APPENDIX

In Vivo Mn-Enhanced MRI Analysis of Mouse Brain Development:

Methodology for Registration-Based 4D Analyses

This chapter is being prepared for publication as:

Jason P. Lerch*, Kamila U. Szulc*, Miriam Friedel, Brian J. Nieman, and Daniel H. Turnbull. *In vivo* Mn-enhanced MRI analysis of mouse brain development: Methodology for registration-based 4D analyses

*Equal contribution

Attribution of Data: This appendix was written primarily by JPL, with editing by MF, DHT, BJN and myself. I designed the imaging experiments, together with DHT and BJN. I was primarily responsible for the acquisition data in this appendix, including optimization of imaging protocols. JPL and MF performed most of the analysis with help from me. JPL and myself provided ideas for analysis in this chapter.

A.1 ABSTRACT

In the companion paper (Part I), a data set was introduced, consisting of longitudinal Mn-enhanced MRI (MEMRI) images of the developing mouse brain, covering the time points between postnatal days 1 to 11. In this paper, we describe the methods used to analyze voxel-wise and region-wise changes in brain volume. The analysis of densely sampled longitudinal brain development data, in which brain growth between early and late time points is sufficiently advanced to make them too different to easily align to each other, poses a unique set of challenges for automated analysis approaches. The core methods consist of image registration to define the transform mapping temporally adjacent scans to each other, transform concatenation to bring all scans into a common space, and computation of the Jacobian determinant of those concatenated transforms to obtain measures of volume change. Linear mixed-effects models are used to compute the relationship between time and local brain growth, and log-likelihood tests between nested models employed to make inferences about factors that describe or influence brain development.

A.2 INTRODUCTION

The advent of *in vivo* MRI opened the unprecedented possibility of non-invasively studying brain development. In both human and small animal studies, we can now repeatedly scan the same individual at different times during their development, starting as early as gestational stages (Berrios-Otero et al., 2012; Nieman and Turnbull, 2010; Studholme, 2011) and continuing into later postnatal stages (Thompson et al., 2009; Voineskos et al., 2011). This data has provided new insights into patterns of cortical development (Raznahan et al., 2011), genetic versus environmental influences on brain development (Lenroot and Giedd, 2011; Schmitt et al., 2009), and covariance patterns in brain development (Alexander-Bloch et al., 2013), to cite just a few examples.

The creation of longitudinal imaging datasets has necessitated the development of novel methods for their analysis. The most prominent methods rely on either tissue classification optionally followed by further processing of individual brain structures (e.g. deformable models for extracting the cortex) (Fischl and Dale, 2000; Kim et al., 2005), or a combination of linear and non-linear image registration approaches wherein development across time is encapsulated in the transform required to deform one image to another (Avants and Gee, 2004; Collins et al., 1995; Studholme et al., 1996). This latter class of registration-based methods is flexible across different imaging contrast mechanisms and especially powerful in analyzing data from animal models (Lerch et al., 2011; Nieman et al., 2011).

Here we describe methods for the analysis of densely-sampled longitudinal image time series of brain development. These methods were created to analyze the data described in the accompanying paper (Szulc et al., submitted), yet are applicable to any dense developmental dataset. Briefly, the data we used consisted of 12 mice imaged with manganese (Mn)-enhanced MRI (MEMRI) every second day from postnatal day (P)1 onwards. The mice were divided into two cohorts, with six mice (the *odd day cohort*) imaged at P1, P3, P5, P7, P9, and P11 and six different mice (the *even day cohort*) imaged at P2, P4, P6, P8, and P10. The MEMRI scans covered the entire brain with exquisite contrast at 100 μm isotropic resolution. The early postnatal time period covered in this dataset, which corresponds roughly to late gestation in human brain development (Clancy et al., 2001), features significant overall brain growth together with region-specific alterations in local brain shape. The cerebellum, for example, acquires its complex folding (foliation) pattern during this period (Sillitoe and Joyner, 2007).

The analysis of the neonatal mouse data described above, - and generally of any dense longitudinal early brain development dataset – holds both challenges and opportunities:

- The brain changes to such an extent between the first and last time point that a direct mapping between them is not possible.
- Developmental patterns vary considerably across the brain; capturing that regional variability is in and of itself interesting.

- Patterns of brain development need to be captured across all mice and how individual mice differ from the group average must be determined.

Our implementation for addressing each of these builds on significant prior work in the literature. In particular, we rely on extensive efforts in designing deformable registration algorithms and their extensions to time-series data (Aljabar et al., 2008; Gogtay et al., 2008; Gogtay et al., 2006; Sadeghi et al., 2010; Thompson et al., 2000). We extend that work by taking it to denser longitudinal data, which both poses additional challenges and brings opportunities for analyses and inferences not hitherto accessible from human brain imaging.

Our approach to analyzing this dense longitudinal brain development data is divided into two components: (1) image processing and registration of the input data and (2) statistical analysis of resulting transforms. Each will be described in detail below.

A.3 METHODS

Animals and longitudinal MEMRI data sets

All mice used in this study were maintained under protocols approved by the Institutional Animal Care and Use Committee of New York University School of Medicine. The mouse handling and image acquisition methods are described in detail in the companion paper (Szulc et al., submitted).

Registration

Our proposed approach to analyzing dense developmental data relies on image registration at its core, where differences between two images (I and J) are captured by the transformation that maps one onto the other. Multiple algorithms have been proposed to optimally perform this alignment (Klein et al., 2009); we adopt the Lagrangian diffeomorphic registration technique implemented in the ANTs toolkit (Avants et al., 2006; Avants et al., 2008). This registration method uses symmetric normalization (SyN) to create a mapping between images I and J. Briefly, SyN weights the contributions of I and J equally in finding the transformation between them. This transformation is invertible and, if both I and J are deformed to the midpoint of the transform, they will be identical. In a study of 14 non-linear image registration methods (Klein et al., 2009), SyN consistently outperformed other optimization strategies.

The alignment of one image to another must be placed within the larger context of analyzing the full developmental dataset consisting of multiple subjects with repeated image acquisitions each. There are several registration strategies that can be pursued

towards the goal of providing insights into both individual and population level trends. A successful strategy needs to (a) map all m images for any individual subject to determine evolution in local brain shape over time, and (b) map all n subjects into a common space so that differences within the population at any one point in time as well as in evolution across time can be captured. For the type of data employed in this study, achieving (b) can be a challenge, as not every subject in the data set can be aligned to every other subject (e.g. it is not possible to accurately register a P1 scan to a P11 scan, even for the same mouse). Below, we describe two approaches for overcoming this challenge and discuss their advantages.

The first registration strategy employed, depicted graphically in Figure A-1 (see Supplemental Video A-1 for the complete set of 3D brain images depicted in Figure A-1) is the *registration chain*. In this strategy, each scan is registered to the next scan in the series for the same mouse. For example, consider Mouse 3 in the odd day cohort. First, its P1 scan is aligned to its P3 scan. Next, the P3 scan is aligned to the P5 scan, followed by P5 to P7 and so on. Similarly for a given mouse in the even day cohort, the P2 scan is aligned to P4, P4 to P6 and so forth. For each pair of scans, registration proceeds first linearly through a 12-parameter alignment, allowing uniform scales, shears, translations and rotations. The second part of the registration process is non-linear. Using the SyN optimization method described above, the scans are brought into correspondence, and the full transform from source to target (including both linear and non-linear components) is saved. In order to find the transformation between two non-adjacent time points for the same mouse, the appropriate transforms can be concatenated (e.g. to find the P1 to P5

transformation for a single mouse, one would concatenate the P1 to P3 and P3 to P5 transforms). Thus, although we cannot directly register two scans that are far apart in the scan series, through transform concatenation we can create a chain of transforms between all time points for each mouse in the study.

