objects and workflows for integrative analysis

Contents

1 Overview of some genetical genomics work

Variation in expression is shown in Cheung et al. (2005) to be associated with SNP genotypes. The multistage investigation used expression measures on approximately 1000 genes and over 700000 SNP markers. SNPs found to be associated with variation in expression for a gene were labeled as *cis*-acting if they were located on the same chromosome as the gene; *trans*-acting otherwise.

Mapping determinants of human gene expression by regional and genome-wide association

Vivian G. Cheung[1,2,3], Richard S. Spielman[2], Kathryn G. Ewens[2], Teresa M. Weber[2,3], Michael Morley[3] & Joshua T. Burdick[3]

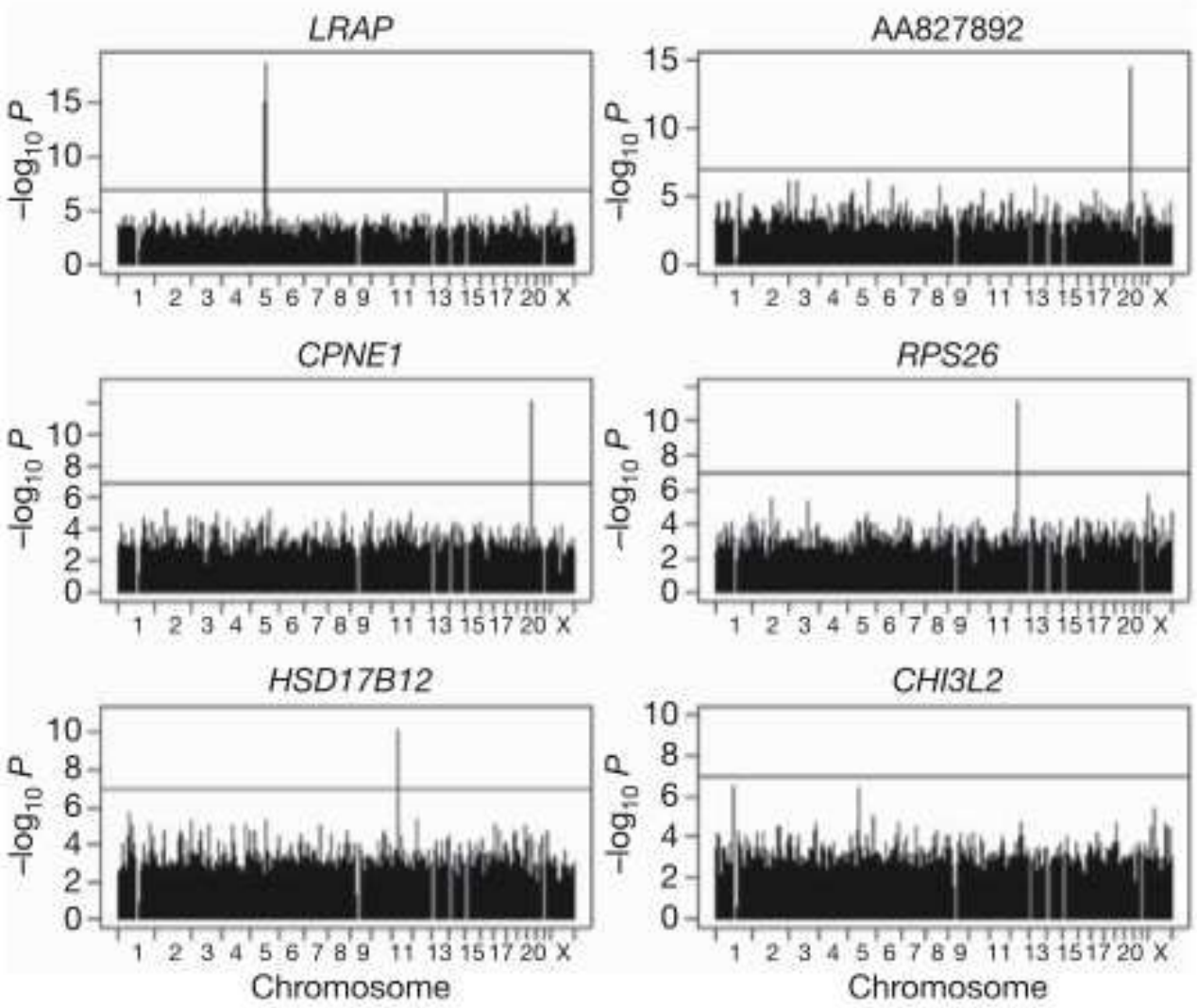


Figure 1 | Results of genome-wide association analysis for six representative phenotypes with *cis* regulators. The horizontal line in each panel corresponds to $P = 0.05$ after Šidák correction.

Table 1 | Genome-wide association results for 27 phenotypes

Phenotype	Location of target gene	Linkage results	GWA results (for peak marker)		
		Peak marker *P*-value (all *cis*)	Marker	Location*	Nominal *P*-value†
LRAP (LOC64167)	5q15	1×10^{-7}	rs2762	58,030	1.98×10^{-19}
AA827892	20q11.23	3×10^{-8}	rs788350	−666	3.67×10^{-15}
PSPHL	7p11.2	3×10^{-11}	rs6593279	−36,903	9.59×10^{-15}
CPNE1	20q11.22	1×10^{-7}	rs6060535	17,327‡	8.35×10^{-13}
CSTB	21q22.3	2×10^{-9}	rs880987	−28,195	2.48×10^{-12}
RPS26	12q13.2	2×10^{-9}	rs2271194	−41,768	7.94×10^{-12}
GSTM2	1p13.3	3×10^{-8}	rs535088	12,699	2.00×10^{-11}
HLA-DRB2	6p21.32	$<10^{-11}$	rs6928482	8,345	6.51×10^{-11}
IRF5	7q32.1	2×10^{-8}	rs2280714	16,731	6.78×10^{-11}
HSD17B12	11p11.2	2×10^{-11}	rs4755741	100,949‡	7.38×10^{-11}
GSTM1	1p13.3	1×10^{-7}	rs535088	−7,052	8.33×10^{-10}
PPAT	4q12	2×10^{-7}	rs227940	*Trans* (Chr 7)	5.29×10^{-9}
PPAT	4q12	2×10^{-7}	rs2139512	25,227‡	2.87×10^{-8}
DDX17	22q13.1	6×10^{-10}	rs10490570	*Trans* (Chr 2)	7.13×10^{-9}
CTSH	15q25.1	7×10^{-9}	rs1369324	−2,298	2.17×10^{-8}
POMZP3	7q11.23	9×10^{-10}	rs1754162	−6,215	7.23×10^{-8}
CGI-96	22q13.2	3×10^{-9}	rs9600337	*Trans* (Chr 13)	2.43×10^{-7}
CHI3L2	1p13.3	3×10^{-11}	rs755467	−91	2.57×10^{-7}
VAMP8	2p11.2	9×10^{-8}	rs10509846	*Trans* (Chr 10)	5.31×10^{-7}
EIF3S8	16p11.2	4×10^{-8}	rs8092794	*Trans* (Chr 18)	7.20×10^{-7}
TM7SF3	12p11.23	$<10^{-11}$	rs11822822	*Trans* (Chr 11)	7.32×10^{-7}
IL16	15q25.1	3×10^{-10}	rs6957902	*Trans* (Chr 7)	9.63×10^{-7}
TCEA1	8q11.23	6×10^{-8}	rs6562160	*Trans* (Chr 13)	1.08×10^{-6}
S100A13	1q21.3	3×10^{-8}	rs3757791	*Trans* (Chr 7)	1.40×10^{-6}
ICAP-1A	2p25.1	$<10^{-11}$	rs10807387	*Trans* (Chr 6)	2.27×10^{-6}
SMARCB1	22q11.23	4×10^{-7}	rs7802273	*Trans* (Chr 7)	2.46×10^{-6}
CTBP1	4p16.3	2×10^{-9}	rs1060043	*Trans* (Chr 19)	5.26×10^{-6}
ZNF85	19p12	9×10^{-9}	rs2168903	*Trans* (Chr 12)	6.51×10^{-6}

*Relative to transcriptional start site of target gene. When the most significant marker is located on a chromosome different from the target gene, it is listed as '*Trans*' and the chromosome is shown.

†Corrected *P*-value of 0.05 corresponds to a nominal *P*-value of 6.7×10^{-8}.

