Submitted to the Annals of Statistics

STATISTICAL INFERENCE IN CELL LINEAGE TREES

By D. G. Hicks[†], T. P. Speed[‡,§], M. Yassin[§,¶], and S. M. Russell[†,§,¶]

Swinburne University of Technology[†], Walter & Eliza Hall Institute of Medical Research[‡], University of Melbourne[§], and Peter MacCallum Cancer Centre[¶]

Cell differentiation is often associated with specific divisions and generations in a lineage tree. The presence of phenotypic noise, however, can make it difficult to observe such patterns. Using the group symmetry representation of a binary tree, it is shown how variation in a lineage can be compactly described by a set of natural variables each of which is labelled by the division at which a type of variation arises and the generation at which it is expressed. This harmonic analysis for a rooted tree provides a disciplined way to aggregate tree-structured data, improving the ability to identify differentiation patterns in noisy lineages. It also allows the proportion of variation of a phenotypic fate associated with each division to be estimated and compared to the proportion of variation expressed at each generation. The method has been applied to T-lymphocyte lineages tracked using time-lapse microscopy over several generations. For comparison, the analysis has been applied to *C. elegans*, a lineage with clear differentiation stages, and to a stationary branching process, which has none.

1. Introduction. In embryonic cell lineages, differentiation is often tightly synchronised to specific cell divisions and generations (Chisholm, 2001; Hadjantonakis and Arias, 2016). Low levels of variability in lineages such as for the roundworm *C. elegans* has allowed unambiguous identification of the developmental patterns (Sulston et al., 1983) without the need for statistical analysis. In contrast, the presence of substantial phenotypic noise in, for example, lymphocyte lineages (Hawkins et al., 2007) may be preventing underlying differentiation patterns from being observed. Whether noise has a functional role in multi-cellular development has been much debated

*This work was supported in part by Australian Research Council (ARC) grant FT140101104 to DGH, the National Health and Medical Research Council of Australia (NHMRC) Program Grant 1054618 to TPS, and NHMRC grants 620500 and APP1099140 and ARC grant FT0990405 to SMR. We thank Alan Rubin for suggesting we test our method on the *C. elegans* lineage.

MSC 2010 subject classifications: Primary 62H99; Secondary 62M99.

Keywords and phrases: Tree-structured data, Patterned covariance matrices, Symmetry invariance, Analysis of variance, Gaussian graphical models

(Balázsi, van Oudenaarden and Collins) with indications that the stability of certain hematopoietic subpopulations arises simply from the law of large numbers, not from programmed differentiation (Gerlach et al., 2013). A statistical method that can detect underlying differentiation patterns in noisy lineage trees would be an important contribution to this debate.

Early approaches to statistical inference in cell lineages (Cowan and Staudte, 1986; Huggins and Staudte, 1994) focused on stationary processes in microbial populations and were not designed to address differentiation patterns. More recently, models have been developed in the context of particular hypotheses, such as chaotic dynamics (Sandler et al., 2015) or state-switching processes (Hormoz et al., 2016) but do not address the general problem of characterising variation in a lineage. The theory of branching processes (Haccou et al., 2005), which has been applied to lymphocyte proliferation and death (Zilman, Ganusov and Perelson, 2010), does not address lineage inference. Phylogenetic inference (Felsenstein, 2003) involves reconstructing an unknown tree structure but is not designed to infer developmental patterns in a known tree structure.

Identifying signal from noisy measurements requires aggregating data, yet for tree-structured data this most elementary of operations is non-trivial. The source of the difficulty is that, with noisy data, daughters are statistically indistinguishable (that is, unindentifiable or exchangeable). This means that their subtrees are indistinguishable too, and so on, recursively through all descendants, leading to a particular pattern of indistinguishability. Now in statistical analysis, one generally aims to aggregate as much of the data as possible to maximise signal-to-noise, but not so much that meaningful associations are lost. Achieving this requires both respecting and exploiting the pattern of indistinguishability.

The optimal aggregation scheme can be found by examining the symmetry invariance of a rooted tree. The indistinguishability of daughters and their subtrees means it is possible to permute family members in ways that preserve the joint distribution over the tree, with the set of all possible permutations forming a group. Group representation theory (Diaconis, 1988; Stiefel and Fässler, 1992) can then be used to identify a linear transformation that not only aggregates tree-structured data optimally but also defines a natural set of variables that is free of the redundancies caused by indistinguishable variables.

This analysis is related to the conventional practice of blocking in nested groups. What is new here is that the observations are distributed at all levels of the nested groups (the tree structure), not just at the lowest level as would be necessary for a conventional analysis of variance (ANOVA) to apply.

This generalisation has profound consequences as the variance components are no longer just scalar eigenvalues, as in an ANOVA (Speed, 1987), but instead form orthogonal subsets of dependent variables where each subset is associated with variation from a particular division.

In this paper this statistical framework is developed and applied to differentiation patterns in various lineage trees, from the highly-ordered structure observed in *C. elegans* to the featureless character of a simulated branching process. The framework is able to characterise the continuum between these two extreme cases of differentiation where most systems of interest, including the T-lymphocytes discussed here, lie. The technique is analogous to the spectral analysis of a noisy time series where spectral lines can exist concurrently with a broadband background. In that case, the lines are interpreted as arising from ordered processes while the flat background is considered to be from unstructured noise. It is in the spectral domain that this distinction between signal and noise becomes clear.

The rest of the paper is organised as follows. Section 2 shows aspects of the 3 lineage types used in this paper. The framework of the model, and how family members are assigned to variables, is given in Section 3. The core of the paper, Section 4, examines ways to improve inference on trees by progressively increasing model constraints until real data can be analysed. Graphical models are used to visualise and interpret the dynamics of variation in a lineage in Section 5. The progression and expression of phenotypic fate is defined and illustrated in Section 6. A discussion about the interpretations and prospects for this analysis is given in Section 7.

2. Lineage Data.

Three types of lineage data are analysed:

T cells Unpublished lineage data on $CD8^+$ T cells from GFP:OT-1 transgenic mice. Naive cells, expressing a T cell receptor for SIINFEKL peptide from ovalbumin, interact with peptide-pulsed bone marrow-derived dendritic cells to activate clonal expansion (Oliaro et al., 2010). Cells and their descendants are tracked using time-lapse fluorescence microscopy and analysed using custom software (Shimoni et al., 2013). Although multiple phenotypic traits were recorded, in this paper the only trait analysed is the average area of a dividing cell over its lifetime. Note that only dividing cells were used in the analysis; cells that die or whose fate is unknown were counted as missing data. 19 replicate families were used.

Worm Published (Santella et al., 2016) embryonic lineage data from the RW10425 transgenic strain of *C. elegans*. In this strain the PHA-4 protein, a marker for pharyngeal and intestinal tissue, is tagged with

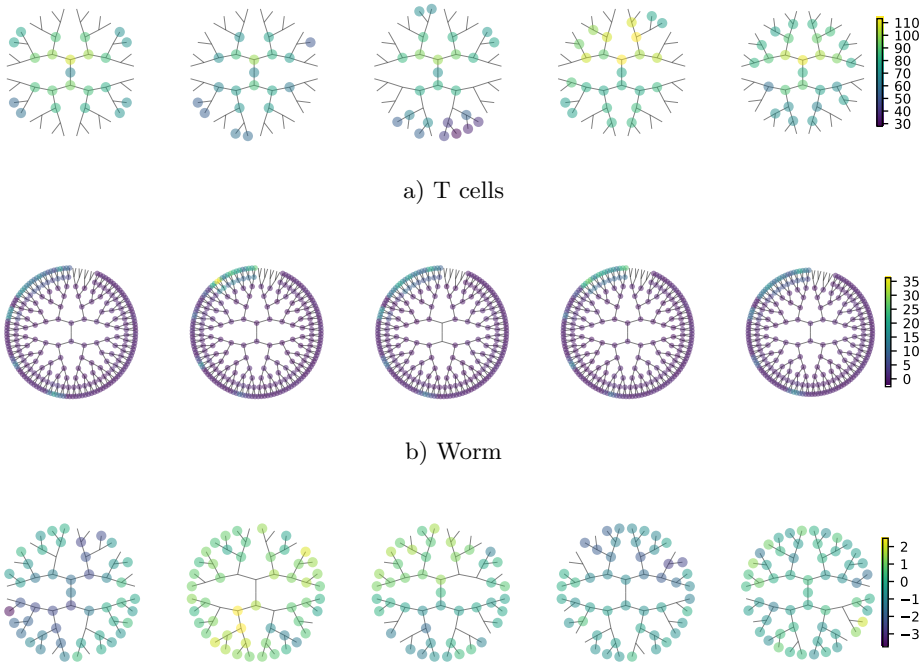
green fluorescent protein. Gut differentiation occurs early during embryogenesis, with PHA-4 expression beginning by generations 7 and 8. There are 10 replicate families.

Branching Process Simulated lineages from a stationary branching process. 20 replicate families are used, with a missing data fraction of 20% assumed. Here we define a branching process to be one with a particularly simple pattern of correlations between family members ς and ς': The mother-daughter correlation is h while the correlation between any two family members ς and ς' separated by a kinship distance $\Delta_{\varsigma\varsigma'}$ is given by $h^{\Delta_{\varsigma\varsigma'}}$. Mother-daughters have $\Delta = 1$, sisters have $\Delta = 2$, cousins $\Delta = 4$ and so on. Importantly, in this scheme daughter-daughter correlations are the square of mother-daughter correlations making daughters independent conditional on their common mother. As will be shown in Section 5, the underlying graphical model for this branching process is a binary tree which is generally not the case for real lineages. For our purposes, this stationary branching process represents a null model since, despite the presence of correlations between family members, there are no preferred differentiation stages.

Sample lineages from these 3 lineage types are shown in Fig. 1 while the expression of each phenotype as a function of generation is shown in Fig. 2.

3. Modelling Framework and Labelling Conventions. As with any statistical model, we must first assign variables to each data point. In general, a lineage measurement, Y_{hijk}, might be indexed by 4 factors: Conditions (h), Family (i), Member (j), and Trait (k). The Condition factor corresponds to the cell type being studied or the experimental arrangement under which a founder cell is chosen or cultured (and may itself consist of multiple factors). Each factor level of Family refers to a particular founder cell and its descendents; each level of Member corresponds to a position in the family tree; and each level of Trait refers to a given phenotype recorded for a cell (such as average size or marker expression intensity).

To focus on the associations among family members, we restrict our attention to modelling a single trait from families subject to the same conditions. A multi-family sample can then be represented by a two-factor array (Y_{ij}), where i has n levels corresponding to the number of families and j has p levels corresponding to the number of members within a family. With no meaningful distinctions among families (they are all of the same cell type and subject to the same conditions) we assume families are independent and identically distributed replicates. The data can thus be represented by a matrix $\boldsymbol{Y}$ with n replicates (rows) and p variables (columns).

