Mouse Pipelines data – structure and analysis

Data Generation

Mice carrying mutations were generated and established.

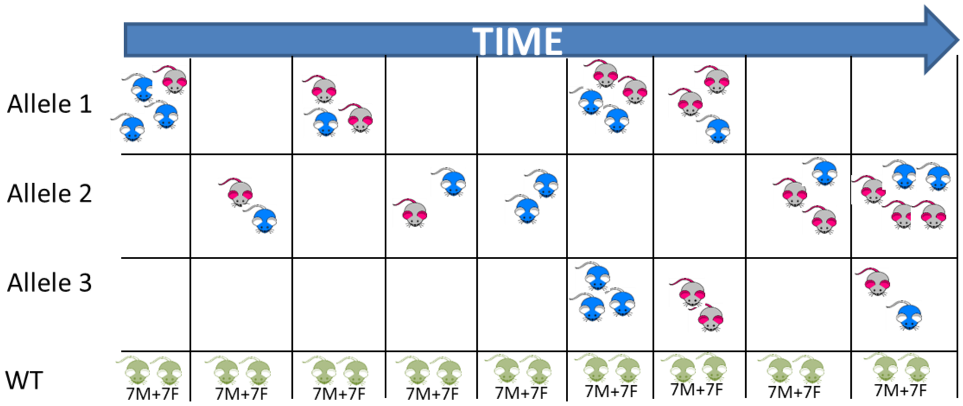
Viability was assessed at postnatal day 14 (P14) by genotyping offspring of heterozygous crosses.

Alleles classed as lethal or subviable (<13% of the expected number of homozygous pups) were further assessed in various developmental screens.

Adult animals from these lines were screened by way of a high-throughput unbiased phenotyping pipeline, utilising a variety of standardised *in vivo* and *ex vivo* assays, designed to cover a broad range of biomedical areas. Homozygous (or hemizygous) mice were used where these were viable, with heterozygotes used in cases of lethality or subviability.

Further details of the assays performed as part of the developmental and adult phenotyping screens can be seen in the Pipelines document.

The majority of tests examined seven male and seven female animals. These were tested in small batches so data for each batch was gathered on different days. These batches were termed 'cohorts' and all animals within one cohort had a date of birth within three days of each other. This design was intentional to counteract the impact that batch (defined here as those readings collected on a particular day) can have on variation in phenotyping variables.
Control or wild-type (WT) animals were processed through the same pipeline each week, to encompass week-to-week variation, and allow the accumulation of a reference range.

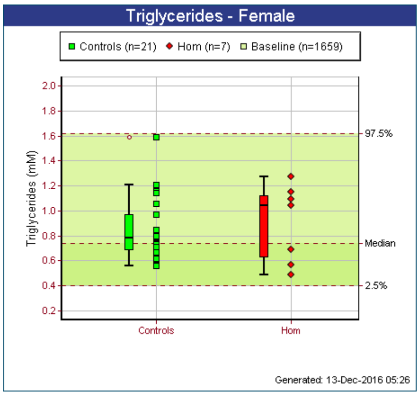


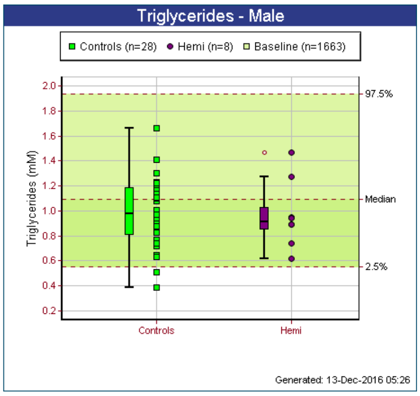
Complementing this functional analysis, a subset of these lines underwent gene expression profiling in adults and embryos (E14.5) using the *lacZ* reporter gene that was introduced as part of the targeting event.

