Statistics 101:

The IMPC Phenotyping Analysis Pipeline

Natasha Karp

nk3@sanger.ac.uk

Last revised: 16th September 2014

This document provides an introduction to the data generated within the IMPC, the statistical analysis used to query the data, and finally the presentation of the raw data and statistical analysis output. Links are provided to examples and tool specific documentation. An online tutorial on how to use the interface is available.

Contents

1. Data generation

Learning objective: This section will explain the how data is collected for a knockout line by various procedures collected together to form a pipeline. We will discuss how the International Mouse Phenotyping Resource of Standardised Screens (IMPReSS) provides information on each procedure.

2. Data QC

Learning objective: This section will explain the QC process that has been implemented on the data. It will highlight the difference between pre and post QC data.

3. Ontology

Learning objective: This section will explain how a standard ontology is used to describe an observed phenotype and where this information is stored for a phenotyping procedure.

4. Version controlled data analysis - PhenStat

Learning objective: This section with explain what PhenStat is and why it is used.

5. Understanding the data types

5.1 Categorical variables

5.2 Continuous variables

5.3 Time course screens

Learning objective: This section will explain how the data can be classed as either continuous or categorical. You will see examples and understand how the each type is visualised. Examples are then given on how time course screens are summarised to produce a continuous variable.

6. Categorical statistical analysis pipeline

6.1 Assessing statistical significance

6.2 Assessing biological significance

6.3 Assemble of baseline control data

6.4 Dataset requirements

Learning objective: This section will explain how phenodeviants are identified in the categorical analysis pipeline. This will include identifying the hypothesis being tested and describing how percentage change is used as a measure of biological effect. You will also learn how the control data has been assembled for the comparison. Finally, this section will cover the analysis requirements and what happens if too few data points are available for statistical interrogation.

7 Continuous statistical analysis pipeline

7.1 Assessing statistical significance

7.2 Classifying the effect

7.3 Assemble of control data

7.4 **Dataset requirements**

7.5 Alternate continuous analysis pipeline

Learning objective: This section will explain how phenodeviants are identified in the continuous analysis pipeline and how the analysis is presented. This will include identifying the hypothesis being tested and describing how percentage change is used as a measure of biological effect. We will discuss how the data is assessed for sexual dimorphism and how the final output gives a classification of the effect observed. You will also learn how the control data has been assembled for the comparison. Finally, this section will cover the analysis requirements, and what happens if too few data points are available or the regression method fails to converge on a solution.

8 Significance threshold

Learning objective: This section will explain what a significance threshold is and why a stringent threshold has been implemented within the IMPC analysis pipeline.

1. Data generation

Systematic broad-based phenotyping is performed by each phenotyping centre collecting data across a variety of screens (e.g. clinical chemistry which monitors various blood characteristics) using standardized procedures collected together to form a pipeline (Figure 1). The phenotyping data is collected on knockout mice (where a gene has been "turned off") and associated control mice. Comparing the data allows assessment into which biological systems are impacted by the gene knockout. Typically pipelines are implemented with control data, from a standard genetic background, collected at regular intervals and a target of seven male and seven female knockout mice being studied per knockout line. In the resulting data processing the mouse is treated as the experimental unit. Further information on the pipelines and associated procedures can be seen at IMPReSS (International Mouse Phenotyping Resource of Standardised Screens). IMPReSS also details for each procedure, the variables collected including both metadata (e.g. instrument used) and variables of interest (e.g. sodium level).

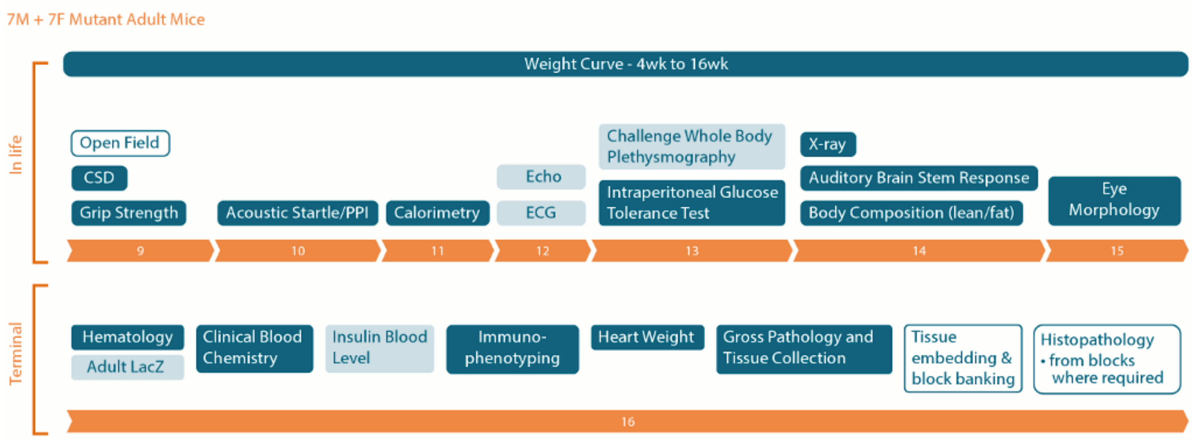

7M + 7F Mutant Adult Mice
Weight Curve - 4wk to 16wk
In life
Open Field
CSD
Grip Strength
Acoustic Startle/PPI
Calorimetry
Echo
ECG
Challenge Whole Body Plethysmography
Intraperitoneal Glucose Tolerance Test
X-ray
Auditory Brain Stem Response
Body Composition (lean/fat)
Eye Morphology
9
10
11
12
13
14
15
Terminal
Hematology
Adult LacZ
Clinical Blood Chemistry
Insulin Blood Level
Immuno-phenotyping
Heart Weight
Gross Pathology and Tissue Collection
Tissue embedding & block banking
Histopathology • from blocks where required
16

Figure 1: Visualisation of the IMPC adult phenotyping pipeline timeline showing the various screens and the age of the mice when the data are collected.

