Improved Data Access and Analysis for the IMPC – International Mouse Phenotyping Consortium

Marina Kan[1], Federico Lopez Gomez[1], Janine Wotton[2], Ewan Selkirk[3], Piia Keskivali-Bond[3], Robert Wilson[1], Sara Wells[3], Jacqui White[2], Helen Parkinson[1]; the IMPC consortium

1 – European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK;
2 – The Jackson Laboratory, Bar Harbor, ME 04609, USA; 3 – Mary Lyon Centre at MRC Harwell, Harwell Campus OX11 7UE, UK

Abstract

IMPC Lines and Data in Numbers

Used in	Analysed	Identified	Produced
8,000+ Publications	9,000+ Genes	111,000+ Phenotypes	98,000,000+ Data Points

IMPC is a collaboration that aims to identify the function of every protein-coding gene in the mouse genome. To achieve this, knock-out mouse lines are extensively phenotyped using highly standardised tests. The results are freely available on the IMPC website, www.mousephenotype.org[1]. IMPC data is released twice yearly and has recently been used to identify new genetic disease in rare disease patients[2,3]and genes with novel functions associated with cardiovascular disease[4]. We have recently improved the processing of late adult data by expanding the window size for the late adult weight covariate sampling from 5 to 11 days, and demonstrated that this increases completeness of the data without compromising accuracy. To make IMPC data easier to access we have developed a Python package for working with the data API and published an on-demand training course that illustrates how to use the package to retrieve the data.

Introduction

IMPC utilises a standardised phenotypic pipeline applied at three life stages.

Embrionyc → Early Adult → Late Adult

After c. 50th Week

To improve explanatory power, body weight (BWT) can be used as a covariate in statistical analyses. The BWT of an adult mouse remains relatively stable. Due to the way data is processed it is not always possible to identify a BWT within the appropriate time window for the analysis.

This study aims to improve data fidelity by expanding the BWT measurement window beyond five days around the experiment date.

Knock-out
7♀ + 7♂

Wild Type

Methods

- Body weight data: IMPC Solr experiment core for data release 21.1.
- IMPC data analysis: R package `OpenStats`[5].
- Calculations: Python `pandas` library.

$$distance = x_2 - x_1$$

$$error = \frac{|\, y_2 - y_1 \,|}{min(y_2,\ y_1)} \cdot 100\%$$

- In the formula $(x_1,\ y_1)$ and $(x_2,\ y_2)$ are two measurements for one specimen. x_1 and x_2 – age in days and y_1 and y_2 – BWT in grams.
- Linear Mixed Model method is used with or without BWT as a covariate in the following equation:

$$Y = G + S + G \times S + BWT$$

- In this formula: Y – Response, G – Genotype, S – Sex, BWT – Body Weight.

Results

Error Distribution by Distance Interval

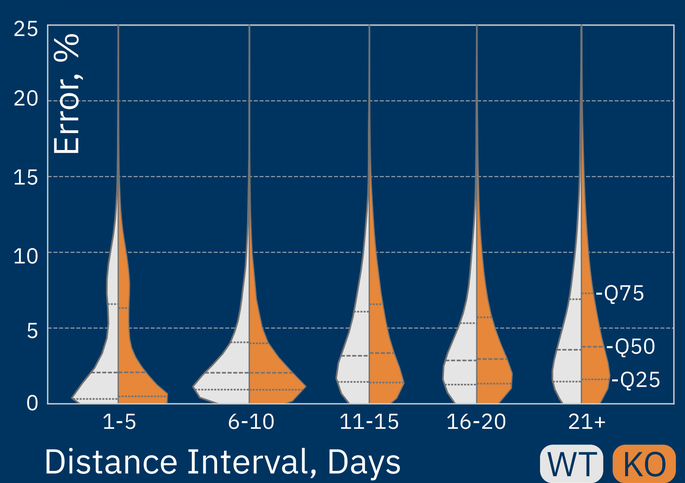

Chart 1. Data were clipped for errors > 25. Starting from days 11–15, the error in both the control and experimental groups diverged, with the median error increasing.

Distance vs. Error

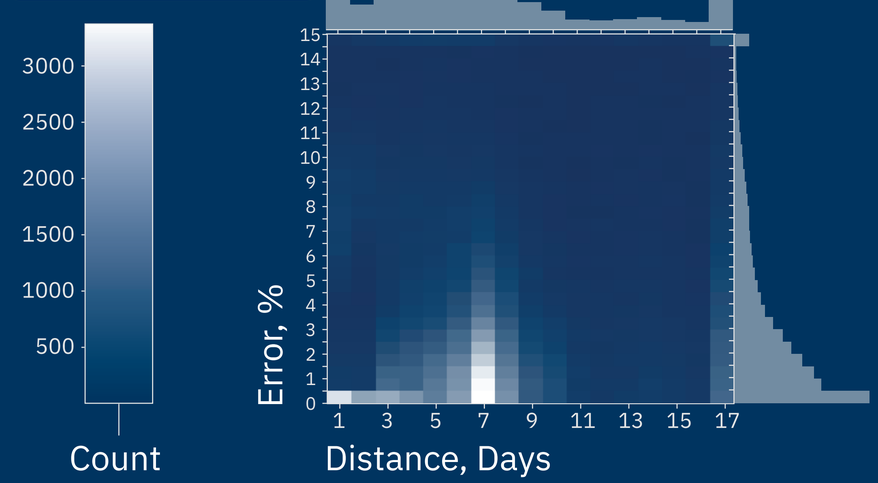

Chart 2. Data were clipped for errors > 15% and distance > 17 days. The highest number of values is at the 7-day interval because tests are usually performed weekly. Total number of observed body weight values is 115,331.

Conclusion

- The BWT window size was increased from ±5 days to ±11 days, which now include BWT values for 305,038 parameters.
- 2,201 values shifted the analysis from 'without BWT' for a ±5-day window to 'with BWT' for an ±11-day window. Most of the p-values maintained their significance level, except for one, which became significant, though this value is very close to the threshold.

Access IMPC Data

A Python package `impc-api` can be utilised to download IMPC data. Use `pip` to install it.

Navigate to our course via the QR code or the provided link to familiarise yourself with the programmatic data access.

Scan QR Code to Learn How to Obtain IMPC Data

https://www.ebi.ac.uk/training/online/courses/impc-solr-api/

References

1. Groza et al. Nucleic Acids Res. 2023;51(D1):D1038-D1045.
2. Smedley et al. Dis Model Mech. 2024;17(6):dmm050604.
3. Cacheiro et al. Preliminary Report. N Engl J Med. 2021;385(20):1868-1880.
4. Spielman et al. Nat Cardiovasc Res. 2022;1(2):157-173.
5. www.bioconductor.org/packages/OpenStats