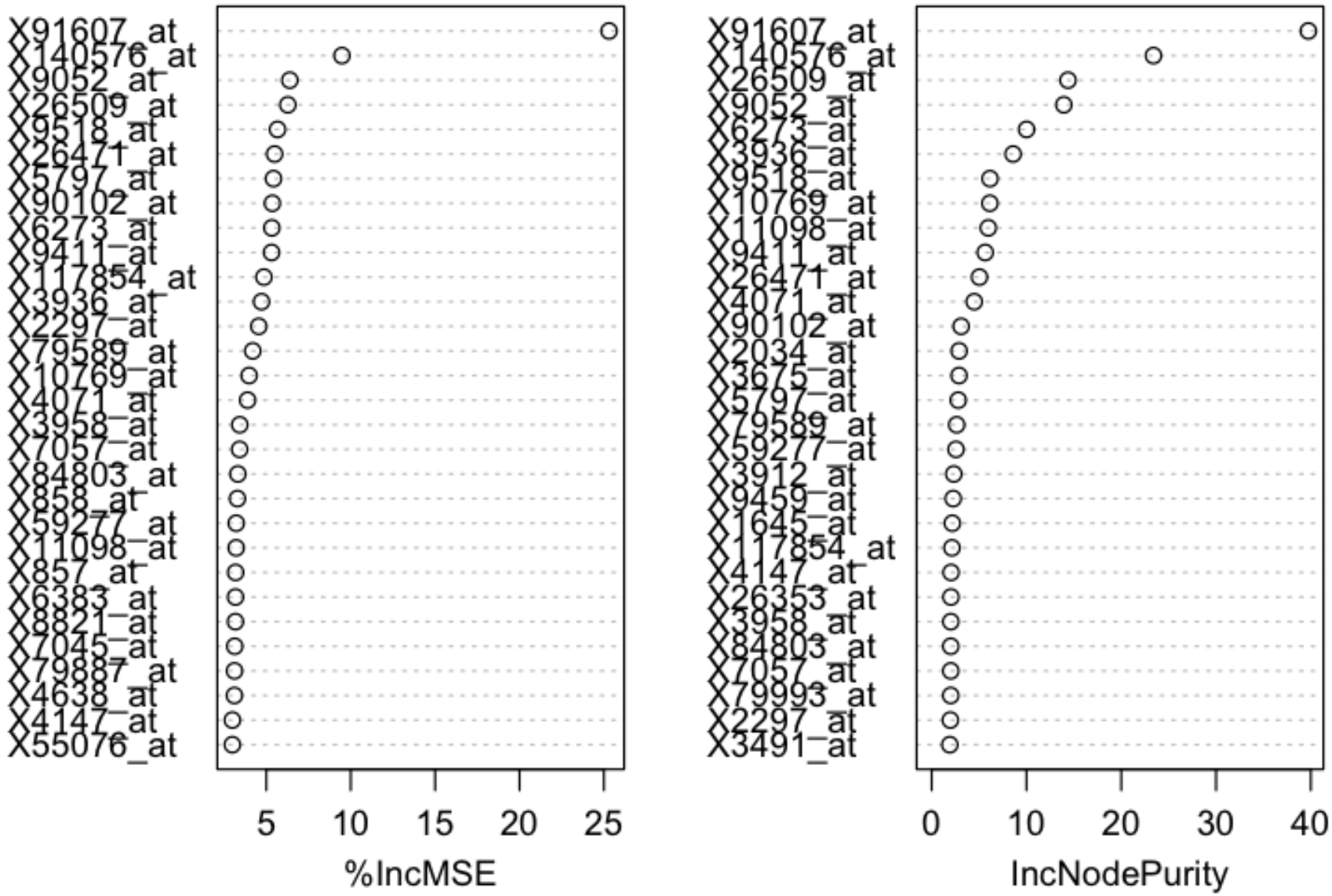
Use of this penalty function has several limitations.[1] For example, in the "large *p*, small *n* problem" case, the LASSO selects at most n variables before it saturates. Also if there is a group of highly correlated variables, then the LASSO tends to select one variable from a group and ignore the others. To overcome these limitations, the elastic net adds a quadratic part to the penalty ($\|\beta\|^2$), which when used alone is ridge regression (known also as Tikhonov regularization). The estimates from the elastic net method are defined by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