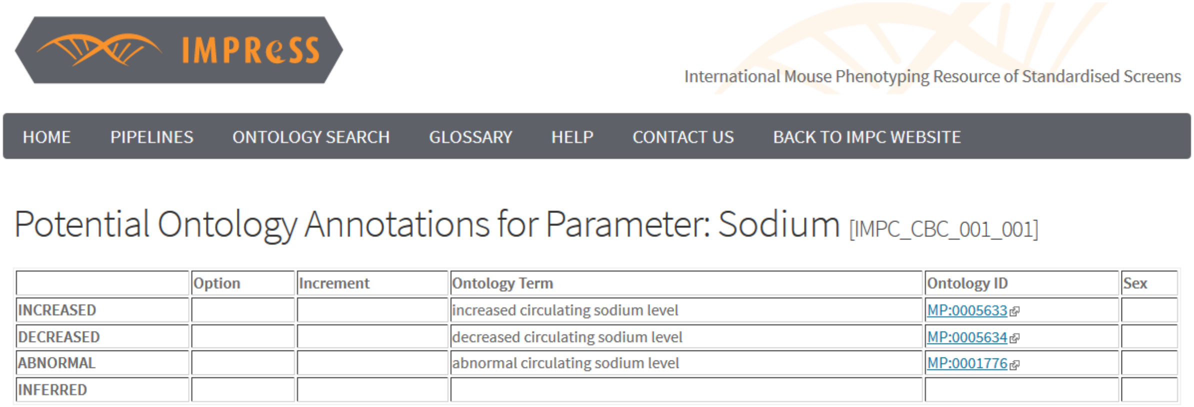
2. QC

A quality control (QC) pipeline investigates the data to ensure accurate data is presented. Concerning data is investigated through collaboration between the inputting data centre and the Data Coordination Centre (DCC). Data can only be QC failed from the dataset if clear technical reasons can be found for a measurement being an outlier. Reasons are provided and this is tracked within the database.

Preliminary statistical analysis is performed at the DCC as soon as enough data is gathered, prior to rigorous quality control checking. This analysis produces results, but due to the preliminary state of the QC checks, the results are considered as pre-QC and not definitive. Once the data has preceded through the QC checks at the DCC, a final definitive statistical test is performed and the MP association made.

3. Ontology

When a significant change in a variable of interest is observed, then the change is described using the mammalian phenotype ontology. The mammalian phenotype terms (MP terms) were development as a community effort to provide standard terms for annotating mammalian phenotypic data. For example, in the plasma chemistry screen the ontology term increased circulating sodium level (MP:0005633) is used when an increase in sodium level is identified (Figure 2). The mammalian phenotype ontology is hosted at Mouse Genomic Informatics (MGI).



Potential Ontology Annotations for Parameter: Sodium [IMPC_CBC_001_001]

	Option	Increment	Ontology Term	Ontology ID	Sex
INCREASED			increased circulating sodium level	MP:0005633	
DECREASED			decreased circulating sodium level	MP:0005634	
ABNORMAL			abnormal circulating sodium level	MP:0001776	
INFERRED					

Figure 2: Example ontology terms stored within IMPReSS. Shown are the ontology terms for the parameter sodium monitored in the plasma chemistry screen. Provided are the Ontology ID and associated term name.

4. Version controlled data analysis - PhenStat

High throughput phenotyping data introduces many challenges in data analysis and is an active area of research. As such the analysis implemented needs to be transparent, reproducible and version controlled. To achieve this goal, we have used the freely available statistical language R to develop a package of statistical tools that can be used interactively on a small scale or in automated application for a large scale use (Figure 3). The resulting package, called PhenStat is freely available for use from Bioconductor, a repository of genomic analysis tools. Further information on PhenStat, the analysis tools available, and how it can be used can be found in its user guide.

As data can be analysed in multiple ways, there are multiple methods implemented to allow data download for further analysis from IMPC and these are discussed within the documentation on accessing the phenotyping data.

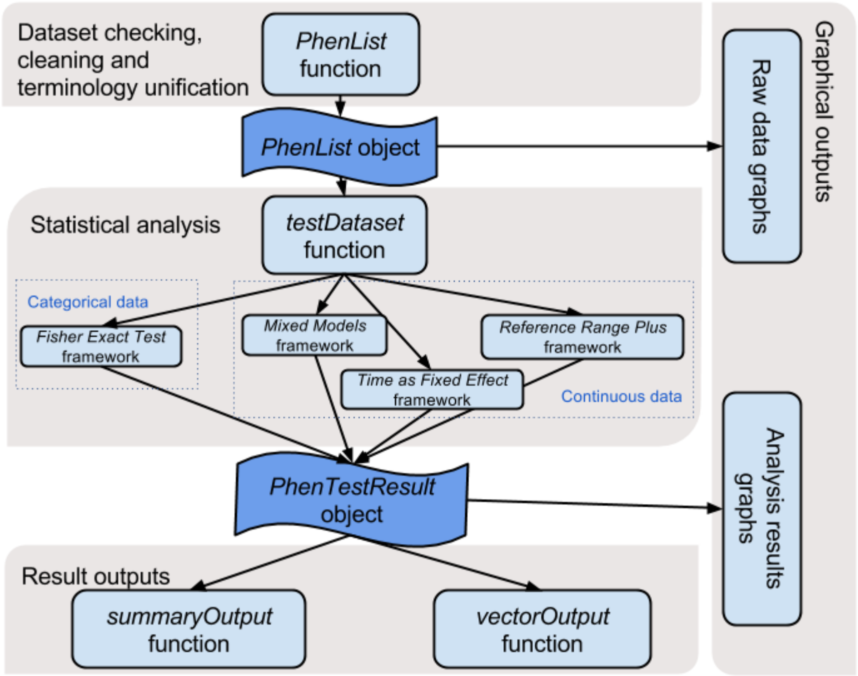


Figure 3: Graphical representation of the three step process within PhenStat providing a standardised easy to use framework for multiple analysis methods producing both graphical and statistical output for identification of phenodeviants.

5. Understanding Continuous and Categorical Data

The variables monitored in the IMPC procedures generate two types of data: categorical and continuous. Each of these data types has their own graphing and analysis pipeline which will be individual discussed in section 6 and 7.

5.1 Categorical variables

A categorical variable is one that can take on one of a limited number of possible values, thus assigning each individual mouse to a particular group or category. Consider the Combined SHIRPA and Dysmorphology screen, which has a goal to assess mice for obvious physical characteristics, behaviours and morphological abnormalities, and therefore as a qualitative assessment returns categorical variables. An example would be the ear variable which as a morphology trait assesses the mice for any structural anomaly of any of the structures involved in the ear or vestibular system. The possible outcomes of the assessment are "as expected" or "not as expected". These are graphed as proportion plots comparing the observed percentages (Figure 4).

