Module 6: Statistical Analysis Methods

Aims

- Current statistical methods used in automated annotation pipelines
- Discussion of issues with these methods
- Introduction of possible alternative

Introduction

The primary goal of high throughput phenotyping projects is to determine the effect of a particular genomic alteration on the phenotype of an organism. This is done by measuring a large number of biological parameters on individuals with and without the genomic alteration. Comparison of these parameters is complicated by the relatively small sample sizes necessarily used (frequently 7 of each gender) as well as the multitude of factors that affect the phenotype. A number of statistical methods have been used to attempt to extract the true signal from the significant noise, and more advanced methods are being worked on for implementation in the next stages of the project.

Current Methods

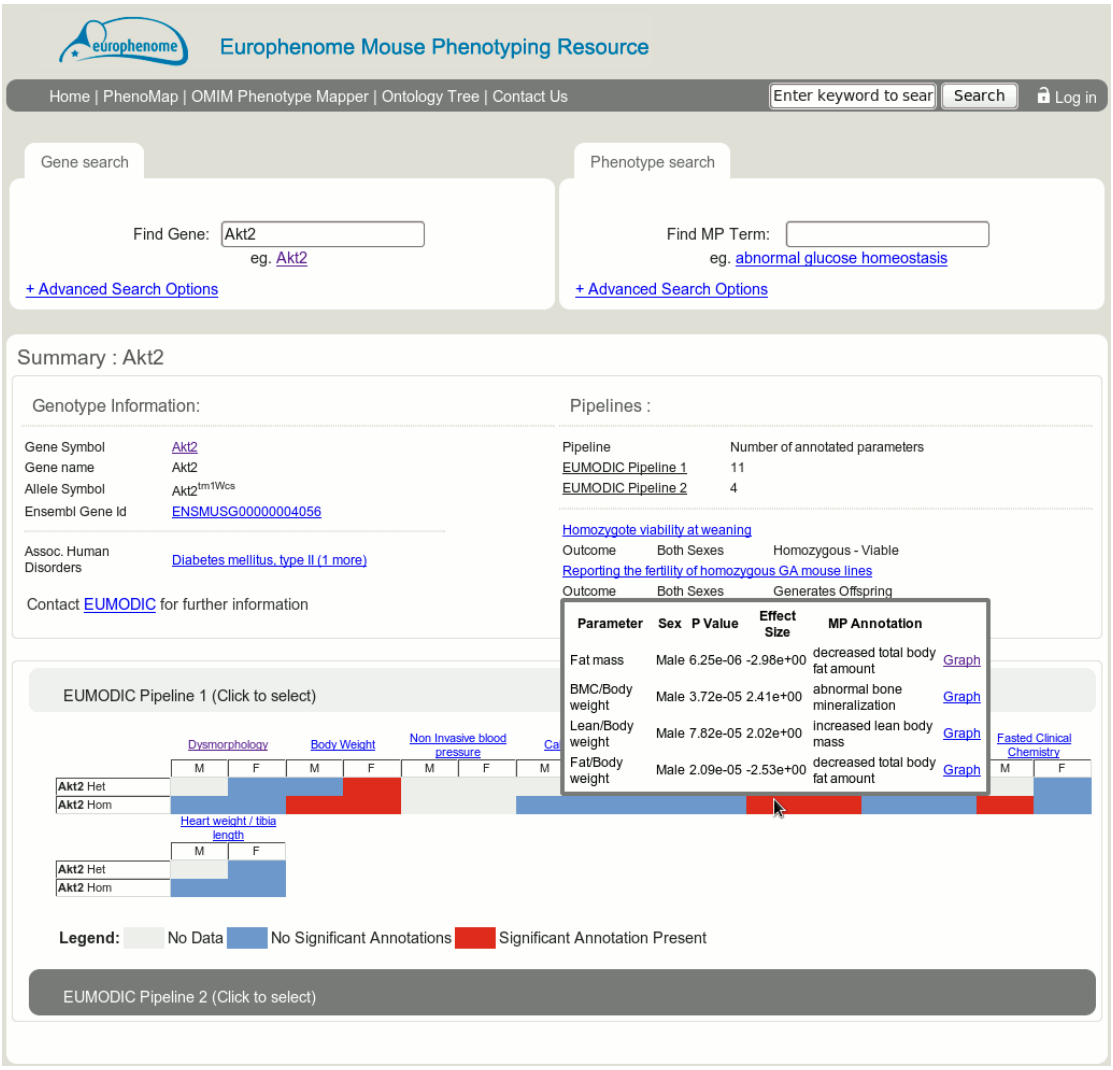
Europhenome Methods

In EuroPhenome all numerical parameters are annotated using the Wilcoxon rank-sum test (also known as the Mann–Whitney U test and the Mann–Whitney–Wilcoxon test). This calculates the probability that given that the 2 samples come from the same population that you would get results this far from the expected equal sample distributions. It makes no assumptions as to the distribution of the population, just that the observations are independent of each other.