Once this has been achieved, a common space is still required to meaningfully compare differences at any point in time. This is done by registering all of the P10 and P11 scans in an iterative, group-wise registration procedure that is described in more detail in (Lerch et al., 2011). Briefly, group-wise registration proceeds as follows: after rigidly aligning all scans into the same coordinate space, each scan is aligned with all other scans in the data set via uniform scales, shears, rotations and translations. From these alignments, the best possible linear average (atlas) of all subjects is created. This atlas provides the starting point for an iterative non-linear registration process; all scans are aligned towards this atlas, resampled with the resulting transforms, and averaged to create a newer, more accurate atlas. This process is repeated multiple times until convergence is achieved. For this particular study, the result is an average of all the P10+P11 brains (henceforth referred to as the P10.5 average), and a transform from this average back to each P10 and P11 brain in the study. By concatenating these transforms with the transforms created by the *registration chain*, we have a way to go from this common space to every scan at every time point.

The second registration strategy relies completely on the *group-wise registration* strategy described above and takes advantage of the balanced study design of the data

described in the companion paper. Overlapping group-wise registrations are performed on all scans from adjacent days within the *odd day cohort* and *even day cohort* as illustrated in Figure A-2. For example, a group-wise average is created from all P3 and P5 scans, and a separate group-wise average from all P5 and P7 scans. The two cohorts are, as in the *registration chain*, brought into a common space by a group-wise registration of the P10 and P11 scans.

The ability to analyze each subject's time series is maintained through overlapping adjacent scans. Each scan can be mapped to any other scan in the series by using appropriate forward and inverse transformations as illustrated in Figure A-3. For example, the following transform concatenations would bring the P7 scan from mouse 2 into the common space (P10.5) created by the group-wise registration of all P10 and P11 scans:

$$T_{P7\rightarrow P10.5} = T_{P7\rightarrow \mathrm{Avg}_{P7+P9}} + T_{\mathrm{Avg}_{P7+P9}\rightarrow P9} + T_{P9\rightarrow \mathrm{Avg}_{P9+P11}} + T_{\mathrm{Avg}_{P9+P11}\rightarrow P11} + T_{P11\rightarrow P10.5} \quad (1)$$

In Equation 1 averages are denoted by Avg_{PI+PJ}. So, the average from the P9+P11 groupwise registration is denoted as Avg_{P9+P11}. Although we explicitly constructed $T_{P7\rightarrow P10.5}$, for much of our analysis we require $T_{P10.5\rightarrow 7}$. This can be attained simply by inverting the transform constructed in Equation 1.

For both the registration chain and group-wise registration strategies, linear registrations were performed using mni_autoreg tools (Collins et al., 1995) and non-linear registrations utilized the ANTs toolkit (Avants et al., 2008), described above.

The registration algorithms we use require sufficient homology between scans to work well. To ascertain which post-natal stages can be accurately registered to which other post-natal stages, and thus inform future study design, we aligned average images from each day to every other day. Accuracy was then computed using a manually segmented P10+P11 average image (described in more detail in Szulc 2014) and the Kappa statistic:

$$k = \frac{2a}{2a + b + c} \tag{2}$$

where a is common to the automatic segmentation and the ground truth, $b + c$ is the sum of the voxels uniquely identified by the segmentation and the ground truth respectively. k takes on values between 0 and 1, with 1 indicating perfect agreement and 0 indicating chance agreement. See (Chakravarty et al., 2013; Collins and Pruessner, 2010; Klein et al., 2009) for further discussion of these measurements of registration accuracy.

Statistical analysis

At the end of the registration process using either of the proposed strategies, a transform exists from each scan to the P10.5 average. As described above, many of these transforms are created by concatenating transforms from multiple individual registrations. Each transform is then inverted so that it originates in the P10.5 average and terminates at

a specific time-point and mouse. From these transforms (encoded by displacement fields), we can compute the Jacobian determinant (in P10.5 space) for each transform. This gives a measure of growth or shrinkage at every voxel for every image in the dataset. We can then analyze developmental patterns either at every voxel or by integrating the Jacobians across all voxels in a particular brain structure to compute the structure volume. The structures we used were obtained by manually segmenting the P10+P11 average and are described in more detail in (Szulc et al., submitted).

Developmental patterns were analyzed using linear mixed effects models via the *lme4* package in R (Bates et al., 2012; Team, 2013). Because each mouse was scanned repeatedly, measurements taken on the same mouse are not independent, and a normal linear model (which requires independence of all measurements) may not be used. Linear mixed effects models allow us to appropriately account for any structure in the data that is due to this repeated scanning so that these random effects do not bias our results. Although repeated measures ANOVA (rANOVA) is sometimes used in studies such as this one, the fact that we have an uneven number of scans for each mouse (6 in the odd day cohort and 5 in the even day cohort) makes it difficult for us to satisfy the assumptions necessary for this method (Pinheiro and Bates, 2000). Here, we introduce the standard formulation for a linear mixed effects model. We then take this general mathematical formalism and explain how we adapted it for this particular study.

The general formula for a linear mixed effects model is as follows (Laird and Ware, 1982):

$$y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1} z_{1ij} + \cdots + b_{iq} z_{qij} + \epsilon_{ij} \quad (3)$$

In Equation 3, y_{ij} is a response variable for the jth measurement in the ith subject. For the purposes of this study, y_{ij} is a measurement in mouse i on the jth day. These measurements correspond to either the value of the Jacobian determinant at a particular voxel, the volume of a particular brain structure, or another imaging measure for comparison (such as image intensity). For example, y_{ij} could be the value of the total brain volume for Mouse 3 measured on day 5. $\beta_i, \cdots, \beta_p$ are fixed effects coefficients corresponding to fixed-effect regressors $x_{1ij}, \cdots, x_{pij}$. Here, these regressors are *cohort* (whether or not a mouse was scanned on even or odd days) and the scan day itself. b_{iq} and z_{qij} are the random effect coefficients and their regressors, respectively. In the present study, we consider the mouse ID to be a random effect; that is, aside from natural variability within the population, the differences in growth trajectories between mice are effectively random, and we need to account for this in our models. ϵ_{ij} is the error term, and is assumed to be independent and normally distributed. We can therefore re-write the above formulation more specifically for our model as:

$$y_{ij} = \beta_1 + \beta_2 c_i + \sum_{k=3}^{l} \beta_k f_k(d_{ij}) + b_{1i} + b_{2i} d_{ij} + \epsilon_{ij} \quad (4)$$

In the above equation, c_i is the cohort (odd or even), which depends only on the mouse i and not the scan day, j. The scan days themselves are denoted d_{ij}. To model growth as a function of time, we used spline interpolation (the *splines* package in R) instead of a simple linear or polynomial fit. This is encapsulated in $\sum_{k=3}^{l} b_k f_k(d_{ij})$ term. Each $f(d_{ij})$ is natural spline basis function and b_k is the corresponding coefficient. For example, if we choose to use a natural spline with three degrees of freedom, this sum would expand to:

$$b_3 f_3(d_{ij}) + b_4 f_4(d_{ij}) + b_5 f_5(d_{ij}) \quad (5)$$

b_{1i} (where $z_{ij} = 1$) is a random effect that corresponds to a separate intercept for each mouse, and b_{2i} (where $z_{ij} = d_{ij}$) is a random effect allows each the slope of the fit to vary for each mouse as well. With these two terms, we can model subtle differences in growth for each mouse, while necessarily separating this random effect from the fixed effects of cohort and day.

Equation 4 is quite general, and relies on several assumptions. Specifically, we assume that the effect of cohort is significant enough to be included in the model, and we also assume that both a random slope and intercept per subject are needed. In order to validate the appropriateness of these assumptions, as well as make inferences on brain developmental patterns, we explicitly tested them using log-likelihood comparisons of nested models.

