

The Jackson Laboratory

## The Mouseion at the JAXlibrary

---

Faculty Research 2022

Faculty Research

---

10-13-2022

### **Mendelian gene identification through mouse embryo viability screening.**

Pilar Cacheiro

Carl Henrik Westerberg

Jesse Mager

Mary E Dickinson

Lauryl M J Nutter

*See next page for additional authors*

Follow this and additional works at: <https://mouseion.jax.org/stfb2022>



Part of the [Life Sciences Commons](#), and the [Medicine and Health Sciences Commons](#)

---

---

## Authors


Pilar Cacheiro, Carl Henrik Westerberg, Jesse Mager, Mary E Dickinson, Lauryl M J Nutter, Violeta Muñoz-Fuentes, Chih-Wei Hsu, Ignatia B Van den Veyver, Ann M Flenniken, Colin McKerlie, Stephen A Murray, Lydia Teboul, Jason D Heaney, K C Kent Lloyd, Louise Lanoue, Robert E Braun, Jacqueline K White, Amie K Creighton, Valerie Laurin, Ruolin Guo, Dawei Qu, Sara Wells, James Cleak, Rosie Bunton-Stasyshyn, Michelle Stewart, Jackie Harrisson, Jeremy Mason, Hamed Haseli Mashhadi, Helen Parkinson, Ann-Marie Mallon, International Mouse Phenotyping Consortium, Genomics England Research Consortium, and Damian Smedley

RESEARCH

Open Access



# Mendelian gene identification through mouse embryo viability screening

Pilar Cacheiro<sup>1</sup>, Carl Henrik Westerberg<sup>2</sup>, Jesse Mager<sup>3</sup>, Mary E. Dickinson<sup>4,5</sup>, Lauryl M. J. Nutter<sup>6</sup>, Violeta Muñoz-Fuentes<sup>7</sup>, Chih-Wei Hsu<sup>4,8</sup>, Ignatia B. Van den Veyver<sup>5,9</sup>, Ann M. Flenniken<sup>10</sup>, Colin McKerlie<sup>6</sup>, Stephen A. Murray<sup>11</sup>, Lydia Teboul<sup>12</sup>, Jason D. Heaney<sup>5</sup>, K. C. Kent Lloyd<sup>13</sup>, Louise Lanoue<sup>13</sup>, Robert E. Braun<sup>11</sup>, Jacqueline K. White<sup>11</sup>, Amie K. Creighton<sup>6</sup>, Valerie Laurin<sup>6</sup>, Ruolin Guo<sup>10</sup>, Dawei Qu<sup>10</sup>, Sara Wells<sup>12</sup>, James Cleak<sup>12</sup>, Rosie Bunton-Stasyshyn<sup>12</sup>, Michelle Stewart<sup>12</sup>, Jackie Harrisson<sup>12</sup>, Jeremy Mason<sup>7</sup>, Hamed Haseli Mashhadi<sup>7</sup>, Helen Parkinson<sup>7</sup>, Ann-Marie Mallon<sup>2</sup>, International Mouse Phenotyping Consortium, Genomics England Research Consortium and Damian Smedley<sup>1\*</sup> 

## Abstract

**Background:** The diagnostic rate of Mendelian disorders in sequencing studies continues to increase, along with the pace of novel disease gene discovery. However, variant interpretation in novel genes not currently associated with disease is particularly challenging and strategies combining gene functional evidence with approaches that evaluate the phenotypic similarities between patients and model organisms have proven successful. A full spectrum of intolerance to loss-of-function variation has been previously described, providing evidence that gene essentiality should not be considered as a simple and fixed binary property.

**Methods:** Here we further dissected this spectrum by assessing the embryonic stage at which homozygous loss-of-function results in lethality in mice from the International Mouse Phenotyping Consortium, classifying the set of lethal genes into one of three windows of lethality: early, mid, or late gestation lethal. We studied the correlation between these windows of lethality and various gene features including expression across development, paralogy and constraint metrics together with human disease phenotypes. We explored a gene similarity approach for novel gene discovery and investigated unsolved cases from the 100,000 Genomes Project.

**Results:** We found that genes in the early gestation lethal category have distinct characteristics and are enriched for genes linked with recessive forms of inherited metabolic disease. We identified several genes sharing multiple features with known biallelic forms of inborn errors of the metabolism and found signs of enrichment of biallelic predicted pathogenic variants among early gestation lethal genes in patients recruited under this disease category. We highlight two novel gene candidates with phenotypic overlap between the patients and the mouse knockouts.

**Conclusions:** Information on the developmental period at which embryonic lethality occurs in the knockout mouse may be used for novel disease gene discovery that helps to prioritise variants in unsolved rare disease cases.

## Background

The rate of molecular diagnosis through genomics approaches continues to improve. However, the diagnostic yield for Mendelian disorders varies significantly, ranging from 25 to 58% [1, 2] depending on the age of the proband, the type of disorder, the criteria for patient

\*Correspondence: d.smedley@qmul.ac.uk

<sup>1</sup> William Harvey Research Institute, Queen Mary University of London, London, UK

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

inclusion (e.g. absence of a clear clinical diagnosis, previous attempts to provide a molecular diagnosis) and the availability of sequence data from family members, e.g. familial versus sporadic cases. Despite this progress, a considerable proportion of patients remain without a diagnosis. Potential strategies to address the challenge of undiagnosed patients and advance our understanding of the molecular basis of these disorders include but are not limited to (i) identifying novel Mendelian disease genes [3]; (ii) developing experimental and computational approaches to assess the pathogenicity of variants of unknown significance in known disease genes; (iii) considering expansion of the phenotype of known disease genes [4]; (iv) investigating noncoding, regulatory variants; (v) assessing the contribution of structural variation [5]; (vi) investigating somatic mosaicism; and (vii) exploring alternative modes of inheritance, i.e. digenic or multigenic [2].

With regard to the first approach, the number of genes currently known to be associated with rare disorders comprises 20–25% of the protein coding genome according to OMIM [6]. There are between 200 and 300 new disease-gene associations published every year [7], with many more to be uncovered. Frameworks such as the Clinical Genome Resource or the Genomics England (GEL) PanelApp, a publicly available knowledgebase containing expert curated gene panels related to human disorders, are key to summarise and assess all curated evidence and provide clinical validation for these gene-disease pairs [8, 9]. The number of additional disease-associated genes yet to be identified is estimated to be high, up to 1.5–3 times the number of currently known causative genes of Mendelian conditions [10].

The main approach to identify genes underlying autosomal recessive (AR) disorders has been homozygosity mapping combined with mutation screening in large consanguineous pedigrees. However, this is infrequent in outbred populations, where recessive disorders likely remain underdiagnosed [11]. The use of large exome and sequence datasets, including information on variant frequency and gene intolerance to variation metrics, has been widely implemented in rare disease diagnostic pipelines. Conversely, in large cohorts such as those from UK Biobank [12] and gnomAD [13], we are unlikely to find homozygous loss-of-function (LoF) variants, i.e. complete knockouts, for many genes [14]. A recent study in the European population estimated that every individual is a carrier of at least 2 pathogenic variants in genes known to be associated with AR disease and consequently up to 1% of couples within this population would be at risk of having a child affected by these disorders. This risk increases for consanguineous couples and skeletal disorders and intellectual disabilities [15].

Additionally, variants associated with AR disorders could result in attenuated phenotypes in heterozygous carriers [16]. Hence, identifying biallelic pathogenic variants in rare disease cohorts like the 100,000 Genomes Project (100KGP) [17] remains a crucial task that requires alternative approaches, including evaluating genes not yet associated with disease.

Combining different sources of information can boost the evidence for new disease-gene associations. Integrating research and clinical datasets has proven to be effective at discovering the molecular basis for genetic disorders [18, 19]. Model organism information on viability and cross-species phenotype comparisons in combination with clinical data constitutes another powerful strategy. Some examples include the automatic detection of mouse models for human disease and phenotype-based variant prioritisation using algorithms such as PhenoDigm and Exomiser [20–22]. Additionally, mouse data on essentiality can be used as a discovery and prioritisation tool [23, 24]. We previously developed a gene prioritisation strategy focused on neurodevelopmental disorders by integrating evidence of intolerance to LoF variation from multiple resources and data from large scale sequencing programmes [25]. Through this approach, combining viability data from mice and human cell line screens, we were able to identify a set of developmentally lethal genes, i.e. genes not essential for cell proliferation but required for organism development, which were enriched for autosomal dominant (AD), developmental disease-associated genes. Investigation of clinical cases with de novo variants in developmental lethal genes and phenotypic overlap between the knockout mouse and affected individuals led us to prioritise a set of 9 candidate genes. Two of these genes have since been validated [26, 27].

To improve and expand these successful strategies to other types of disorders, here we again leverage evidence from high-throughput mouse phenotype screens conducted by the International Mouse Phenotyping Consortium (IMPC) to further explore the spectrum of intolerance to LoF variation. For genes with null alleles that result in a lethal phenotype in a primary viability screen (i.e. no live homozygous animals identified between 14 days of age and weaning), the IMPC performs a secondary embryo viability screen to determine a ‘window of lethality’ (WoL) by examining the survival of homozygous null mutants at different embryonic developmental time points [24]. In the present study, we further dissected this set of lethal genes in the mouse with the primary aim of investigating how they can inform human disease gene discovery.

First, we explored these WoL and show how they relate to essentiality inferred from human cell proliferation assays,

gene expression across development, intolerance to variation metrics and duplication events. Secondly, we investigated these WoL in the context of human Mendelian disease and found that early-gestation lethal genes in the mouse are correlated with AR disease-associated genes, in particular those involved in inherited metabolic disorders, resulting mainly from enzyme deficiencies [28]. Finally, we developed two gene prioritisation strategies to identify novel candidate genes for this type of disorders: one based on gene similarity to biallelic inborn errors of metabolism (BIEM) genes, a broad category of genes that function in metabolism and impact, or are impacted by most cellular processes [29], and the other based on enrichment of biallelic predicted pathogenic variants among unsolved metabolic disorder cases from the 100KGP [17].

## Methods

### Data sources

#### IMPC mouse data

Mouse primary and secondary viability data were obtained from the IMPC resource [30].

Primary viability data: <http://ftp.ebi.ac.uk/pub/databases/imp/all-data-releases/release-15.0/results/viability.csv.gz> (DR15) [Downloaded 28.09.21].

Phenotype annotations [31, 32]: DR15.0 / DR16.0.

Embryonic viability data: Detailed information on the primary and secondary viability pipelines, including definitions, procedures and protocols, can be found at <https://www.mousephenotype.org/impres/index>. These include the following: Viability Primary Screen, Viability E9.5 Secondary Screen, Viability E12.5 Secondary Screen, Viability E14.5-E15.5 Secondary Screen, Viability E18.5 Secondary Screen, Homozygote Viability at Weaning Screen. A full description of the WoL is available (File S1 [33]).

#### Entire set of human protein coding genes with the corresponding mouse orthologues

One-to-one human orthologues were obtained from the HUGO Gene Nomenclature Committee (HGNC) resource [34]: [http://ftp.ebi.ac.uk/pub/databases/genenames/hgnc/tsv/locus\\_groups/protein-coding\\_gene.txt](http://ftp.ebi.ac.uk/pub/databases/genenames/hgnc/tsv/locus_groups/protein-coding_gene.txt) [Downloaded 28.09.21].

All other gene features used in this study correspond to human orthologue gene annotations. Gene symbols, Ensembl and Uniprot identifiers were converted into HGNC unique identifiers. Where there was any ambiguity about gene id mapping, the annotation was discarded.

#### Human cell proliferation scores

CRISPR knockout screens from the Achilles pipeline (release 21Q3) for 902 cell lines and the corresponding cell line information were obtained from the DepMap portal [35]: <https://depmap.org/portal/download/all/>

(Achilles\_gene\_effect\_CERES.csv) [Downloaded 28.09.21]. Gene effect scores are direct estimates of the effect of a gene knockout on viability. Thus, a more negative CERES score indicates more depletion in the cell line. Average scores per gene were computed. In order to establish a binary threshold to classify genes as cellular essential and non-essential, previous data on cell essentiality, based on 11 cell lines from 3 different studies, was used to compute F1 scores derived from confusion matrices generated when considering different CERES mean scores and the classification from these 3 studies, and mean score cut-offs of  $-0.40$ ,  $-0.45$ , and  $-0.55$  were found to maximise the F1 scores across the different datasets, similar to the  $-0.45$  threshold estimated with information from 485 cell lines [25, 30].

