© 2021. Published by The Company of Biologists Ltd | Development (2021) 148, dev192955. doi:10.1242/dev.192955

TECHNIQUES AND RESOURCES

RESEARCH ARTICLE

LAMA: automated image analysis for the developmental phenotyping of mouse embryos

Neil R. Horner[1], Shanmugasundaram Venkataraman[2], Chris Armit[2,3], Ramón Casero[1], James M. Brown[4], Michael D. Wong[5], Matthijs C. van Eede[5], R. Mark Henkelman[5], Sara Johnson[1], Lydia Teboul[1], Sara Wells[1], Steve D. Brown[1], Henrik Westerberg[1] and Ann-Marie Mallon[1,*]

ABSTRACT

Advanced 3D imaging modalities, such as micro-computed tomography (micro-CT), have been incorporated into the high-throughput embryo pipeline of the International Mouse Phenotyping Consortium (IMPC). This project generates large volumes of raw data that cannot be immediately exploited without significant resources of personnel and expertise. Thus, rapid automated annotation is crucial to ensure that 3D imaging data can be integrated with other multi-dimensional phenotyping data. We present an automated computational mouse embryo phenotyping pipeline that harnesses the large amount of wild-type control data available in the IMPC embryo pipeline in order to address issues of low mutant sample number as well as incomplete penetrance and variable expressivity. We also investigate the effect of developmental substage on automated phenotyping results. Designed primarily for developmental biologists, our software performs image pre-processing, registration, statistical analysis and segmentation of embryo images. We also present a novel anatomical E14.5 embryo atlas average and, using it with LAMA, show that we can uncover known and novel dysmorphology from two IMPC knockout lines.

KEY WORDS: Automated, Computational, Embryo, Micro-CT, Mouse, Phenotyping

INTRODUCTION

A major goal in biomedical research is to assign functional roles to all genes in order to shed light on disease mechanisms, and to identify disease-associated genes and novel drug targets. However, almost two decades since the human and mouse draft genomes were published (Lander et al., 2001; Waterston et al., 2002), the proportion of genes in the dark genome, defined as those having minimal gene function or disease-association annotations, remains high at over 30% (Oprea, 2019). The International Mouse Phenotyping Consortium (IMPC) is a high-throughput functional genomics project tasked with generating a genome-wide catalogue of gene function by phenotyping gene knockouts on a uniform genetic background (Brown and Moore, 2012; Lloyd et al., 2020). Phenotype annotations for over 7000 genes are currently available on the IMPC web portal (mousephenotype.org), data that have already contributed to the identification of many novel candidate disease genes and new mouse models of human disease (Cacheiro et al., 2020; Meehan et al., 2017; Bowl et al., 2017; Moore et al., 2018).

Postnatal lethality or subviability is observed in approximately one-third of knockout mouse lines from both the IMPC (Dickinson et al., 2016) and its precursor EUMODIC (Hrabě de Angelis et al., 2015). These classes of genes provide important insights into developmental processes and disorders. The IMPC seeks to phenotype these classes of gene through the embryo phenotyping pipeline at key embryonic developmental stages (E14.5, E15.5 and E18.5) via the generation and analysis of high resolution, whole embryo, 3D images (Adams et al., 2013). There are currently several thousand 3D images across hundreds of mutant lines at the IMPC, which will be impractical to manually annotate by domain experts as this can take several hours per image (Wilson et al., 2016). Therefore, a high-throughput, generalisable analysis method is needed to extract phenotype associations from these data. An automated method will also mitigate any user bias that may negatively affect reproducibility.

One approach is to automate phenotypic annotation using voxel-based morphometry (VBM) in which 3D images are spatially aligned to allow voxel intensities or deformation fields to be statistically analysed in order to identify morphological differences. This approach, originally developed for human brain MRI images (Wright et al., 1995), proved to be suitable for the analysis of MRI whole-embryo images of E15.5 mice (15.5 days post coitum) which was able to identify morphological differences between wild-type inbred strains (Zamyadi et al., 2010). This work was expanded to the analysis of micro-CT images of mutant E15.5 embryos, showing that known dysmorphology could also be identified using this approach (Wong et al., 2014).

Statistical parametric heatmaps obtained from VBM analysis can be overlaid onto the registered images in order to highlight regions of dysmorphology, facilitating manual annotation by an expert anatomist. But in order to automate the assignment of anatomical phenotypes, an atlas is required. An atlas consists of an image volume where visible anatomical structures have been manually delineated and identified, which enables the automatic calculation of organ volume when combined with VBM. An atlas can be derived from a single reference image, but a population average consensus reference image formed from spatially normalising multiple specimens provides increased signal to noise and contrast (Holmes et al., 1998) making it easier to segment. The population average also provides a less biased registration target from which to

[1]Medical Research Council Harwell Institute, Harwell OX11 0RD, UK. [2]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine (IGMM), University of Edinburgh, Edinburgh EH4 2XU, UK. [3]BGI Hong Kong, 26/F, Kings Wing Plaza 2, 1 On Kwan Street, Shek Mun, New Territories, Hong Kong. [4]School of Computer Science, University of Lincoln, Lincoln LN6 7TS. [5]Mouse Imaging Centre, Hospital for Sick Children, Toronto, Ontario M5T 3H7, Canada.

*Author for correspondence (a.mallon@har.mrc.ac.uk)

N.R.H., 0000-0002-0550-4537; S.V., 0000-0002-3200-2698; R.C., 0000-0002-5684-1242; J.M.B., 0000-0001-7636-4554; M.C.v.E., 0000-0001-9590-7016; S.J., 0000-0001-6308-6401; L.T., 0000-0002-2789-8637; S.W., 0000-0002-0572-0600; H.W., 0000-0002-7204-6900; A.-M.M., 0000-0003-4047-4019

Handling Editor: Liz Robertson
Received 19 May 2020; Accepted 21 December 2020

Development (2021) 148, dev192955. doi:10.1242/dev.192955

propagate anatomical labels. Mouse embryo population average models and associated atlases have been developed for the E15.5 developmental stage using MRI (Cleary et al., 2011) and micro-CT (Wong et al., 2012) where six and 48 anatomical structures were segmented, respectively.

Sample sizes of both mutant and wild-type embryos are important considerations in any phenotyping experiment. Wong et al. (2012, 2014) proposed using eight wild types and eight mutants in their pipeline; however, knockout lines submitted to the IMPC often have much lower sample sizes. Despite the fact that knockout lines are generated from isogenic inbred mice, they frequently exhibit incomplete penetrance and variable expressivity of phenotypes (Wilson et al., 2017; Dickinson et al., 2016), reducing the statistical power to uncover gene-level phenotypes. One solution to this problem is to increase the number of control specimens in order to increase statistical power. However, there are two features of the previous method (Wong et al., 2014) that place restrictions on increasing sample numbers. First, the groupwise registration steps increase computational cost exponentially with increased sample number, and second, the wild-type controls must be registered along with specimens from each mutant line into a unique coordinate space and so cannot be reused for the analysis of other genes.

Another complicating factor in the study of mouse embryos is the presence of inter- and intra-litter variability in developmental stage and in associated morphological differences, which are frequently observed even with inbred wild-type mice (Miyake et al., 1996; Geyer et al., 2017). Indeed, fertilisation time is only estimated by the presence of vaginal plugs, and therefore may vary within a litter along with rate of development. Therefore, embryos harvested from a litter will present a range of developmental sub-stages that can span up to several hours, which corresponds to significant morphological differences in embryos. Therefore, to avoid spurious annotations or masking of real dysmorphology, it is essential to control for developmental stage. These issues of both partial penetrance and of developmental stage variability have yet to be studied in the context of automated phenotyping using 3D images.

In this article, we introduce a new automated phenotyping pipeline (LAMA) that is designed to address the issues arising from the high-throughput analysis of 3D mouse embryo data from the IMPC pipeline. One of the main differences in LAMA when compared with previous work (Wong et al., 2014) is the use of a registration strategy where all baseline and mutant specimens are registered directly towards a population average target in a pairwise manner with no groupwise registrations. This allows a large increase in the number of wild-type controls that can be used when analysing mutant lines, which greatly increases statistical power. Using E14.5 embryo images, a timepoint yet to be subject to this form of automated annotation, we show that with this increase in power, LAMA is able to identify sex differences using a low sample number, and that previously known and novel phenotypes can be uncovered from two knockout lines. Importantly, in one of these lines, LAMA is able to assign known phenotypes to individual specimens, which has not been shown previously for the automated analysis of mouse embryos. We report the results of the effect of developmental substage on automated analysis and include an updated statistical model to account for this. To accompany the pipeline, we present a novel, highly detailed anatomical atlas of an E14.5 population average with 184 labels and associated Mouse Developmental Anatomy Ontology (EMAPA) terms (Hayamizu et al., 2013), which is the most-detailed atlas of a micro-CT population average embryo that is currently available. Finally, we have made LAMA open source and simple to install on all major operating systems and have included tools for preprocessing of data, distributed computing to speed up analysis, and the production of various plots and reports to help users understand the registration process and statistical analyses. These tools, resources and insights presented here will greatly increase the ability to uncover useful phenotype information from embryo imaging data and it is currently being optimised to work with IMPC data from other developmental timepoints.

RESULTS

Overview of the LAMA phenotyping pipeline

LAMA is a voxel-based morphometry (VBM) approach to automate the detection of anatomical dysmorphology in mouse embryos. It is written in the Python programming language and is distributed as a PyPi package (https://pypi.org/) enabling installation with a single command (see Materials and Methods). It features spatial normalisation of images, using a registration process to iteratively align micro-CT embryo images to a population average image, putting them into the same coordinate space. Internally, it uses elastix (Klein et al., 2010; Shamonin et al., 2014) for multi-resolution 3D image registration. First, LAMA constructs a population average model embryo from wild-type derived images (Fig. 1A) if one does not exist already. This serves as a target image for the subsequent spatial normalisation of the phenotyping images, and as the template for a hand-labelled atlas of mouse organs and anatomical structures (described below) (Fig. 1A). The next step involves spatially normalising each wild-type and mutant image that will be used in the downstream statistical analysis, by registering it towards the population average (Fig. 1B), resulting in morphologically similar images with homologous anatomical structures occupying identical coordinates and existing in the same coordinate space as the atlas.