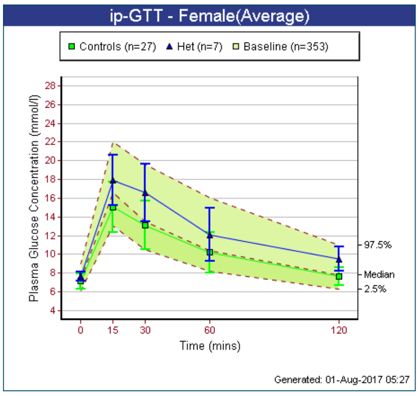
How the data was internally visualised and assessed

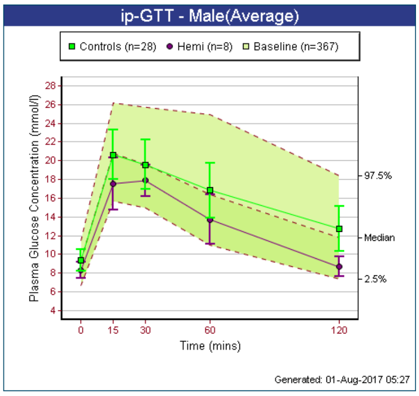
Internal graphs:

Continuous/time course data:









Blue/red/purple data points = mutant data

- All genotype confirmed data from mutant samples. This can be heterozygous (blue), homozygous (red) or hemizygous (purple), depending on the allele

Green data points = local controls

- All wildtype data run during the same week as the mutant data points

Green background = baseline = reference range

- Presented as the median and 95% confidence interval (from 2.5% to 97.5%)
- Created from all wildtype data from the same genetic background, pipeline, protocol (plus any metadata splits such as anaesthetic or analyser)
- Note that baselines were frozen at the point in time that the mutant and local controls were genotype confirmed, so will not include any animals processed beyond this date

Box plot

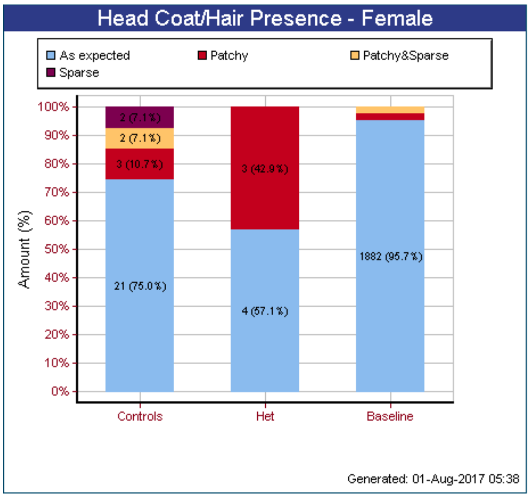
- The box represents the 25th and 75th percentile (with the 50th percentile as a line in the middle – the median)
- The whiskers are the lowest and highest data point still within 1.5x the IQR (inter quartile range)
- Anything outside of 1.5x the IQR is classed as an outlier (and has a little circle symbol)

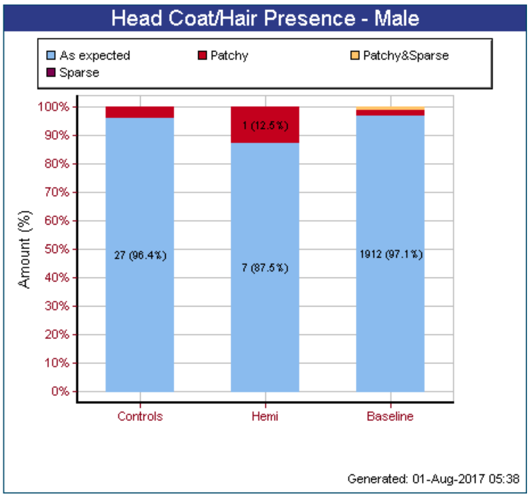
Categorical data:

Presented as bar charts with separate bars for:

- Controls - All wildtype data run during the same week as the mutant data points
- Het/Hom/Hemi - All genotype confirmed data from mutant samples
- Baseline - All wildtype data from the same genetic background, pipeline, protocol (plus any metadata splits). Note that baselines were frozen at the point in time that the mutant and local controls were genotype confirmed, so will not include any animals processed beyond this date.