#### Gene expression across development

Human gene expression (RPKM) across development for brain, cerebellum, heart, kidney, liver, ovary and testis was obtained from Cardoso-Moreira et al. [36] <https://apps.kaessmannlab.org/evodevoapp/> [Downloaded 10.08.21].

Data on comparison of temporal trajectories between human genes and their orthologues in mouse for brain and cerebellum was obtained from Cardoso-Moreira et al. [37].

#### Intolerance to variation scores

gnomaAD v2.1.1 constraint metrics [13] (LOEUF, pLI and pRec) and DOMINO scores [38]: <https://gnomad.broadinstitute.org/downloads#v2constraint>; <https://www.fbm.unil.ch/domino/> [Downloaded 10.08.21] SCoNeS [39] and RVIS [40] scores.

#### Gene duplicates

Annotation of paralogues of human genes was obtained from Ensembl Biomart (Ensembl Genes 104) [41] <https://www.ensembl.org/biomart/martview/>. Only protein coding paralogues with HGNC ids and % amino acid identity  $\geq 20\%$  were considered [Downloaded 10.08.21].

#### Protein-protein interactions

Human protein network data (scored links between proteins) were obtained from STRING [42] [https://stringdb.org/cgi/download?sessionId=%24input%3E%7BsessionId%7D&species\\_text=Homo+sapiens](https://stringdb.org/cgi/download?sessionId=%24input%3E%7BsessionId%7D&species_text=Homo+sapiens) [Downloaded 13.08.21].

#### Pathways

Lowest level pathways were obtained from Reactome [43] <https://reactome.org/download/current/UniProt2Reactome.txt> and <https://reactome.org/download/current/ReactomePathways.txt> [Downloaded 10.08.21].

#### Protein families

PFAM protein families [44] were obtained through Ensembl biomart (Ensembl Genes 104) <https://www.ensembl.org/biomart/martview/> [Downloaded 10.08.21].

**Protein complex**

Corum protein complex information [45] was accessed at: <https://mips.helmholtzmuemchen.de/corum/#download> [Downloaded 13.08.21].

**Disease features****Mendelian disease genes, disease category and mode of inheritance**

Diagnostic grade ‘green’ genes with sufficient evidence for disease association and their corresponding modes of inheritance were obtained from GEL PanelApp, a publicly available knowledge base containing gene panels related to human disorders [9]. A total number of 313 gene panels (excluding additional findings) were investigated. Information on allelic requirement and level of evidence of disease causation was retrieved for our analysis. Genes from 186 gene panels containing level 2 disease category information (21 categories) were used for the analysis based on disease classification <https://PanelApp.genomicsengland.co.uk/panels/> [Downloaded 10.08.21].

**Human Phenotype Ontology annotations**

Phenotypes were obtained from the Human Phenotype Ontology (HPO) (genes to phenotypes) [46] and mapped to the top level of the ontology, broadly corresponding to the physiological system affected. Co-occurrence with the most frequent systems affected (neurological and musculoskeletal) were computed for early lethal genes (EL) versus non early lethal genes (NEL). <https://hpo.jax.org/app/download/annotation>; <https://raw.githubusercontent.com/obophenotype/human-phenotype-ontology/master/hp.obo> [Downloaded 23.08.21, HPO notes: format-version: 1.2 data-version: hp/releases/2021-08-02].

**Prenatal and perinatal lethal genes in humans**

A set of 624 genes associated to prenatal and perinatal lethality based on OMIM records obtained from Dawes et al. [6, 47] were used for the analysis. OMIM text fields across the database were queried through the API for terms associated with early lethality, before or shortly after birth. A total of 86 search terms were queried, including ‘early death’, ‘fetal death’, ‘lethal AND prenatal’, ‘lethal AND perinatal’, ‘lethal AND neonatal’ among others. The clinical descriptions for each of the initial hits were reviewed to exclude genes with no explicit evidence.

**Prediction of early lethal genes**

Several genes have undergone the IMPC primary viability assessment, but the embryonic stage at which lethality occurs has not yet been investigated. To increase the pool of potential candidate early lethal genes, we built a classifier using human cell proliferation scores from 902 lines as predictor variables. For that we used the R

implementation of Generalized Additive Model Selection, *gamselect* [48]. The training set consisted of 895 genes, 430 early-lethal (EL) and 465 non-early lethal (NEL). Imputation of missing values was performed via nuclear-norm regularisation implemented in the *softImpute* [49] R package. Cross validation (5-fold) ROC-AUCs and accuracy were computed to assess the performance of the model. A set of 33 genes externally assessed as EL [50] was used as additional validation (File S2 [33]).

**Gene similarity approach**

Similarity with known genes associated to biallelic forms of inherited metabolic disorders (biallelic inborn error of metabolism green genes from PanelApp, BIEM) was assessed according to 5 attributes (5ps): (p1) being a paralogue of a known BIEM gene according to Ensembl genes 104 and a threshold of % amino acid identity of 20% [41]; (p2) sharing a Reactome pathway (lowest level) with a BIEM gene [43]; (p3) belonging to the same Corum protein complex of a BIEM gene [45]; (p4) being a direct interactor within the protein-protein interaction network (high confidence cut-off 0.7) of a BIEM gene according to STRING [42]; and (p5) sharing a PFAM protein family with a BIEM gene [44]. The number of features shared was computed for every early lethal gene—assessed and predicted (File S3 [33]).

**Investigation of cases from the 100KGP**

To investigate the occurrence and enrichment of homozygous LoF variants in cases from the 100KGP among our set of EL genes in the mouse, we searched for variants in those genes in 35,422 families, 631 of which were recruited under the categories of interest (‘undiagnosed metabolic disorders’ and ‘mitochondrial disorders’). One important caveat is that these are not healthy population controls, and we cannot rule out that patients recruited under other categories show similar metabolic phenotypes, which means that these ratios can be an underestimation. The number of observed homozygous LoF and missense variants prioritised by Exomiser based on variant scores [20] were compared between cases and *pseudo* controls to compute observed versus expected ratios (File S4) [33].

**Statistical analysis and software**

R software [51] including the following packages were used for data integration and analysis: *tidyverse* [52], *matrixStats* [53], *epitools* [54], *DescTools* [55], *oddsratio* [56]; data visualisation: *waffle* [57], *ggridges* [58], *alluvial* [59], *cowplot* [60], *upSetR* [61]; ontologies: *ontologyIndex*



[62]; modelling and prediction: *softImpute* [49], *gamsel* [48], *pROC* [63]. To test for significant differences in the proportions of cellular essential genes, genes with no paralogues, paralogues properties, Mendelian disease genes, modes of inheritance and disease categories across the 3 WoL and to perform pairwise comparisons, Pearson's chi-squared (two-sided) was used as implemented in *prop.test* and *pairwise.prop.test* functions. In the case of continuous variables, to test for significant differences between the three WoL, we used the non-parametric Kruskal-Wallis test to compare the three groups and post hoc Dunn's test for pairwise comparisons (two-sided) using their R implementations: *kruskal.test* and *dunnTest* functions. The null hypotheses being that the distribution of the CERES depletion scores for the different tissues, the levels of gene expression across development and several intolerance to variation metrics is the same across WoL. For each cell lineage, all the gene individual scores were used to assess statistical significance. 95% CI for the median score for each window and cell line were computed using the exact method implemented in the *MedianCI* function in the *DescTools* R package. The thresholds for statistical significance after multiple testing corrections are specified in Additional file 1: Tables S1-S4. Odds ratios (OR) were calculated by unconditional maximum likelihood estimation (Wald) and confidence intervals (CI) using the normal approximation, with the corresponding adjusted *P* values (Benjamini-Hochberg, BH) for the test of independence using the *oddsratio* function (Additional file 1: Table S5). To evaluate the performance of our approach to identify candidate genes, *F*-scores were computed for our strategy based on EL genes and alternative ones based on pRec, DOMINO, SCoNeS and LOEUF scores. Precision and recall were estimated based on the number of predicted recessive genes using the suggested thresholds for the different scores and the number of BIEM genes in each of these sets of candidate genes. A multiple logistic regression model was fitted using EL and these other metrics as predictors of BIEM genes and the ORs associated with each predictor for specific increment steps were estimated as implemented in the *or\_glm* function (Additional file 1: Tables S6-S7).

## Results

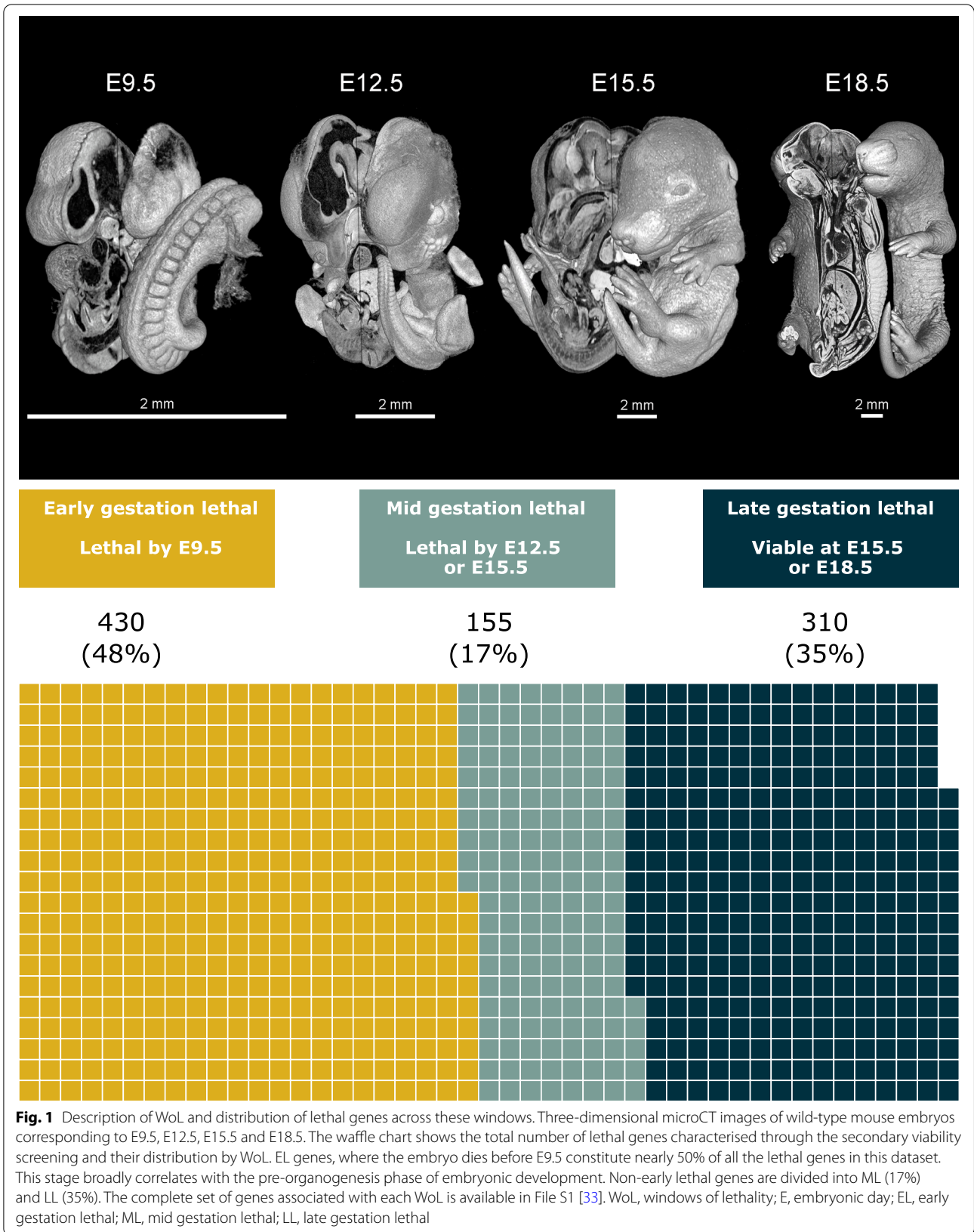
### Gaining functional knowledge from WoL

The IMPC measures viability between 14 days of age and weaning and, for lethal strains, employs a high-throughput embryonic phenotyping pipeline to examine embryonic viability and phenotypes at embryonic day (E) E9.5, E12.5, E15.5 and/or E18.5. The developmental period during which lethality occurs in the mouse can be summarised by establishing a set of WoL. A WoL for a gene was defined by the interval between the latest developmental

