Pseudotime
and
Trajectory Inference

Stefania Giacomello

The basics

Cells display a continuous spectrum of states (i.e. activation and/or differentiation process)

Individual cells are executing through a gene expression program in an unsynchronized manner → each cell is a snapshot of the transcriptional program under study

sc-omics technologies allow to model biological systems

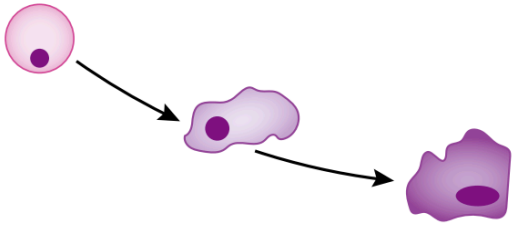
The basics

Discrete classification of cells is not appropriate

Summary of the continuity of cell states in the data
→ Trajectory Inference (TI) (or pseudotemporal ordering)

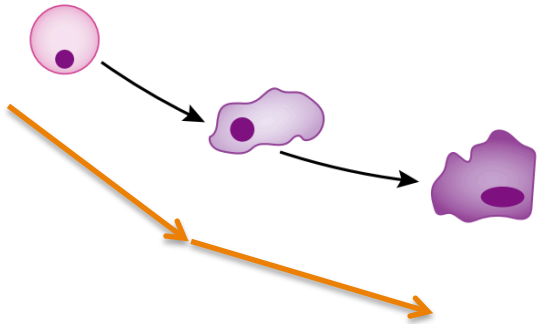
What is a trajectory?

Sequence of gene expression changes each cell must go through as part of a dynamic biological process



What is a trajectory?

Sequence of gene expression changes each cell must go through as part of a dynamic biological process

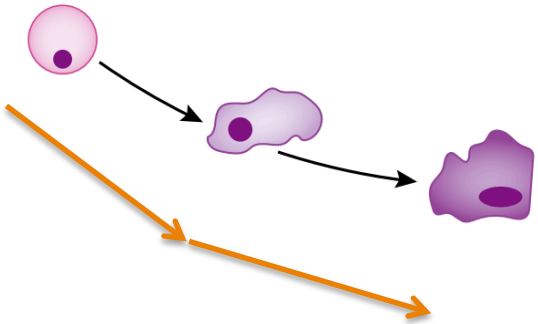


Track changes in gene expression:

- function of time
- function of progress along the trajectory

What is a trajectory?

Sequence of gene expression changes each cell must go through as part of a dynamic biological process



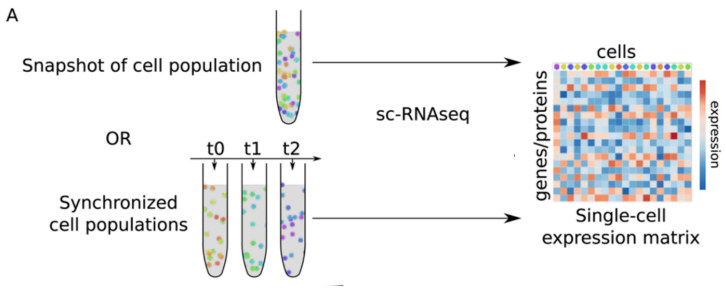
Track changes in gene expression:

- function of time
- function of progress along the trajectory

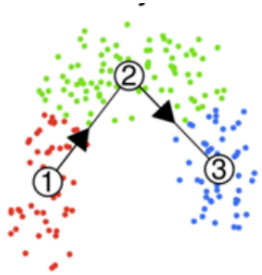
Pseudotime → abstract unit of progress:
distance between a cell and the start of the trajectory

How do TI tools work?

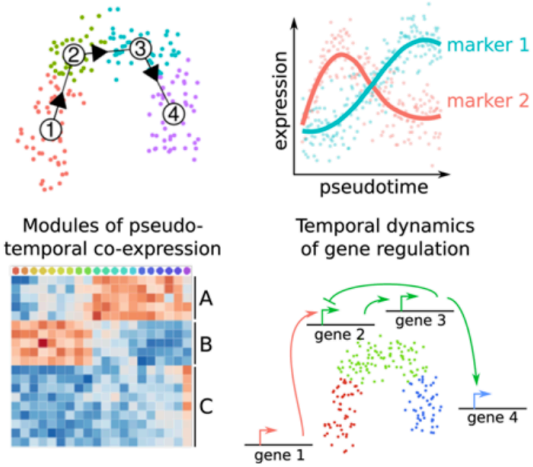
1. Population of single cells → different stages



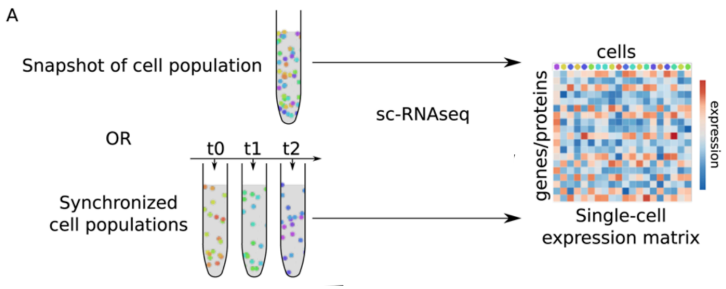
2. Computational tools to order cells along a trajectory topology
Automatic reconstruction of a cellular dynamic process by structuring individual cells sampled and profiled from that process



