# Final Project: GLANCE Study

Wendy Wang and Ning Li

Tuesday, March 10, 2015

# Setting up some options

Let's first turn on the cache for increased performance and improved styling

```
# Set some global knitr options
library("knitr")
opts_chunk$set(tidy=TRUE, size='Huge',tidy.opts=list(blank=
```

# Reference

**Navarro S, White E, Kantor ED, Zhang Y, Rho J, Song X, Milne GL, Lampe PD, Lampe JW. Randomized trial of glucosamine and chondroitin supplementation on inflammation and oxidative stress biomarkers and plasma proteomics profiles in healthy humans. PLoS One. 2015; 10(2):e0117534**

# Outline

- Background
- Study Design and Data Collection
- Statistical Analysis Plan
- Data Analysis and Results
- Limitations
- Conclusions

# Background

**Glucosamine and Chondroitin (G&C)**

- One of the most popular dietary supplements in the US
- Mostly taken for osteoarthritis (OA)
- Associated with a 27-35% lower incidence of colorectal cancer, a 26-28% lower incidence of lung cancer, 17% lower overall mortality, and a 13% lower cancer mortality.
- Safe supplements, with no known major adverse side effects.
- Regular dose
- 1500 mg/d glucosamine hydrochloride (GHCl) + 1200 mg/d chondroitin sulfate (CS)

# Background

**Mechanisms of G&C on cancer prevention**

- Several lines of evidence from *in vitro* and preclinical studies support a possible role for G&C in reducing inflammation.
- Potentially through inhibition of nuclear factor kappa B pathway.
- No human intervention trials have been evaluated the effect of G&C on reducing inflammation or altering other pathways in healthy individuals

# Study Design

- Randomized, double-blinded, placebo-controlled crossover trial
- Ranomized on the order of treatment or placebo period
- Each intervention period lasted 28 days with a 28 day washout period in-between

# Participants

- 18 subjects: 9 males and 9 females
- Healthy, overweight(25<BMI<=32.5), non-smoking, aged 20-55 years individuals in the greater Seattle area.
- Strict exclusion criteria (e.g. medical history, medication use, large weight change, alcohol intake, supplemental use, vegetarian dietary and abnormal laboratory values)

# Data collection

- Blood samples were collected after each 28-day intervention period in the morning after a minimum of a 12-hour overnight fast.
- Blood was drawn into a tube containing EDTA for plasma. All samples were aliquoted and stored at -80°C.

# Proteomics Analysis

- Plasma samples were evaluated on a customized antibody array populated with ~3,000 full-length antibodies, printed in triplicates.
- Protein (200 µg) from a pool of albumin and IgG-depleted plasma were labeled with Cy5. Each sample was combined with reference sample labeled with Cy3.
- Unbound proteins were removed by washing and the slides scanned for Cy3 and Cy5 fluorescence in an Axon Genepix 4000B scanner.
- The Cy5/Cy3 ratio determined the relative concentration of protein compared to reference.
- Most (>85%) antibodies on the array had coefficients of variation, for triplicates, of less than 10%. Antibodies with CV>10% were excluded for further analysis.

# Array Analysis and Normalization

- The array image was scanned using a GenePix 4000B (Axon Instruments) scanner.
- For each antibody, fold-change of the signal (red channel) was compared to the reference (green channel) after background correction, the ratio noted as M value.
- Experimental variation was normalized using within-array print-tip loess and between-array quartile normalization.
- Triplicate features were summarized using their median. M values were standardized such that the mean value and standard deviation of the placebo groups were set to zero and one, respectively.
- After all processing, data were available for analysis on a total of 2938 antibodies.

# Statistical Analysis

**Overview of data**

- We have 18 treatment and 18 placebo samples
- We delete any genes with more than 50% missing values
- Each has data on a 2938 antibodies
- About 1100 of the antibodies are duplicates

*Would this be of concern? (Stay tuned!)*

# Statistical Analysis Plans

**Plan 1. Replicate the reported results**

- Using linear regression to adjust for batch effects
- Paired t-test comparing treatment and placebo
- GSEA analysis using Wilcoxon Test