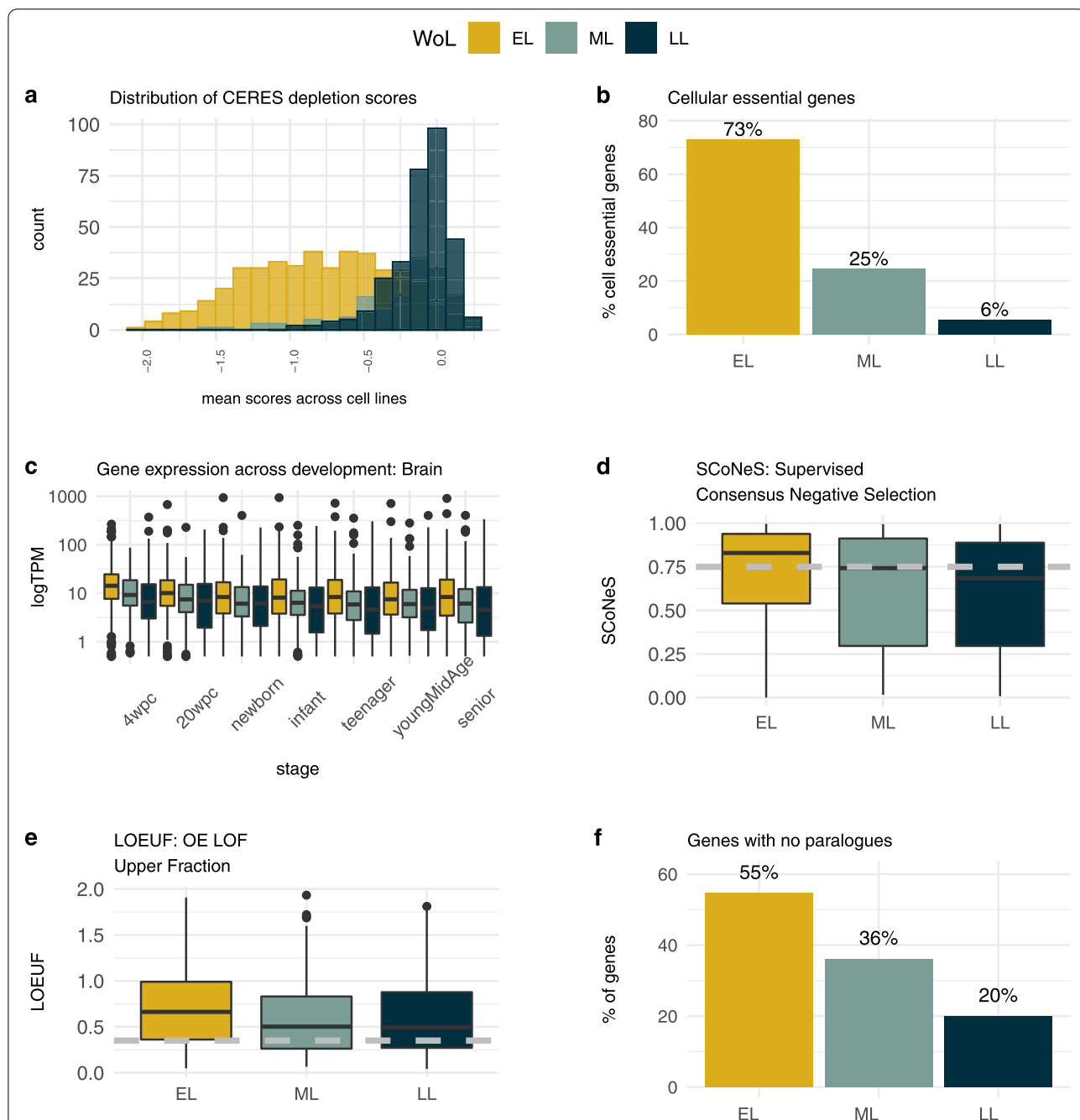
stage at which live homozygous null embryos (mice) are identified and the earliest stage at which no live homozygous embryos are found [24]. Complete lethality by E9.5 was classified as early-gestation lethal (EL), by E12.5 or E15.5 as mid-gestation lethal (ML), and viability at E15.5 or E18.5 as late-gestation lethal (LL). These WoL approximately correlate with the pre-organogenesis, organogenesis and post-organogenesis phases of mouse embryonic development, while also providing sufficient sample sizes to perform downstream statistical analyses. Among 895 embryonic lethal genes with one-to-one human orthologues assessed in the IMPC to date, nearly half (430, 48%) are EL, 155 (17%) ML, and 310 (35%) are LL. A full description of the WoL is available (File S1 [33]) and the distribution of lines per window can be found in Fig. 1.

### Human cellular essential genes correlate with mouse EL genes

We previously reported that EL genes show a considerable overlap with human cellular essential genes [25]. The CERES dependency scores obtained from CRISPR knockout screens through the Achilles pipeline [35] compute the depletion effect on cell proliferation. A lower and more negative value is the result of greater depletion of cancer cells upon genetic perturbation and indicates higher essentiality [64]. Plotting median proliferation scores and the corresponding 95% CI of genes for different human cell lines across tissues, we observed a clear distinction between the three WoL. The set of EL genes stands alone as a distinctive category from the ML and LL genes that have closer median values (Additional file 2: Fig. S1). The differences in score distribution are consistent and statistically significant across cell lineages (*P* value < 2.2e−50), with a few exceptions when comparing ML and LL sets (Additional file 1: Table S1). Considering the average CERES score across 902 cell lines, we observed that only EL genes are found in the bins with lowest scores, and that the percentage of ML and LL genes within bins increased with higher values of this score (Fig. 2a, Additional file 2: Fig. S2a). When cellular essentiality is considered as a binary property after categorising the mean scores using a cut-off of −0.45 (≤ −0.45: 'cellular essential', >−0.45: 'cellular non-essential'; see the 'Methods' section), 73% of EL genes are essential in human cell lines, compared to 25% of ML genes and only 6% of LL genes (*P* value < 2.2e−50) (Fig. 2b, Additional file 1: Table S1). Alternative thresholds are considered in Additional file 2: Fig. S2b-2c and show a similar enrichment. Cell line essentiality was previously explored for mouse viable genes and showed that > 99% are non-essential in human cell lines [25]. We additionally examined individual cell lines to discard any potential cell line specific effect, and the percentage of EL genes found







**Fig. 2** WoL and gene features. **a** Distribution of mean CERES depletion scores. Histograms represent the probability distribution of mean CERES scores across cell lines for each WoL. **b** WoL and cellular essential genes. Percentage of EL, ML and LL genes considered cellular essential when a mean CERES depletion score across cell lines of  $-0.45$  is considered as threshold. **c** Gene expression in brain. Boxplots show the distribution of human gene expression values for genes within each WoL across selected developmental stages for human brain. **d** SCoNeS scores. Boxplots show the distribution of SCoNeS scores, the predicted probability of a given gene being AR. The dashed grey line represents a threshold ( $SCoNeS > 0.75$ ) used to identify genes underlying AR disorders. **e** LOEUF scores. Boxplots show the distribution of LOEUF scores across WoL. Low LOEUF scores indicate strong selection against predicted loss-of-function (pLoF) variation in a gene. The dashed grey line represents a threshold ( $LOEUF < 0.35$ ) used to identify genes that are constrained against pLoF variation. **f** WoL and paralogues. Barplots represent the percentage of genes with no paralogues (singletons) across WoL, with the proportion of genes with no duplicates decreasing across development stages. Tests for differences between WoL available in Additional file 1: Table S1-S3. For plots **a-f**, the data shown correspond to gene annotations for the human orthologues. WoL, windows of lethality; EL, early gestation lethal; ML, mid gestation lethal; LL, late gestation lethal; LOEUF, LoF observed/expected upper bound fraction; SCoNeS, supervised consensus negative selection; AR, autosomal recessive

to be essential in each cell line based on this threshold ranges from 58 to 79% with a mean value of 72% (ML mean 24%, range 15–34%; LL mean 9%, range 5–25%).

#### ***EL genes consistently show higher levels of human gene expression across developmental stages***

Examination of human gene expression data [36] showed a consistent pattern of expression in brain across developmental stages with the human orthologues of mouse EL genes being expressed at higher levels, on average, compared to the orthologues of mouse ML and LL genes, and with the differences in gene expression between EL and LL genes being statistically significant across most developmental stages (Fig. 2c, Additional file 2: Fig. S3a, Additional file 1: Table S2). A similar pattern was observed for other organs with data available, including cerebellum, heart, kidney, liver, ovary and testis (data not shown). High levels of expression may help identify key developmental processes. To that end, gene expression patterns during early human development have been used to predict essential genes lacking a known human disease association [65]. To assess whether the organ development trajectories for these genes differ substantially between mouse and human, we investigated the similarity of spatiotemporal gene expression profiles for the two species. We found that 78 and 82% of the entire set of genes studied showed the same trajectory for cerebellum and brain respectively, with no significant differences observed between WoL and in concordance with what was observed for the entire set of genes with data available [37] (Additional file 2: Fig. S3b). Similarities in gene expression do not always imply conserved phenotypes between mouse and human, but can serve as a proxy for how translatable the findings for these genes are to human disease.

#### ***Intolerance to LoF variation differs across WoL***

EL genes are more likely to underlie an AR condition, based on higher Supervised Consensus Negative Selection scores (SCoNeS) [39], a metric that estimates the predicted probability of a gene being AR, particularly when compared to LL genes (unadjusted  $P$  value =  $5.45e-07$ ; Fig. 2d). When the LoF observed/expected upper bound fraction (LOEUF) [13], a quantitative measure of the observed depletion of LoF variation compared to a null mutational model, was investigated, we observed an inverted pattern, with EL genes showing higher mean values of this score (weaker selection against predicted LoF variants) compared to ML and LL genes (unadjusted  $P$  values  $6.70e-03$  and  $2.69e-04$  respectively; Fig. 2e). Albeit only nominally statistically significant, this observation agrees with our previous findings that developmental lethal genes, those genes that are not essential

for cell survival but required for organism development, and that broadly correlate with ML or LL genes, are more intolerant to heterozygous LoF variation compared to cellular lethal genes, those found to be essential in human cell lines and lethal in the mouse, and more likely to be EL [25]. Additional constraint metrics were explored, including pLI and pRec [13], RVIS [40] and DOMINO [38] (Additional file 2: Fig. S3c–3f). DOMINO scores represent a gene level metric based on a machine learning approach that extracts discriminant information from a broad set of features and computes the probability for a gene to carry dominant mutations. Based on this measurement, EL genes were also more likely to be linked to AR disease compared to LL genes (unadjusted  $P$  value =  $3.39e-05$ ; Additional file 2: Fig. S3g). The results for the statistical tests of significance are shown in Additional file 1: Table S3.

#### ***Gene duplicates and time of duplication event are distinctive features of EL genes***

EL genes have the highest proportion of genes with no paralogues (singletons). This proportion decreases gradually from ML to LL genes (unadjusted  $P$  value =  $1.41e-20$ ; Fig. 2f). Not only are EL genes more likely to be singletons, but also, for those genes that do have paralogues, the number of paralogues is lower and the paralogues are more likely to be older, with longer times since the duplication event when compared to ML or LL genes, which suggests more time to evolve new functions (Additional file 2: Fig. S4a, S4b). Thus, not only do gene duplications, or the lack thereof, seem to play a role in essentiality but so do the number of paralogues and the time of the duplication event. Similar observations were made by others using different species and/or definitions of essentiality [66, 67]. Paralogues of EL genes are also more likely to be EL, and similarly paralogues of ML/LL genes are more likely to be ML/LL (Additional file 2: Fig. S4c). This implies that paralogues are predominantly essential at the same developmental stage, potentially reflecting similar key functions at the cellular level and early stages of organism development. The differences in all these metrics are statistically significant when comparing EL vs LL genes (Additional file 1: Table S3). Additionally, by dividing genes into singletons and duplicates, we explored the proportion of genes that are cellular essential among these two sets of genes for the three WoL (Additional file 2: Fig. S4d). Previous studies investigating the relationship between essentiality, developmental expression and gene duplication have suggested that timing of developmental expression influences the ability of a gene in a paralogue pair to compensate for the loss of function of the other gene [68].

