Exploring 1000 Genomes with Bioconductor

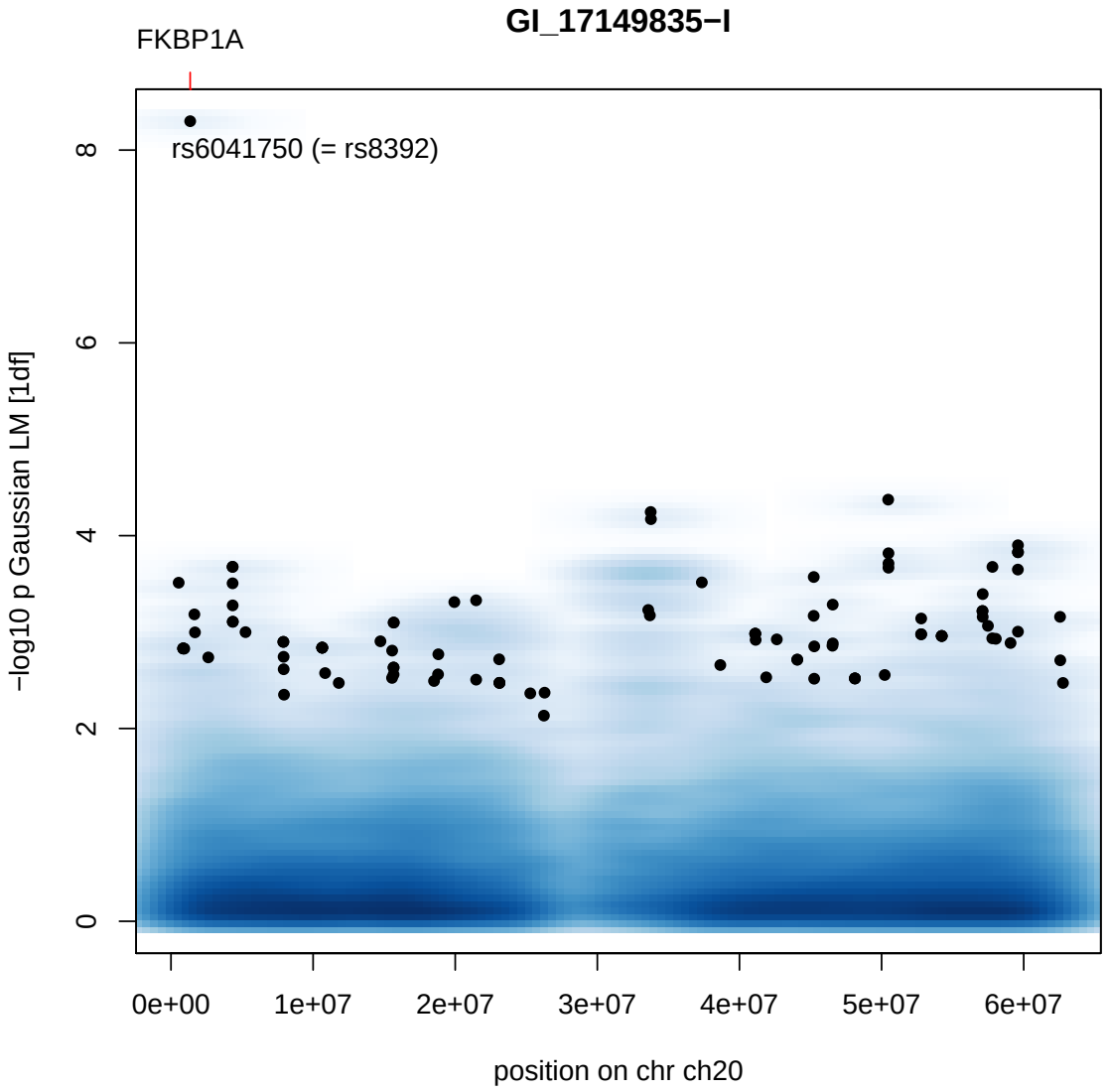
Vince Carey
Channing Lab
Harvard Medical School