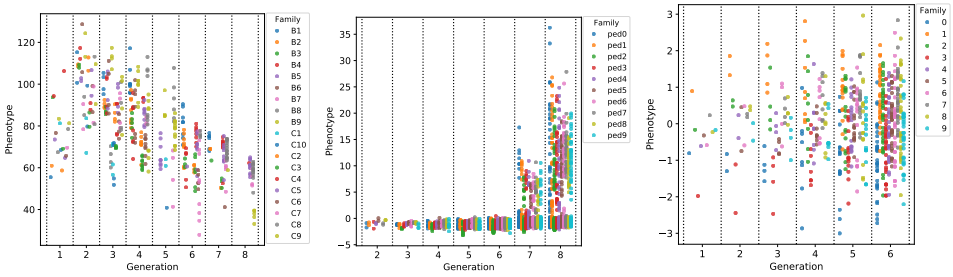


a) T cells

b) Worm

c) Branching process

FIG 1. *Comparison of some sample lineages. Colouring of the nodes reflects quantification of the phenotype under analysis (average area over lifetime for T cells, PHA-4 expression for C. elegans). The absence of a node on a branch represents a missing data point. Note that for the T cell lineage the root node is the naive cell while for the worm lineage the root node is the zygote (labelled P0 in the C. elegans naming convention).*



a) T cells b) Worm c) Branching process

FIG 2. *Expression of each phenotype as a function of generation. For T cells the measured phenotype is the average cell area in μm^2; for C. elegans it is PHA-4 fluorescence intensity.*

Each of the p dimensions corresponds to a family member. We use a binary number to label each family member so that, for example, the first 3 generations are labelled as founder (`1`), daughters (`10`, `11`), and granddaughters (`100`, `101`, `110`, `111`), where each label thus encodes the family member position.

The group symmetry methodology described in Section 4.2.5 demands a clear distinction between the terms generation and division: generation refers to the depth of a family member in a tree while division refers to a process occurring between two adjacent generations. A cell thus belongs to a generation but arises from a division. It will be necessary to assign each division a unique two-factor index (ℓ, τ) with ℓ referring to the longitudinal coordinate of the division and τ to the transverse coordinate. In our convention, a cell in generation g always arises from a division with $\ell = g$. We choose the convention that the founder cell is in generation 1. This means that the first division within the family is actually $\ell = 2$ since it produces daughters in $g = 2$. The 'division' $\ell = 1$ refers to a process, occurring outside the family, that gave rise to the founder cell. Variation attributed to $\ell = 1$ therefore refers to inter-family variation. Fig. 3 gives a summary of these definitions and conventions.

Often in lineage measurements there are many more members of a family (p) than there are families (n). Thus $p \gtrsim n$, with the disparity getting exponentially worse with the number of generations studies. Performing reliable inference when $p/n > 1$ is an open research question (Hastie, Tibshirani and Wainwright, 2015). Best results are achieved when prior knowledge of the problem can be incorporated.

In the next section we describe increasingly more sophisticated steps to reduce the effective dimensionality of the inference calculation, first by exploiting known symmetry properties and then by using observed sparsity properties. Our practical goal is to identify a scheme where the data requirement of the model (the number of replicates required to infer its parameters) is independent of the number of generations studied.

4. Statistical Inference. Our objective is to infer the joint probability distribution $\mathcal{P}(\boldsymbol{y})$ from a sample of size n where $\boldsymbol{y}$ is a p-dimensional random variable representing the single trait for each family member in the lineage. In this study we discuss inference for a multivariate Gaussian since the traits we examine for T cells and *C. elegans* are continuous and approximately marginally Gaussian. Thus,

$$\mathcal{P}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right],$$

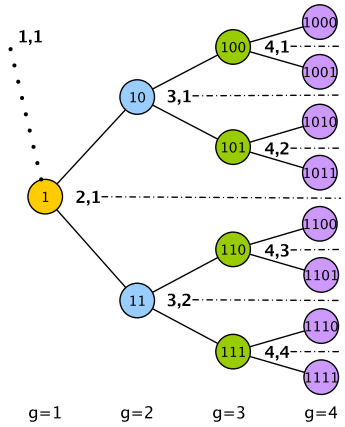


FIG 3. *Labeling convention for family members, generations, and divisions. Each family member is identified with a binary number. The founder cell is defined to be at generation* $g = 1$. *Each division is uniquely identified with a 2-index label written as* (ℓ, τ) *where* ℓ *refers to the longitudinal coordinate and* τ *refers to the transverse coordinate. 'Division'* $(1, 1)$ *actually represents the process that distinguishes different founder cells and will be use to label inter-family variation. According to group representation theory, each division is a potential source of variation. Importantly, in this study, only the longitudinal division coordinates are distinguishable.*

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix and $\boldsymbol{\mu}$ is the multivariate mean.

Our preliminary goal is to infer the maximum-likelihood estimate $\hat{\boldsymbol{\Sigma}}$ and its inverse, the precision matrix $\hat{\boldsymbol{K}} = \hat{\boldsymbol{\Sigma}}^{-1}$. Since there are many shared associations between family members it will turn out to be more useful to estimate the covariance matrix for a natural set of variables. These will be determined from group symmetry arguments.

4.1. *Unstructured Gaussian.* A naive method for finding $\hat{\boldsymbol{\Sigma}}$ is to assume no prior structure on $\boldsymbol{\Sigma}$, allowing for all possible associations between family members. Then, if the sample mean $(\overline{\boldsymbol{y}})$ and (biased) sample covariance $(\boldsymbol{S})$ are given by the usual

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y}_i, \quad \boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y}_i \boldsymbol{Y}_i' - \overline{\boldsymbol{y}}\,\overline{\boldsymbol{y}}' \tag{1}$$

where $\boldsymbol{Y}_i$ is the data vector from family i, the mean and variance-covariance estimates are given explicitly by

$$\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{y}}, \quad \hat{\boldsymbol{\Sigma}} = \boldsymbol{S} \tag{2}$$

Using this to analyse the first G generations, the effective number of dimensions p_{eff}, the number of unknown variance-covariance parameters $\mathcal{N}_\Sigma$, and the minimum number of replicates $n_{\min}$ required to ensure existence of the MLE are:

$$p_{\text{eff}} = p, \quad \mathcal{N}_\Sigma = p(p+1)/2, \quad n_{\min} = p+1. \tag{3}$$

where the number of family members $p = 2^G - 1$. Although $p_{\text{eff}} = p$ for this unstructured case, with group symmetries $p_{\text{eff}} < p$.

Note how $n_{\min}$ increases exponentially with the number of generations G being studied, making this simple model impractical for analyzing trees. Nevertheless it provides a reference for our improved models which aim to make $n_{\min}$ independent of G. For each model we will examine the reduction in p_{eff}, $\mathcal{N}_\Sigma$ and $n_{\min}$, as will be shown.

4.2. *Symmetry.* To reduce $n_{\min}$ requires identifying constraints. Constraints on $\mathbf{\Sigma}$ can be found from the group symmetry properties of a tree. We first demonstrate how this pattern can be determined by inspection alone and then use group representation theory to identify the natural set of variables associated with this pattern.

4.2.1. *Shared Parameters.* To reduce the number of unknowns in the model, we start by identifying a pattern of shared parameters in the covariance matrix. The shared parameters arise from the indistinguishability of daughters and their subtrees.

For example, consider the pair of cells `10` and `110` which have `1` as their Most Recent Common Ancestor (MRCA). We can uniquely identify the covariance matrix element for this pair by using the 3-index 231 to specify the generation of each cell (2 and 3) and the generation of their MRCA (1). Now because of daughter indistinguishability, the association between a different cell pair, `11` and `101`, must have the same 3-digit label 231. We can proceed to give a 3-index label to each covariance matrix element, leading to the following patterned covariance matrix $\mathbf{\Sigma}_{\mathcal{G}}$ for the first 3 generations:

$$
\begin{array}{c} \\ \boldsymbol{\Sigma}_{\mathcal{G}} = \end{array}
\begin{array}{c}
\begin{array}{cccccccc} 1 & 10 & 11 & 100 & 101 & 110 & 111 & \end{array} \\
\left(\begin{array}{c|cc|cccc}
\sigma_{111} & \sigma_{121} & \sigma_{121} & \sigma_{131} & \sigma_{131} & \sigma_{131} & \sigma_{131} \\
\hline
\sigma_{121} & \sigma_{222} & \sigma_{221} & \sigma_{232} & \sigma_{232} & \sigma_{231} & \sigma_{231} \\
\sigma_{121} & \sigma_{221} & \sigma_{222} & \sigma_{231} & \sigma_{231} & \sigma_{232} & \sigma_{232} \\
\hline
\sigma_{131} & \sigma_{232} & \sigma_{231} & \sigma_{333} & \sigma_{332} & \sigma_{331} & \sigma_{331} \\
\sigma_{131} & \sigma_{232} & \sigma_{231} & \sigma_{332} & \sigma_{333} & \sigma_{331} & \sigma_{331} \\
\sigma_{131} & \sigma_{231} & \sigma_{232} & \sigma_{331} & \sigma_{331} & \sigma_{333} & \sigma_{332} \\
\sigma_{131} & \sigma_{231} & \sigma_{232} & \sigma_{331} & \sigma_{331} & \sigma_{332} & \sigma_{333}
\end{array}\right)
\begin{array}{c} 1 \\ 10 \\ 11 \\ 100 \\ 101 \\ 110 \\ 111 \end{array}
\end{array}
\tag{4}
$$

where lines separate different generations. We use the subscript $\mathcal{G}$ to indicate when a matrix has this patterned structure. Because a generation denotes the lineage distance to the founder cell, a meaningful quantity in cell differentiation, we need to specify the generations of the cell pair and its MRCA, not just generational differences. For example, sisters in generation 3 may have different statistical associations to sisters in generation 2.

The MLE of a Gaussian with this form of $\boldsymbol{\Sigma}$ can be found explicitly. To understand why, consider the log-likelihood of the multivariate Gaussian

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{K}; \boldsymbol{Y}) &= \ln \mathcal{P}(\boldsymbol{Y}; \boldsymbol{\mu}, \boldsymbol{K}) \\
&= \frac{n}{2} \left[\ln \det \boldsymbol{K} - \operatorname{tr}(\boldsymbol{S}\boldsymbol{K}) - p \ln(2\pi) \right]
\end{aligned}
\tag{5}
$$

over positive definite matrices $\boldsymbol{K}$. If the pattern is in the covariance matrix (Anderson, 1973), maximum likelihood (ML) optimisation is not in general convex; if the pattern is in the precision matrix, ML optimization is convex and leads to explicit estimates for $\boldsymbol{\Sigma}$ (Hojsgaard and Lauritzen, 2008; Szatrowski, 1980). Since, as we will show in the next section, this particular pattern arises from group symmetries in the binary tree, the pattern is invariant under the action of an inverse and is thus the same for covariance and precision matrices. Convex optimisation is thus guaranteed.

To constrain the structure of $\boldsymbol{K}$ we represent it as a linear combination of matrices. The resulting patterned inverse covariance is then

$$
\boldsymbol{K}_{\mathcal{G}} = \sum_{\alpha} a_{\alpha} \boldsymbol{A}_{\alpha} \tag{6}
$$

where each $\boldsymbol{A}_{\alpha}$ is a matrix of 0's and 1's which has the same dimensions as $\boldsymbol{K}_{\mathcal{G}}$, with 1's identifying a particular shared parameter and a_{α} giving the value of that parameter.