Each individual specimen image is then automatically segmented by applying the inverse of its registration transformation to the atlas labels (Fig. 1B). Whole-embryo volumes (WEVs) and organ volumes are calculated from these segmentations. Each set of organ segmentations also serves as a specimen-specific visual atlas that can aid in identifying structures in the original inputs and visually detecting segmentation errors. Organ volumes are fitted to a linear model organ volume/WEV~genotype+WEV. In this model, the organ volumes are first normalised by WEV to control for overall embryo size and the WEV fixed effect accounts for differences due to developmental substage (see the section ‘Developmental substage’). Modelling organ volumes is orders of magnitude less complex than with voxel-based deformation measures (hundreds of data points per line instead of millions), allowing us to employ a more robust permutation-based method for multiple testing correction (see Materials and Methods and Hrabě de Angelis et al., 2015). For this article, we focus on the organ volume analysis as this is linked to the atlas, resulting in automated phenotype calls at the organ level that are more robust and interpretable than those at the individual voxel level. However, we also include statistical parametric heat maps from the voxel-level Jacobian determinants analysis (which indicate local volume shrinkage/expansion during spatial normalisation) for illustrative purposes. The statistical analysis is carried out by combining specimens with the same gene deletion to give gene-level phenotype calls. Each specimen is also analysed individually to give a specimen-level phenotype call, which aims to uncover phenotypes with variability in penetrance or expressivity (see Materials and Methods). During analysis, various plots and information files are generated, including registration metric plots to aid in registration optimisation, organ volume

Development (2021) 148, dev192955. doi:10.1242/dev.192955

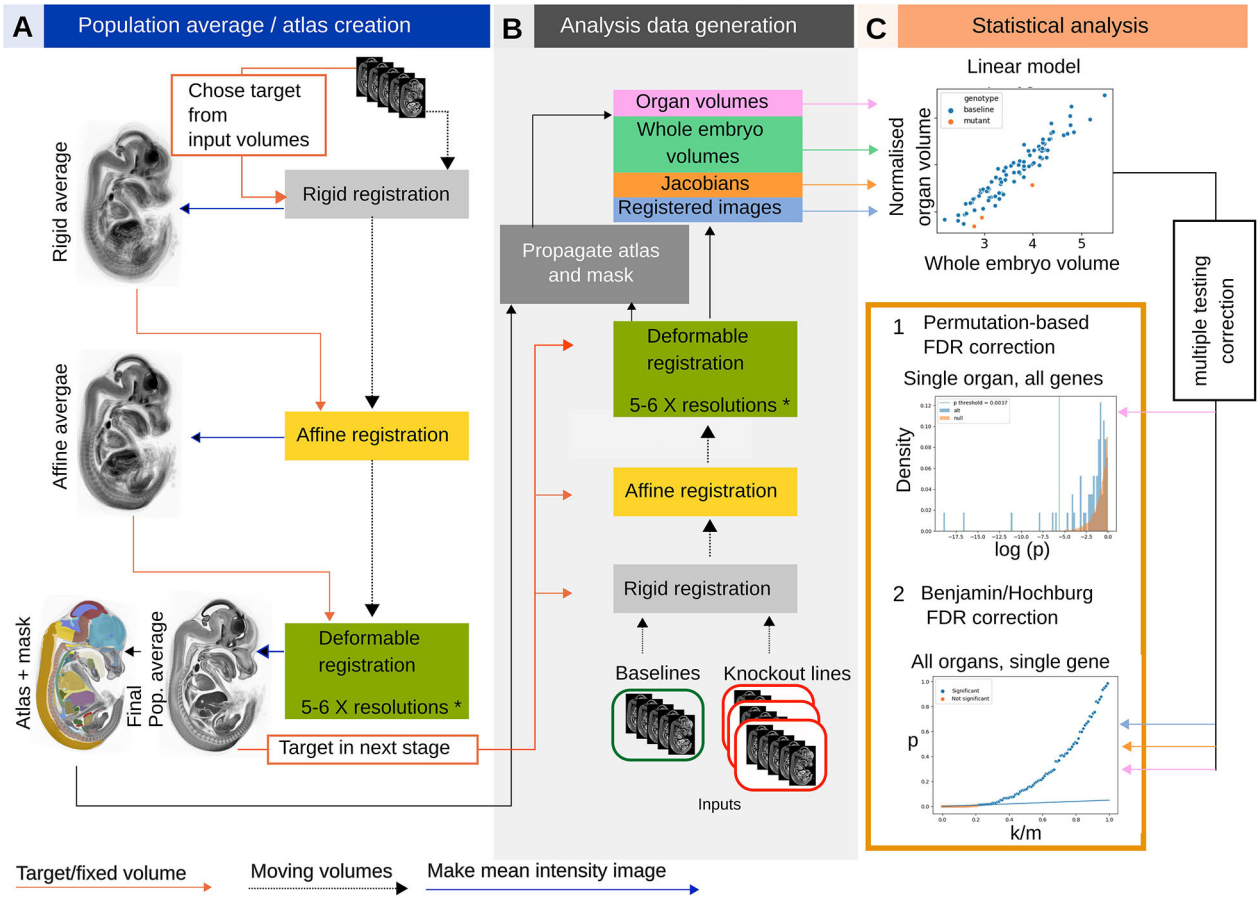

Fig. 1. LAMA pipeline workflow. (A) Population average construction. A random initial image is used as a target to rigidly align all other images, creating a rigid population average. This is repeated with affine and deformable registration, refining the population average and using it as the target for the next level. (B) Generation of data for phenotype detection. Each test image is registered to the final population average obtained in A, but using the same rigid/affine/B-spline registration levels. (C) Jacobian determinant volumes, registered images and organ volumes are statistically analysed. Top panel: organ volumes (shown here) or voxel values are fitted to a linear model by genotype and whole embryo volume. The resulting genotype effect *P*-values are corrected for multiple testing in one of two ways. (1) permutation-based FDR correction (organ volumes only). The orange histogram shows the null distribution from permuting the wild-type organ volumes and the blue histogram is the alternative distribution derived from testing the organ volume from all mutant lines tested. The vertical green line indicates the calculated *P*-value threshold for this organ, with values lower than this annotated as significant. (2) Benjamini Hochberg FDR correction for voxel-level or organ volume data. k/m=*P*-value rank divided by number of values. The straight blue line indicates the threshold under which values are annotated as significant.

plots and heatmaps that give an overview of the statistical results, and QC report montages showing a snapshot of the automated segmentation results.

Creation of a novel E14.5 mouse embryo atlas

The E14.5 population average used in this study was created from 16 specimens (eight male and eight female), with a resulting crown-rump length of 9.18 mm (s.d. 0.52 mm) and an isotropic voxel size of 14 µm^3 (Fig. 2A). The visible organs were segmented using a mixture of manual and semi-automatic segmentation (see Materials and Methods) producing an E14.5 atlas containing 184 unique labels (Fig. 2B,D; Table S1; Movie 1). The 184 labels were mapped to gross anatomy terms from the EMAPA developmental ontology (Hayamizu et al., 2013), where a one to one relationship existed, allowing the automatic integration of the resulting gene-to-phenotype data from LAMA with other data sources that also use this ontology, such as the other IMPC pipelines and MGI (Bult et al., 2019). Through visual inspection of registration results, a number of labels, including blood vessels, nerves and small muscles, were identified as being too small or too thin, leading to difficulty in assessing their registration accuracy. With this in mind, we identified such labels in the atlas (see Materials and Methods) to exclude them from downstream analyses, resulting in a final set of 103 labels that were used in the current analysis (Table S1). These 103 labels were distributed across the majority of the EMAPA high-level organ system terms (Fig. 2C) and range in size from the largest (forebrain at 7.0 mm^3) to smallest (metatarsals 0.04 mm^3).

Developmental substage

As the developmental substage (DSS) of E14.5 embryos has been shown to be an important consideration when manually phenotyping embryos (Geyer et al., 2017), we did a series of experiments to gauge the effect that DSS has on our automated phenotyping results. To account for overall embryo size, we normalised organ volumes, derived from automatically segmented labels, to whole embryo volume (WEV). Similarly, we removed rigid and affine transformations from the analysis of Jacobian determinants so that the determinants correspond to local deformable transformations (see Materials and Methods) as described previously (Wong et al., 2012) for E15.5 embryos. To assess whether these normalisations are

Development (2021) 148, dev192955. doi:10.1242/dev.192955

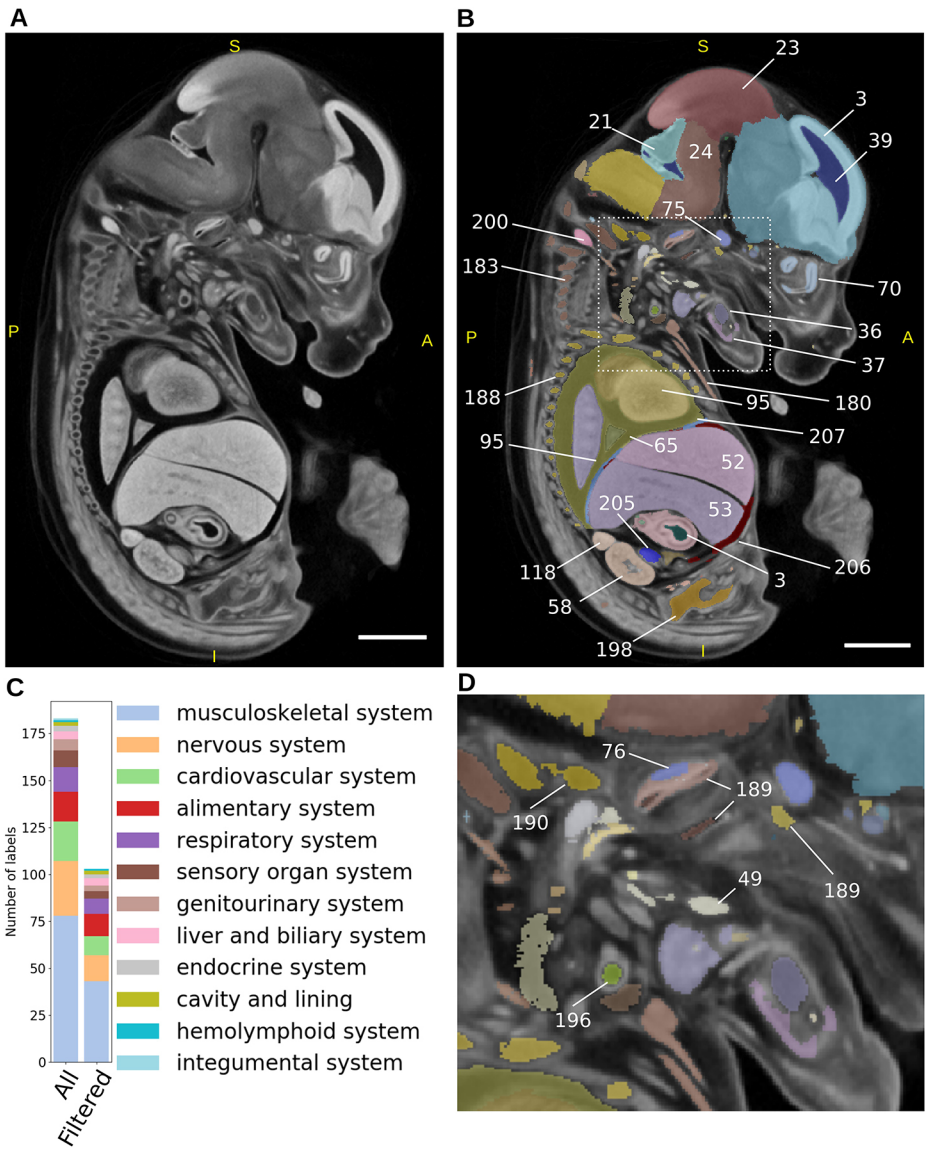

Fig. 2. E14.5 atlas creation. (A) Sagittal view of the E14.5 population average created from 16 wild-type female and male E14.5 C57BL/6 mice. (B) Sagittal view of the E14.5 atlas consisting of 184 individual structures overlaid on the population average. Only a subset of labels is shown for illustrative purposes. Numbers indicate the labels assigned to segmented organs, which can be looked up in Table S1. (C) Breakdown of labels into corresponding organ systems for the whole atlas (All), and those remaining after filtering small or spindly labels (Filtered) (see Materials and Methods). (D) Expanded region from area outlined in B, highlighting more labels. Scale bars: 1 mm. S, superior; I, inferior; A, anterior; P, posterior.