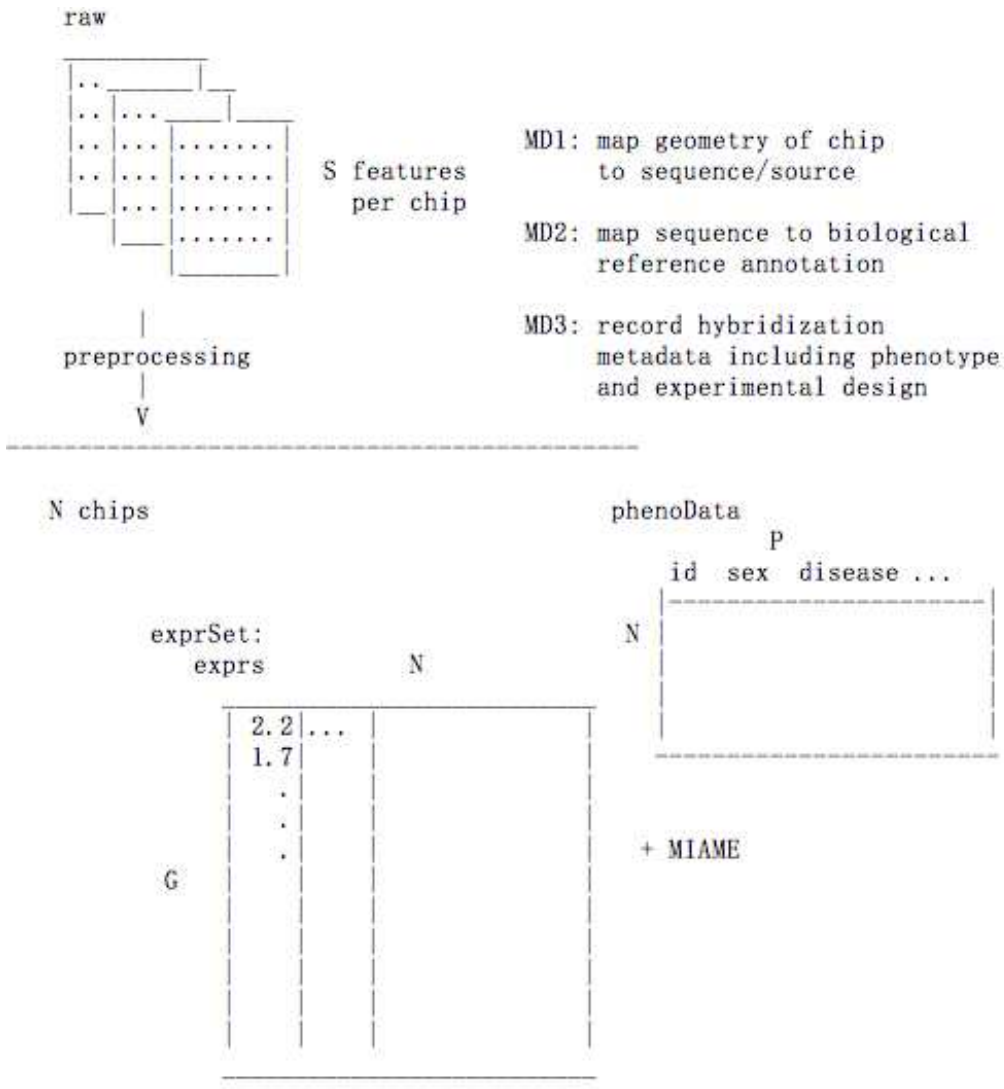
‡Marker is within genomic extent of target gene.

basic findings

- cis and trans (off chromosome) type determinants exist
- locations of cis determinants seem equally balanced between 3' and 5' regions
- findings are possible with modest sample size

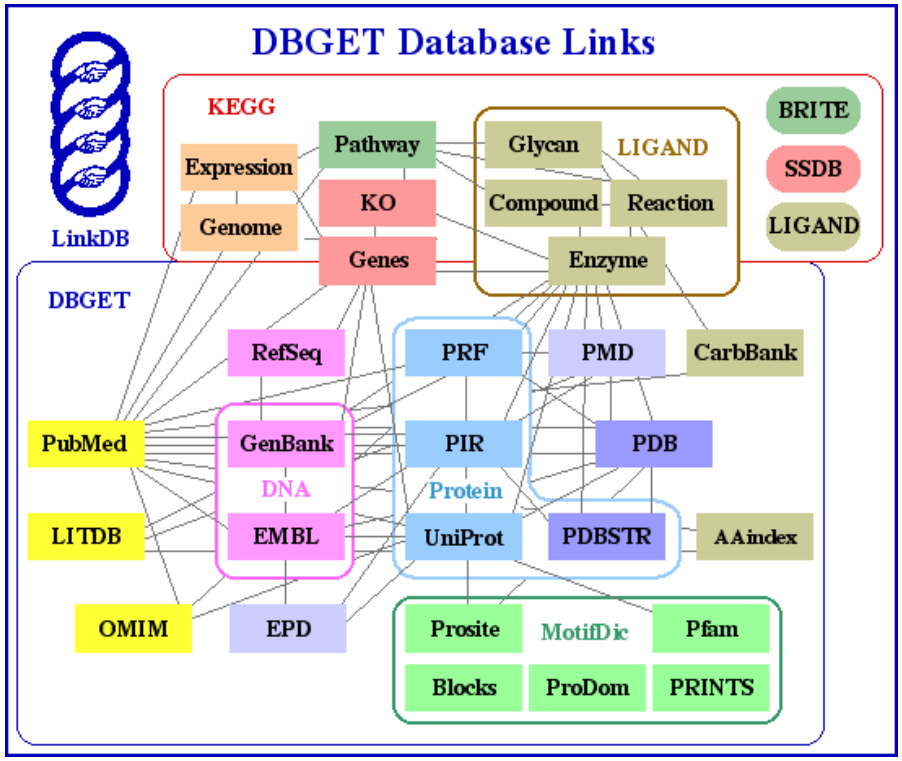
2 Sources of complexity and anxiety

Primitive schematic



Metadata complex

Reporter materials have context in genomic sequence and in biological knowledge. Some of the resources that can be used to specify context are depicted in the following schematic from KEGG (Kanehisa, 1997; Kanehisa et al., 2004):



Reproducibility issues

Prediction of cancer outcome with microarrays: a multiple random validation strategy

Stefan Michiels, Serge Koscielny, Catherine Hill

Summary

Background General studies of microarray gene-expression profiling have been undertaken to predict cancer outcome. Knowledge of this gene-expression profile or molecular signature should improve treatment of patients by allowing treatment to be tailored to the severity of the disease. We reanalysed data from the seven largest published studies that have attempted to predict prognosis of cancer patients on the basis of DNA microarray analysis.

Methods The standard strategy is to identify a molecular signature (ie, the subset of genes most differentially expressed in patients with different outcomes) in a training set of patients and to estimate the proportion of misclassifications with this signature on an independent validation set of patients. We expanded this strategy (based on unique training and validation sets) by using multiple random sets, to study the stability of the molecular signature and the proportion of misclassifications.

Findings The list of genes identified as predictors of prognosis was highly unstable; molecular signatures strongly depended on the selection of patients in the training sets. For all but one study, the proportion misclassified decreased as the number of patients in the training set increased. Because of inadequate validation, our chosen studies published overoptimistic results compared with those from our own analyses. Five of the seven studies did not classify patients better than chance.

Interpretation The prognostic value of published microarray results in cancer studies should be considered with caution. We advocate the use of validation by repeated random sampling.

“Preferred” methods

Research

Open Access

Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset

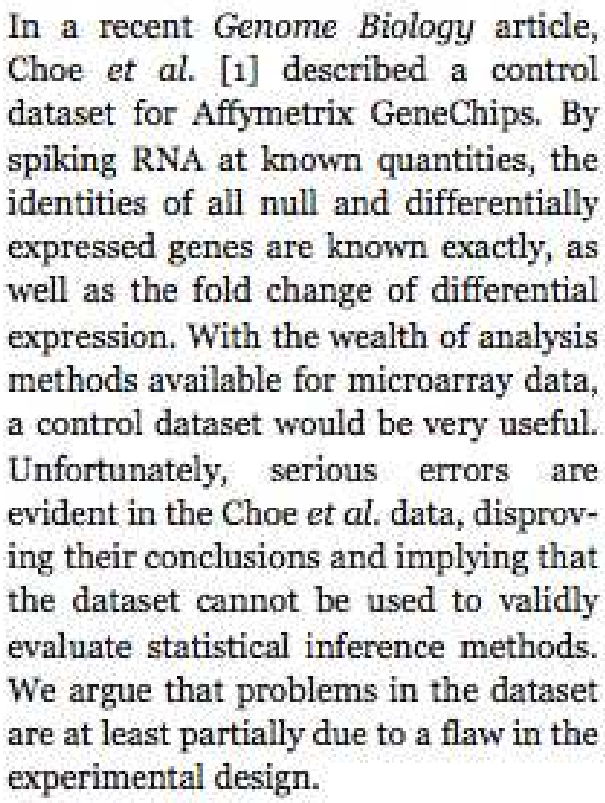
Sung E Choe[*†], Michael Boutros[*¥], Alan M Michelson[*†‡], George M Church[*] and Marc S Halfon[†§¶]

Correspondence

A reanalysis of a published Affymetrix GeneChip control dataset

Alan R Dabney and John D Storey

A response to **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset** by SE Choe, M Boutros, AM Michelson, GM Church and MS Halfon. *Genome Biology* 2005, 6:R16.



