eQTL analysis – an approach with Bioconductor

Vincent Carey
Channing Division of Network Medicine
Harvard Medical School

Road map

- Background on scope and implications of eQTL studies
- Data structures suitable for current studies (RNA-seq and genome-wide DNA sequencing/ imputation [SNP only])
- Demonstration with Bioconductor
  - RNA-seq FPKM in geuvPack (special for this course)
  - 1000 Genomes VCF in an Amazon S3 bucket
  - “Batch effect” correction and analysis

Genetic analysis of genome-wide variation in human gene expression

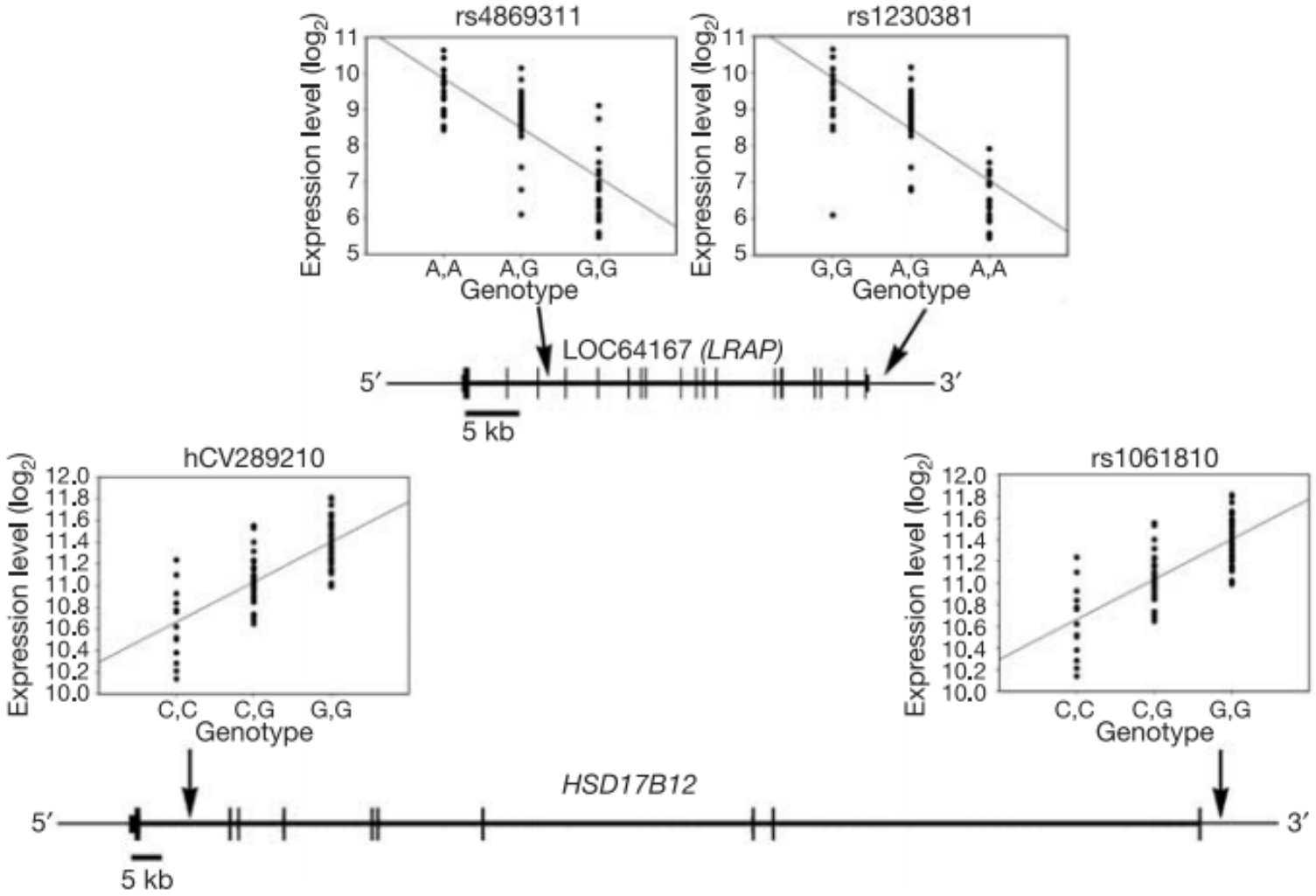
Michael Morley[1,3]*, Cliona M. Molony[2]*, Teresa M. Weber[1,3], James L. Devlin[2], Kathryn G. Ewens[2], Richard S. Spielman[2] & Vivian G. Cheung[1,2,3]

[1]*Department of Pediatrics and* [2]*Department of Genetics, University of Pennsylvania,*
[3]*The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA*

* These authors contributed equally to this work

Natural variation in gene expression is extensive in humans and other organisms, and variation in the baseline expression level of many genes has a heritable component. To localize the genetic determinants of these quantitative traits (expression phenotypes) in humans, we used microarrays to measure gene expression levels and performed genome-wide linkage analysis for expression levels of 3,554 genes in 14 large families. For approximately 1,000 expression phenotypes, there was significant evidence of linkage to specific chromosomal regions. Both *cis*- and *trans*-acting loci regulate variation in the expression levels of genes, although most act *in trans*. Many gene expression phenotypes are influenced by several genetic determinants. Furthermore, we found hotspots of transcriptional regulation where significant evidence of linkage for several expression phenotypes (up to 31) coincides, and expression levels of many genes that share the same regulatory region are significantly correlated. The combination of microarray techniques for phenotyping and linkage analysis for quantitative traits allows the genetic mapping of determinants that contribute to variation in human gene expression.

Nature 2004, PMID 15269782



ssion of expression phenotype of LOC64167 and *HSD17B12* on nearby SNPs. For LOC64167, the distance between SNP markers rs4869311 and rs123038
the distance between hCV289210 and rs1061810 is 172 kb.