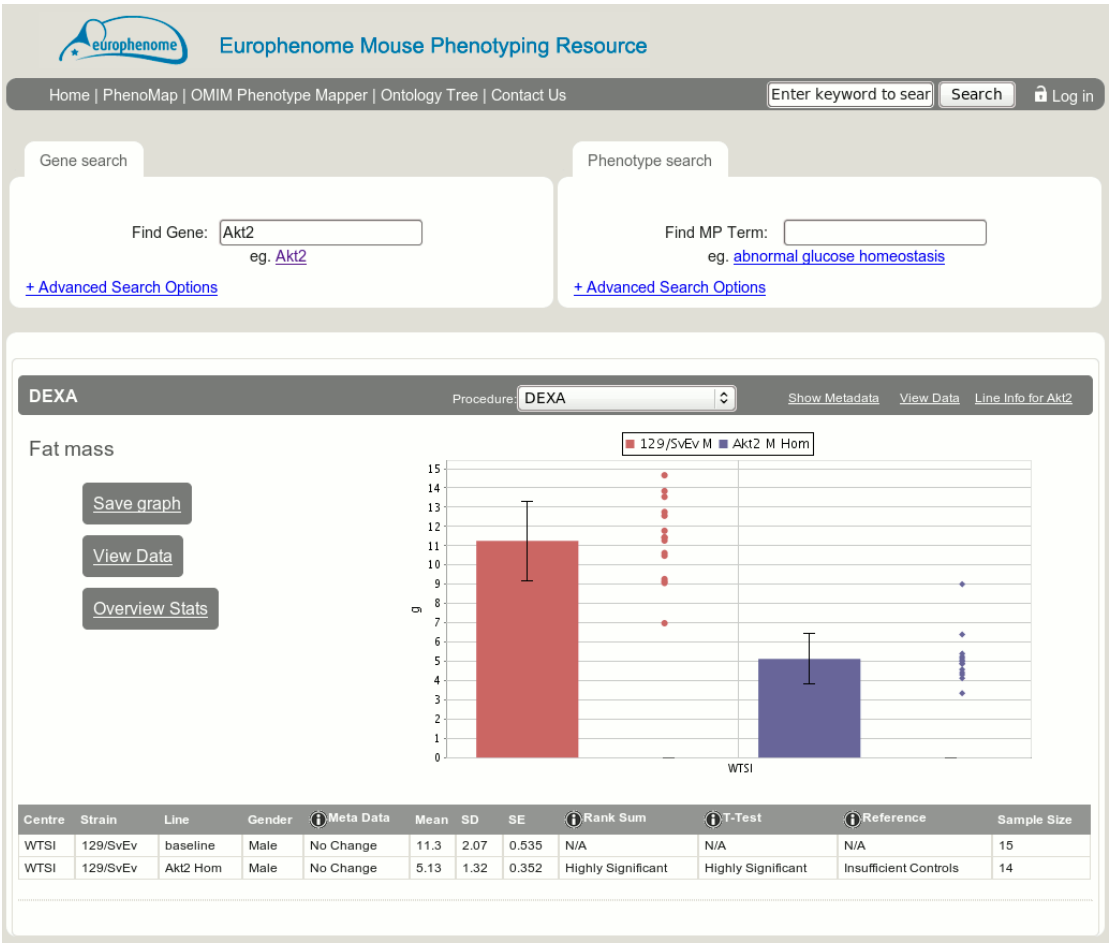
This could be performed by arranging all the observations into a single ranked series. Taking the group with the lowest results, take each observation and count the observations on the other group that have lower values. The sum of these counts is the test statistic U. From this the p-value can be determined.