sufficient to account for differences due to DSS, we made two wild-type datasets from the 93 wild-type controls where either the smallest or largest specimens (WEV mean z-score of −1.7 and 1.9, respectively) were relabelled as ‘mutant’. We applied the LAMA Jacobian determinant analysis to each dataset using the linear model deformable Jacobians~genotype (Fig. 3A), simulating the scenario of two mutant lines containing embryos at early or late E14.5 DSS. Both tests returned significant Jacobian determinant voxels, suggesting that the relabelled wild types had morphological differences that were dependent on the developmental stage of the specimens. WEV was then included as a fixed effect in the following model deformable Jacobians~genotype+WEV to act as surrogate for DSS. This experiment returned no significant genotype effect voxels, showing that DSS variability can be thus controlled for. To gain a more detailed view on the voxel-level DSS-dependent relative size differences, Jacobian determinants from 93 wild-type specimens we fitted to the linear model deformable Jacobians~WEV (Fig. 3B), which further highlighted regions that are proportionally larger at later stages (red) or proportionally smaller at later stages (blue). The equivalent test using organ volume analysis organ volume/WEV~WEV similarly resulted in significant calls for 78/103 labels (Table S2), including organs that were proportionally larger (n=58) later in development such as thymus and lung lobes (Fig. 3C,D), and those that were proportionally smaller (n=23), including brain ventricles and trigeminal glands (Fig. 3E,F). To summarise, normalising the Jacobian determinants or organ volumes before statistical analysis is not sufficient to account for DSS, and failure to model DSS can lead to false-positive results. Our method resolves this issue by regressing out the DSS effect in the statistical analysis.

Optimal sample size for phenodeviance testing

In order to validate LAMA with a positive control we applied it to wild-type embryos where females were relabelled as ‘mutant’. After removing specimens with indeterminate sex, our test data set contained 49 males and 40 females, in which we expected gross morphological differences between the two sets to be located at the gonads only (see Fig. S1A for example gonad images). These data were fitted to the linear model organ volume/WEV~sex+WEV and we found that, along with the gonad (FDR-corrected q-value=1.3e^{-25}) (Fig. 4A), the lens of the eye unexpectedly also had a significant sex effect (FDR-corrected q-value=0.028) (Fig. 4B; Fig. S1B,C). We next wanted to address the effect of sample size on

Development (2021) 148, dev192955. doi:10.1242/dev.192955

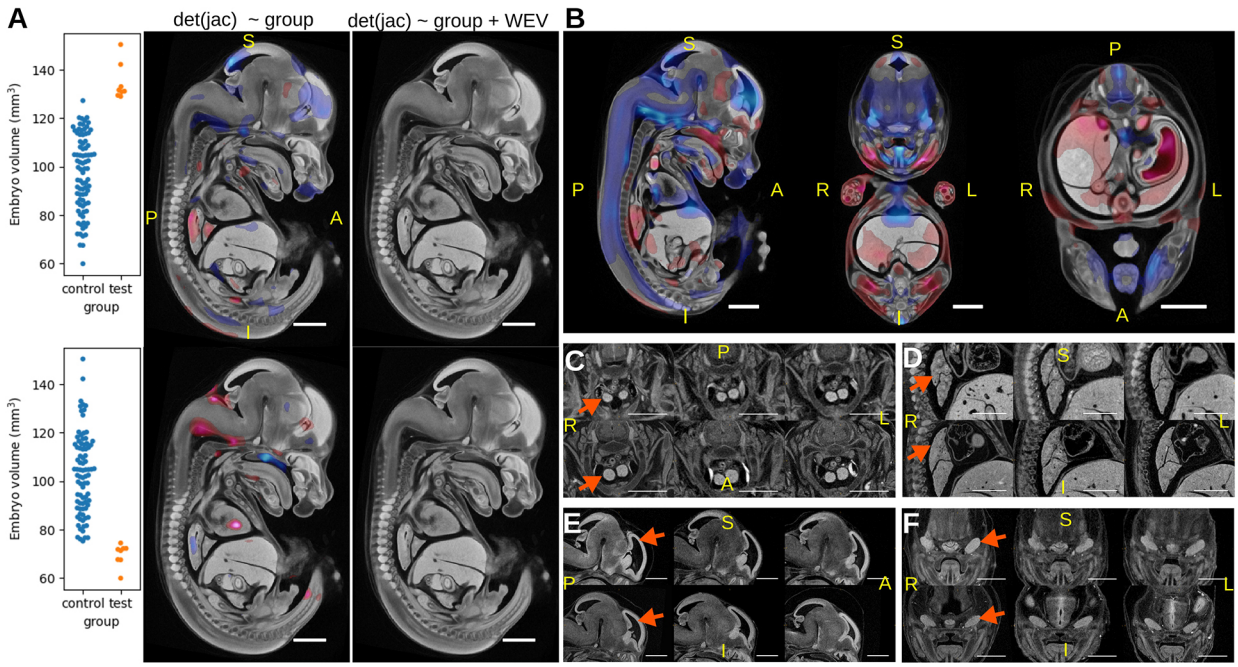

Fig. 3. Effect of E14.5 developmental substage on local volume changes detected by LAMA. (A) Simulation of the analysis of mutant lines with large (top panels) or small (bottom panels) wild types relabelled as mutants. Genotype effect t-statistics where q values (FDR-corrected *P*-values)<0.05 were overlaid on the E14.5 population average. Left images show the results from the model deformable Jacobians~genotype, with red voxels highlighting regions that are significantly larger in the test group and blue voxels highlighting regions that are significantly smaller in the test group. Right images show results from the model: deformable Jacobians~genotype+WEV, where no voxels passed the q<0.05 threshold. (B) Result from fitting 93 wild-type specimens to the linear model deformable Jacobians~WEV, indicating regions where the normalised organ volume is positively correlated (red) or negatively correlated (blue) to WEV. (C-F) Illustrative examples of differences in organ size relative to embryo volume. Each panel shows three representative wild types from the smallest set (top) and largest set (bottom) of specimens. Images are affinely registered towards the population average to account for overall embryo volume. Arrows indicate relevant anatomy. Thymus (C) and lung (D) show larger relative sizes in larger specimens. Lateral ventricles (E) and trigeminal gland (F) show smaller relative sizes in larger specimens. Scale bars: 1 mm. S, superior; I, inferior; A, anterior; P, posterior; L, left; R, right.

the sensitivity of phenodeviance detection (in this case the ability to differentiate between male and female gonads and lenses). To do this, the previous experiment was repeated, but with varying numbers of male or female specimens, in this way replicating the effect of testing mutant lines containing various sample numbers and with different baseline control sample numbers (ranging from two to eight females and from 10 to 49 males). Each experiment was repeated 50 times with random specimen selection and permutation-based multiple testing correction (see Materials and Methods for further details). Owing to the large difference between male and female gonad sizes, significant gonad volume differences were identified in almost every replication of each experiment (Fig. 4C), and significant Jacobian determinant voxels were identified within, or close to, the gonad, with significant voxels covering a larger area with increasing male sample size (Fig. 4E,F). For the lens of the eye, significant volume differences were detected only with a male sample size of 32 or over, with the maximum male and female sample size (49 male and eight female) resulting in significant hits in over half of the tests (Fig. 4D). To assess the rate of false-positive detection, any significant organ volumes other than gonad or lens of the eye were classed as false positives. The rate of false positives was found to be well controlled with only 1 out of 103 organs called as significant in more than 1% of tests (epiglottis in 1.6% of all the replications), and with a mean false-positive rate of 0.07% per label. These experiments show that LAMA is able to identify sex-specific differences in wild-type embryos and that even with a low mutant sample size, differences in morphology can be detected, although the detection becomes more reliable as the control sample size is increased.

Automated identification of developmental phenotypes in E14.5 mice embryos

The initial aim of LAMA was to automatically identify dysmorphology from IMPC-generated data. To demonstrate its effectiveness on IMPC-generated data, we have chosen two exemplar mutant lines that illustrate its use in embryos with multiple dysmorphologies throughout the body and embryos with very specific, localised abnormalities.

The first example is *Wfdc2*, which encodes a protease inhibitor protein that is expressed in several tissues during mouse development prior to E14.5 (Lizio et al., 2015), including intestines, lungs and pancreas. *WFDC2* plays a role in cancer development (Bingle et al., 2002; Li et al., 2013) and two articles have recently shown *Wfdc2* homozygous mutant mice display severe pulmonary phenotypes, including collapsed lungs at perinatal day 1.5 (P1.5) (Nakajima et al., 2019), and alveolar abnormalities, dyspnea and reduced blood oxygen saturation at birth (Zhang et al., 2020), but are otherwise anatomically normal. The IMPC viability screen reported that *Wfdc2*$^{-/-}$ animals were viable at E18.5 but displayed a partially penetrant preweaning lethality with 5.5% of the alive pups being homozygous for the mutation. Our analysis of four E14.5 *Wfdc2*$^{-/-}$ specimens uncovered significantly smaller bronchi and trachea at the gene level (using all four mutants in the analysis) (Fig. 5A) but no significant specimen-level differences (analysing each specimen individually) were observed for this gene. The Jacobian determinant analysis identified two significant regions that largely overlap with one of the bronchi after the FDR-corrected *P*-value threshold was raised to 0.1 (Fig. 5B).

