Reproducible workflow concepts and tools in genome-scale statistics

VJ Carey, Ph.D.
Harvard Medical School

CSAMA 2012, Bressanone

Road map

- Six slides on scope
- Two concepts of reproducibility
- Basic concerns:
  - Experimental design
  - Archive design and management
  - Reproducible interpretation and reporting
  - Audit support
- Examples
- Exercises

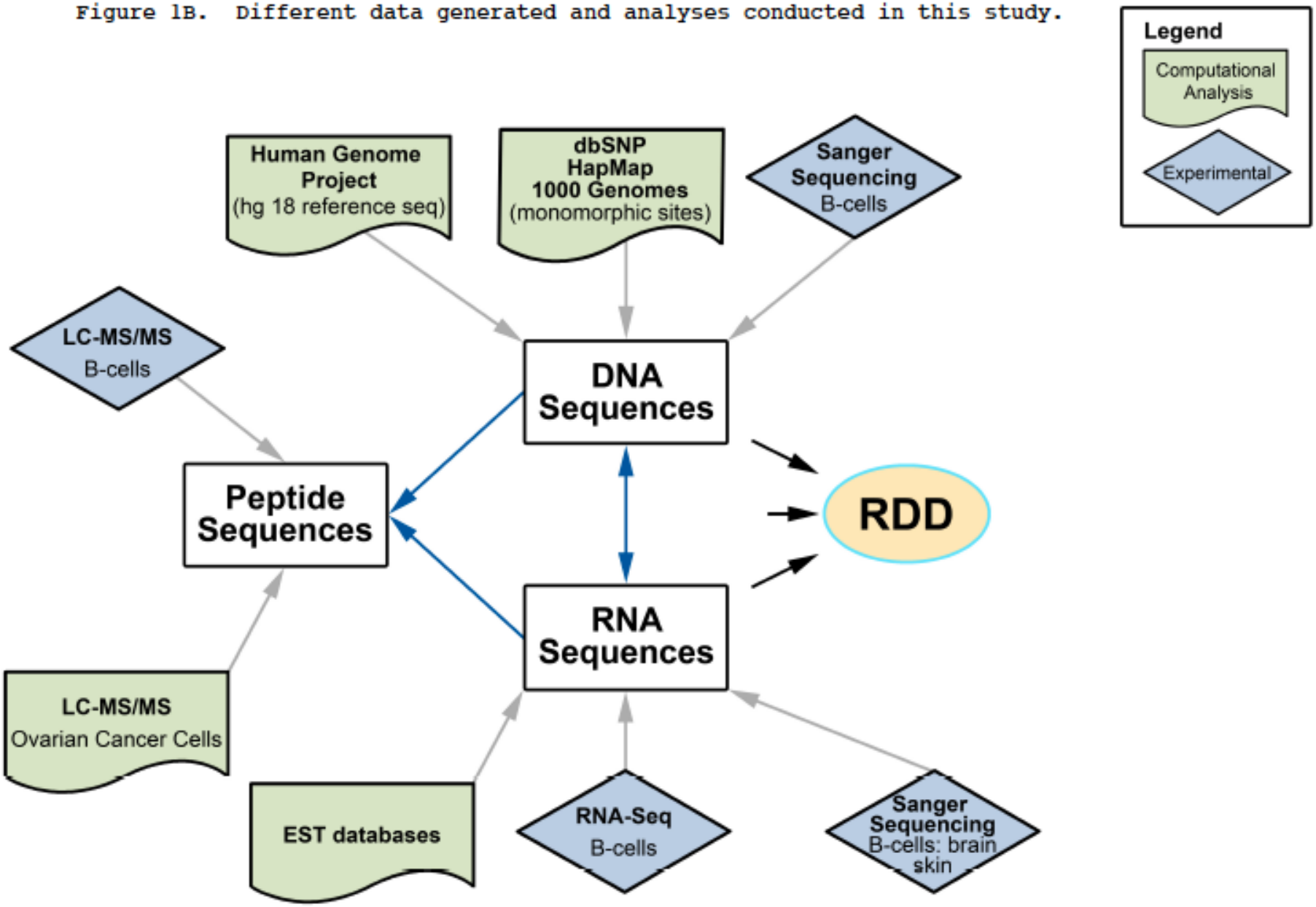
A three-phase rubric for large genomic data activities

- QAN: Quality assessment and normalization
  - Data will be discarded
  - Data will be changed to facilitate comparison
- A4I: Assembly for interrogation
  - Data will be labeled with ad hoc tokens
  - Samples must be coordinated with various forms of metadata
- DRF: Discovery and reporting of findings
  - You will try many things and need to exhibit them all
  - Whatever works will be worthy of packaging for reuse

Some processes and attributes, tools

		Define	Reference	Optimize	Scale	Verify	*Packages*, `classes`
Genotyping arrays	QAN	✓	✓		✓	✓	*GWAStools*, *snpStats*
	A4I	✓	✓	✓	✓	✓	*crlmm*, `SnpMatrix`
	DRF	✓	✓	✓	✓	✓	*GWAStools*, *GGtools*
NGS for rare and structural variants	QAN	✓	✓		✓		*ShortRead*, *Rsamtools*
	A4I	✓	✓		✓		`SummarizedExperiment`
	DRF	✓	✓				*VariantAnnotation*
ChIP-seq	QAN	✓	✓		✓		as above
	A4I	✓	✓		✓		*chipseq*
	DRF	✓	✓				*BayesPeak*, *ChIPpeakAnno*
RNA-seq	QAN	✓	✓		✓		as above
	A4I	✓	✓		✓		*Genominator*
	DRF	✓	✓				*DEseq*, *edgeR*
mRNA abundance arrays	QAN	✓	✓	✓	✓	✓	*affy*, *oligo*, *limma*, *vsn*, *lumi*
	A4I	✓	✓		✓		*Biobase*, `ExpressionSet`
	DRF	✓	✓			✓	*limma*, *MLInterfaces*, *SpikeIn*

Figure 1B. Different data generated and analyses conducted in this study.