Colours for each category within the bar are defined in the legends above the plots.



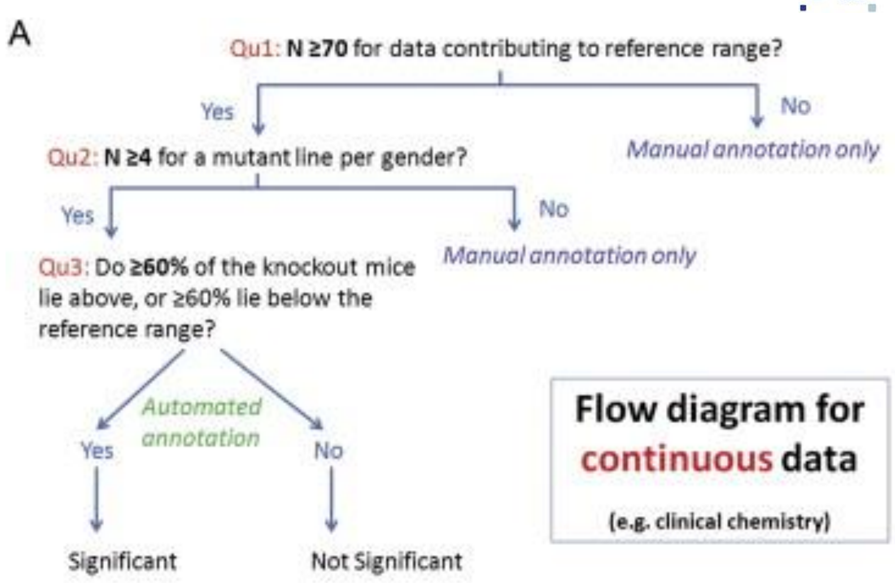


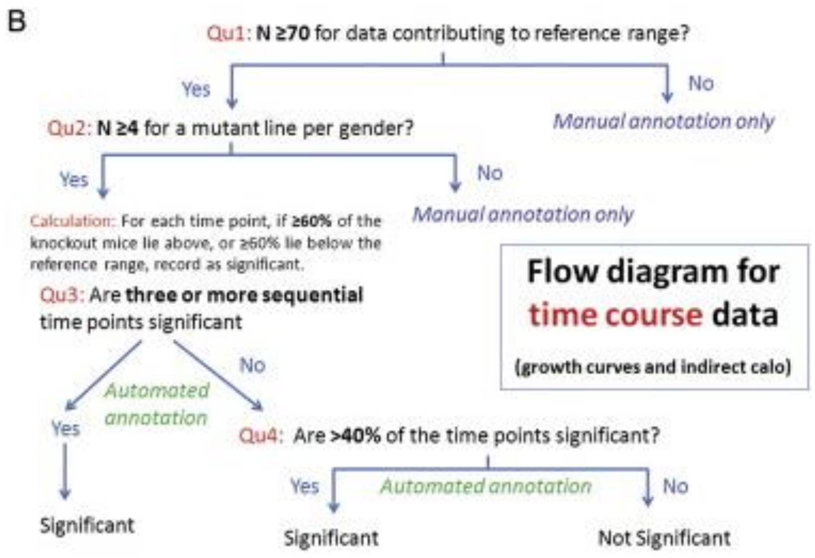
Internal calls of significance

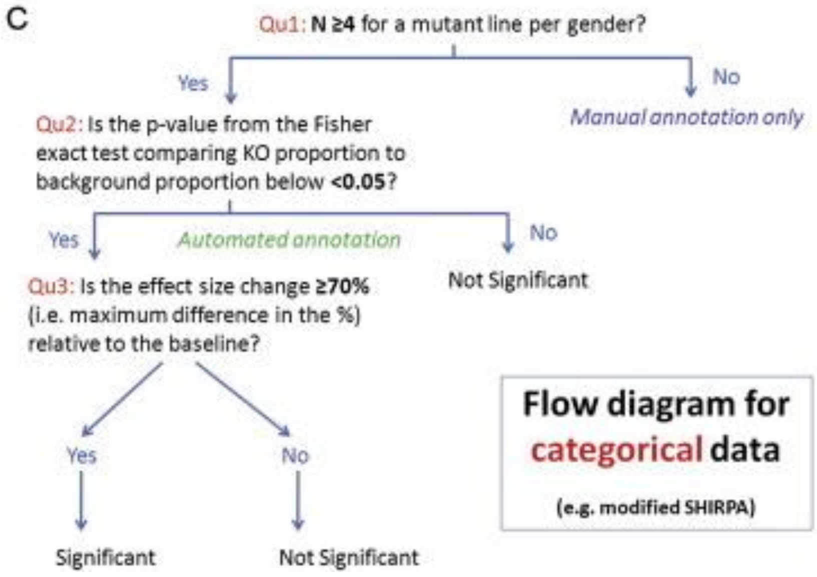
Mutant data were compared to the relevant baseline, composed of all animals of the same sex, age and genetic background that had been screened using that particular experimental protocol. Phenotypes were determined in two ways.

Firstly, a standardised set of rules was applied to the data, which allowed phenodeviants to be automatically highlighted by the software, as described in the flow charts below for the three main data types.

The philosophy of this approach was to discover robust phenotypes with large effect sizes, and resulted in conservative calls that minimised false positives.


A
Qu1: N ≥70 for data contributing to reference range?
Yes
No
Manual annotation only
Qu2: N ≥4 for a mutant line per gender?
Yes
No
Manual annotation only
Qu3: Do ≥60% of the knockout mice lie above, or ≥60% lie below the reference range?
Automated annotation
Yes
No
Significant
Not Significant
Flow diagram for continuous data
(e.g. clinical chemistry)


B
Qu1: N ≥70 for data contributing to reference range?