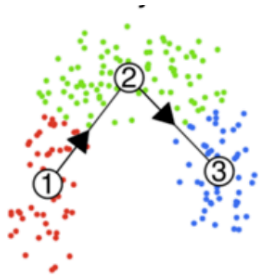
3. Identify the different stages in the dynamic process and their interrelationships



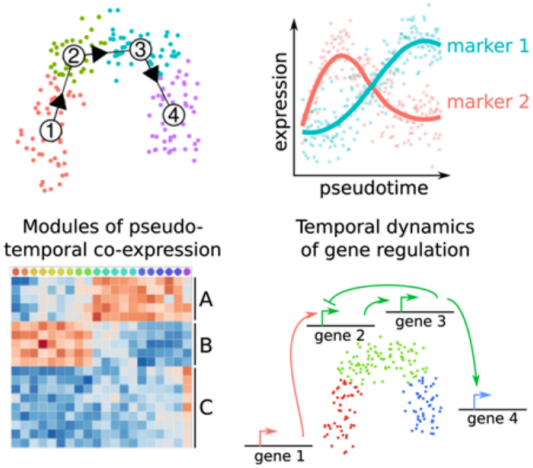
What TI offers



- Unbiased and transcriptome-wide understanding of a dynamic process

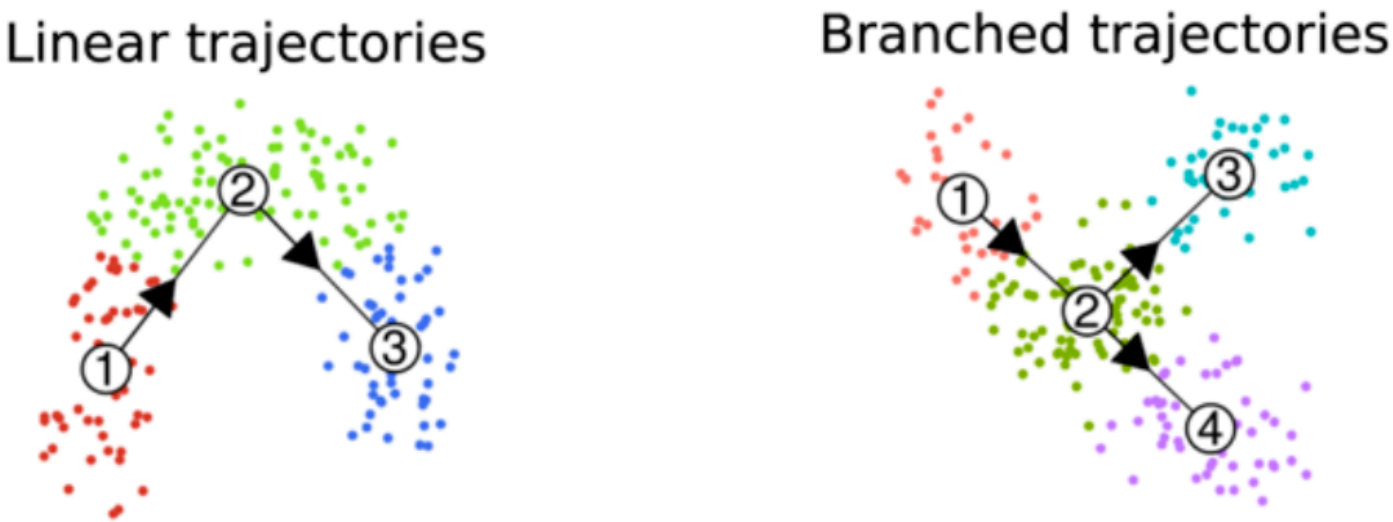


- They allow the objective identification of new subsets of cells



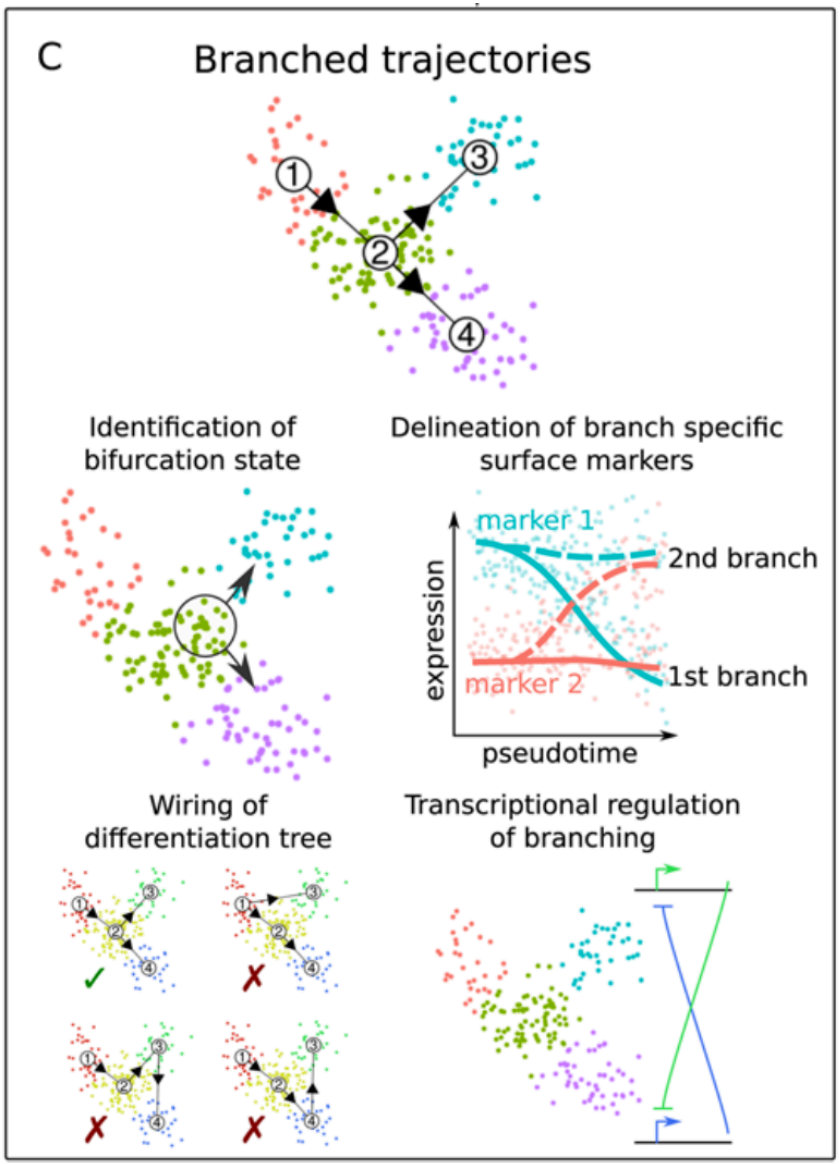
Type of trajectories

Trajectory's total length: total amount of transcriptional change that a cell undergoes at it moves from the starting to the end state



Linear, branched, or a more complex tree or graph structure

Type of trajectories



- Delineation of a differentiation tree
- Inference of regulatory interaction responsible for one or more bifurcations

Type of input data

- Transcriptome-wide data
- Starting cell from which the trajectory will originate
- Set of important marker genes, or even a grouping of cells into cell states.

Input data – potential risks

Providing prior information:

can help the method to find the correct trajectory among many, equally likely, alternatives

IF available, can bias the trajectory towards current knowledge

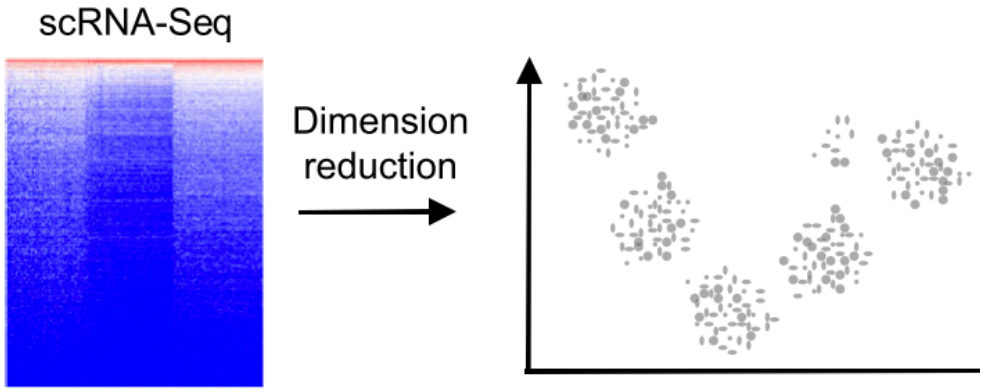
How TI tools usually work

1. conversion of data to a simplified representation using:
    - dimensionality reduction
    - clustering
    - graph building

2. ordering the cells along the simplified representation:
    - identify cell states