In a recent *Genome Biology* article, Choe *et al.* [1] described a control dataset for Affymetrix GeneChips. By spiking RNA at known quantities, the identities of all null and differentially expressed genes are known exactly, as well as the fold change of differential expression. With the wealth of analysis methods available for microarray data, a control dataset would be very useful. Unfortunately, serious errors are evident in the Choe *et al.* data, disproving their conclusions and implying that the dataset cannot be used to validly evaluate statistical inference methods. We argue that problems in the dataset are at least partially due to a flaw in the experimental design.

In a JHU technical report, Irizarry, Cope and Wu address the same dataset:

Introduction

In [1] a spike-in experiment is described which the authors use to compare expression measures for Affymetrix GeneChip technology. Two sets of triplicates were created to represent control (C) and experimental (S) samples. In [2] and [3] we describe a benchmark for such measures based on experiments developed by Affymetrix and a GeneLogic. These datasets are described in detail in [2]. A web-based implementation of the benchmark, is available at `affycomp.biostat.jhsph.edu`. There are various inconsistencies between the conclusions reached by [1] and [3]. In this letter we describe certain characteristics of the feature-level data produced by [1] which we believe explain these inconsistencies. These can be divided into 1) induced by the experimental design and 2) an artifact.

Experimental design

There are three characteristics of the experimental design described by [1] make the resulting data inappropriate for assessment. Below we enumerate these problems and explain how they lead to unfair assessments. Other problems with the experimental design are described by [4].

1. The spike-in concentrations are unrealistically high. In [3] we demonstrate that background noise makes it harder to detect differentially expression for genes that are present in low concentrations. In [3] we point out that in the Affymetrix spike-in experiments the concentrations for spiked-in features are artificially high but that a large number of these are actually in a usable range (See Figure 1A). Figure 1B demonstrates that in a typical experiment, features related to differentially expressed genes show intensities with a similar range as the rest of the genes. However, Figures 1C-D suggest that none of the genes spiked-in by [1] are in a usable range since less than 1% of the data would reach the intensity levels seen for the spiked-in genes. This implies that expression measure assessments based on this dataset only apply to unlikely situations where we expect differentially expressed genes to be in the top 1% of overall expression.

2. A large percentage of the genes (about 10%) are spiked-in to be differentially expressed and all of these are expected to be up-regulated. This design makes this spike-in data very different from those produced by typical experiments where at least one of the following assumptions is expected to hold: 1) a small percentage of genes are differentially expressed, 2) there is a balance between up and down regulation, and 3) the gene expression distribution across arrays is roughly the same. Most preprocessing algorithms implement normalization routines motivated by one or more of these assumptions, thus we should not expect existing expression measure methodology to perform well with the Choe et al. data.

3. A careful look at Table 1 in [1] shows that nominal concentrations and fold change sizes are confounded. This is better demonstrated by a graphical representation (Figure 2). This problem will not permit us to distinguish ability to detect small fold changes from the ability to detect differential expression when concentration is low. [3] show why this distinction is important.

upshots

- assertions about "preferred" methods, even if methods are transparent, must be taken with caution
- users should execute multiple "preferred" methods and understand sources of discrepant conclusions
- concrete reproducibility of research is useful to support reuse and extension of useful methods

3 Genetical genomics using chr 7,15, 20

Provenance:

- hgfocus expression data for N=58 CEPH CEU unrelated individuals provided by Vivian Cheung and Richard Spielman at the 2006 Cold Spring Harbor course on Integrative Data Analysis for High-throughput Biology.
- high-density SNP genotypes from HapMap, matched by CEU NAnnnnn number to the expression samples, for only 48 individuals
- key results of Cheung and Spielman qualitatively reproducible with the N=48 subsample

```
> library(GGtools)

> data(c20GGceu)
> c20GGceu

GG Expression Set (exprSet catering for many SNP attributes) with
        8793 genes
        48 samples
There are  114666  attributes; names include:
rs4814683 rs6076506 rs6139074 rs1418258 rs7274499

> pData(c20GGceu)[1:4, 1:4]

        rs4814683 rs6076506 rs6139074 rs1418258
NA11829         0         2         0         0
NA11830         1         2         1         1
NA11831         0         2         0         0
NA11832         1         2         1         1
```

The value in the i, j element of the phenoData is the count of rare alleles found in the genotype on snp j in individual i. Any missing call leads to a missing record.

Cheung, Spielman et al. report genome-wide association (GWA) results for gene CPNE1 in connection with rs6060535. The *RSNPper* package gives us some curated information about the SNP:

```
> library(RSNPper)
Loading required package: XML
> SNPinfo("6060535")
SNPper SNP metadata:
     DBSNPID     CHROMOSOME POSITION   ALLELES VALIDATED
[1,] "rs6060535" "chr20"    "33698936" "C/T"   "Y"
There are details on 4 populations
and 10 connections to gene features
SNPper info:
     SOURCE           VERSION              GENOME DBSNP
[1,] "*RPCSERV-NAME*" "$Revision: 1.38 $" "hg17" "123"
```

```
bb = SNPinfo("6060535")
> popDetails(bb)
            PANEL   SIZE MAJOR.ALLELE MINOR.ALLELE   majorf    minorf
1        Japanese sanger            C            T 0.918605 0.0813954
2     Han_Chinese sanger            C            T  0.94186 0.0581395
3 Yoruba-30-trios sanger            C            T    0.925     0.075
4   CEPH-30-trios sanger            C            T      0.9       0.1

> geneDetails(goo)
    HUGO LOCUSLINK                          NAME      MRNA   ROLE RELPOS AMINO
1  CPNE1      8904                     copine I NM_003915   Exon -14677  <NA>
2  CPNE1      8904                     copine I NM_152925   Exon -14677  <NA>
3  CPNE1      8904                     copine I NM_152926   Exon -14677  <NA>
4  CPNE1      8904                     copine I NM_152927   Exon -14677  <NA>
5  CPNE1      8904                     copine I NM_152928   Exon -14677  <NA>
6  CPNE1      8904                     copine I NM_152929   Exon -14677  <NA>
7  CPNE1      8904                     copine I NM_152930   Exon -14677  <NA>
8  CPNE1      8904                     copine I NM_152931   Exon -14677  <NA>
9  RBM12     10137 RNA binding motif protein 12 NM_006047 3' UTR   7722  <NA>
10 RBM12     10137 RNA binding motif protein 12 NM_152838 3' UTR   7722  <NA>
```

```
> geneInfo("CPNE1")
       snpper.ID             NAME           CHROM          STRAND
         "12438"          "CPNE1"         "chr20"             "-"
         PRODUCT        LOCUSLINK            OMIM         UNIGENE
      "copine I"           "8904"         "604205"      "Hs.166887"
       SWISSPROT            NSNPS       REFSEQACC          MRNAACC
        "Q9NTZ6"            "189"             " "      "NM_152931"
TRANSCRIPT.START  CODINGSEQ.START   TRANSCRIPT.END    CODINGSEQ.END
      "33677382"       "33677577"       "33716262"       "33684259"

> geneInfo("RBM12")
                     snpper.ID                        NAME
                       "12440"                     "RBM12"
                         CHROM                      STRAND
                       "chr20"                         "-"
                       PRODUCT                    LOCUSLINK
"RNA binding motif protein 12"                      "10137"
                          OMIM                      UNIGENE
                       "607179"                         " "
                       SWISSPROT                      NSNPS
                             " "                      "113"
                     REFSEQACC                      MRNAACC
                             " "                  "NM_152838"
               TRANSCRIPT.START               CODINGSEQ.START
                     "33700295"                    "33703860"
                 TRANSCRIPT.END                 CODINGSEQ.END
                     "33716252"                    "33706658"
```

Both genes are antisense on chromosome 20 in the vicinity of 33.7M. Note that all *RSNPper* responses provide a toolInfo attribute describing the underlying database versions.