- Prologue: What is an eQTL?
- Sketches: 1000 genomes; Bioconductor
- Imputation to the 1000 genomes SNP panel
- Expression arrays, RNA-seq, and eQTL identification

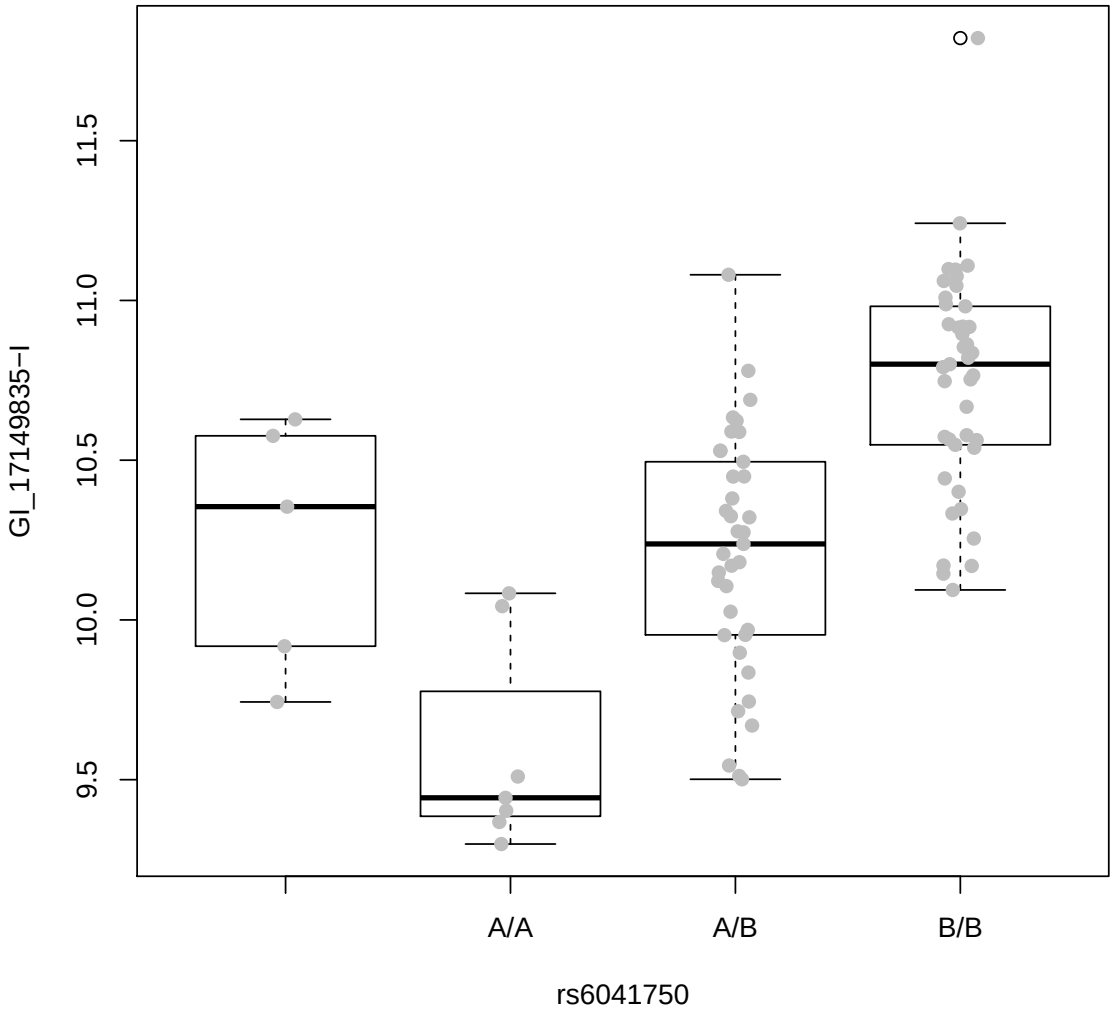
Prologue: What is an eQTL (expression quantitative trait locus)?