Summary

- Contemporary experiments involve complex protocols, large data volumes, extensive transformation and interactive analysis
- Management requirements are substantial but distributed – wet lab, IT (data and software support), intellectual collaborations
- Genome scale studies attract particular concern owing to risk, cost, impact

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.)

by defi
their ra
to bioc
sive sp
major
in the
trend c
sive int
As
with n
believe
cies—a
an aca
the glo
of the r
to biol
many l
IUCN
(*9*), pra
use a s
of inva
arrival
biosecu
campai
suppor

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

Two concepts of reproducibility

- Concrete reproducibility: Given the data, party B can compute the tables and figures of the paper by party A
- Substantive reproducibility: Given the experimental design/protocol of the paper by party A, party B can reach compatible conclusions on independently collected data
- N.B.
  - Neither are guaranteed even when A and B are in the same lab
  - A and B may be the same person at different times

Issues with these notions of reproducibility

- Concrete reproducibility seems empty – says nothing about correctness/reliability
  - If the work IS correct, achieving concrete reproducibility adds value, ensuring reusability and extensibility of original components
- Substantive reproducibility is a *sine qua non* for scientific progress, but not formalized
- After a discussion of design and substantive reproducibility, we'll focus on reducing barriers to concrete reproducibility

Key quantities for scientific measurement protocols

- Bias: difference between average of measured quantities and the biologic constant to be estimated (e.g., actual fold change or rate)
- Variance: average squared departure of individual measurements from the biologic constant to be estimated
  - Technical variance: errors of measurement, only of metrological interest
  - Biologic variance: ``true'' fluctuations in properties between individuals, over time, in response to treatment … this kind of variance is our key concern

Experimental design

- Good experimental design is fundamental to substantive reproducibility
- Genome-scale experiments have **widely distributed** design components (sample collection, wet lab(s), assay execution, digitization, normalization, filtering)
- Experimental design is the prospective **systematic analysis** of sources of bias and variance affecting experiment interpretation leading to a measurement protocol whose outcomes have good properties – **low technical bias and variance**, good recovery of biological variation of interest

Reducing technical bias

- Blocking and randomization
- Example: 16 samples, 8 treated, 8 controls, two 8 array chips
- What do we need to assume if we want an unbiased estimate of difference if all treated samples are assigned to chip 1 and all controls to chip 2? Suggest a formalism and show how different allocations allow weaker assumptions.

Simple formalism

- Y_i is measurement on sample i, i = 1, …, 16
- $Y_i = \mu + \delta t_i + \theta c_i + \varepsilon_i$ where t_i is 1 if sample i is treated, 0 otherwise, c_i is 1 is sample i is allocated to chip 1 and 0 otherwise, and ε_i is a random disturbance with mean zero and variance σ^2
- Bind columns 1, t and c together and you get the ``design matrix”
- Give precise definitions of δ and θ and consider how properties of estimates of these parameters depend on the structure of the design matrix

Before treatment

- Suppose the samples are apparently homogeneous
- How should we choose the ones to be treated?
- How should we administer the treatment?

Within each chip

- Suppose we decide to place 4 controls and 4 cases on each chip.
- What do we need to assume about location effects to obtain unbiased estimates of treatment effects?
- How can we compare placement schemes with respect to bias reduction?

Design and reproducibility

- A confounded design leads to estimates of effects of interest that are sensitive to extraneous aspects of execution of the experiment
- Failure to record details of design may not interfere with concrete reproducibility, but questions of interpretability will persist
- Upshot: Take care with design and record and propagate as much information about it as possible (MAGE/MIAME/etc. are models)

Archive design and management

- How do you store your data? How do you document it? If you leave, how easy is it for co-workers to continue your progress? If you stop for a while, how easy is it to restart?
- Tradeoff: complex structures for archiving (e.g., MAGE-OM) are self-documenting but costly to learn/deploy; simple structures may not be recoverable when loosely coupled documentation on formatting is misplaced

Early Bioconductor strategy

- Persistent and performant data archives for microarray experiments can be values of R variables, so that any desired transformation of the experimental data can be programmed as a `straight' R function
- ExpressionSet class, instances respond to exprs(), pData(), fData(), annotation(), lmFit(), MLearn() and so on

Gaining mileage with container and workflow designs

- ExpressionSets for classic experiments (Spellman (yeast time course), Gasch (yeast stress), harbChIP (yeast ChIP), Golub (expression in cancer), Neve (expression + aCGH), MAQC, hmyriB36 (expression + hapmap), …)
- Using R packages for distribution allows coordination with documentation, particularly detailed **vignettes**

Bioconductor experimental data set archive

Bioconductor version 2.10 (Release)

- Software (553)
- AnnotationData (626)
- ExperimentData (124)
  - Cancer (20)
  - ChIPchipData (1)
  - ChIPseqData (3)
  - EColiData (1)
  - HapMap (7)
  - HighThroughputSequencingData (3)
  - HIV (1)
  - MassSpectrometryData (1)
  - NormalTissue (1)
  - RNAExpressionData (2)
  - RNAseqData (8)
  - StemCells (1)
  - Yeast (9)