### WoL and Mendelian disease

It is well established that there is an association between lethal genes in the mouse and human disease genes [24, 47]. Our previous study showed that this enrichment was mainly driven by developmental lethal genes [25] so we hypothesised that the distribution of disease genes across WoL may not be uniform and that information about WoL could reveal additional correlations. When translating our WoL to relevant developmental stages in humans, the EL mouse category broadly correlates with the human pre-organogenesis stage occurring during the first 2 weeks of development. The ML class relates to human organogenesis occurring during the embryonic period from weeks 3 through 8, and ending in the first trimester, around week 9 of gestation. Lastly, the LL category aligns with the human foetal stage, from week 9 until birth [69].

We used PanelApp as the source of Mendelian genes to perform subsequent analyses [9]. Genes are rated according to level of evidence to support the phenotype association: ‘green’ means high level of evidence from several unrelated families and/or strong additional functional data, ‘amber’ moderate evidence and ‘red’ not enough evidence. The advantages of using this source of diagnostic genes include the high-level disease categorisation and allelic requirement annotations that allows for tailored analysis, the categorisation of genes according to the level of evidence for the gene-disease association and the potential to map directly to patient data recruited in the 100KGP.

### Disease category and mode of inheritance are not uniformly distributed across WoL

Although the three WoL are all enriched for Mendelian disease genes, their properties differ. The proportion of genes associated with rare disorders is lowest among the EL, followed by the ML and LL genes (Fig. 3a). When allelic requirement is considered, this trend is reversed for AR disorder-associated genes, where the EL fraction showed a significantly higher number of biallelic genes compared to LL genes (unadjusted  $P$  value =  $5.16e-06$ ;

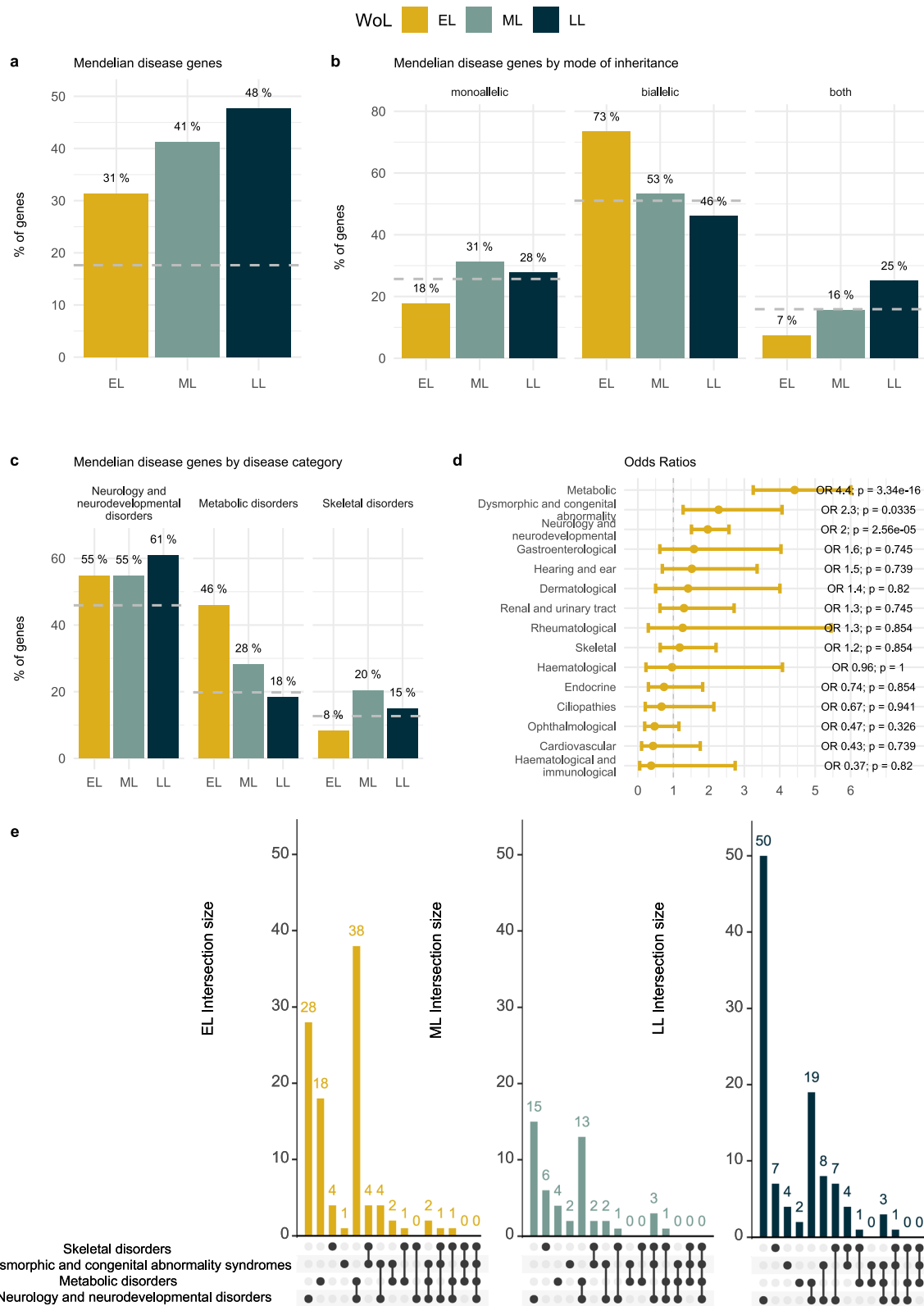
Fig. 3b; results for the statistical tests of significance in Additional file 1: Table S4).

Further dissection of disease genes according to PanelApp high level disease categories showed that (1) the proportion of neurodevelopmental disorder associated genes is higher than expected among the three WoL compared to baseline, with the highest percentage among LL genes; (2) the proportion of genes associated to metabolic disorders follows the inverse pattern, with EL genes showing the highest percentage of inherited metabolic disease genes (46%), followed by ML (28%) and showing the lowest percentage among the LL (18%) (unadjusted  $P$  value =  $2.7e-06$ ); most notably, this is the only disease category with a higher percentage of disease genes among the EL compared to ML and LL genes; (3) a higher percentage of skeletal disorder genes is found in ML set, although this association is only nominally significant; and (4) for the remaining disease categories, the frequency of disease genes among the EL genes shows values comparable to baseline or even lower, indicative of depletion of these disease categories among the EL genes (Fig. 3c, Additional file 2: Fig. S5a, Additional file 1: Table S4). In order to assess the strength of the association between EL genes and the different disease categories, OR were computed using the entire set of non-EL genes, i.e. all those genes with IMPC data on viability, including ML, LL, subviable and viable categories (see the ‘Methods’ section). Three disease categories showed a positive association (with a lower bound of the 95% CI for the OR > 1): metabolic disorders (OR = 4.4; adjusted  $P$  value =  $3.34e-16$ ), dysmorphic and congenital abnormality syndromes (OR = 2.3; adjusted  $P$  value = 0.034) and neurology and neurodevelopmental disorders (OR = 2; adjusted  $P$  value =  $2.56e-05$ ) (Fig. 3d).

Given that most inborn errors of metabolism (IEM) show neurological manifestations, and neurodevelopmental disorders are still the most predominant disease category across the three WoL, we further explored the gene overlap between neurodevelopmental and metabolic disease categories to assess any potential confounding effect. The combination of genes associated with both metabolic and neurodevelopmental disorders was found

(See figure on next page.)

**Fig. 3** WoL and human disease. **a** Mendelian disease genes. Barplots represent the percentage of rare disease associated genes in each WoL according to PanelApp, only ‘green’ genes with a high level of evidence for the gene-disease association were included. **b** Mode of inheritance. Barplots represent the percentage of Mendelian genes by associated allelic requirement across WoL, only monoallelic or biallelic genes were included. **c** Disease category. Mendelian genes by disease type according to PanelApp level 2 disease categories, with the bars indicating the percentage of PanelApp genes mapping each disease class for the 3 WoL. For plots **a–c**, the dashed grey line represents the baseline percentage for the entire set of protein coding genes (19,197 genes according to HGNC, **a**) or PanelApp ‘green’ genes (3384 genes, **b, c**). **d** Disease categories OR and BH adjusted  $P$  values for EL genes compared to ANEL genes: this included mid and late gestation lethal genes as well as subviable and viable categories. **e** Disease category overlap. Overlap between genes associated with the most frequent disease categories across WoL for EL, ML and LL genes respectively. Tests for differences between WoL are available in Additional file 1: Table S4. WoL, windows of lethality; EL, early gestation lethal; ML, mid gestation lethal; LL, late gestation lethal; HGNC, HUGO Gene Nomenclature Committee; ANEL, all non-early gestation lethal genes; OR, odds ratio; BH, Benjamini-Hochberg



**Fig. 3** (See legend on previous page.)

to be predominant among the EL class, opposite to what we observed among the ML and LL classes, where neurodevelopmental only genes are the prevalent disease class, thus providing additional evidence for the IEM association with EL genes (Fig. 3e).

The analysis of HPO phenotypes associated with known inborn error of metabolism genes showed that the five most frequent physiological systems affected are nervous system, followed by musculoskeletal, metabolism/homeostasis, growth abnormality, and digestive. An enrichment analysis showed no significant differences in the frequency of any of these phenotypes among EL genes when compared to ML and LL genes (Additional file 2: Fig. S5b, Additional file 1: Table S5).

#### **Evidence of prenatal and perinatal lethality in humans**

Among the wide range of Mendelian phenotypes observed in humans, prenatal lethality poses a unique challenge in terms of providing a molecular diagnosis. Development failure may occur at any point between fertilisation and birth. Estimates suggest that 20–30% of implanted embryos fail to develop beyond week 6 [70]; similarly early embryo losses occurring between implantation and clinical recognition could be around 10–25% [71]. A proportion of first trimester miscarriages where no chromosomal abnormalities are detected could have a Mendelian or polygenic origin [72, 73].

We previously hypothesised that many human genes contributing to prenatal lethality are likely unidentified and not captured in current disease databases due to early embryo losses and miscarriages either being unnoticed, or when they are detected, the difficulty in determining the molecular basis of this extreme phenotype. Here, we used a set of 624 genes associated with early lethality in humans curated from OMIM [6, 47]. We found that 19% of EL disease-associated genes are linked to pre- and perinatal lethality. For LL genes, this percentage is 31% (Additional file 1: Fig. S5c). Based on our hypothesis that most genes associated to early-gestation lethality in humans remain unrecognised, the set of EL genes in the mouse constitutes a source of candidates of interest in the field of foetal precision medicine.