- Arises from a basic form of integrative genome-scale data analysis
- On a cohort of N individuals
  - SNP-chip yields allele counts for S SNP, $S \approx 10^6$
  - Expression array yields mRNA abundance measures for G genes, $G \approx 20000$
- perform $G \times S$ association tests of H_{ogs} : mean expression of g is independent of allele count for s
- the best hits are eQTL

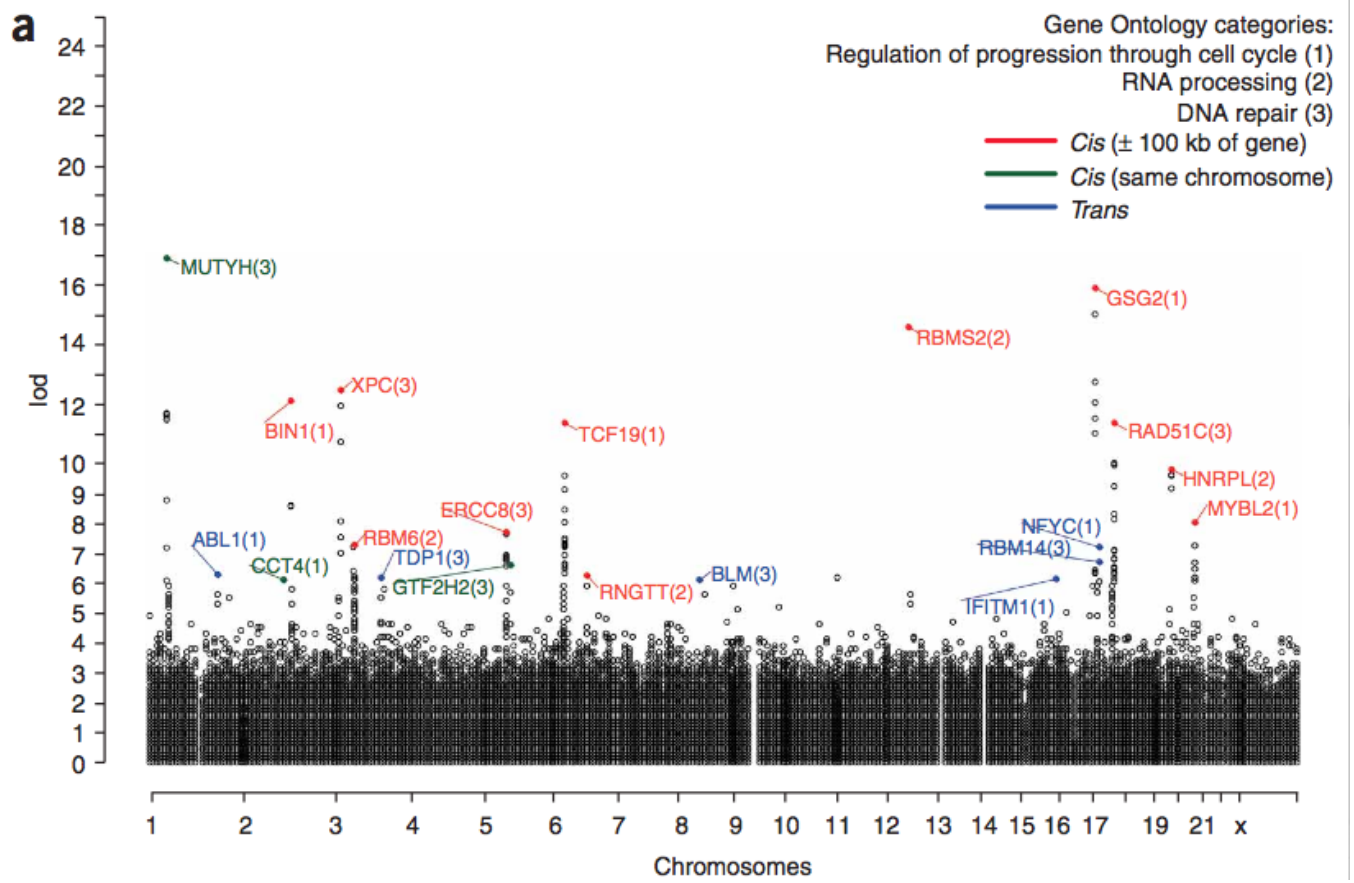
A chromosome-wide scan for a single gene