    - constructing a trajectory through the different states
    - projecting cells back to the trajectory

Dimensionality reduction step

Convert high-dimensional data to a more simplified representation, while maintaining the main characteristics of the data in the original space.



Dimensionality reduction step

Dimensionality reduction techniques:

- PCA (linear projection of the data such that the variance is preserved in the new space)
- independent component analysis (ICA)

- t-stochastic neighbor embedding (t-SNE)
- diffusion maps

} able to detect nonlinear relationships between cells

- Graph-based techniques

  cells = nodes in a graph

  edges =connect transcriptionally similar cells

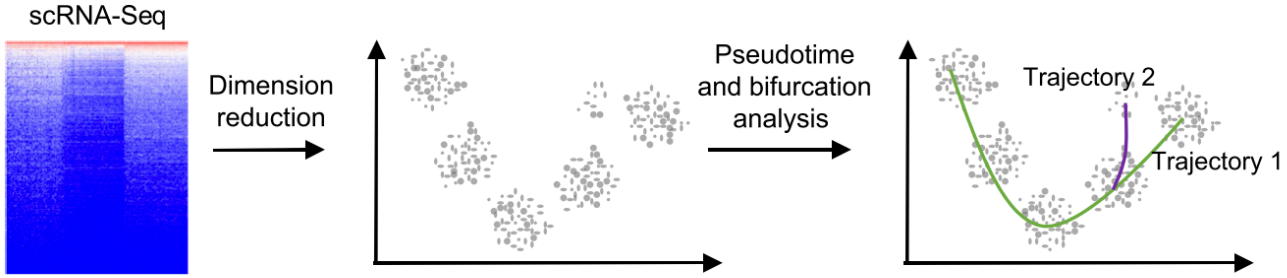
  It retains the most important edges in the graph → scales well to large numbers of cells (n > 10 000)

Trajectory modeling step

Many TI methods use graph-based techniques

1. simplified graph representation as input to find a path through a series of nodes (i.e. individual cells or groups of cells)

2. different path-finding algorithms are used by different algorithms



- "starting cell" by the user → representative for cells at the start of the process (e.g. the most immature cell in the case of a cell developmental process) used as a reference cell to compare all other cells against

- longest connected path in a sparsified graph → all cells are projected onto that path

Tools available

59 methods - unique combination of characteristics:

- required input
- methodology used
- produced outputs (topology fixing and trajectory type)

Method	Date	Most complex trajectory type	Fixes topology	Prior required	Prior optional	Evaluated	Reference
Monocle ICA	01/04/2014	Tree	Parameter	# branches	None	Yes	[13]
Wanderlust	24/04/2014	Linear	Fixed	Start cell(s)	None	Yes	[14]
SCUBA	30/12/2014	Tree	Free	None	Time course, Marker genes	Yes	[15]
Sincell	27/01/2015	Tree	Free	None	None	Yes	[16]
NBOR	08/06/2015	Linear	TBD	TBD	TBD	No[a i]	[6]
Waterfall	03/09/2015	Linear	Fixed	None	None	Yes	[17]
gpseudotime	15/09/2015	Linear	TBD	TBD	TBD	No[c]	[18]
Embeddr	18/09/2015	Linear	Fixed	None	None	Yes	[19]
ECLAIR	12/01/2016	Tree	TBD	TBD	TBD	No[f]	[20]
DPT	08/02/2016	Bifurcation	Fixed	None	Marker genes	Yes	[21]
Pseudogp	05/04/2016	Linear	Fixed	None	None	Yes	[22]
SLICER	09/04/2016	Graph	Free	Start cell(s)	End cell(s), Marker genes	Yes	[23]
SCell	19/04/2016	Linear	TBD	TBD	TBD	No[e]	[24]
Wishbone	02/05/2016	Bifurcation	Parameter	Start cell(s), # end states	Marker genes	Yes	[25]
TSCAN	13/05/2016	Tree	Free	None	None	Yes	[26]
SCOUP	08/06/2016	Multifurcation	Parameter	Start cell(s), Cell grouping, # end states	None	Yes	[27]
DeLorean	17/06/2016	Linear	TBD	TBD	TBD	No[g]	[28]
StemID	21/06/2016	Tree	Free	None	None	Yes	[29]
Ouija	23/06/2016	Linear	Fixed	Marker genes	None	Yes	[30]
Mpath	30/06/2016	Tree	Free	Cell grouping	None	Yes	[31]
cellTree	13/08/2016	Tree	Free	None	Cell grouping	Yes	[32]
WaveCrest	17/08/2016	Linear	TBD	Time course	None	No[f]	[33]
SCIMITAR	04/10/2016	Linear	Fixed	None	None	Yes	[34]
SCORPIUS	07/10/2016	Linear	Fixed	None	None	Yes	[35]
SCENT	30/10/2016	Linear	TBD	TBD	TBD	No[d]	[36]
k-branches	15/12/2016	Tree	TBD	TBD	TBD	No[h]	[37]
SLICE	19/12/2016	Tree	Free	None	Cell grouping, Marker genes	Yes	[38]
Topslam	13/02/2017	Linear	Fixed	Start cell(s)	None	Yes	[39]
Monocle DDRTree	21/02/2017	Tree	Free	None	# end states	Yes	[40]
Granatum	22/02/2017	Tree	TBD	TBD	TBD	No[e]	[41]
GPfates	03/03/2017	Multifurcation	Parameter	# end states	None	Yes	[42]
MFA	15/03/2017	Multifurcation	Parameter	# end states	None	Yes	[43]
PHATE	24/03/2017	Tree	TBD	TBD	TBD	No[h]	[44]
TASIC	04/04/2017	Tree	TBD	TBD	TBD	No[a e]	[45]
SOMSC	05/04/2017	Tree	TBD	TBD	TBD	No[a]	[46]
Slingshot	19/04/2017	Tree	Free	None	Start cell(s), End cell(s)	Yes	[47]
scTDA	01/05/2017	Linear	TBD	TBD	TBD	No[f]	[48]
UNCURL	31/05/2017	Linear	TBD	TBD	TBD	No[f]	[49]
reCAT	19/06/2017	Cycle	Fixed	None	None	Yes	[50]
FORKS	20/06/2017	Tree	TBD	Start cell(s)	None	No[f j]	[51]
MATCHER	24/06/2017	Linear	TBD	TBD	TBD	No[j]	[52]
PhenoPath	06/07/2017	Linear	Fixed	None	None	Yes	[53]
HopLand	12/07/2017	Linear	TBD	TBD	TBD	No[a j]	[54]
SoptSC	26/07/2017	Linear	TBD	Start cell(s)	None	No[a j]	[55]
PBA	30/07/2017	Multifurcation	TBD	TBD	TBD	No[j]	[56]
BGP	01/08/2017	Bifurcation	TBD	TBD	TBD	No[j]	[57]
scanpy	09/08/2017	Bifurcation	TBD	TBD	TBD	No[j]	[58]
B-RGPs	01/09/2017	Acyclic graph	TBD	TBD	TBD	No[j]	[59]
WADDINGTON-OT	27/09/2017	Graph	TBD	TBD	TBD	No[b j]	[60]
AGA	27/10/2017	Disconnected graph	TBD	TBD	TBD	No[j]	[61]
GPseudoRank	30/10/2017	Linear	TBD	TBD	TBD	No[a j]	[62]
p-Creode	15/11/2017	Tree	TBD	TBD	TBD	No[j]	[63]
iCpSc	30/11/2017	Linear	TBD	TBD	TBD	No[d j]	[64]
GrandPrix	03/12/2017	Multifurcation	TBD	Time course	None	No[j]	[65]
Topographer	21/01/2018	Tree	TBD	None	Start cell(s)	No[j]	[66]
CALISTA	31/01/2018	Graph	TBD	None	None	No[j]	[67]
scEpath	05/02/2018	Tree	TBD	TBD	TBD	No[a j]	[68]
MERLoT	08/02/2018	Tree	TBD	TBD	TBD	No[j]	[69]
ElPiGraph.R	04/03/2018	Graph	TBD	TBD	TBD	No[j]	