As an example of a log-likelihood comparison, consider the Jacobian determinant of a single voxel, measured in all subjects on all days. Alternatively, consider the volume of the hippocampus, measured in all subjects on all scan days. For both of these measurements, we'd like to find the best possible model for growth. To begin, we can compare a linear fit (first order natural spline) to a quadratic fit (second order natural spline). These fits are modeled as:

$$\begin{aligned} y_{ij,L} &= \beta_1 + \beta_2 c_i + \beta_3 f_3(d_{ij}) + b_{1i} + b_{2i} d_{ij} + \epsilon_{ij} \\ y_{ij,Q} &= \beta_1 + \beta_2 c_i + \beta_3 f_3(d_{ij}) + \beta_4 f_4(d_{ij}) + b_{1i} + b_{2i} d_{ij} + \epsilon_{ij} \end{aligned} \quad (6)$$

In the above equations, L and Q stand for the linear and quadratic fits respectively. Note that the only difference between them is the presence of the $b_4 f_4(d_{ij})$ term in the quadratic equation. How much does this additional term improve the model of growth? We can answer this question using the standard log-likelihood test

$$-2\log\frac{L(\hat{q}_{\mathbf{L}};\mathbf{y})}{L(\hat{q}_{\mathbf{Q}};\mathbf{y})} \quad (7)$$

In Equation 7, $L(\hat{q}_{\mathbf{L}};\mathbf{y})$ is the maximum likelihood of the linear model, subject to a set of optimized parameters $\hat{q}_{\mathbf{L}}$. The analogous definition applies for $L(\hat{q}_{\mathbf{Q}};\mathbf{y})$. Here, the vector $\mathbf{y}$ indicates that we are considering the joint probability distribution for all measurements y_{ij} in the construction of L. For example, $\mathbf{y}$ could represent the volume of hippocampus, measured for each subject i across multiple days (j). Note that

Equation 7 follows the c^2 distribution with 1 degree of freedom, which corresponds to the difference in the number of parameters between the two models. From this distribution and the value of the ratio given in Equation 7, we can find the appropriate p-value for comparing these models. $p < 0.05$ generally indicates that the term in the denominator (in this case, the quadratic model) is a better fit. In every voxel in the brain and for each brain structure we used the log-likelihood test to compare linear and quadratic growth curves, as described here, as well comparing quadratic and cubic growth patterns (second vs third order spline fits) and cubic vs fourth order growth patterns (third vs fourth order).

In addition to testing growth patterns, we also used the log-likelihood test to explicitly test whether or not there is a cohort effect. That is, does it matter if mice were first scanned on P1 or P2. In this instance, the models we compare are:

$$\begin{aligned} y_{ij,NC} &= \beta_1 + \beta_3 f_3(d_{ij}) + \beta_4 f_4(d_{ij}) + b_{1i} + b_{2i} d_{ij} + \epsilon_{ij} \\ y_{ij,C} &= \beta_1 + \beta_2 c_i + \beta_3 f_3(d_{ij}) + \beta_4 f_4(d_{ij}) + b_{1i} + b_{2i} d_{ij} + \epsilon_{ij} \end{aligned} \quad (8)$$

The key difference in these equations is the presence of the $b_2 c_i$ term in the second, but not the first. By using the log-likelihood ratio defined in Equation 7 (using the appropriate substitutions for $y_{ij,NC}$ and $y_{ij,C}$) we can assess whether or not cohort does have an effect on voxel and brain structure volumes.

Similarly, we can nest random effects to assess whether each brain region or voxel requires a random intercept per subject or a random intercept plus a random slope

per subject. To compare these models, one would simply omit $b_{2i}d_{ij}$ from Equation 4, which is the term that corresponds to including a random slope. Comparing the resulting equation to Equation 4 using the log-likelihood test would validate whether or not the random slope is needed.

In addition, we can explore the developmental curve while staying with regression splines by using general additive models (GAM) (Wood, 2004, 2011). Here the standard linear model formalism

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (9)$$

is modified as in equation (10) to allow explicit incorporation of smooth functions, where the degree of smoothness is estimated as part of the fitting.

$$g(y) = b_0 + f(x_1) + f(x_2) + \ldots + f(x_p) \quad (10)$$

This provides the benefit of not having to determine the spline degrees of freedom beforehand, which could be especially beneficial given multiple voxels/brain structures. Instead, the effective degrees of freedom are estimated based on a penalized smoothing spline with degrees of freedom selected using generalized cross-validation (Wood, 2000).

Alternatively, where we can apply prior knowledge about the type of brain growth expected, we can also describe our data using non-linear models. More specifically, we used two non-linear, sigmoidal functions to model growth, the Gompertz function

$$y(x) = b_0 exp(- b_1 exp(b_2^x)) \quad (11)$$

and the four-parameter logistic model

$$y(x) = f_1 \frac{f_2 - f_1}{1 + exp\left[(f_3 - x)/f_4\right]} \tag{12}$$

Although both of these models are more complex than the linear model described in Equation 4, we gain the benefit of more interpretable terms in the models. For instance, in the Gompertz function, b_2 is the rate of growth.

We note that the p-values resulting from the log-likelihood test are slightly anti-conservative (Faraway, 2006; Stram and Lee, 1994), and can be replaced with an accurate p-value using the parametric bootstrap. Here the null hypothesis of the two models being the same is tested by generating the base model, H_0, and using its parameters to simulate new data n (we picked 1000) times. For each such simulation the log-likelihood value of the real data versus the simulated data is computed, thus generating a new distribution of log-likelihood values. The actual log-likelihood value from comparing the alternate model, H_1, to H_0 is then compared against this distribution to ascertain whether it is sufficiently extreme to reject the null hypothesis and thus declare that H_1 provides the better fit.

To account for the multiple comparisons created by testing at every voxel or every segmented brain structure we use the False Discovery Rate (Genovese et al., 2002).

Software

The core algorithms described in these two papers are open source and freely available; an overview can be found at http://wiki.mouseimaging.ca. The registration pipelines are implemented in pydpiper (Friedel et al., 2014) (source code: https://github.com/mfriedel/pydpiper), and mixed effects models in R are exposed through RMINC (https://github.com/mcvaneede/RMINC).

A.4 RESULTS

A.4.1 Comparison of registration strategies

We introduced two different registration strategies for handling data sets of this type: the *registration chain* and the *overlapping group-wise registration*. A natural question that arises is whether or not these registration strategies give the same quantitative results. To assess this, we calculated brain structure volumes for the following brain structures (also depicted in Figure A-7 in the companion paper): whole brain, cortex, cerebellum, hippocampus, olfactory bulbs, caudate putamen, and the superior and inferior colliculi. These volumes were calculated twice at every time point for every mouse: once using the Jacobian determinants from the *registration chain* and once using the Jacobian determinants from the *overlapping group-wise registration*. For each volume, growth curves were fit using a third order natural spline and are shown in Figure A-4. For all brain structure volumes considered, the growth curves calculated by each method overlapped to within the calculated 95% confidence intervals. Moreover, for the larger structures (whole brain, cortex, and cerebellum) the growth curves were virtually identical.

A.4.2 Modeling growth

We assessed three different strategies for understanding growth patterns in the brain in this early post-natal dataset. Results are shown in Figures A-5 and A-6 as well as Table 3-1. Brain growth between post-natal day 1-11 for the most part can be effectively modeled by a four parameter logistic regression which, unlike the Gompertz model, gives

realistic parameter estimates for the starting and ending volumes, which is particularly evident for the cerebellar compartments. Remaining brain structures, such the colliculi, follow a more clearly linear pattern between P1 and P11. Convergence of the logistic model in those cases is a problem, particularly evident when attempting to model alterations at every voxel (not shown), where 31% failed to converge. General additive models indicate that most brain compartments are best fit with approximately 4 degrees of freedom, though log-likelihood tests of different spline degrees of freedom show that, except for the colliculi which follow a linear pattern, a more conservative third order spline is adequate. Comparing growth patterns shows that the cerebellum clearly develops differently from the rest of the brain, but there is only weak evidence of differences within cerebellar compartments or within cerebral compartments. The cerebellum matures later, with slower initial growth followed by more rapid acceleration, as evidenced by the uniformly higher f_3 parameter and generally lower f_4 parameters from the logistic regression.

