Single-cell lineage tracing is fundamentally different from inference of traditional lineage trees

TracerSeq is SO COOL!

Let me back up for a second. In the Maehr lab, we use single-cell RNA-seq to study development. We try to keep up with what's going on in the field, so I recently hosted a discussion in the lab on two blockbuster papers that had sequenced the whole zebrafish embryo at multiple timepoints.

Between those two papers, the zebrafish lineage tree is reconstructed using three different datasets and at least four algorithms. The results differ in substantial ways, and that's a little worrisome, so we wanted to understand why they differ. This blog post digs into details of one technology behind these papers. It's called TracerSeq.

What is TracerSeq?

TracerSeq is a scalable lineage tracing method based on Tol2, which is a naturally active transposon from a kind of fish called a medaka fish. (Transposons or "jumping genes" are genes that make proteins that then go back to the genes and edit, move, or copy them.) Tol2 is a small stretch of DNA -- 4.7kb -- and it's really cool because it is modular.

- It's got sequences that encode a transposase protein. Edit these out and there's no way for it to do its thing -- unless you provide transposase separately.
- It's got little flanking sequences (150 to 200 bases) that allow transposase to recognize it and move it around. Leave these out and it's just another more-or-less stationary gene.
- It's got enough space in the middle to carry up to 11kb of foreign DNA. Wagner et al. inserted a 20bp random barcode (for lineage tracing) and a beta-actin promoter (so that the barcode gets transcribed into RNA and sequenced).

When the embryo is still a single cell, Wagner et al. inject a boatload of these randomly barcoded Tol2's. (I have no clue how they physically do this. Probably magic.) Free-floating barcodes get divvied up by mitosis, so every time the cell divides, the daughter cells contain a different mix of barcodes. The barcodes also gradually get patched into the genome due to the nature of Tol2. Once a given Tol2 molecule integrates into the genome, the barcode is copied at every cell division, so all the daughter cells share that barcode. You can use shared barcodes to infer shared ancestry between cells. Duuuuuuude.

What do the results look like?

Wagner et al. sequenced TracerSeq embryos after 24 hours. On a human scale, that's sort of like 4 weeks. (For developmental biologists, it's at ~30 somites.) The zebrafish has a wide variety of specialized tissues -- brain, spinal cord, gut tube, tail bud, epidermis -- and the little fishy heart is just starting to beat. In the first part of their paper, Wagner et al. distinguish these specialized tissues using plain scRNA-seq (no TracerSeq). So, does TracerSeq tell us the same story about how the tissues arise?

No. TracerSeq results actually have little to do with tissue type. From the paper:

> "Tracer-seq lineage groups tend to be organized by position (e.g., along the [anterior-posterior] axis) rather than strictly by germ layer/tissue origin (e.g., neural, epidermal, mesodermal)."

This was a big surprise to me. I've never taken a dev bio class, but since joining the Maehr lab, I have absorbed some of the field's orthodoxy. I thought that the first major split in the lineage tree of any vertebrate would be between the three *germ layers*: ectoderm gives rise to skin and nerves; mesoderm gives rise to bones, muscles, and the heart; and endoderm gives rise to internal organs such as the stomach, liver, and pancreas. The TracerSeq results produce a completely different partitioning of the organism!

Why?

Here's the only plausible explanation I can come up with. It rests on two theses.

- Claim 1: There are different ideas of the concept of a lineage tree: clonal and state-based. Single-cell RNA-seq measures state and TracerSeq measures clonality.
- Claim 2: TracerSeq places heavy emphasis on what happens very early in development, prior to the separation of endoderm, mesoderm, and ectoderm. This emphasizes differences between clonal trees and state trees.

These two claims, in combination, can explain the divergence in results between the two techniques.

Claim 1: clonal lineage trees versus Waddington lineage trees

In the paper, Wagner et al. ask "how clonal relationships compared with cell state relationships." When I read the paper, I thought these were two different methods of measuring the same lineage tree. Below, I give a formal definition of clonal relationships and state relationships. My current understanding is that these trees are distinct but related.