The MLE is found by differentiating Eq. 5 with respect to each a_α and setting $d\mathcal{L}/da_\alpha = 0$, giving

$$\frac{d}{da_\alpha} \ln\det \boldsymbol{K}_\mathcal{G} = \frac{d}{da_\alpha} \mathrm{tr}(\boldsymbol{S}\boldsymbol{K}_\mathcal{G}). \tag{7}$$

Substituting Eq. 6 gives

$$\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_\mathcal{G}\boldsymbol{A}_\alpha) = \mathrm{tr}\left(\boldsymbol{S}\boldsymbol{A}_\alpha\right). \tag{8}$$

Since the matrices are symmetric we can equate the inner products of the matrices:

$$\langle \hat{\boldsymbol{\Sigma}}_\mathcal{G}, \boldsymbol{A}_\alpha \rangle = \langle \boldsymbol{S}, \boldsymbol{A}_\alpha \rangle. \tag{9}$$

Thus the MLE of each shared parameter in $\hat{\boldsymbol{\Sigma}}_\mathcal{G}$ is found by averaging the corresponding elements in $\boldsymbol{S}$ (Hojsgaard and Lauritzen, 2008). The result, $\boldsymbol{S}_\mathcal{G}$, is thus the MLE of the patterned covariance:

$$\hat{\boldsymbol{\Sigma}}_\mathcal{G} = \boldsymbol{S}_\mathcal{G} \tag{10}$$

This aggregation of selected elements improves the signal-to-noise in just the right way, enhancing information about all possible associations in the tree by averaging over the indistinguishability pattern.

Note that the MLE of the mean, $\hat{\boldsymbol{\mu}}$, is also given explicitly for a multivariate Gaussian with group symmetries (Gehrmann and Lauritzen, 2012). For the case of a binary tree $\hat{\boldsymbol{\mu}}$ is found by pooling data from family members sharing the same generation, following the pattern in the variances shown on the diagonal of $\boldsymbol{\Sigma}_\mathcal{G}$ (Eq. 4).

While the shared parameters have reduced the number of variance and covariance elements, it is difficult to evaluate how many replicates, $n_{\min}$, are now required to ensure that $\boldsymbol{S}$ is positive definite (Uhler, 2012). The answer will become apparent when we examine the invariant subspaces of the group representation.

4.2.2. *Indistinguishability and Symmetry Invariance.* The shared parameter pattern, found above by inspection, arises directly from the symmetry invariance properties of the tree. These symmetries have deep implications, beyond just optimal data aggregation.

An object possesses a symmetry if it remains invariant under the actions of a group. Symmetries are possible in a binary tree because daughters are statistically indistinguishable. We can thus exchange the two subtrees descended from any ancestor without altering the joint distribution over the

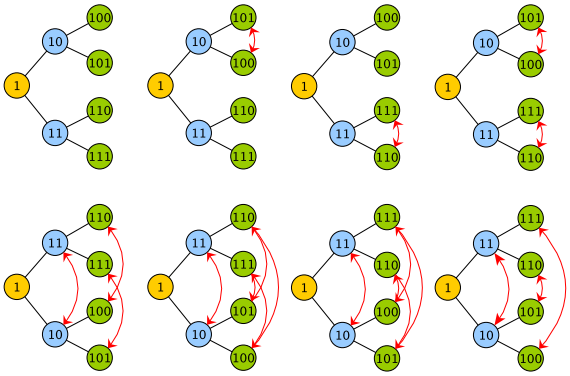


FIG 4. *Group symmetry in a binary tree arises from the possible permutations of family members that preserve their mutual relatedness. In this example, the 8 permutations of a tree with 3 generations are shown.*

tree. Here we use the term ancestor to mean a branch node, i.e. any family member that has daughters.

The complete set of group actions is found by considering all the ancestors together, allowing each to be in one of two 'states': having its subtrees exchanged or not. For a tree with G generations and thus $\mathcal{A} = 2^{G-1} - 1$ ancestors, there are $2^{\mathcal{A}}$ unique configurations of all ancestor states that keep the family relationships invariant. These configurations form the complete set of elements in the group of order $2^{\mathcal{A}}$.

For example, a tree consisting of the first 3 generations has $\mathcal{A} = 3$ (corresponding to members `1`, `10`, and `11`). The order of the group is thus $2^3 = 8$. Each of the 8 permutations is shown in Fig. 4.

4.2.3. *Group-Averaged MLE.* Symmetry constrains the joint distribution because each relationship-preserving configuration of family members is a permutation of the p variables that keeps $\mathcal{P}(\boldsymbol{y})$ invariant. Thus if $\boldsymbol{D}_s$ is the p-dimensional permutation matrix representing an action s of the group $\mathcal{G}$, symmetry invariance (referred to as $\mathcal{G}$-invariance) requires $\mathcal{P}(\boldsymbol{D}_s\boldsymbol{y}) = \mathcal{P}(\boldsymbol{y})$, $\forall s \in \mathcal{G}$. For a multivariate Gaussian, this means that the covariance matrix belongs to the set

$$\mathcal{W}_{\mathcal{G}} = \{\boldsymbol{M} \in \mathbb{R}^{p\times p} |\ \boldsymbol{D}_s\boldsymbol{M}\boldsymbol{D}_s' = \boldsymbol{M} \quad \forall s \in \mathcal{G}\}, \tag{11}$$

referred to as the fixed point subspace of the group $\mathcal{G}$. Note that a $\mathcal{G}$-invariant covariance matrix implies a $\mathcal{G}$-invariant precision matrix since if $\boldsymbol{D}_s\boldsymbol{\Sigma}\boldsymbol{D}_s' = \boldsymbol{\Sigma}$ then $\boldsymbol{D}_s\boldsymbol{\Sigma}^{-1}\boldsymbol{D}_s' = \boldsymbol{\Sigma}^{-1}$.

A fundamental technique for transforming an unconstrained matrix $\boldsymbol{M}$ into one that is symmetry invariant is the group-average or Reynolds operator given by:

$$\mathbb{P}_{\mathcal{G}}(\boldsymbol{M}) = \frac{1}{|\mathcal{G}|} \sum_{s \in \mathcal{G}} \boldsymbol{D}_s \boldsymbol{M} \boldsymbol{D}_s', \quad \mathbb{P}_{\mathcal{G}} : \mathbb{R}^{p \times p} \to \mathcal{W}_{\mathcal{G}} \tag{12}$$

where $|\mathcal{G}|$ is the order of the group. This projects the matrix onto the fixed point subspace by averaging with respect to the orbits of the group. It is straightforward to demonstrate that the pattern that arises from $\mathbb{P}_{\mathcal{G}}(\boldsymbol{\Sigma})$, when $\mathcal{G}$ is the symmetry group of the tree, is the same as that shown in Eq 4 that was found by inspection.

The requirement that $\boldsymbol{\Sigma} \in \mathcal{W}_{\mathcal{G}}$ places a constraint on the log-likelihood optimisation (Eq. 5). The resulting constrained MLE is given by (Shah and Chandrasekaran, 2012)

$$\hat{\boldsymbol{\Sigma}}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{s \in \mathcal{G}} \boldsymbol{D}_s \boldsymbol{S} \boldsymbol{D}_s', \tag{13}$$

or just $\hat{\boldsymbol{\Sigma}}_{\mathcal{G}} = \mathbb{P}_{\mathcal{G}}(\boldsymbol{S})$ from Eq. 12. Thus the MLE is found by projecting the sample covariance onto the fixed point subspace of the group, a result that necessarily agrees with Eq. 9 since the underlying pattern is the same. $\mathbb{P}_{\mathcal{G}}()$ is the general operation for pooling data with group symmetries that preserves associations between variables.

This method for finding the shared parameters is only practical when the order of the group, $|\mathcal{G}|$, is small. In a binary tree, $|\mathcal{G}|$ increases super-exponentially, as $2^{\mathcal{A}}$ where $\mathcal{A} = 2^{G-1} - 1$. Thus, even with 4 generations, $|\mathcal{G}| = 128$, and the group-averaging approach of Eq. 13 is awkward to implement in practice.

4.2.4. *Decomposition into Irreducible Components.* Deeper insight, and computational benefit, arises from examining the invariant subspaces of the group representation. This identifies a set of orthogonal components that are a particularly meaningful and concise way of describing tree-structured variation. In this subsection we briefly summarise the general theory for decomposition according to group symmetries, following (Stiefel and Fässler, 1992; Shah and Chandrasekaran, 2012). In the next subsection we apply it to a binary tree.

Let $\vartheta : s \to \boldsymbol{D}_s \ \forall s \in \mathcal{G}$ be the representation of $\mathcal{G}$ on the vector space $\mathcal{V} \in \mathbb{R}^p$. From Maschke's theorem, and by induction, a linear representation ϑ of a finite group is a direct sum of irreducible representations. Accordingly,

$\vartheta = \bigoplus_{\omega=1}^{\mathcal{I}} m_\omega \vartheta^{(\omega)}$ where each inequivalent irreducible representation $\vartheta^{(\omega)}$ has a multiplicity m_ω and a dimensionality d_ω such that $p = \sum_{\omega=1}^{\mathcal{I}} m_\omega d_\omega$. This ensures that any matrix $\boldsymbol{M} \in \mathcal{W}_\mathcal{G}$ can be decomposed as

$$\boldsymbol{T}^\dagger \boldsymbol{M} \boldsymbol{T} = \begin{pmatrix} \mathcal{C}^{(1)} & & 0 \\ & \ddots & \\ 0 & & \mathcal{C}^{(\mathcal{I})} \end{pmatrix} \tag{14}$$

where each $\mathcal{C}^{(\omega)} \in \mathbb{R}^{m_\omega d_\omega \times m_\omega d_\omega}$ corresponds to an isotypic component. Here $\boldsymbol{T}$ is a unitary change-of-basis matrix that transforms $\boldsymbol{M}$ to a symmetry-adapted basis where its block diagonal form is revealed.

Furthermore, Schur's lemma states that, since $\mathcal{C}^{(\omega)}$ and ϑ_ω commute, the isotypic components themselves decompose and can be written as a direct sum of repeated subblocks,

$$\mathcal{C}^{(\omega)} = \begin{pmatrix} M_\Omega^{(\omega)} & & 0 \\ & \ddots & \\ 0 & & M_\Omega^{(\omega)} \end{pmatrix} \tag{15}$$

where there are d_ω repeated subblocks of $M_\Omega^{(\omega)} \in \mathbb{R}^{m_\omega \times m_\omega}$. Note that this requires the appropriate arrangement of symmetry-adapted basis vectors within each isotypic component as specified in $\boldsymbol{T}$ (Stiefel and Fässler, 1992).

The fundamental decomposition of $\boldsymbol{M} \in \mathcal{W}_\mathcal{G}$ is thus

$$\boldsymbol{T}^\dagger \boldsymbol{M} \boldsymbol{T} = \bigoplus_{\omega=1}^{\mathcal{I}} \left[\bigoplus_{\nu=1}^{d_\omega} M_\Omega^{(\omega)} \right], \quad M_\Omega^{(\omega)} \in \mathbb{R}^{m_\omega \times m_\omega}, \tag{16}$$

where ν indexes the repeated subblocks $M_\Omega^{(\omega)}$. This expression highlights how the $M_\Omega^{(\omega)}$ are orthogonal building blocks of $\boldsymbol{M}$. Each $M_\Omega^{(\omega)}$ is referred to as an irreducible block while the set of m_ω variables from which it is composed is an irreducible component. Irreducible blocks are thus the main objects of inference and, as we will see for the case of a binary tree, represent the variation arising from each division.

Since $\boldsymbol{M}$ is either $\boldsymbol{\Sigma}_\mathcal{G}$ or $\boldsymbol{K}_\mathcal{G}$ in this paper, Eq. 16 states that a model with p variables decomposes into $\mathcal{I}$ unique irreducible components each with m_ω variables. This reduces the number of pairwise associations in the model since Eq. 16 only permits associations between variables *within* an irreducible component. It also reduces the total number of unique variables since only one of the d_ω identical copies of each irreducible component needs to be considered.