Yes
No
Manual annotation only
Qu2: N ≥4 for a mutant line per gender?
Yes
No
Manual annotation only
Calculation: For each time point, if ≥60% of the knockout mice lie above, or ≥60% lie below the reference range, record as significant.
Qu3: Are three or more sequential time points significant
No
Automated annotation
Yes
Significant
Qu4: Are >40% of the time points significant?
Yes
Automated annotation
No
Significant
Not Significant
Flow diagram for time course data
(growth curves and indirect calo)

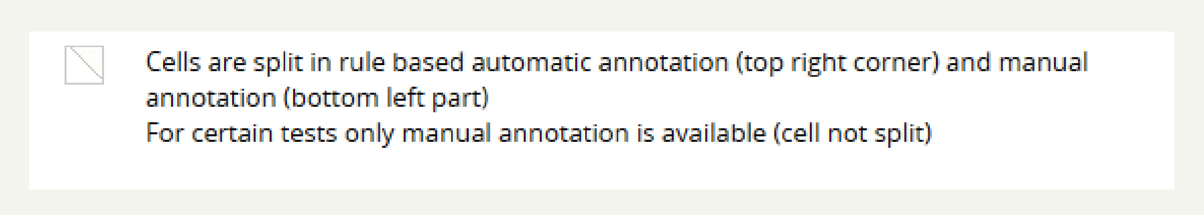

C
Qu1: N ≥4 for a mutant line per gender?
Yes
No
Manual annotation only
Qu2: Is the p-value from the Fisher exact test comparing KO proportion to background proportion below <0.05?
Yes
Automated annotation
No
Not Significant
Qu3: Is the effect size change ≥70% (i.e. maximum difference in the %) relative to the baseline?
Yes
No
Significant
Not Significant
Flow diagram for categorical data
(e.g. modified SHIRPA)

A second review of the data was performed by an area expert who would either agree with the automated call made or override that call based on their knowledge or appraisal of the pipeline data as a whole for that line.

These calls of significance can be viewed in the heat map directory.