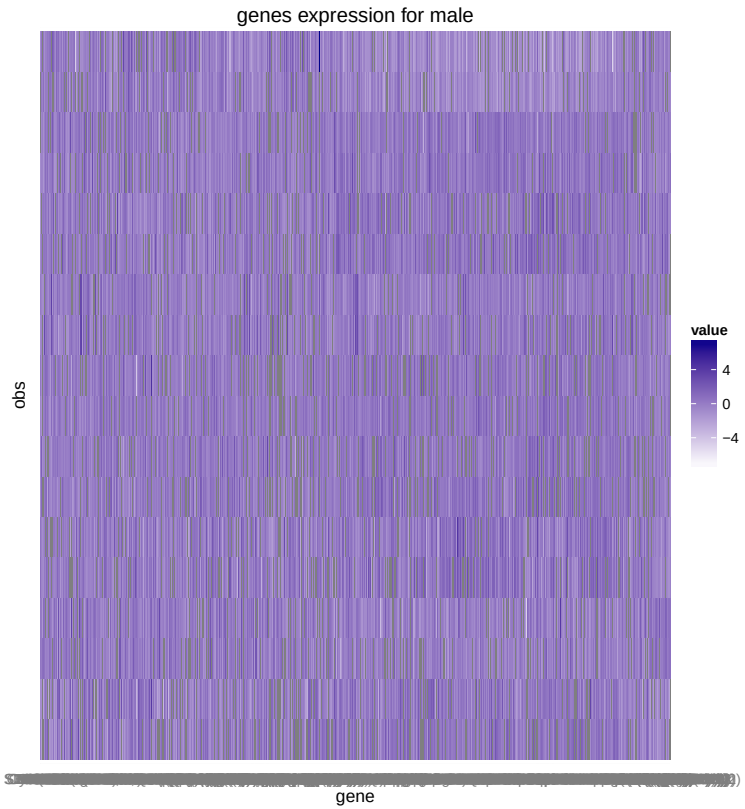
# Statistical Analysis Plans

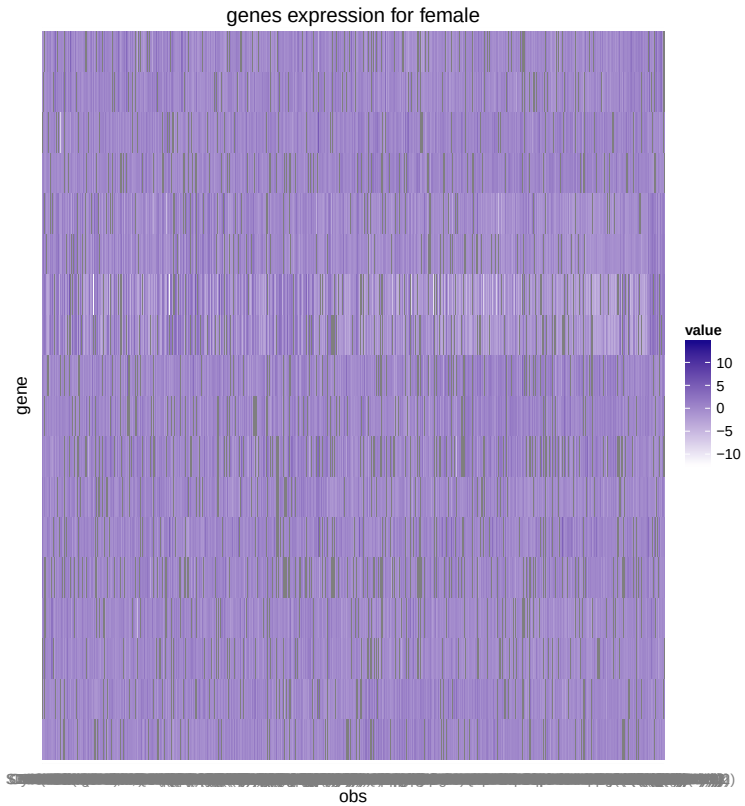
**Plan 2. Alternative analysis plan**

- Clean duplicated antibodies
- (Same as plan 1) Using linear regression to adjust for batch effects
- Moderated T test using R Limma package
- GSEA using GEGG and GO pathway databases and CAMERA

# Statistical Analysis Results -Plan 1

- Batch effects check(graph)