The ‘best SNP’ discriminates mean expression



Dixon 2007 Nat Genet 'global map'



Why do this? 1: Mechanisms of transcriptional control

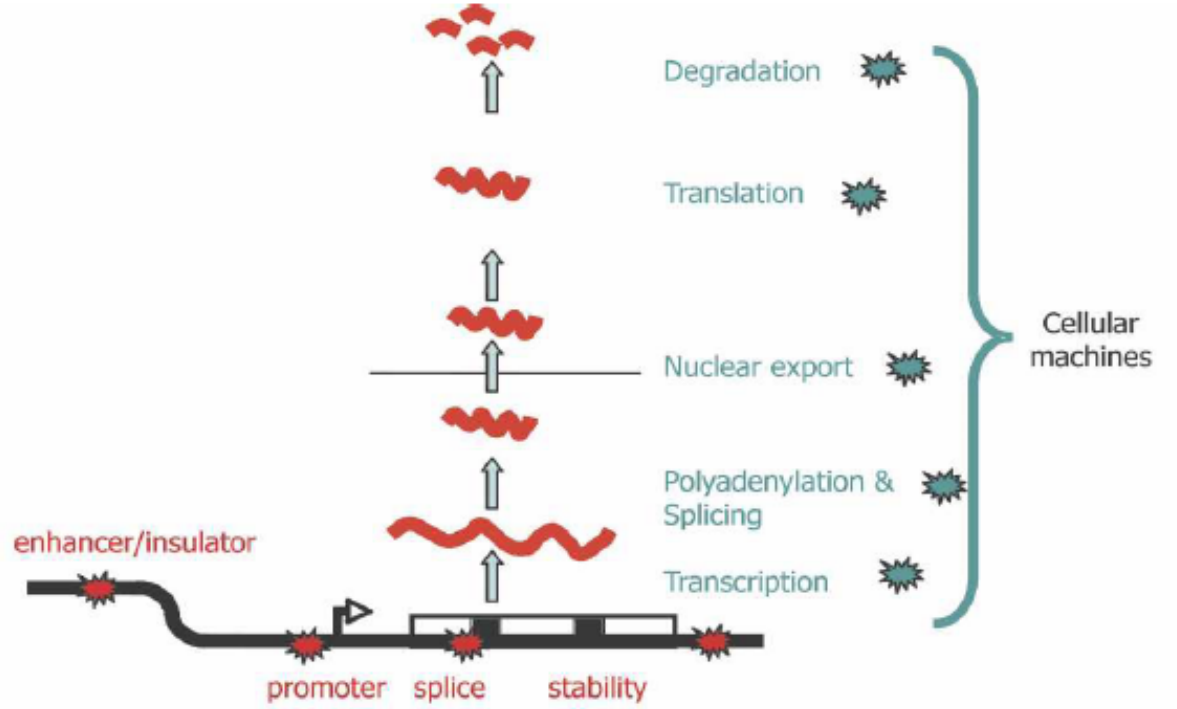
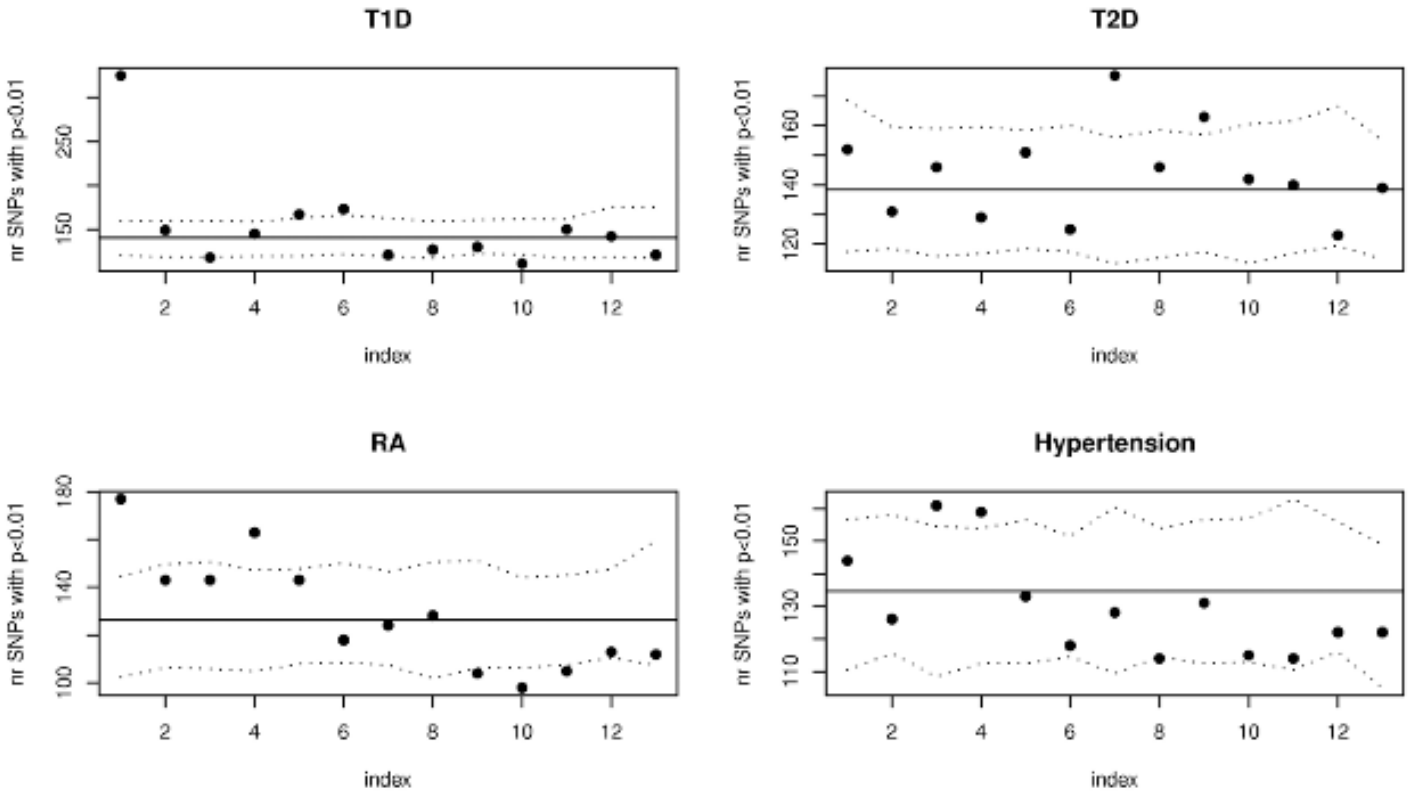