Topology of the trajectory:

- fixed by design

Early methods
Mainly focused on correctly ordering the cells along the fixed topology

- inferred computationally

Increased difficulty of the problem
Broadly applicable on more use cases
Topology inference still in the minority

Tool classification

TI methods classified also on a set of algorithmic components:

- Performance
- Scalability
- Output data structures

Monocle 2

Monocle introduced the concept of pseudotime

Now it has a complete new version - has been rated one of the most performing methods

Trajectory inference workflow:

1. Choosing genes to order the data

2. Reducing dimensionality of the data

3. Ordering cells in pseudotime

Trajectory inference workflow:

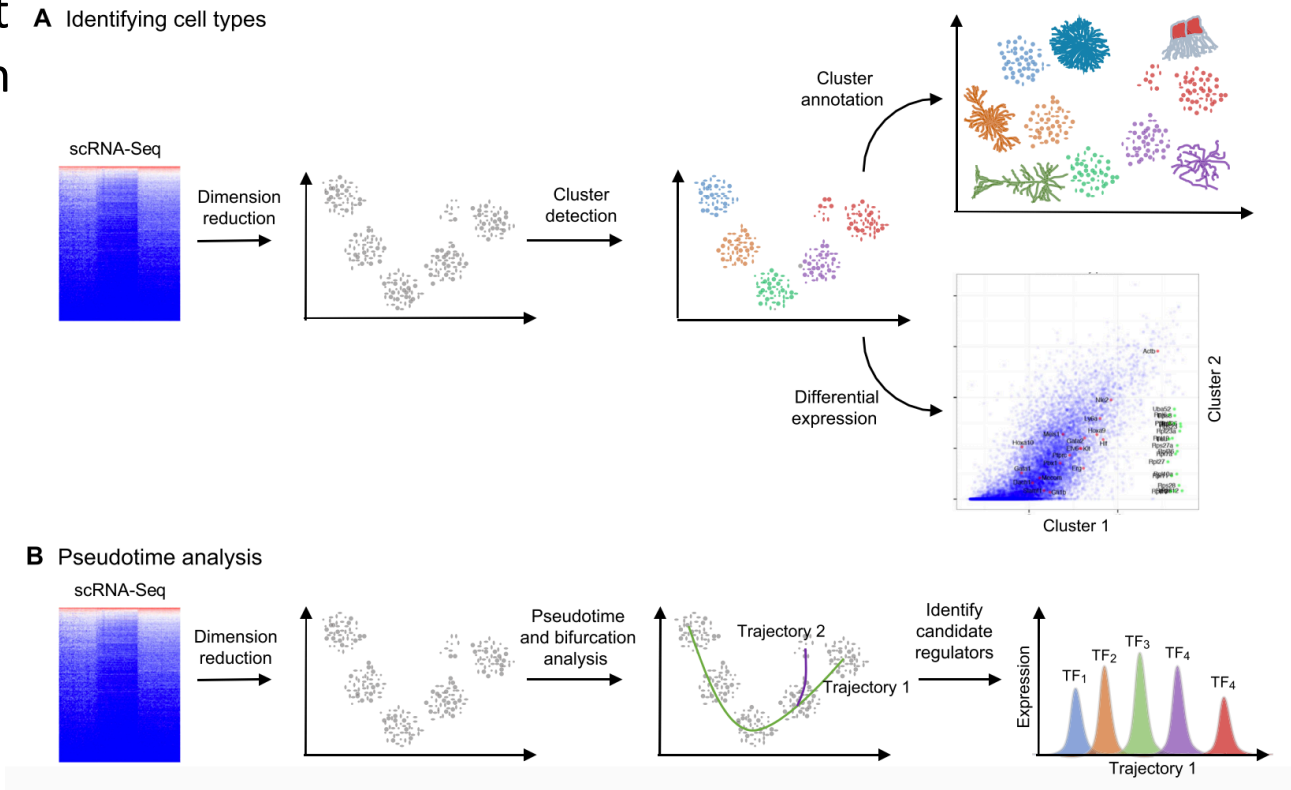
1. Choosing genes to order the data → look for genes that increase or decrease in expression during the functional process and use them to structure the data
    - unsupervised dpFeature → desirable approach to avoid biases
    - semi-supervised → genes that co-vary with marker genes
    - if we have time points → find differentially expressed genes between start and end
    - genes selected based on high dispersion among cells (gene's variance usually depends on its mean → careful how genes are selected based on variance, i.e. mean expression)

Monocle 2 – gene identification (dpFeature)

tSNE often groups cells into clusters that do not reflect their progression through the process

DE genes of cells in different clusters are informative markers of cell's progress in the trajectory

tSNE finds genes that vary over the trajectory but not the trajectory itself



Monocle 2 – gene identification (dpFeature)

1. Exclude genes expressed in very few cells (usually 5%)

2. PCA on remaining genes → components explaining variance in the data

3. Use identified PCs in tSNE

4. Apply density peak clustering to the 2D tSNE
   → takes into account cells density and distance to cells with higher density
   → density peaks = cells with high local density and far away from other high density cells
   → density peaks = clusters

5. Identify genes that differ between clusters

Trajectory inference workflow:

2. Reducing dimensionality of the data → Reversed Graph Embedding

3. Ordering cells in pseudotime → It assumes a tree structure with root and leaves and it fits the best tree to the data (manifold learning)

Monocle 2 – dimensionality reduction – learning the structure

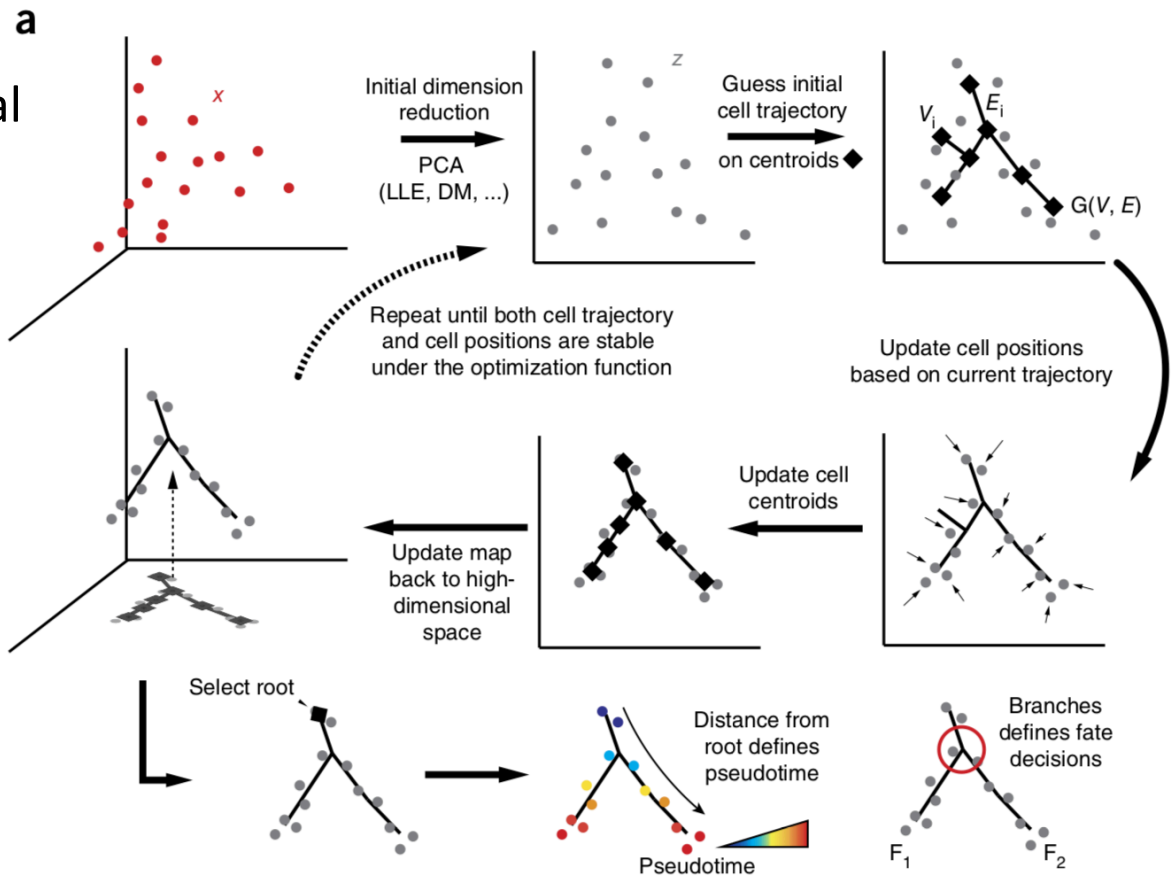
Monocle 2 uses reverse graph embedding to learn the data structure

It simultaneously:

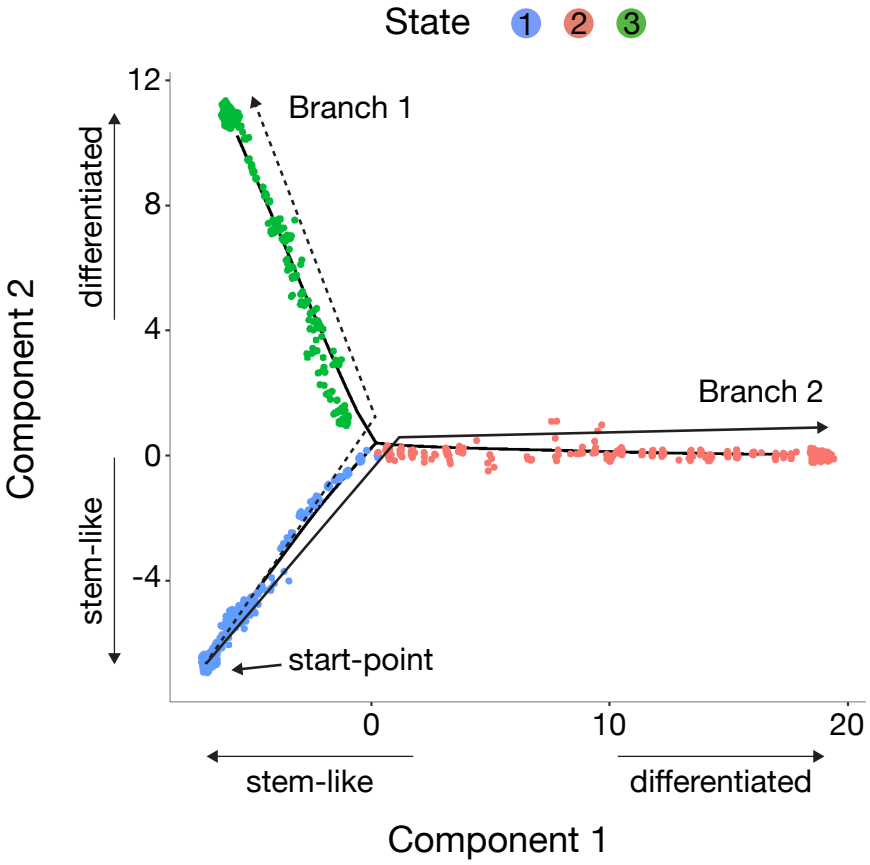
1. Reduces high-dimensional expression data into a lower dimensional space