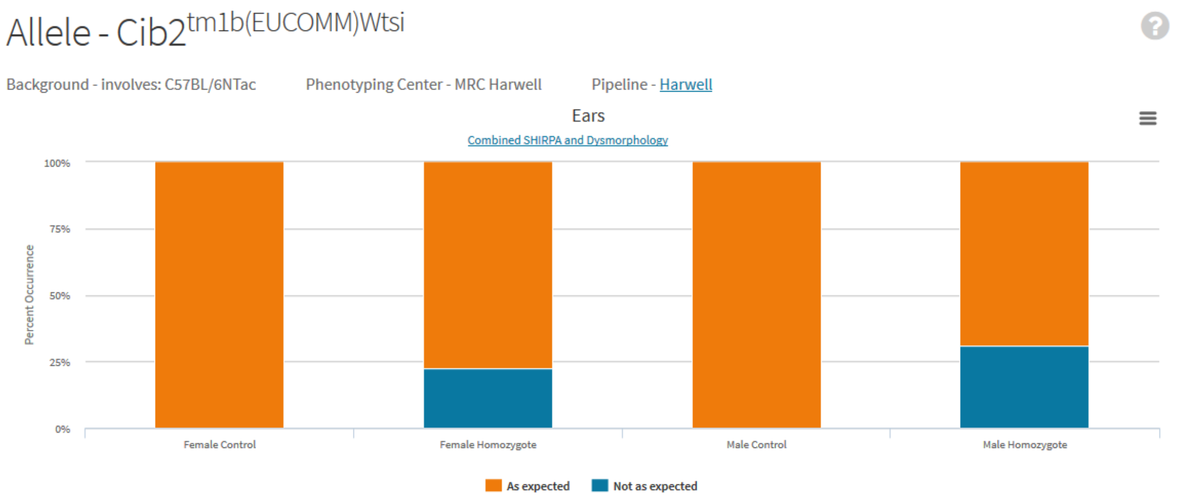


Figure 4: Example categorical plot proportion plot. Shown is the data for the assessment of ear morphology for *Cib2*. Here a higher proportion of "not as expected" is observed in the male and female homozygote compared to the control data.

5.2 Continuous variables

A continuous variable is a variable that can take on any value between its minimum value and its maximum value. Consider the Clinical Blood Chemistry screen, which has a goal to determine biochemical parameters in plasma including enzymatic activity, specific substrates and electrolytes, and returns continuous variables. An example would be the trait; glucose which is a measure of the circulating glucose level in the plasma. These continuous variables are with boxplots and scatterplots allowing you to comparing the distribution for each genotype tested for each sex (Figure 5).

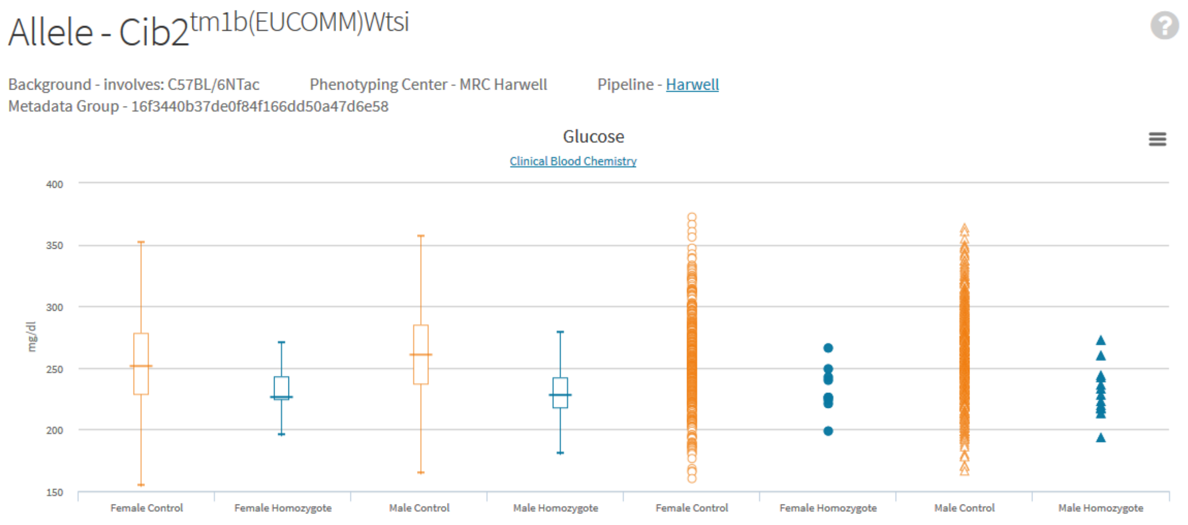
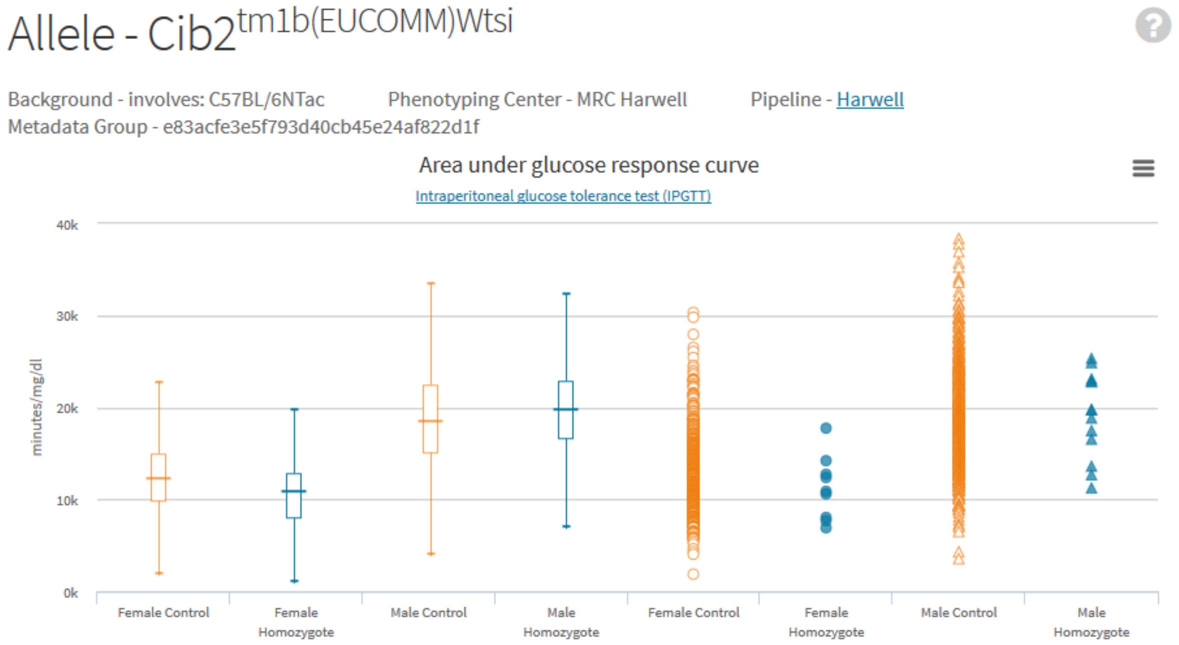


Figure 5: Example visualisation of continuous data. Shown is the data for the assessment of circulating glucose for *Cib2*. Here statistically significantly lower levels of glucose were observed in both the male and female homozygote mice compared to the control mice. The boxplot is a five point summary of the data. The central tendency shows the median in the centre where the median is the middle number if the data is ordered. The lower edge of the box is the first quartile (value seen at the 25 percentile) and the upper edge of the box is the third quartile (value seen at the 75 percentile). The whiskers show the range reached by the

addition/subtraction of 1.5 times the inter-quartile range (IQR) to the quartiles, where the IQR is the difference between the third and first quartile.