ARTICLE

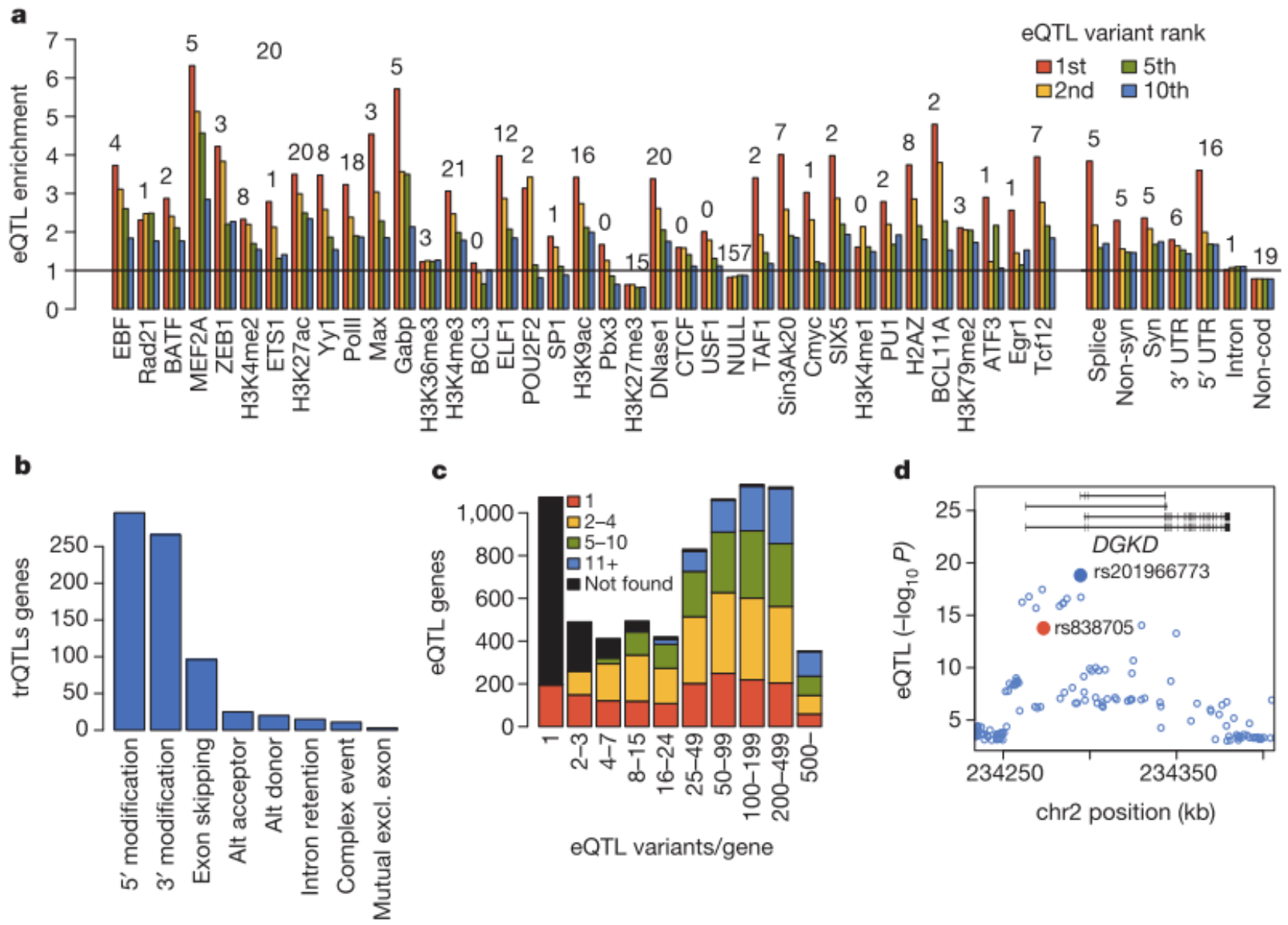
PMID 24037378

doi:10.1038/nature12531

Transcriptome and genome sequencing uncovers functional variation in humans

Tuuli Lappalainen[1,2,3], Michael Sammeth[4,5,6,7]†*, Marc R. Friedländer[5,6,7,8]*, Peter A. C. 't Hoen[9]*, Jean Monlong[5,6,7]*, Manuel A. Rivas[10]*, Mar Gonzàlez-Porta[11], Natalja Kurbatova[11], Thasso Griebel[4], Pedro G. Ferreira[5,6,7], Matthias Barann[12], Thomas Wieland[13], Liliana Greger[11], Maarten van Iterson[9], Jonas Almlöf[14], Paolo Ribeca[4], Irina Pulyakhina[9], Daniela Esser[12], Thomas Giger[1], Andrew Tikhonov[11], Marc Sultan[15], Gabrielle Bertier[5,6], Daniel G. MacArthur[16,17], Monkol Lek[16,17], Esther Lizano[5,6,7,8], Henk P. J. Buermans[9,18], Ismael Padioleau[1,2,3], Thomas Schwarzmayr[13], Olof Karlberg[14], Halit Ongen[1,2,3], Helena Kilpinen[1,2,3], Sergi Beltran[4], Marta Gut[4], Katja Kahlem[4], Vyacheslav Amstislavskiy[15], Oliver Stegle[11], Matti Pirinen[10], Stephen B. Montgomery[1]†, Peter Donnelly[10], Mark I. McCarthy[10,19], Paul Flicek[11], Tim M. Strom[13,20], The Geuvadis Consortium‡, Hans Lehrach[15,21], Stefan Schreiber[12], Ralf Sudbrak[15,21]†, Ángel Carracedo[22], Stylianos E. Antonarakis[1,2], Robert Häsler[12], Ann-Christine Syvänen[14], Gert-Jan van Ommen[9], Alvis Brazma[11], Thomas Meitinger[13,20,23], Philip Rosenstiel[12], Roderic Guigó[5,6,7], Ivo G. Gut[4], Xavier Estivill[5,6,7,8] & Emmanouil T. Dermitzakis[1,2,3]

Genome sequencing projects are discovering millions of genetic variants in humans, and interpretation of their functional effects is essential for understanding the genetic basis of variation in human traits. Here we report sequencing and deep analysis of messenger RNA and microRNA from lymphoblastoid cell lines of 462 individuals from the 1000 Genomes Project—the first uniformly processed high-throughput RNA-sequencing data from multiple human populations with high-quality genome sequences. We discover extremely widespread genetic variation affecting the regulation of most genes, with transcript structure and expression level variation being equally common but genetically largely independent. Our characterization of causal regulatory variation sheds light on the cellular mechanisms of regulatory and loss-of-function variation, and allows us to infer putative causal variants for dozens of disease-associated loci. Altogether, this study provides a deep understanding of the cellular mechanisms of transcriptome variation and of the landscape of functional variants in the human genome.