Of course we use a computer to do this work for us. You can derive this result for yourself using the following procedure:

Find a line of interest, Akt2:



Select an annotation, eg. Fat Mass. Click on Graph, and click overview stats:



You can mouse over the statistical results to get the details, but we shall calculate them ourselves:
Click View Data, Get as CSV, Save Page
Now you need a statistical tool. I recommend R, available from http://cran.r-project.org/.
Load R, and load the file we have just saved:
(you may need to change R's working directory, something like:

```
setwd("Downloads")
```

Then load the file:

```
akt2Fat=read.csv("akt2Fat.csv")
```

We can check this has worked by looking at the summary of the file:

```
> summary(akt2Fat)
  Centre         Strain       Genotype  Zygosity  Gender
Parameter
 WTSI:29   129/SvEv:29   Akt2    :14      :15   Male:29   Fat mass:29
                         baseline:15   Hom:14


      Metadata       Animal       Increment          Value
 No Change:29   Min.   :19491   Mode:logical   Min.   : 3.350
```

```
                1st Qu.:19498   NA's:29        1st Qu.: 4.990
                Median :19663                  Median : 9.010
                Mean   :39741                  Mean   : 8.297
                3rd Qu.:84489                  3rd Qu.:11.420
                Max.   :84654                  Max.   :14.660
```

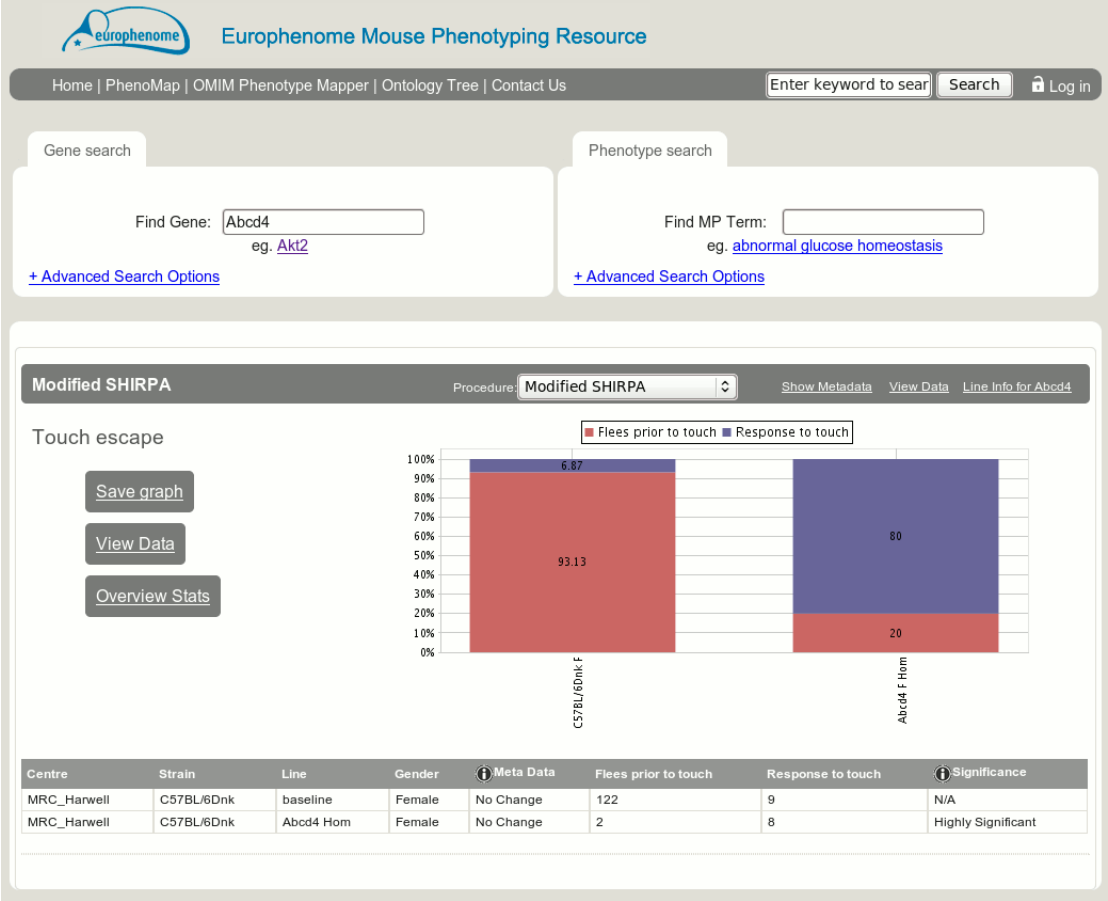
To run the rank sum test we use:

```
> wilcox.test(Value~Genotype, data = akt2Fat)

      Wilcoxon rank sum test with continuity correction

data:  Value by Genotype
W = 1, p-value = 6.252e-06
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(4.42, 9.01, 6.4, 4.99, 4.13, 4.31,  :
  cannot compute exact p-value with ties
```