As a result, the elastic net method includes the LASSO and ridge regression: in other words, each of them is a special case where $\lambda_1 = 1, \lambda_2 = 0$ or $\lambda_1 = 0, \lambda_2 = 1$. Meanwhile, the naive version of elastic net method finds an estimator in a two-stage procedure : first for each fixed λ_2 it finds the ridge regression coefficients, and then does a LASSO type shrinkage. This kind of estimation incurs a double amount of shrinkage, which introduces unnecessary extra bias and outcomes with bad prediction performance. To improve the prediction performance, the authors rescale the coefficients of the naive version of elastic net by multiplying the estimated coefficients by $(1 + \lambda_2)$.[1]

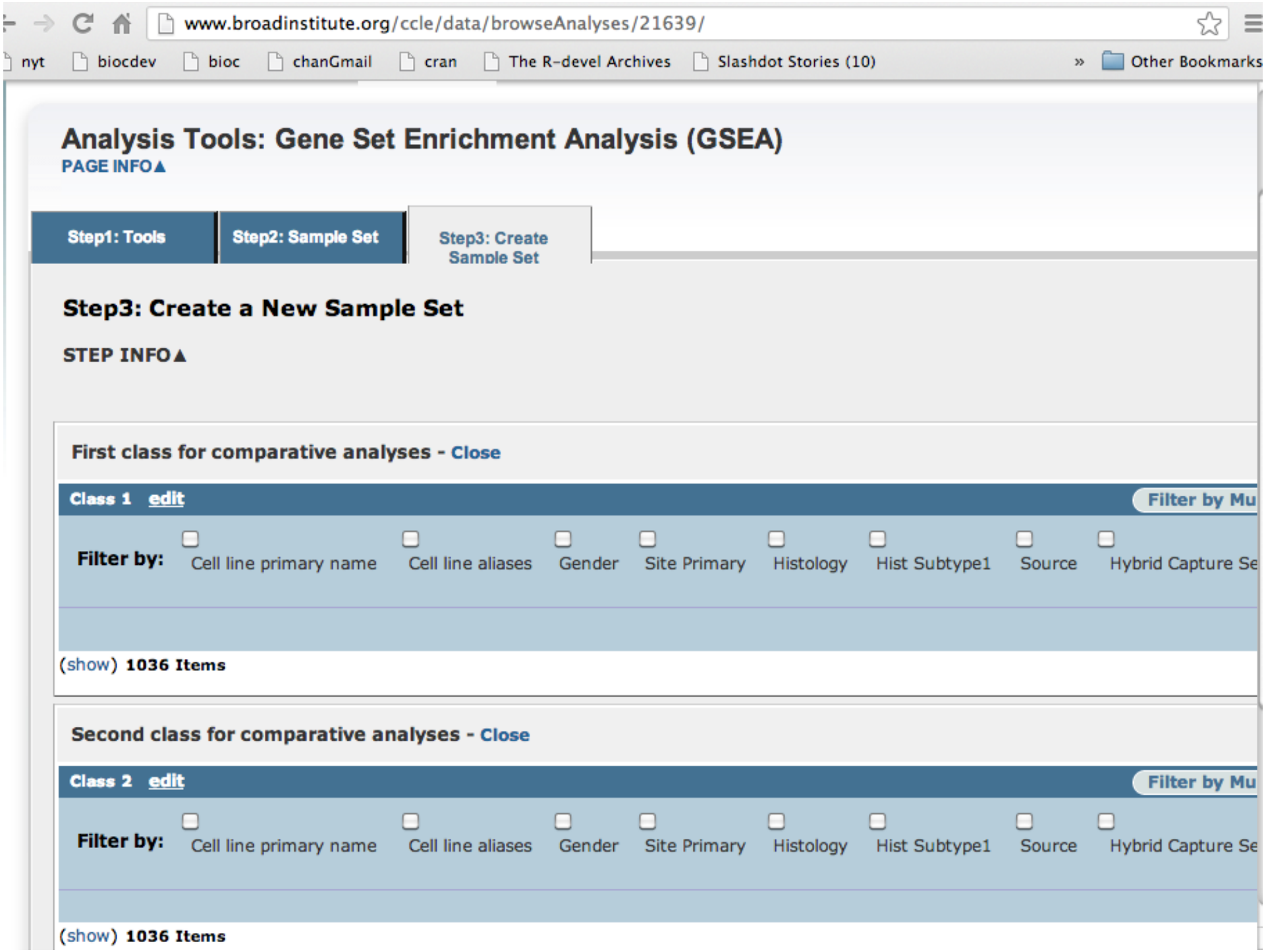

Standardized Coefficients
6
4
2
0
−2
0.0
0.2
0.4
0.6
0.8
1.0
|beta|/max|beta|
91607_at
50814
11043
55078


lam = .1
Standardized Coefficients
6
4
2
0
-2
0.0
0.2
0.4
0.6
0.8
1.0
|beta|/max|beta|
91607_at
55076_at


rf1
X91607_at
X140576_at
X9052_at
X26509_at
X9518_at
X26471_at
X5797_at
X90102_at
X6273_at
X9411_at
X117854_at
X3936_at
X2297_at
X79589_at
X10769_at
X4071_at
X3958_at
X7057_at
X84803_at
X858_at
X59277_at
X11098_at
X857_at
X6383_at
X8821_at
X7045_at
X79887_at
X4638_at
X4147_at
X55076_at
5
10
15
20
25
%IncMSE
X91607_at
X140576_at
X26509_at
X9052_at
X6273_at
X3936_at
X9518_at
X10769_at
X11098_at
X9411_at
X26471_at
X4071_at
X90102_at
X2034_at
X3675_at
X5797_at
X79589_at
X59277_at
X3912_at
X9459_at
X1645_at
X117854_at
X4147_at
X26353_at
X3958_at
X84803_at
X7057_at