Figure 1. Plausible sites of action for genetic determinants of mRNA levels. Genetic variations influencing gene expression may reside within the regulatory sequences, promoters, enhancers, splice sites, and secondary structure motifs of the target gene and so be genetically in *cis* (red stars), or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

(RBH Williams et al 2007 Genome Resch)

Why do this? 2: Filtering SNP for efficient GWAS

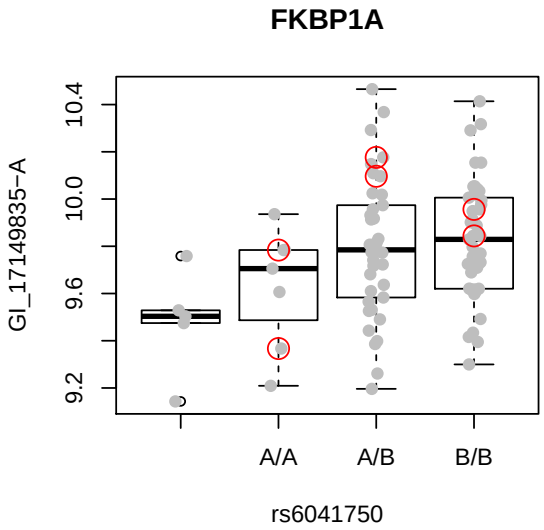
- SNPs binned left to right in decreasing order of expression regulatory capacity
- y axis: proportion SNP in bin associated with macro phenotype in WTCCC