A.4.3 Effect of imaging cohort

There were two separate cohorts of six mice each; the *odd day cohort* was imaged on P1, P3, P5, P7, P9, and P11, whereas the *even day cohort* was imaged on P2, P4, P6, P8, and P10. We used log-likelihood tests to assess whether the two cohorts differed in regional volumes and/or developmental patterns. As seen in Figure A-7, there was a trend towards overall increase in brain volume in the *even day cohort* ($q<0.06$), with significant cohort effects identified in the Cerebellum as a whole ($q<0.02$) and vermis lobules VII

(q<0.03), VIII(q<0.02), and IX (q<0.02). Comparing an additive model (cohort + day) to an interaction model (cohort * day) revealed no significant effect of cohort on developmental patterns, though the hippocampus and vermis lobules VIII, IX, and X reached trend levels (q<0.06).

A.4.4 Registration accuracy

To test how accurately brain images can be registered across days an average image from each day was registered to every other day, and the registration accuracy computed by comparing the best available labels for each day to those from each registration. As can be seen in Figure A-8, accurate registrations can be obtained to within two days of the early scans (P1-P5, and increasing in range for the later days (P6-P11).

A.5 DISCUSSION

In this paper we described the data analysis methods that allow for inferences on densely sampled longitudinal brain development data. The need for novel methodology originated from the need to analyze *in vivo* early postnatal mouse brain MEMRI images described in the companion article (Szulc et al., submitted). Differentiating this dataset from previous longitudinal brain imaging studies described in the literature is the fact that it involves multiple datasets per animal over a time-period where the brain undergoes large changes in shape.

The analysis methods we have implemented rely on image registration to bring homologous points into correspondence. This is inherently a 4D (time and space) problem, similar to registration challenges involving DCE-MRI (Hodneland et al., 2014). In either case, we deconstruct the problem into a series of 3D optimization problems either between adjacent scans in the time-series or between an individual scan and an evolving average anatomy of all scans at that age. Continuity in the time dimension is never explicitly optimized, but instead assured through a combination of concatenating appropriate transforms and adding a smoothing function - a spline - on time in the final statistical analyses.

Alternate solutions proposed in the literature are either similar to our use of 3D registration between adjacent time points (Gogtay et al., 2008; Gogtay et al., 2006; Thompson et al., 2000; Wang et al., 2013), or assume the existence of homology between

all scans in the time-series. Examples of the latter include the FreeSurfer longitudinal module, wherein intermediate representations are created to avoid bias towards any one scan being treated as baseline (Reuter and Fischl, 2011), as well as the aBEAT toolkit (Dai et al., 2013), which requires the simultaneous registration of an atlas to all longitudinal images (thus assuming homology). While many aspects of these techniques, in particular their use of deformable models and tissue segmentation, could be incorporated into the analysis of the mouse postnatal MEMRI dataset described here, they could not be used without significant modifications akin to what we implemented to account for the differences between the beginnings and ends of the time-series. Similarly, it needs to be remembered that the mouse brain is smooth, lacking cortical folds, which makes registration both within and across animals more powerful than the equivalent would be in human datasets given the highly idiosyncratic cortical folding patterns (Mangin et al., 2010).

Statistical analyses were carried out using massively univariate (i.e. separate test at every voxel or for every brain structure) linear mixed effects models and corrected for multiple comparisons using the False Discovery Rate (Genovese et al., 2002). To make inferences on the data we chose to compare nested models using log-likelihood tests. This allows us to frame any inference on the basis of whether the simplest fit of the data, that voxel/volume size is modeled by a linear effect of time with a separate intercept per subject, can be improved by additional terms. The advantage of this approach is that it can be used for both fixed (number of spline degrees of freedom, cohort effects) and random effects (whether a separate slope is needed per subject). It also circumvents the

challenges surrounding how to determine degrees of freedom for linear mixed effects models. The use of the chi-squared distribution for assessing the outcome of the log-likelihood test is, however, somewhat anti-conservative. To assure that we weren't biasing our results, we used the parametric bootstrap (Faraway, 2006; Stram and Lee, 1994) on a subset of our results. While much more computationally expensive than log-likelihood followed by a chi-squared test, the parametric bootstrap method is free from this bias. In the end, we found that simple log-likelihood testing and the parametric bootstrap produced the same results, and we thus kept the computationally simpler log-likelihood tests.

A core aspect of understanding brain development data from longitudinal acquisitions is correctly modeling time. We assayed three approaches. The first, and in the end most powerful, approach was to model time as a natural spline within the linear mixed effects models. We used log-likelihood tests to assess the spline degrees of freedom required, and found that it varies in space, with the inferior colliculus, for example, showing linear growth whereas the cerebellum followed a more cubic pattern which differed significantly from the rest of the brain. The upside of using splines is their simplicity; the downside is the need for selecting degrees of freedom as well as the relative difficulty in interpreting individual spline components in the model. To overcome these limitations we attempted to both use general additive models as well as non-linear mixed effects models. General additive models also use splines to model time, yet determine the degrees of freedom needed from the data (Wood, 2000). On our data, however, they appeared to consistently overestimate the required degrees of freedom.

Additionally, it is difficult to model interactions when using general additive models, somewhat limiting their applicability for a subset of the inferences we sought to make. Of the two non-linear models we used, the Gompertz model and the four parameter logistic model, the latter provided more believable parameter estimates than the former. The advantage of these non-linear models is the intuitive meaning that can be assigned to each of the model parameters; the major downside we found was the lack of convergence in all brain structures. This was especially the case when modeling voxel-wise data, where lack of convergence became a significant issue across much of the brain. In the end, using mixed effects models with a spline term for modeling time remains the most recommended method, though the judicious use of non-linear models should be encouraged.

Future studies of early post-natal mouse brain development face several choices in study design. The first is how densely to sample in time. Estimating the accuracy with which each post-natal stage can be registered to every other indicate that, between P1 and P5, at most one day can be skipped. Later days, however, feature neuroanatomy with more homology, thereby allowing for increased scan spacing while still enabling accurate registrations between those stages. Additional concerns relate to the numbers of mice necessary. A recent paper of ours (Lerch et al., 2012) explored the tradeoff between number of mice and numbers of scans per mouse in great detail, and so we will not belabor that point here. Lastly, the issue of Mn toxicity, effects of anesthesia, and stress related to maternal separation need to be taken into account as discussed in (Szulc et al., submitted) and also identified by the differences in anatomy between the even and odd

day cohorts. If that concern is crucial for phenotyping a novel mouse-line then complementing in-vivo MEMRI with ex-vivo diffusion weighted images (Portmann et al., 2014; Zhang et al., 2006) should be encouraged.

Of the two registration approaches described, the registration chain and overlapping group-wise registration, the chain is the more general as it does not have the same requirements of equally spaced scans for all subjects. If applied to human data it also has the advantage of not having the same requirement for homology across all subjects. The downside of the registration chain is that it loses the advantages inherent in group-wise registrations, including greater robustness to noise and less bias towards any one baseline scan or atlas (Fonov et al., 2011; Guimond et al., 2000; Reuter and Fischl, 2011). In our case we also found that the two strategies produced almost identical results, thus leaving the decision of which strategy to use to be based on the more practical considerations outlined above.