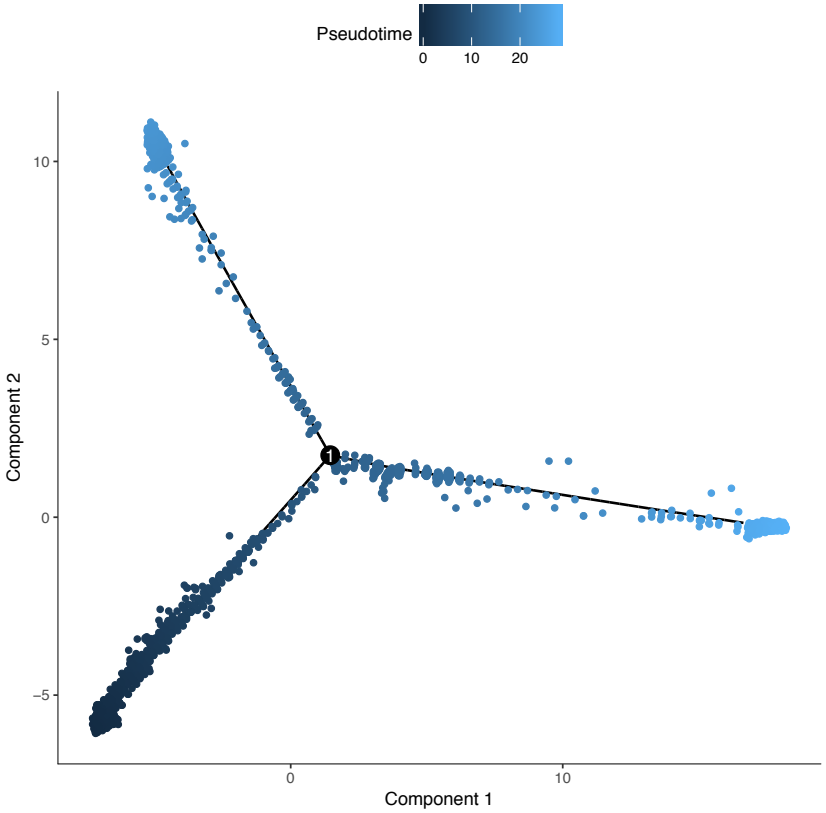
2. Learns a manifold that generates the data –
No a priori knowledge of the tree structure