nscriptome QTLs. **a**, Enrichment of EUR exon eQTLs in notations for the first, second, fifth and tenth best associating per gene, relative to a matched null set of variants denoted by the . The numbers are $-\log_{10}(P$ values) of a Fisher test between the caused by transcript ratio QTLs. **c**, The rank of the best Omni 2.5 the significant EUR eQTL variants per gene. **d**, The *DGKD* gene an intronic SNP rs838705 is associated with calcium levels (red eQTL variant 21 kb downstream (blue) is a very likely causal va

doi: 10.1093/hmg/ddu226

Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types

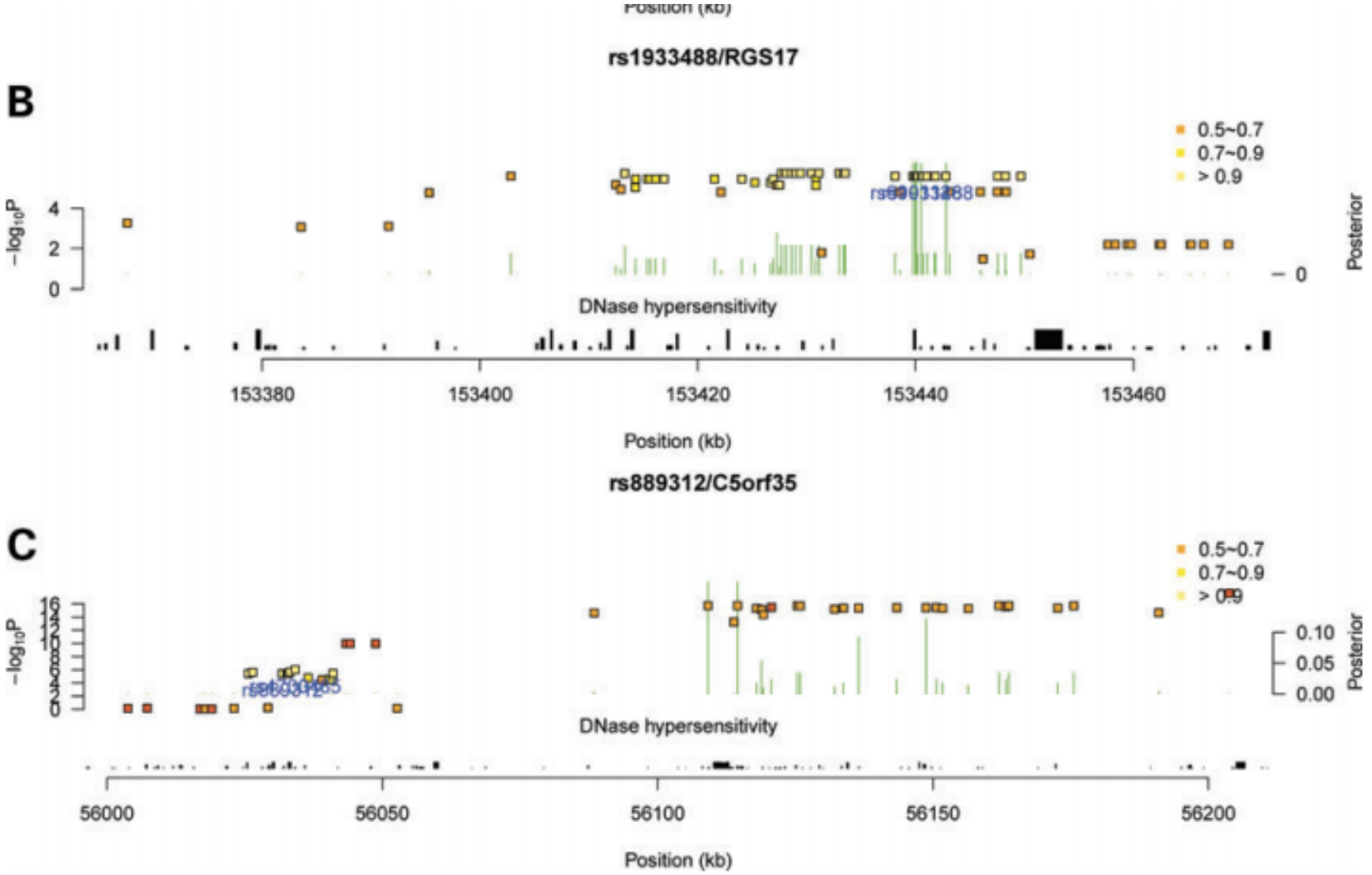
Qiyuan Li[1,2,3], Alexander Stram[4], Constance Chen[5], Siddhartha Kar[6], Simon Gayther[7], Paul Pharoah[6], Christopher Haiman[4], Barbara Stranger[8], Peter Kraft[5] and Matthew L. Freedman[1,3,*]

[1]Department of Medical Oncology, The Center for Functional Cancer Epigenetics, Dana Farber Cancer Institute, Boston, MA, USA, [2]Medical College of Xiamen University, Xiamen, China, [3]Program in Medical and Population Genetics, The Broad Institute, Cambridge, MA, USA, [4]University of Southern California, Los Angeles, CA, USA, [5]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA, [6]Strangeways Research Laboratory, University of Cambridge, Cambridge, UK, [7]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Log Angeles, CA, USA and [8]Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL, USA

Received February 17, 2014; Revised April 17, 2014; Accepted May 6, 2014

The majority of trait-associated loci discovered through genome-wide association studies are located outside of known protein coding regions. Consequently, it is difficult to ascertain the mechanism underlying these variants and to pinpoint the causal alleles. Expression quantitative trait loci (eQTLs) provide an organizing principle to address both of these issues. eQTLs are genetic loci that correlate with RNA transcript levels. Large-scale data sets such as the Cancer Genome Atlas (TCGA) provide an ideal opportunity to systematically evaluate eQTLs as they have generated multiple data types on hundreds of samples. We evaluated the determinants of gene expression (germline variants and somatic copy number and methylation) and performed *cis*-eQTL ana-