# Statistical Analysis Results -Plan 1

- Batch effects check(multiple tests)

```
round(Table, digits = 3)
```

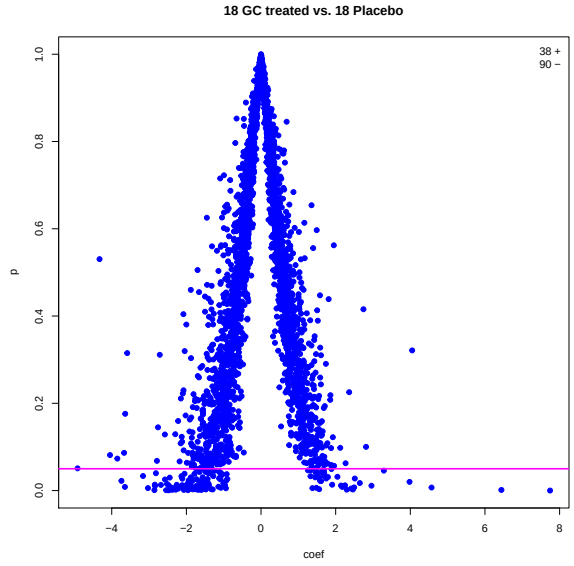
```
##                          Hybrid.day   box position gender
## prop(p-value<.05)             0.316 0.525    0.016  0.316
## prop(bonferroni<.05)          0.028 0.016    0.000  0.028
```

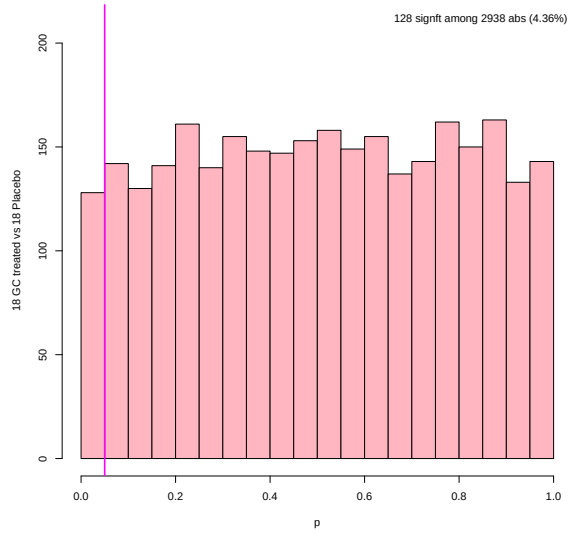
# Statistical Analysis Results -Plan 1

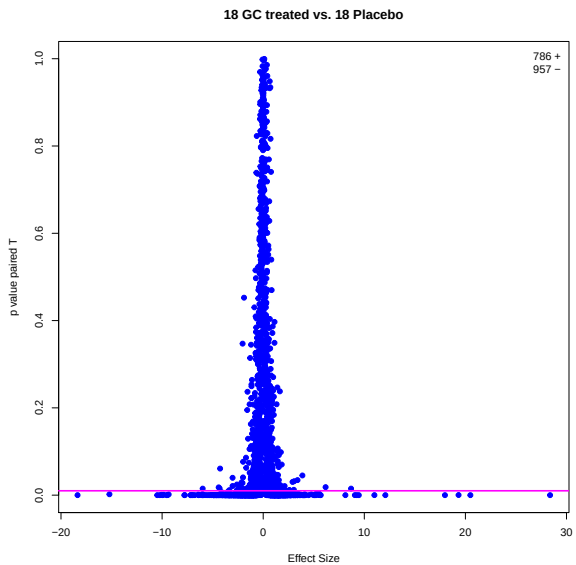
- Using linear regression to adjust for batch effects *m~case+Hybrid.day+box+position+gender*
- Paired t-test comparing treatment and placebo
- The therethold value for q value and bonferroni test is 0.001.

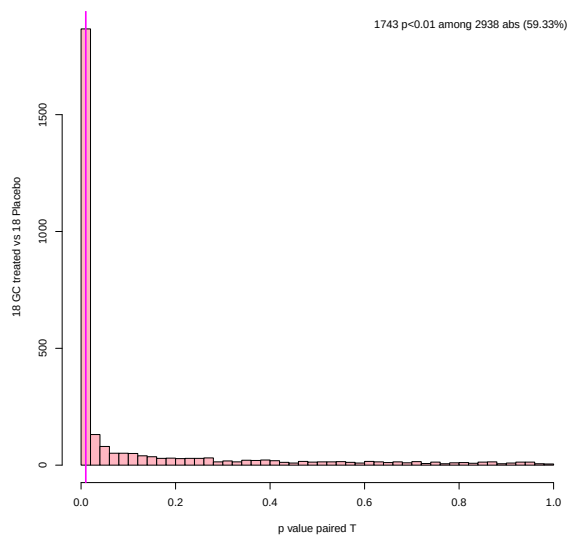
```
##
## TRUE
##   36

##    aliquot Array.number Slide.ID Hybe.day  Hybe.date Pri
## 20    1033             3 10013621   3-redo 11/22/2013
##    Freezer.box.row Freezer.box.column Case.status Hybric
## 20               A                  3           ?
##                                Array.name number    id inter
## 20 s03_1033_slide10013621_h112213redo      3 80002
##    position batch gender case
## 20        3     1      1    0

##
```