5.3 Time Course screens

A number of the screens are time course studies monitoring variables of interest with time. For example, the intraperitoneal glucose tolerance test (IPGTT) measures the clearance of an intraperitoneally injected glucose load from the body to detect disturbances in glucose metabolism that can be linked to human conditions such as diabetes or metabolic syndrome. As such, the glucose concentration is measured five times for a single mouse from the beginning of the study to 2 hours post glucose injection.



Allele - Cib2$^{tm1b(EUCOMM)Wtsi}$

Background - involves: C57BL/6NTac Phenotyping Center - MRC Harwell Pipeline - Harwell
Metadata Group - e83acfe3e5f793d40cb45e24af822d1f

Figure 6: Example plot of a summary measure comparison for the intraperitoneal glucose tolerance test. Shown is the summary measure "area under the glucose response curve" arising from the monitoring of the glucose level in the *Cib2* knockout study. In this dataset, the genotype was not found to have a statistically significant impact on the data (p=0.1598).

6. Categorical statistical analysis pipeline

6.1 Assessing statistical significance

A Fisher Exact Test is used as a statistical test to compare the observed proportions between the wildtype and the knockout data for a sex of mice to test the hypothesis that the proportions are the same. This statistical test is ideal for datasets which are monitoring rare events with small sample sizes. The test will return a *p*-value and if this *p* value is below the significance threshold then we reject the hypothesis that the proportions are the same providing indirect evidence that the proportions are different between the wildtype and knockout dataset.

Consider the abnormal ear morphology seen in the *Cib2* example (Table 1) where the abnormal phenotype ("not as expected") was not observed in the large control dataset whilst within the male homozygous mice we observed 4 mice out of 13 as "not as expected". This change in proportion was found to statistically significant returning a *p* value from the Fisher Exact Test of 4.553E-8.

	As expected		**Not as expected**	
	Number	Percentage	Number	Percentage
Controls	772	100	0	0
Homozgyote	9	69	4	31

Table 1: The observed counts and percentage seen for each category for the control and knockout animals. Shown is the data for the assessment of ear morphology for *Cib2* male mice. Here a higher proportion of "not as expected" is observed in the male homozygote compared to the control data.

6.2 Assessing biological significance

To give a measure of biological significance, a maximum effect size is reported, which indicates percentage penetrance of the abnormality within the knockout mice data compared to the control data. The maximum effect size is the maximum percentage change seen for an observation type (Table 2 for example calculation).

	As expected	**Not as expected**
	Percentage	Percentage
Controls	100	0
Homozgyote	69	31
Change	|100-69|=31	|0-31|=31

Table 2: Calculating the maximum effect size as a measure of biological significance. For each observation type, the absolute percentage change is calculated from the difference in percentage observed in the knockout mice against the control mice. The maximum value observed is returned as the maximum effect size, 31% in this example.

6.3 Assemble of baseline control data

To increase sensitivity of the test, all control data for a variable is combined into a baseline dataset, to give greater confidence in the abnormality rate in the wildtype population. Control data is only combined if it arose from the same institute, same genetic background, same pipeline and the same procedure id. Certain screens can have additional metadata rules that are also used in data assemble. Consider the Combined SHIRPA and Dysmorphology Screen, there is one metadata variable, the size of squares in arena, that can influence data assemble as shown as being "required for analysis" within IMPReSS. This approach of combining control data assumes that temporal and litter are minimal sources of variation and takes no account of these variables in the analysis.

6.4 Dataset requirements

The categorical analysis pipeline requires at least one data points per genotype sex group.

7. Continuous statistical analysis pipeline

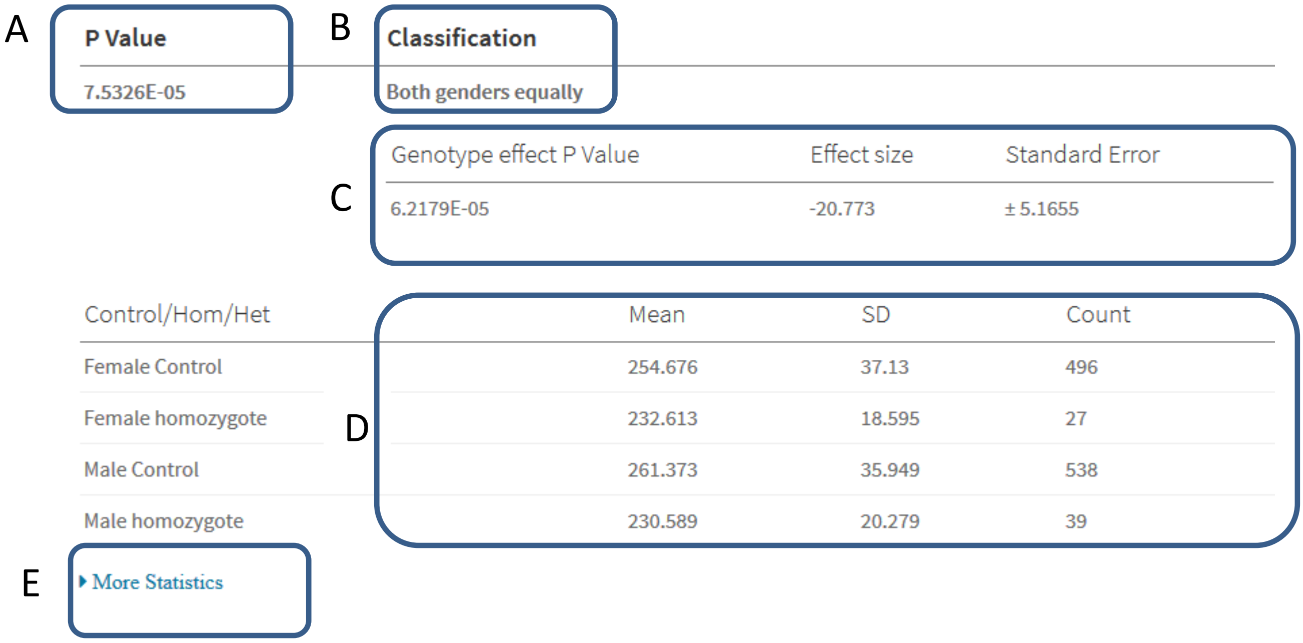
7.1 Assessing statistical significance

The continuous analysis pipeline uses regression analysis to interrogate the data. The pipeline starts with equation 1 which includes a number of factors that could explain the variation in the variable of interest.

depVariable ~ Genotype+Sex+Genotype*Sex+(1|Batch) [Equation 1]

To fit a final model that is most appropriate to the data, a model optimisation process is followed before assessing for a genotype effect. The model optimisation focuses first on global model issues: such as the type of model and whether the variance (variability) is consistent. There are two types of models considered: linear model or a mixed model. The mixed model is used when temporal variation is found to be a significant source of variation in the variable of interest. After assessing the general model characteristics, the optimisation tests whether sex should be included and whether the genotype effect shows evidence of sexual dimorphism and should be assessed for each sex separately. Further details on the model optimisation can be found in Karp *et al* (PLOS One 2012) and the PhenStat User Guide (section 3.2).