PMID 24907074



ation of the fine-mapping candidates of three cancer risk loci based on an integrated posterior probability combining association
sents a correlated germline variant of the initially reported risk locus (labeled by blue text); the height of the points corresponds
$\log_{10} P$ values); the DNaseI HS scores are shown beneath the posterior; the green bars show the posterior probabilities. (**A**) Xp11
tate cancer; (**B**) 6q25.2/rs1933488 with*RGS17* in prostate cancer and (**C**) 5q11.2/rs889312 with *C5orf35*in ER-positive breast c

www.sciencemag.org SCIENCE VOL 343 7 MARCH 2014
Published by AAAS

RESEAR

Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells

Mark N. Lee,[1,2,3]* Chun Ye,[1]* Alexandra-Chloé Villani,[1,2] Towfique Raj,[1,2,4] Weibo Li,[1,3] Thomas M. Eisenhaure,[1,3] Selina H. Imboywa,[2] Portia I. Chipendo,[2] F. Ann Ran,[1,5,6,7,8] Kamil Slowikowski,[9] Lucas D. Ward,[1,10] Khadir Raddassi,[11] Cristin McCabe,[1,4] Michelle H. Lee,[2] Irene Y. Frohlich,[2] David A. Hafler,[8] Manolis Kellis,[1,10] Soumya Raychaudhuri,[1,2,12,13] Feng Zhang,[6,7,8] Barbara E. Stranger,[14,15] Christophe O. Benoist,[2] Philip L. De Jager,[1,2,4] Aviv Regev,[1,16,17]† Nir Hacohen[1,2,3]†

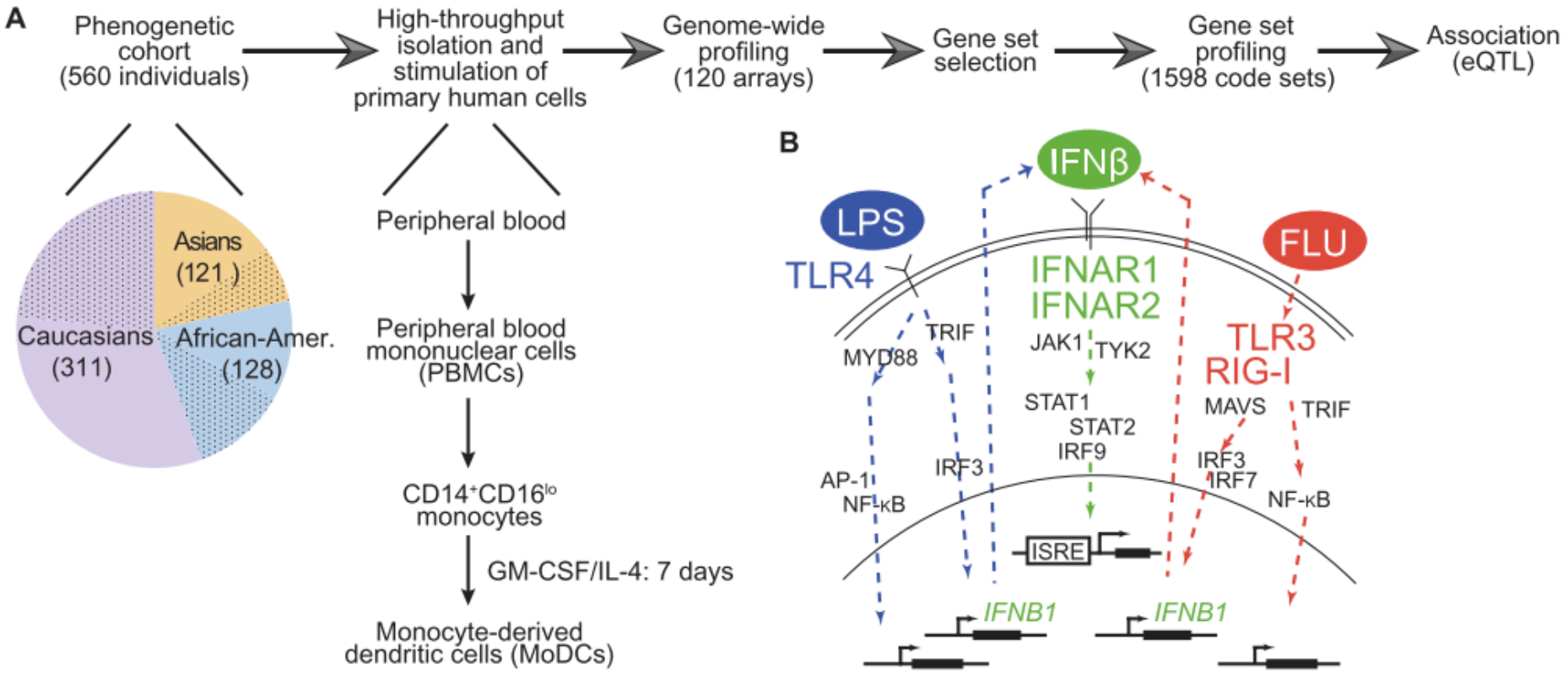
Little is known about how human genetic variation affects the responses to environmental stimuli in the context of complex diseases. Experimental and computational approaches were applied to determine the effects of genetic variation on the induction of pathogen-responsive genes in human dendritic cells. We identified 121 common genetic variants associated in cis with variation in expression responses to *Escherichia coli* lipopolysaccharide, influenza, or interferon-β (IFN-β). We localized and validated causal variants to binding sites of pathogen-activated STAT (signal transducer and activator of transcription) and IRF (IFN-regulatory factor) transcription factors. We