Biologists often think about development using a "landscape" metaphor popularized by a man named Waddington. In this metaphor, the landscape corresponds to all the possible states of a cell. In the world of single-cell RNA seq, people might think of it as the set of all possible gene expression patterns (even though this leave out important aspects of cell state -- all the dynamic aspects of proteins, DNA, and subcellular spatial organization). Position on the landscape represents the state of a cell. (Many drawings and discussions of the

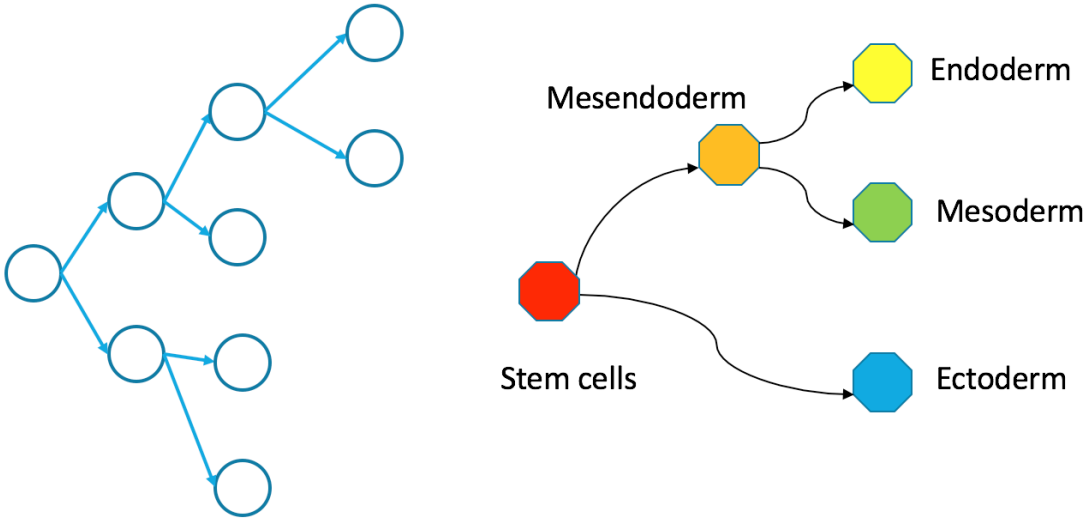
Waddington landscape reserve one dimension -- usually the vertical axis -- to measure "developmental potential", but that complicates things immensely and it is not useful in this discussion.)

For this discussion, I'll define a Waddington lineage tree as the subset of the Waddington landscape that is occupied during natural development, and I will assume it can be discretized into a finite number of states. A typical cell might develop by traversing a path within the Waddington tree. For this discussion, I will assume a Waddington lineage tree has no loops, but my claims can be extended to the case where it is a directed graph with no cycles.

The "clonal lineage tree" is much simpler to explain. It's a binary tree. The root is the zygote (fertilized egg). The children (in the graph-theory sense) are the two daughter cells after this zygote divides. So on and so forth.

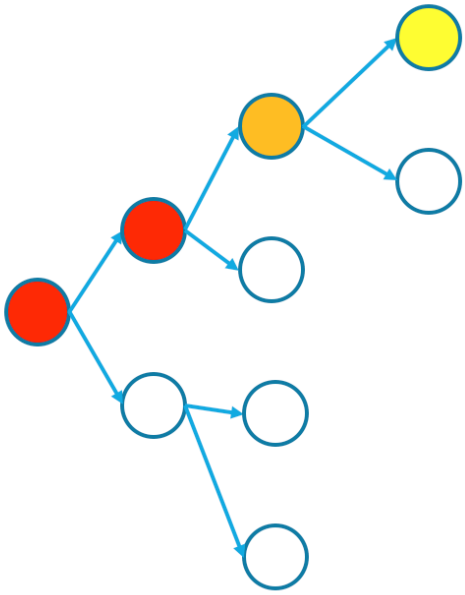
In summary:

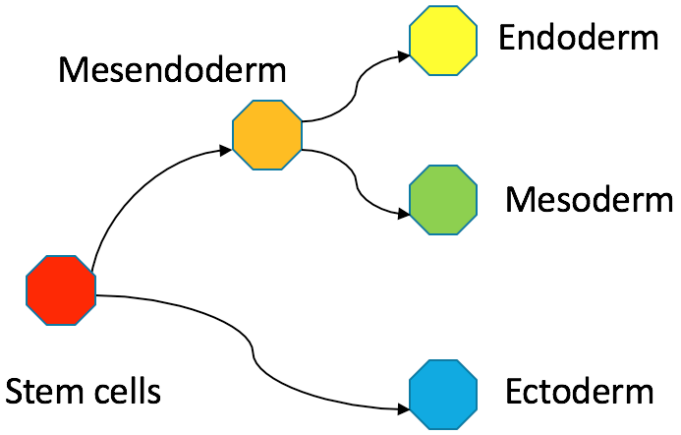
- For clonal trees, nodes are actual cells. Edges are literal mother-daughter relationships.
- For Waddington trees, nodes are regions within transcriptome space, not actual cells. Edges are pairs of abutting regions.



