GENOME 569

Class 2: Vibecoding with GitHub Co-pilot

Science

RESEARCH ARTICLES

Cite as: J. S. Packer *et al.*, *Science*
10.1126/science.aax1971 (2019).

A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution

Jonathan S. Packer[1]*, Qin Zhu[2]*, Chau Huynh[1], Priya Sivaramakrishnan[3], Elicia Preston[3], Hannah Dueck[3]†, Derek Stefanik[4], Kai Tan[3,5,6,7], Cole Trapnell[1], Junhyong Kim[4]‡, Robert H. Waterston[1]‡, John I. Murray[3]‡

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Genomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, PA, USA. [3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA. [5]Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [6]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [7]Department of Cell and Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

*These authors contributed equally to this work.

†Present address: National Cancer Institute, Bethesda, MD, USA.

‡Corresponding author. Email: junhyong@sas.upenn.edu (J.K.); watersto@uw.edu (R.H.W.); jmurr@pennmedicine.upenn.edu (J.I.M.)

Caenorhabditis elegans **is an animal with few cells, but a striking diversity of cell types. Here, we characterize the molecular basis for their specification by profiling the transcriptomes of 86,024 single embryonic cells. We identify 502 terminal and pre-terminal cell types, mapping most single-cell transcriptomes to their exact position in *C. elegans*' invariant lineage. Using these annotations, we find that: 1) the correlation between a cell's lineage and its transcriptome increases from mid to late gastrulation, then falls dramatically as cells in the nervous system and pharynx adopt their terminal fates; 2) multilineage priming contributes to the differentiation of sister cells at dozens of lineage branches; and 3) most distinct lineages that produce the same anatomical cell type converge to a homogenous transcriptomic state.**