# Statistical Analysis Results -Plan 1



# Statistical Analysis Results -Plan 1

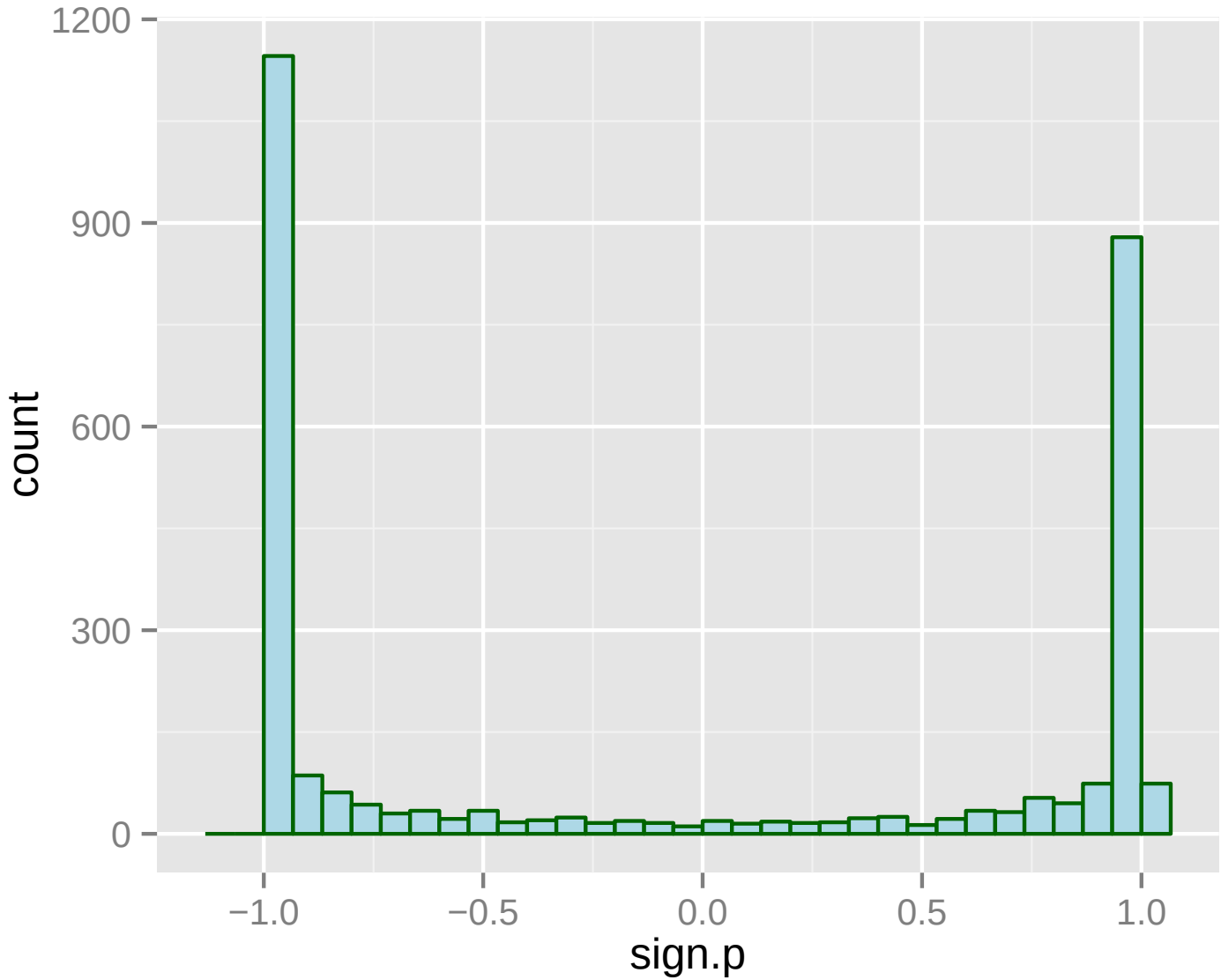
```
head(out)
```

```
##                                        Name effect.size
## 2901 Plate9_O12_CEACAM1(Std)_Sigma   -2.324371 2.288379e
## 511         Plate2_A22_ITGA5 _Abcam   -1.741159 2.665558e
## 1570            Plate5_M9_itgb4_R&D   -1.962622 6.898708e
## 1373         Plate5_E10_C1orf38_SDI    4.036857 1.141251e
## 120          Plate1_F12_MUC3B_Aviva    4.140094 3.204087e
## 819         Plate2_P5_GADD45A_Abcam   -1.137069 4.832897e
##               bonfer
## 2901 6.723256e-11
## 511  7.831409e-11
## 1570 2.026840e-10
## 1373 3.352996e-10
## 120  9.413608e-10
## 819  1.419905e-09
```