Development (2021) 148, dev192955. doi:10.1242/dev.192955

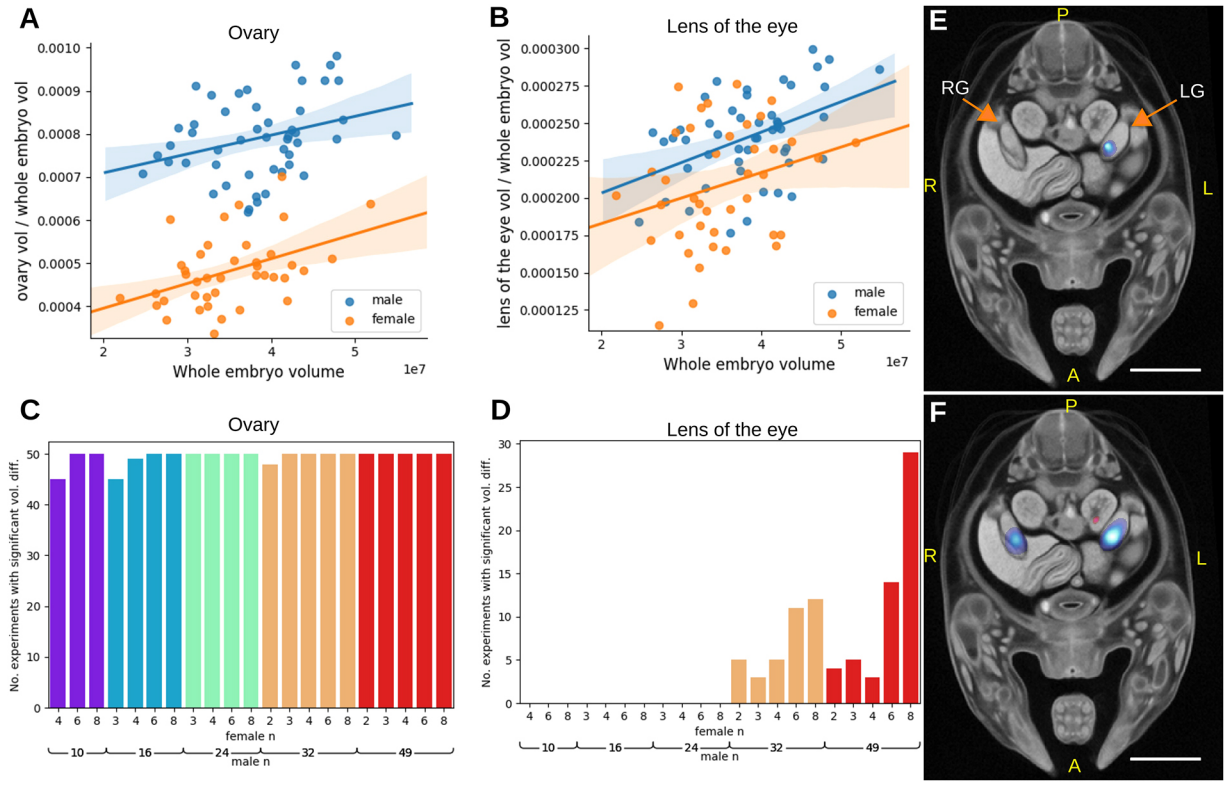

Fig. 4. Identification of sex differences in wild-type E14.5 mice. (A,B) Plots showing WEV normalised organ volume against WEV for ovary (A) and lens (B). (C,D) A series of statistical tests was carried out with various combinations of male and female wild-type sample size (with females relabelled as 'mutant'). Each test was repeated 50 times with randomly selected specimens. The values reported are the number of times the gonad volume was reported as significantly different (P<organ p threshold) in each set of tests for gonad (C) and lens of the eye (D). (E,F) Example results of Jacobian determinant analysis using eight females and eight males (E), and eight females and 46 males (F). Scale bars: 1 mm. RG, right gonad; LG, left gonad; A, anterior; P, posterior; L, left; R, right.

This means that, for this gene, the whole organ volume statistics were more sensitive than the Jacobian determinants. In all four mutants, the trachea and bronchus are visibly smaller in diameter (Fig. 5C,D), but otherwise appear normal.

Acan encodes for the protein aggrecan, which is the primary proteoglycan in articular cartilage, is present in the extracellular matrix of long bone epiphyseal growth plates and is required for normal bone development. *Acan*$^{-/-}$ mice exhibit phenotypes associated mainly with abnormal bone morphology, including long bones, ribs and vertebrae, as well as enlarged liver and pulmonary hypoplasia (Table 1). Human diseases associated with *ACAN* mutations include osteochondritis (Stattin et al., 2010) and skeletal dysplasia (Tompson et al., 2009). The IMPC viability screen reports that *Acan*$^{-/-}$ embryos are viable up to E18.5, but have a completely penetrant preweaning lethality phenotype. Other IMPC-assigned phenotypes include a reduced bone area composition and increased circulating cholesterol levels in adult heterozygous animals. We analysed six E14.5 *Acan*$^{-/-}$ specimens with LAMA, identifying 28 statistically significantly gene-level organ volume differences (Fig. 6A; Fig. S2A). Twelve of the significant organs are bones, including all those present in the Mouse Genome Informatics (MGI) annotations for this gene (Table 1). These include smaller cervical vertebrae, scapula, humerus, ribs and exoccipital bones (Fig. 6C,D). Importantly, LAMA was able to assign statistical significance to organ volume differences for 15 of the gene-level annotated organs to individual specimens. Of these organs, most were bones and had the largest mean volume difference, relative to wild-type controls. The specimen-level annotations are similar across all specimens, but with some exceptions, e.g. specimen 15.1a appeared largely unaffected (Fig. 6A). From a total of 42 specimen-level calls, there were four significantly different organ volumes highlighted by the specimen-level analysis that were not present at the gene level, including a tail vertebra annotation (Fig. 6A). The significant Jacobian determinant voxels indicated a smaller Meckel's cartilage in the mutants, which is not identified in previous literature or by our organ volume difference test, but is visibly smaller in these mutants (Fig. 6D). We did not find a significantly smaller lung volume difference for any of the lung lobes that would indicate hypoplasia as previously reported (Houghton et al., 1989), but visual inspection of the lung lobe segmentation labels indicated acceptable registration accuracy (Fig. S2B). The position of the lungs within the thoracic cavity, however, looked altered, possibly owing to changes in the thoracic cavity size. We did not identify the remaining previously reported phenotype of tracheal cartilage morphology either.

DISCUSSION

In this paper we presented a new automated computational phenotyping pipeline for dysmorphology detection in mutant mouse embryos along with a new E14.5 anatomical atlas. We have undertaken validation of the pipeline and provide insights on the effect of developmental substage and sample size on phenotyping results, as well as showing that LAMA can uncover previously reported and novel phenotypes from E14.5 IMPC knockout mice embryos. This pipeline and atlas will accelerate the automatic analysis of 3D embryo data at this developmental stage within the IMPC (where it is being adapted for other developmental

Development (2021) 148, dev192955. doi:10.1242/dev.192955

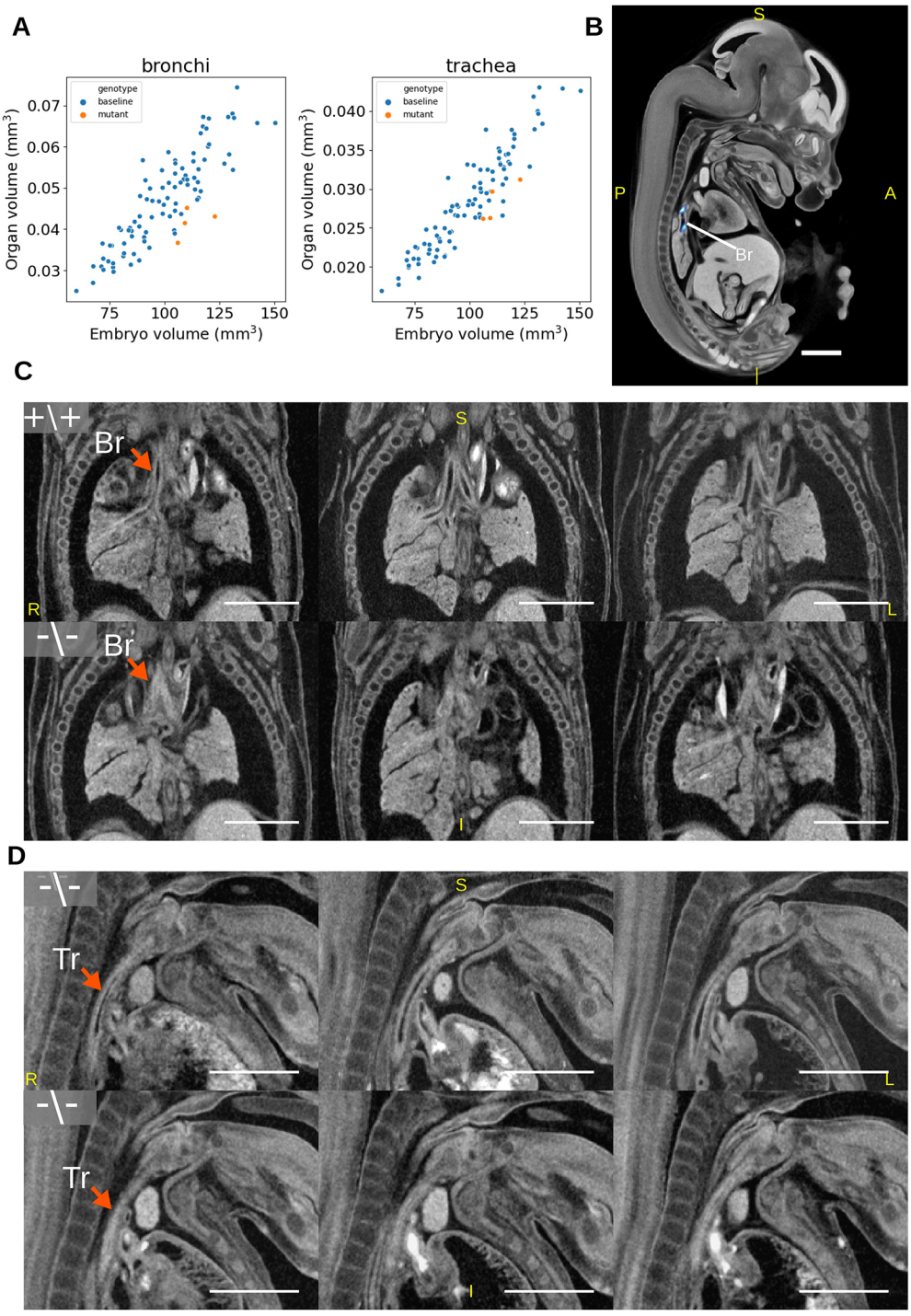

Fig. 5. Automated identification of pulmonary system phenotypes in a *Wfdc2* knockout mouse line. (A) Organ volume plots of statistically significant organs. (B) Jacobian determinant analysis t-statistics (FDR corrected to q<0.1) overlaid on the E14.5 population average. Blue regions indicate smaller bronchi in the mutants. (C,D) Rigidly aligned sections of *Wfdc2* mutants (bottom) and whole-embryo volume-matched wild-type rigidly aligned specimens (top), with arrows indicating the location of affected organs. (C) Coronal sections highlighting the bronchi. (D) Sagittal sections highlighting the trachea. Scale bars: 1 mm. A, anterior; P, posterior; S, superior; I, inferior; L, left; R, right; Tr, trachea; Br, bronchus.

stages), other large scale projects, and smaller challenge-led projects, driving forward the use of disease models in scientific discovery.