Packages

Package	Maintainer	
affycompData	Harris Jaffee	affycomp
affydata	Harris Jaffee	Affymetri Purpose
AffymetrixDataTestFiles	Henrik Bengtsson	Affymetri PSI) for t
ALL	Robert Gentleman	A data pa
ALLMLL	B. M. Bolstad	A subset lymphobl
AmpAffyExample	Rafael A. Irizarry	Example
beadarrayExampleData	Mark Dunning	Example
BeadArrayUseCases	Mike Smith	Analysing data usin
beta7	Jean Yang	Rodrigue Expressio Cells Bea Integrin a
bladderbatch	Jeffrey T. Leek	Bladder g batch eff
breastCancerMAINZ	Markus Schroeder, Benjamin	Gene exp Schmidt

Drilling to RNAseqData

Bioconductor version 2.10 (Release)

- Software (553)
- AnnotationData (626)
- ExperimentData (124)
  - Cancer (20)
  - ChIPchipData (1)
  - ChIPseqData (3)
  - EColiData (1)
  - HapMap (7)
  - HighThroughputSequencingData (3)
  - HIV (1)
  - MassSpectrometryData (1)
  - NormalTissue (1)
  - RNAExpressionData (2)
  - RNAseqData (8)
  - StemCells (1)
  - Yeast (9)

Packages

Package	Maintainer	Title
cheung2010	Vince Carey	resources for genetics of gene expression bas 2010
GSVAdata	Robert Castelo	Data employed in the vignette of the GSVA pa
leeBamViews	VJ Carey	leeBamViews -- multiple yeast RNAseq sampl 2009
pasilla	Alejandro Reyes	Data package with per-exon and per-gene rea samples of Pasilla knock-down by Brooks et a 2011.
pasillaBamSubset	H. Pages	Subset of BAM files from "Pasilla" experiment
RnaSeqTutorial	Nicolas Delhomme	RNA-Seq Tutorial (EBI Cambridge UK, Octobe
tweeDEseqCountData	Juan R Gonzalez	RNA-seq count data employed in the vignette package
yeastRNASeq	J. Bullard	Yeast RNA-Seq Experimental Data from Lee e

Drilling to a tutorial data package

RnaSeqTutorial

RNA-Seq Tutorial (EBI Cambridge UK, October 2011)

Bioconductor version: Release (2.10)

A selection of RNA-Seq data to get familiar with the related Bioconductor core packages and the easyRNASeq package.

Author: Nicolas Delhomme, Ismael Padioleau

Maintainer: Nicolas Delhomme <delhomme at embl.de>

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("RnaSeqTutorial")
```

To cite this package in a publication, start R and enter:

```
citation("RnaSeqTutorial")
```

Documentation

PDF	R Script	RNA-Seq Tutorial
PDF		Reference Manual

Details

Drilling to the vignette

RNA-Seq Tutorial (EBI, October 2011)

Nicolas Delhomme

March 31, 2012

Contents

1 Introduction **2**

2 Walk through a single sample use case **4**
2.1 Reading the data 5
2.2 Filtering the data 10
2.3 Loading the annotation 11
2.4 Summarizing read counts per feature 16
2.5 Conclusion 18

3 Using easyRNASeq **19**
3.1 easyRNASeq 19

4 Advanced usage **22**
4.1 Normalizing counts 22
4.2 De-multiplexing samples 22

5 Visualizing the data **23**
5.1 exporting the coverage 23

A workflow schema in the vignette

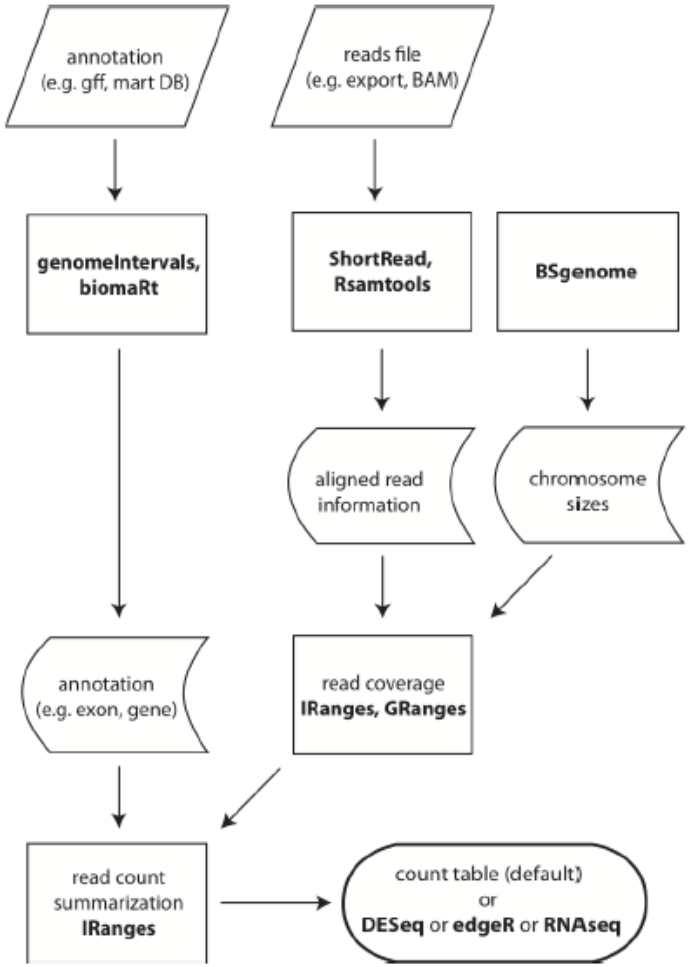


Figure 1: RNA-Seq Procedure Overview. The R packages used for the different steps are emphasized in bold face.

Illustration of a container

ShortRead *ShortRead* was the first NGS package developed to read in NGS data and is able to read almost every sequencer's manufacturer proprietary formats (with the notable exception of ABI color-space). First, an Illumina "export" file produced by a GenomeAnalyzer GAIIx will be read in; and then the same data set, in BAM format.

Illumina export The export file is read in using the `readAligned` function, and the resulting object displayed.