# Statistical Analysis Results -Plan 1

- GSEA analysis using Wilcoxon Test
- a Wilcoxon test to test for diserences in the distribution of p values between SNPs within the gene set under test and a control set of SNPs.
- The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ

# Statistical Analysis Results -Plan 1



# Statistical Analysis Results -Plan 1

- GSEA analysis using GO pathways

```
## [1] "Number of unique genes:" "1272"
```

```
## [1] "Number of Gene Sets:" "1353"
```

```
## NUCLEOPLASM
##          25
```

```
## NUCLEOPLASM
##           8
```

```
## [1] 33
```

# Statistical Analysis Results -Plan 1

- GSEA analysis using GO pathways

```
head(gene.set)
```

```
##                                                                MasterGr
## EXTRACELLULAR_REGION
## NUCLEOPLASM
## INSOLUBLE_FRACTION
## GTPASE_REGULATOR_ACTIVITY
## ER_GOLGI_INTERMEDIATE_COMPARTMENT
## RNA_POLYMERASE_II_TRANSCRIPTION_FACTOR_ACTIVITY
##
## EXTRACELLULAR_REGION
## NUCLEOPLASM
## INSOLUBLE_FRACTION
## GTPASE_REGULATOR_ACTIVITY
## ER_GOLGI_INTERMEDIATE_COMPARTMENT
## RNA_POLYMERASE_II_TRANSCRIPTION_FACTOR_ACTIVITY RNA_POLY
```

# Statistical Analysis Results -Plan 1

- Repeat GSEA analysis using KEGG pathways

```
## [1] "Number of unique genes:" "890"
```

```
## [1] "Number of Gene Sets:" "178"
```

```
## KEGG_GLYCOLYSIS_GLUCONEOGENESIS
##                               8
```

```
## KEGG_GLYCOLYSIS_GLUCONEOGENESIS
##                              19
```

```
## [1] 27
```

# Statistical Analysis Results -Plan 1

- Repeat GSEA analysis using KEGG pathways

```
head(gene.set)
```

```
##                                                                MasterGr
## KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS
## KEGG_OXIDATIVE_PHOSPHORYLATION
## KEGG_NOTCH_SIGNALING_PATHWAY
## KEGG_HOMOLOGOUS_RECOMBINATION
## KEGG_RNA_DEGRADATION
## KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY
##
## KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS
## KEGG_OXIDATIVE_PHOSPHORYLATION
## KEGG_NOTCH_SIGNALING_PATHWAY
## KEGG_HOMOLOGOUS_RECOMBINATION
## KEGG_RNA_DEGRADATION
## KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY KEGG_NATU
```

# Statistical Analysis Results -Plan 2

- Clean duplicated antibodies
- we have 1892 genes and 36 samples

# Statistical Analysis Results -Plan 2

- (Same as plan 1) Using linear regression to adjust for batch effects

```
## [1] "est"    "result"
```

# Statistical Analysis Results -Plan 2

- Moderated T test using R Limma package

```
# So we run everything in the orginal codes before paired
# instead of pairted t test to fit the model
mm <- model.matrix(~0 + id + case, data = info.temp)
mm <- mm[, -11]
mm <- mm[, -(19:24)]
fit <- lmFit(est.m, mm)
```

```
## Warning: Partial NA coefficients for 464 probe(s)
```

