IMPC annotation pipeline

Annotation pipeline in the Internationa Mouse Phenotyping Consortium (IMPC) is an existing data assignment project with a goal to associate phenotipic observations to the genetic modidification. Here we explain the steps that are taken to select the best Mamalian Phenotype (MP) term to the genetic modification in mice when a significant difference (typically at the level of 0.0001) from the baselines observed.

Annotation pipeline and the analysis framework

The IMPC annotation pipeline (IMPC-AP) assigns MP terms to the the significant genetic effect. The genetic effect at the IMPC is specified by three statistical analysis platforms that are designed in the IMPC statistical pipeline through OpenStats software. Here we break the the annotation pipeline by the type of the input data and the analysis frameworks.

Annotation table

The annotation pipeline in the IMPC require a reference table that summarizes the available terms for an IMPC parameter. This can be retrieved from IIMPReSS however to remove the dependency to the live servers, the IMPC-AP utilised an offline version of the file called *Annotation Indexer* in this document. This file is available from XXXXX.

Continuous data – Linear mixed model

Continuous data such as tail length, tibia length etc. in IMPC is analysed by linear mixed model, implemented in the software package OpenStats. The continuous measurements are more informative that the other types in that aspect that the direction of change can be determined by the effect size. Here we summarised the steps to assign MP terms to the continuous measurements.

<table>
<tr><th></th><th>From the statistical results</th><th>From the Annotation Indexer</th></tr>
<tr><td></td><td>1. Overal effect (both sexes)
• if pvalue ≥ threshold → Assign no MP term
• if pvalue < threshold (II)
i. If effect size>0 → Increase term
ii. if effect size<0 → Decrease term
iii. if effect size=0 → Steady term
• similar steps in 1 apply for Male effect (III)
• similar steps in 1 apply for Female effect (IV)</td><td>Filter for
1. Pipeline_stable_id
2. Procedure_group
3. Parameter_stable_id

Get available MP terms (I)</td></tr>
<tr><td></td><td colspan="2">Find matches between I and II, III, IV</td></tr>
<tr><td>Notes</td><td colspan="2">• If increase or decrease effect detected then ignore ABNORMAL MP term.
• Generally accepted threshold by the IMPC consortium is 0.0001</td></tr>
</table>

Continuous data – Reference Range plus

Due to the complexity of the data not all continues data can be analysed by linear mixed model. Alternatively, there are many cases in the IMPC that are analysed by Reference Range plus (RR+) method implemented in the OpenStats software pachage. RR+ is a hurstic method that works on the basis of discritising baseline data into low/normal/high categories. The mutants aer then assigned a class based on the reference categories. Finally, Fisher Exact test applies to specify any significant devuatuoin from the normal category. Here we explain the MP term assignment algorithm for the results from the RR+ frameork.

	From the statistical results	From the Annotation Indexer
	1. Overal effect (do not consider gender) • if pvalue.low ≥ threshold & pvalue.high ≥ threshold → Assign no MP term • for each pvalue(.low/.high) < threshold then assign label 'ABNORMAL', 'INCREASED', 'DECREASED' to the search criteria • Remove any Low.INCREASE and High.DECREASE from the labels **(II)** • Apply a similar step to 1 to Male effect **(III)** • Apply a similar step to 1 to Female effect **(IV)**	Filter for 2. Pipeline_stable_id 3. Procedure_group 4. Parameter_stable_id Get available MP terms **(I)**
	Find matches between **I** and **II, III, IV *but ignore the term Low and High.***	
Notes	• IF LOW and High MP terms detected then select ABNORMAL term Generally accepted threshold by the IMPC consortium is 0.0001	

Categorical data

Categorical data in the IMPC encomapsses a range of qualitative measurements such as abnormality in eye, ear, tail and are analysed using Fisher's Exact test implemented in the R package OpenStats. The output MP term for this type of data is a single term Abnormal phenotype if the test is significant. Here we explain the algorithm:

Step	From the statistical results	From the Annotation Indexer
	1. Overall effect (do not consider gender) • if pvalue ≥ threshold → Assign no MP term • if pvalue < threshold search for the MP term **(II)** • Apply a similar step to 1 to Male effect **(III)** • Apply a similar step to 1 to Female effect **(IV)**	Filter for 5. Pipeline_stable_id 6. Procedure_group 7. Parameter_stable_id Get available MP terms **(I)**
	Find matches between **I** and **II, III, IV**	
Note	• Generally accepted threshold by the IMPC consortium is 0.0001	

Schematic view of the IMPC-AP