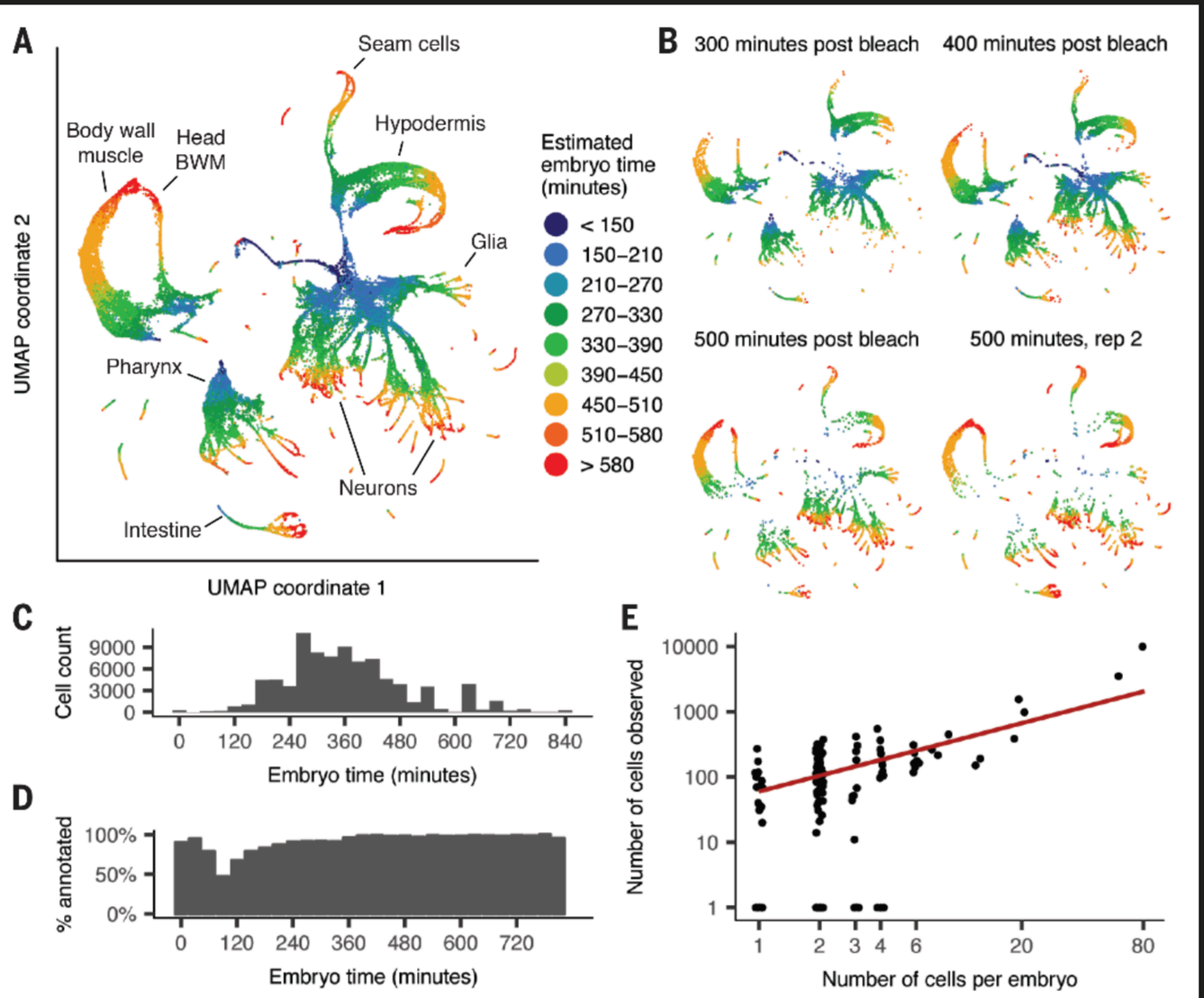

Fig. 1. UMAP projection shows tissues and developmental trajectories in *C. elegans* embryogenesis. (A) UMAP projection of the 81,286 cells from our sc-RNA-seq dataset that passed our initial QC. This UMAP does not include 4,738 additional cells that were initially

Downloading short read data

Where is the data?

ACKNOWLEDGMENTS

We thank members of the Murray, Waterston, and Kim labs, and Ben Lehner and Meera Sundaram for providing critical comments on the manuscript. We also thank A. Zacharias, D. Vafeados, M. Corson, R. Terrell, L. Gevirtzman, and P. Weisdepp for their contributions to the EPiC database. **Funding:** This work was funded by NIH grants U41HG007355 and R01GM072675 to RHW, and R35GM127093 and R21HD085201 to JIM. This work was also funded in part by Commonwealth of Pennsylvania Health Research Formula Funds and RM1HG010023 to JK, by U2C CA233285 to KT, by the William H. Gates Chair of Biomedical Sciences (RHW), and by the Allen Discovery Center for Lineage Tracing (JSP, CT). **Author contributions:** JP, CH, JK, RW, and JM conceived and designed the study; CH, PS, EP, HD, and DS performed the experiments; JP, QZ, RW, and JM did the analyses; CT, JK, RW, and JM supervised analyses; JK, JM, and KT supervised the development of VisCello; JP, QZ, JK, RW, and JM wrote the paper. **Competing interests:** The authors have no competing interests. **Data and materials availability:** The raw data have been deposited with the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE126954. Source code of VisCello (with *C. elegans* data) has been deposited at Github (https://github.com/qinzhu/VisCello.celegans) and Zenodo (*50*). Source code of VisCello for hosting other single cell data has been deposited at Github (https://github.com/qinzhu/VisCello) and Zenodo (*51*). Gene expression movies used in the annotation but not previously published have been deposited at Dryad (doi: 10.5061/dryad.7tg31p7) (*52*). This article was prepared while HD was employed by the University of Pennsylvania. The opinions expressed in this article are the authors' own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

Series GSE126954

Query DataSets for GSE126954

Status	Public on Mar 01, 2019
Title	A lineage-resolved molecular atlas of C. elegans embryogenesis at single cell resolution
Organism	Caenorhabditis elegans
Experiment type	Expression profiling by high throughput sequencing
Summary	We sequence the transcriptomes of 86,024 single cells from C. elegans embryos, spanning from gastrulation to the beginning of cuticle synthesis. We identify the lineage (from the invariant C. elegans cell lineage) and approximate developmental age of each cell in the single cell data. Using these annotations, we investigate the competing influences of cell lineage and cell fate on gene expression.
Overall design	Single cell RNA-seq profiles of cells from C. elegans embryos at varying developmental stages (~100-650 minutes post first cleavage). Please note that the GSM2599701 (in GSE98561) raw data was re-analyzed and the resulting (processed) data is linked to the GSM4318946 records, which is duplicated sample record of GSM2599701 (with the re-analysis details) for the convenient retrieval of the complete raw data from SRA