# Statistical Analysis Results -Plan 2

- Moderated T test using R Limma package

```
ebay <- eBayes(fit)
top <- topTable(ebay, coef = "case1", number = Inf, sort.by
head(top)
```

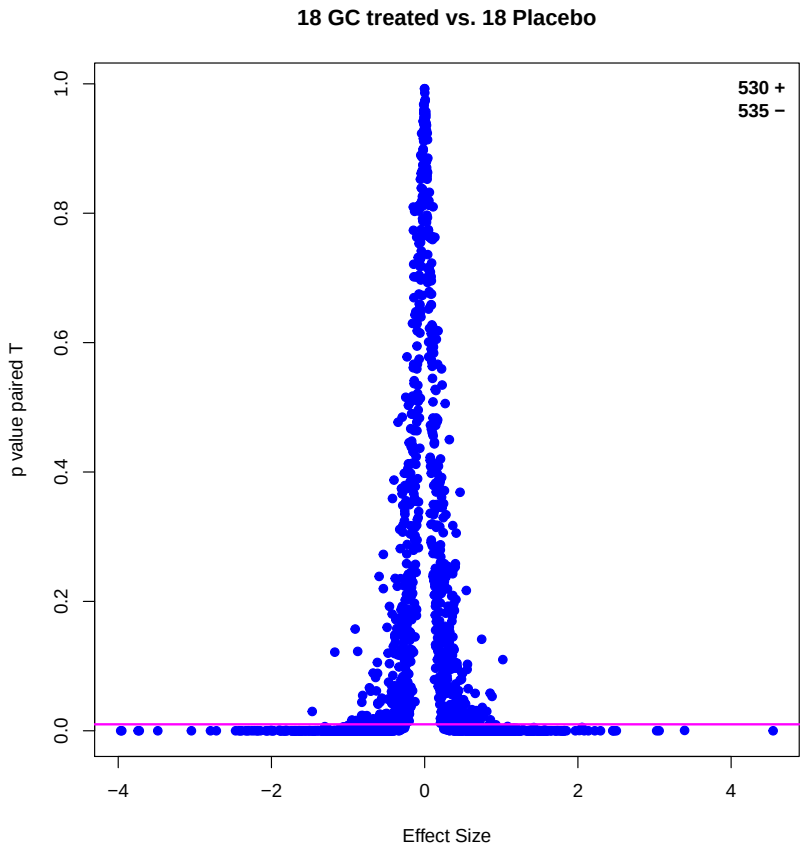
```
##              logFC     AveExpr         t      P.Value    ad
## 199    2.2948168 -0.03505333  23.24965 2.468302e-15 4.670
## 1145   2.4549708 -0.72395072  23.90062 5.343684e-15 5.055
## 906   -1.3556617  0.08320615 -18.98122 3.438212e-14 2.168
## 407   -0.9629058  0.25137131 -17.39349 6.944883e-14 3.284
## 991    2.5016576 -1.34374942  20.12273 1.036342e-13 3.719
## 16     1.4329576  0.53263945  16.92802 1.179410e-13 3.719
```

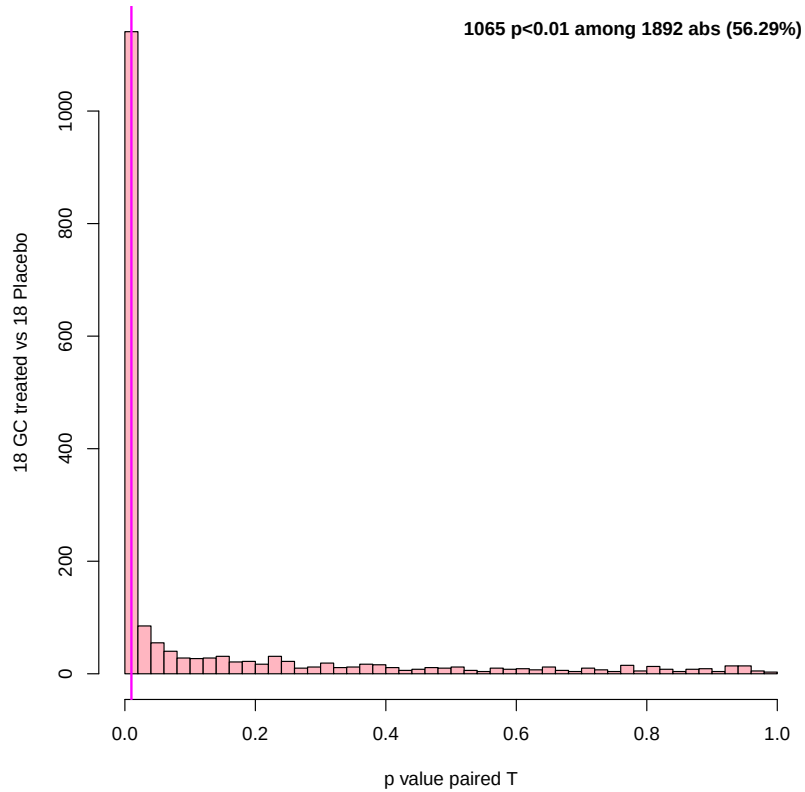
```
mean(top$adj.P.Val < 0.01)
```

```
## [1] 0.5628964
```