This p-value agrees with what is shown on the web site. We can also calculate the t test result while we are at it:

```
> t.test(Value~Genotype,var.equal = TRUE, data = akt2Fat)

      Two Sample t-test

data:  Value by Genotype
t = -9.4185, df = 27, p-value = 5.052e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.462350 -4.792602
sample estimates:
    mean in group Akt2 mean in group baseline
              5.127857              11.255333
```

Catagorical values are easier to pull in from the summary table. Choose a line and test, in this case Abcd4, Touch escape:



These values can be loaded into R with:

```
abcd4Touch=matrix(c(122,9,2,8),2)
```

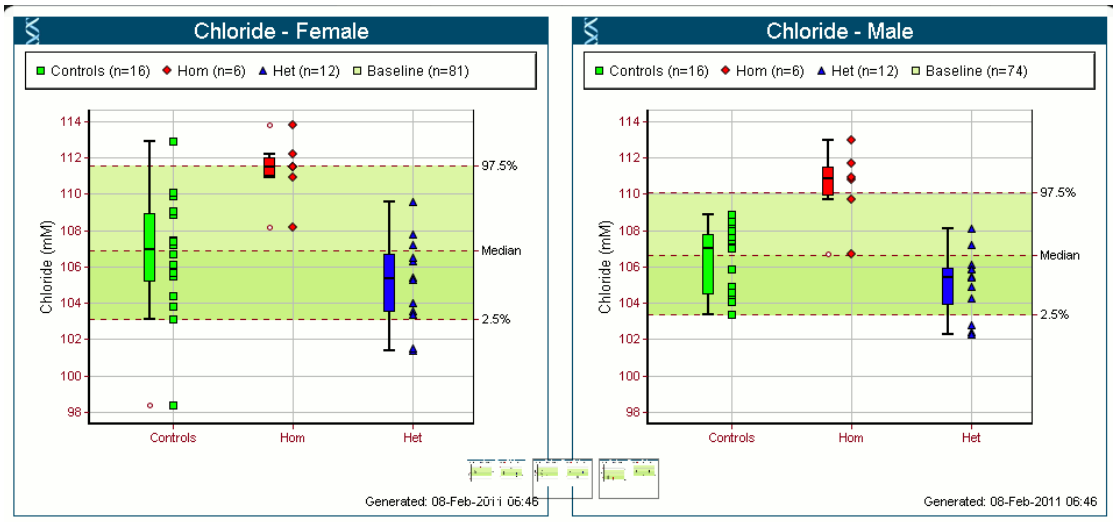
And the fisher exact test is run with:

```
> fisher.test(abc4Touch)

        Fisher's Exact Test for Count Data

data:  abc4Touch
p-value = 3.052e-07
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   8.491575 550.552750
sample estimates:
odds ratio
  50.40908
```

In all cases a default threshold of 0.0001 is used to call a phenodeviant. This is based on there being 400 measured parameters in the pipeline, so this gives a whole pipeline false positive rate of approximately 4%. This is currently under consideration and may change in the future, and in many cases can be changed by the user.

MGP Methods

The Welcome Trust Sanger Institute uses different methods. For numerical parameters they use a reference range comparison, where all appropriate control animals are grouped and the 95% reference range is calculated. If 60% of the mutant animals are outside of this range then the line is called as a phenodeviant in that parameter. This is visible on their web site by selecting a gene (Mysm1), then a phenotype area (homeostasis/metabolism) then a protocol (Energy Expenditure), then a parameter (Energy Expenditure). This results in the below graph:



view data download data

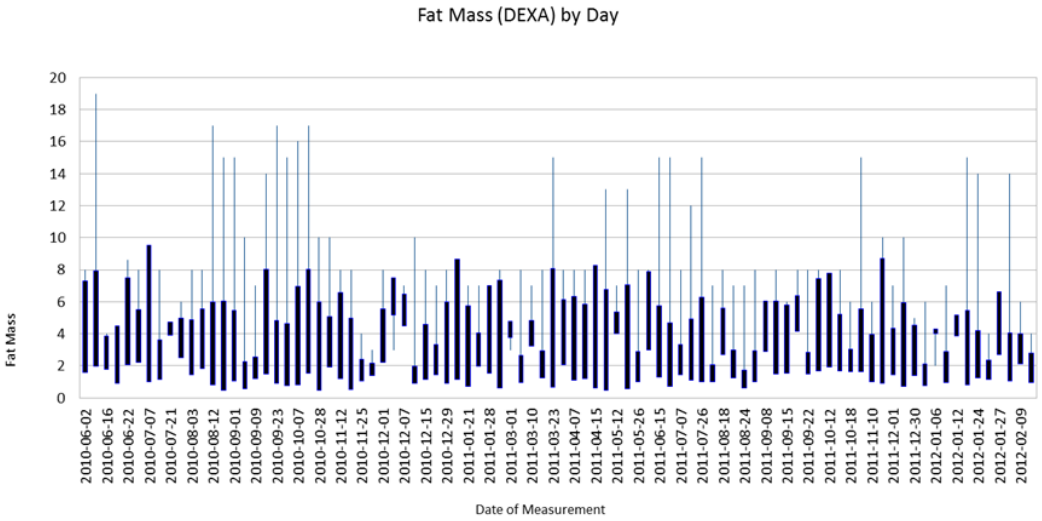
Protocol	Parameter	Female Het	Female Hom	Male Het	Male Hom	population parameter	MP Annotation
Plasma Chemistry (Avertin)	Chloride						MP:0003019 - increased circulating chloride level

For categorical parameters they also use the fishers exact test, but they use a threshold of $p < 5\%$ with an additional requirement that the absolute percentage change must be 60% or greater. So if the baseline have 10% abnormal the mutants must have 70% or more abnormal.

Problems with these methods

The principle problem with these methods is that there appears to be local structure in the data, *i.e.* animals that are measured on the same day are likely to have more similar results than animals measured on different days. The cause of this structure is currently undetermined, and is assumed to be multifactorial encompassing similarity from litter, operator and equipment, and probably other unidentified causes.

This structure is visible in the graph below. You can see that some days have much larger variation than others, and also some days have higher results than others. It is considered that this variation is greater than that expected by chance.



This problem is particularly acute as in many cases the mutant animals are measured on 1 or a small number of days. This means that tests that assume independence are likely to under estimate the false positive rate, *i.e.* call a phenodeviant when not present.