The image registration approach in LAMA provides significant advantages for high-throughput phenotyping compared to previous work (Wong et al., 2012, 2014). The major difference is the registration strategy in which all data is registered at once, directly towards a pre-made population average and atlas, which puts all the specimens from all mutant lines and controls into the same coordinate space. This contrasts with the approach of Wong et al. (2014) where specimens from a mutant line and a small number of wild types are registered into a unique coordinate space. With our approach, we were able to dramatically increase the number of wild type control embryos used in the phenotype analysis stage of the pipeline. This is due to the lack of a groupwise registration stage, which decreases computational expense, as well as the ability to reuse wild-type registered specimens in statistical analysis across many mutant lines. A further advantage of our approach is that it facilitates the distributed processing of images as each registration requires only the fixed population average and the moving specimen images.

We have shown that even after normalising organ volumes by whole-embryo volume (WEV), there remains a significant WEV effect, which we interpret as substage-dependent differential organ

Development (2021) 148, dev192955. doi:10.1242/dev.192955

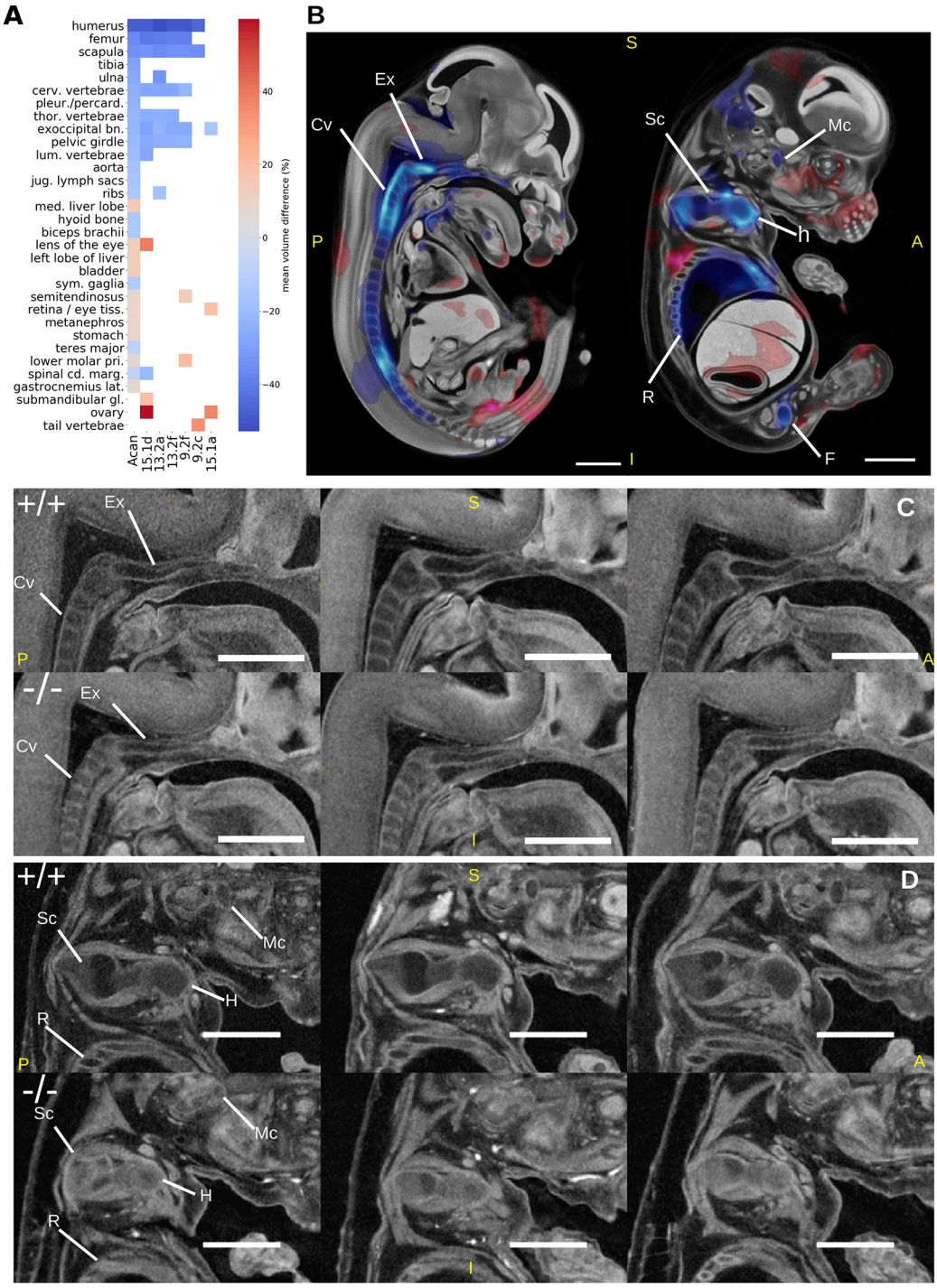

Fig. 6. Analysis of *Acan*$^{-/-}$ mutants by LAMA. (A) Heatmap showing statistically significant organ volume differences at the gene level and of individual specimens. The first column (Acan) are gene-level results and the remaining columns are results from individual specimens. Statistically significant organ volume differences for genotype effect are coloured according to mean normalised volume difference between wild type and mutants. White cells indicate no significant genotype effect. (B) Jacobian determinant t-statistics for genotype effect (FDR corrected to q<0.05) overlaid onto the E14.5 population average. (C,D) Illustrative sagittal slices from rigidly aligned *Acan*$^{-/-}$ mutants (bottom) and whole embryo volume-matched wild-type specimens (top) highlighting identified dysmorphology. Scale bars: 1 mm. Mc, Meckels's cartilage; Ex, exoccipital bone; Cv, cervical vertebrae; H, humerus; Sc, scapula; R, ribs; F, femur; A, anterior; P, posterior; S, superior; I, inferior.

growth rates. This variation in organ volume after normalising by WEV agrees with findings by Wong et al. (2012), who reported an organ volume standard deviation of 8-13% among wild-type E15.5 embryos after normalising organ volumes by WEV. Developmental substage variability is prevalent in our control set because we sought to include as many wild-type specimens as possible in order to increase the power of our analyses. However, we have shown that we can control for developmental substage variability by adding a WEV term in our linear regression model. Without this substage correction, the statistical model can lead to false-positive results

Development (2021) 148, dev192955. doi:10.1242/dev.192955

Table 1. Correspondence of phenotypes for *Acan*$^{-/-}$ mutant mice assigned by LAMA with phenotypes reported in the MGI database

Allelic composition	Genetic background	Annotated term	Identified with LAMA organ volume analysis	Identified with LAMA Jacobian analysis	Phenotype summary category	MGI publication reference
Acan$^{cmd\text{-}Bc}$/Acan$^{cmd\text{-}Bc}$	BALB/cGaBc-Acan$^{cmd\text{-}Bc}$	Cleft palate	No atlas label	Yes*	Craniofacial	J:58755
Acan$^{cmd\text{-}Bc}$/Acan$^{cmd\text{-}Bc}$	BALB/cGaBc-Acan$^{cmd\text{-}Bc}$	Short snout	No atlas label	Yes*	Craniofacial	J:58755
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Short basicranium	Smaller organ volume (exoccipital)*	Yes*	Craniofacial	J:30795
Acan$^{cmd\text{-}Bc}$/Acan$^{cmd\text{-}Bc}$	BALB/cGaBc-Acan$^{cmd\text{-}Bc}$	Distended abdomen	No atlas label	Yes*	Growth/size/ body region	J:58755
Acan$^{cmd\text{-}Bc}$/Acan$^{cmd\text{-}Bc}$	BALB/cGaBc-Acan$^{cmd\text{-}Bc}$	Protruding tongue	No	No	Growth/size/ body region	J:58755
Acan$^{cmd\text{-}Bc}$/Acan$^{cmd\text{-}Bc}$	BALB/cGaBc-Acan$^{cmd\text{-}Bc}$	Small thoracic cavity	Yes (pleural and pericardial cavities)*	Yes*	Growth/size/ body region	J:65316
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Flattened snout	Not in atlas	Yes*	Growth/size/ body region	J:5952
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Abnormal limb morphology	Yes*	Yes*	Limbs/digits/tail	J:30795
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Absent caudal vertebrae	Yes (specimen level)*	Yes*	Limbs/digits/tail	J:30795
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Brachydactyly	No (digits not analysed)	Some significant voxels in forelimb phalanges*	Limbs/digits/tail	J:30795
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Enlarged liver	Yes*	Yes*	Liver/biliary system	J:5952
Acan$^{cmd\text{-}Bc}$/Acan$^{cmd\text{-}Bc}$	BALB/cGaBc-Acan$^{cmd\text{-}Bc}$	Pulmonary hypoplasia	No	No	Respiratory system	J:65316
Acan^{cmd}/Acan^{cmd}	involves: STOCK T t^{low} Itpr3^{tf}	Abnormal lung morphology	No	Yes*	Respiratory system	J:23353
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Abnormal tracheal cartilage morphology	No	Yes	Respiratory system	J:5952
Acan$^{tm1b(EUCOMM)Hmgu}$/ Acan^{+}	C57BL/6N-Acan$^{tm1b(EUCOMM)Hmgu}$/H	Abnormal bone structure	Yes*	Yes*	Skeleton	J:211773
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Abnormal cartilage development	Yes (abnormal bones)*	Yes*	Skeleton	J:30795
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Abnormal craniofacial bone morphology	No	Yes*	Skeleton	J:30795
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Abnormal rib morphology	Yes*	Yes*	Skeleton	J:30795
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Abnormal trabecular bone morphology	Yes*	Yes*	Skeleton	J:30795
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Abnormal vertebrae morphology	Yes*	Yes*	Skeleton	J:5952
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Decreased length of long bones	Yes*	Yes*	Skeleton	J:5952
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Short femur	Yes*	Yes*	Skeleton	J:5952
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Short humerus	Yes*	Yes*	Skeleton	J:5952
Acan^{cmd}/Acan^{cmd}	STOCK T t^{low} Itpr3^{tf}	Short vertebral column	Yes*	Yes*	Skeleton	J:5952

*Concordance with recorded phenotype.
A subset of MGI annotations is included (see Materials and Methods).