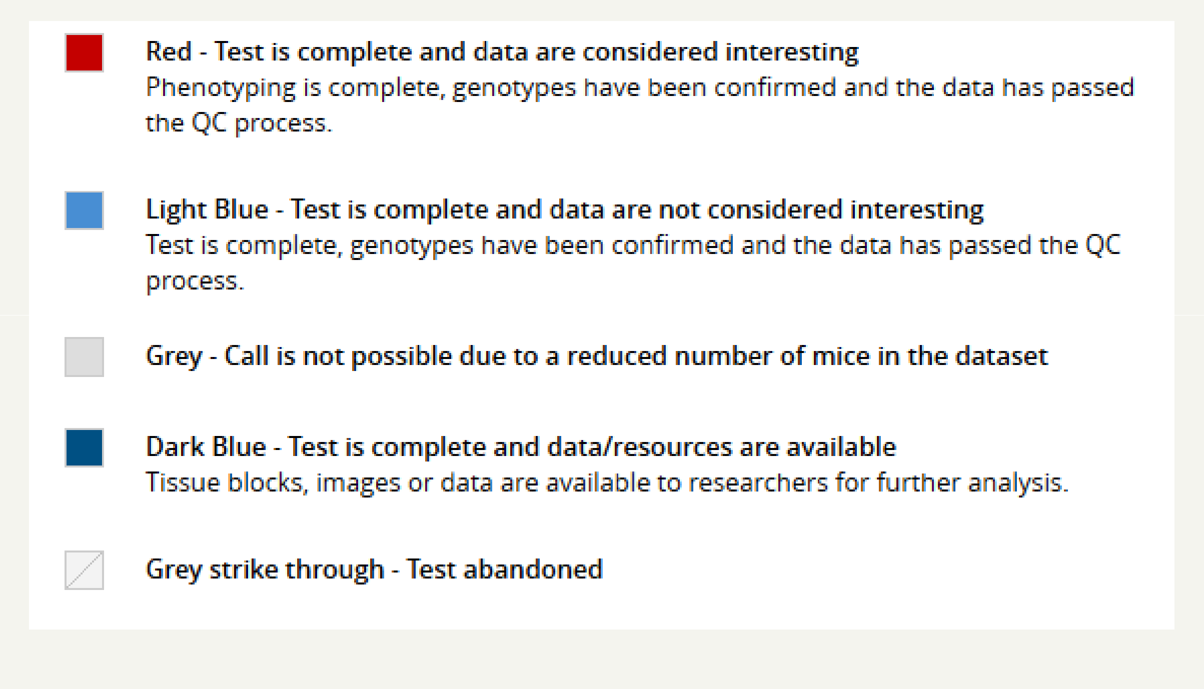
The downloadable Excel version of the heat map only contains the manual call made by the area expert.

If you have access to the internal database version of the heat map, this displays both the automated and manual annotation as described below.



⧅ Cells are split in rule based automatic annotation (top right corner) and manual annotation (bottom left part)
For certain tests only manual annotation is available (cell not split)

Both versions of the heatmap use the colour coding described in the legend below.



🟥 Red - Test is complete and data are considered interesting
Phenotyping is complete, genotypes have been confirmed and the data has passed the QC process.

🟦 Light Blue - Test is complete and data are not considered interesting
Test is complete, genotypes have been confirmed and the data has passed the QC process.

⬜ Grey - Call is not possible due to a reduced number of mice in the dataset

🟦 Dark Blue - Test is complete and data/resources are available
Tissue blocks, images or data are available to researchers for further analysis.

⧄ Grey strike through - Test abandoned

How to use the data for your own analysis and factors to be aware of

How to assemble data

Go to the Populations directory. Use the colonies_to_pipelines_mapping document to determine which pipeline(s) your gene of interest was processed through. Find your gene of interest within the correct pipeline folder and download the file.

This will give you the list of mutant animals tested plus the local controls (as defined above – the controls processed during the same weeks as the mutants). The unique identifiers for each animal are the Mouse Name or Mouse Barcode.