Solution to this problem

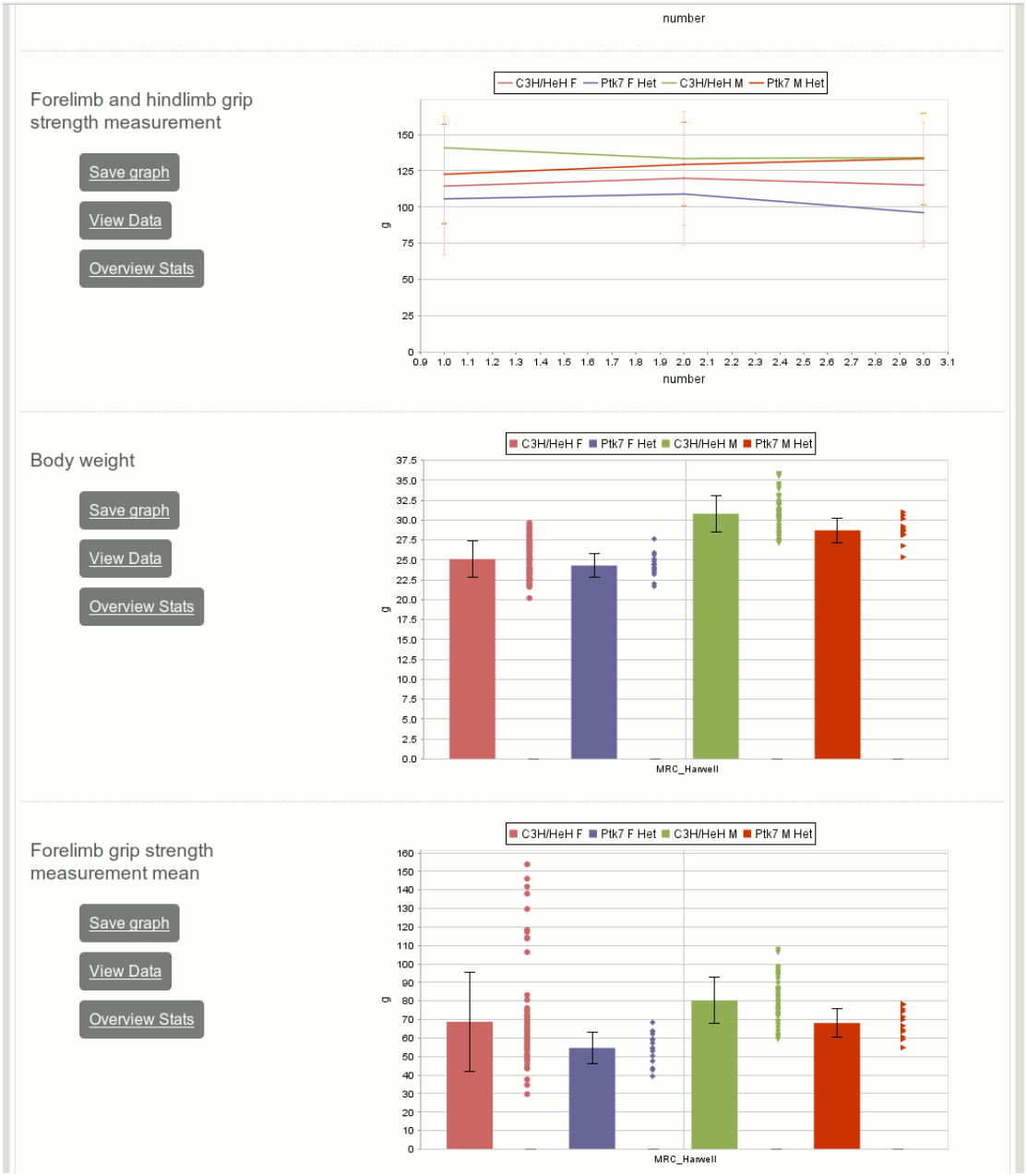
Work is ongoing as to how to most accurately detect true phenodeviants and avoid calling false positives. The currently favoured method is the Linear Mixed Model. This models each data point as the sum of 2 normal distributions, one representing the day to day variation and the other representing the animal to animal variation. To this sum are added a number of fixed terms, representing the influence of various factors that can influence the value, such as the gender of the mouse and the value of other parameters like body weight that are measured and are known to affect the result, as well as the effect of the genomic alteration and the possible gender specific effect of the genomic alteration. This calculation can be represented as a formula:

$$Y_{ij} = \beta_o + \beta_1 Genotype1_{ij} + \beta 2 Sex1ij + \beta_3 Weight1_{ij} + \beta_4 Genotype1_{ij} Sex1_{ij} + u_j + e_i$$

Maximum likelihood estimations are made for all the terms in the formula, and the final p value is given as the probability of getting distributions that different given that the true value of the genotype term is zero.

We shall work through an example of the mixed model estimations of the p value:

Select a line (Ptk7), a procedure (Grip-Strength):



Chose a parameter (Forelimb grip strength measurement mean) and click View Data and Get as CSV, then save page.