#### **Predicting new EL genes in the mouse**

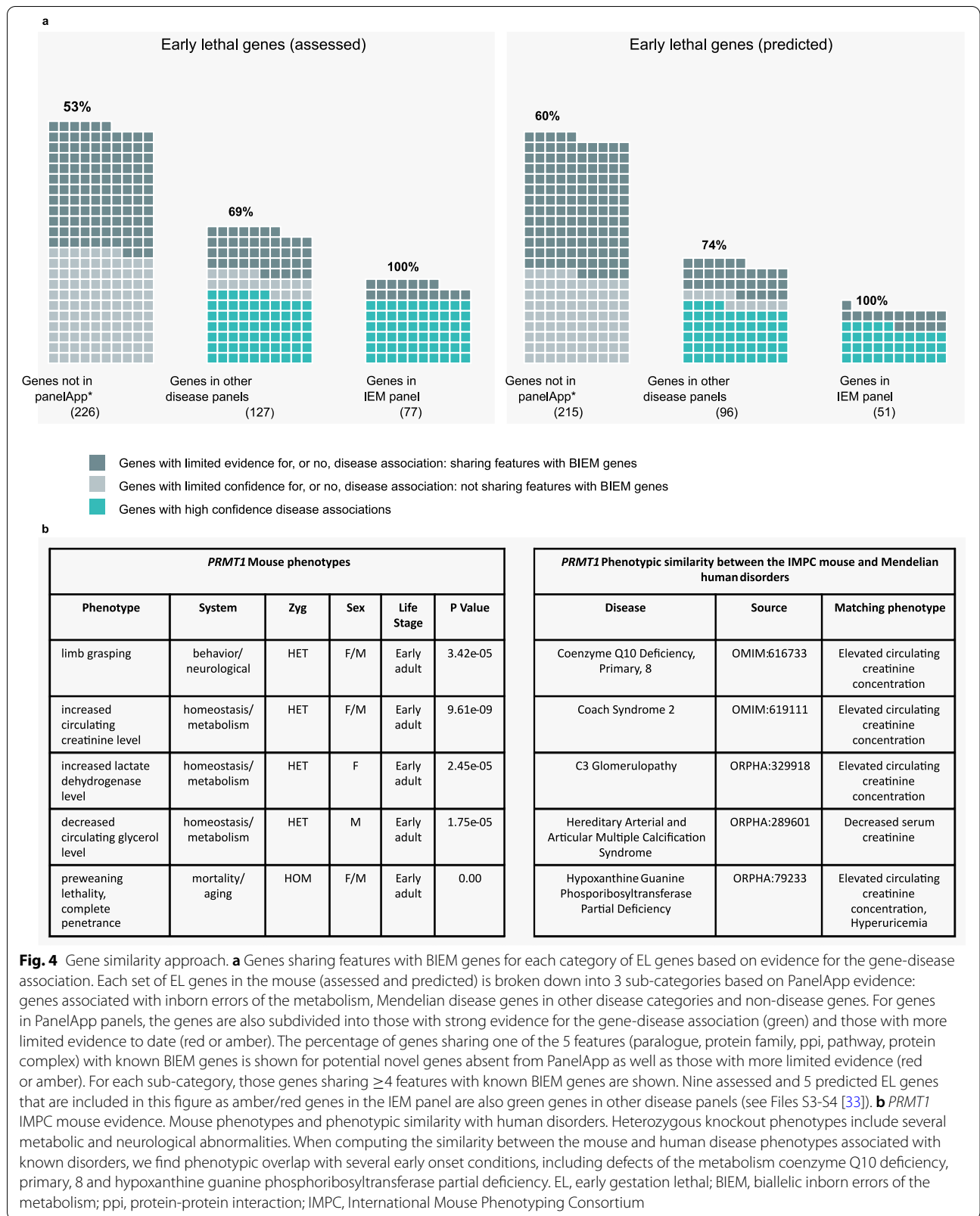
Since the number of IMPC mouse lines that have undergone the primary viability assessment is higher than those with a secondary evaluation to identify the embryonic stage at which lethality occurs, we tried to predict additional EL genes among lethal genes without secondary viability data to have a larger pool of candidate genes. For this, we used a penalised likelihood approach to fit a generalised additive model using proliferation (essentiality) scores from multiple human cell lines as predictors

[35] and subsequently used that model to make the predictions. This added a further set of 362 predicted EL genes (out of 725 lethal genes with no secondary viability assessment) to the previous 430 EL genes assessed through embryo viability screening. Details on the model, predictive accuracy, and predictor variables are described in the ‘Methods’ section and Additional file 2: Fig. S6. Of 33 genes in our prediction set that were externally assessed as EL [50], 29 were correctly predicted by the classifier (87.9%) [33]. CRISPR knockout screens to identify those genes affecting cell survival across hundreds of genomically characterised cancer cell lines [74] can consequently assist with the identification of early-gestation lethal lines in the mouse.

#### **Similarity with known BIEM genes**

A gene similarity strategy was applied to 792 (assessed and predicted) human orthologues to mouse EL genes based on features shared with 552 diagnostic-grade BIEM genes from PanelApp. This approach was based on the unknown gene sharing at least one of 5 attributes: (p1) being a paralogue of a known BIEM gene; (p2) sharing a pathway with a BIEM gene; (p3) belonging to the same protein complex as a known BIEM gene; (p4) interacting with a known BIEM gene; and/or (p5) sharing a PFAM protein family with a known BIEM gene. This gene ranking approach served a dual purpose: (1) to identify completely novel disease genes and (2) to bring additional proof for those genes in PanelApp that are not considered diagnostic-grade genes, i.e. ‘amber’ and ‘red’ genes. Among novel EL genes not associated with any disease in PanelApp, 53–60% share at least one of the above five attributes with a BIEM gene. This percentage increases to 69–74% when the non-diagnostic-grade genes in PanelApp excluding the IEM panel are examined and to 100% for the non-diagnostic-grade genes on the IEM panel (Fig. 4a).

Ten of the EL non-disease-associated genes are of particular interest as they share 4 of the 5 attributes with BIEM genes: *CHKA*, *FDX1*, *GGPS1*, *GLRX3*, *HMGCS1*, *MGATI* and *SLC39A10* are paralogous and direct interactors as well as belonging to the same protein family(ies) and pathway(s) while *MRPS25*, *PRMT1* and *RPA1* are interactors, share a protein family(ies) and pathway(s) and are also part of the same protein complex(es). The complete gene list and annotations are provided in [33]. Four of these genes, *Ggps1*, *Mrps25*, *Prmt1* and *Rpa1*, show abnormal metabolic phenotypes in the heterozygous viable mouse [31]. *MRPS25* is a member of the human mitochondrial ribosomal protein gene family, with evidence from mouse embryos indicating compromised mitochondrial function [75]. Several other mitochondrial ribosomal small (MRPS) and large (MRPL)





subunit genes are associated with different metabolic disorders, and many of the remaining MRPS genes are also potentially associated with disease [76]. Evidence of pathogenicity of homozygous missense variants in this gene has been reported [77]. In the case of *PRMT1*, encoding a member of the protein arginine N-methyltransferase (PRMT) family, additional neurological phenotypes found in the IMPC knockout of the orthologous *Prmt1* imply a high phenotypic similarity with neonatal disorders including several defects of the metabolism as computed by PhenoDigm [32] (Fig. 4b). Emerging evidence supports the role of this family of enzymes in skeletal muscle and metabolic disease [78].

To evaluate this approach, and whether EL genes not associated with Mendelian disorders are more likely to share attributes with BIEM genes compared to non-EL and non-disease associated genes, we computed the ORs to obtain a measure of this association. Importantly, we found a significant association between sharing any of these 5 attributes with a BIEM gene and being EL (1.64 fold-increase, adjusted  $P$  value =  $2.7e-06$ ). When these attributes were considered separately, the strongest association was observed for being part of the same protein complex as a BIEM gene (13.9 fold-increase, adjusted  $P$  value =  $6.5e-20$ ). Significant results were also obtained for sharing a pathway and interacting with a BIEM gene. EL genes were less likely to be a paralogue of a BIEM gene (OR = 0.49, adjusted  $P$  value = 0.018), which can be explained by the enrichment for singletons among this set of genes (Additional file 2: Fig. S7).

Disaggregating the set of EL genes by disease association showed that the closer to the IEM disease class, the higher the percentage of genes in that category sharing attributes with BIEM genes. Consistently, EL genes are more likely to share attributes with BIEM genes compared to non-EL genes.

#### **Undiagnosed cases of inherited metabolic disorders from the 100KGP**

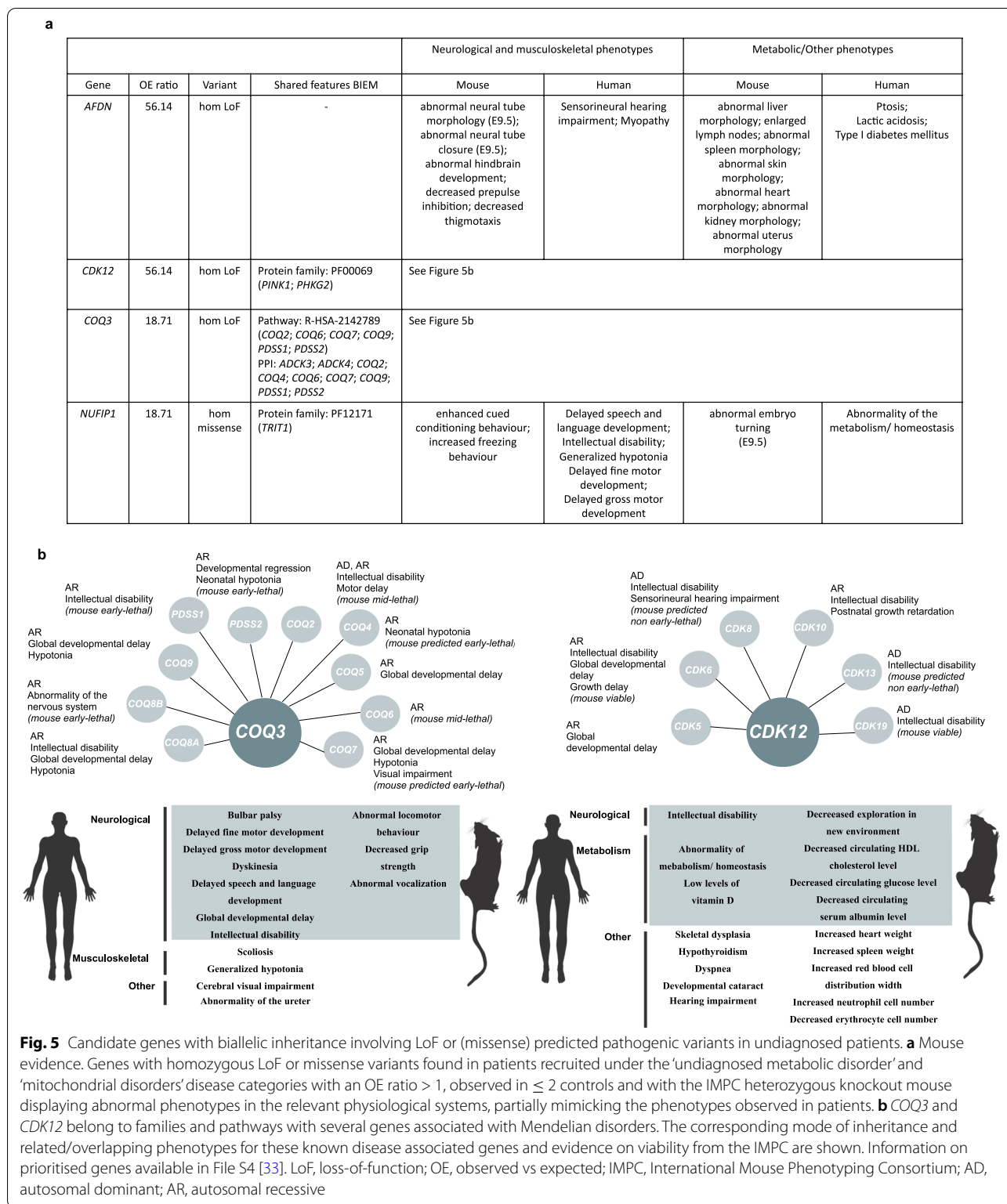
An alternative approach, based on patient data, was also used to identify potential metabolic disease genes among the set of EL genes in the mouse. Cases recruited under the ‘undiagnosed metabolic disorder’ and ‘mitochondrial disorders’ categories in the 100KGP were investigated for rare, segregating and biallelic LoF or predicted pathogenic missense variants in EL genes, using the Exomiser variant prioritisation tool [20]. Observed versus expected (OE) ratios per gene were computed by comparing the number of biallelic variants observed in these patients to those observed on a set of *pseudo controls*, i.e., patients recruited under other disease categories. Predicted homozygous or compound heterozygous pathogenic variants were found in 21 EL genes (13

assessed, 8 predicted) with OE ratios > 1 and observed in  $\leq 2$  controls. None of the 21 genes showed enrichment of synonymous variants by these same criteria. Out of the 21 genes, 3 involved biallelic LoF, 6 had biallelic LoF/missense and 12 had biallelic missense variants. Five of these genes are already classified as diagnostic grade genes in the IEM panel (*COQ4*, *ELAC2*, *MRPL44*, *MSTO1* and *SKIV2L*) and three others are diagnostic grade genes in different neurology and neurodevelopmental disorder gene panels (*EIF2B4*, *ELP1*, *EXOSC8*). *ALG2*, *NDUFA8* and *RNASEH2A* are classified as amber or red in the IEM panel. For the cases associated with these 11 known disease genes, only those associated with *MRPL44* and *ALG2* biallelic variants have been diagnosed with these variants so far, with the others currently classified as variants of uncertain significance. For the remaining 10 genes (*AFDN*, *CDK12*, *COQ3*, *GINS4*, *GPATCH1*, *INTS11*, *KIF2C*, *NUFIP1*, *PTPMT1*, *RCC1*), there is no current evidence for a disease association in PanelApp or OMIM. The complete set of genes is provided in File S4 [33].

For two of the amber or red genes in the IEM panel, *ALG2* and *NDUFA8*, IMPC heterozygous knockout mice have neurological and metabolic phenotypes [31], providing additional evidence to validate this gene-disease association. In addition, *ALG2* shares 4 features with known BIEM genes: protein family (2 genes), pathway (10 genes), paralogue (1 gene) and protein-protein interaction (9 genes). Similarly, *NDUFA8* shares 3 features: protein complex (17 genes), pathways (44 genes) and protein-protein interaction (28 genes).