```
> library(ShortRead)
> aln<-readAligned(
+                  system.file("extdata",package="RnaSeqTutorial"),
+                  pattern="subset_export",type="SolexaExport")
> show(aln)

class: AlignedRead
length: 100000 reads; width: 36 cycles
chromosome: NM 1:0:0 ... chr2R chr2L
position: NA NA ... 20555556 13903608
strand: NA NA ... + -
alignQuality: NumericQuality
alignData varLabels: run lane ... filtering contig
```

Upshots

- Mature analyses involve substantial data reduction and interesting analyses and visualizations
- These can be organized into R packages with self-contained documentation according to established protocols
- Inclusion with Bioconductor supports portability and distribution
- Even if you won't distribute your data package, using the discipline or one functionally equivalent has substantial benefits

Technical steps for a data package (public or private)

- Choosing cooked data representation and its incorporation into a package
  - If a class instance in a .rda serialization, could go stale
  - Could store raw and use .onLoad to populate container obeying current definition
- Building package and documentation
- Building a good vignette

What is a vignette?

- Two basic criteria for a document
  - Narrates a multistep, multicomponent analysis, providing details of computation and interpretation
  - Is computable, can raise error conditions, has verifiable output
- Typically a vignette is composed in Sweave (LaTeX + R) and placed in a specific package folder
- Alternatives to LaTeX are available if necessary

Constructing a package and a vignette

- package.skeleton() will create folders and templates for documentation
- Composition with LaTeX involves a markup that can be complex, can have a tutorial session if desired
- Bridging to Sweave involves ``literate programming'', where LaTeX narration is interlarded with escapes to R code

```
\documentclass[a4paper]{article}

\begin{document}

<<echo=false,results=hide>>=
library(lattice)
library(xtable)
data(cats, package="MASS")
@

\section*{The Cats Data}

Consider the \texttt{cats} regression example from Venables \& Ripley
(1997). The data frame contains measurements of heart and body weight
of \Sexpr{nrow(cats)} cats (\Sexpr{sum(cats$Sex=="F")} female,
\Sexpr{sum(cats$Sex=="M")} male).

A linear regression model of heart weight by sex and gender can be
fitted in R using the command
<<>>=
lm1 = lm(Hwt~Bwt*Sex, data=cats)
lm1
@
```

The Cats Data

Consider the `cats` regression example from Venables & Ripley (1997). The data frame contains measurements of heart and body weight of 144 cats (47 female, 97 male).

A linear regression model of heart weight by sex and gender can be fitted in R using the command

```
> lm1 = lm(Hwt ~ Bwt * Sex, data = cats)
> lm1

Call:
lm(formula = Hwt ~ Bwt * Sex, data = cats)