3.1 Reproducing Cheung and Spielman on CPNE1

We can plot the available data on CPNE1 expression and the rare allele counts in the N=58 individuals:

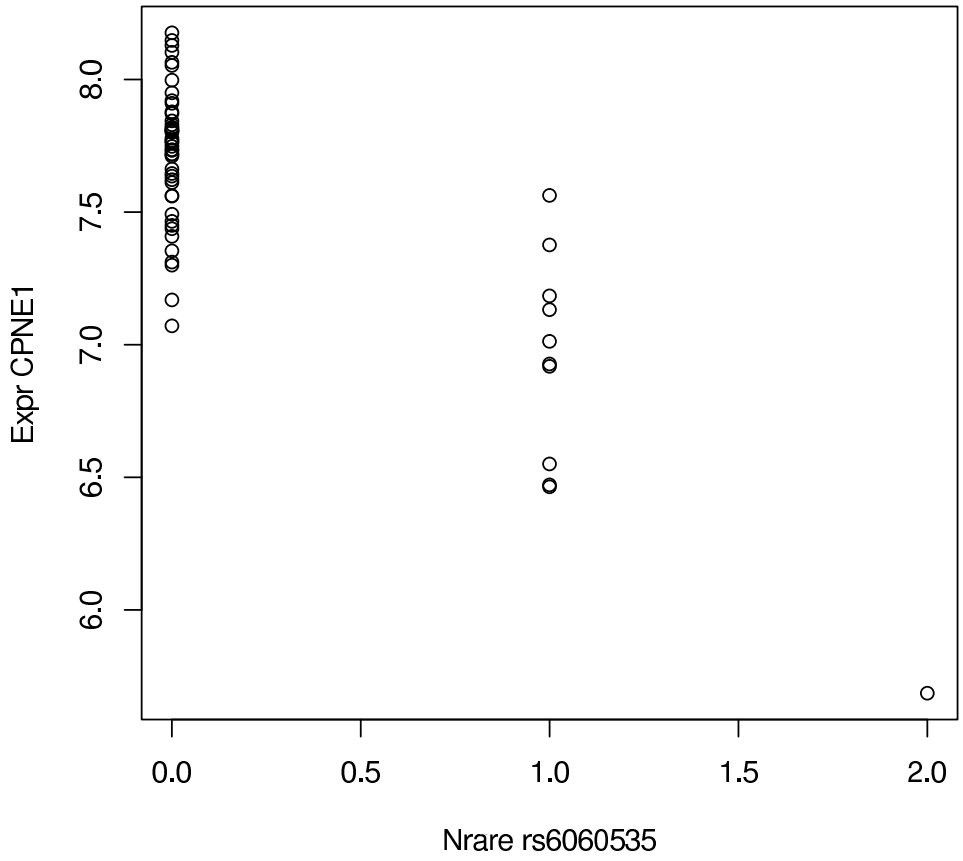


Table 1 of the paper of Cheung, Spielman et al. presents p-values for regression hypotheses about data configurations like the one displayed above.

Table 1 | Genome-wide association results for 27 phenotypes

Phenotype	Location of target gene	Linkage results	GWA results (for peak marker)		
		Peak marker *P*-value (all *cis*)	Marker	Location*	Nominal *P*-value†
LRAP (LOC64167)	5q15	1×10^{-7}	rs2762	58,030	1.98×10^{-19}
AA827892	20q11.23	3×10^{-8}	rs788350	−666	3.67×10^{-15}
PSPHL	7p11.2	3×10^{-11}	rs6593279	−36,903	9.59×10^{-15}
CPNE1	20q11.22	1×10^{-7}	rs6060535	17,327‡	8.35×10^{-13}
CSTB	21q22.3	2×10^{-9}	rs880987	−28,195	2.48×10^{-12}
RPS26	12q13.2	2×10^{-9}	rs2271194	−41,768	7.94×10^{-12}
GSTM2	1p13.3	3×10^{-8}	rs535088	12,699	2.00×10^{-11}
HLA-DRB2	6p21.32	$<10^{-11}$	rs6928482	8,345	6.51×10^{-11}
IRF5	7q32.1	2×10^{-8}	rs2280714	16,731	6.78×10^{-11}
HSD17B12	11p11.2	2×10^{-11}	rs4755741	100,949‡	7.38×10^{-11}
GSTM1	1p13.3	1×10^{-7}	rs535088	−7,052	8.33×10^{-10}
PPAT	4q12	2×10^{-7}	rs227940	*Trans* (Chr 7)	5.29×10^{-9}
PPAT	4q12	2×10^{-7}	rs2139512	25,227‡	2.87×10^{-8}
DDX17	22q13.1	6×10^{-10}	rs10490570	*Trans* (Chr 2)	7.13×10^{-9}
CTSH	15q25.1	7×10^{-9}	rs1369324	−2,298	2.17×10^{-8}
POMZP3	7q11.23	9×10^{-10}	rs1754162	−6,215	7.23×10^{-8}
CGI-96	22q13.2	3×10^{-9}	rs9600337	*Trans* (Chr 13)	2.43×10^{-7}
CHI3L2	1p13.3	3×10^{-11}	rs755467	−91	2.57×10^{-7}
VAMP8	2p11.2	9×10^{-8}	rs10509846	*Trans* (Chr 10)	5.31×10^{-7}
EIF3S8	16p11.2	4×10^{-8}	rs8092794	*Trans* (Chr 18)	7.20×10^{-7}
TM7SF3	12p11.23	$<10^{-11}$	rs11822822	*Trans* (Chr 11)	7.32×10^{-7}
IL16	15q25.1	3×10^{-10}	rs6957902	*Trans* (Chr 7)	9.63×10^{-7}
TCEA1	8q11.23	6×10^{-8}	rs6562160	*Trans* (Chr 13)	1.08×10^{-6}
S100A13	1q21.3	3×10^{-8}	rs3757791	*Trans* (Chr 7)	1.40×10^{-6}
ICAP-1A	2p25.1	$<10^{-11}$	rs10807387	*Trans* (Chr 6)	2.27×10^{-6}
SMARCB1	22q11.23	4×10^{-7}	rs7802273	*Trans* (Chr 7)	2.46×10^{-6}
CTBP1	4p16.3	2×10^{-9}	rs1060043	*Trans* (Chr 19)	5.26×10^{-6}
ZNF85	19p12	9×10^{-9}	rs2168903	*Trans* (Chr 12)	6.51×10^{-6}

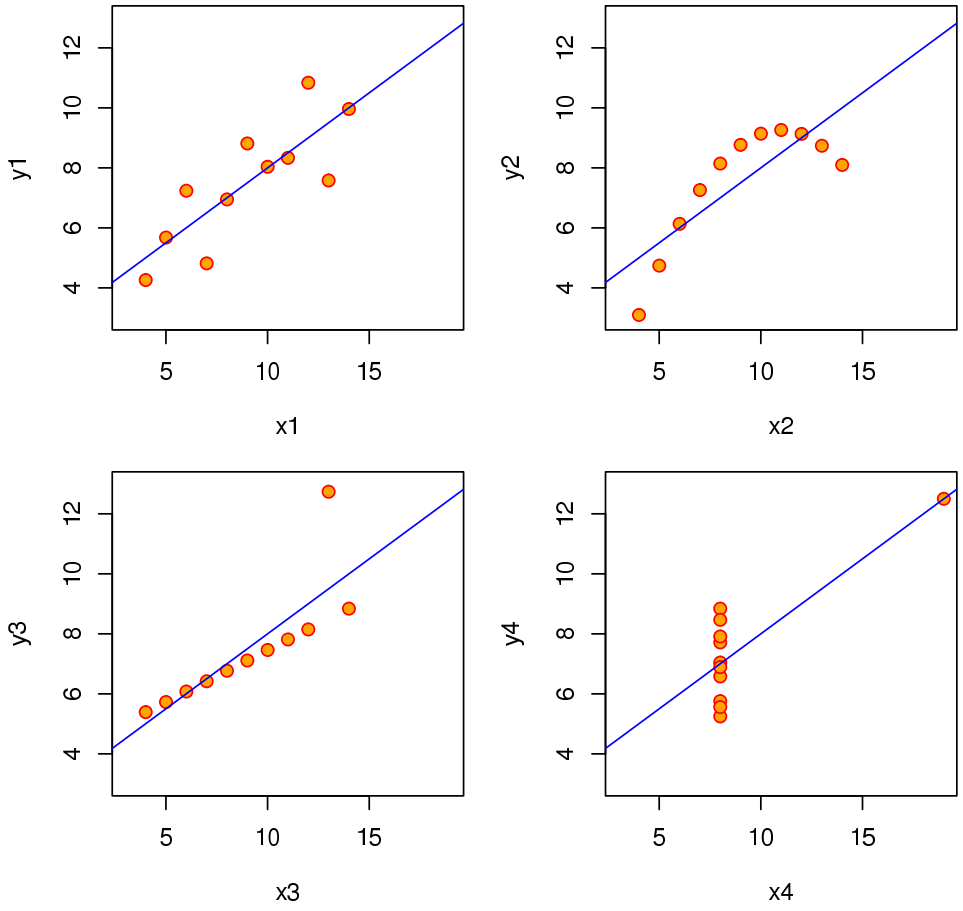
*Relative to transcriptional start site of target gene. When the most significant marker is located on a chromosome different from the target gene, it is listed as '*Trans*' and the chromosome is shown.