In summary, we have outlined a framework for the analysis of dense longitudinal brain imaging data, combining image registration, transform concatenation, and statistics on the resulting deformation fields to make inferences about regional brain growth. This approach is applicable to any similar dataset be it from animal models or human studies. Combined with the acquisition of early mouse brain developmental data described in (Szulc et al., submitted) we envision a plethora of interesting scientific questions on normal and abnormal brain development that can be answered using these methods.

A.6 ACKNOWLEDGEMENTS

This research was supported by NIH grant R01NS038461 (DHT) as well as by the Ontario Brain Institute's Province of Ontario Neurodevelopmental Disorders Network (JPL). We would like to thank Drs. John Sled and Mark Henkelman for many informative discussions regarding image registration and imaging statistics.

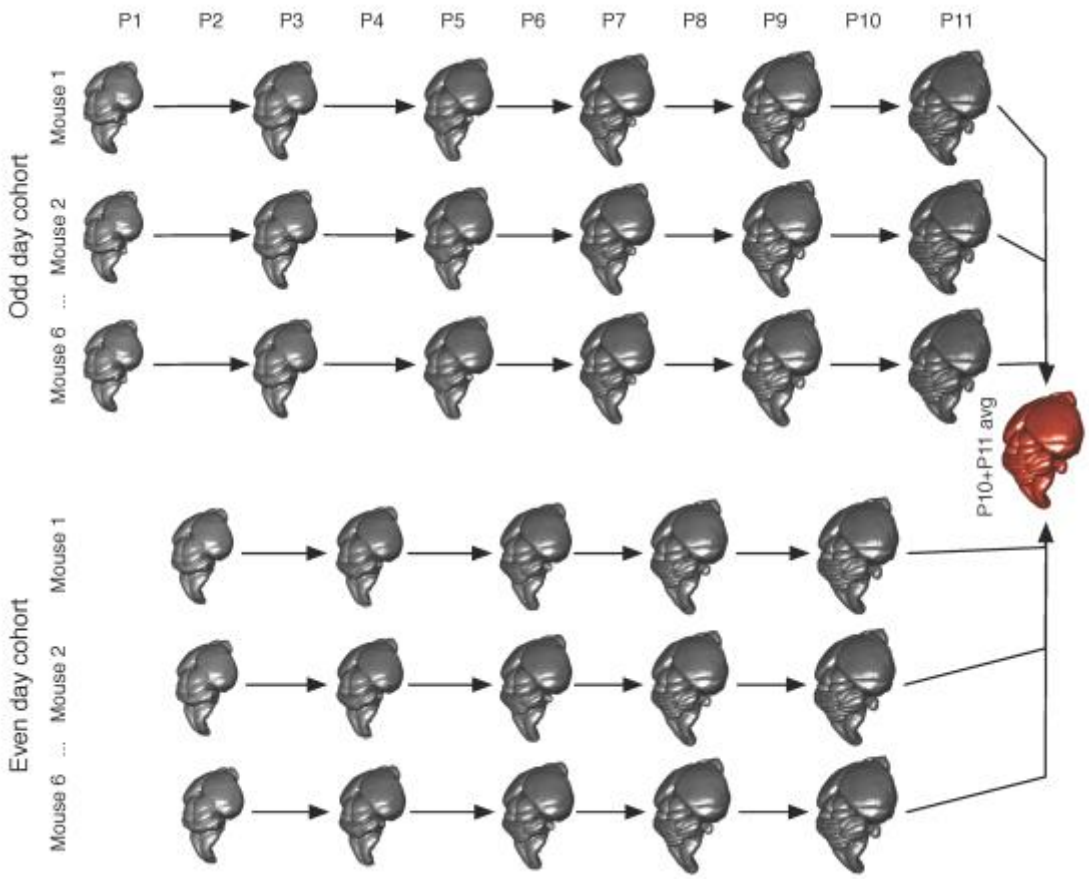


Figure A-1. Longitudinal data analysis through a registration chain. Here each scan from each individual mouse in the dataset is linearly and non-linearly registered to the next scan in the series for that mouse. For example, the P3 scan for Mouse 2 in the Odd day cohort is aligned to the P5 scan of Mouse 2 in the Odd day cohort. The final scans of all mice, acquired on days P10 and P11 for the Even day cohort and Odd day cohort respectively, are then aligned using group-wise registrations (shown in red.)

Figure A-2

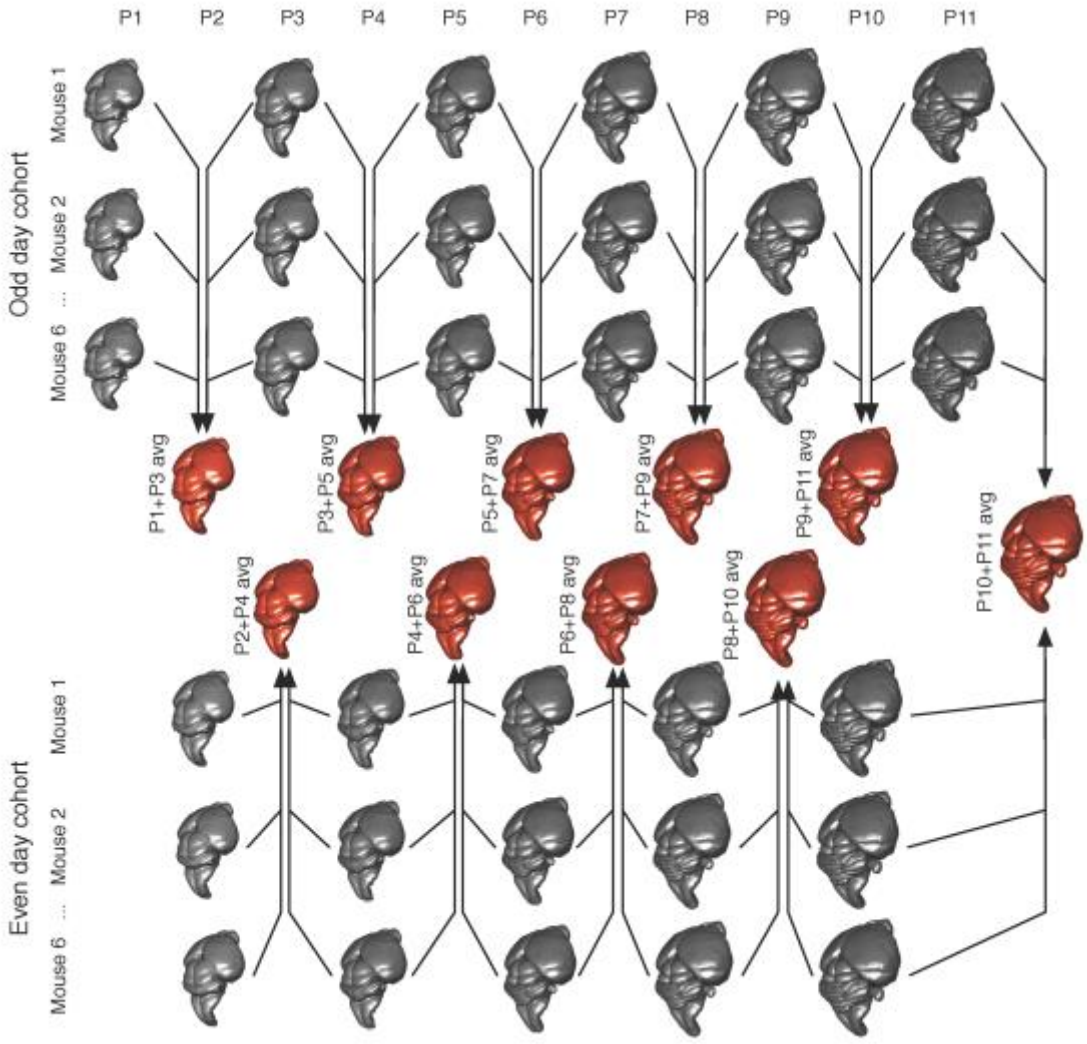

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
Odd day cohort
Mouse 1
Mouse 2
...
Mouse 6
P1+P3 avg
P3+P5 avg
P5+P7 avg
P7+P9 avg
P9+P11 avg
P10+P11 avg
P2+P4 avg
P4+P6 avg
P6+P8 avg
P8+P10 avg
Even day cohort
Mouse 1
Mouse 2
...
Mouse 6

Figure A-2. Longitudinal data analysis through overlapping group-wise registrations. Here we take advantage of the improved registration performance and numerical stability of groupwise registrations; adjacent days for each cohort are aligned together, so that every brain, with the exception of those at the beginning of the series, participates in two groupwise registrations. For example, the P5 scans are incorporated in both a registration of all P3 and P5 scans and in another registration of all P5 and P7 scans. The even and odd day cohorts are joined by registering the P10 and P11 scans together.