Coefficients:
(Intercept)          Bwt         SexM     Bwt:SexM
      2.981        2.636       -4.165        1.676
```

Tests for significance of the coefficients are shown in Table 1, a scatter plot including the regression lines is shown in Figure 1.

Fostering reproducibility and extensibility of analysis work

- Use well-annotated containers to manage complexity of inputs
- Use versioned metadata to manage effects of external evolution of biological knowledge
- Drive the analysis with one or more Sweave-based vignettes, so that the main computations are scripted and runnable via Sweave()
- Use caching to reduce recomputation of complex objects when the steps leading to their creation are sound

Recap

- Data/analysis flows from genome scale experiments are complex and require detailed management
- Good experimental design minimizes bias and extraneous variation: substantive reproducibility
- Container/package/vignette disciplines reduce organizational and recovery complexity, foster concrete reproducibility and extensibility

A high profile paper and some reproduction/extensibility exercises

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

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

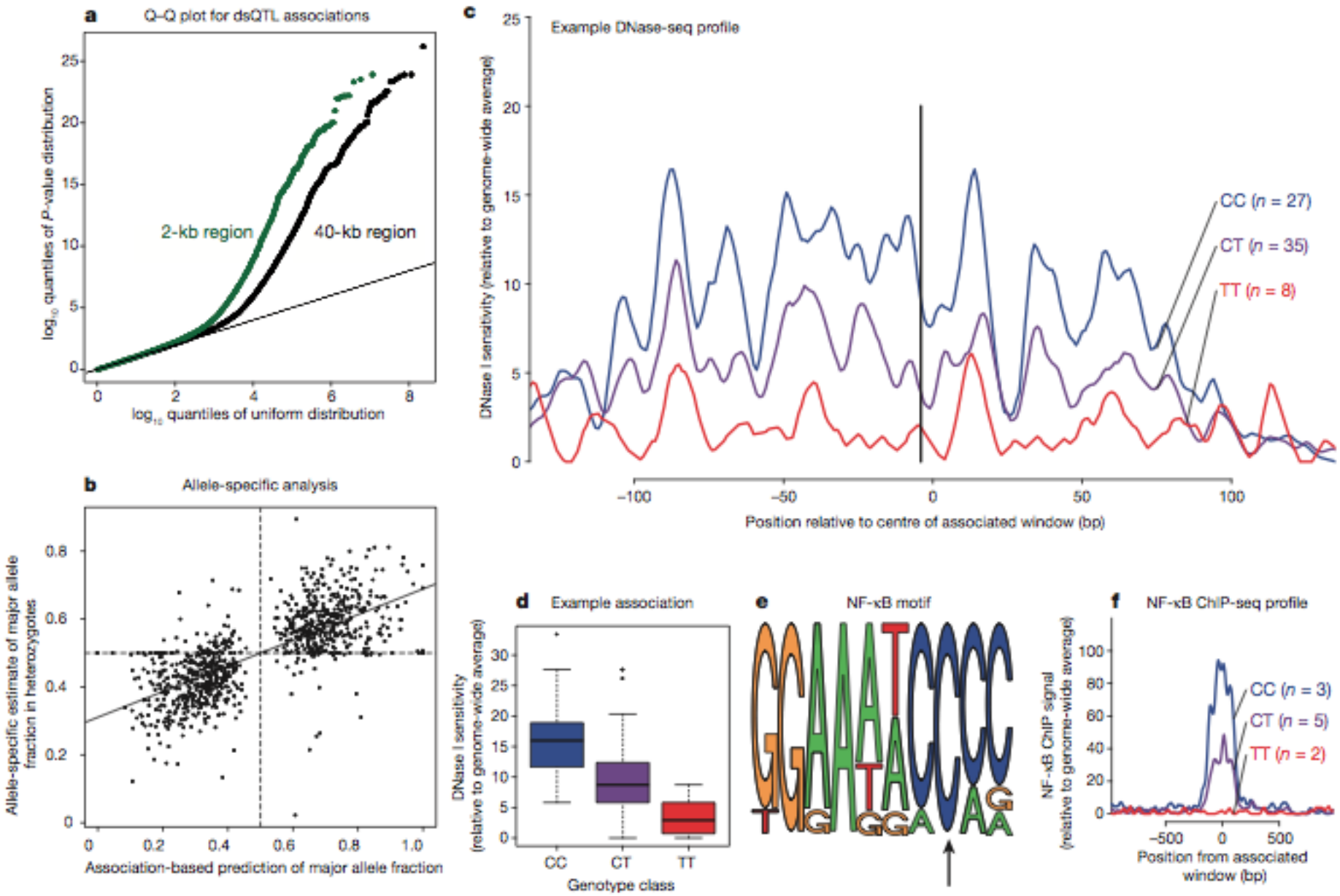


Figure 1 | Genome-wide identification of dsQTLs and a typical example. **a**, Q–Q plots for all tests of association between DNase I cut rates in 100-bp windows, and variants within 2-kb (green) and 40-kb (black) regions centred dsQTL (rs4953223). The black line indicates the position of the associated SNP. **d**, Box plot showing that rs4953223 is strongly associated with local chromatin accessibility ($P = 3 \times 10^{-13}$). **e**, The T allele, which is associated with low

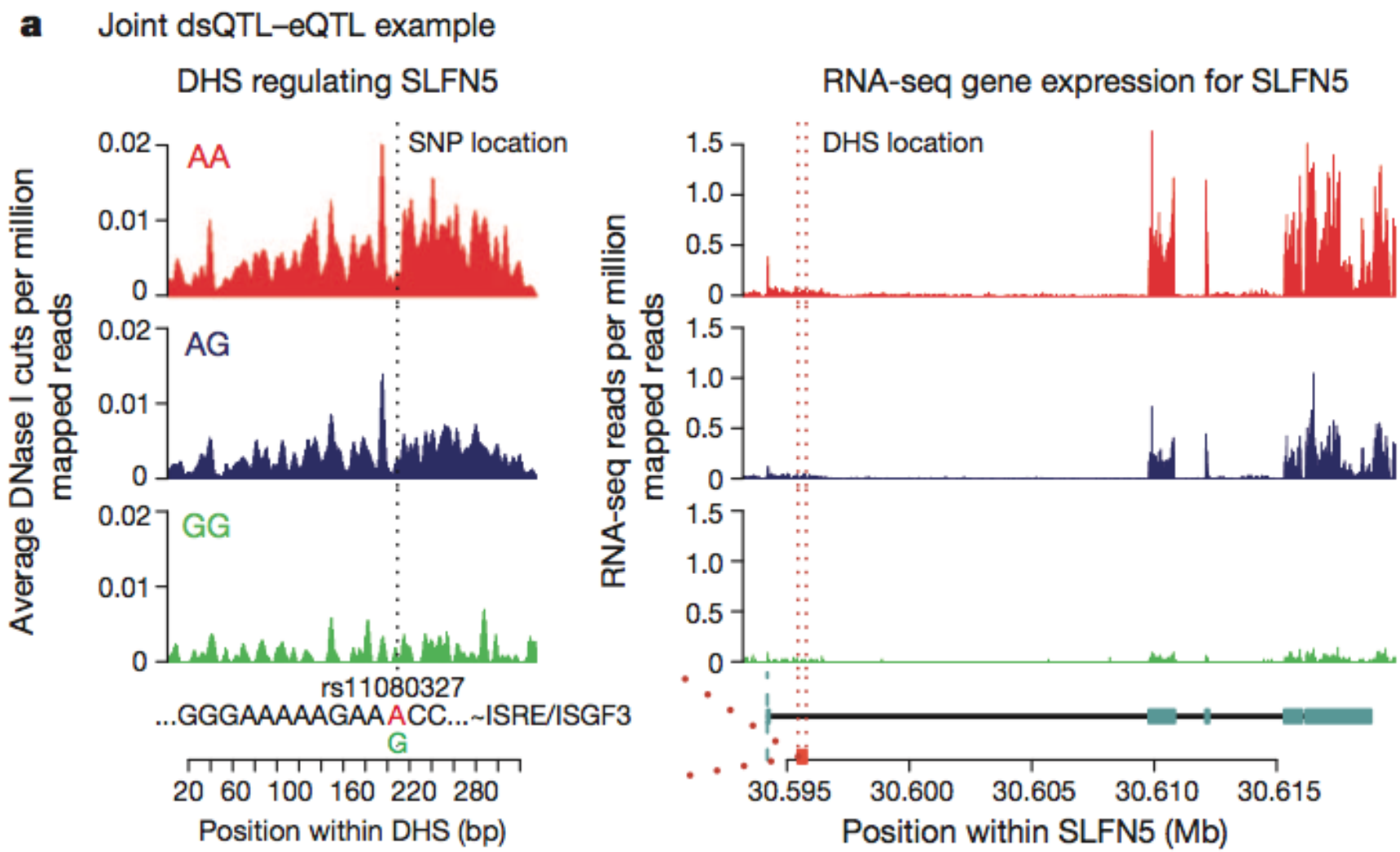
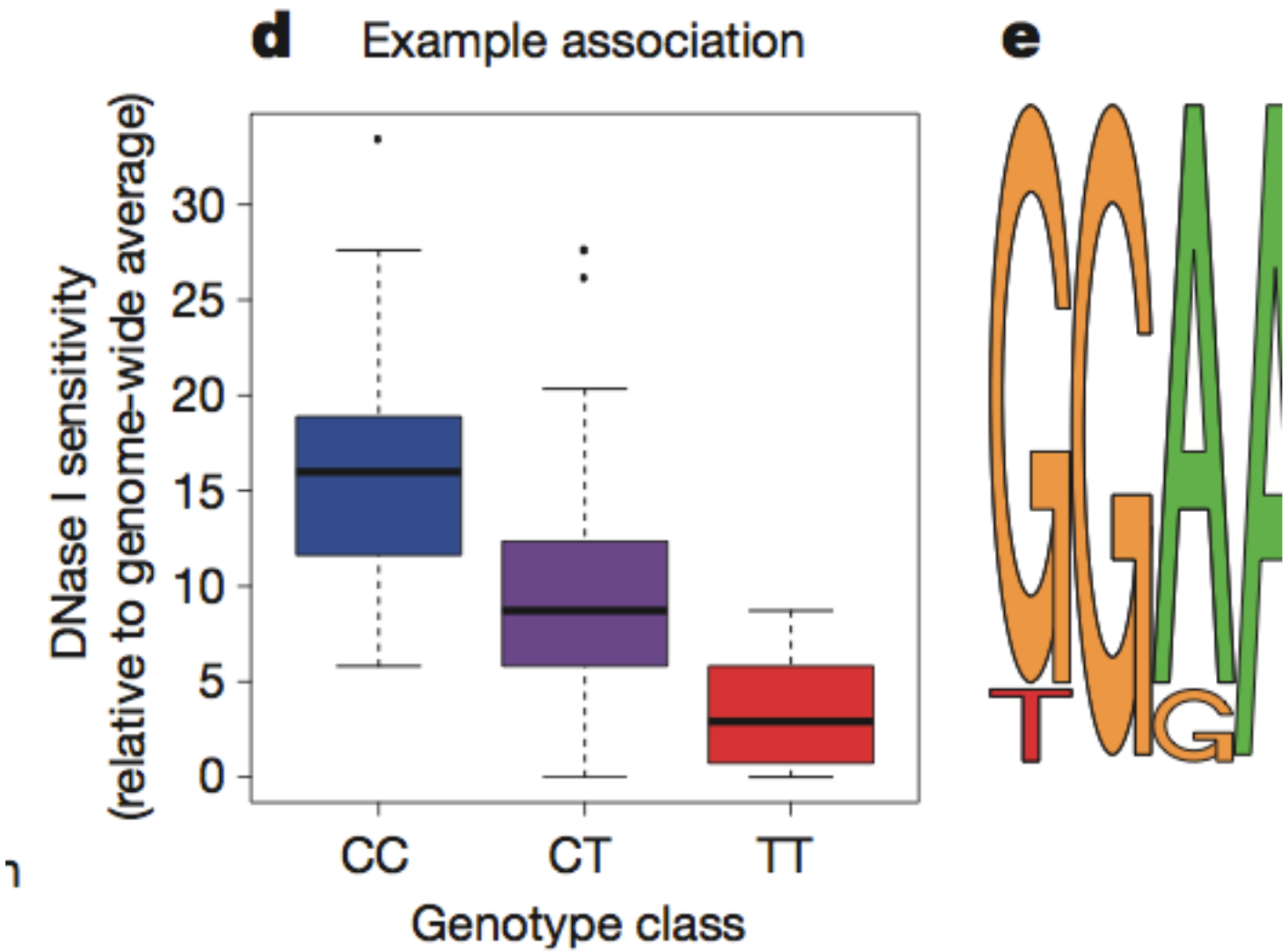