(Fig. 3A,B). WEV is a convenient staging metric as it can be easily calculated after embryo spatial normalisation, and is correlated with crown-rump length and whole-body weight, both previously used to stage embryos (Dagg, 1963; Peterka et al., 2002). Alternative methods that rely on the appearance of external features of the embryo (Theiler, 1989; Boehm et al., 2011; Geyer et al., 2017) may be more accurate, but these methods have yet to be automated for whole embryo 3D images. An alternative approach to automated staging involves spatially and temporally normalising embryos to a 4D (3D+time) dimensional population average, with developmental stage as the temporal dimension (Wong et al., 2015). This method was shown to provide high-resolution staging information and could identify developmental asynchrony across all organs. This 4D atlas was not used in this study as the latest developmental stage in the atlas, E14.0, does not overlap with our E14.5 dataset. The generation of a new 4D micro-CT atlas covering the E14.5 stage

Development (2021) 148, dev192955. doi:10.1242/dev.192955

would require the breeding of a large number of embryos at various gestational timepoints, which is not currently possible. With our analysis, we identified 81 organs that show a statistically significant WEV effect (Table S2), which likely reflects different relative growth rates of various organs at different E14.5 developmental substages. This represents the most detailed embryo-wide data of E14.5 substage-specific organ growth rates that we are currently aware of.

Stratifying our wild-type specimens by sex enabled us to test the ability of LAMA to identify specific anatomical differences, as males and females are anatomically similar except for clear gonad differences. We found that we were able to uncover statistically significant gonad volume differences while keeping false positives low. To assess the performance on mutant data, we tested LAMA on two IMPC-generated knockout lines. The first (*Wfdc2*$^{-/-}$) was predicted to display specific pulmonary abnormalities; the second (*Acan*$^{-/-}$) to produce severely dysmorphic phenotypes across the whole embryo. Our automated analysis of E14.5 *Wfdc2*$^{-/-}$ embryos revealed two significantly smaller organ volumes: those of the trachea and bronchi, which are novel findings for this gene. These overlap broadly with the locations of the previously reported pulmonary-specific abnormalities in *Wfdc2*$^{-/-}$ mice, including the absence of mature club cells from the bronchi and trachea, postnatally collapsed lung, reduced lung surfactant levels (Nakajima et al., 2019) and alveoli abnormalities (Zhang et al., 2020). The novel phenotypes reported here bring forward the time when gross abnormalities due to loss of *Wfdc2* first become visible during embryo development (previously postnatally), and therefore add new temporal information to the role of *Wfdc2* in pulmonary development. In addition, LAMA could recapitulate the majority of previously reported *Acan*$^{-/-}$ phenotypes, which will greatly speed up the annotation of this severely affected mutant. There were four significant organ volume differences for *Acan*$^{-/-}$ that have not previously been reported. It is possible that these are novel phenotypes, but it is difficult to confirm this by inspection of the micro-CT images alone. It is possible that the apparent abnormality is due to proximity to actual severe dysmorphology, which, during the registration process was warped along with the abnormal organ. For example, two of these organs without previous reports, sympathetic ganglia and the spinal cord marginal layer, are located close to the affected vertebra. Another reason for this discrepancy could be the different background strains used in this study and the previous studies.

As efforts are under way to reduce the numbers of animals used in scientific experiments, we wanted to test whether LAMA could identify dysmorphology with low mutant sample numbers. In addition, being able to use low sample numbers would let us investigate the effects of incomplete penetrance and variable expressivity, as well as providing phenotype data from mutant lines where many specimens do not reach the developmental stage being tested. To begin to answer this, in the sex difference test we show that increasing the control sample size from 10 to 49 greatly increases the power to detect anatomical differences, and that by using many controls, it is possible to sometimes uncover phenotype information even with a low mutant sample size of two. We also show that for six *Acan*$^{-/-}$ specimens, 42 specimen-level organ difference annotations were generated (Fig. 6A). Support for these specimen-level annotations comes from the fact that they mostly overlapped with the gene-level annotated organs (only 4/42 did not). These specimen-level annotations were strongly enriched for organs that had the largest volume difference relative to the control mean, which is expected due the reduced power of these tests being unable to detect smaller volume differences. There were differences between the annotations of the individual specimens, with specimens 15.1a and 9.2c lacking many of the bone annotations (Fig. 6A). These results show that, by using a large wild-type control sample number, useful phenotype information can be obtained when analysing low *N* mutant lines or even specimens individually.

Developmental delay is a common phenotype of knockout mice and such animals could pose a problem for this analysis as they could be smaller than the smallest mice in the baseline controls. In future, one way round this would be to include some E14.0 or E13.5 animals into our wild-type control set to extend the range to where we might expect to see developmental delay and improve the reliability of the model. In lieu of this, the whole embryo volume z-score (standard deviations from the wild-type mean) of each specimen, which is reported by LAMA, can be used to identify, and exclude, potentially delayed animals from analysis.

The choice of registration parameters can involve a compromise of balancing good registration accuracy on some organs with misregistration at others. We found that gonad registration, for example, could be improved by removing most of the registration constraints, but this led to over-warping at the heart. One solution to this could be to use multiple sets of registration parameters, each optimised to different parts of the atlas. Alternatively, approaches that directly segment organs without registration have been previously proposed (Yan et al., 2017; Ashish and Brusniak, 2018) but only on a limited number of organs, and these have yet to be applied to embryonic mice. LAMA is able to perform statistical analysis on the voxel intensities of the spatially normalised images, but we found that the image data used in this study contained large differences in intensity profiles that were possibly due to the different users and imaging equipment involved in image acquisition over a number of years. For this reason, we have concentrated our current analysis on organ volume differences and Jacobian determinant analysis, which are both more robust to varying intensity profiles. Future work will look towards employing more sophisticated image normalisation methods and exploring the use of other image features, such as textures, that may be less susceptible to intensity profile differences. Current work includes optimising registration parameters for E15.5 and E18.5 developmental timepoints (see Fig. S3 for current population average images), the latter being a key developmental time point for the analysis of gene mutations that result in perinatal lethality and subviability. Earlier stages, such as E12.5, may also be amenable to this analysis, but even earlier stages such as E9.5 may prove difficult for a registration-based approach due to the rapid developmental changes at this time point and extreme dysmorphology that is often caused by mutations that are lethal early in development. LAMA has been applied successfully to mouse bones that were dissected and scanned separately (N.R.H., unpublished) and can readily be adapted to other model systems where good registration between subjects is possible. This requires no software changes to the pipeline and only the registration parameters need to be optimised. Finally, we believe that the tools, resources and insights introduced in this article will accelerate the use of the rapidly increasing amounts of mouse embryo image data at the IMPC and within the wider mouse developmental biology community.

MATERIALS AND METHODS

Mice

All animals were housed and maintained in the Mary Lyon Centre at the MRC Harwell Institute under specific pathogen-free (SPF) conditions in individually ventilated cages adhering to environmental conditions, as outlined in the Home Office Code of Practice. All animal studies were

Development (2021) 148, dev192955. doi:10.1242/dev.192955

licensed by the Home Office under the Animals (Scientific Procedures) Act 1986 Amendment Regulations 2012 (SI 4 2012/3039), UK, and additionally approved by the Institutional Ethical Review Committee. The Acan^{tm1b} allele was obtained by cre deletion of C57BL/6N-Acan$^{tm1a(EUCOMM)Hmgu}$/H (EM:10224) mice as described previously (Birling et al., 2019 preprint). Homozygous mutants are named *Acan*$^{-/-}$ here. The C57BL/6NTac-Wfdc2$^{em1(IMPC)H}$/H (EM:11407, homozygous mutants named *Wfdc2*$^{-/-}$ here) was obtained by genome editing as described previously (Mianné et al., 2017). Lines were maintained by crossing heterozygous animals with inbred C57BL/6N wild-type animals. Mice were euthanised by Home Office Schedule 1 methods.

Micro-CT imaging of whole embryos

E14.5 female mice were sacrificed by cervical dislocation and the uterine horns removed into ice-cold phosphate-buffered saline (PBS). Embryos were extracted and a piece of yolk sac collected for genotype analysis. Embryos were fixed in 4% paraformaldehyde (PFA) at 4°C and left overnight. After fixation, the samples were washed and stored in PBS at 4°C. For staining, samples were rinsed in distilled H_2O for 10 min before being submerged in 50% Lugol's solution and protected from light. Embryos were then left in the contrast agent for 2 days. Following staining, embryos were washed in distilled H_2O for at least 1 h, embedded in 1% agarose (in distilled H_2O) and left at room temperature for a minimum of 2 h.

High resolution micro-CT images (SkyScan 1172, Bruker) of agarose-embedded embryos were acquired at a source voltage of 70 kV, with the current set at maximum (~100 mA). Specimens were imaged, in a standard orientation, at 3 µm with a 0.5 mm aluminium filter. X-ray projections were acquired at 0.25° increments, and reconstructed using the Feldkamp algorithm (Feldkamp et al., 1984) provided by NRecon (Bruker). Ring artefact corrections were applied as necessary. Reconstructions were automatically cropped to remove background and scaled to 14 µm isotropic voxels using the HARP software (Brown et al., 2018).

Phenotyping pipeline implementation

The image registration pipeline was written in the Python programming language (Python 3.6+), adapting a modular design that allows for individual components (registration, inversion, statistics, etc.) to be run either sequentially or independently using simple TOML configuration files. Individual image registrations are performed using the elastix toolkit (Klein et al., 2010; Shamonin et al., 2014). The linear model analysis is implemented in R. All code is available on Github (https://github.com/mpi2/LAMA) and is tested to work on Ubuntu versions 18.04 and 20.4, as well as Windows 10. The use of interactive shell scripts that show how to use LAMA on a real dataset is described at https://github.com/mpi2/LAMA/wiki/walkthroughs. To make the installation of LAMA as easy as possible and to help data reproducibility, LAMA is available via the PyPi Python package repository (https://pypi.org/project/lama-phenotype-detection).

Population average construction

Micro-CT images from 16 specimens of both sexes were used in the creation of the population average through a groupwise multi-level and multi-resolution registration process. For the first level, an initial fixed image was chosen at random and all other images were rigidly registered onto it. The registered images were averaged creating a rigidly aligned blurry average. This population average is invariant to the choice of the initial fixed image, because the composition of rigid transformations entails only a change of pose. For the second registration level, the rigid registration outputs were affinely registered onto the blurry average. A new (less) blurry average was computed from the outputs of this affine step. For the last registration levels, the process of alignment to the group average followed by recalculation of the average was repeated, using B-spline transformations, which allow local nonlinear deformations. We used a five-level Gaussian pyramid with increasing resolution for the images and five corresponding B-spline levels with decreasing control point spacing, with a final grid spacing of 8 voxels, to sequentially align coarser to finger anatomical structures (Fig. 1A). The parameter file for our population average and the final population average image available for download at https://www.doi.org/10.5281/zenodo.4559800.

