Causal Challenge 2016 - Mouse gene knockouts: prediction and causal inference

Team name: The Noncompliers

K. McGregor[1,2] and G. Simoneau[1]

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University [2] Lady Davis Institute for Medical Research, Jewish General Hospital

Background

- The International Mouse Phenotype Consortium (IMPC) is an international collaboration aimed at discovering functional insight for every gene through the systematic phenotyping of 20,000 knockout mouse strains.
- The knockout procedure turns off the activity of a mouse gene in order to assess what biological systems are impacted.
- Available data: 22 phenotypic measurements on 614 mice from 190 litters, representing 14 genotypes (wild type and 13 different knockout conditions).
- Phenotypic measurements are likely to vary across litters and genotypes.
- Missing data: 5 knockout conditions were randomly selected for which all observations from a randomly selected variable (different for each condition) were removed. There are 67 missing values in total.

Objective

1. To infer missing data from 5 phenotypic measurements: red blood cell distribution width, lymphocyte differential count, basophil differential count, neutrophil differential count, monocyte cell count.
2. To produce a causal interpretation of the available data by taking advantage of the experimental perturbations.

Methods

Predictions for the 67 missing values were obtained with **Multiple Imputation using Chained Equations** (MICE) [1]:

- The unknown missing values were replaced by 30 independent simulated sets of values drawn from the posterior predictive distribution of the missing data conditional on the observed data.
- Predictions were obtained by averaging the imputed values across the 30 simulated datasets.
- Assumes missing data are missing at random (MAR) or completely at random (MCAR).
- Lymphocyte differential was predicted via a direct calculation (lymphocyte count/total WBC count). Neutrophil differential predictive model formed from OLS on complete observations.

Causal relationships were investigated separately for each of the 22 phenotypic measurements:

- A simple linear regression model was applied to estimate the effect of the knockout conditions on each phenotype. For phenotypes with missing values, the MICE algorithm was used to incorporate uncertainty of the predicted values.
- The **Sandwich Estimator** [2] was used to estimate the variance of the effect size to account for clustering by litter.
- Significant causal relationships were identified after a Bonferroni correction (i.e. at level 0.05/22).
- Prior biological knowledge was used to infer possible causal relationships between the phenotypic measurements.

Results

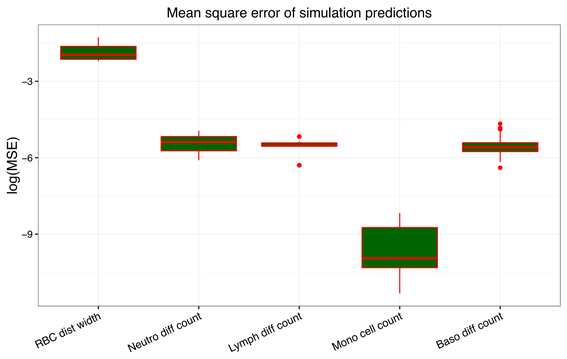


Figure 1. Results from simulation with 100 replications. In each replication five variables were deleted for five different knockout conditions and the MSE of the imputed values was calculated.

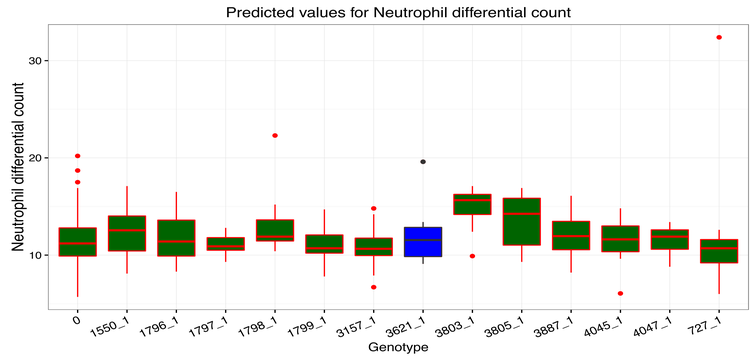


Figure 2. Boxplots of neutrophil differential count by knockout condition. The green boxplots correspond to observed data and the blue boxplots correspond to predicted data.

Table 1. Significant knockout-phenotype relationships

Knockout	Phenotypic measurements (direction of association)
1550_1	Hematocrit (+), Mean cell volume (+), Hemoglobin (–), MCHC[1] (–), Mean platelet volume (–)
1796_1	MCHC (+), Hematocrit (–), Mean cell volume (–), Monocyte diff.[2] count (–), LUC[3] count (–), LUC diff. count (–)
1797_1	WBC[4] count (+), Hemoglobin (+), MCHC (+), Mean platelet volume (+), Lymphocyte cell count (+)
1798_1	MCHC (+), Hematocrit (+), Mean cell volume (+)
1799_1	Hematocrit (+), Mean cell volume (+), LUC count (+), LUC diff. count (+), Mean corpuscular hemoglobin (–), MCHC (–), Mean platelet volume (–), Eosinophil diff. count (–), Eosinophil cell count (–)
3157_1	Mean platelet volume (+), Lymphocyte diff. count (+), Eosinophil diff. count (–), Eosinophil cell count (–)
3621_1	LUC diff. count (+), Monocyte cell count (+), RBC[5] count (–), Hemoglobin (–), Hematocrit (–), Mean cell volume (–), Mean corpuscular hemoglobin (–), MCHC (–), Platelet count (–), Monocyte diff. count (–)
3805_1	Mean cell volume (+), LUC count (+), LUC diff. count (+), MCHC (–)
3887_1	Mean cell volume (+), Neutrophil cell count (+), Mean platelet volume (–), LUC diff. count (–), Monocyte cell count (–)
4045_1	Mean platelet volume (–), LUC diff. count (–)
4047_1	Hematocrit (+), Mean cell volume (+), Mean platelet volume (–)
727_1	Monocyte diff. count (+), Mean corpuscular hemoglobin (–), MCHC (–)