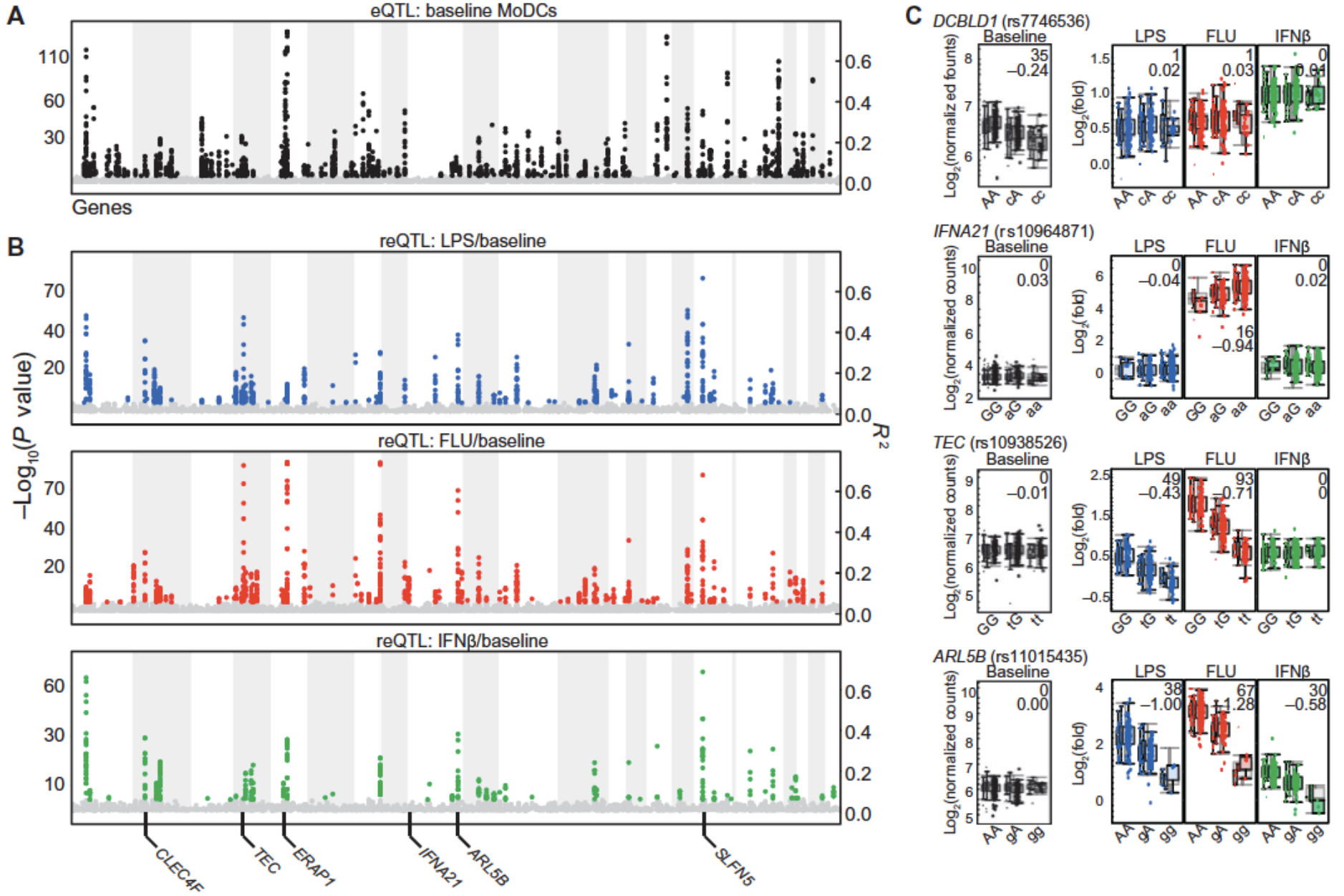
cleus to ind
including ir
gages the ty
the expressi
Genetic stu
iants near r
risk of diffe
DCs also p
immune res
eases (*11–1*
wide associa
(*14–17*), espe

Results

Assessing t Pathogen S

We develop
computatio
variability i
this variabi
First, we op
to isolate pri
human bloo


A
Phenogenetic cohort (560 individuals)
High-throughput isolation and stimulation of primary human cells
Genome-wide profiling (120 arrays)
Gene set selection
Gene set profiling (1598 code sets)
Association (eQTL)
Asians (121)
Caucasians (311)
African-Amer. (128)
Peripheral blood
Peripheral blood mononuclear cells (PBMCs)
CD14+CD16lo monocytes
GM-CSF/IL-4: 7 days
Monocyte-derived dendritic cells (MoDCs)
B
IFNβ
LPS
TLR4
FLU
IFNAR1
IFNAR2
TRIF
MYD88
JAK1
TYK2
TLR3
RIG-I
STAT1
STAT2
IRF9
MAVS
TRIF
IRF3
IRF7
AP-1
NF-κB
IRF3
NF-κB
ISRE
IFNB1
IFNB1


A
eQTL: baseline MoDCs
Genes
B
reQTL: LPS/baseline
reQTL: FLU/baseline
reQTL: IFNβ/baseline
−Log10(P value)
R2
CLEC4F
TEC
ERAP1
IFNA21
ARL5B
SLFN5
C
DCBLD1 (rs7746536)
IFNA21 (rs10964871)
TEC (rs10938526)
ARL5B (rs11015435)
Baseline
LPS
FLU
IFNβ
Log2(normalized counts)
Log2(fold)
AA
cA
cc
GG
aG
aa
tG
tt
gA
gg

Summary

- 2004: Microarray methods can expose genetic variants associated with expression variation
- 2011-2013: RNA-seq applications expose genetic sources of detailed variations in transcription (e.g., alternate splicing, allele-specific expression)
- 2014+: Tissue- and environment-specificity of eQTL associations to the fore
- Upshot – more eQTL experiments, reference data, detailed annotation for integrative use

Major concerns of this tutorial

- How can I do my own eQTL inferences?
- How can I use eQTL inferences obtained elsewhere?
- These questions imply concern with
  - Efficient representation of inputs to eQTL analyses
  - Efficient computation of eQTL inferences
  - Efficient representation of eQTL inferences

Difficult feature of eQTL analyses

- For the comprehensive analysis of trans-eQTL associations, we are considering |G| x |S| tests, which ranges into the hundreds of billions
  - You are *creating* big data even if what you are starting with is modest in volume
  - *After you've done the tests*, interactive general query resolution into tables with millions of records requires computing methods that are somewhat unusual relative to classical statistical computing
  - You will want to use scalable computing strategies as the standard tests are embarassingly parallel
    - A typical approach (see MatrixEQTL) is to discard information on apparently insignificant associations, given a prior threshold, as soon as possible in the analysis