Four non-disease-associated genes have IMPC data for null alleles with heterozygous mouse mimicking some of the clinical features observed in patients. *AFDN* and *NUFIP1* show neurological phenotypes in the orthologous mouse embryo or early adult [31, 32]. *COQ3* and *CDK12* also show neurological and other physiological system phenotypes [31, 32] shared between the undiagnosed patients and the knockout mouse. Detailed information on the phenotypes observed in the patients is shown in Fig. 5a, b. They are of particular interest as several other genes from the same family have already been associated with similar disorders, and the IMPC lines are the first reported mouse models with abnormal phenotypes observed in the early adult heterozygous knockout [79].

*COQ3* (coenzyme Q3, methyltransferase) is one of the genes required for the biosynthesis of Coenzyme Q10, which has many vital functions. Several genes involved in this pathway are associated with Primary CoQ10 Deficiency, including *PDSS1*, *PDSS2*, *COQ2*, *COQ4*, *COQ5*, *COQ6*, *COQ7*, *COQ8A*, *COQ8B* and *COQ9* [80]. The heterozygous *Coq3* IMPC mouse shows several neurological/behavioural phenotypes including abnormal



locomotor behaviour, abnormal vocalisation and decreased grip strength. No homozygous LoF variants have been observed for this gene according to gnomAD

(pLI = 0; pRec = 0.283; DOMINO = very likely recessive). The homozygous frameshift variant observed in the 100KGP cohort is present in gnomADv2.1.1

(p.Lys366SerfsTer2), with an allele frequency of  $6.04e-04$  but with no homozygous individuals for that allele. The OE ratio in our 100KGP study cohort is 18.7, with the other two different variants found in the set of *pseudo controls* recruited under the ‘unexplained sudden death in the young’ and ‘ultra-rare undescribed monogenic disorders’.

*CDK12* (cyclin dependent kinase 12) is one of the cyclin-dependent kinases with a key role in molecular processes relevant during development. Several other protein kinases are involved in developmental disorders: *CDK5*, *CDK6*, *CDK8*, *CDK10*, *CDK13* and *CDK19* [81]. The phenotypic abnormalities observed in heterozygous *Cdk12* IMPC mice include cardiac, haematopoietic, metabolic (decreased circulating HDL cholesterol level) and neurological features (decreased exploration in new environment) (Fig. 5b). The homozygous splice acceptor variant (c.1047-2A>G) is present in gnomADv2.1.1, with an allele frequency of  $4.06e-4$  and one homozygote observed in the South Asian population. This gene is in fact predicted to be highly intolerant to heterozygous LoF variation (pLI = 1; pRec = 0; DOMINO = very likely dominant). The OE ratio computed with biallelic variants in our GEL study cohort for this gene is 56.14 with no variants meeting the criteria described found in controls.

A note of caution is needed when interpreting the impact of these two homozygous LoF variants in *COQ3* and *CDK12* identified in the 100KGP cohort due to their position on the transcript (near the end of the transcript and into a NAGNAG sequence, which may indicate a frame-restoring splice site, respectively), as indicated by gnomAD. Where available, data on gene expression across development for the aforementioned genes (*AFDN*, *NUFIP1*, *COQ3* and *CDK12*) confirmed similar developmental gene expression profiles across time points from early organogenesis to adulthood in brain and cerebellum between mouse and human, which supports the translatability of the findings in the knockout mouse for these genes [37].

## Discussion

Many predicted LoF variants identified in Mendelian disease sequencing studies are found in genes not previously associated with disease, making assessment of pathogenicity particularly challenging. High-throughput mouse standardised phenotyping screens including viability assessment contribute to acquiring new knowledge about orthologues of such genes with limited functional data [82, 83]. By also exploring correlations between abnormal phenotype(s) in the knockout mouse and disease features in the human orthologues, we were able to identify novel candidates for Mendelian conditions.

Previously, we developed a successful framework to prioritise gene candidates for neurodevelopmental disorders using mouse phenotyping data, with two of the top nine candidate genes, *VPS4A* and *SPTBN1*, having been recently validated. In both cases, a causal link has been found between heterozygous, predominantly de novo mutations and distinctive developmental syndromes [25–27]. Here we present another example of how the IMPC data resource can be combined with other sources of evidence to develop a tailored approach for disease-gene discovery and variant prioritisation to assist the diagnosis of inherited metabolic disorders.

The requirement of a gene for the survival of an organism, i.e. gene essentiality, can be disaggregated into more granular categories/WoL according to the embryonic period during which lethality occurs. In the present study, we show that these categories correlate with different gene features, including gene expression across development and intolerance to LoF variation. Higher levels of gene expression among cellular essential genes compared to non-essential genes have been previously reported across developmental stages [84]. Human embryonic gene expression data, integrated with other gene features has been used to identify essential genes, suggesting that gene-specific expression changes during early development could be particularly relevant [65]. Importantly, housekeeping genes, defined as those genes being stably expressed irrespective of tissue and developmental stage, are not necessarily essential, and the genes that are both essential and invariably expressed may differ across organisms [85]. Additionally, the distribution of singleton and duplicated genes across these WoL supports hypotheses about the ability of paralogues for functional compensation at the cellular level [86]. EL genes are more likely to be singletons, and when paralogues exist, they tend to have originated earlier, suggesting more time to evolve new and/or distinct functions [66, 67]. Parologue functional compensation is not a universal ability, and physical and functional dependencies of the paralogues could reduce their buffering capacity [87]. Studies of synthetic lethality between parologue pairs suggest which gene features may be associated with the ability to compensate for each other’s function [88].

By looking at different features of human orthologous disease genes across the WoL, two observations stand out. First, the set of lethal genes in the mouse is enriched for Mendelian disease genes [24], but the proportion of genes associated with disease is not consistent across WoL with this enrichment mainly driven by LL genes. The lower proportion of disease genes among the EL compared to LL genes was previously reported when comparing cellular lethal with developmental lethal genes [25], as well as other categorisations of essential

genes [47, 89]. Second, we identified a strong association between EL genes and inherited metabolic disorders. This includes genes that are needed to maintain the metabolic machinery required to provide energy and basic components for cell survival. Most of the EL lines die prior to implantation or gastrulation, and differentiation into disease-associated tissues occurs at a later stage. This could explain why non-metabolic disease categories are underrepresented among the set of EL genes.

Building on this finding, we focused on the EL genes and gathered additional information on similarity with known disease genes associated with BIEM disorders. It is already known that members of paralogous gene families where one gene is associated with human disease are more likely to be associated with Mendelian disorders themselves [90]. Similarly, disease-associated variants are enriched at sites conserved among paralogues [91, 92]. We used these and other observations to identify the EL genes showing most similarity to existing BIEM genes and, hence, most likely to be novel BIEM disease genes.

Inherited metabolic disorders comprise a large group of ~1450 disorders in which the primary alteration of a biochemical pathway leads to a set of biochemical, clinical and/or pathophysiological features [93]. The majority manifest in new-borns, show predominantly neurological manifestations and can lead to sudden premature death [94]. By investigating patients recruited under this disease category from the 100KGP and looking at human orthologues of EL genes in the mouse for evidence of enrichment of biallelic LoF or predicted pathogenic missense variants, we were able to identify a set of candidate genes where the heterozygous knockout mouse mimicked some neurological and/or metabolic phenotypes observed in patients.

Two of the genes identified through our analysis, *COQ3* and *CDK12*, belong to pathways and extended gene families associated with similar disorders, which strongly supports their involvement in the disease process. Further functional characterisation of these and other predicted pathogenic variants, together with the identification of additional probands with biallelic variants segregating with similar phenotypes, is still needed to establish a causal link, and to confirm that the candidate LoF variants result in the lack of protein product and/or have a discernible clinical phenotypic effect.

The approach described here is based on the premise that biallelic LoF in a gene leads to early embryonic lethality in mice but that biallelic LoF or missense variants in humans lead to recessively inherited metabolic disorders with related phenotypes in humans. In fact, for the four highlighted candidate genes identified in the GEL cohort, it is the heterozygous mouse model which is mimicking the phenotypes observed in patients carrying

biallelic mutations. This somehow counterintuitive observation has been reported for other IEM disorders [95, 96]. Most metabolic disorders represent a spectrum of phenotypes. According to OMIM clinical records, more than a third of BIEM genes are associated with lethality before or soon after birth, indicating that a considerable proportion of these conditions in humans are life threatening, leading to early death if untreated. And this proportion is likely an underestimation, given the limited sources of genes linked to prenatal and neonatal lethality in humans. Consistent with this observation, several genes in the same pathway or gene family of our candidate genes (*COQ2*, *COQ4*, *COQ9*, *PDSS2*) [97–100] have been associated with early lethality in humans.

Comparing lethality outcomes between mouse and human presents several limitations. Monoallelic mutations required for early development (dominant lethals) are missing from our set of mouse embryonic lethal knockouts since they would not result in lines, introducing a bias towards recessive lethal genes. Similarly, while in the mouse knockouts the observed phenotype is most likely due to the loss of protein function, other types of mutation may lead to different molecular mechanisms and thus different phenotypic outcomes. True loss of protein function in these genes may be early embryonic lethal in humans whereas postnatal phenotypes could be caused by hypomorphic variants leading to partial LoF [101, 102]. Other explanations include potential mechanisms of compensation through other genes in the pathway in humans or differences in essentiality between the two species. Given the number of genes associated with lethality in the mouse (35% of the knockout lines are classified as lethal or subviable according to IMPC primary viability screening) [24, 25], monogenic factors could explain a proportion of the high and often understated level of occurrence of miscarriages in human [72, 103]. This, together with the potential lack of molecular diagnosis for confirmed miscarriages, leads to an underestimation in current disease databases of embryonic lethality as a Mendelian phenotype [104]. Even when gene essentiality does not perfectly correlate between the two species, the set of lethal genes in the mouse provides knowledge on the molecular functions and biological processes [105] and constitutes an invaluable resource to identify relevant genes in humans, including those for which LoF variation may lead to pregnancy loss and other severe phenotypes with an early manifestation [47].

In summary, the embryonic stage at which lethality occurs in the mouse can be used to inform human disease. Several intolerance to variation scores inferred from human population sequencing data and a broad set of gene features estimate the predicted probability of a gene underlying AR conditions. Our target was a particular



subgroup of those genes, associated with BIEM disorders, and in this context our approach outperformed other potential strategies based on existing metrics (Additional file 1: Tables S6-S7). Integration of multi-species datasets and the extended use of standardised phenotypes is key to building novel Mendelian gene discovery approaches [3, 106]. This, coupled with the availability of data from large-scale sequencing programmes that allow for bespoke computational and statistical analysis for variant prioritisation, constitutes a powerful instrument for increasing the molecular diagnostic rate [17]. Additionally, the set of genes essential for embryonic development in the mouse may constitute an additional source of evidence for diagnosis of lethal foetal disorders [47, 107, 108]. Whether this is the only observable outcome or the most extreme phenotype within a wider range of clinical features observed in patients, it will be crucial to catalogue these genes. Several efforts are being made in this area. The foetal medicine community and ontologists are currently working to extend the HPO to cover the prenatal phenotypic manifestations of disease, and including data on the time course of these manifestations, including death will allow further comparisons between mouse and human phenotypes and discrimination between prenatal and postnatal phenotypes [109]. Additionally, we are collating all the information available from OMIM clinical records [6] and the literature to catalogue Mendelian disease genes into lethality categories.

## Conclusions

We have shown cross-species data integration and gene similarity approaches can complement other strategies to identify novel genes underlying Mendelian conditions. In particular, information on knockout mouse embryo lethality can be used to prioritise candidate genes associated with particular types of disorders. Access to unsolved cases from rare disease genome sequencing programmes allows the screening of those genes for potentially pathogenic variants that will hopefully lead to a diagnosis and potentially new treatment options.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-022-01118-7>.

**Additional file 1: Table S1.** Gene features: Human cellular essential genes. **Table S2.** Gene features: Gene expression in human brain. **Table S3.** Gene features: Intolerance to variation metrics and paralogues. **Table S4.** Disease features. **Table S5.** HPO phenotypes Odds Ratios. **Table S6.** Comparison of our approach based on EL genes with other strategies based on standard scores thresholds: F-score. **Table S7.** Odds Ratios and 95% CI from multiple logistic regression analysis.