# Statistical Analysis Results -Plan 2

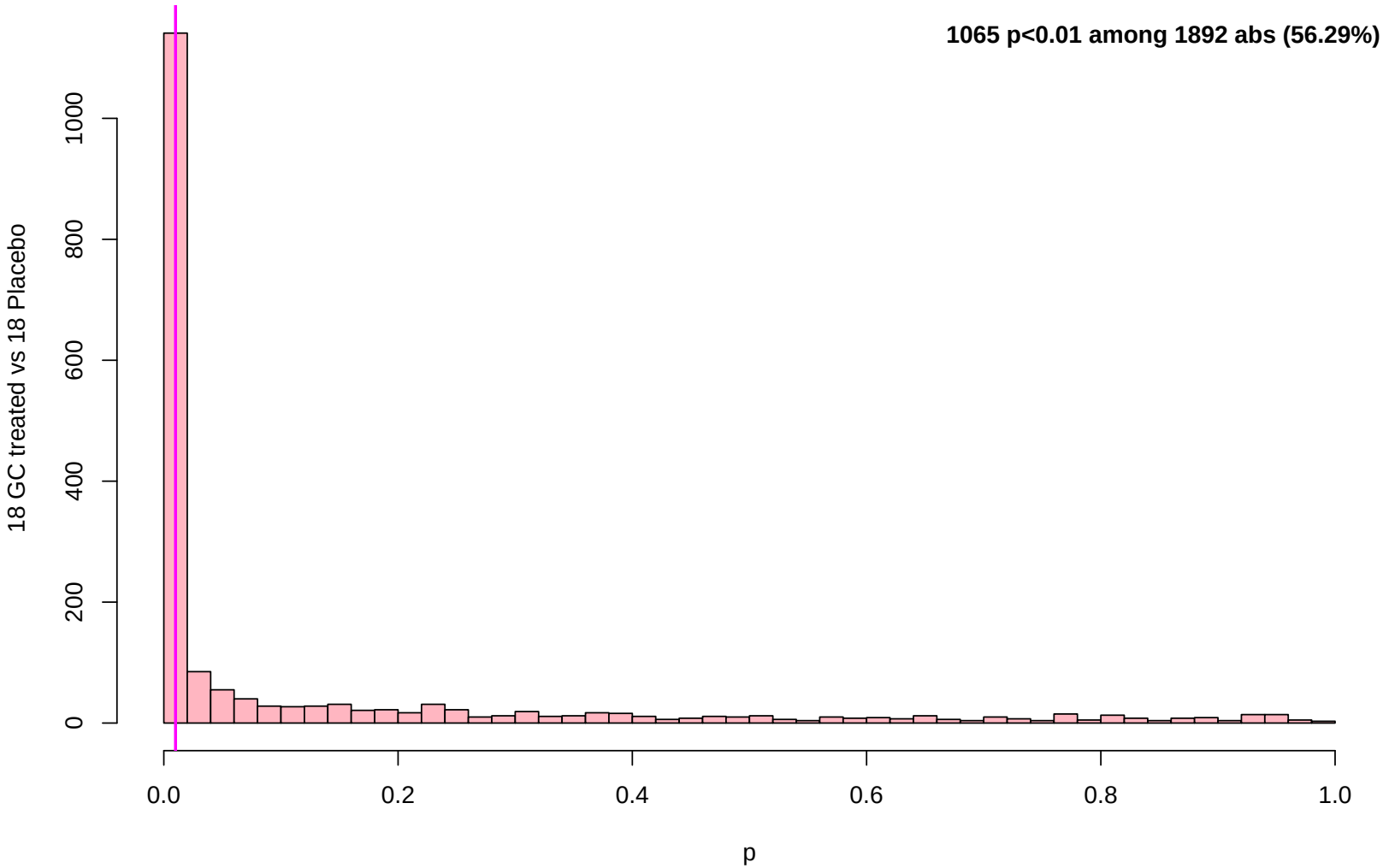
- Moderated T test using R Limma package *Plots*