Representing the input data on Expression

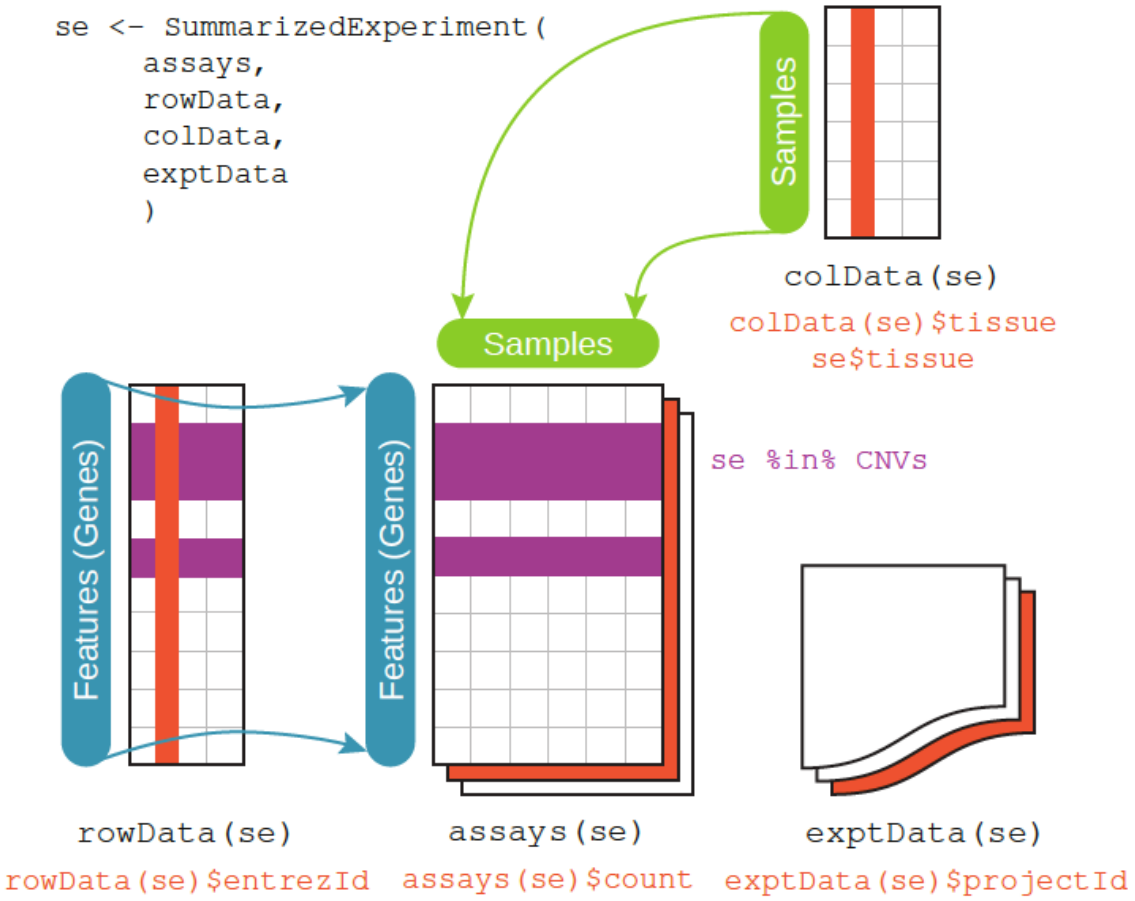


Figure 4: Elements from which *SummarizedExperiment* objects are composed.

The “FPKM” representation of expression in GEUVADIS samples

```
Console ~/
R » library(geuvPack)
R » data(geuFPKM)
R » geuFPKM
class: SummarizedExperiment
dim: 23722 462
exptData(2): MIAME constrHist
assays(1): exprs
rownames(23722): ENSG00000152931.6 ENSG00000183696.9 ...
  ENSG00000257337.1 ENSG00000177494.5
rowData metadata column names(18): source type ... tag ccdsid
colnames(462): HG00096 HG00097 ... NA20826 NA20828
colData names(0):
R » exptData(geuFPKM)$MIAME
Experiment data
  Experimenter name: Lappalainen T
  Laboratory: NA
  Contact information:
  Title: Transcriptome and genome sequencing uncovers functional variation in humans.
  URL:
  PMIDs: 24037378
```

Moderately self-documenting; currently lacks ‘sample info’

Quiz questions

- Q1. Use dim(assay(geuFPKM)) to determine G (number of 'genes') and N (number of samples).
- Q2. Use rowData(geuFPKM)… to determine the name of the gene measured in the first row of the assay data in geuFPKM. Note the type of gene.
- Q3. Tabulate the gene types assayed in the experiment. Hint – use mcols()
- Q4. Obtain a histogram of the distribution of expression values for this gene. Is a Gaussian model reasonable for this sample?

Q5. How can you verify the coordinates of DGKD using information in geuFPKM?

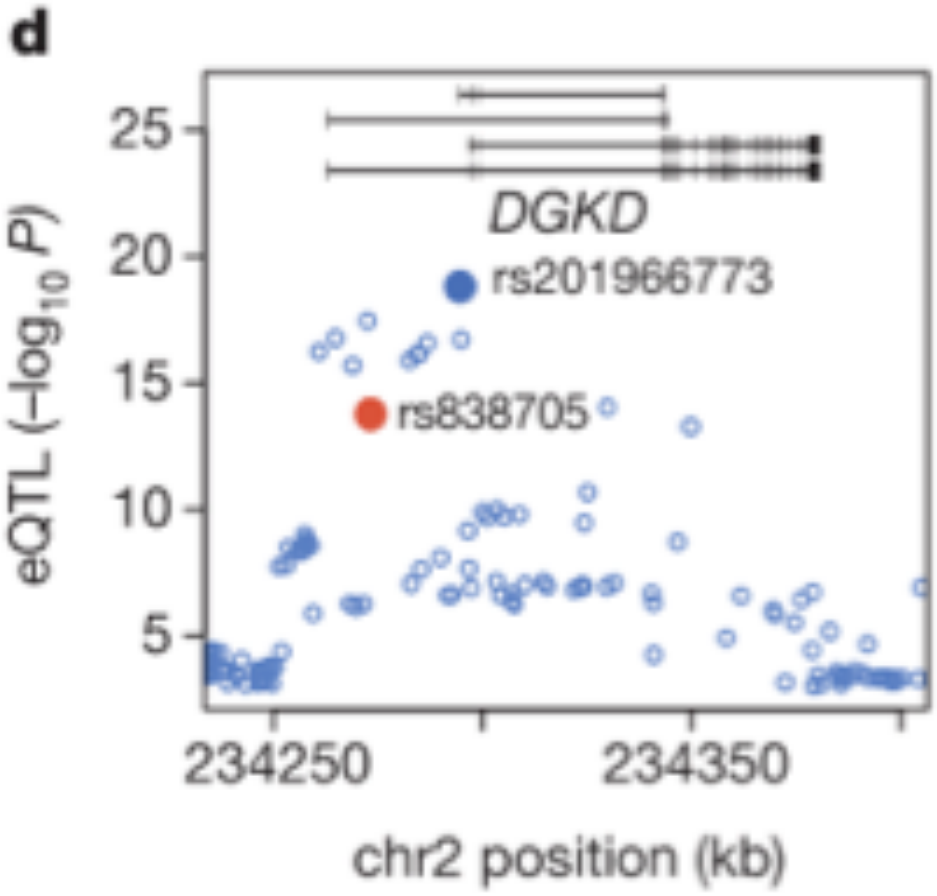


Figure 2d from the Geuvadis paper, PMID 24037378

Tools for approximate reproduction of Figure 2d

- We find the GEUVADIS quantification of DGKD in geuFPKM
- http://1000genomes.s3.amazonaws.com/release/20110521/ALL.chr2.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz is the URL of a Tabix-indexed VCF with genotype calls from 1000 genomes
- GGtools cisAssoc() will use these two resources to compute association tests