Figure 3 | Relationship between dsQTLs and eQTLs. a, Example of a dsQTL SNP that is also an eQTL for the gene *SLFN5*. The SNP disrupts an interferon- (right) measurer genotype at the p

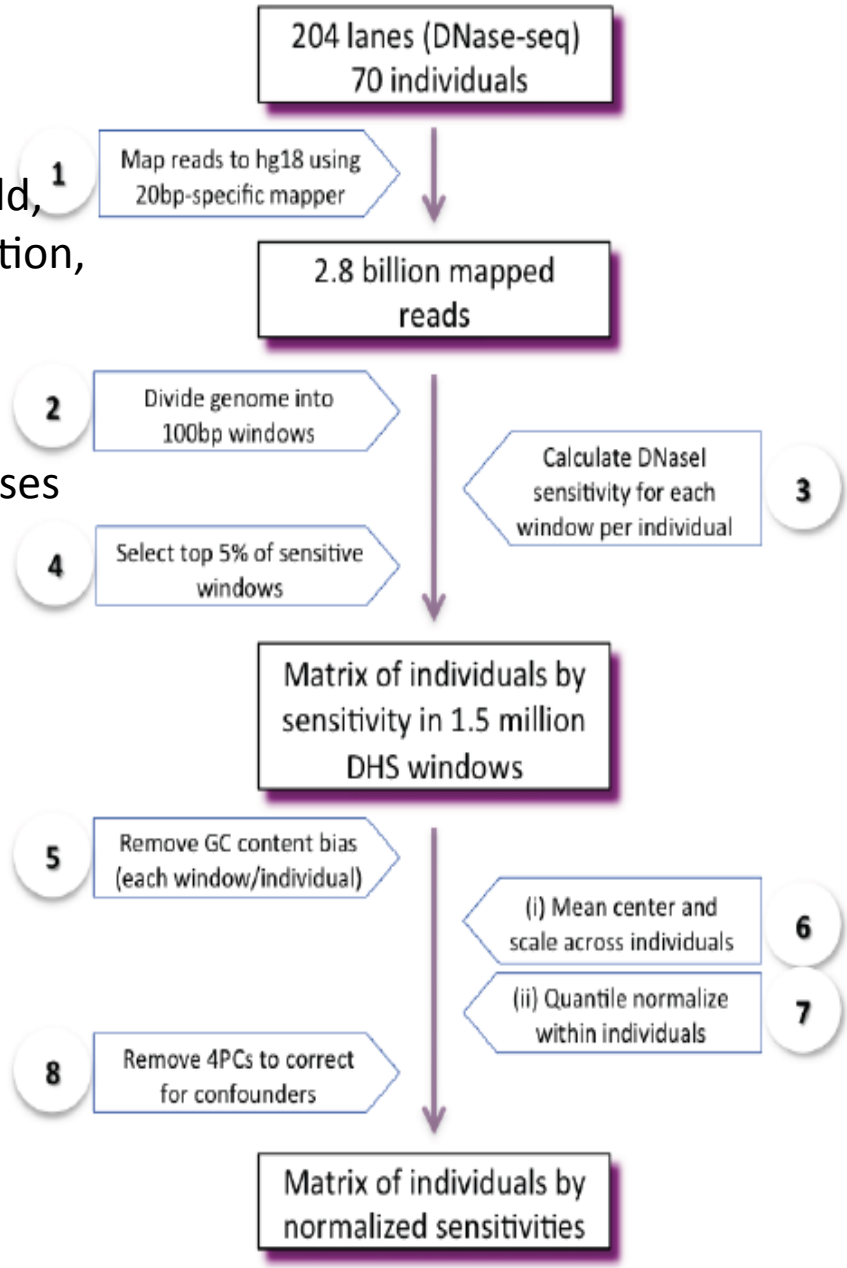
What can we do to make this finding concretely reproducible? Extensible?



and a typical example. dsQTL (rs4953223).

Vulnerabilities: hg18,
100bp window,
5% sensitivity threshold,
GC content bias reduction,
4 PC (SVA-like),
cis radius

Some sensitivity analyses described



New directions in feature/test volume?

- DNase-seq read-counts were assembled in a 100bp tiling of the genome, so 30 million scores per individual
- dsQTL analysis involves associating ~30 million imputed SNP with each of these scores; cis filtering reduces volume considerably
- How should we manage the basic quantities?
  - Stage 1: SummarizedExperiment demo
  - Stage 2: integrative DHS+genotype container permitting very high-volume testing with small footprint

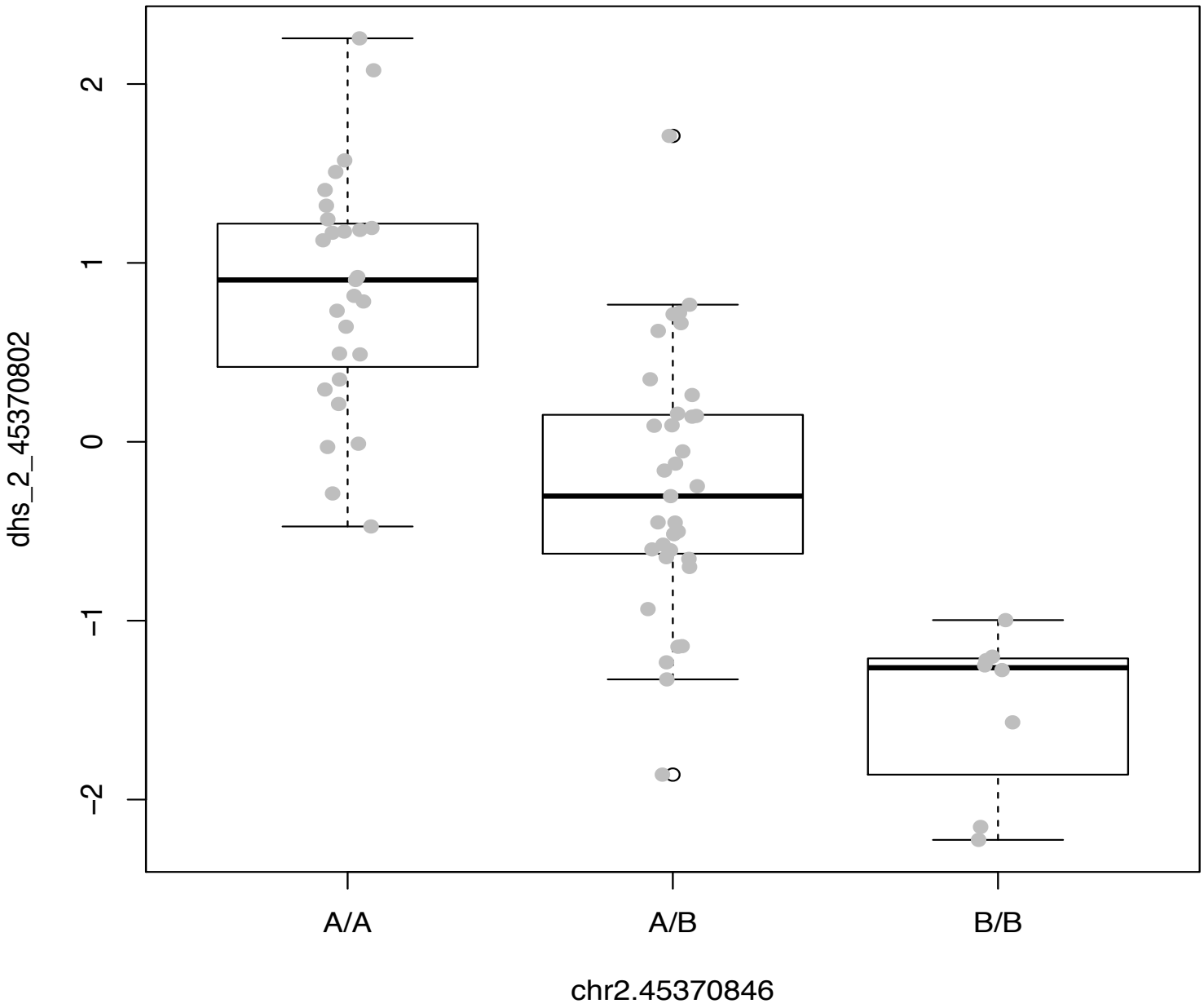
The dsQTL experimental data package