4.2.5. *Decomposition for a Binary Tree.* Here we show the results of applying group representation theory (Serre, 1977; Diaconis, 1988; Stiefel and Fässler, 1992) to the vector space defined by all members of a binary tree. Detailed calculations (see Supplement S1) show that for a completely reducible representation on a binary tree with G generations, the number of active irreducible representations $\mathcal{I}$ (here indexed by $1 \leq \omega \leq \mathcal{I}$), their dimensionalities d_ω and multiplicities m_ω are

$$\mathcal{I} = G \tag{17}$$

$$d_\omega = \begin{cases} 1, & \text{if } \omega = 1, 2 \\ 2^{\omega-2}, & \text{if } \omega \geq 3 \end{cases} \tag{18}$$

$$m_\omega = G - \omega + 1. \tag{19}$$

As required, this satisfies $p = \sum_{\omega=1}^{\mathcal{I}} m_\omega d_\omega$ when $p = 2^G - 1$.

The change-of-basis matrix $\boldsymbol{T}$ is calculated using the standard method for identifying the symmetry-adapted basis (Stiefel and Fässler, 1992). For example, with 3 generations:

$$\boldsymbol{T} = \begin{array}{c} \begin{array}{ccc|ccc|c} \textit{1,1} & \textit{1,1} & \textit{1,1} & \textit{2,1} & \textit{2,1} & \textit{3,1} & \textit{3,2} \\ \textit{1} & \textit{2} & \textit{3} & \textit{2} & \textit{3} & \textit{3} & \textit{3} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{-1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{-1}{2} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{-1}{2} & 0 & \frac{-1}{\sqrt{2}} \end{pmatrix} \end{array} \begin{array}{l} \texttt{1} \\ \texttt{10} \\ \texttt{11} \\ \texttt{100} \\ \texttt{101} \\ \texttt{110} \\ \texttt{111} \end{array} \tag{20}$$

This unitary (and orthonormal) matrix defines a set of symmetry-adapted, or natural, variables in columns based on linear combinations of the standard variables in rows. This transformation of data to its natural basis defined by a group symmetry is referred to simply as spectral analysis by Diaconis, see Chapter 8 (Diaconis, 1988).

The parts of Haar transformation matrices (Strang, 1993) for each generation can be recognised and are outlined in blue for generation 2 and green for generation 3. Because representation theory requires the natural bases be grouped into isotypic components, shown here separated by vertical lines, $\boldsymbol{T}$ is not simply a direct sum of Haar matrices. The column headers give

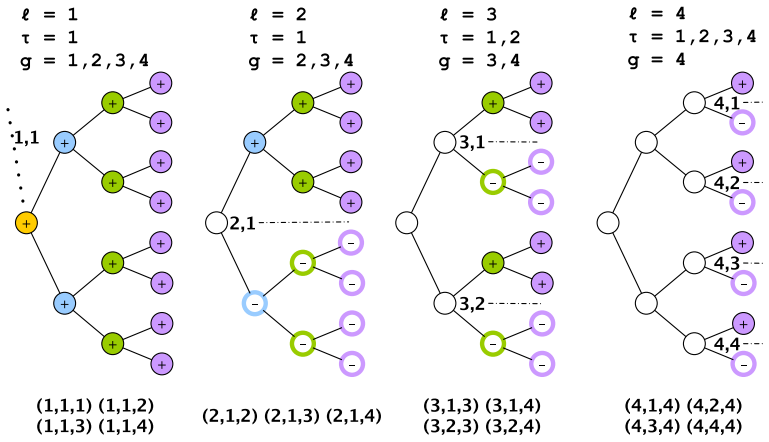


FIG 5. *Construction of the natural variables for a tree with 4 generations. The + and − on each family member indicates how members of a generation are combined. Each natural variable is identified by the division (ℓ, τ) on which the asymmetry is centered and the generation g on which the variation is observed. The 15 natural variables thus defined by the 3-tuple (ℓ, τ, g) are listed in the bottom row. Since the transverse division coordinates are arbitrary and there is no way to distinguish between them, only 10 of these variables (those with $\tau = 1$, say) are unique.*

the natural variables in terms of divisions and generations, following the interpretation to be described next.

How the natural variables are constructed from standard variables is illustrated in Fig. 5 for the case of 4 generations. Examining these combinations suggests labelling each natural variable with a 3-integer tuple (ℓ, τ, g) where

ℓ: Longitudinal division The longitudinal coordinate of a division, ℓ, corresponds to an isotypic component, ω. The fact that there are G longitudinal division coordinates is consistent with there being $\mathcal{I} = G$ isotypic components (Eq. 17).

τ: Transverse division The transverse coordinate of a division, τ, corresponds to an irreducible component ν within an isotypic component. The fact that there is 1 transverse division coordinate for $\ell = 1, 2$ and $2^{\ell-2}$ coordinates for $\ell \geq 3$ (see e.g. Fig. 5) is consistent with the dimensionalities d_ω found for each irreducible representation (Eq. 18).

g: Generation The variables within each irreducible component correspond to the generations g at which variation arising from division (ℓ, τ) can be observed. Given that variation from a division can only be observed in members after that division, g is restricted to $\ell \leq g \leq G$ and thus has $G - \ell + 1$ values. This is consistent with the multiplicity

of each irreducible representation being $m_\omega = G - \omega + 1$ (Eq. 19). Importantly, since each accessible g occurs exactly once in each irreducible component, the irreducible component consists of an ordered sequence and is thus a time series.

The most important aspect of the natural variables is their second order properties. It is straightforward to check that, using $\boldsymbol{T}$ from Eq. 20, the natural decomposition Eq. 16 transforms the patterned covariance, $\boldsymbol{\Sigma}_\mathcal{G}$ (Eq. 4), into its block-diagonal form, $\boldsymbol{\Sigma}_\Omega$ (see Supplement S2). Just to be clear on the subscript notation, the covariance matrix (or its inverse) can take on 3 forms: unstructured $\boldsymbol{\Sigma}$, patterned $\boldsymbol{\Sigma}_\mathcal{G}$, and transformed into its natural basis $\boldsymbol{\Sigma}_\Omega$. Then for the 3-generation tree,

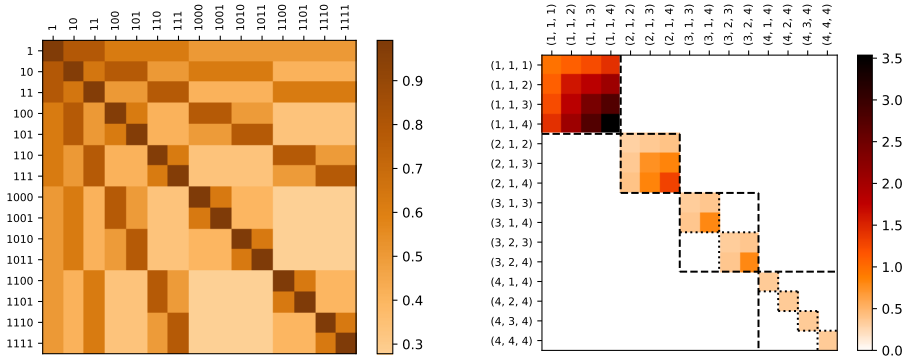
$$\boldsymbol{\Sigma}_\Omega = \boldsymbol{T}^\dagger \boldsymbol{\Sigma}_\mathcal{G} \boldsymbol{T} \tag{21}$$

$$= \begin{pmatrix} \xi^{(1)}_{11} & \xi^{(1)}_{12} & \xi^{(1)}_{13} & \cdot & \cdot & \cdot & \cdot \\ \xi^{(1)}_{12} & \xi^{(1)}_{22} & \xi^{(1)}_{23} & \cdot & \cdot & \cdot & \cdot \\ \xi^{(1)}_{13} & \xi^{(1)}_{23} & \xi^{(1)}_{33} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \xi^{(2)}_{22} & \xi^{(2)}_{23} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \xi^{(2)}_{23} & \xi^{(2)}_{33} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \xi^{(3)}_{33} & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 & \xi^{(3)}_{33} \end{pmatrix} \tag{22}$$

Matrix elements within each isotypic block are labelled with a superscript (ℓ) and 2 subscripts referencing the interacting generations. Each division ℓ forms an isotypic block. Starting at $\ell = 3$, degeneracy occurs and repeated eigenvalues appear. Elements outside the isotypic blocks are zero and labelled with a dot; elements inside an isotypic block but outside an irreducible block are zero and labelled with a 0. For the case of 4 generations, $\boldsymbol{\Sigma}_\Omega$ is shown as a heat map in Fig. 6, alongside $\boldsymbol{\Sigma}_\mathcal{G}$

Each division (ℓ, τ) thus represents an independent source of variation. This means that variation at generation g can be decomposed into independent contributions from all divisions where $\ell \leq g$. This is simply the traditional concept of variance components in our language of divisions and generations. We will use this to estimate the contributions of each division to the pattern of fate expression (see Section 6).

The decomposition has thus taken a model with associations between all variables and partitioned it into a set of independent ordered sequences each representing the effects of variation from a division (ℓ, τ). In summary: (i) Each natural variable (ℓ, τ, g) represents variation originating at division (ℓ, τ) and observed at generation g, (ii) Each division is a source of varia-



a) $\mathbf{\Sigma}_{\mathcal{G}}$ b) $\mathbf{\Sigma}_{\Omega} = \boldsymbol{T}^{\dagger}\mathbf{\Sigma}_{\mathcal{G}}\boldsymbol{T}$

FIG 6. *Heat map of (a) the patterned covariance $\mathbf{\Sigma}_{\mathcal{G}}$, and (b) the corresponding transformed covariance $\mathbf{\Sigma}_{\Omega}$ for the first 4 generations, taken from the branching process. Natural variables along the axes of $\mathbf{\Sigma}_{\Omega}$ are given in the format (ℓ, τ, g). Isotypic blocks are bounded by dashed squares and correspond to a longitudinal division coordinate, ℓ. Irreducible blocks correspond to a unique division (ℓ, τ) and are bounded by a dotted square. Each irreducible component has a dimension m_ℓ and is repeated d_ℓ times in its isotypic component ℓ. For $\ell = 1, 2$ the isotypic and irreducible blocks coincide since $d_\ell = 1$.*

tion that can influence generations after it, (iii) Variation originating from different divisions is independent.

Note that because the transverse division coordinates identify identical irreducible components we only need to consider one irreducible component per isotypic block, say $(\ell, 1)$. Variation from different transverse divisions τ with the same ℓ are thus identically distributed. This is a consequence of the degeneracy of eigenvalues arising from the non-commutative symmetry group. Physically this is because transverse division coordinates cannot be distinguished. Thus, in practice we will only decompose variation into longitudinal division components, not transverse division components.

These natural variables provide an intuitive language for characterising differentiation patterns in a tree. Fig. 7 shows examples of some differentiation 'collective modes' (to use a term from physics) and their corresponding ℓ and g. Since τ's are not distinguishable we will ignore their values.