Conceptual setup

- Perform an adjustment to expression data to remove technical variation
  - GEUVADIS used “PEER” to remove 10 latent factors; we’ll use 10 PC … should be followed by sensitivity analysis; SVA is also of interest
- Compute association statistics of interest (in this case, we’ll use a cis radius of 50k)
- Correct p-values for multiple testing (we’ll use parametric testing; see the eQTL workflow for nonparametric approach)

Obtain cis eQTL for DGKD and one other gene on chr2

bfac are 10 PC from expression FPKM

```
R » load("bfac.rda")
R » data(geuFPKM)
R » colData(geuFPKM) = cbind(colData(geuFPKM), DataFrame(bfac))
R » source("buckpath.R")
R » cd = cisAssoc(geuFPKM[c(14894,42)], TabixFile(buckpath(2)), lbmaf=.05,
    cisradius=50000, rhs=~PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10)
```

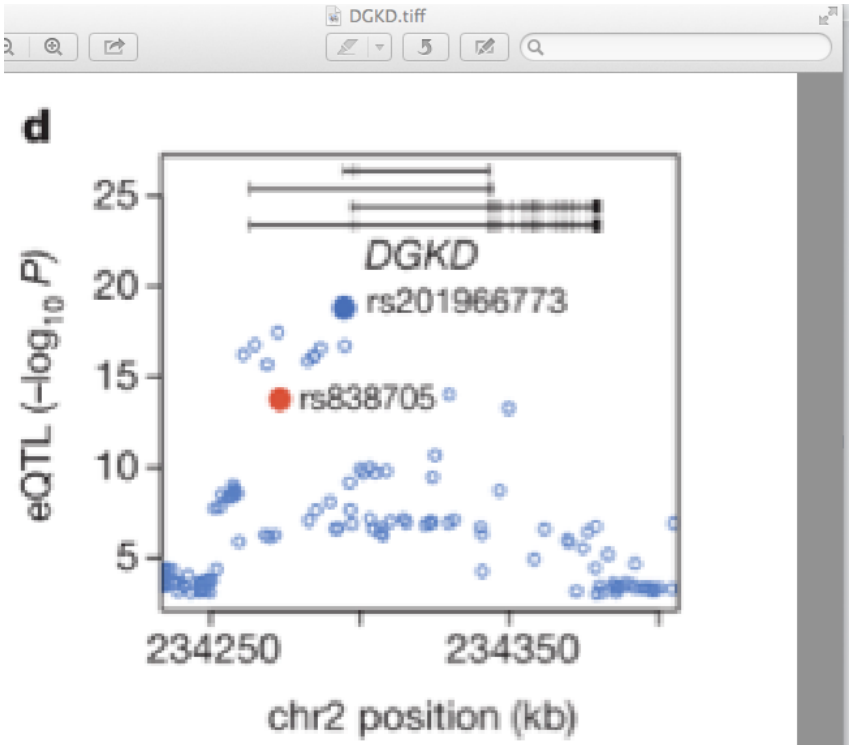
Output of cisAssoc (first 3 records)

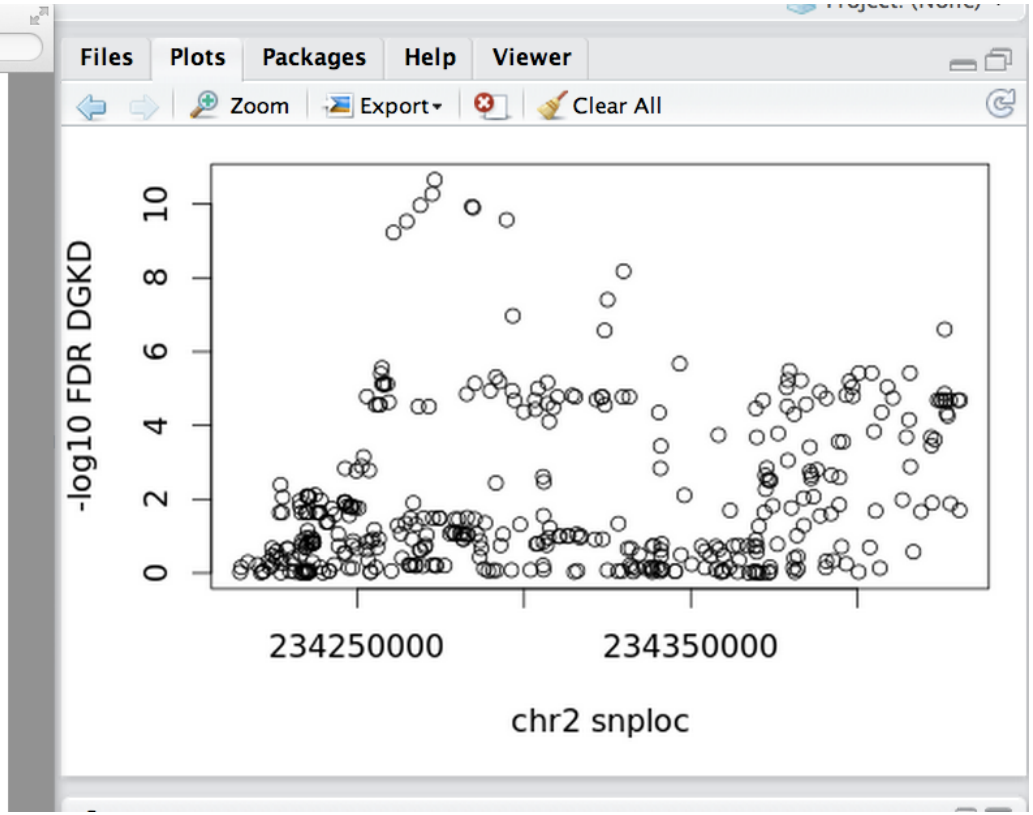
```
R » options(digits=3)
R » cd[1:3]
GRanges with 3 ranges and 10 metadata columns:
      seqnames                 ranges strand |      paramRangeID             REF             ALT
         <Rle>              <IRanges>  <Rle> |          <factor> <DNAStringSet> <CharacterList>
  [1]        2 [234214947, 234214947]      * | ENSG00000077044.5               T               C
  [2]        2 [234215226, 234215226]      * | ENSG00000077044.5               G               T
  [3]        2 [234217249, 234217249]      * | ENSG00000077044.5               T               A
         chisq permScore_1 permScore_2 permScore_3         snp       MAF           probeid
     <numeric>   <numeric>   <numeric>   <numeric> <character> <numeric>          <factor>
  [1]    0.0046      0.4842        0.72        2.03   rs7588509    0.1675 ENSG00000077044.5
  [2]    0.1487      0.0326        2.55        1.80   rs6719241    0.0618 ENSG00000077044.5
  [3]    0.4409      0.0102        2.90        1.05  rs78571834    0.1425 ENSG00000077044.5
  ---
  seqlengths:
    2
   NA
```