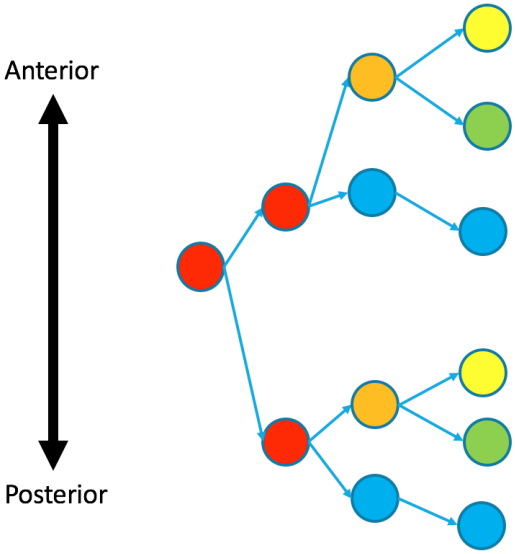
How do clonal and lineage trees relate?

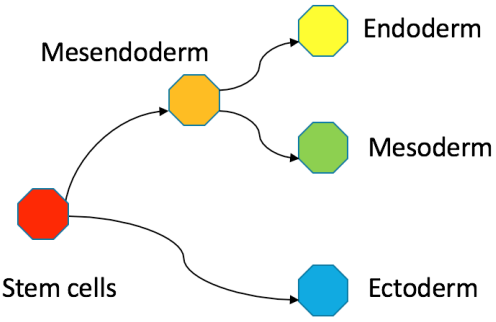
If you trace a descending path through a clonal lineage tree, it will form a continuous path through the corresponding Waddington tree. That's true because the Waddington tree includes any state occupied during natural development, and because natural development is effectively continuous.





Even though individual cells follow contiguous paths through each tree, the two types of trees are not necessarily superimposable. In an appendix, I define precisely what I mean by superimposability and provide a counterexample. For now, just take a look at my counterexample.





Correspondence between nodes is given by color, with blue mapping to blue, et cetera. The ectodermal subtree of the Waddington tree (the blue node on the bottom right) does not have a single origin in the clonal graph. It contains two separate subtrees of the clonal tree. Biologically, this means that the germ layers have separated independently in the anterior and posterior parts of the embryo, so the clonal tree does not look like the Waddington tree.

I want to emphasize that these trees are still biologically compatible. Cells may indeed vary independently in

terms of both transcriptional identity and clonal origin. An extreme example of this would be to take the two subtrees in the diagram not as anterior and posterior parts of the same embryo, but as monozygotic twins. Their shared clonal tree will have two subtrees that are heterogeneous, each containing all three germ layers. This should not be taken as evidence against the orthodox view that differentiation begins with the separation of germ layers.

(Do zebrafish have twins? I can't imagine trying to get their names right.)

Claim 2: TracerSeq's pre-gastrulation emphasis

In the first section of this discussion, I set out the difference between Waddington and clonal lineage trees. Here, I discuss how this applies to practical analysis of TracerSeq data. In particular, I will argue that TracerSeq places a major emphasis on extremely early development (prior to separation of germ layers). Thus, it is especially able to infer aspects of the clonal tree that cannot be superimposed onto a typical zebrafish Waddington tree.

Re-reading the paper, this is not a surprise. In the authors' words,

> "... [T]he current timing of TracerSeq integrations encompasses the transition from unrestricted pluripotency to the first fate restriction events appearing in the zebrafish embryo."

But, I hope I can provide some additional food for thought. The key to my argument is that per-cell Tol2 counts will be cut in half with every cell division. Thus, they will definitely decrease dramatically during development (due to degradation and DNA doubling). By gastrulation, there will be thousands of cells and thus thousands of times fewer integration events per minute for each copy of the genome.

To make this precise, let me consider an oversimplified model of the situation using the following assumptions and notation.