†Corrected *P*-value of 0.05 corresponds to a nominal *P*-value of 6.7×10^{-8}.

‡Marker is within genomic extent of target gene.

Are p-values a good index for this problem?

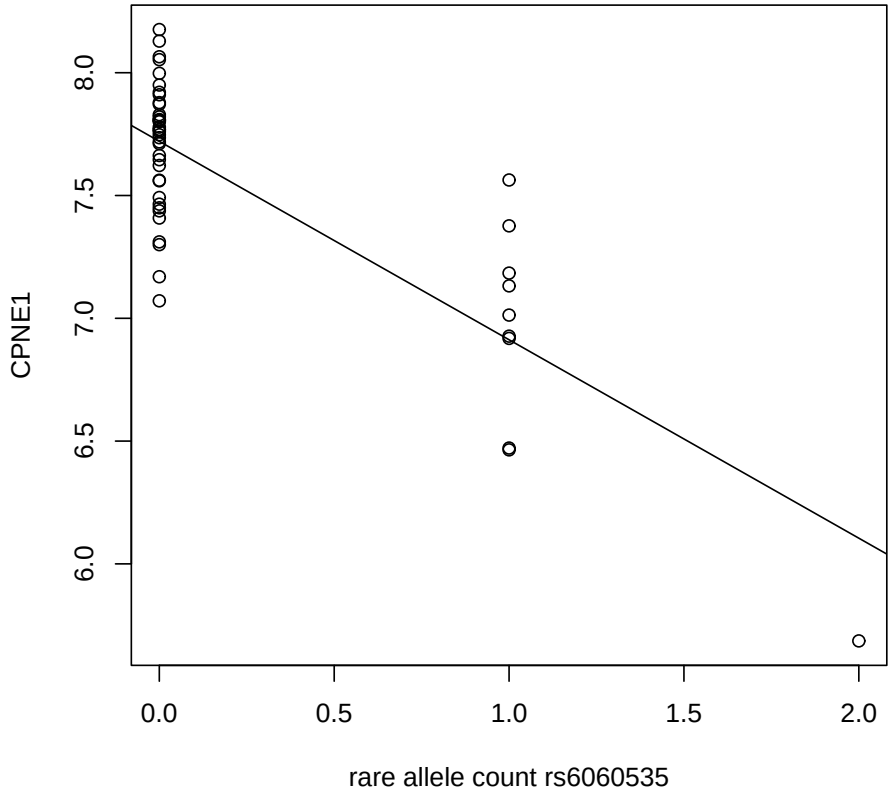
Anscombe's 4 Regression data sets



3.2 Probing around with GGtools

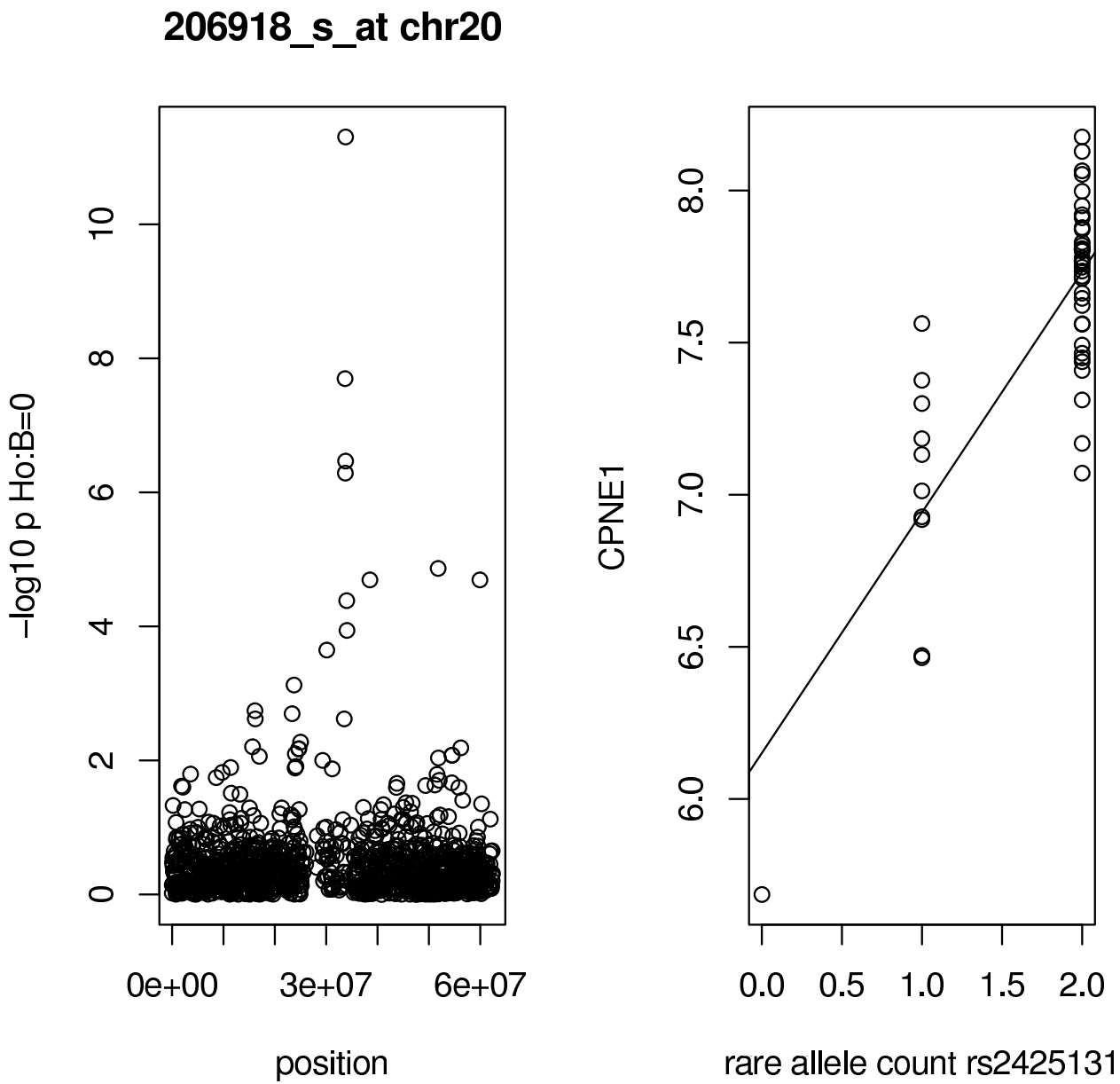
Based on the 48 that I could find, we have

```
> mcpne1 = ggrplot(c20GGceu, "CPNE1", "rs6060535")
```