Note that if we only had data from a single generation, G, the symmetry group would still be the same as for the case of the entire tree. However, the group would be represented on a vector space with dimension 2^{G-1}, given by the number of leaves of the tree. In this case $\boldsymbol{T}$ reduces to the Haar transformation matrix (Strang, 1993), and $\mathbf{\Sigma}_{\Omega}$ is diagonal with eigenvalues given by classical nested ANOVA (Speed, 1987). $\mathbf{\Sigma}_{\Omega}$ is then the spectral covariance.

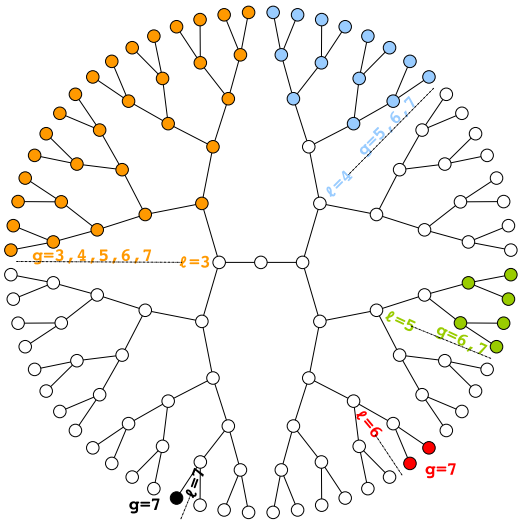


FIG 7. *Examples of asymmetric expression patterns in a tree and the ℓ-divisions to which they correspond. Each colour identifies an irreducible component associated with division ℓ and observed at some generations $\{g\}$. For example, the green asymmetry arises at division 5 but is not observed until generations 6 and 7. Note that transverse division coordinates, τ, cannot be diagnosed and are ignored.*

What is new here is that we have examined the group representation on *all* the generations up to and including G, a vector space with dimension $2^G - 1$. The abstract group is same in both cases; the difference is in its representation.

We remark that these variables are called natural because the underlying group symmetry of the tree required the standard variables be aggregated in just this way. It is satisfying that the groupings correspond to the those one would intuitively want to define in a nested ANOVA if we were to consider each generation separately.

4.2.6. *MLE from Irreducible Components.* The linear decomposition of a matrix into block-diagonal form reduces the single MLE calculation over all p variables into several smaller independent MLE calculations. To see this, let $\boldsymbol{K} \in \mathcal{W}_{\mathcal{G}}$. Then using Eq. 16 to decompose $\boldsymbol{K}$ into irreducible blocks

$K_\Omega^{(\ell)}$ gives:

$$\begin{aligned} \ln\det \boldsymbol{K} &= \ln\det \boldsymbol{T}^\dagger \boldsymbol{K}\boldsymbol{T} = \ln\det \boldsymbol{K}_\Omega \\ &= \sum_{\ell=1}^{\mathcal{I}} d_\ell \ln\det K_\Omega^{(\ell)} \end{aligned} \tag{23}$$

and

$$\begin{aligned} \operatorname{tr}(\boldsymbol{S}\boldsymbol{K}) &= \operatorname{tr}\left(\boldsymbol{T}^\dagger \boldsymbol{S}\boldsymbol{T}\boldsymbol{T}^\dagger \boldsymbol{K}\boldsymbol{T}\right) = \operatorname{tr}(\boldsymbol{S}_\Omega \boldsymbol{K}_\Omega) \\ &= \sum_{\ell=1}^{\mathcal{I}} \sum_{\tau=1}^{d_\ell} \langle S_\Omega^{(\ell,\tau)}, K_\Omega^{(\ell)} \rangle \end{aligned} \tag{24}$$

Note that $K_\Omega^{(\ell)}$ has the same dimensionality as $S_\Omega^{(\ell,\tau)}$ but is independent of τ. Substituting Eqs. 23 and 24 in Eq. 5, it is thus apparent that each irreducible block can be treated as an independent MLE calculation. The result,

$$\hat{\boldsymbol{\Sigma}}_\Omega^{(\ell)} = \frac{1}{d_\ell} \sum_{\tau=1}^{d_\ell} S_\Omega^{(\ell,\tau)} = \overline{S_\Omega^{(\ell)}} \tag{25}$$

states that $\hat{\boldsymbol{\Sigma}}_\Omega^{(\ell)}$ is found by averaging the d_ℓ irreducible subblocks in the transformed sample covariance. This procedure ensures that elements of $\boldsymbol{S}_\Omega$ that are outside the block diagonal are ignored.

The resulting $\hat{\boldsymbol{\Sigma}}$ can be reconstructed by substituting Eq. 25 in Eq. 16 and transforming back to the original basis:

$$\hat{\boldsymbol{\Sigma}}_\Omega = \bigoplus_{\ell=1}^{\mathcal{I}} \left[\bigoplus_{\tau=1}^{d_\ell} \hat{\Sigma}_\Omega^{(\ell)} \right] \tag{26}$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{T}\hat{\boldsymbol{\Sigma}}_\Omega \boldsymbol{T}^\dagger \tag{27}$$

The procedure for finding $\hat{\boldsymbol{\Sigma}}$ for a $\mathcal{G}$-invariant covariance is thus as follows:

1. Transform $\boldsymbol{S}$ into the symmetry-adapted basis.
2. Zero the elements outside the irreducible blocks.
3. Average the irreducible blocks within each isotypic block (if there is more than one).
4. Transform back to the original basis.

This result is necessarily the same as the projection of $\boldsymbol{S}$ onto the fixed point subspace of the group (Eq. 13) but $\hat{\boldsymbol{\Sigma}}_\Omega$ provides a more compact and informative way of describing tree-structured variation.

Importantly, Eq. 25 gives an explicit MLE of the irreducible block that involves simply re-arranging terms in the sufficient statistic $\boldsymbol{S}$. Each of these blocks is thus a descriptive statistic for tree-structured data, involving linear combinations of data points, sums of squares and no parameters.

4.2.7. *Complexity of MLE calculation.* Transforming $\boldsymbol{\Sigma}$ to the symmetry-adapted basis makes it clear how much group symmetry reduces the complexity of the model. The effective number of dimensions, p_{eff}, is found by summing the dimensions of each unique irreducible component. The number of free parameters in the covariance matrix, $\mathcal{N}_\Sigma$, is found by summing the number of parameters in each unique irreducible block. The minimum number of replicates required, $n_{\min}$, is found from the dimensionality of the largest irreducible block ($\ell = 1$). Thus

$$p_{\text{eff}} = \sum_{\ell=1}^{G}(G-\ell+1) = \frac{G(G+1)}{2} = \mathcal{O}(G^2) \tag{28}$$

$$\mathcal{N}_\Sigma = \frac{1}{2}\sum_{\ell=1}^{G}(G-\ell+1)(G-\ell+2) = \frac{G}{6}(G+1)(G+2) = \mathcal{O}(G^3) \tag{29}$$

$$n_{\min} = G+1 = \mathcal{O}(G) \tag{30}$$

The group-symmetric model is thus significantly more constrained than the unstructured model (compare Eq. 3), with the number of parameters growing polynomially with G instead of exponentially. Note how $p_{\text{eff}} < p$ (when $G \geq 3$), a reduction in the effective number of dimensions that was not apparent from the fixed-point subspace perspective. Even with these symmetry constraints however, $n_{\min}$ still grows with G, albeit linearly (Eq. 30) instead of exponentially. This means that for a fixed set of n replicates there will always be a limit to the number of generations that can be analysed. An additional constraint is required.

4.3. *Sparsity.* The additional constraint comes from recognising that each irreducible component ℓ is a time series from generation ℓ to G (see Section 4.2.5). Together, the irreducible components form a set of G independent time series each starting at a different generation but all ending at G. A standard procedure for restricting the complexity of a time series is to consider a fixed order Markov process. This restricts each irreducible block of the transformed precision matrix to having non-zero values along a diagonal band (the tri-diagonal in the case of a 1st order Markov process). Remember that here it is the structure of each $\boldsymbol{K}_\Omega^{(\ell)}$ that is sparse; the precision matrix itself, $\boldsymbol{K}$, is not particularly sparse.

A restricted-order Markov chain is a simple case of a decomposable Gaussian graphical model (Speed and Kiiveri, 1986; Lauritzen, 1996) and thus yields an explicit MLE whose calculation we briefly summarise here. Following the standard procedure for a decomposable model, variables in the block are organised into cliques and separators, a straightforward exercise for a Markov chain of any order. The corresponding sub-blocks within $S_\Omega^{(\ell)}$ are defined as

$$S_{\Omega,c_i}^{(\ell)},\ i=1,...,\mathcal{N}_C; \quad S_{\Omega,s_i}^{(\ell)},\ i=2,...,\mathcal{N}_C$$

where the subscript c_i refers to a clique, s_i refers to a separator, and $\mathcal{N}_C$ is the number of cliques in the irreducible block. The MLE for an irreducible block is then given explicitly by (Lauritzen, 1996)

$$\hat{\boldsymbol{K}}_\Omega^{(\ell)} = \sum_{i=1}^{\mathcal{N}_C} \left\{ \left[S_{\Omega,c_i}^{(\ell)} \right]^{-1} \right\}^0 - \sum_{i=2}^{\mathcal{N}_C} \left\{ \left[S_{\Omega,s_i}^{(\ell)} \right]^{-1} \right\}^0 \tag{31}$$

$$\hat{\boldsymbol{\Sigma}}_\Omega^{(\ell)} = \left[\hat{\boldsymbol{K}}_\Omega^{(\ell)} \right]^{-1} \tag{32}$$

where the expression $\{\Upsilon\}^0$ denotes a matrix with the dimensions of $\hat{\boldsymbol{K}}_\Omega^{(\ell)}$ which has its appropriate sub-block occupied by Υ and zeros elsewhere.

This expression makes it clear that, since it is the inverse of the clique and separator sub-blocks that are required, it is only these sub-blocks that need to be positive definite. The mininum number of replicates required for positive definiteness is thus set by the order $\mathcal{M}$ of the Markov process, which is fixed, rather than by the size of the irreducible block, which grows linearly with G. In general, $n_{\min} = 2 + \mathcal{M}$ and we have finally achieved our goal of having the data requirements be independent of the number of generations being analysed. Note that restricting the non-zero parameters in the precision matrix to be on the diagonal band means that $\mathcal{N}_\Sigma \sim \mathcal{O}(G^2)$, down from the cubic dependence in Eq. 29. p_{eff} remains unchanged.

Inspection of the T-cell and worm lineage data show that, at least up to generation 4, non-zero values in $\boldsymbol{K}_\Omega^{(\ell)}$ are indeed primarily confined to the tri-diagonal. This justifies the (first-order) Markov process assumption, and we hereafter use it to extend the analysis to higher generations.

4.4. *Missing Data.* The MLE calculations for the models described above assume complete data. In reality, some measurements are missing, often because data collection is imperfect but also because cells die and have no descendants. This creates problems. Having to marginalise over the missing variables would mean that the MLE calculations are no longer convex

and the explicit expressions for the MLE no longer apply. Also, with such partially-balanced data we would not be able to perform the similarity transformation to the symmetry-adapted basis.

A simple solution is to apply the Expectation-Maximization (EM) Algorithm (Dempster, Laird and Rubin, 1977). This iteratively improves the estimate of the covariance matrix, generating expected values of the sufficient statistic at each step. In the E-step, the current estimate of the mean $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ are used to calculate the expected sufficient statistic for each replicate, conditioned on the observed data. The average sufficient statistic $\hat{\boldsymbol{S}}$ over all replicates is then calculated. In the M-step, $\hat{\boldsymbol{S}}$ is used in the MLE calculation of the irreducible blocks (as described above) to update the estimate $\hat{\boldsymbol{\Sigma}}$. The E and M steps are then repeated until $\hat{\boldsymbol{\Sigma}}$ converges.