Image segmentation/E14.5 Atlas creation

Key anatomical structures within the E14.5 population average were identified manually by referencing the online digitised mouse atlas (Graham et al., 2015), which itself is based on *The Atlas of Mouse Development* (Kaufman, 1992). Structures that could be identified were restricted to those that showed good contrast and resolution within the population average. ITK-SNAP was used (Yushkevich et al., 2006; www.itksnap.org) to create the segmentations, using a variety of semi-automated and manual methods suited to the size and complexity of each of these anatomical structures, and combined into a single label file. These structures were then merged with some previous segmentations of brain structures derived from an E15.5 atlas (Wong et al., 2012) to give a total of 184 anatomical components. Small, spindly labels in the atlas were flagged by calculating a 3D euclidean distance transform for each label using the Python package edt (https://github.com/seung-lab/euclidean-distance-transform-3d) and flagging labels with a value <1.5. The resulting atlas and associated metadata files are available for download at https://www.doi.org/10.5281/zenodo.4559800.

Generation of data for phenotype detection

Baseline and mutant specimens were registered onto the previously created population average image. The outputs of this registration include the rigid, similarity, affine and B-spline spatial transformations, co-registered images and the Jacobian determinants det(jac) of the composition of the B-spline transformations, a scalar field that describes the local volume change in each voxel. Statistical analysis of these outputs (described below) produces statistical parametric heat maps that can be overlaid onto the population average image or superimposed onto the input images by applying the inverse of the spatial transformations. The statistical parametric heat maps can be viewed with the Volume Phenotype Viewer (VPV) (Brown et al., 2018; https://github.com/mpi2/vpv).

Statistical analysis

Multiple linear regression analysis was conducted in R (www.R-project.org) using the lm() function from the MASS package (Venables and Ripley, 2002). Benjamini-Hochberg FDR (BH-FDR) correction was carried out using the padjust R package (Benjamin and Hochberg, 1995).

Voxel-level data

Jacobian determinants, generated at each voxel within the population average mask, provide information about how the registration has behaved locally. The scalar value of the Jacobian determinant at a given location is the factor by which that region has expanded [$\det(J_F)>1$] or shrunk [$\det(J_F)<1$] in volume during registration. This approach, known as tensor-based morphometry, can be used to reveal biologically significant localised shape or size changes within a population (Ashburner and Friston, 2000). To account for small registration inaccuracies, a Gaussian blur of full-width-half-maximum (FWHM) 100 µm is applied to voxel-level data. Each voxel is fitted to a linear model voxel~genotype+WEV, where WEV stands for whole embryo volume. We use WEV as a proxy for developmental stage, and the addition of it as a fixed effect controls for changes that are due to developmental stage only. To account for multiple testing, the resulting *P*-value maps are corrected using the Benjamini Hochberg method. The final parametric heat maps are made by thresholding the t-static volume at $q>0.05$. This is output as a 3D image that can be overlaid onto the target or registered image in VPV.

Whole-organ volume analysis

Organ volumes and whole-embryo volumes are derived for each specimen. Significant volume differences are detected by using an organ-specific linear model organ volume/WEV~genotype+WEV. We apply a permutation-based approach (for each organ) for multiple testing correction described previously by Hrabě de Angelis et al. (2015). To summarise, organ-specific null distributions are generated by sampling synthetic mutants from the baseline data in such a way as to match the distribution of the number of mutant specimens per line. For example, if there are 40 mutant lines and 10 of these have $n=3$ and the other 10 have $n=4$, we create 50% synthetic lines with $n=3$ and 50% with $n=4$. Synthetic mutants and baseline controls are fitted to the linear model as described

Development (2021) 148, dev192955. doi:10.1242/dev.192955

above and the genotype effect *P*-values are computed. Alternative distributions are made by computing genotype effect *P*-values from testing the real mutants of each line. To obtain a dataset-wide *P*-value threshold per organ, combined null and alternative *P*-values for the organ are ranked and a descending *P*-value threshold search is conducted starting at $P=0.05$ until a threshold is found where the proportion of alternative *P*-values under the threshold divided by the proportion of null *P*-values under the threshold is <0.05. Mutant *P*-values below this threshold are assigned as significant, which sets the organ-specific FDR to 5%.

Detection of sex-specific differences

For the initial experiment that tested for a sex effect on organ volume (Fig. 4A,B), organ volumes from all available males and females were fitted to the linear model organ volume/WEV∼sex+WEV. False discovery correction was performed across all organs using the Benjamini Hochberg method, as it was not possible to permute the data as all the data was fitted to the model. For the organ volume analysis that tested varying sample numbers (Fig. 4C,D) organ-specific *P*-value thresholds were generated using the permutation based method described above. Only experiments where combinations of male and female sample size allowed at least 500 unique null permutations were included. Jacobian determinant analysis (Fig. 4E,F) was carried out as described above.

Variable penetrance and low *N*

To identify potentially variable expressivity or incomplete penetrance of organ volume phenotypes, LAMA routinely performs what we term 'specimen-level analysis'. After the analysis of a mutant line as a group, each individual specimen is analysed singly, i.e. the organ volume, or voxel value, for a single specimen is fitted to a linear model along with baseline controls to obtain a specimen-level *P*-value for the genotype effect. Owing to the reduction in power from using one specimen, we have set the FDR threshold for the specimen-level analysis to 20%. The voxel-level data are processed similarly to the gene-level voxel data and the voxels are thresholded to an FDR of 5%.

Optimisation and quality control

At each level and resolution of the registration process, the similarity metric output by elastix is plotted against iteration number, allowing the user to visually decide when the optimisation process has converged, and set an optimal number of iterations in future.

Image registration can sometimes fail to produce acceptable results, e.g. when the moving image is over fitted to the fixed image, producing unrealistic warping. To check for issues such as these, after each registration stage an additional HTML report was generated that contains mid-sagittal slices for each registered image for rapid quality control. Another issue that can be encountered using B-spline-based image registration is folding of the deformation field, which prevents topology preservation and has no inverse transformation. This problem can be detected by the presence of negative Jacobian determinants. In that case, only pixels with negative Jacobian determinants are displayed, allowing the user to quickly identify problematic regions within the registered images. The registration stage of the analysis is the most time-consuming part of the pipeline, and so it is a requirement to be able to optimise registration parameters, especially within a high-throughput context.

Comparing *Acan* phenotypes found by LAMA to known phenotypes

Tables of known phenotypes were generated by querying MGI phenotype pages for the gene of interest *Acan:* www.informatics.jax.org/marker/phenotypes/MGI:99602. Only phenotypes generated from homozygous null strains were included to aid in the comparison. Duplicate and redundant phenotypes (e.g. abnormal bone structure if more specific bone phenotypes were present) were removed. Phenotypes that might not translate to a gross anatomical dysmorphology that could be potentially detected by LAMA (e.g. deafness) were also removed.

Acknowledgements

We thank James Cleak and Zsombor Szoke-Kovacs for generating image data and helpful discussions; George Nicholson and Hugh Morgan for statistics advice; and the husbandry team from the Mary Lyon Centre for generation and maintenance of all the mouse specimens used in the study.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: N.R.H., R.C., M.D.W., M.C.v.E., R.M.H., L.T., S.W., S.D.B., H.W., A.-M.M.; Methodology: N.R.H., S.V., R.C., M.D.W., M.C.v.E., R.M.H., H.W.; Software: N.R.H., J.M.B.; Validation: N.R.H., J.M.B., H.W.; Formal analysis: N.R.H.; Investigation: N.R.H., S.V., C.A., S.J.; Resources: N.R.H.; Data curation: N.R.H., S.V., C.A.; Writing - original draft: N.R.H.; Writing - review & editing: N.R.H., S.V., R.C., J.M.B., L.T., H.W., A.-M.M.; Visualization: N.R.H.; Supervision: H.W., A.-M.M.; Project administration: A.-M.M.; Funding acquisition: A.-M.M.

Funding

The work was supported by a National Institutes of Health grant (U54 HG006370-01), by a Medical Research Council Strategic Award and by a Medical Research Council-funded programme (MC_U142684171). Deposited in PMC for immediate release.

Data availability

The E14.5 population average, label map metadata and LAMA configuration files are available for download from https://www.doi.org/10.5281/zenodo.4559800.

Supplementary information

Supplementary information available online at
https://dev.biologists.org/lookup/doi/10.1242/dev.192955.supplemental

Peer review history

The peer review history is available online at
https://dev.biologists.org/lookup/doi/10.1242/dev.192955.reviewer-comments.pdf

References

Adams, D., Baldock, R., Bhattacharya, S., Copp, A. J., Dickinson, M., Greene, N. D. E., Henkelman, M., Justice, M., Mohun, T., Murray, S. A. et al. (2013). Bloomsbury report on mouse embryo phenotyping: recommendations from the IMPC workshop on embryonic lethal screening. *Dis. Model. Mech.* **6**, 571-579. doi:10.1242/dmm.011833

Ashburner, J. and Friston, K. J. (2000). Voxel-based morphometry - The methods. *NeuroImage* **11**, 805-821. doi:10.1006/nimg.2000.0582

Ashish, N. and Brusniak, M.Y. (2018). Automated mouse organ segmentation: a deep learning based solution. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 236-243. New York, NY, USA: Association for Computing Machinery. doi:10.1145/3233547.3233552

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bingle, L., Singleton, V. and Bingle, C. D. (2002). The putative ovarian tumour marker gene HE4 (WFDC2), is expressed in normal tissues and undergoes complex alternative splicing to yield multiple protein isoforms. *Oncogene* **21**, 2768-2773. doi:10.1038/sj.onc.1205363

Birling, M.-C., Yoshiki, A., Adams, D. J., Ayabe, S., Beaudet, A. L., Bottomley, J., Bradley, A., Brown, S. D. M., Bürger, A., Bushell, W. et al. (2019). A resource of targeted mutant mouse lines for 5,061 genes. *bioRxiv*.

Boehm, B., Rautschka, M., Quintana, L., Raspopovic, J., Jan, Ž. and Sharpe, J. (2011). A landmark-free morphometric staging system for the mouse limb bud. *Development* **138**, 1227-1234. doi:10.1242/dev.057547

Bowl, M. R., Simon, M. M., Ingham, N. J., Greenaway, S., Santos, L., Cater, H., Taylor, S., Mason, J., Kurbatova, N., Pearson, S. et al. (2017). A large scale hearing loss screen reveals an extensive unexplored genetic landscape for auditory dysfunction. *Nat. Commun.* **8**, 886. doi:10.1038/s41467-017-00595-4

Brown, S. D. M. and Moore, M. W. (2012). The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. *Mamm. Genome* **23**, 632-640. doi:10.1007/s00335-012-9427-x

Brown, J. M., Horner, N. R., Lawson, T. N., Fiegel, T., Greenaway, S., Morgan, H., Ring, N., Santos, L., Sneddon, D., Teboul, L. et al. (2018). A bioimage informatics platform for high-throughput embryo phenotyping. *Brief. Bioinform.* **19**, 41-51.

Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., Richardson, J. E., Anagnostopoulos, A., Asabor, R., Baldarelli, R. M., Beal, J. S., Bello, S. M. et al. (2019). Mouse Genome Database (MGD) 2019. *Nucleic Acids Res.* **47**, D801-D806. doi:10.1093/nar/gky1056

Cacheiro, P., Muñoz-Fuentes, V., Murray, S. A., Dickinson, M. E., Bucan, M., Nutter, L. M. J., Peterson, K. A., Haselimashhadi, H., Flenniken, A. M., Morgan, H. et al. (2020). Human and mouse essentiality screens as a resource

for disease gene discovery. *Nat. Commun.* **11**, 655. doi:10.1038/s41467-020-14284-2

Cleary, J. O., Modat, M., Norris, F. C., Price, A. N., Jayakody, S. A., Martinez-Barbera, J. P., Greene, N. D. E., Hawkes, D. J., Ordidge, R. J., Scambler, P. J. et al. (2011). Magnetic resonance virtual histology for embryos: 3D atlases for automated high-throughput phenotyping. *Neuroimage* **54**, 769-778. doi:10.1016/j.neuroimage.2010.07.039

Dagg, C. P. (1963). The interaction of environmental stimuli and inherited susceptibility to congenital deformity. *Integr. Comp. Biol.* **3**, 223-233. doi:10.1093/icb/3.2.223

Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., Meehan, T. F., Weninger, W. J., Westerberg, H., Adissu, H. et al. (2016). High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508-514. doi:10.1038/nature19356

Feldkamp, L. A., Davis, L. C. and Kress, J. W. (1984). Practical cone-beam algorithm. *J. Optic. Soc. Am. A* **1**, 612-612. doi:10.1364/JOSAA.1.000612

Geyer, S. H., Reissig, L., Rose, J., Wilson, R., Prin, F., Szumska, D., Ramirez-Solis, R., Tudor, C., White, J., Mohun, T. J., et al. (2017). A staging system for correct phenotype interpretation of mouse embryos harvested on embryonic day **14** (E14.5). *Journal of Anatomy* **230**, 710-719. doi:10.1111/joa.12590

Graham, E., Moss, J., Burton, N., Armit, C., Richardson, L. and Baldock, R. (2015). The atlas of mouse development eHistology resource. *Development* **142**, 1909-1911. doi:10.1242/dev.124917

Hayamizu, T. F., Wicks, M. N., Davidson, D. R., Burger, A., Ringwald, M. and Baldock, R. A. (2013). EMAP/EMAPA ontology of mouse developmental anatomy: 2013 update. *J. Biomed. Semantic.* **4**, 15-15. doi:10.1186/2041-1480-4-15

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W. and Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *J. Comput. Assist. Tomogr.* **22**, 324-333. doi:10.1097/00004728-199803000-00032

Houghton, M. J., Carey, J. C. and Seegmiller, R. E. (1989). Pulmonary hypoplasia in mice homozygous for the cartilage matrix deficiency (Cmd) gene: a model for a human congenital disorder. *Pediatr. Pathol.* **9**, 501-512. doi:10.3109/15513818909026909

Hrabě de Angelis, M., Nicholson, G., Selloum, M., White, J. K., Morgan, H., Ramirez-Solis, R., Sorg, T., Wells, S., Fuchs, H., Fray, M. et al. (2015). Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. *Nat. Genet.* **47**, 969-978. doi:10.1038/ng.3360

Kaufman, M. H. (1992). *The Atlas of Mouse Development*. Academic Press.

Klein, S., Staring, M., Murphy, K., Viergever, M. A. and Pluim, J. P. W. (2010). elastix: a toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging* **29**, 196-205. doi:10.1109/TMI.2009.2035616

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921. doi:10.1038/35057062

Li, J., Chen, H., Mariani, A., Chen, D., Klatt, E., Podratz, K., Drapkin, R., Broaddus, R., Dowdy, S. and Jiang, S.-W. (2013). HE4 (WFDC2) promotes tumor growth in endometrial cancer cell lines. *Int. J. Mol. Sci.* **14**, 6026-6043. doi:10.3390/ijms14036026

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S. et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22. doi:10.1186/s13059-014-0560-6

Lloyd, K. C. K., Adams, D. J., Baynam, G., Beaudet, A. L., Bosch, F., Boycott, K. M., Braun, R. E., Caulfield, M., Cohn, R., Dickinson, M. E. et al. (2020). The Deep Genome Project. *Genome Biol.* **21**, 18. doi:10.1186/s13059-020-1931-9

Meehan, T. F., Conte, N., West, D. B., Jacobsen, J. O., Mason, J., Warren, J., Chen, C.-K., Tudose, I., Relac, M., Matthews, P. et al. (2017). Disease model discovery from 3,328 gene knockouts by the International Mouse Phenotyping Consortium. *Nat. Genet.* **49**, 1231-1238. doi:10.1038/ng.3901

Mianné, J., Codner, G. F., Caulder, A., Fell, R., Hutchison, M., King, R., Stewart, M. E., Wells, S. and Teboul, L. (2017). Analysing the outcome of CRISPR-aided genome editing in embryos: Screening, genotyping and quality control. *Methods* **121-122**, 68-76. doi:10.1016/j.ymeth.2017.03.016

Miyake, T., Cameron, A. M. and Hall, B. K. (1996). Detailed staging of inbred C57BL6 mice between Theiler's [1972] stages 18 and 21 (11-13 days of gestation) based on craniofacial development. *J. Craniofac. Genet. Dev. Biol.* **16**, 1-31.

Moore, B. A., Leonard, B. C., Sebbag, L., Edwards, S. G., Cooper, A., Imai, D. M., Straiton, E., Santos, L., Reilly, C., Griffey, S. M. et al. (2018). Identification of genes required for eye development by high-throughput screening of mouse knockouts. *Commun. Biol.* **1**, 236. doi:10.1038/s42003-018-0226-0

Nakajima, K., Ono, M., Radović, U., Dizdarević, S., Tomizawa, S.-I., Kuroha, K., Nagamatsu, G., Hoshi, I., Matsunaga, R., Shirakawa, T. et al. (2019). Lack of whey acidic protein (WAP) four-disulfide core domain protease inhibitor 2 (WFDC2) causes neonatal death from respiratory failure in mice. *Dis. Model. Mech.* **12**, dmm040139. doi:10.1242/dmm.040139

Oprea, T. I. (2019). Exploring the dark genome: implications for precision medicine. *Mamm. Genome* **30**, 192-200. doi:10.1007/s00335-019-09809-0

Peterka, M., Lesot, H. and Peterková, R. (2002). Body weight in mouse embryos specifies staging of tooth development. *Connect. Tissue Res.* **43**, 186-190. doi:10.1080/03008200290000673

Shamonin, D. P., Bron, E. E., Lelieveldt, B. P. F., Smits, M., Klein, S. and Staring, M. (2014). Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Front. Neuroinform.* **7**, 50. doi:10.3389/fninf.2013.00050

Stattin, E.-L., Wiklund, F., Lindblom, K., Önnerfjord, P., Jonsson, B.-A., Tegner, Y., Sasaki, T., Struglics, A., Lohmander, S., Dahl, N. et al. (2010). A missense mutation in the aggrecan C-type lectin domain disrupts extracellular matrix interactions and causes dominant familial Osteochondritis dissecans. *Am. J. Hum. Genet.* **86**, 126-137. doi:10.1016/j.ajhg.2009.12.018

Theiler, K. (1989). *Procedure, in: The House Mouse*, pp. 1-2. Berlin, Heidelberg: Springer Berlin Heidelberg.

Tompson, S. W., Merriman, B., Funari, V. A., Fresquet, M., Lachman, R. S., Rimoin, D. L., Nelson, S. F., Briggs, M. D., Cohn, D. H. and Krakow, D. (2009). A recessive skeletal dysplasia, SEMD aggrecan type, results from a missense mutation affecting the C-type lectin domain of aggrecan. *Am. J. Hum. Genet.* **84**, 72-79. doi:10.1016/j.ajhg.2008.12.001

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th edn. New York: Springer.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562. doi:10.1038/nature01262

Wilson, R., McGuire, C., Mohun, T., Adams, D., Baldock, R., Bhattacharya, S., Collins, J., Fineberg, E., Firminger, L., Galli, A. et al. (2016). Deciphering the mechanisms of developmental disorders: phenotype analysis of embryos from mutant mouse lines. *Nucleic Acids Res.* **44**, D855-D861. doi:10.1093/nar/gkv1138

Wilson, R., Geyer, S. H., Reissig, L., Rose, J., Szumska, D., Hardman, E., Prin, F., McGuire, C., Ramirez-Solis, R., White, J. et al. (2017). Highly variable penetrance of abnormal phenotypes in embryonic lethal knockout mice. *Wellcome Open Res.* **1**, 1-1. doi:10.12688/wellcomeopenres.9899.2

Wong, M. D., Dorr, A. E., Walls, J. R., Lerch, J. P. and Henkelman, R. M. (2012). A novel 3D mouse embryo atlas based on micro-CT. *Development* **139**, 3248-3256. doi:10.1242/dev.082016

Wong, M. D., Maezawa, Y., Lerch, J. P. and Henkelman, R. M. (2014). Automated pipeline for anatomical phenotyping of mouse embryos using micro-CT. *Development* **141**, 2533-2541. doi:10.1242/dev.107722

Wong, M. D., Van Eede, M. C., Spring, S., Jevtic, S., Boughner, J. C., Lerch, J. P. and Mark Henkelman, R. (2015). 4d atlas of the mouse embryo for precise morphological staging. *Development* **142**, 3583-3591. doi:10.1242/dev.125872

Wright, I. C., McGuire, P. K., Poline, J. B., Travere, J. M., Murray, R. M., Frith, C. D., Frackowiak, R. S. J. and Friston, K. J. (1995). A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *NeuroImage* **2**, 244-252. doi:10.1006/nimg.1995.1032

Yan, D., Zhang, Z., Luo, Q. and Yang, X. (2017). A novel mouse segmentation method based on dynamic contrast enhanced micro-CT images. *PLoS ONE* **12**, e0169424. doi:10.1371/journal.pone.0169424

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C. and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* **31**, 1116-1128. doi:10.1016/j.neuroimage.2006.01.015

Zamyadi, M., Baghdadi, L., Lerch, J. P., Bhattacharya, S., Schneider, J. E., Henkelman, R. M. and Sled, J. G. (2010). Mouse embryonic phenotyping by morphometric analysis of MR images. *Physiol. Genomics* **42 A**, 89-95. doi:10.1152/physiolgenomics.00091.2010

Zhang, T., Long, H., Li, J., Chen, Z., Wang, F. and Jiang, S.-W. (2020). WFDC2 gene deletion in mouse led to severe dyspnea and type-I alveolar cell apoptosis. *Biochem. Biophys. Res. Commun.* **522**, 456-462. doi:10.1016/j.bbrc.2019.11.011