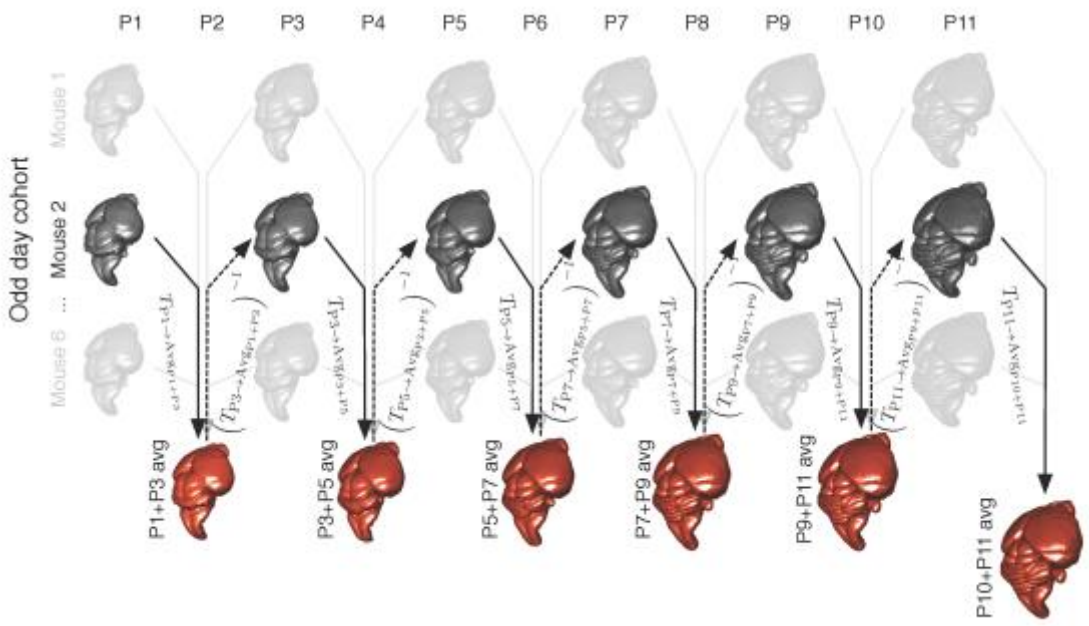


Figure A-3. Example concatenations necessary to take the P1 scan of one mouse to the consensus P10+P11 space. The overlapping groupwise registrations allows one to take any brain into the space of any other brain through appropriate transform concatenations and inversions. The figure above illustrates the process of taking the P1 scan of Mouse 2 in the Odd day cohort into the final space of all P10 and P11 scans. This is accomplished by taking the forward transform of every groupwise registration and concatenating with the inverse transform of the second timepoint in each groupwise registration.

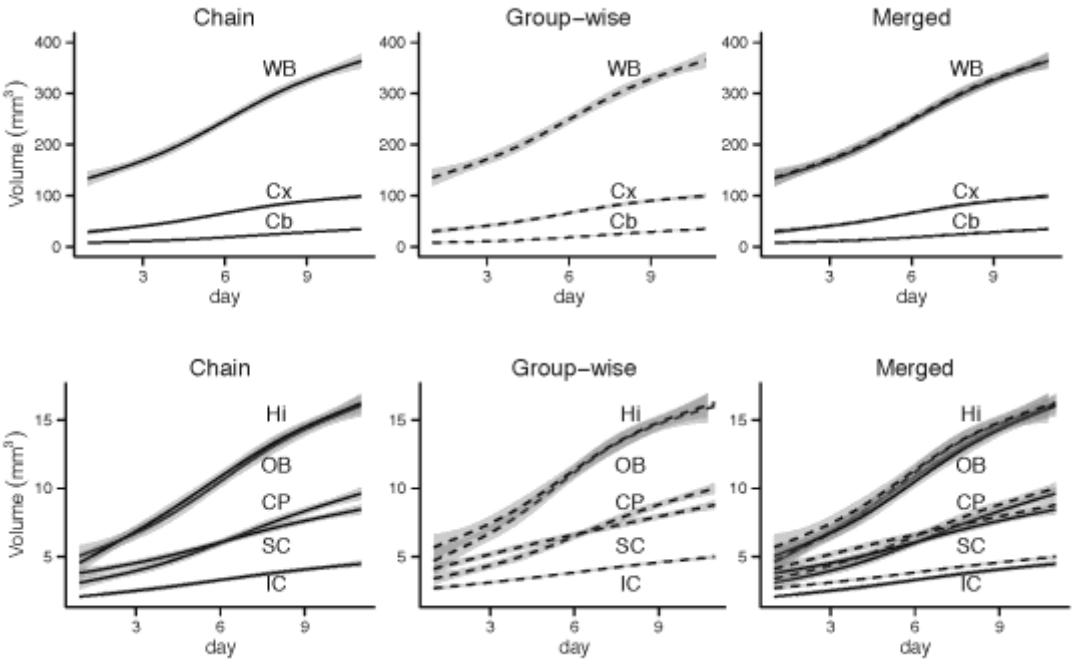


Figure A-4. Comparison of brain structure volumes: registration chain vs. overlapping groupwise registration methods. In this figure, we compare growth curves for various brain structure volumes. The top row shows growth curves for brain structure volumes for the whole brain (WB), cortex (Cx) and cerebellum (Cb). The bottom row depicts growth curves for the hippocampus (hi), olfactory bulbs (OB), caudate putamen (CP), superior colliculus (SC) and inferior colliculus (IC). From left to right, we show results from the registration chain, group-wise registration, and both methods superimposed. For many structures, the growth curves generated from the two different methods are quite similar. The grey bars around each curve represent 95% confidence intervals.