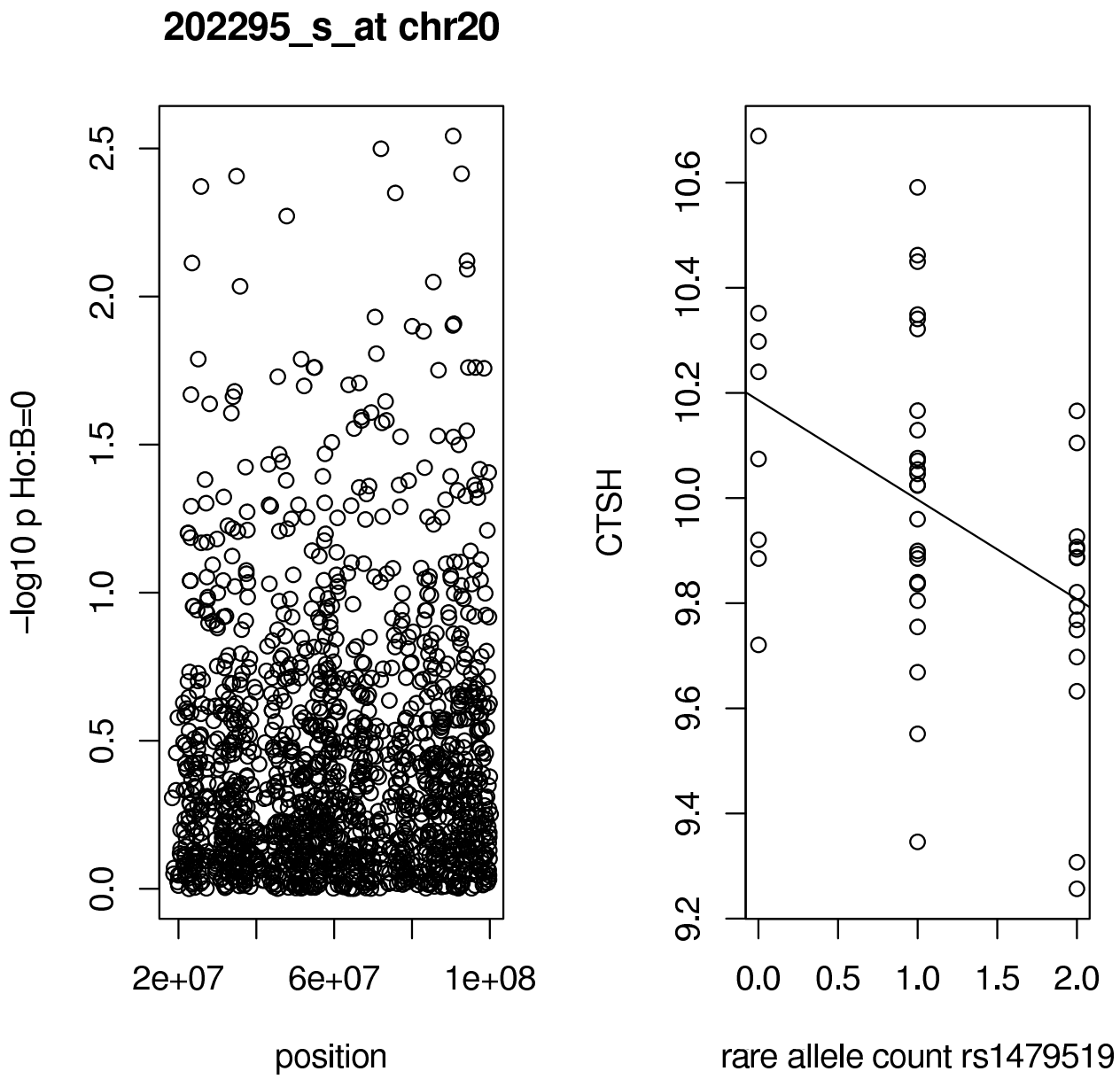
detScreen: Want SNPs on a sequence of locations checked for association with expression of a given gene

```
> dput(detScreen)

function (gge = c20GGceu, psn = "206918_s_at", chrmeta = chr20meta,
    chr = "chr20", gran = 50, gene = "")
{
    opar = par()
    cpn = regseq(gge, psn, seq(1, ncol(gge@phenoData@pData),
        gran), chrmeta, chr)
    par(mfrow = c(1, 2))
    plot(cpn$locs, -log10(cpn$pva), main = paste(psn, chr), xlab = "positi
        ylab = "-log10 p Ho:B=0")
    bot = which.min(cpn$pva)
    ggrplot(gge, gene, names(bot))
    par(opar)
    invisible(list(bot = bot, cpn = cpn))
}
```

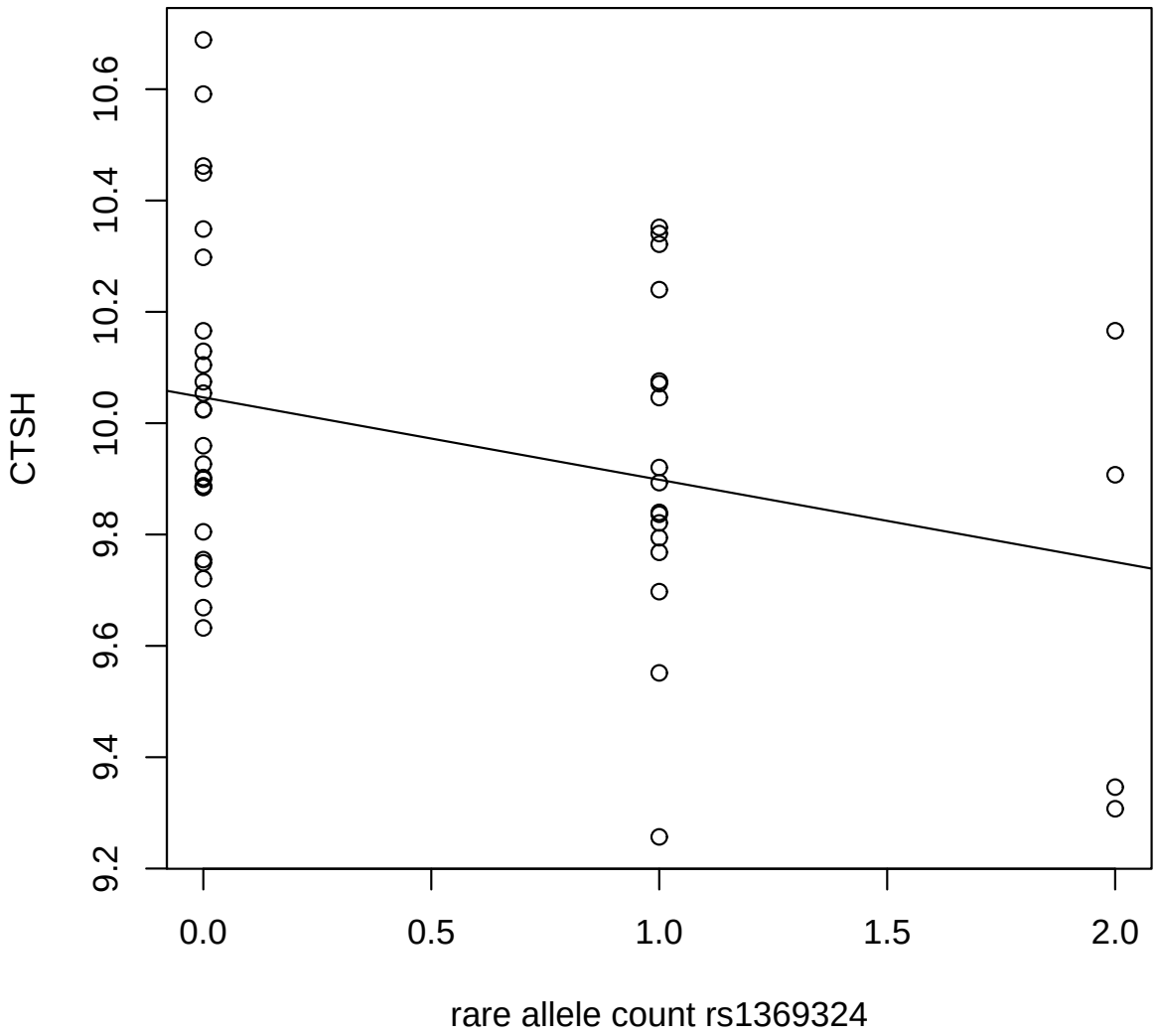

206918_s_at chr20
−log10 p Ho:B=0
0
2
4
6
8
10
0e+00
3e+07
6e+07
position
CPNE1
6.0
6.5
7.0
7.5
8.0
0.0
0.5
1.0
1.5
2.0
rare allele count rs2425131

Another cis example (1/50 available snps sampled):

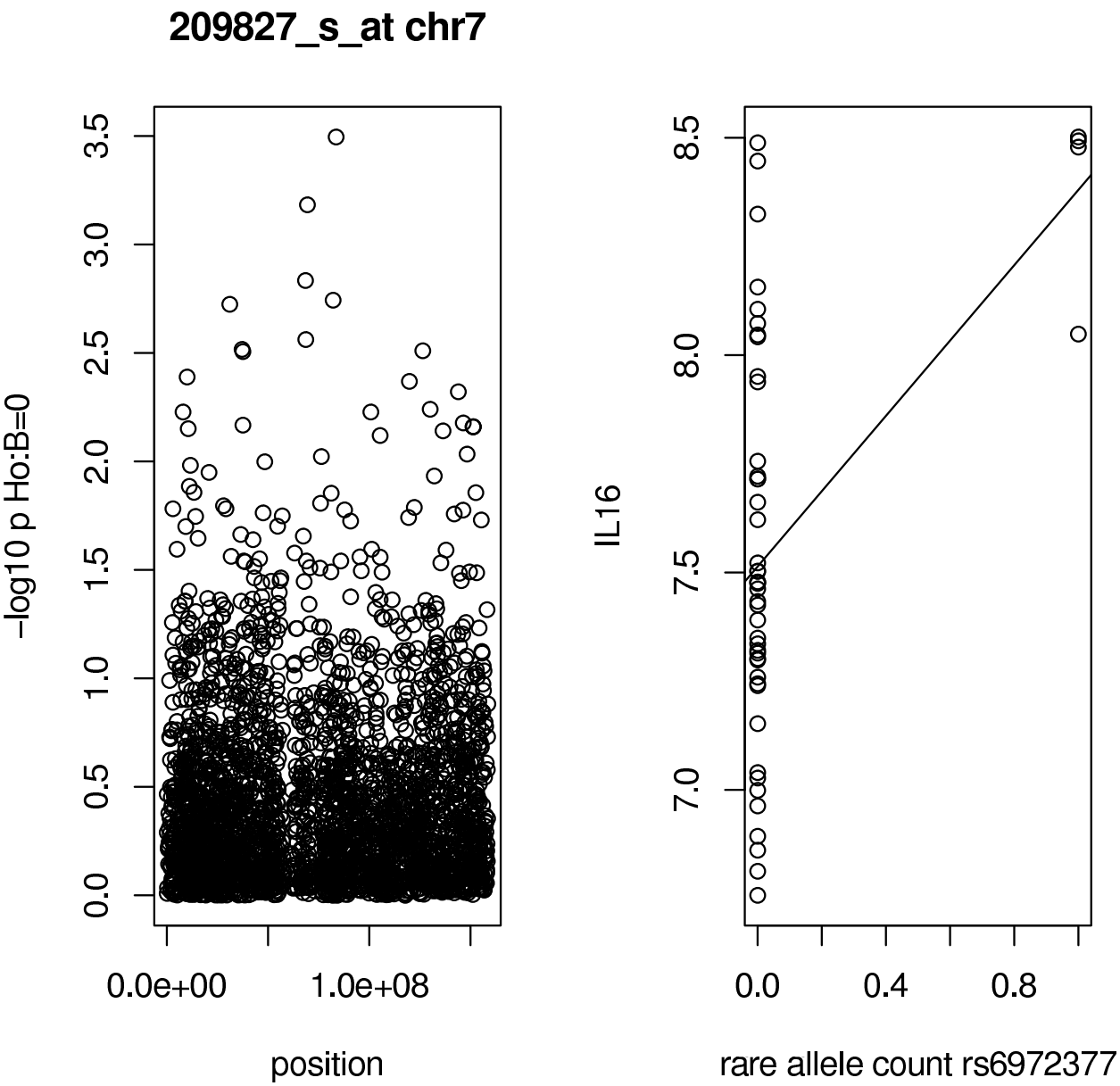
```
> detScreen(c15GGceu, psn = "202295_s_at", chrmeta = chr15meta,
+     chr = "chr20", gene = "CTSH")
```