**Additional file 2: Fig. S1.** WoL and cell essentiality scores. **Fig. S2.** WoL and cell essentiality categorisation. **Fig. S3.** WoL and additional gene features. **Fig. S4.** WoL and paralogues features. **Fig. S5.** WoL and additional disease features. **Fig. S6.** Prediction of early lethal genes. **Fig. S7.** Enrichment analysis of genes sharing attributes with a BIEM gene among the EL category.

## Acknowledgements

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project (<http://www.genomicsengland.co.uk>). The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support and we are grateful to both for making this available.

### Collaborators: International Mouse Phenotyping Consortium

John R. Seavitt<sup>5</sup>, Angelina Gaspero<sup>5</sup>, Uche Akoma<sup>4</sup>, Audrey Christiansen<sup>4</sup>, Sowmya Kalaga<sup>4</sup>, Lance C. Keith<sup>4</sup>, Melissa L. McElwee<sup>4</sup>, Leeyean Wong<sup>4</sup>, Tara Rasmussen<sup>4</sup>, Uma Ramamurthy<sup>4,14,15</sup>, Kiran Rajaya<sup>14</sup>, Panithee Charoenrattanakul<sup>14</sup>, Qing Fan-Lan<sup>6</sup>, Lauri G. Lintott<sup>6</sup>, Ozge Danisment<sup>6</sup>, Patricia Castellanos-Penton<sup>6</sup>, Daniel Archer<sup>12</sup>, Sara Johnson<sup>12</sup>, Zsombor Szoke-Kovacs<sup>12</sup>, Kevin A. Peterson<sup>11</sup>, Leslie O. Goodwin<sup>11</sup>, Ian C. Welsh<sup>11</sup>, Kristina J. Palmer<sup>11</sup>, Alana Luzzio<sup>11</sup>, Cynthia Carpenter<sup>11</sup>, Coleen Kane<sup>11</sup>, Jack Marcucci<sup>11</sup>, Matthew McKay<sup>11</sup>, Crystal Burke<sup>11</sup>, Audrie Seluke<sup>11</sup>, Rachel Urban<sup>11</sup>

<sup>14</sup> Office of Research Information Technology, Baylor College of Medicine, Houston, TX, USA

<sup>15</sup> Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA

### Collaborators: Genomics England Research Consortium

John C. Ambrose<sup>16</sup>, Prabhu Arumugam<sup>16</sup>, Roel Bevers<sup>16</sup>, Marta Bleda<sup>16</sup>, Freya Boardman-Pretty<sup>1,16</sup>, Christopher R. Boustred<sup>16</sup>, Helen Brittain<sup>16</sup>, Matthew A. Brown, Mark J. Caulfield<sup>1,16</sup>, Georgia C. Chan<sup>16</sup>, Greg Elgar<sup>1,16</sup>, Adam Giess<sup>16</sup>, John N. Griffin<sup>16</sup>, Angela Hamblin<sup>16</sup>, Shirley Henderson<sup>1,16</sup>, Tim J. P. Hubbard<sup>16</sup>, Rob Jackson<sup>16</sup>, Louise J. Jones<sup>1,16</sup>, Dalia Kasperaviciute<sup>1,16</sup>, Melis Kayikci<sup>16</sup>, Athanasios Kousathanas<sup>16</sup>, Lea Lahnstein<sup>16</sup>, Sarah E. A. Leigh<sup>16</sup>, Ivonne U. S. Leong<sup>16</sup>, Javier F. Lopez<sup>16</sup>, Fiona Maleady-Crowe<sup>16</sup>, Meriel McEntagart<sup>16</sup>, Federico Minneci<sup>16</sup>, Jonathan Mitchell<sup>16</sup>, Loukas Moutsianas<sup>1,16</sup>, Michael Mueller<sup>1,16</sup>, Nirupa Murugaesu<sup>16</sup>, Anna C. Need<sup>1,16</sup>, Peter O'Donovan<sup>16</sup>, Chris A. Odhams<sup>16</sup>, Christine Patch<sup>1,16</sup>, Mariana Buongiorno Pereira<sup>16</sup>, Daniel Perez-Gil<sup>16</sup>, John Pullinger<sup>16</sup>, Tahrima Rahim<sup>16</sup>, Augusto Rendon<sup>16</sup>, Tim Rogers<sup>16</sup>, Kevin Savage<sup>16</sup>, Kushmita Sawant<sup>16</sup>, Richard H. Scott<sup>16</sup>, Afshan Siddiq<sup>16</sup>, Alexander Sieghart<sup>16</sup>, Samuel C. Smith<sup>16</sup>, Alona Sosinsky<sup>1,16</sup>, Alexander Stuckey<sup>16</sup>, Mélanie Tanguy<sup>16</sup>, Ana Lisa Taylor Tavares<sup>16</sup>, Ellen R. A. Thomas<sup>1,16</sup>, Simon R. Thompson<sup>16</sup>, Arianna Tucci<sup>1,16</sup>, Matthew J. Welland<sup>16</sup>, Eleanor Williams<sup>16</sup>, Katarzyna Witkowska<sup>1,16</sup>, Suzanne M. Wood<sup>1,16</sup>, Magdalena Zarowiecki<sup>16</sup>

<sup>16</sup>Genomics England, London, UK

## Authors' contributions

PC. and D.S. contributed to the conceptualisation, data analysis, presentation and interpretation of the results and writing the manuscript. C.H.W. contributed to the data analysis, interpretation, reviewing and editing the manuscript. J.M. contributed to the data generation, interpretation, reviewing and editing the manuscript. M.E.D., V.M.-F., I.B.V.d.V. and J.D.H. contributed to the data interpretation, reviewing and editing the manuscript. L.M.J.N. and A.M.F. supervised the research generating the embryos, viability and windows of lethality data and reviewed and edited the manuscript. C.-W.H. contributed to the data analysis and presentation. C.M. contributed to reviewing and editing the manuscript. S.A.M., K.C.K.L. and L.L. contributed to the data collection, interpretation and editing and reviewing the manuscript. L.T. and S.W. contributed to the conceptualisation, data generation and data interpretation. A.K.C., V.L., R.G. and D.Q. performed the research generating embryos, viability and windows of lethality data. J.C., R.B.-S., M.S. and J.H. contributed to the data generation and interpretation. J.M. and H.H.M. contributed to the data analysis and software development. M.E.D., C.M., J.D.H., K.C.K.L., R.E.B., J.K.W., A.-M.M., H.P. and D.S. are PIs of the key programmes who contributed to the management

and execution of the work. The additional IMPC members all contributed to the data acquisition and data handling. The authors read and approved the final manuscript.

#### Funding

This work was supported by NIH grant U54 HG006370 (P.C., C.H.W., V.M.-F., J.M., H.M.M., H.P., A.-M.M., D.S.). Other National Institute of Health grants include R01 HD083311 (J.M.), UM1 HG006348 (M.E.D., C.H., J.D.H., L.T., S.W.), UM1 OD023221 (C.M., K.C.K.L., L.L.), UM1 OD023221-09S1 (C.M.), UM1 OD0023222 (S.A.M.), U42 OD011174 and 5UM1 HG006348-10 (L.T., S.W., J.C., R.B.S., M.S., J.H.). Additional support was provided by the Medical Research Council, Strategic Award A410-53658 (L.T., S.W., J.C., R.B.S., M.S., J.H.).

#### Availability of data and materials

All the results presented in the manuscript are available in Supplementary Information (Additional files 1 and 2). All the data supporting the findings of this study are made publicly available in the following repository: <https://doi.org/10.5281/zenodo.5796621> [33].

Full viability reports and additional files containing mouse embryo and adult phenotype associations are available through the IMPC web portal (<https://www.mousephenotype.org/>). Data can be accessed directly through the search box in the homepage, through the batch query tool, via API or via FTP repository (<http://ftp.ebi.ac.uk/pub/databases/impc/>). More detailed information on how to access and use data and images can be found here: <https://www.mousephenotype.org/understand/accessing-the-data/>.

#### Declarations

##### Ethics approval and consent to participate

The IMPC Consortium collects data from international member institutes who collect phenotyping data guided by their own ethical review panels, licences and accrediting bodies that reflect the national and/or geo-political constructs in which they operate (Institutional Animal Care and Usage Committee, Baylor College of Medicine; Animal Welfare and Ethical Review Body (AWERB), MRC Harwell; Animal Care Committee (ACC) of The Centre for Phenogenomics; The Jackson Laboratory Institutional Animal Care and Use Committee (IACUC); UC Davis Institutional Animal Care and Use Committee (IACUC)).

All the information regarding animal ethics approval of mouse production, breeding and phenotyping, including study design, experimental procedures, housing and husbandry and sample size can be found in the following links: <https://www.mousephenotype.org/about-impc/animal-welfare/> <https://www.mousephenotype.org/about-impc/animal-welfare/arrive-guide-lines/>

All efforts were made to minimise suffering by considerate housing and husbandry. All phenotyping procedures were examined for potential refinements that were disseminated throughout the Consortium. Animal welfare was assessed routinely for all mice involved.

All patient data used from the 100,000 Genomes Project were accessed through the research environment provided by Genomics England and conforming to their procedures. All participants in the 100KGP have provided written consent to provide access to their anonymised clinical and genomic data for research purposes and all research conformed to the principles of the Helsinki Declaration.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>William Harvey Research Institute, Queen Mary University of London, London, UK. <sup>2</sup>MRC Harwell Institute, Harwell, Oxfordshire, UK. <sup>3</sup>Department of Veterinary and Animal Sciences, University of Massachusetts, Amherst, MA, USA. <sup>4</sup>Department of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, TX, USA. <sup>5</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. <sup>6</sup>The Hospital for Sick Children, The Centre for Phenogenomics, Toronto, Canada. <sup>7</sup>European Molecular Biology Laboratory-European Bioinformatics Institute, Hinxton, UK. <sup>8</sup>Department of Education, Innovation and Technology, Baylor College of Medicine,

Houston, TX, USA. <sup>9</sup>Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, TX, USA. <sup>10</sup>Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, The Centre for Phenogenomics, Toronto, Canada. <sup>11</sup>The Jackson Laboratory, Bar Harbor, ME, USA. <sup>12</sup>The Mary Lyon Centre, MRC Harwell Institute, Harwell, Oxfordshire, UK. <sup>13</sup>Mouse Biology Program, University of California Davis, Davis, CA, USA.

Received: 18 January 2022 Accepted: 26 September 2022

Published online: 13 October 2022

#### References

- Fung JLF, et al. A three-year follow-up study evaluating clinical utility of exome sequencing and diagnostic potential of reanalysis. *NPJ Genom Med.* 2020;5:37.
- Posey JE. Genome sequencing and implications for rare disorders. *Orphanet J Rare Dis.* 2019;14:153.
- Seaby EG, Rehm HL, O'Donnell-Luria A. Strategies to uplift novel Mendelian gene discovery for improved clinical outcomes. *Front Genet.* 2021;12:674295.
- Chong JX, et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet.* 2015;97:199–215.
- Seaby EG, Ennis S. Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Brief Funct Genomics.* 2020;19:243–58.
- Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* 2019;47:D1038–43.
- Boycott KM, et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am J Hum Genet.* 2017;100:695–705.
- Strande NT, et al. Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the clinical genome resource. *Am J Hum Genet.* 2017;100:895–906.
- Martin AR, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet.* 2019;51:1560–+.
- Bamshad MJ, Nickerson DA, Chong JX. Mendelian gene discovery: fast and furious with no end in sight. *Am J Hum Genet.* 2019;105:448–55.
- Ropers HH. New perspectives for the elucidation of genetic disorders. *Am J Hum Genet.* 2007;81:199–207.
- Sudlow C, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12:e1001779.
- Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581:434–43.
- Minikel EV, et al. Evaluating drug targets through human loss-of-function genetic variation. *Nature.* 2020;581:459–64.
- Fridman H, et al. The landscape of autosomal-recessive pathogenic variants in European populations reveals phenotype-specific effects. *Am J Hum Genet.* 2021;108:608–19.
- Barton AR, Huijoe ML, Mukamel RE, Sherman MA, Loh P-R. A spectrum of recessiveness among Mendelian disease variants in UK Biobank. *Am J Hum Genet.* 2022;109(7):1298–307.
- Smedley D, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med.* 2021;385:1868–80.
- Kaplanis J, et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature.* 2020;586:757–62.
- Bertoli-Avella AM, et al. Combining exome/genome sequencing with data repository analysis reveals novel gene-disease associations for a wide range of genetic disorders. *Genet Med.* 2021;23:1551–68.
- Smedley D, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc.* 2015;10:2004–15.
- Cacheiro P, et al. New models for human disease from the International Mouse Phenotyping Consortium. *Mamm Genome.* 2019;30:143–50.
- Meehan TF, et al. Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nat Genet.* 2017;49:1231–8.