Note that there may be some instances where the data is represented as 0/1. In these cases, 0 converts to No, and 1 to Yes.

The raw data for each test can be accessed via the Pipelines directory. Select the pipeline stated in the filename for your population then go to your test of interest.

The index files in the Index directory will allow you to determine which file(s) contain your measurements of interest.

If you would also like to assemble a baseline, the population file should specify a baseline ID e.g. the ID in the following file *acbd5_tm1b(eucomm)wtsi_(ID-12488_BaselineID-94795)_PMAJ_(57-ROWS).csv* is 94795. These baseline files can be found within the Baselines directory.

Examples of constructing datasets are detailed in the Examples file in the Documentation subdirectory.

Statistics

As the data was not generated in a traditional manner of all treated and control animals being processed on the same day, it does not meet the assumptions for analysis by such statistical tests as Student's t-test and ANOVA.

A large proportion of Sanger Mouse Pipelines data is available to view on the IMPC website (www.mousephenotype.org). The analysis implemented here uses the statistical methods available in the R package PhenStat. The package was developed for data generated by high throughput phenotyping pipelines to help specifically address the issue of temporal variation.

We recommend the use of PhenStat for analysis of Mouse Pipelines data and the main basic frameworks implemented within it are described below.

Once data is loaded, the software determines the most appropriate analysis method:

- Categorical data:
  - Fisher Exact test – standard approach for comparing proportion data. Most categorical abnormalities are rare events, important for the sensitivity to detect differences using a small number of mutant animals when compared to controls. Batch is not considered in this analysis method.

- Continuous data:
    - Reference Range Plus – a conservative method that categorises values as low, normal or high, based on the natural variation in the values of the control animals. The proportions in each category are then compared using a Fisher Exact test. This method can be used in cases of low numbers or a single batch (with no concurrent controls).
    - Time as a Fixed Effect – a regression approach treating batch (or time) as a fixed effect, accepting up to five batches of mutant animals with concurrent controls. This method can include or exclude body weight as a covariate in the model below.

      $$\mathit{Variable} = \mathit{Genotype} + \mathit{Sex} + \mathit{Genotype}*\mathit{Sex} + \mathit{Weight} + \mathit{Batch}$$

    - Mixed model – a linear mixed model treating batch as a random effect, used in cases with multiple batches of both mutants and controls, not necessarily concurrent. This method can also include or exclude body weight as a covariate in the model below.

      $$\mathit{Variable} = \mathit{Genotype} + \mathit{Sex} + \mathit{Genotype}*\mathit{Sex} + \mathit{Weight} + (1|\mathit{Batch})$$

For further information on this methodology and how to run the code, please refer to the following publication and documentation.

https://doi.org/10.1371/journal.pone.0131274
https://www.bioconductor.org/packages/release/bioc/html/PhenStat.html

Metadata

Much additional information was collected at the same time as the experimental variables. This includes fields such as 'Assay Date' and 'Sex', which can be used in analysis methods as described above, but also other data such as 'Anaesthetic' and 'Equipment Model' or 'Instrument'. These fields should be considered as potential confounders and it may not be valid to combine measurements taken under different metadata conditions.

In general, data found in different folders within a pipeline should not be combined. E.g. in the MGP Select pipeline, data in Haematology_ABC and Haematology_ABC-plus_10-03-2016 should not be analysed together as they were collected on different analysers.

Within an individual folder/test, the data may also need to be divided on other factors as mentioned above. Internally, we split data for separate analysis based on the metadata conditions in the following table for each test. Baselines were also split on these conditions.

Procedure	**Important metadata**
Open Field	Equipment, surface area
Grip Strength	Equipment, Grid model
Hot Plate	Light intensity, equipment
Indirect calorimetry	Equipment
ipGTT	Fasting length, Strip type, equipment
DEXA	Equipment
Ophthalmoscope	Topical agents(s)
Haematology	Equipment, fasting status, anaesthesia
Clinical chemistry	Equipment, fasting status, anaesthesia
PBL flow cytometry	Equipment