In more detail (Little and Rubin, 2002), the first and second order statistics are calculated for each replicate i by partitioning the variables into observed sets, labelled o_i, and unobserved sets, labelled u_i. Members of each set usually differ from one replicate to the next. The vector of unobserved values in each replicate is then filled by its expected value conditioned on the vector of observed values:

$$\begin{aligned} \boldsymbol{Y}_{i,u_i} &= \mathbb{E}(\boldsymbol{Y}_{i,u_i}|\boldsymbol{Y}_{i,o_i}) \\ &= \hat{\boldsymbol{\mu}}_{u_i} + \hat{\boldsymbol{\Sigma}}_{u_i,o_i}\hat{\boldsymbol{\Sigma}}^{-1}_{o_i,o_i}\left(\boldsymbol{Y}_{i,o_i} - \hat{\boldsymbol{\mu}}_{o_i}\right) \end{aligned} \tag{33}$$

These combined with the observed values completes the first order statistic, $\boldsymbol{Y}_i = \{\boldsymbol{Y}_{i,o_i}, \boldsymbol{Y}_{i,u_i}\}$ for i.

The second order statistic $(\boldsymbol{YY})_i$ for each replicate i, partitioned into observed and unobserved sections, is found from

$$\begin{aligned} (\boldsymbol{YY}')_{i,o_io_i} &= \boldsymbol{Y}_{i,o_i}\boldsymbol{Y}'_{i,o_i} \\ (\boldsymbol{YY}')_{i,u_io_i} &= \boldsymbol{Y}_{i,u_i}\boldsymbol{Y}'_{i,o_i} \\ (\boldsymbol{YY}')_{i,o_iu_i} &= \boldsymbol{Y}_{i,o_i}\boldsymbol{Y}'_{i,u_i} \\ (\boldsymbol{YY}')_{i,u_iu_i} &= \boldsymbol{Y}_{i,u_i}\boldsymbol{Y}'_{i,u_i} + \hat{\boldsymbol{\Sigma}}_{u_iu_i|o_io_i}, \end{aligned} \tag{34}$$

where

$$\hat{\boldsymbol{\Sigma}}_{u_iu_i|o_io_i} = \hat{\boldsymbol{\Sigma}}_{u_i,u_i} - \hat{\boldsymbol{\Sigma}}_{u_i,o_i}\hat{\boldsymbol{\Sigma}}^{-1}_{o_i,o_i}\hat{\boldsymbol{\Sigma}}_{o_i,u_i}$$

is the residual covariance of the unobserved variables after conditioning on the observed variables.

Once this exercise has been completed for all replicates, the sample mean and covariance are calculated from the usual

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y}_i, \quad \hat{\boldsymbol{S}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{Y}\boldsymbol{Y}')_i - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}' \tag{35}$$

The estimated sample covariance, $\hat{\boldsymbol{S}}$, is then used in the procedures described in the previous sections to calculate a new estimate, $\hat{\boldsymbol{\Sigma}}$.

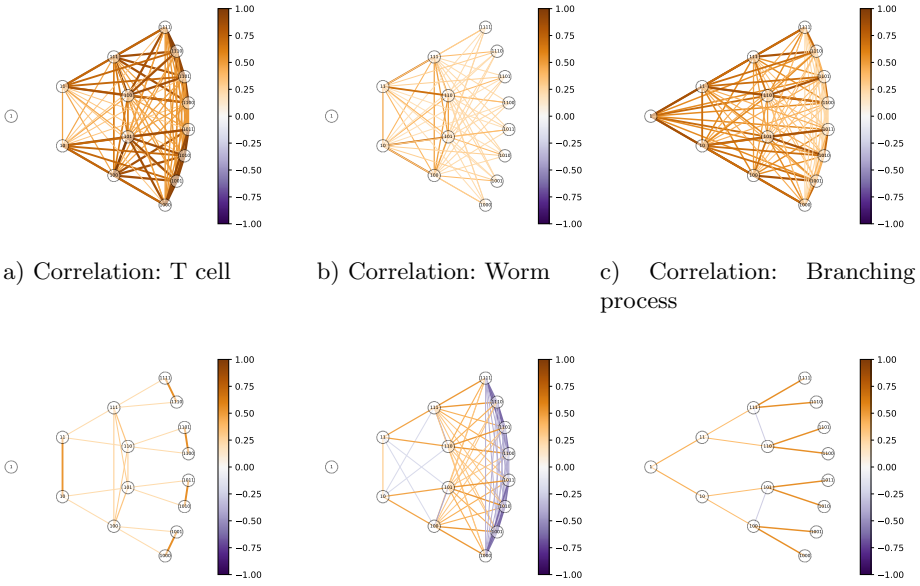
Iterating these steps gives the following algorithm:

1. Initialize $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$.
2. Expectation step to determine the expected value of the sufficient statistics for each replicate. Use Eqs. 33, 34, 35 to calculate the updated estimate $\hat{\boldsymbol{\mu}}$ and the estimated sample covariance, $\hat{\boldsymbol{S}}$
3. Maximization step to find $\hat{\boldsymbol{\Sigma}}$ from $\hat{\boldsymbol{S}}$.
   (a) Find $\hat{\boldsymbol{S}}_\Omega = \boldsymbol{T}^\dagger \hat{\boldsymbol{S}} \boldsymbol{T}$.
   (b) Set elements outside the diagonal blocks to zero.
   (c) Find the average of the repeated irreducible blocks in each isotypic component.
   (d) For each unique irreducible block, find $\hat{\boldsymbol{\Sigma}}_\Omega^{(\ell)}$ from $\hat{\boldsymbol{S}}_\Omega^{(\ell)}$ using Eq. 31 and 32, assuming a Markov chain of given order $\mathcal{M}$.
   (e) Recover $\hat{\boldsymbol{\Sigma}} = \boldsymbol{T}\hat{\boldsymbol{\Sigma}}_\Omega \boldsymbol{T}^\dagger$
4. Return to Step 2 until convergence.

5. Graphical Models of a Lineage. Having estimated the parameters in the model we can visualise and interpret the results using graphical models. Here we use two types of graphs: an undirected graph in the standard basis where the individual family members are nodes, and a directed graph in the symmetry-adapted basis where the natural variables are nodes. As will become apparent, the natural basis is a more compact way of viewing all the parameters at once.

5.1. *Standard Basis, Undirected Graph.* To visualise the network of statistical associations between different family members we use an undirected graph (Speed and Kiiveri, 1986; Lauritzen, 1996). Here we examine the graphs defined either by marginal or by conditional associations. The first is found from $\boldsymbol{\Sigma}$ while the second is found from $\boldsymbol{K}$.

For the network of marginal associations the strength of an edge between a pair of variables is defined by the Pearson correlation coefficient, $\rho_{jj'} = \sigma_{jj'}/\sqrt{\sigma_{jj}\sigma_{j'j'}}$ where $\sigma_{jj'}$ is an element of $\hat{\boldsymbol{\Sigma}}$. For the network of conditional



a) Correlation: T cell b) Correlation: Worm c) Correlation: Branching process

d) Partial correlation: T cell e) Partial correlation: Worm f) Partial correlation: Branching process

FIG 8. *Undirected graphs in the standard basis. The colour of edges in each graph corresponds to the correlation (top row) or partial correlation (bottom row) between pairs of family members. To avoid clutter only the first 4 generations are shown. Note how the graph of partial correlations (8f) is a binary tree for the simulated branching process but not for the real lineages. This emphasises how the graphical model of real lineages has the symmetry properties of a tree but not the sparsity properties.*

associations the strength of an edge is determined by the partial correlation $\varrho_{jj'|V\setminus\{j,j'\}} = -\kappa_{jj'}/\sqrt{\kappa_{jj}\kappa_{j'j'}}$ where $\kappa_{jj'}$ is an element of $\hat{\boldsymbol{K}}$, and $V\setminus\{j,j'\}$ refers to the set of variables excluding j and j'.

Both types of undirected graphs are shown in Fig. 8 for the 3 lineage types. The network of conditional associations identifies direct interactions between variables, conditioned on all other variables, and, as expected, provides a sparser representation than does the network of marginal associations.

Note how a binary tree is revealed in the graph of partial correlations for the branching process (Fig. 8f). This is expected since the branching process was designed so that daughters were independent when conditioned on their common mother. In the network of partial correlations this assumption reveals itself as the lack of an edge between sisters. In contrast, in the partial correlation graphs for T-cell (Fig. 8d) and worm (Fig. 8e) lineages, sisters are often joined by edges.

We emphasise that the inferred undirected graph for all lineages has the *symmetry* structure of a binary tree but not necessarily its *sparsity* structure. We are thus able to examine how the inferred network of *statistical* relationships compares to the known network of *familial* relationships; though the latter is a binary tree, the former may not be.

5.2. *Natural Basis, Directed Graph.* One problem with representing each family member as a node is that the graph appears cluttered since that there are many edges and nodes with similar parameters. This problem gets exponentially worse with increasing generations. Such redundancies disappear when examining the tree over its natural, or symmetry-adapted, variables, where the indistinguishabilities have been removed.

Since the natural variables in each irreducible component are ordered by generation they can be represented by a directed graph, with each variable conditioned on the past (Wermuth, 1980; Kiiveri, Speed and Carlin, 1984; Pearl, 1988). Each irreducible component is a chain and the tree is represented by G independent chains.

The structural equation model, sometimes called a causal model, underlying each chain is a non-stationary time series given by the following system of equations:

$$z_j = \sum_{j'=\ell}^{j-1} \beta_{jj'} z_{j'} + \varepsilon_j, \quad \text{for } \ell \le j \le G \tag{36}$$

Note that each irreducible component is represented by its own system of equations but we avoid the superscripts ℓ to reduce index clutter. Here z_j is a natural variable corresponding to a generation j, $\beta_{jj'}$ is the regression

coefficient of generation j on j', and ε_j is an independent zero-bias random variable representing the noise originating at generation j. Defining a lower-triangular coefficient matrix $\boldsymbol{B} = (b_{jj'})$ gives the system of equations in matrix form:

$$\boldsymbol{B}\boldsymbol{z} = \boldsymbol{\varepsilon},$$

$$b_{jj'} = \begin{cases} 1, & \text{if } j = j' \\ 0, & \text{if } j - j' < 0 \text{ or } j - j' > \mathcal{M} \\ -\beta_{jj'}, & \text{otherwise.} \end{cases} \tag{37}$$

To find the parameters in the structural equation model, $\beta_{jj'}$ and $\mathbb{E}(\varepsilon_j^2)$, from the inferred $\hat{\boldsymbol{\Sigma}}_\Omega$ we use the modified Cholesky decomposition:

$$\boldsymbol{\Sigma}_\Omega = \boldsymbol{L}\boldsymbol{\Phi}\boldsymbol{L}' \tag{38}$$

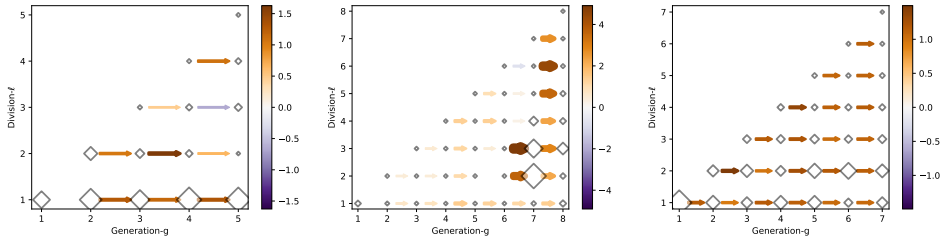
where $\boldsymbol{\Phi} = (\varphi_{jj'})$ is diagonal and $\boldsymbol{L}$ is lower triangular. Then since $\mathbb{E}(\boldsymbol{z}\boldsymbol{z}') = \boldsymbol{\Sigma}_\Omega$, we find that $\boldsymbol{L}^{-1} = (b_{jj'})$. Thus $\beta_{jj'}$ can be found from Eq. 37 while the noise terms are found directly from $\mathbb{E}(\varepsilon_j^2) = \varphi_{jj}$.