```
> data(DSQ_2)
> DSQ_2
class: SummarizedExperiment
dim: 96024 70
exptData(2): MIAME annotation
assays(1): normDHS
rownames(96024): dhs_2_1202 dhs_2_1602 ... dhs_2_242737902
  dhs_2_242739902
rowData values names(0):
colnames(70): NA18486 NA18498 ... NA19239 NA19257
colData names(9): naid one ... male isFounder
> exptData(DSQ_2)[["MIAME"]]
Experiment data
  Experimenter name: Degner JF
  Laboratory: Department of Human Genetics, University of Chicago, Chicago,
nois 60637, USA.
  Contact information:
  Title: DNaseI sensitivity QTLs are a major determinant of human expressio
iation.
  URL:
  PMIDs: 22307276

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

```
  abstract: A 252 word abstract is available. Use 'abstract' method.
> assays(DSQ_2)[["normDHS"]][1:5,1:5]
               NA18486     NA18498    NA18499    NA18501    NA18502
dhs_2_1202 -0.2684343 -0.78076674 -0.4840237  2.3894003 -1.0813642
dhs_2_1602 -1.4445813  0.92170439  0.5812017  0.8627376  0.5186581
dhs_2_2002  0.7624075 -0.12340745 -1.1821308  1.4253179  0.3125592
dhs_2_7502  0.1242963  0.60788505  0.6754706 -0.0452303  0.4876332
dhs_2_8802 -0.9554503 -0.06016578 -0.1990696  1.9383937 -1.3758668
> rowData(DSQ_2)[1:5,]
GRanges with 5 ranges and 0 elementMetadata cols:
             seqnames       ranges strand
                <Rle>    <IRanges>  <Rle>
  dhs_2_1202     chr2 [1202, 1301]      *
  dhs_2_1602     chr2 [1602, 1701]      *
  dhs_2_2002     chr2 [2002, 2101]      *
  dhs_2_7502     chr2 [7502, 7601]      *
  dhs_2_8802     chr2 [8802, 8901]      *
  ---
  seqlengths:
   chr2
     NA
```

```
> subsetByOverlaps( rowData(DSQ_2), GRanges("chr2", IRanges(1000,2000)))
GRanges with 2 ranges and 0 elementMetadata cols:
             seqnames       ranges strand
                <Rle>    <IRanges>  <Rle>
  dhs_2_1202     chr2 [1202, 1301]      *
  dhs_2_1602     chr2 [1602, 1701]      *
  ---
  seqlengths:
   chr2
     NA
> DSQ_2[ which(rowData(DSQ_2) %in% GRanges("chr2", IRanges(1000,2000))),
+   which(colData(DSQ_2)$male == TRUE) ]
class: SummarizedExperiment
dim: 2 28
exptData(2): MIAME annotation
assays(1): normDHS
rownames(2): dhs_2_1202 dhs_2_1602
rowData values names(0):
colnames(28): NA18501 NA18504 ... NA19223 NA19239
colData names(9): naid one ... male isFounder
>
```


dhs_2_45370802
2
1
0
−1
−2
A/A
A/B
B/B
chr2.45370846

Recap

- Tight binding of metadata to assay data for many millions of features per sample
- Fast, idiomatic query resolution using genomic coordinates
- X[G, S] has values for selected features and samples, responds to any method on X
- Relax restrictions on the “back end” when the resources are really massive
- Often the cooked resources are manageable and can reside in such containers, facilitating easy distribution and uptake: extensibility

Remaining issues

- I have focused on adding value by providing readily distributed and manipulated images of complete analyses – bioc experimental data packages
- These are relatively costly to generate but simplify checking, understanding, perturbing **what was done**
- **How early in analysis should we be asking that such images be present? Cost vs. benefit**
- Scripts+files+text documents can provide equivalent information but momentum can be hard to develop