202295_s_at chr20
−log10 p Ho:B=0
0.0
0.5
1.0
1.5
2.0
2.5
2e+07
6e+07
1e+08
position
CTSH
9.2
9.4
9.6
9.8
10.0
10.2
10.4
10.6
0.0
0.5
1.0
1.5
2.0
rare allele count rs1479519

```
> cs2 = ggrplot(c15GGceu, "CTSH", "rs1369324")
> summary(cs2[[3]])

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64164 -0.16139 -0.04057  0.17366  0.64218

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.04638    0.05467 183.753   <2e-16 ***
X           -0.14788    0.06594  -2.243   0.0298 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2947 on 46 degrees of freedom
Multiple R-Squared: 0.09857,        Adjusted R-squared: 0.07897
F-statistic:  5.03 on 1 and 46 DF,  p-value: 0.02977
```

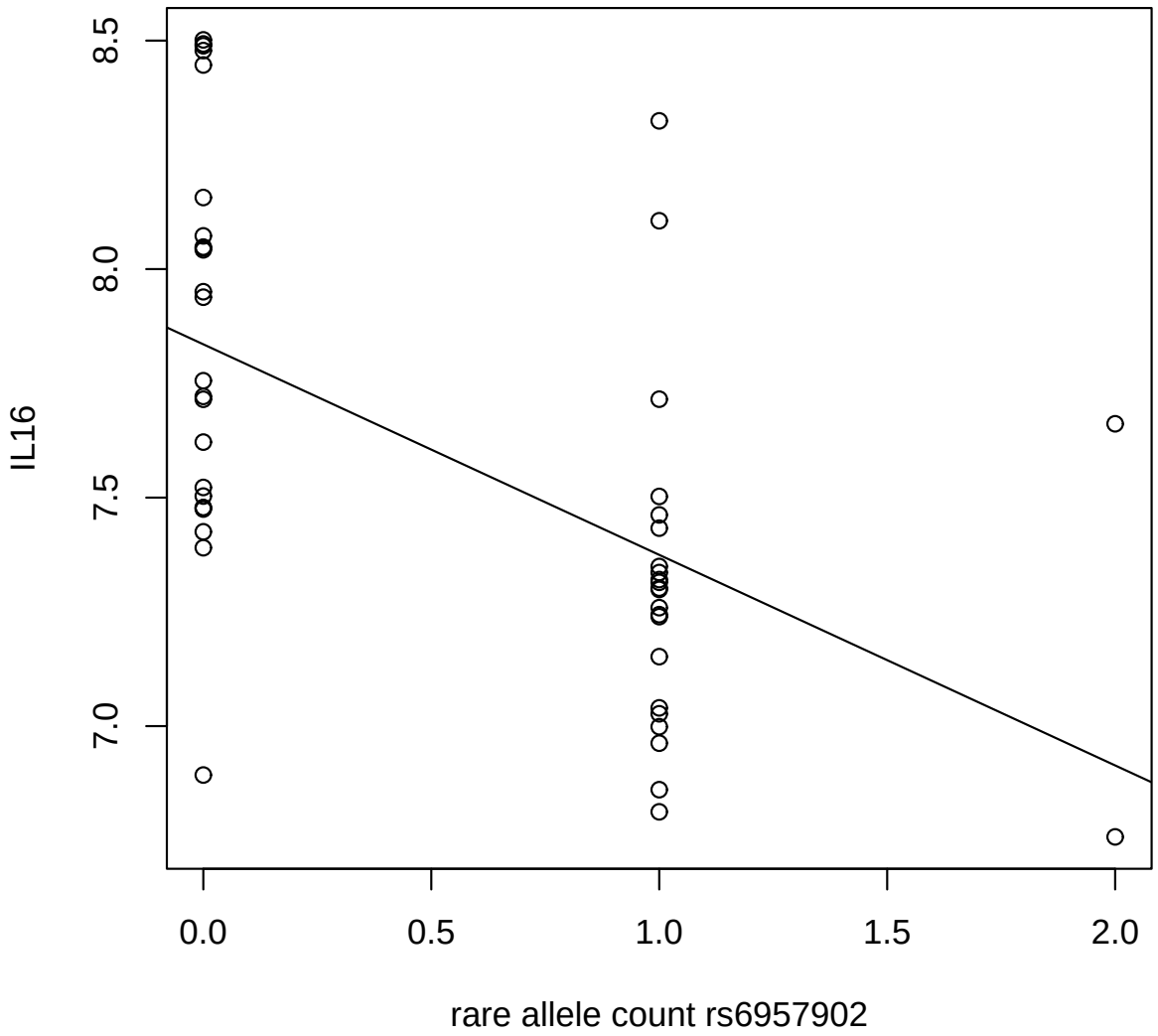

CTSH
9.2
9.4
9.6
9.8
10.0
10.2
10.4
10.6
0.0
0.5
1.0
1.5
2.0
rare allele count rs1369324

a trans example (IL16 [resident on chr15], determinant on chr 7]), random set of snps (1/50)