Color coding: Phenotypic measurements related to **white blood cells**, **red blood cells** or **platelets**.
[1] Mean corpuscular hemoglobin concentration, [2] Differential, [3] Large Unstained Cell, [4] White blood cell, [5] Red blood cell

- Each phenotypic variable falls into one of three categories: red blood cells, white blood cells, and platelets. [4]
- Many associations appear due to construction of variables (i.e. lymphocyte cell count and lymphocyte differential count; red blood cell count and mean cell volume)
- Difficult to determine whether gene knockout acts directly on a phenotype vs. through one or more other phenotypes.
- Size of the relative percentage of a white blood cell type can influence whether a change in that cell type's differential count is associated with genotype.
- Hemoglobin and red blood cell measures appear to be closely related, though implying more meaningful causal relationships would likely require more investigation.

Table 2. Significant knockout-phenotype relationships for knockout 3803_1

Knockout	Phenotypic measurements	Effect size (p-values)
3803_1	White blood cell count	-2.99 (< 0.0001)
	Neutrophil cell count	-0.19 (< 0.0001)
	Lymphocyte cell count	-2.63 (< 0.0001)
	Eosinophil cell count	-0.11 (< 0.0001)
	Basophil cell count	-0.02 (< 0.0001)
	Monocyte cell count	-0.03 (< 0.0001)
	Neutrophil differential count	0.39 (< 0.0001)
	Lymphocyte differential count	-0.39 (< 0.0001)
	Red blood cell count	-0.42 (< 0.0001)
	Mean cell volume	2.48 (< 0.0001)
	Mean corpuscular hemoglobin	0.55 (< 0.0001)
	Platelet count	-240 (< 0.0001)

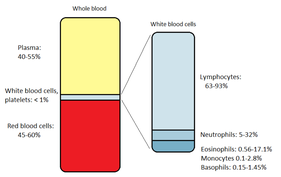
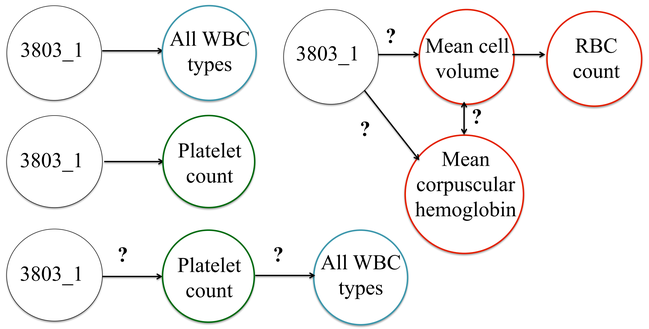


Figure 3. Range of proportions of components of mouse blood [4]

Direct (or indirect?) causal relationships



All analyses were performed using R statistical software.

Discussion

- We have predicted missing values for 5 phenotypes, using prior biological knowledge of the phenotypic measurements, simple OLS, and the MICE algorithm. Moreover, we have investigated possible causal interpretations of the available data by taking advantage of the experimental perturbations of the mouse gene knockouts.
- MICE assumed that the missing data were MAR or MCAR suggesting that missing values can be imputed from the observed data. This assumption holds given the way the data were deleted.
- To find genotype-phenotype causal associations, we had to assume that gene knockout was assigned randomly.
- Prior biological knowledge of the studied phenotypes was incorporated both in the prediction and causal interpretation analyses. This constituted a major strength of our general approach.
- The MICE algorithm was used to investigate causal relationships for phenotypic measurements with missing values. This approach took into account uncertainty in the predicted values and prevented the finding of spurious associations. This also constituted a strength of our method.
- It was not clear whether the causal relationships found by our method were direct or indirect i.e. other measured (or unmeasured) biological elements may be on the causal pathway.
- Our approach did not allow inferring causal relationships between the 22 phenotypic measurements. However, a few interesting associations were observed.
- Temporality was an issue for drawing causal relationships among phenotypes.

Conclusion

We inferred causal links between knockout conditions and phenotypic measurements, but relationships between phenotypes require more information or outside knowledge. Consequently, we chose not to infer a global underlying DAG.

Acknowledgements

GS would like to thank her PhD supervisors, Dr. Erica Moodie and Dr. Robert Platt, for giving her the opportunity to participate in this competition. KM would like to thank his PhD supervisors Dr. Celia Greenwood and Dr. Aurélie Labbe. We would also like to thank Claudia Kleinman for her precious insights on the biological interpretations of the studied phenotypes.

References

[1] White, I. R., Royston, P., & Wood, A. M. (2011). *Multiple imputation using chained equations: issues and guidance for practice*. Statistics in medicine, *30*(4), 377-399.

[2] Freedman, David A. (2012). The American Statistician.

[3] Blood Basics. htt. *On the so-called "Huber sandwich estimator" and "robust standard errors"*p://www.hematology.org/Patients/Basics/#a6

[4] McGarry, Michael P, Cheryl A. Protheroe, and James J. Lee. (2010) *Mouse Hematology: A Laboratory Manual*. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press. Print.