The directed graph can then be defined with edge weights given by $\beta_{jj'}$ and node strengths given by $\mathbb{E}(\varepsilon_j^2)$. The edges represent transmission of variation while the nodes represent innovations. If $|\beta_{jj'}| < 1$ then transmission is regressive, with descendants gradually losing memory of previous generations. However, if $|\beta_{jj'}| > 1$ then variation from that division observed at generation j' is *amplified* during transmission to generation j. Thus variation can arise directly from a noisy generation (node) or it can by amplified by strong transmission between generations (edge), or both.

The directed graphs for the 3 lineage types are shown in Fig. 9. It is striking how $\beta_{jj'}$ is particularly high for certain generations and divisions in the worm lineage. For the T cells division $\ell = 1$ has large innovations and high transmission between generations whereas the other divisions are fairly quiet. The branching process is largely featureless across all generations and divisions, as expected.

6. Explaining Fate. The directed graphs shown in Fig. 9 give a detailed summary of variation throughout lineage. We now examine what the results mean for our understanding of cell fate, addressing two questions in particular: How much each division contributes to cell fate, and how well an ancestor's phenotype predicts the fate of its descendants.

In this study, fate is defined to be the phenotype y observed at the latest generation modelled, G. This practical definition allows us to develop the



a) T cells b) Worm c) Branching process

FIG 9. *Directed graphs in the natural basis. These graphs show the dynamics of variation arising from each longitudinal division. The colour (and thickness) of an edge between node j and j' in division ℓ corresponds to the transmission strength, $\beta_{jj'}^{(\ell)}$. The size of the node corresponds to the innovation strength, $\mathbb{E}(\varepsilon_j^2)$.*

quantitative framework for understanding fate in terms of explained variance. Of course fate at generation G may not represent the ultimate fate of cells in a lineage, especially if G is an early generation. Furthermore, cell fate is often conceptualised in terms of cell types and thus represented by discrete states. However cell types are usually identified by the continuous expression levels of certain underlying phenotypes. It seems reasonable then to define fate directly in terms of the measured (continuous) traits rather than the derived (discrete) cell types.

With this definition, a statistical analysis of cell fate involves explaining the variability of the phenotype at generation G in terms of other variables. If a set of variables can be shown to account for most of the variability at generation G then we can say that those variables explain fate (in the statistical sense). First we will explain a cell's fate by the *hidden* contributions from each division; then we will explain it by the *observed* phenotypes of the cells' ancestors. The first we interpret as a measure of fate progression (or commitment) while the second we interpret as a measure of fate expression.

6.1. *Contributions to Fate.* Whether certain divisions are more important than others, and how committed cells are to particular fates at different stages in the lineage are questions of obvious scientific interest.

It is straightforward to estimate the contributions to cell fate at generation G from each division ℓ, where $1 \leq \ell \leq G$. This is just the standard problem of estimating the variance components for nested groups of family members in a single generation, a trivial exercise now that we have estimated the transformed covariance.

Consider the variance of a cell in generation G, represented by a diagonal

element in $\mathbf{\Sigma}_{\mathcal{G}}$ and, in the 3-index notation used in Eq. 4, denoted by σ_{GGG}. This can be written in terms of the elements $\xi^{(\ell)}_{gg'}$ of $\mathbf{\Sigma}_{\Omega}$ (from Eq. 22) by applying the similarity transformation $\mathbf{\Sigma}_{\mathcal{G}} = \boldsymbol{T}\mathbf{\Sigma}_{\Omega}\boldsymbol{T}^{\dagger}$ to $\mathbf{\Sigma}_{\Omega}$. Supplement S2 gives a specific example. The result shows that σ_{GGG} is just the sum of independent contributions from each division (ℓ, τ) :

$$\sigma_{GGG} = \frac{1}{N_{\text{div}}} \sum_{\ell=1}^{G} \sum_{\tau=1}^{d_\ell} \xi^{(\ell)}_{GG} = \frac{1}{N_{\text{div}}} \sum_{\ell=1}^{G} \xi^{(\ell)}_{GG} d_\ell, \tag{39}$$

where $N_{\text{div}} = \sum_{\ell=1}^{G} d_\ell = 2^G$ is the total number of divisions (which is equal to the number of members of generation G). As before, d_ℓ is the dimension of the ℓ-th irreducible eigenspace or, equivalently, the number of transverse divisions for the ℓ-th division (Section 4.2.5).

The components of variance, $\xi^{(\ell)}_{GG} d_\ell / N_{\text{div}}$, given in Eq. 39 are the normalised eigenvalues of a classical ANOVA (Speed, 1987). In a classical ANOVA for nested groups, the data are confined to the lowest group level. As mentioned in Section 4.2.5, the resulting covariance matrix is that for a single generation of the tree and is diagonalisable with the common variance satisfying Eq. 39.

The resulting proportion of variance, η^2, for a cell in generation G that is attributable to the ℓ-th division is:

$$\eta^2(G, \ell) = \frac{\xi^{(\ell)}_{GG} d_\ell}{\sum_{\ell'=1}^{G} \xi^{(\ell')}_{GG} d_{\ell'}}, \qquad \ell \leq G \tag{40}$$

It will also be useful to calculate the cumulative proportion of total variance attributable to divisions from 1 to ℓ, inclusive:

$$\eta^2_{\text{cml}}(G, \ell) = \frac{\sum_{\ell'=1}^{\ell} \xi^{(\ell')}_{GG} d_{\ell'}}{\sum_{\ell'=1}^{G} \xi^{(\ell')}_{GG} d_{\ell'}}, \qquad \ell \leq G \tag{41}$$

which is related to the intraclass correlation.

An obvious question is how the results would differ if we had simply performed an ANOVA on the single generation G, ignoring measurements in the other generations. With complete data, this would give the identical result to a classical ANOVA calculation: our approach using a decomposable model for a Markov chain ensures that estimates of diagonal elements in $\mathbf{\Sigma}_{\Omega}$ are given by their corresponding sufficient statistics. With incomplete data however, data from other generations provide a better estimate of missing data in generation G and thus improve the estimate of $\mathbf{\Sigma}_{\Omega}$.

6.2. *Predictability of Fate.* The question of how well a cell's fate can be predicted from the phenotypes of its ancestors is a measure of how much information about a given phenotypic fate is being expressed in earlier generations.

We define the predictability of fate to be the proportion of variance of a family member in generation G that can be explained by the phenotypes of its ancestors, where here the 'explaining' is by linear regression. Given a family member in generation G and its direct ancestor in generation g, the proportion of explained variance is just the squared correlation coefficient, or coefficient of determination,

$$R^2(G,g) = \frac{\sigma_{gG}^2}{\sigma_{gg}\sigma_{GG}} = \rho_{gG}^2, \quad g < G. \tag{42}$$

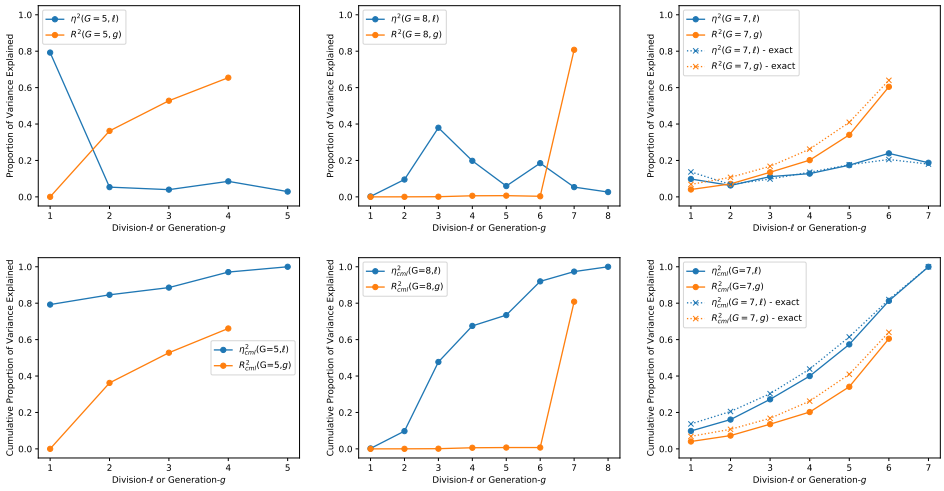
In the subscripts we have simplified the 3-index notation from Eq. 4 by ignoring the third index. This does not cause confusion since in this context we are only concerned with direct ancestors; the third index, associated with the MRCA, is thus redundant. For example, σ_{34} is understood to be the covariance between a mother in generation 3 and one of its daughters in generation 4, not any other daughter.

Generalising to prediction using multiple generations of direct ancestors up to and including that in generation g gives

$$R^2_{\text{cml}}(G,g) = \frac{\boldsymbol{\Sigma}_{G\boldsymbol{g}}\boldsymbol{\Sigma}_{\boldsymbol{gg}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{g}G}}{\sigma_{GG}} \tag{43}$$

where $\boldsymbol{g}$ represents a vector of direct ancestors of the cell in generation G that are from generations 1 to g inclusive. Note that Eq. 43 accounts for possible dependencies in the variation between ancestors. Unlike for the case of components of variance, variation between ancestral generations is not (in general) orthogonal.

6.3. *Comparing Explanations of Variance.* The two explanations of phenotypic variability are complementary: η^2 explains variance in terms of shared ancestral divisions, compares members within the same generation, and involves the transformed covariance matrix $\boldsymbol{\Sigma}_\Omega$; in contrast, R^2 explains variance in terms of ancestral generations, compares members across generations, and involves the covariance matrix $\boldsymbol{\Sigma}$. The two quantities are shown in Fig. 10 for the 3 lineage types with the top row giving the explained variance and the bottom row giving the cumulative explained variance. Note that because of the first order Markov process assumption (Section 4.3),



a) T cell b) Worm c) Branching process

FIG 10. *Explained variance (top row) and the cumulative explained variance (bottom row) for different lineages. η^2 (blue) measures the contribution to the variance in generation G by each ℓ-division. R^2 (orange) measures how well a generation-G cell's phenotype is predicted by its ancestor.*

$R^2 = R^2_{\text{cml}}$ in these plots. For the case of the simulated branching process the exact result is also shown.

$\eta^2(G,\ell)$ (blue line, top row) gives the contributions to fate at generation G from each of the earlier divisions ℓ. For the worm, $\ell = 3, 4, 6$ are particularly important divisions for explaining fate (at $G = 8$) while $\ell = 1$ is irrelevant. For the T cell $\ell = 1$ is by far the important division for explaining fate (at $G = 5$) while higher divisions are unimportant. As expected, for the branching process, contributions for all divisions are comparable.