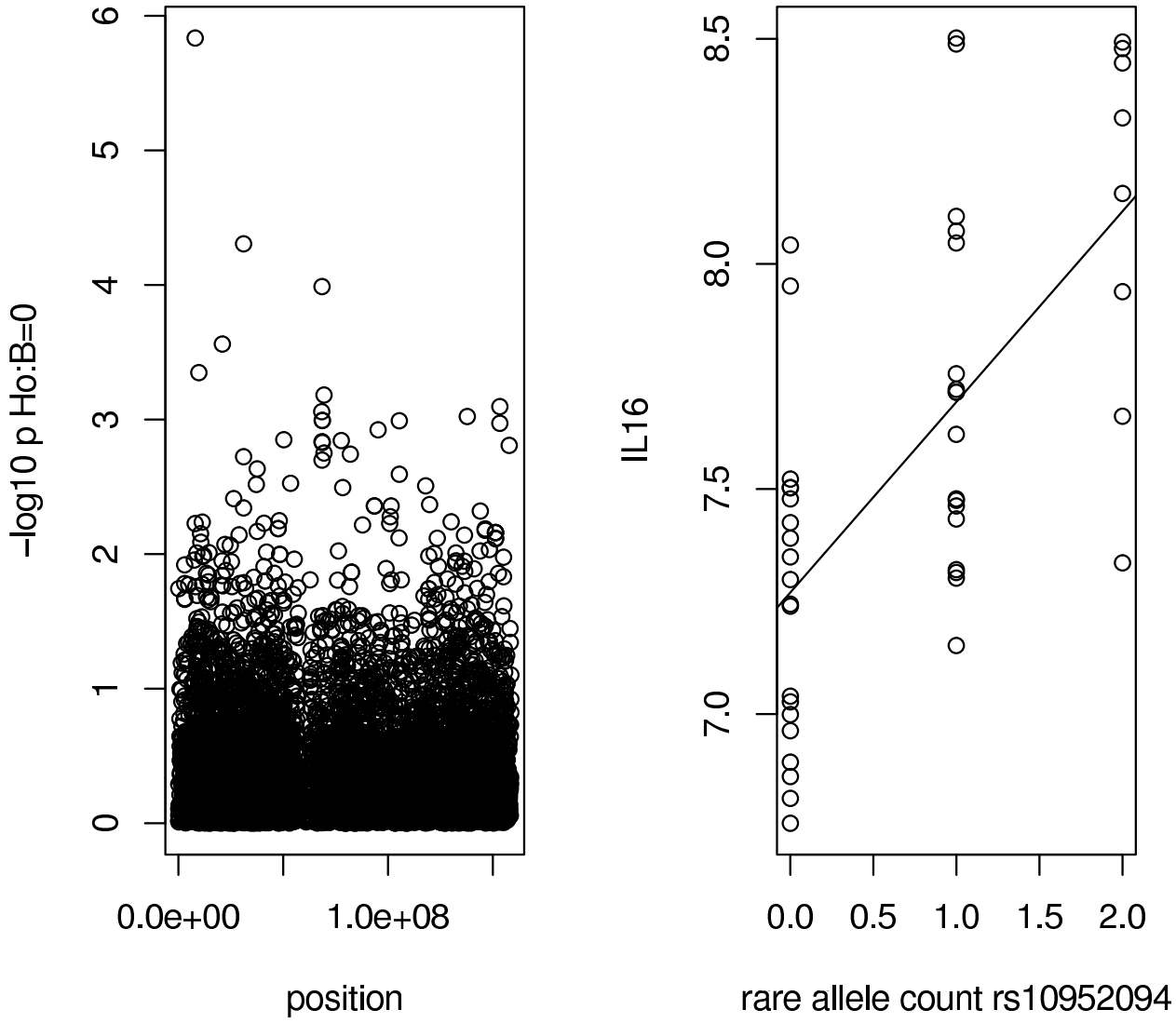

209827_s_at chr7
−log10 p Ho:B=0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0e+00
1.0e+08
position
IL16
7.0
7.5
8.0
8.5
0.0
0.4
0.8
rare allele count rs6972377

if we focus on the finding of CS:

```
> csil16 = ggrplot(c7GGceu, "IL16", "rs6957902")
> summary(csil16[[3]])

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9424 -0.3328 -0.0742  0.2112  0.9496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.83542    0.08127  96.408  < 2e-16 ***
X           -0.46072    0.10280  -4.482 4.89e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4101 on 46 degrees of freedom
Multiple R-Squared: 0.3039,        Adjusted R-squared: 0.2888
F-statistic: 20.08 on 1 and 46 DF,  p-value: 4.889e-05
```

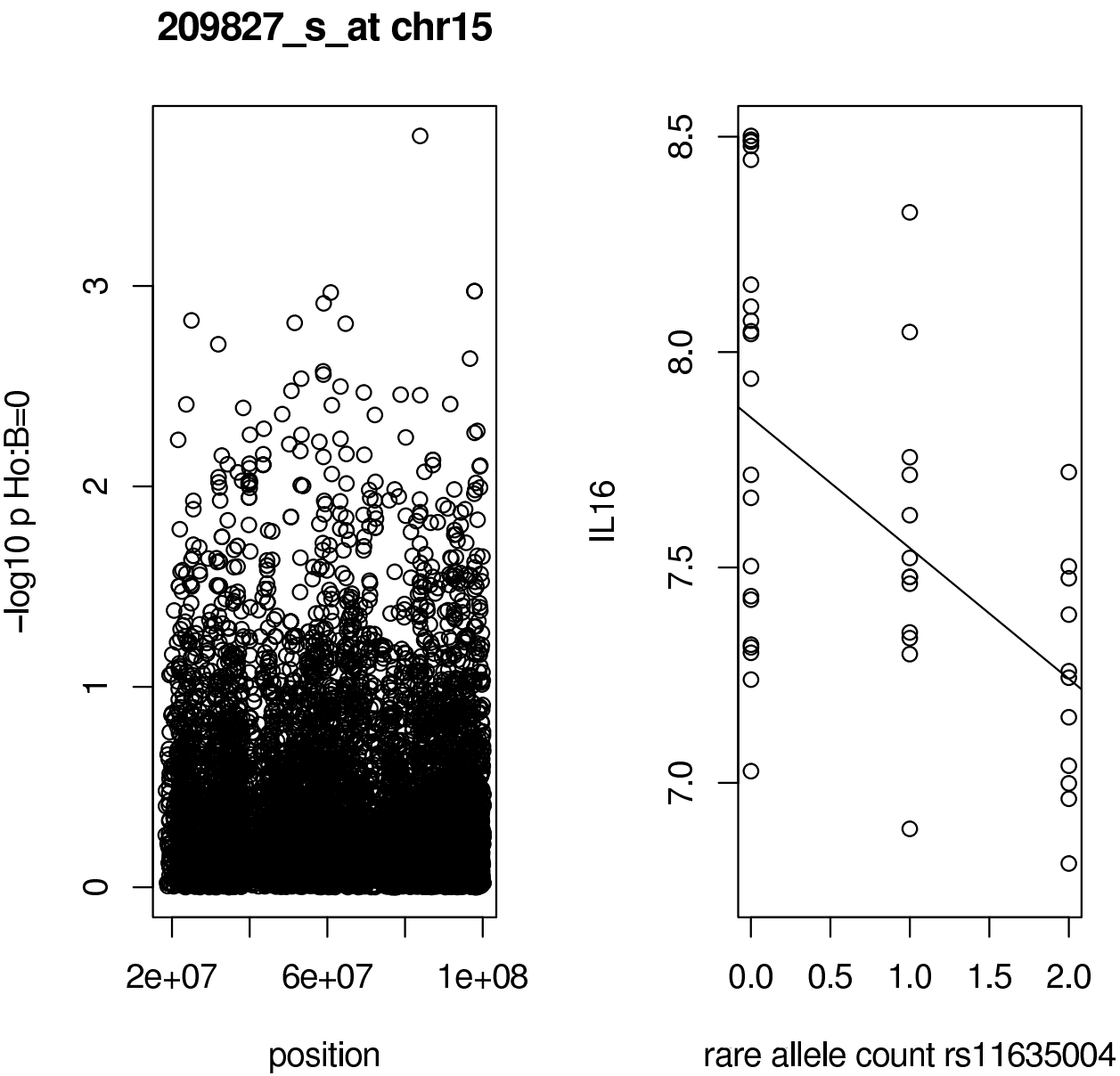

IL16
8.5
8.0
7.5
7.0
0.0
0.5
1.0
1.5
2.0
rare allele count rs6957902

screen at a fraction of 1/20 snps on chr7:



```
> vcil16
SNPper SNP metadata:
     DBSNPID      CHROMOSOME POSITION  ALLELES VALIDATED
[1,] "rs10952094" "chr7"     "8011051" "A/C"   "Y"
There are details on 3 populations
and 3 connections to gene features



> csil16
SNPper SNP metadata:
     DBSNPID     CHROMOSOME POSITION   ALLELES VALIDATED
[1,] "rs6957902" "chr7"     "68383269" "C/T"   "Y"
There are details on 4 populations
and 1 connections to gene features

SNPper info:
     SOURCE           VERSION              GENOME DBSNP
[1,] "*RPCSERV-NAME*" "$Revision: 1.38 $" "hg17" "123"

> geneDetails(vcil16)
  HUGO LOCUSLINK                                NAME
```

```
1 ICA1      3382 islet cell autoantigen 1 isoform 1
2 ICA1      3382 islet cell autoantigen 1 isoform 3
3 ICA1      3382 islet cell autoantigen 1 isoform 2
(Intron role noted)
```


209827_s_at chr15
−log10 p Ho:B=0
0
1
2
3
2e+07
6e+07
1e+08
position
IL16
7.0
7.5
8.0
8.5
0.0
0.5
1.0
1.5
2.0
rare allele count rs11635004

notes

- compact representation of assay data (expr+snp) feasible, leads to simple workflow
- detScreen function should be configurable (alternatives to OLS with 0-1-2 genotype representation)
- competitive trans determinants easily discoverable
- linking trans findings to target gene via networks? other organizations?

4 Bibliography

References

V. G. Cheung, R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, and J. T. Burdick. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437(7063):1365–9, 2005. 1476-4687 (Electronic) Journal Article.

M. Kanehisa. A database for post-genome analysis. *Trends in Genetics*, 13: 375–376, 1997.

M. Kanehisa, S. Goto, S. Kawashima, et al. The KEGG resources for deciphering the genome. *Nucleic Acids Res*, 32:D277–D280, 2004.