Following model optimisation, the contribution of the genotype effect to differences in the data from the knockout mice compared to the controls is assessed by a likelihood ratio test comparing a treatment model which includes the genotype component against a null model where the genotype element is absent. This tests the hypothesis that the genotype is not significant in explaining variation. When the *p* value is below the significance threshold then we have statistical evidence that the genotype is a significant source of variation in the variable of interest. The genotype effect is then estimated from the final fitted model and returned along with an error measure and a *p* value for a contribution test within the final model (see Figure 7). The assignment of abnormality for MP term assignment is driven by the global test of genotype being significant to the phenotype. Therefore this is the first measure to assess and then if significant you look at the estimate and associated errors of the genotype effect.



P Value	Classification
7.5326E-05	Both genders equally

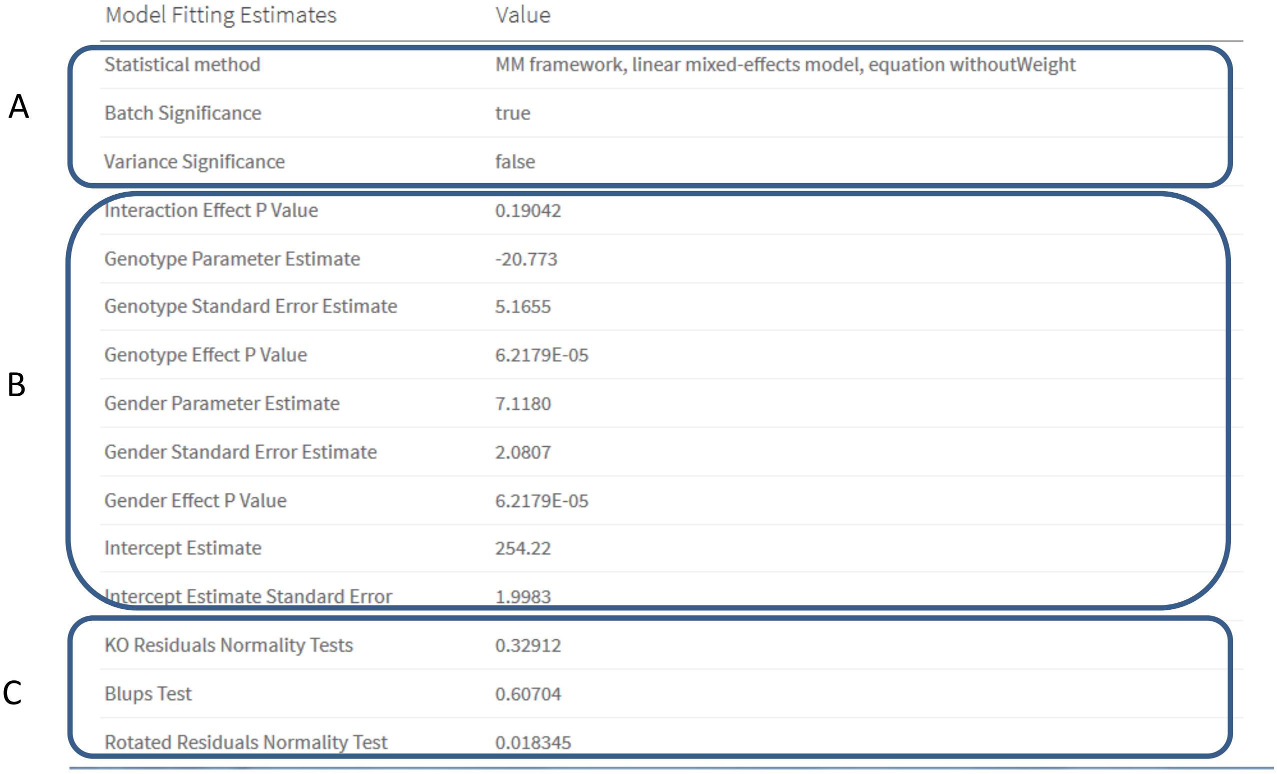
Genotype effect P Value	Effect size	Standard Error
6.2179E-05	-20.773	± 5.1655

Control/Hom/Het	Mean	SD	Count
Female Control	254.676	37.13	496
Female homozygote	232.613	18.595	27
Male Control	261.373	35.949	538
Male homozygote	230.589	20.279	39

Figure 7: The statistical summary output for the *Cib2* circulating glucose measure.

- A: The *p* value from the global test of genotype. Here it was found to be highly significant indicating that a statistically significant genotype difference was present in the dataset.
- B: The genotype effect was classified as affecting both sexes equally. This means there was no evidence of sexual dimorphism.
- C: The estimated values for the genotype effect in the final fitted model are reported. In this example we can see that genotype within a model was found to be highly significant (*p* value: 6.2e-5) and was estimated at the circulating glucose being 20.73±5.1 mg/dl lower in the knockout mice.
- D: The summary values (mean, standard deviation (SD) and count (number of measurements)) are provided for each genotype and sex group. In this example we can see that the circulating glucose is lower in both the female and male homozygous mice compared to the control mice.
- E: More Statistics: this is a link that will access further information from the model fitting process (see Figure 8).

PhenStat outputs additional information (Figure 8). For example it includes information on the final model fitted after the optimisation process. It also includes information on the significant variables included in the model in how they affect the variable of interest. Finally it includes tests on the final quality of the model as model diagnostics. Further information on these can be found in PhenStat User Guide (sections 3.2.4 and 5.2).



Model Fitting Estimates	Value
Statistical method	MM framework, linear mixed-effects model, equation withoutWeight
Batch Significance	true
Variance Significance	false
Interaction Effect P Value	0.19042
Genotype Parameter Estimate	-20.773
Genotype Standard Error Estimate	5.1655
Genotype Effect P Value	6.2179E-05
Gender Parameter Estimate	7.1180
Gender Standard Error Estimate	2.0807
Gender Effect P Value	6.2179E-05
Intercept Estimate	254.22
Intercept Estimate Standard Error	1.9983
KO Residuals Normality Tests	0.32912
Blups Test	0.60704
Rotated Residuals Normality Test	0.018345

Figure 8: Further information found under the 'More Statistics' Link. Shown is the output for *Cib2* circulating glucose level.