X79993_at
X2297_at
X3491_at
0
10
20
30
40
IncNodePurity

Comments

- Qualitative effects of tuning the elastic net vs. lasso … organize with CV … performance of a single fit not so great
- Please consider models beyond "main effects" – randomForests is a black-box approach that accommodates one approach to variable interaction
- What about batch effects? Has the published expression data been properly adjusted?


www.broadinstitute.org/ccle/data/browseAnalyses/21639/
nyt
biocdev
bioc
chanGmail
cran
The R-devel Archives
Slashdot Stories (10)
Other Bookmarks
Analysis Tools: Gene Set Enrichment Analysis (GSEA)
PAGE INFO▲
Step1: Tools
Step2: Sample Set
Step3: Create Sample Set
Step3: Create a New Sample Set
STEP INFO▲
First class for comparative analyses - Close
Class 1 edit
Filter by Mu
Filter by:
Cell line primary name
Cell line aliases
Gender
Site Primary
Histology
Hist Subtype1
Source
Hybrid Capture Se
(show) 1036 Items
Second class for comparative analyses - Close
Class 2 edit
Filter by Mu
Filter by:
Cell line primary name
Cell line aliases
Gender
Site Primary
Histology
Hist Subtype1
Source
Hybrid Capture Se
(show) 1036 Items

Quiz

- How should we organize data from integrative experiments (expression, CNV, genotype, drug sensitivity)?
- Ordering genes measured across tumor types with respect to association between expression and drug sensitivity: what methods are preferred? Can we distinguish sensitive from insensitive tumor types?
- How to test whether a given mutation distinguishes sensitivities within a tumor class?