23. Georgi B, Voight BF, Bucan M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet*. 2013;9:e1003484.
24. Dickinson ME, et al. High-throughput discovery of novel developmental phenotypes. *Nature*. 2016;537:508–14.
25. Cacheiro P, et al. Human and mouse essentiality screens as a resource for disease gene discovery. *Nat Commun*. 2020;11:655.
26. Rodger C, et al. De novo VPS4A mutations cause multisystem disease with abnormal neurodevelopment. *Am J Hum Genet*. 2020;107:1129–48.
27. Cousin MA, et al. Pathogenic SPTBN1 variants cause an autosomal dominant neurodevelopmental syndrome. *Nat Genet*. 2021;53:1006–21.
28. Agana M, Frueh J, Kamboj M, Patel DR, Kanungo S. Common metabolic disorder (inborn errors of metabolism) concerns in primary care practice. *Ann Transl Med*. 2018;6:469.
29. DeBerardinis RJ, Thompson CB. Cellular metabolism and disease: what do metabolic outliers teach us? *Cell*. 2012;148:1132–44.
30. Munoz-Fuentes V, et al. The International Mouse Phenotyping Consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. *Conserv Genet*. 2018;19:995–1005.
31. IMPC. Data Release 15.0 <http://ftp.ebi.ac.uk/pub/databases/impc/all-data-releases/release-15.0/>. Accessed 29 May 2022.
32. IMPC. Gene Page. Data Release 16.0 <https://www.mousephenotype.org/data/genes/>; <http://ftp.ebi.ac.uk/pub/databases/impc/all-data-releases/release-16.0/>. Accessed 29 May 2022.
33. Cacheiro P, Smedley D. Mendelian gene identification through mouse embryo viability screening [Data set]. Zenodo. 2022. <https://doi.org/10.5281/zenodo.5796621>.
34. Tweedie S, et al. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res*. 2021;49:D939–46.
35. Meyers RM, et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet*. 2017;49:1779–84.
36. Cardoso-Moreira M, et al. Gene expression across mammalian organ development. *Nature*. 2019;571:505–9.
37. Cardoso-Moreira M, et al. Developmental gene expression differences between humans and mammalian models. *Cell Rep*. 2020;33:108308.
38. Quinodoz M, et al. DOMINO: using machine learning to predict genes associated with dominant disorders. *Am J Hum Genet*. 2017;101:623–9.
39. Rapaport F, et al. Negative selection on human genes underlying inborn errors depends on disease outcome and both the mode and mechanism of inheritance. *Proc Natl Acad Sci U S A*. 2021;118:e2001248118.
40. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*. 2013;9:e1003709.
41. Howe KL, et al. Ensembl 2021. *Nucleic Acids Res*. 2021;49:D884–91.
42. Szklarczyk D, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*. 2017;45:D362–8.
43. Jassal B, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48:D498–503.
44. Mistry J, et al. Pfam: the protein families database in 2021. *Nucleic Acids Res*. 2021;49:D412–9.
45. Giurgiu M, et al. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*. 2019;47:D559–63.
46. Kohler S, et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res*. 2021;49:D1207–17.
47. Dawes R, Lek M, Cooper ST. Gene discovery informatics toolkit defines candidate genes for unexplained infertility and prenatal or infantile mortality. *NPJ Genom Med*. 2019;4:8.
48. Chouldechova A, Hastie T, Spinu V. gamsel: fit regularization path for generalized additive models; 2018.
49. Hastie T, Mazumder R. softImpute: matrix completion via iterative soft-thresholded SVD; 2021.
50. Mager J. A Catalog of Early Lethal KOMP Phenotypes; 2021. <https://blogs.umass.edu/mager/>.
51. R Core Team. R: a language and environment for statistical computing; 2021.
52. Wickham H, et al. Welcome to the tidyverse. *J Open Source Softw*. 2019;4:1686.
53. Bengtsson H. matrixStats: functions that apply to rows and columns of matrices (and to vectors); 2021.
54. Aragon TJ. epitools: epidemiology tools. R package version 0.5-10.1; 2020.
55. Signorell Aea. DescTools: Tools for Descriptive Statistics. R package version 0.99.45; 2022.
56. Schratz P. R package 'oddsratio': odds ratio calculation for GAM(M)s & GLM(M)s, version: 1.0.2; 2017. <https://doi.org/10.5281/zenodo.1095472>.
57. Rudis B, Gandy D. waffle: create waffle chart visualizations in R; 2017.
58. Wilke CO. ggrridges: ridgeline plots in 'ggplot2'; 2021.
59. Bojanowski M, Edwards R. alluvial: R package for creating alluvial diagrams; 2016.
60. Wilke CO. cowplot: streamlined plot theme and plot annotations for 'ggplot2'; 2020.
61. Gehlenborg N. UpSetR: a more scalable alternative to Venn and Euler diagrams for visualizing intersecting sets. R package version 1.4.0; 2019.
62. Greene D, Richardson S, Turro E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics*. 2017;33:1104–6.
63. Robin X, et al. pROC: an open-source package for R and S plus to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
64. Wang WY, et al. Combined gene essentiality scoring improves the prediction of cancer dependency maps. *Ebiomedicine*. 2019;50:67–80.
65. Penon-Portmann M, et al. Human embryonic expression identifies novel essential gene candidates. *bioRxiv*. 2020:2020.08.15.252338.
66. Shakhnovich BE, Koonin EV. Origins and impact of constraints in evolution of gene families. *Genome Res*. 2006;16:1529–36.
67. De Kegel B, Ryan CJ. Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. *PLoS Genet*. 2019;15:e1008466.
68. Kabir M, Wenlock S, Doig AJ, Hentges KE. The essentiality status of mouse duplicate gene pairs correlates with developmental co-expression patterns. *Sci Rep*. 2019;9:3224.
69. Zhai J, Xiao Z, Wang Y, Wang H. Human embryonic development: from peri-implantation to gastrulation. *Trends Cell Biol*. 2021;32:18–29.
70. Shahbazi MN. Mechanisms of human embryo development: from cell fate to tissue shape and back. *Development*. 2020;147:dev190629.
71. Jarvis GE. Early embryo mortality in natural human reproduction: what the data say. *F1000Res*. 2016;5:2765.
72. Colley E, et al. Potential genetic causes of miscarriage in euploid pregnancies: a systematic review. *Hum Reprod Update*. 2019;25:452–72.
73. Agenor A, Bhattacharya S. Infertility and miscarriage: common pathways in manifestation and management. *Womens Health (Lond)*. 2015;11:527–41.
74. Tsherniak A, et al. Defining a Cancer Dependency Map. *Cell*. 2017;170:564–76.
75. Cheong A, et al. Nuclear-encoded mitochondrial ribosomal proteins are required to initiate gastrulation. *Development*. 2020;147:dev188714.
76. Gopisetty G, Thangarajan R. Mammalian mitochondrial ribosomal small subunit (MRPS) genes: a putative role in human disease. *Gene*. 2016;589:27–35.
77. Bugiardini E, et al. MRPS25 mutations impair mitochondrial translation and cause encephalomyopathy. *Hum Mol Genet*. 2019;28:2711–9.
78. vanLieshout TL, Ljubicic V. The emergence of protein arginine methyltransferases in skeletal muscle and metabolic disease. *Am J Physiol Endocrinol Metab*. 2019;317:E1070–80.
79. Bult CJ, et al. Mouse Genome Database (MGD) 2019. *Nucleic Acids Res*. 2019;47:D801–6.
80. Hargreaves I, Heaton RA, Mantle D. Disorders of human coenzyme Q10 metabolism: an overview. *Int J Mol Sci*. 2020;21:6695.
81. Colas P. Cyclin-dependent kinases and rare developmental disorders. *Orphanet J Rare Dis*. 2020;15:203.
82. Brown SDM, et al. High-throughput mouse phenomics for characterizing mammalian gene function. *Nat Rev Genet*. 2018;19:357–70.
83. Lloyd KCK, et al. The Deep Genome Project. *Genome Biol*. 2020;21:18.
84. Chen H, et al. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. *Brief Bioinform*. 2020;21:1397–410.
85. Joshi CJ, Ke W, Drangowska-Way A, O'Rourke EJ, Lewis NE. What are housekeeping genes? *bioRxiv*; 2021.
86. Wang T, et al. Identification and characterization of essential genes in the human genome. *Science*. 2015;350:1096–101.

87. Dandage R, Landry CR. Paralog dependency indirectly affects the robustness of human cells. *Mol Syst Biol*. 2019;15:e8871.
88. De Kegel B, Quinn N, Thompson NA, Adams DJ, Ryan CJ. Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines. *Cell Syst*. 2021;12:1144–1159 e6.
89. Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol*. 2014;10:733.
90. Paine I, et al. Paralog studies augment gene discovery: DDX and DHX genes. *Am J Hum Genet*. 2019;105:302–16.
91. Lal D, et al. Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. *Genome Med*. 2020;12:28.
92. Perez-Palma E, et al. Identification of pathogenic variant enriched regions across genes and gene families. *Genome Res*. 2020;30:62–71.
93. Ferreira CR, Rahman S, Keller M, Zschocke J, Grp IA. An international classification of inherited metabolic disorders (ICIMD). *J Inher Metab Dis*. 2021;44:164–77.
94. Saudubray JM, Garcia-Cazorla A. An overview of inborn errors of metabolism affecting the brain: from neurodevelopment to neurodegenerative disorders. *Dialogues Clin Neurosci*. 2018;20:301–25.
95. Balakrishnan B, et al. A novel phosphoglucomutase-deficient mouse model reveals aberrant glycosylation and early embryonic lethality. *J Inher Metab Dis*. 2019;42:998–1007.
96. Nyman LR, et al. Homozygous carnitine palmitoyltransferase 1a (liver isoform) deficiency is lethal in the mouse. *Mol Genet Metab*. 2005;86:179–87.
97. Diomedì-Camassei F, et al. COQ2 nephropathy: a newly described inherited mitochondriopathy with primary renal involvement. *J Am Soc Nephrol*. 2007;18:2773–80.
98. Chung WK, et al. Mutations in COQ4, an essential component of coenzyme Q biosynthesis, cause lethal neonatal mitochondrial encephalomyopathy. *J Med Genet*. 2015;52:627–35.
99. Danhauser K, et al. Fatal neonatal encephalopathy and lactic acidosis caused by a homozygous loss-of-function variant in COQ9. *Eur J Hum Genet*. 2016;24:450–4.
100. Lopez LC, et al. Leigh syndrome with nephropathy and CoQ10 deficiency due to decaprenyl diphosphate synthase subunit 2 (PDSS2) mutations. *Neurology*. 2007;68:A202.
101. Beecroft SJ, et al. Biallelic hypomorphic variants in ALDH1A2 cause a novel lethal human multiple congenital anomaly syndrome encompassing diaphragmatic, pulmonary, and cardiovascular defects. *Hum Mutat*. 2021;42:506–19.
102. Blackburn PR, et al. Expanding the clinical and phenotypic heterogeneity associated with biallelic variants in ACO2. *Ann Clin Transl Neurol*. 2020;7:1013–28.
103. Jarvis GE. Early embryo mortality in natural human reproduction: What the data say. *F1000Res*. 2016;5:2765.
104. Shamseldin HE, et al. Identification of embryonic lethal genes in humans by autozygosity mapping and exome sequencing in consanguineous families. *Genome Biol*. 2015;16:116.
105. Liao BY, Zhang JZ. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A*. 2008;105:6987–92.
106. Baldrige D, et al. Model organisms contribute to diagnosis and discovery in the undiagnosed diseases network: current state and a future vision. *Orphanet J Rare Dis*. 2021;16:206.
107. Filges I, Friedman JM. Exome sequencing for gene discovery in lethal fetal disorders - harnessing the value of extreme phenotypes. *Prenat Diagn*. 2015;35:1005–9.
108. Vaiman D. Genetics of Early Miscarriages. In eLS, John Wiley & Sons, Ltd (Ed.); 2016. <https://doi.org/10.1002/9780470015902.a0025043>.
109. Dhombres F, et al. Prenatal phenotyping: a community effort to enhance the Human Phenotype Ontology. *Am J Med Genet C: Semin Med Genet*. 2022.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