# Statistical Analysis Results -Plan 2

- Moderated T test using R Limma package *Plots*



# Statistical Analysis Results -Plan 2

- GSEA using KEGG pathway database and CAMERA

```
kegg <- getGmt("Data/c2.cp.kegg.v4.0.symbols.gmt")
gene_ids <- geneIds(kegg)
```

```
if (exists("ids2indices")) {
    sets_indices <- ids2indices(gene_ids, unique.gene)
}
if (exists("symbols2indices")) {
    sets_indices <- symbols2indices(gene_ids, unique.gene)
}
cont_matrix <- makeContrasts("case1", levels = mm)
# gsea <- camera(est_new, sets_indices, design=mm, cont_ma
# due to the NA's So we change all the NA's to zero
est_narm <- est_new
est_narm[is.na(est_narm)] <- 0
gsea_kegg <- camera(est_narm, sets_indices, design = mm, co
```

# Statistical Analysis Results -Plan 2

- GSEA using KEGG pathway database and CAMERA

```
head(gsea_kegg)
```

```
##                                          NGenes  Correlation
## KEGG_PRION_DISEASES                          18 -0.00834646
## KEGG_ARGININE_AND_PROLINE_METABOLISM          6 -0.10347407
## KEGG_HEMATOPOIETIC_CELL_LINEAGE              38 -0.00839779
## KEGG_ADHERENS_JUNCTION                       28 -0.00197926
## KEGG_WNT_SIGNALING_PATHWAY                   63  0.01882816
## KEGG_HUNTINGTONS_DISEASE                     24 -0.01163709
##                                            PValue          
## KEGG_PRION_DISEASES                   0.007455228 0.6167
## KEGG_ARGININE_AND_PROLINE_METABOLISM  0.012645655 0.6167
## KEGG_HEMATOPOIETIC_CELL_LINEAGE       0.013293206 0.6167
## KEGG_ADHERENS_JUNCTION                0.024288061 0.6167
## KEGG_WNT_SIGNALING_PATHWAY            0.034002868 0.6167
## KEGG_HUNTINGTONS_DISEASE              0.034243671 0.6167
```

# Statistical Analysis Results -Plan 2

- GSEA using GO pathway database and CAMERA

```
go <- getGmt("Data/c5.all.v4.0.symbols.gmt")
gene_ids_go <- geneIds(go)
if (exists("ids2indices")) {
    sets_indices_go <- ids2indices(gene_ids_go, unique.gene
}
if (exists("symbols2indices")) {
    sets_indices_go <- symbols2indices(gene_ids_go, unique
}
cont_matrix <- makeContrasts("case1", levels = mm)
gsea_go <- camera(est_narm, sets_indices_go, design = mm, c
```

# Statistical Analysis Results -Plan 2

- GSEA using GO pathway database and CAMERA

```
head(gsea_go)
```

```
##                                                                 NGe
## TRANSMEMBRANE_RECEPTOR_PROTEIN_PHOSPHATASE_ACTIVITY
## SH2_DOMAIN_BINDING
## MICROTUBULE_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS
## TRANSLATION
## CHROMOSOMEPERICENTRIC_REGION
## INTRA_GOLGI_VESICLE_MEDIATED_TRANSPORT
##                                                                 Dir
## TRANSMEMBRANE_RECEPTOR_PROTEIN_PHOSPHATASE_ACTIVITY
## SH2_DOMAIN_BINDING
## MICROTUBULE_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS
## TRANSLATION
## CHROMOSOMEPERICENTRIC_REGION
## INTRA_GOLGI_VESICLE_MEDIATED_TRANSPORT
```

# Limitation

- Small sample size
- Only detecting proteints that have antibodies on the microarray

# Conclusion and Discussion

- Results for individual proteins are similar
- Results for pathway analysis are very different.

**Why?** - We set all missing values to 0 - Different methods: paired t-test vs. LIMMA; Wilcoxon test vs. CAMERA

# Questions?