- A: These provide information on the final fitted model characteristics following model optimisation. The Statistical method, tells us that the output arises from the use of the mixed model (MM) framework within PhenStat. The reference to "Equation without weight" tells us that the starting model did not include a covariate for body size (i.e. Equation 1 was used). The second string "linear-mixed effect model" gives the final model type used in the analysis. This is confirmed by the output where it states batch was found to be significant source of variation and therefore a mixed regression model was used. As variance was found not to be significant, then a homogenous model was used.
- B: In this section, the estimates are shown for any variable thought to have an impact on the variable of interest. In this example, we can see that sex was included in the model and tells us that being a male mouse was found to increase the measurement of the variable of interest by 7.12±2.1mg/dl.
- C: Model diagnostics: In this section, outputs from automatic tests of model diagnostics are reported. A model is a good fit, when the residuals (difference between estimated in the model and the reality) are normally distributed, and KO Residuals Normality Tests the hypothesis that the residuals are normally distributed in the knockout dataset. Therefore a low *p* value would raise concerns that the model was potentially not fitting the data well. The remaining two tests, consider the assumption that batch is a normally distributed variable and again test this distribution. In this case one of the two is significant and could be explored further by the graphical model diagnostic tools available within PhenStat.

7.2 Classifying the effect

The rich output of the model fitting process obtained in a regression analysis enables a classification of the outcome in addition to the identification of a phenodeviant (Figure 9). During the model optimisation process, if there is statistical evidence of sexual dimorphism, then the genotype effect is estimated for each sex separately. The global test of genotype impact tells us in the presence of sexual dimorphism that somewhere the genotype is significant in explaining the variation in the variable of interest. By looking then at the model estimates from the finally fitted model, we can assess which of the sexes was contributing to this genotype effect. Potentially it could be both, or one. Figure 9 details a decision tree where the effect is classified and this gives a variety of possible tag – for example "Males only" or "Different sizes, males greater". The classification tag is estimated regardless of whether the genotype effect was found to be globally significant, and therefore it is important that you first assess the global *p* value for significance and then look at the output to assess how this arose.

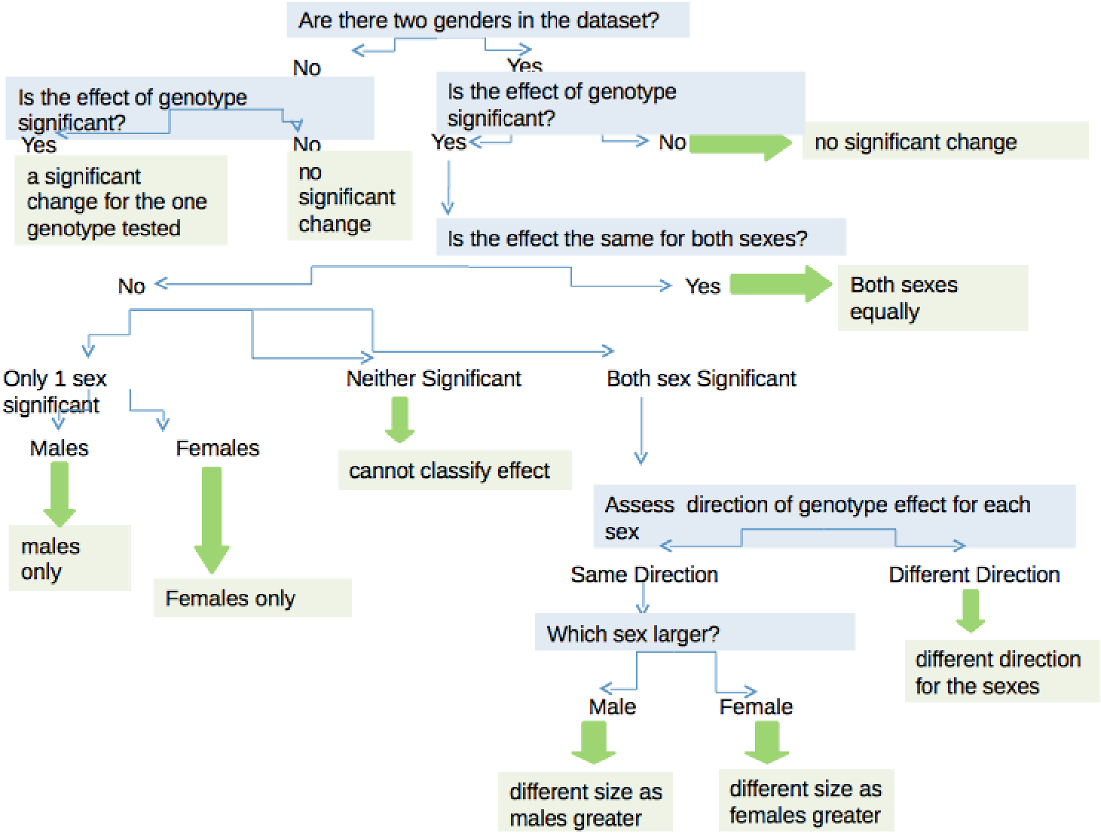
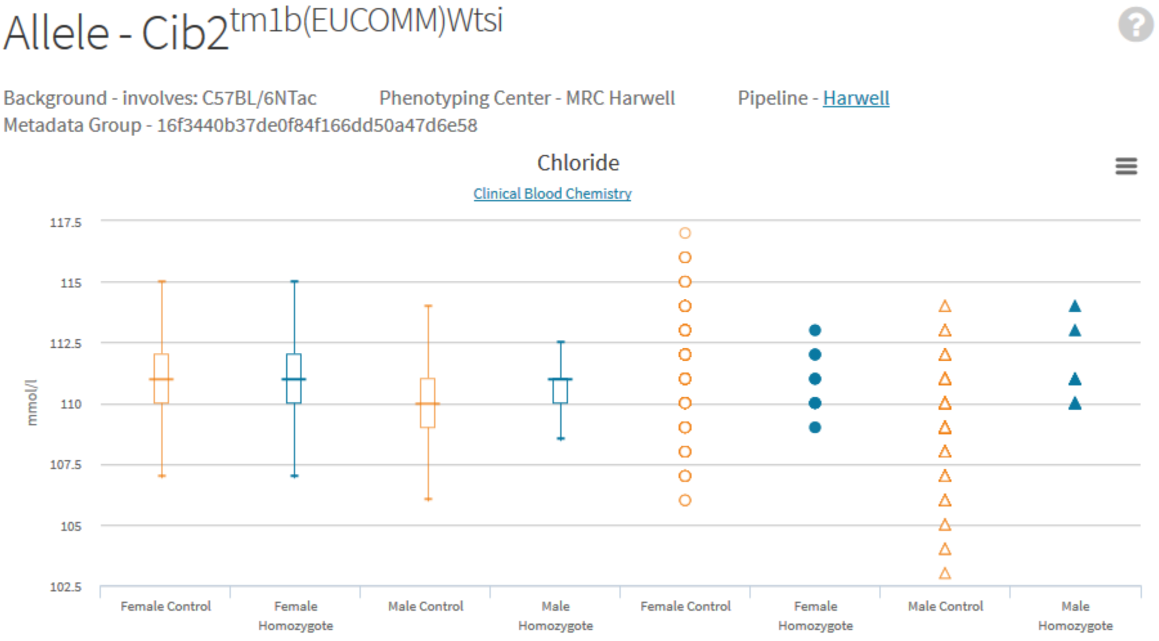

Are there two genders in the dataset?
No
Yes
Is the effect of genotype significant?
Yes
No
a significant change for the one genotype tested
no significant change
Is the effect of genotype significant?
Yes
No
no significant change
Is the effect the same for both sexes?
No
Yes
Both sexes equally
Only 1 sex significant
Neither Significant
Both sex Significant
Males
Females
cannot classify effect
males only
Females only
Assess direction of genotype effect for each sex
Same Direction
Different Direction
Which sex larger?
different direction for the sexes
Male
Female
different size as males greater
different size as females greater

Figure 9: Determining the classification tag from the rich regression output to assess the impact of the genotype effect across the sexes tested.