Creating the manhattan plot

```
R » pv = 1-pchisq(cd$chisq,1)
R » library(multtest)
R » adp = mt.rawp2adjp(pv, proc="BY")
R » adpo = adp$adjp[order(adp$index)]
R » plot(start(cd)[1:404], -log10(adpo)[1:404])
R » plot(start(cd)[1:404], -log10(adpo)[1:404],
+   xlab="chr2 snploc", ylab="-log10 FDR DGKD"
+ )
```

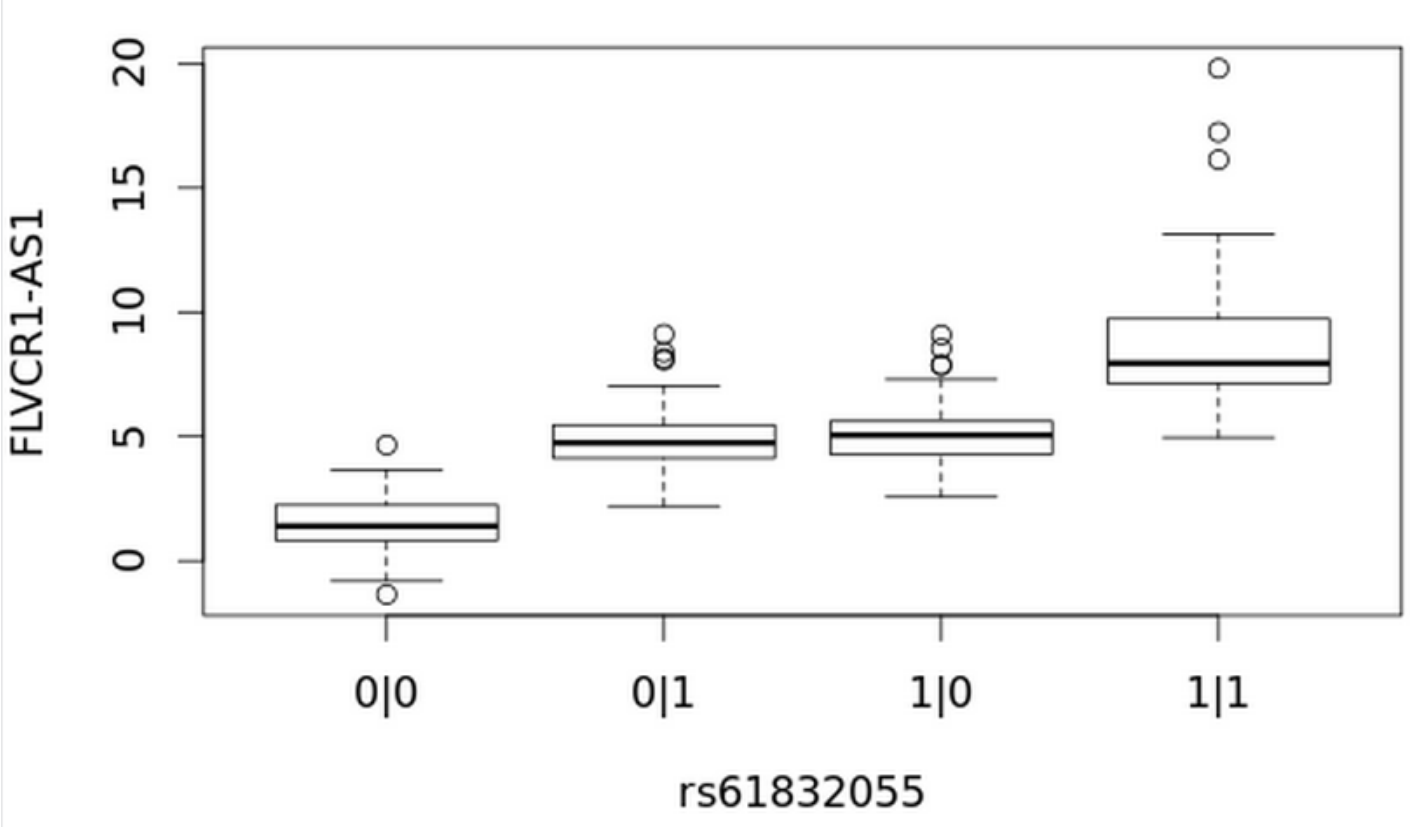
Roughly in agreement…





Broader question: SNP associated with abundance of lncRNA transcripts

- Q1: create the subset of geuFPKM confined to genes assigned to type ‘lincRNA’
- Q2: confine the subset to genes on chr1
- Q3: create the TabixFile reference to the 1000 genomes VCF for chr1
- Q4: use cisAssoc and plotOne to create the visualization



Creating the test statistics for FLVCR1-ASV

```
data(geuFPKM)
type = mcols(geuFPKM)$gene_type
islnc = which(type=="lincRNA")
length(islnc)
gll = geuFPKM[islnc,]
gl1 = gll[which(seqnames(gll)=="chr1"),]
t1 = TabixFile(buckpath(1))
seqlevelsStyle(gl1) = "NCBI"
library(VariantAnnotation)
library(BiocParallel)
cl1 = cisAssoc(gl1[3:4,], t1, lbmaf=.05, cisradius=50000)
```

To create the plot

- Determine the range associated with the SNP of interest
- Check the args of plotOne and supply the elements you've used to do the testing with the SNP range

Summary

- Genome-wide searches for cis-eQTL are easily carried out with
  - RNA-seq data in a SummarizedExperiment
  - Genotype data in a tabix-indexed VCF
  - Ggtools cisAssoc, with various approaches to FDR computation available
- See the ‘eQTL workflow’ at Bioconductor for more details on nonparametric inference, sensitivity analysis, and functional assessment