3. Assigns each cell to its position on that manifold



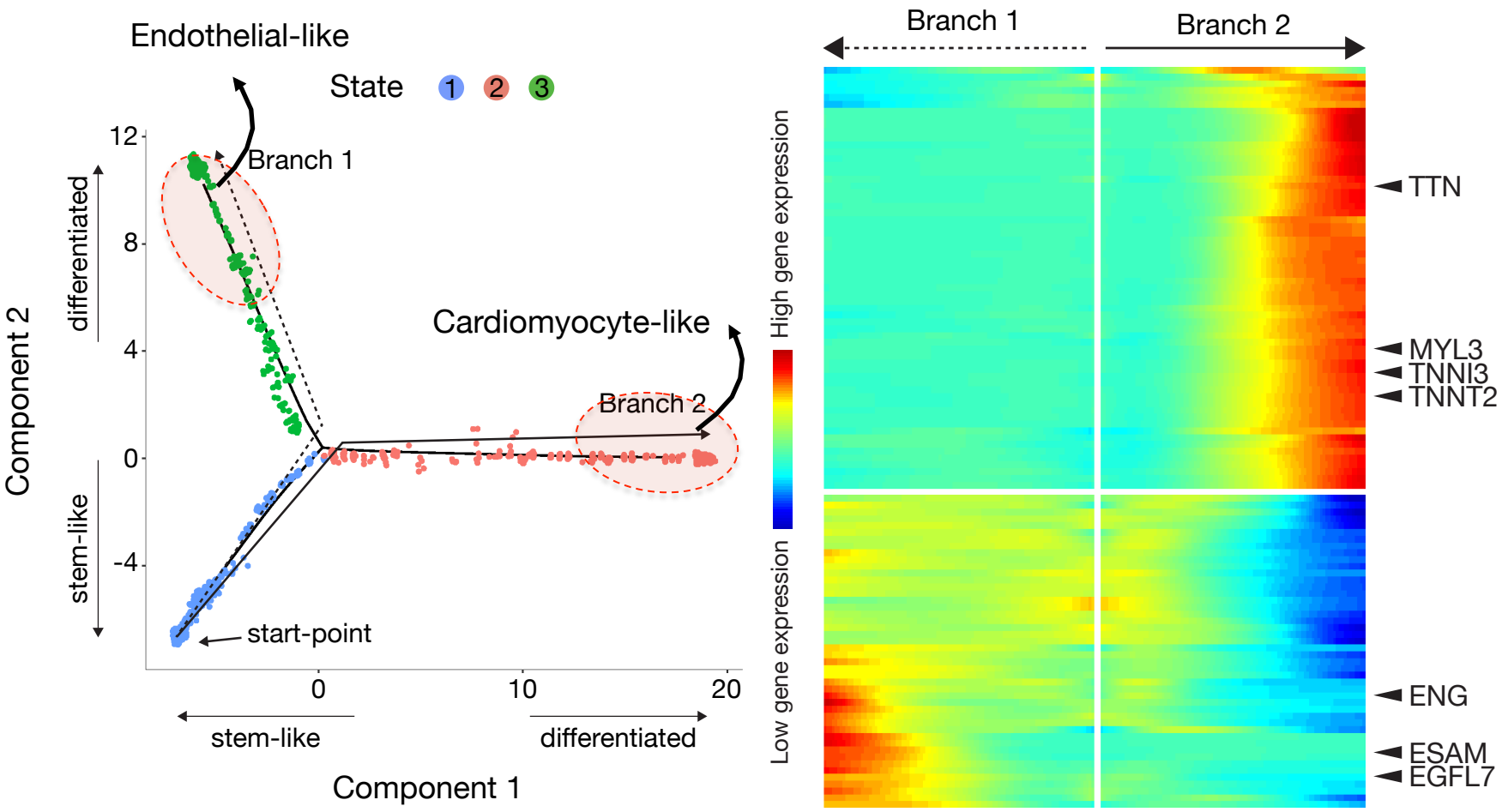
Fates of human fetal heart cells





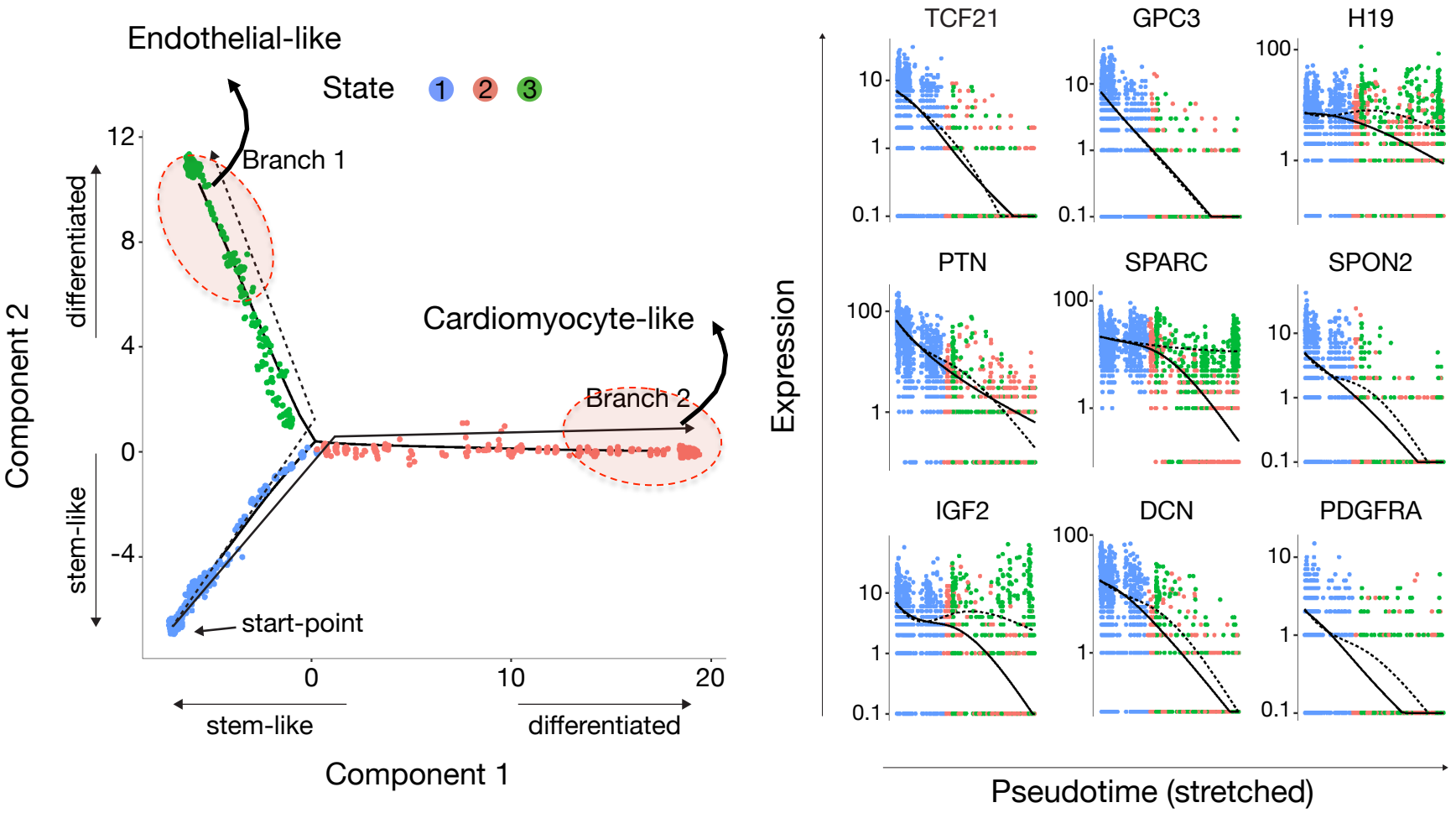
SciLifeLab

Fates of human fetal heart cells



SciLifeLab

Fates of human fetal heart cells



SciLifeLab