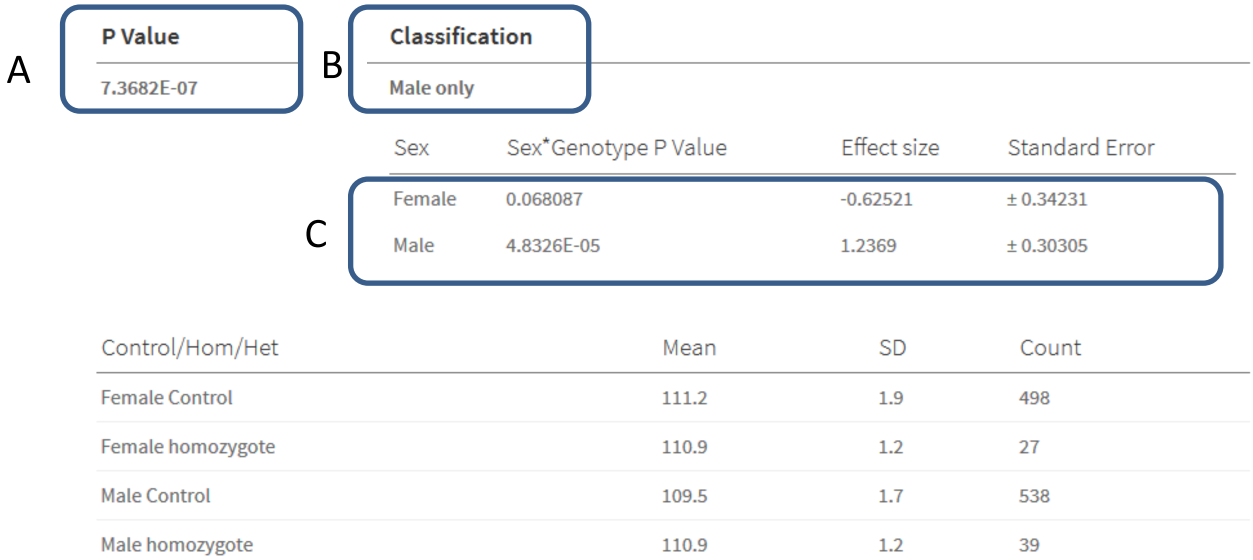
Let's consider an example from *Cib2* where the model fitting process found evidence of sexual dimorphism (Figure 10 and 11).



Allele - Cib2$^{tm1b(EUCOMM)Wtsi}$

Background - involves: C57BL/6NTac Phenotyping Center - MRC Harwell Pipeline - Harwell
Metadata Group - 16f3440b37de0f84f166dd50a47d6e58

Figure 10: Visualisation of the circulating chloride measurements for the *Cib2* knockout line. Looking at control data we can see that there is a sex effect with the male mice having a lower chloride reading. In the knockout mice the female mice have comparable readings to the control group whilst the male mice have higher readings.



P Value
7.3682E-07

Classification
Male only

Sex	Sex*Genotype P Value	Effect size	Standard Error
Female	0.068087	-0.62521	± 0.34231
Male	4.8326E-05	1.2369	± 0.30305

Control/Hom/Het	Mean	SD	Count
Female Control	111.2	1.9	498
Female homozygote	110.9	1.2	27
Male Control	109.5	1.7	538
Male homozygote	110.9	1.2	39

Figure 11: The statistical summary output for the *Cib2* circulating chloride measure which demonstrates a sexual dimorphic call. As the model optimisation process found evidence of sexual dimorphism, then the genotype effect was estimated for each sex separately. The global test of genotype contribution was highly significant (A) and it was classed as Male only effect (B). Looking at the final fitted model estimates, we can see that within the model, the genotype effect for male mice was found to be highly significant but not the female genotype effect (C).

7.3 Assemble of control data

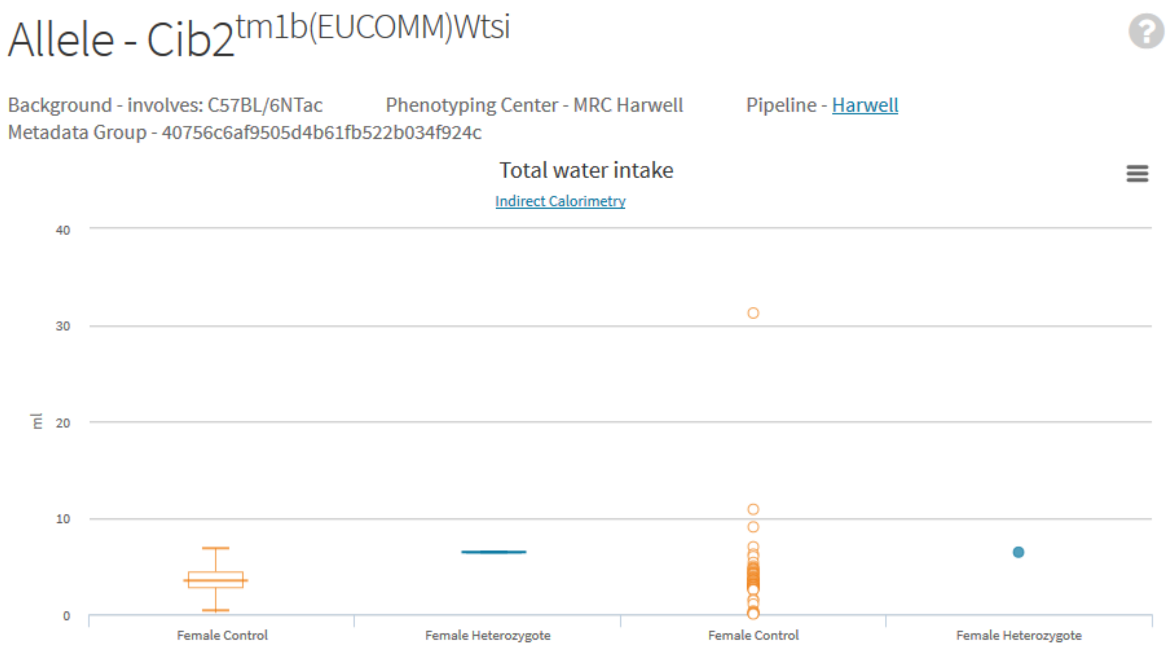
The mixed model analysis pipeline is treating batch as a variable that is normally distributed and that it adds to the variance of the variable of interest. This modelling process behaves well when the knockout data is split into multiple batches (Karp *et al* PLOS One 2014). When the knockout data is in few batches then the method can give false positives and have low power. To manage this, the knockout data is assessed, and if the data for a line arises from one batch with concurrent controls then only the concurrent controls collected on the same genetic background, for an institute and pipeline are submitted with the knockout data. As there is no variation in the batch variable, PhenStat will automatically revert to a linear regression and will continue the model optimisation to build the final model (as discussed in section 7.1). If there are multiple batches or no concurrent controls, then the control data is assembled by collecting all control data that has been collected for a pipeline for an institute on that genetic background.