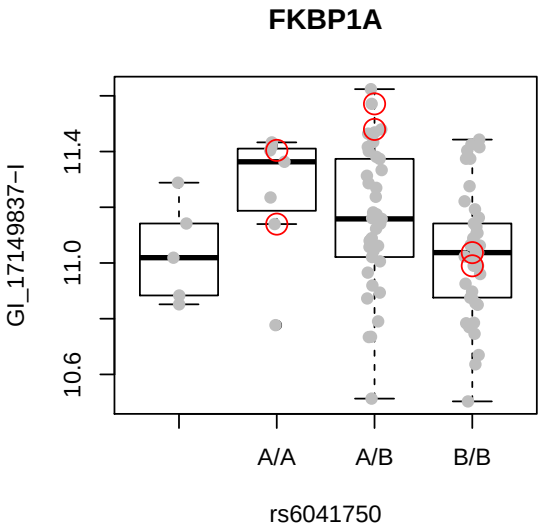


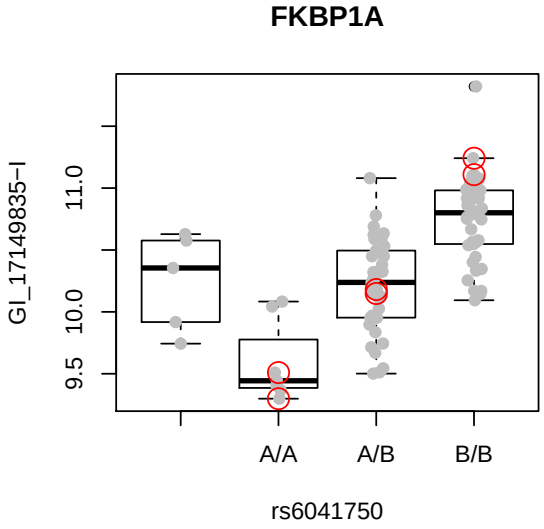
(D Nicolae et al 2010 PLoS Genetics)

Upshots

- eQTL catalogs seem useful; can efficiencies for individual studies be gained by imputing denser SNP panels using results of institutional deep sequencing?
- How can higher-resolution measures of mRNA abundance add to value from eQTL concepts: eQTL searches based on RNA-seq/DNA-seq?
- Under the hood, things may not be so nice...


