Predicting Novel Abnormal Circadian Phenotypes in Mouse

JA Williams\textsuperscript{1,2*}, S Sethi\textsuperscript{1}, K Condon\textsuperscript{1}, PM Nolan\textsuperscript{1}, MM Simon\textsuperscript{1}, GV Gkoutos\textsuperscript{1}, A-M Mallon\textsuperscript{1}

\textsuperscript{1}Mammalian Genetics Unit, MRC Harwell Institute, Harwell Campus, OX11 0RD, UK
\textsuperscript{2}Institute of Cancer and Genomic Sciences, Centre for Computational Biology, University of Birmingham, B15 2TT, Birmingham, UK

*To whom correspondence should be addressed: joe@nonsequitur.edu

I. Introduction
Disruptions in circadian rhythms have been associated with disorders ranging from schizophrenia to bipolar disorder, implicating mutations in known circadian clock genes as potential therapeutic targets or biomarkers of mental illness (1). In mouse models, gene function can be discovered by reverse genetic screens by analyzing phenotypes of knockout or mutagenized mouse lines. Current large phenotype programs such as the International Mouse Phenotype Consortium test for a variety of behavioral phenotypes, but circadian related assays are not widespread. To prioritize candidates for circadian phenotype screens, we used machine learning approaches to predict the abnormal circadian rhythm phenotype in mouse genes.

II. Results
We analyzed RNA-Sequencing libraries from two mouse tissues: the central pacemaker clock of the suprachiasmatic nucleus, and a peripheral clock tissue liver. We detected signature circadian oscillations in each tissue’s expression, indicative of circadian characteristics. Abundance of expression in the central clock and relative expression distributions throughout 26 tissues were also shown to be indicative of potential circadian phenotype. Using protein-protein interaction graphs, we constructed a diffusion kernel on known interactions, scoring nodes from short random walks around known circadian genes. These were used as features in a RUSBoost ensemble tree classifier. 91 genes with annotated abnormal circadian rhythm phenotypes were used as positive targets. A total of 275 out of 12502 protein coding genes expressed in the suprachiasmatic nucleus were predicted to have novel abnormal phenotypes (Table 1). Highly ranked are known clock genes have no abnormal circadian mouse phenotypes in mouse, but present circadian phenotypes in human orthologs. Known circadian genes Npas2 and Fbxw11 were successfully predicted to contribute to circadian phenotypes (Figure 1).

<table>
<thead>
<tr>
<th></th>
<th>No Phenotype</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Phenotype</td>
<td>12136</td>
<td>24</td>
</tr>
<tr>
<td>Phenotype</td>
<td>275</td>
<td>67</td>
</tr>
<tr>
<td>Total</td>
<td>12411</td>
<td>91</td>
</tr>
</tbody>
</table>

Table 1. 275 novel abnormal circadian phenotype predictions, classed as false positives, narrow the search space for protein coding circadian related genes. Table represents predicting on all data.
Figure 1. Npas2 predicted to have circadian phenotype. Npas2 shown with interacting protein partners expressed in suprachiasmatic nucleus. Shading (dark to light) indicates probability of circadian phenotype. Diamond = known circadian phenotype. Dark borders = cycling expression.

III. Conclusions

Machine learning predicts genes which, if disturbed, may contribute to abnormal circadian phenotypes in mouse. Potential novel circadian functions will require experimental validation. Our findings highlight bias in mammalian phenotype annotations, reflecting both what has been studied to date in mouse, and the redundancy in the mammalian clock. Inferring mammalian phenotypes has potential to improve whole-phenome approaches to disease prioritization in humans. Using a combination gene expression and protein graph metrics, our machine learning method reduces the search space for phenotype annotations.

IV. REFERENCES