Go back to R, and load this file:

```
ptk7GS=read.csv("ptk7GS.csv")
```

You also need to load the library that does the work:

```
require(nlme)
```

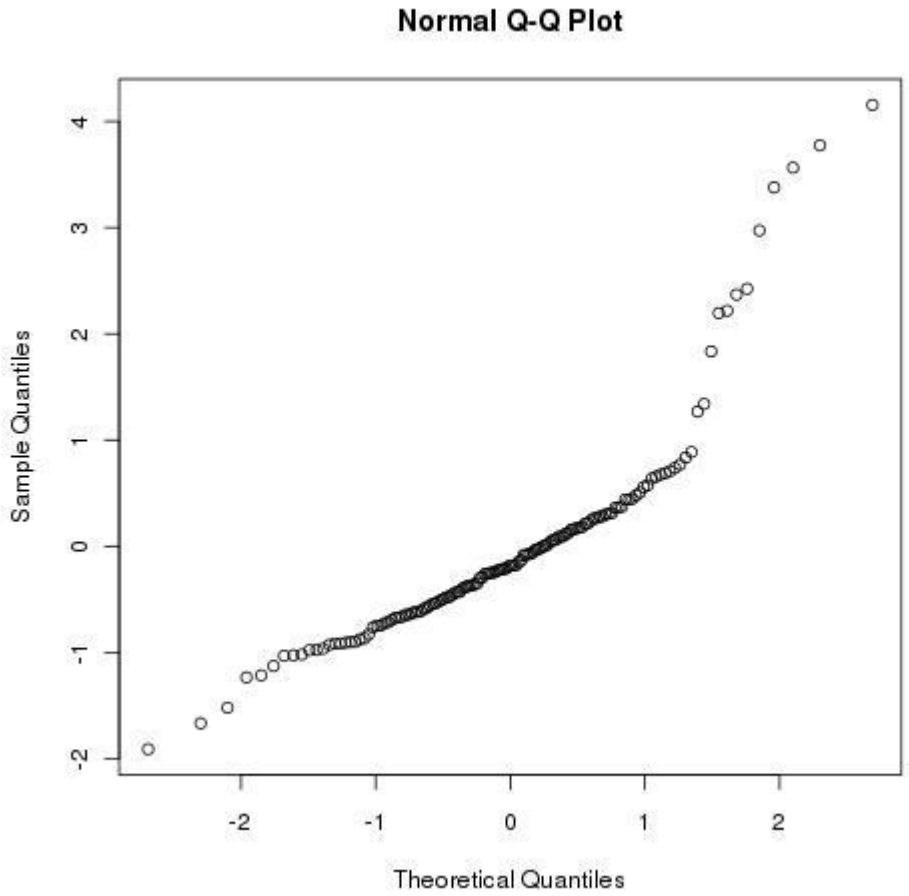
First we shall do a more simple Linear model, that does not take into account the day to day variation, but does have the advantage of taking into account both genders at the same time:

```
ptk7GSLM=lm(Value~Genotype + Gender + Genotype*Gender, ptk7GS,
na.action="na.omit")
summary(ptk7GSLM)
Call:
lm(formula = Value ~ Genotype + Gender + Genotype * Gender, data =
ptk7GS,
    na.action = "na.omit")

Residuals:
    Min      1Q  Median      3Q     Max
-39.110 -12.518  -3.763   5.735  85.090

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               68.777      2.475  27.794  < 2e-16 ***
GenotypePtk7             -14.134      5.891  -2.399  0.01777 *
GenderMale                11.454      4.011   2.855  0.00497 **
GenotypePtk7:GenderMale    1.987      8.966   0.222  0.82496
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.7 on 136 degrees of freedom
Multiple R-squared: 0.1222,   Adjusted R-squared: 0.1028
F-statistic: 6.311 on 3 and 136 DF,  p-value: 0.0004862
```

This gives us a p value of 0.01777. We can see how well this model fits the data by seeing how well the residuals fit a normal distribution:

```
ptk7GSLMRes<-residuals(ptk7GSLM)
qqnorm(scale(ptk7GSLMRes))
```

This produces the graph below, which shows the predicted values of the inter-animal difference against those expected if it was normally distributed:



Because this is not very close to a straight line we consider this model to not be a good fit, and so shall try another.

We want to introduce litter as a batch effect, and since this is a harwell line we can deduce the litter of each animal from its name (note: in IMPC we shall capture this information so this text manipulation will not be required). We know that the litter is the part of the name excluding everything after the number after the period, so if the mouse name is “CHUZHOI/114.6d_4196473” then the litter is “CHUZHOI/114.6”. To get this we load the file into excel, and extract the litter id into a new column (headed litter) with the function “=LEFT(H2,(SEARCH("_",H2)-2))”.

This file is then saved and loaded into R:

```
ptk7GSLitter=read.csv("ptk7GSLitter.csv")
```

We can then do a mixed model using the litter as a random effect:

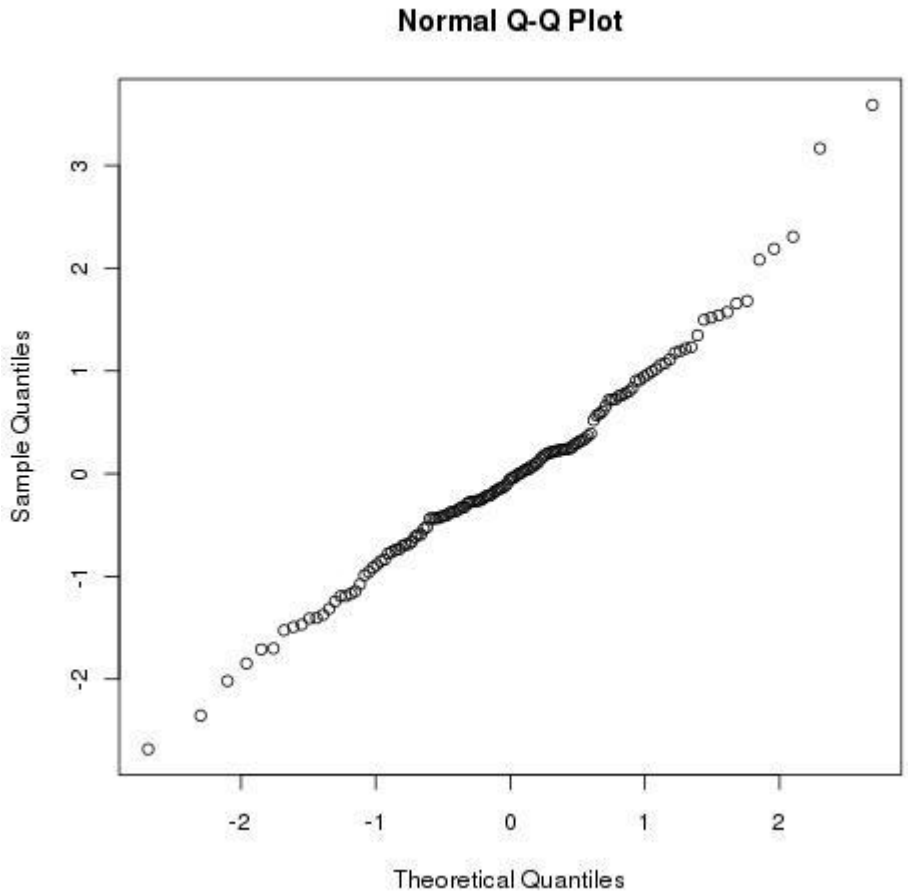
```
ptk7GSMM=lme(Value~Genotype + Gender +
Genotype*Gender,random=~1|Litter, ptk7GSLitter, na.action="na.omit")
summary(ptk7GSMM)

 Linear mixed-effects model fit by REML
 Data: ptk7GSLitter
       AIC      BIC   logLik
  1115.302 1132.778 -551.651
```

```
Random effects:
 Formula: ~1 | Litter
        (Intercept) Residual
StdDev:    18.59526 8.529573

Fixed effects: Value ~ Genotype + Gender + Genotype * Gender
                            Value Std.Error DF   t-value p-value
(Intercept)              67.02067  3.377184 85 19.845137  0.0000
GenotypePtk7            -12.05973  7.461470 85 -1.616267  0.1097
GenderMale               12.59607  4.403984 85  2.860154  0.0053
GenotypePtk7:GenderMale   1.42342  8.819061 85  0.161403  0.8722
 Correlation:
                        (Intr) GntyP7 GndrMl
GenotypePtk7            -0.402
GenderMale              -0.461  0.185
GenotypePtk7:GenderMale  0.305 -0.658 -0.534

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.17258586 -0.47955226 -0.04337395  0.47750461  2.90866001

Number of Observations: 140
Number of Groups: 52
```

This gives a p-value of 0.11, which is not significant. We can examine the fit of the model as before:

```
ptk7GSMMRes<-residuals(ptk7GSMM)
qqnorm(scale(ptk7GSLMRes))
```

Which gives:



This is a much better fit to a straight line, so we can accept this as an accurate result, and conclude there is no evidence to reject the null hypothesis that there is not a genotypic effect.

Tasks

1: Look at Nodal, Fasted Clinical Chemistry (pipeline 1), Free Fatty acids. Calculate the Rank Sum p-value for females. Try and fit a linear model and look at the residules. What do you think about the fit? Try a mixed model. What do you think about the fit? What conclusions can you draw?

2: Pick a line of interest and analyse the data yourself. Do you agree with the results shown in the web portal?

3: Go to http://www.surveymonkey.com/s/YLKJPC3 and complete the survey, telling us about what you would want the new web interface to do!

Answers

1. Nodal:
    a. Rank Sum P Value = 0.08169
    b. I think the fit is not so good under a linear model
    c. I think the fit is better under a mixed model
    d. The mixed model has a p value of 0.0014, so I can reject the null hypothesis that the genotype has no effect on the Glycerol blood concentration