$R^2(G,g)$ (orange line, top row) gives the predictability of fate at generation G using each of the earlier generations g. For the worm R^2 is zero until $g = 7$, when differentiation, and thus expression of fate, has started to occur and successive generations start to resemble each other. Strikingly then, none of the structure in η^2 from $1 \leq \ell \leq 6$ is reflected in R^2. For the T cell, even though most of the variation is explained by division $\ell = 1$, R^2 is low for $g = 1, 2$ indicating that the phenotypes of those ancestors contain little information about their descendants despite their fate having largely been set. For both these lineages then, cell fate is being determined in early generations but is not being expressed until later.

Such 'hidden' fate progression is best visualised in the cumulative ex-

plained variance shown in the bottom row of Fig. 10. For the worm, η^2_{cml} increases unevenly with each division while $R^2_{\text{cml}}(G, g)$ remains zero. For the T cell, even though η^2_{cml} starts high at $\ell = 1$, $R^2_{\text{cml}}(G, g)$ starts at zero. Contrast both these lineages with the branching process where η^2_{cml} and R^2_{cml} both start near zero and increase steadily in a similar fashion. This would be expected since all the variation in the branching process is expressed. Clearly a T cell lineage cannot be modelled by a branching process.

Based on these observations we interpret η^2_{cml} as a measure of fate progression (or commitment) and R^2_{cml} as a measure of fate expression. In the two real lineages fate progression is always higher than fate expression, with the difference representing hidden fate progression. This reflects cells committing to a particular fate before they express individual markers of that fate.

7. Discussion. We have developed a general method for inferring the structure of lineage heterogeneity. Symmetry invariance, invoked to constrain the joint probability distribution, identifies a set of natural variables which compactly describe tree-structured variation. To apply the method to real measurements we employed a Markovian constraint and used the EM algorithm to account for missing data.

The inferred parameters were interpreted in two ways. First, directed graphs over the irreducible components gave a fine-grained, causal view of all variation throughout the lineage. Second, variation in a late generation was explained in terms of earlier divisions and generations, allowing the progression and expression of cell fate to be inferred. Comparing the two methods distills how much of the fine-grained behaviour seen in the directed graphs actually affects later generations: ancestral variation only influences fate if it is effectively transmitted through intermediate generations. This makes it possible to distinguish between variation that influences fate and variation that is inconsequential noise.

Examining the differences between the worm and branching process highlights how regularities in noisy lineage patterns can be distinguished from noise. The sharp features, analogous to spectral lines, seen in the 'fate spectrum' of the worm, Fig. 10b, are associated with asymmetric divisions that give rise to well-known differentiation structure. Supplement S3 shows a detailed worm lineage with each family member annotated by its standard label, highlighting the specific divisions contributing to the spectral features. In contrast the branching process seen in Fig. 10c has a comparatively featureless structure, analogous to a white noise spectrum. The T cell lineage in Fig. 10a lies somewhere in between, showing a single peak at $\ell = 1$, but

no other structure.

We emphasise that lineage patterns do not have to be invariant (as in *C. elegans*) in order to be detected. In mammalian lineages, where cell differentiation is more likely to occur over clusters of closely-related divisions and generations rather than at specific ones, peaks in η^2 would be spread over neighbouring ℓ. This would be analogous to the broadening of spectral lines. We note that the concept of a fate map has largely been a deterministic one, describing at what stage a particular fate is specified (Chisholm, 2001). This method can thus be regarded as the first steps towards a probabilistic representation of the fate map.

The development of new lineage measurement techniques (Amat et al., 2014; Frieda et al., 2016) has been increasing the need for statistical methods to analyse lineage variation. Moreover, since the method can be viewed as a generalisation of ANOVA or multilevel modelling, it may find similarly broad use. The ubiquity of binary trees in both natural and human-designed systems suggest that potential applications exist in areas beyond cell lineages.

SUPPLEMENTARY MATERIAL

Supplement to 'Statistical Inference in Cell Lineage Trees': (sup1.pdf). Additional information is provided in a separate file. This covers the derivation of the group representation for a binary tree (Supplement S1), explicit expressions for the decomposition of a tree-structured covariance matrix (Supplement S2), and a figure showing *C. elegans* lineage with standard notation (Supplement S3).

References.

Amat, F., Lemon, W., Mossing, D. P., McDole, K., Wan, Y., Branson, K., Myers, E. W. and Keller, P. J. (2014). Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nature Methods* **11** 951 EP -.

Anderson, T. W. (1973). Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure. *Ann. Statist.* **1** 135–141.

Balázsi, G., van Oudenaarden, A. and Collins, J. J. Cellular Decision Making and Biological Noise: From Microbes to Mammals. *Cell* **144** 910–925.

Chisholm, A. D. (2001). Cell Lineage. In *Encyclopedia of Genetics* (S. Brenner and J. H. Miller, eds.) 302 - 310. Academic Press, New York.

Cowan, R. and Staudte, R. (1986). The Bifurcating Autoregression Model in Cell Lineage Studies. *Biometrics* **42** 769-783.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39** 1-38.

Diaconis, P. (1988). *Group Representations in Probability and Statistics. Lecture notes-monograph series*. Institute of Mathematical Statistics.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer.

Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K.-H. K., Singer, Z. S., Budde, M. W., Elowitz, M. B. and Cai, L. (2016). Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541** 107 EP -.

Gehrmann, H. and Lauritzen, S. L. (2012). Estimation of means in graphical Gaussian models with symmetries. *Ann. Statist.* **40** 1061–1073.

Gerlach, C., Rohr, J. C., Perié, L., van Rooij, N., van Heijst, J. W. J., Velds, A., Urbanus, J., Naik, S. H., Jacobs, H., Beltman, J. B., de Boer, R. J. and Schumacher, T. N. M. (2013). Heterogeneous Differentiation Patterns of Individual CD8+ T Cells. *Science* **340** 635–639.

Haccou, P., Jagers, P., Vatutin, V. A. and for Applied Systems Analysis, I. I. (2005). *Branching Processes: Variation, Growth, and Extinction of Populations. Cambridge Studies in Adaptive Dynamics*. Cambridge University Press.

Hadjantonakis, A.-K. and Arias, A. M. (2016). Single-Cell Approaches: Pandora's Box of Developmental Mechanisms. *Developmental Cell* **38** 574 - 578.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hawkins, E. D., Turner, M. L., Dowling, M. R., van Gend, C. and Hodgkin, P. D. (2007). A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proceedings of the National Academy of Sciences* **104** 5032-5037.

Hojsgaard, S. and Lauritzen, S. L. (2008). Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 1005–1027.

Hormoz, S., Singer, Z. S., Linton, J. M., Antebi, Y. E., Shraiman, B. I. and Elowitz, M. B. (2016). Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell Systems* **3** 419 - 433.e8.

Huggins, R. M. and Staudte, R. G. (1994). Variance Components Models for Dependent Cell Populations. *Journal of the American Statistical Association* **89** 19-29.

Kiiveri, H., Speed, T. P. and Carlin, J. B. (1984). Recursive causal models. *Journal of the Australian Mathematical Society. Series A. Pure Mathematics and Statistics* **36** 30–52.

Lauritzen, S. L. (1996). *Graphical Models. Oxford Statistical Science Series*. Clarendon Press.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics*. Wiley.

Oliaro, J., Van Ham, V., Sacirbegovic, F., Pasam, A., Bomzon, Z., Pham, K., Ludford-Menting, M. J., Waterhouse, N. J., Bots, M., Hawkins, E. D., Watt, S. V., Cluse, L. A., Clarke, C. J. P., Izon, D. J., Chang, J. T., Thompson, N., Gu, M., Johnstone, R. W., Smyth, M. J., Humbert, P. O., Reiner, S. L. and Russell, S. M. (2010). Asymmetric Cell Division of T Cells upon Antigen Presentation Uses Multiple Conserved Mechanisms. *The Journal of Immunology* **185** 367–375.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann series in representation and reasoning*. Morgan Kaufmann Publishers.

Sandler, O., Mizrahi, S. P., Weiss, N., Agam, O., Simon, I. and Balaban, N. Q. (2015). Lineage correlations of single cell division time as a probe of cell-cycle dynamics. *Nature* **519** 468–471.

Santella, A., Kovacevic, I., Herndon, L. A., Hall, D. H., Du, Z. and Bao, Z. (2016). Digital development: a database of cell lineage differentiation in C. elegans with lineage phenotypes, cell-specific gene functions and a multiscale model. *Nucleic Acids*

Research **44** D781-D785.

Serre, J. P. (1977). *Linear Representations of Finite Groups. Graduate Texts in Mathematics*. Springer New York.

Shah, P. and Chandrasekaran, V. (2012). Group symmetry and covariance regularization. *Electron. J. Statist.* **6** 1600–1640.

Shimoni, R., Pham, K., Yassin, M., Gu, M. and Russell, S. M. (2013). TACTICS, an interactive platform for customized high-content bioimaging analysis. *Bioinformatics* **29** 817-818.

Speed, T. P. (1987). What is an Analysis of Variance? *Ann. Statist.* **15** 885–910.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov Distributions over Finite Graphs. *Ann. Statist.* **14** 138–150.

Stiefel, E. and Fässler, A. (1992). *Group Theoretical Methods and Their Applications*. Birkhäuser Boston.

Strang, G. (1993). Wavelet transforms versus Fourier transforms. *Bull. Amer. Math. Soc.* **28** 288-305.

Sulston, J. E., Schierenberg, E., White, J. G. and Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology* **100** 64 - 119.

Szatrowski, T. H. (1980). Necessary and Sufficient Conditions for Explicit Solutions in the Multivariate Normal Estimation Problem for Patterned Means and Covariances. *The Annals of Statistics* **8** 802-810.

Uhler, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *Ann. Statist.* **40** 238–261.

Wermuth, N. (1980). Linear Recursive Equations, Covariance Selection, and Path Analysis. *Journal of the American Statistical Association* **75** 963-972.

Zilman, A., Ganusov, V. V. and Perelson, A. S. (2010). Stochastic Models of Lymphocyte Proliferation and Death. *PLoS ONE* **5** e12775.

D. G. Hicks,
Centre for Micro-Photonics and
Dept. of Physics and Astronomy,
Swinburne University of Technology,
Hawthorn, VIC 3122, Australia
E-mail: dghicks@swin.edu.au

T. P. Speed,
Bioinformatics Division,
Walter & Eliza Hall Institute
of Medical Research,
Parkville, VIC 3052, Australia
E-mail: terry@wehi.edu.au
URL: https://www.wehi.edu.au/people/terry-speed
and
School of Mathematics & Statistics,
University of Melbourne,
Parkville, VIC 3050, Australia

M. Yassin
Peter MacCallum Cancer Centre,
Parkville, VIC 3052, Australia
and
Department of Pathology
University of Melbourne,
Parkville, VIC 3050, Australia
E-mail: Mohammed.Yassin@petermac.org

S. M. Russell,
Peter MacCallum Cancer Centre,
Parkville, VIC 3052, Australia
and
Centre for Micro-Photonics,
Swinburne University of Technology,
Hawthorn, VIC 3122, Australia
and
Department of Pathology and
Sir Peter MacCallum Department of Oncology,
University of Melbourne,
Parkville, VIC 3050, Australia
E-mail: sarah.russell@petermac.org