Figure A-5

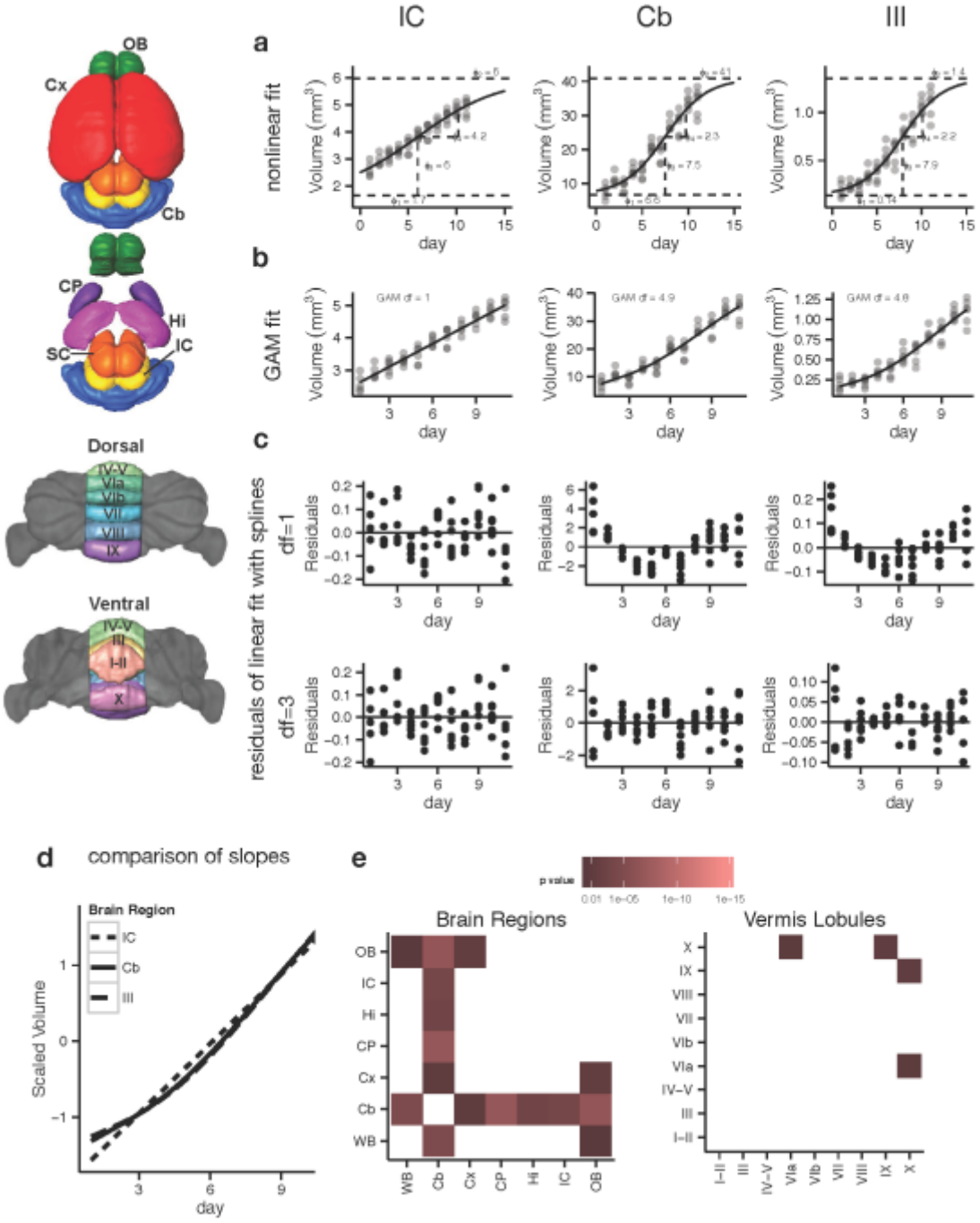

IC
Cb
III
a
nonlinear fit
Volume (mm³)
day
b
GAM fit
GAM df = 1
GAM df = 4.9
GAM df = 4.8
c
residuals of linear fit with splines
df=1
df=3
Residuals
day
OB
Cx
Cb
CP
Hi
IC
SC
Dorsal
IV-V
VIa
VIb
VII
VIII
IX
Ventral
IV-V
III
I-II
X
d
comparison of slopes
Brain Region
IC
Cb
III
Scaled Volume
day
e
p value
0.01
1e-05
1e-10
1e-15
Brain Regions
OB
IC
Hi
CP
Cx
Cb
WB
Vermis Lobules
X
IX
VIII
VII
VIb
VIa
IV–V
III
I–II

Figure A-5. Slope fitting. Different fits are illustrated for three regions, starting with the four-parameter logistic model, followed by the general additive model. The residuals of a linear with model with either one or three degrees of freedom on the spline fit further illustrate how the IC follows a relatively linear growth during the first 11 days of postnatal life, whereas the cerebellum and folia VIa need a higher order spline to model the "S" curve growth they undergo. To fully compare growth curves we performed pairwise comparisons of all segmented structures after scaling the values by subtracting the mean and dividing by the standard deviation separately for each structure. Pairwise comparisons were computed by fitting with and without a by structure interaction term, and the p-value of a log-likelihood test of those two models retained. A quick examination of the heat map of those p-value shows that the cerebellum clearly grows differently from the rest of the brain, but that the cerebellar subregions follow a relatively homogenous pattern.

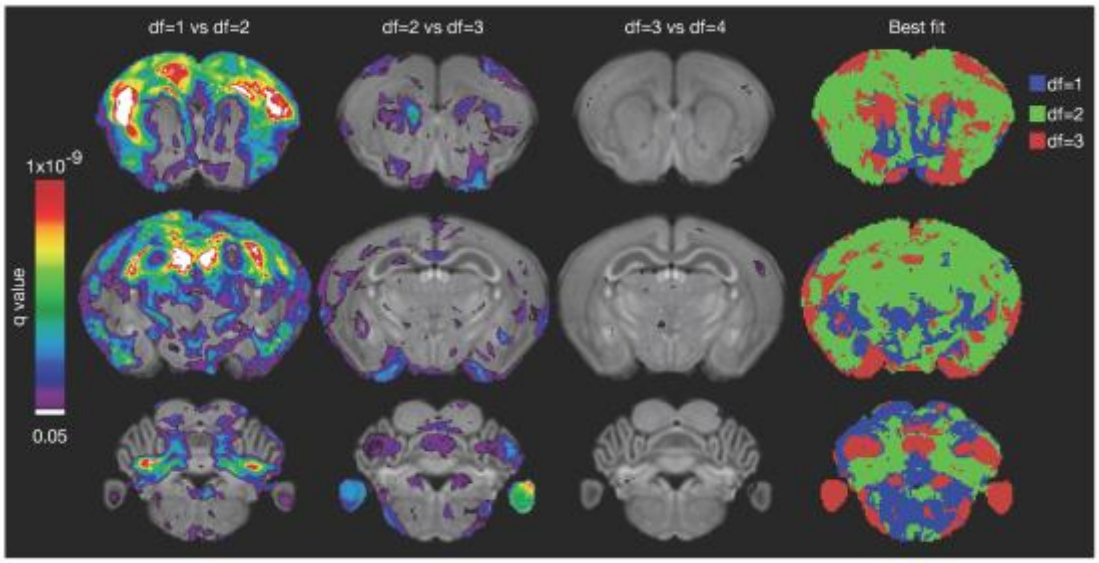


Figure A-6. Assessing the type of curve that best describes regional brain development. Brain development patterns were tested by fitting four different linear models at every voxel using natural splines with increasing degrees of freedom, and at each voxel fits compared using model comparison tests. The first column shows areas of the brain were there is significant evidence that a spline with two degrees of freedom provides a better fit than a spline with a single degree of freedom (i.e. a line), the next column compares two degrees of freedom to three degrees of freedom, and the third column 3 vs 4 degrees of freedom. Color bar is on a -log10 scale. The final column shows the resulting map of best fit degrees of freedom.