- there are initially N free copies of Tol2
- in each cell, there are r integrations per Tol2 before the next division (with no randomness).
- N is extremely large and r is small, so that losses due to integration are negligible.
- the rate of degradation of Tol2 is negligible.
- mitosis divides the available Tol2 evenly.
- all cells divide simultaneously at even intervals.

After N divisions, a typical cell will contain rN integrations from the single-cell stage, $\frac{rN}{2}$ integrations from the two-cell stage, and so on up to $\frac{rN}{2^N}$ divisions from the most recent stage. I am no expert in zebrafish embryo stages, but from a layperson's reading, the germ layers separate when N is greater than 12. Thus, a little calculation indicates that in any given cell, at any stage of development, *more than 99.97% of observed barcodes will have integrated prior to separation of the germ layers*. In other words, 99.97% of the barcodes

found in each cell will encode information that has little to do with the Waddington tree for zebrafish development.

The quantitative details above are definitely wrong. Later cell divisions ought to happen much more slowly than earlier ones, increasing the amount of time that Tol2 is able to integrate into the genome. I dimly remember that early cell divisions result in a finer partitioning of the human embryo with little increase in volume. If this is true in zebrafish, it could effectively increase the concentration of Tol2 barcodes around each genome, thus increasing the rate of integration at that stage of development. (This makes intuitive sense if we consider only duplication of the genome without the rest of mitosis: the reaction rate would be proportional to $X_T X_G$ where X_T is the concentration of Tol2 and X_G is the concentration of genomic DNA. The reaction rate goes up every time X_G doubles.)

I downloaded the Tracer-seq datasets, and I was planning to back this up with some analysis. Do most barcodes in a given cell actally fall towards the start of the clonal tree? But, the barcode detection is very sparse: of the 1113 tol2 barcodes in clone 1, most appear in 0 cells or 1, and of the 5753 cells, most have no Tol2 barcode or only one. This makes it impossible to infer anything like a complete clonal tree. The analyses that would make the most sense are (no surprise) already in the paper: highlighting of partial clonal trees on the Waddington graph (figure 4) and assessment of clonal coupling across germ layers groups (figure 5).

Appendices

Graph theory terms

- A *directed graph* G is a finite set of nodes $N(G)$ and a set of edges $E(G)$. The edges consist of ordered pairs of nodes; each edge connects two nodes. Not all pairs appear; in fact, the edge set may be almost or completely empty.
- A graph is a *tree* if it is connected and has no cycles. A lack of cycles means for any set of nodes $n_1, n_2, \ldots n_C$, if the edge set $E(G)$ contains $(n_1, n_2), (n_2, n_3), \ldots, (n_{C-1}, n_C)$, then it must lack (n_C, n_1). "Connected" means what you think it means: you can move from any node to any other node by following edges.
- A subgraph G' is a graph whose nodes are a subset of $N(G)$ and whose edges are a subset of $E(G)$ containing only edges with both ends in $N(G')$. Formally, an edge (n_1, n_2) belongs to $E(G')$ if and only if n_1 and n_2 are both in $N(G')$.
- A subtree is a subgraph of a tree that is also a tree.

Superimposing lineage trees

For a precise statement of my argument, I must define what it means to superimpose trees. For readers not trained in graph theory, I will make use of only a few simple concepts. I hope you will be able to read straight through with no problems, but if not, there is an appendix with definitions.

Let a Waddington tree W be represented as a mathematical graph with nodes consisting of lists of numbers (formally, $N(W) \in \mathbb{R}^n$). Suppose for the sake of this discussion that the Waddington tree has no cycles.

Let a clonal tree C be represented as a graph with integers for nodes (formally, $N(C)$ in $\mathbb{Z}$). Edges $E(C)$ are ordered pairs of mother and daughter cells.

The clonal tree C is defined to be *superimposable onto W* if each cell state in W arises only once in C. Formally, if S is a subtree of $N(W)$, then the corresponding nodes in S must also be a subtree.

In the illustrated counterexample, correspondence between nodes is given by color, with blue mapping to blue, et cetera. The ectodermal subtree of the Waddington tree (the blue node on the bottom right) does not have a single origin in the clonal graph. It contains two separate subtrees of the clonal tree. Biologically, this means that the germ layers have separated independently in the anterior and posterior parts of the embryo, so the clonal tree does not look like the Waddington tree.