FKBP1A
GI_17149835−A
10.4
10.0
9.6
9.2
A/A
A/B
B/B
rs6041750


FKBP1A
GI_17149837−I
11.4
11.0
10.6
A/A
A/B
B/B
rs6041750

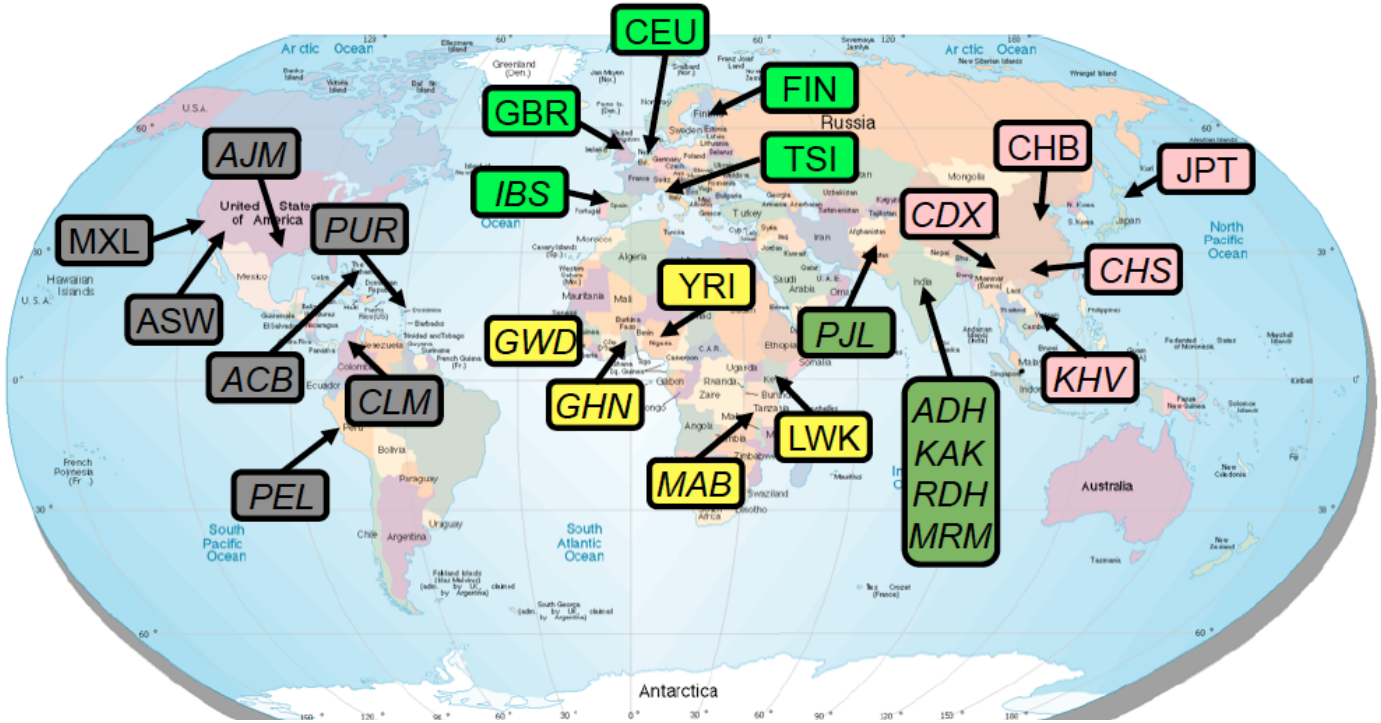

FKBP1A
GI_17149835−I
11.0
10.0
9.5
A/A
A/B
B/B
rs6041750

Sketch: 1000 genomes

Pilot project 180 samples

Extension to 1,100 samples summer 2010

1900 samples end 2010, 2500 samples end 2011



Sketch: 1000 genomes

- DNA sequencing to various depths; high-level interfaces via browsers
- public release to aligned read level: many many BAM files
- more tractable: SNP and variant 'calls': VCF files
- while focus is on DNA variation, availability of cell lines permits measurement of various microscopic phenotypes
- summary:
  - archive of genetic sequence
  - institutional data reductions
  - resource for inference on genetic hypotheses and for methods development

Sketch: Bioconductor

- open-source repository for R-based software targeting genome-scale data analysis
- progress to date
  - preprocessing/annotation/analysis
  - important methods support for affy and illumina expression and genotyping arrays
  - interfaces to GEO/ArrayExpress/SRA for rapid import
  - support for high-performance GWAS and eQTL searches
  - exploit innovations in R: multicore, "disk as RAM", "orchestrator"
  - efforts in sequencing: QC, annotation, analysis (particularly RNA-seq)
  - for 1000 genomes, we have ind1KG, ceu1KG

ceu1kg-package package:ceu1kg R Documentation

CEU (N=60) genotypes from 1000 genomes pilot phase I

Description:

CEU genotypes from 1000 genomes pilot phase I (approx 8 million SNP); includes wellcome trust GENEVAR expression for 41 individuals

Details:

Package:	ceu1kg
Version:	0.0.0
Depends:	R (>= 2.11.1), snpMatrix (>= 1.13.1), GGBase (>= 3.9.0)
License:	Artistic-2.0
LazyLoad:	yes
Built:	R 2.12.0; ; 2010-07-01 01:14:27 UTC; unix

Index:

ceu1kg-package 60 hapmap CEU samples, 47K expression, 8mm 1000 genomes SNP

There are three basic data resources provided here.