7.4 Dataset requirements

The continuous analysis pipeline requires four data points per genotype sex group, except for the data analysis of the Auditory Brain Stem Response screen where only 2 data points per genotype sex group are required. If there are too few readings then the data is graphed and the analysis defaults to the alternate continuous analysis pipeline.

7.5 Alternate continuous analysis pipeline

If the mixed model methodology fails to return a model fit, for example for failing to converge on a solution or there are two few data points for the analysis to run, then the analysis defaults to an alternate analysis pipeline (Figure 12 and 13). In this pipeline the controls and knockouts are analysed for each sex independently using a Wilcoxon rank sum test. This statistical test compares the rank distribution between the two groups and is testing the null hypothesis of equivalence in distribution. The effect size is calculated as the difference in median (the mid-point value of the dataset) between the two datasets.



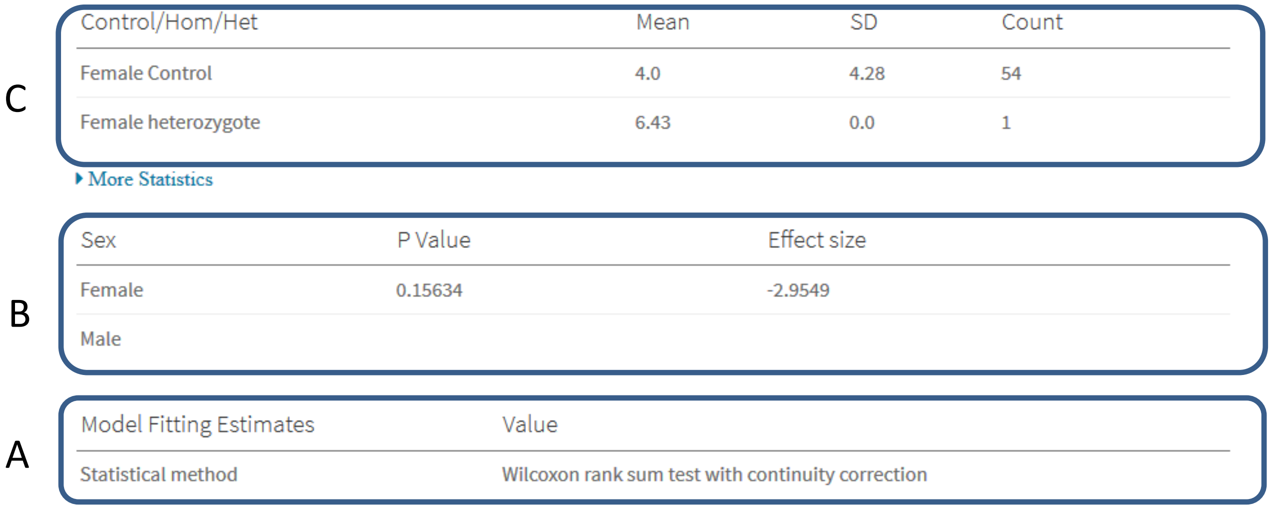
Allele - Cib2$^{tm1b(EUCOMM)Wtsi}$

Background - involves: C57BL/6NTac   Phenotyping Center - MRC Harwell   Pipeline - Harwell
Metadata Group - 40756c6af9505d4b61fb522b034f924c

Total water intake

Indirect Calorimetry

Figure 12: Visualisation of the total water intake variable measured during the Calorimetry Screen for the *Cib2* knockout line. Looking at the data we can see one point for the Female heterozgote which could either indicate there is only one data point or there is very low variation in the readings obtained.



Control/Hom/Het	Mean	SD	Count
Female Control	4.0	4.28	54
Female heterozygote	6.43	0.0	1

▸ More Statistics

Sex	P Value	Effect size
Female	0.15634	-2.9549
Male		

Model Fitting Estimates	Value
Statistical method	Wilcoxon rank sum test with continuity correction

Figure 13: The statistical summary output for the *Cib2* total water intake measure. The output (A) confirms that statistical method applied, the statistical output shown (B) and the summary measures (C). In the statistical output (B), we have a non-significant *p* value (0.156) indicating there is no statistical evidence of a difference between the heterozygous and control animals for the female mice. In the summary measures we can confirm that there was only one mouse in the female heterozgote group as the count equals 1.

<u>8. Significance threshold</u>

The *p* value calculated from a statistical test is the probability of getting the results you did (or more extreme results) given that the null hypothesis is true. For example in the Fisher Exact Test, it is the probability that you will see the difference or larger differences when the proportions are the same between the groups tested. Therefore the *p* value is a measure of the strength of evidence for the null hypothesis and varies between 0 and 1.

- A small *p* value indicates strong evidence against the null hypothesis so you accept the alternate. In our pipeline this would result in a variable for a knockout line being classed as abnormal and classed as a phenodeviant.
- A large *p* value indicates weak evidence against the null hypothesis, so do not reject the null hypothesis. In this case, we would conclude there is no evidence of phenotypic change associated with the genotype change.

An artificial cut off point is chosen, called the significance threshold, and the result is called statistically significant if the *p* value is less than the threshold. Consider the commonly used significance threshold of 0.05, the chance of sampling leading to the difference is low and controls the false positive rate to 5%.

However within IMPC, we are conducting not just one test but many thousands of tests. This introduces the multiple testing problem, where false positives can accumulate simply because so many statistical tests are conducted. Consider the case, where you have 20 hypotheses to test and you used a significance threshold of 0.05. We can calculate the probability of observing at least one significant result just due to chance?

$$\begin{aligned} \text{P(at least one significant result)} &= \text{1-P(no significant results)} \\ &= 1-(1-0.05)^{20} \\ &= 0.64 \end{aligned}$$

So with 20 tests we have a 64% chance of observing at least one significant result even if all the tests are not significant.

Therefore, within the IMPC statistical pipeline, we have pre-set a conservative significance threshold (0.0001) that is used to identify phenodeviants and result in an MP term being associated with a knockout line.