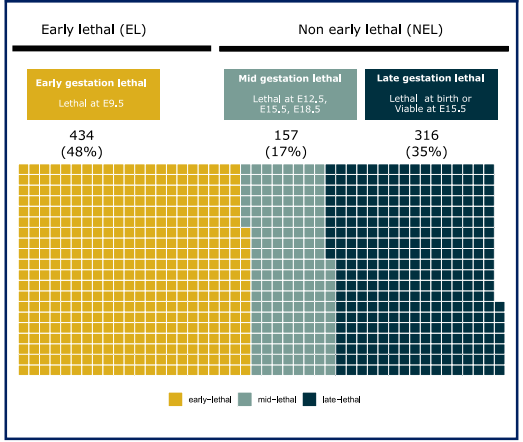
Predicting early lethal genes in the mouse

Pilar Cacheiro[1], Henrik Westerberg[2], Jesse Manger[3], Damian Smedley[1], the International Mouse Phenotyping Consortium

[1] William Harvey Research Institute, Queen Mary University of London, UK; [2] Medical Research Council Harwell Institute, Harwell, UK; [3] Department of Veterinary and Animal Sciences, University of Massachusetts, Amherst, USA

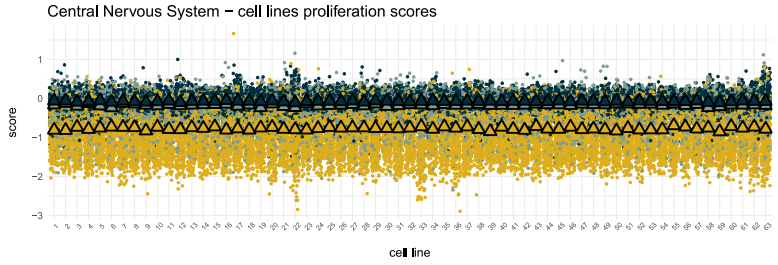
1

The International Mouse Phenotyping Consortium performs different viability screens. First, a primary viability screen to assess the postnatal viability, sub-viability, and lethality of mice. For those lines found to be lethal, a secondary screen is conducted, measuring the viability at 9.5, 12.5, 15.5, 18.5 days of gestation and at birth. The embryonic stage at which lethality occurs is summarised in a set of windows of lethality, where a single live homozygous embryo at any time point defines a window. Complete lethality at E9.5 is therefore classified as early-lethal, at E12.5, E15.5 or E18.5 as mid-lethal, and lethality at birth or viability at E15.5 as late-lethal. With information for 907 genes, nearly half of embryonic lethal lines (434, 48%) are early-gestation lethal, 17% (157) mid-gestation lethal and 35% (316) late-gestation lethal.

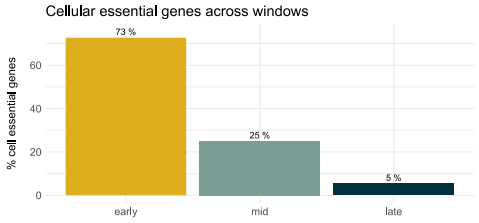


2

Previous results showed that early lethal genes are highly correlated with cellular essential genes. Human cell essentiality scores from CRISPR knockout screens (DepMap release 21Q2, project Achilles) for 859 cell lines corresponding to different tissues are available for analysis.



Distribution of proliferation scores for different Central Nervous System cell lines across windows of lethality. Higher values indicate more proliferation, i.e. less essential. Triangles represent median values of gene expression per window and cell line

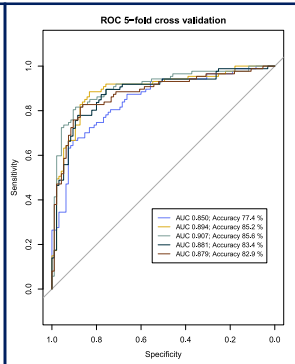


Percentage of EL, ML and LL genes considered cellular essential when a threshold for cellular essentiality is considered

3

Since the number of lines that have undergone the primary viability assessment is higher than those with a secondary screen, we tried to predict additional early lethal genes from this pool. We used a penalised likelihood approach to fit a generalised additive model using the R implementation of GAM selection, *gamsel*.

The training set consisted of 434 EL and 473 NEL genes. Cross validation (5-fold) was used to assess the performance of the model. Using this model, predictions for a total number of 681 lethal genes with no secondary viability assessment were made: 296 (43.5%) genes were predicted as EL and the remaining 385 (56.5%) genes were predicted as NEL.



	DepMap_ID	lineage
1	ACH−000045	blood
2	ACH−000141	bile_duct
3	ACH−000257	lung
4	ACH−000272	kidney
5	ACH−000318	esophagus
6	ACH−000459	kidney
7	ACH−000524	ovary
8	ACH−000537	liver
9	ACH−000573	breast
10	ACH−000618	upper_aerodigestive
11	ACH−000677	lung
12	ACH−000688	ovary
13	ACH−000760	central_nervous_system
14	ACH−000851	lung
15	ACH−000969	colorectal
16	ACH−001041	lung
17	ACH−001061	colorectal
18	ACH−001163	kidney
19	ACH−001518	uterus
20	ACH−001628	ovary
21	ACH−001736	blood
22	ACH−001737	blood
23	ACH−001786	colorectal

CV ROC-AUCs ranged from 0.850 to 0.907. The accuracy ranged from 77.4 to 85.6% of instances correctly classified as EL and NEL.

Only 23 (out of 859) predictor variables (non-zero effects) were selected in the final model

Gene	Predicted Probability	Predicted Class
Kcna10	0.544	EL
Cdipt	0.921	EL
Ipo11	0.851	EL
Ddx59	0.810	EL
Ipo7	0.915	EL
Dnajc13	0.470	NEL
Ltv1	0.823	EL
Nup205	0.920	EL
Eif5b	0.108	NEL
Ncapd3	0.954	EL
Elf1	0.854	EL
Nars	0.976	EL
Exosc1	0.793	EL
Lin52	0.578	EL
Taf5	0.938	EL
Osbp	0.463	NEL
Tubgcp2	0.973	EL
Cct7	0.551	EL
Pop5	0.984	EL
Isy1	0.963	EL
Thoc3	0.904	EL
Rad54l2	0.321	NEL
Uap1	0.463	NEL
Dhx8	0.977	EL
Nifk	0.963	EL
Ppat	0.756	EL

Out of 26 genes that were externally assessed as EL, 21 were correctly predicted by the classifier (80.8%), while the predicted probabilities for 3 out of 5 genes that were incorrectly labelled were very close to the 0.51 probability threshold that maximised the F1 score. External dataset: ***https://blogs.umass.edu/jmager/***