First, the 1000 genomes SNP calls for 60 CEU individuals were extracted from the pilot 1 VCF files distributed at <URL: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_03/p

Second, metadata 'GRanges-class' instances are provided in chromosome-specific containers.

Third, an 'smlSet' is provided for 41 individuals in the 1000 genomes CEU SNP call set for whom expression data are available via the Sanger GENEVAR distribution (<URL: ftp://ftp.sanger.ac.uk/pub/genevar/CEU_parents_norm_march2007.zip>).

```
> library(ceu1kg)
> data(ceu1KG.sml)

> sapply(ceu1KG.sml, dim)
```

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10
[1,]	60	60	60	60	60	60	60	60	60	60
[2,]	605756	664326	556362	567547	499164	518645	451004	429055	328069	396487

	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18	chr19	chr20
[1,]	60	60	60	60	60	60	60	60	60	60
[2,]	381826	365883	293253	254837	210540	238117	196327	225279	157182	174484

	chr21	chr22
[1,]	60	60
[2,]	109143	101568

- derived from the VCF representation of all calls for CEU
- distributed as a 700MB gzipped tabix-indexed file
- `GGtools:::vcf2sm` imports record-at-a-time decompressing on the fly

```
> ceu1KG.sml[[1]][1:2, 1:5]

A snp.matrix with  2 rows and  5 columns
Row names:  NA06985 ... NA06986
Col names:  chr1:533 ... rs2462492

> as(ceu1KG.sml[[1]][1:2, 1:5], "matrix")

        chr1:533 chr1:41342 chr1:41791 chr1:44449 rs2462492
NA06985       01         01         01         01        01
NA06986       01         02         01         01        01

> as(ceu1KG.sml[[1]][1:2, 1:5], "character")

        chr1:533 chr1:41342 chr1:41791 chr1:44449 rs2462492
NA06985 "A/A"    "A/A"      "A/A"      "A/A"      "A/A"
NA06986 "A/A"    "A/B"      "A/A"      "A/A"      "A/A"
```

```
> ceu1kg

snp.matrix-based genotype set:
number of samples:  41
number of chromosomes present:  22
annotation: illuminaHumanv1.db
Expression data dims: 47293 x 41
Phenodata: An object of class "AnnotatedDataFrame"
  sampleNames: NA06985, NA06994, ..., NA12874  (41 total)
  varLabels and varMetadata description:
    famid: hapmap family id
    persid: hapmap person id
    ...: ...
    male: logical TRUE if male
    (7 total)

> dim(exprs(ceu1kg))

[1] 47293    41

> summary(smList(ceu1kg)[[20]])

$rows
   Call.rate Heterozygosity
 Min.   :1   Min.   :0.2114
 1st Qu.:1   1st Qu.:0.2285
```

```
 Median :1   Median :0.2339
 Mean   :1   Mean   :0.2358
 3rd Qu.:1   3rd Qu.:0.2430
 Max.   :1   Max.   :0.2555

$cols
     Calls      Call.rate      MAF              P.AA             P.AB
 Min.   :41   Min.   :1   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
 1st Qu.:41   1st Qu.:1   1st Qu.:0.03659   1st Qu.:0.3659   1st Qu.:0.07317
 Median :41   Median :1   Median :0.12195   Median :0.7317   Median :0.19512
 Mean   :41   Mean   :1   Mean   :0.16550   Mean   :0.6361   Mean   :0.23579
 3rd Qu.:41   3rd Qu.:1   3rd Qu.:0.28049   3rd Qu.:0.9268   3rd Qu.:0.39024
 Max.   :41   Max.   :1   Max.   :0.50000   Max.   :1.0000   Max.   :0.92683

      P.BB             z.HWE
 Min.   :0.00000   Min.   : -6.4031
 1st Qu.:0.00000   1st Qu.: -0.1525
 Median :0.02439   Median :  0.1746
 Mean   :0.12815   Mean   :  0.0892
 3rd Qu.:0.14634   3rd Qu.:  0.5578
 Max.   :1.00000   Max.   :  5.4731
                   NA's   :8173.0000
```