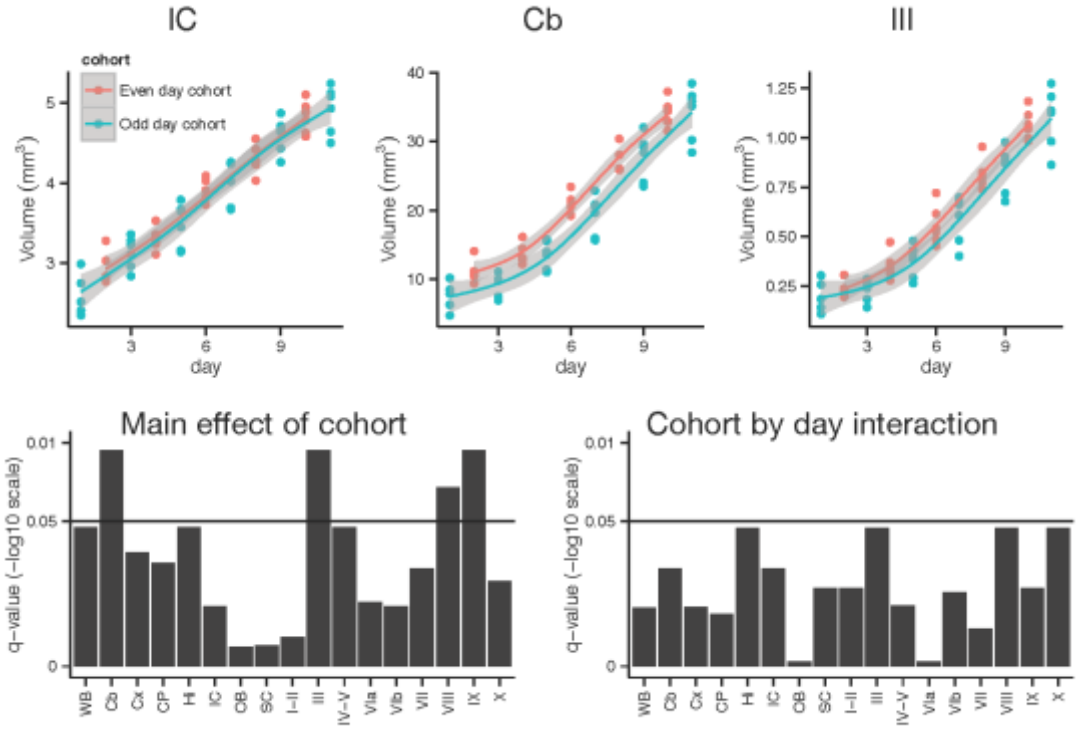


Figure A-7. Assessing the effect of imaging starting at P1 vs starting at P2. Starting imaging at P1 reduces global brain volume as compared to starting imaging at P2. These effects are exaggerated in the cerebellum compared to other structures like the inferior and superior colliculi.

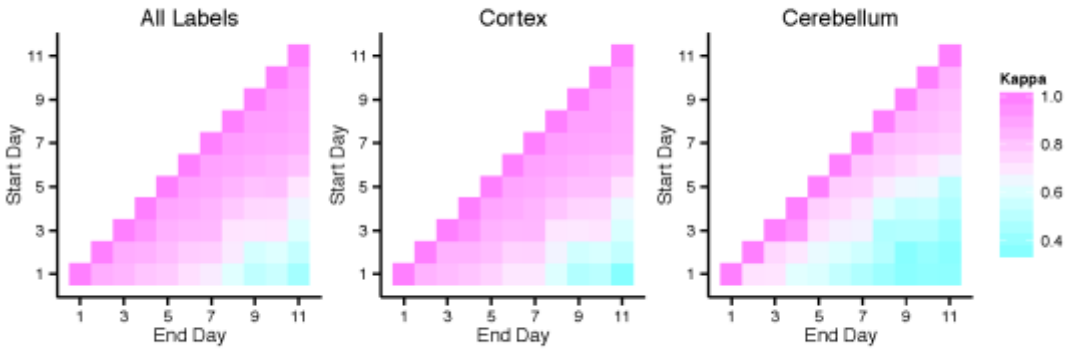
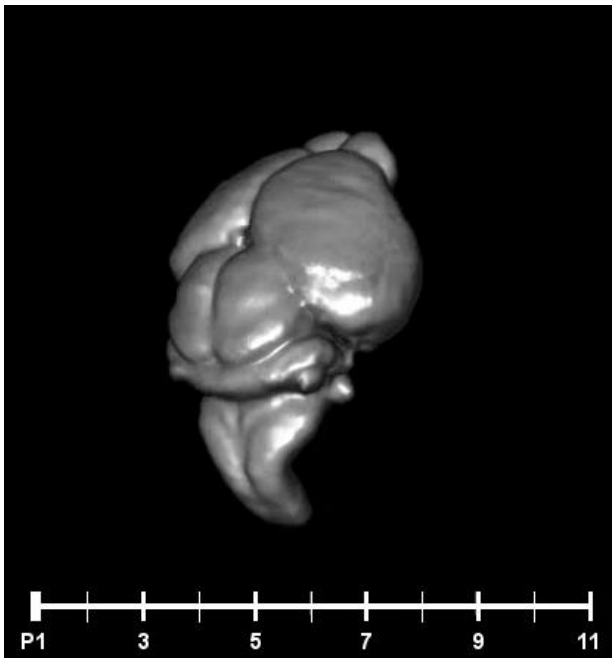


Figure A-8. Registration accuracy. The accuracy of registering images from different postnatal stages are illustrated for all 7 labels combined as well as the cerebral cortex and the cerebellum individually, with the colors indicating the Kappa statistics obtained. A P1 scan, for example, can be aligned with reasonable accuracy in the cerebellum to P2 and P3, but not to later stages, whereas a P7 scan can be aligned accurately to P8-P11. Cortical alignment across days has increased accuracy compared to cerebellar alignment.

structure	DF	1 vs 2	2 vs 3	3 vs 4	a_0	b_0	b_1	ϕ_1	ϕ_2	ϕ_3	ϕ_4
Brain	5.04	0.119	1e-05	0.391	655.94	1.81	0.90	96.89	408.93	6.14	2.68
Cb	4.91	1e-10	3e-05	0.319	136.81	3.18	0.92	6.56	40.81	7.49	2.25
Cx	4.26	0.971	3e-05	0.456	152.90	1.96	0.87	20.29	108.85	5.84	2.40
CP	3.82	0.005	2e-04	1.000	21.04	2.07	0.91	2.24	11.58	6.55	2.83
Hi	4.11	0.717	6e-05	0.336	25.34	2.03	0.86	1.65	19.40	5.67	2.96
Ic	1.00	0.770	0.052	0.113	9.75	1.40	0.94	1.66	5.97	5.98	4.22
Ob	7.34	0.683	2e-04	0.126	26.67	1.82	0.88	3.83	18.46	5.92	2.60
SC	1.00	0.517	0.550	0.349	13.13	1.29	0.90				
I-II	4.27	5e-10	0.033	0.145	15.19	4.28	0.97	0.18	1.03	8.36	3.09
III	4.83	2e-11	2e-04	0.187	4.47	3.66	0.91	0.14	1.35	7.85	2.24
IV-V	4.81	7e-11	5e-05	0.218	8.00	3.34	0.91	0.33	2.67	7.61	2.26
VIa	4.80	5e-06	7e-05	0.273	1.26	3.01	0.90	0.09	0.55	7.10	1.85
VIb	4.07	2e-07	0.009	0.046	2.66	3.55	0.90	0.06	1.10	7.94	2.56
VII	4.47	2e-09	0.008	0.018	3.44	3.43	0.90	0.06	1.50	8.16	2.90
VIII	3.99	1e-11	0.036	0.752	10.58	4.09	0.94	0.21	1.58	8.14	2.30
IX	4.79	2e-07	6e-06	1.000	7.58	3.03	0.92	0.54	2.44	7.06	1.67
X	4.87	<2e-16	0.049	0.438				0.18	1.09	8.66	2.39

Table A-1. Results of different curve fitting strategies. The first column shows the estimated degrees of freedom needed to fit a spline as determined by General Additive Models. To test whether higher order splines are indeed required, log likelihood tests of four different fits (with DF of 1 through 4) were carried out; the p-values are kept in the next three columns. The three parameter estimates from the Gompertz function and the four parameter estimates from the four-parameter logistic model follow, with entries left blank if the fitting did not converge.



Supplemental Video A-1. Brain development between P1 and P11. Brain development between P1 and P11 using the same viewing angle and rendering as in Figures 1 - 3.