GSM4318946 c.elegans cells (L2 stage) [re-analysis]

Relations

BioProject	PRJNA523834
SRA	SRP186643

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE126954_RAW.tar	54.1 Mb	(http)(custom)	TAR (of TSV, TXT)
GSE126954_cell_annotation.csv.gz	2.7 Mb	(ftp)(http)	CSV
GSE126954_gene_annotation.csv.gz	173.1 Kb	(ftp)(http)	CSV
GSE126954_gene_by_cell_count_matrix.txt.gz	249.2 Mb	(ftp)(http)	TXT

SRA Run Selector

Raw data are available in SRA

Processed data are available on Series record

Processed data provided as supplementary file

GSM4318946 c.elegans cells (L2 stage) [re-analysis]

Relations

BioProject	PRJNA523834
SRA	SRP186643

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE126954_RAW.tar	54.1 Mb	(http)(custom)	TAR (of TSV, TXT)
GSE126954_cell_annotation.csv.gz	2.7 Mb	(ftp)(http)	CSV
GSE126954_gene_annotation.csv.gz	173.1 Kb	(ftp)(http)	CSV
GSE126954_gene_by_cell_count_matrix.txt.gz	249.2 Mb	(ftp)(http)	TXT

SRA Run Selector

Raw data are available in SRA

Processed data are available on Series record

Processed data provided as supplementary file

SRA Run Selector

Filters List

1 AvgSpotLen
2 Bases
3 BioSample
4 Bytes
5 Developmental_stage
6 Experiment
7 GEO_Accession
8 ReleaseDate
9 Sample Name
10 source_name
11 strain

Accession PRJNA523834 Search

Common Fields

BioProject	PRJNA523834
Consent	PUBLIC
Assay Type	RNA-Seq
Center Name	GEO
DATASTORE filetype	FASTQ, RUN.ZQ, SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.us-east1, ncbi.public, s3.us-east-1
Instrument	NextSeq 500
LibrarvLayout	PAIRED

Select

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	966	122.02 Gb	288.53 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

Found 966 Items

Search within results Clear

1 1 20

	Run	BioSample	AvgSpotLen	Bases	Bytes	Developmental_stage	Experiment	GEO_Accession	ReleaseDate	create_date	Sample Name	source_name
1	SRR11102139	SAMN14126930	70	113.13 M	47.48 Mb	L2	SRX7740729	GSM4318946	2020-04-24	2020-02-19 02:42:00Z	GSM4318946	whole c.elegans (L2 st
2	SRR11102140	SAMN14126930	70	122.22 M	51.82 Mb	L2	SRX7740729	GSM4318946	2020-04-24	2020-02-19 02:42:00Z	GSM4318946	whole c.elegans (L2 st
3	SRR11102141	SAMN14126930	70	209.35 M	87.73 Mb	L2	SRX7740729	GSM4318946	2020-04-24	2020-02-19 02:57:00Z	GSM4318946	whole c.elegans (L2 st
4	SRR11102142	SAMN14126930	70	125.96 M	52.57 Mb	L2	SRX7740729	GSM4318946	2020-04-24	2020-02-19 02:43:00Z	GSM4318946	whole c.elegans (L2 st
5	SRR11102143	SAMN14126930	70	96.16 M	40.45 Mb	L2	SRX7740729	GSM4318946	2020-04-24	2020-02-19 03:04:00Z	GSM4318946	whole c.elegans (L2 st
6	SRR11102144	SAMN14126930	70	128.87 M	53.77 Mb	L2	SRX7740729	GSM4318946	2020-04-24	2020-02-19 03:04:00Z	GSM4318946	whole c.elegans (L2 st
7	SRR11102145	SAMN14126930	70	115.29 M	47.97 Mb	L2	SRX7740729	GSM4318946	2020-04-24	2020-02-19 02:55:00Z	GSM4318946	whole c.elegans (L2 st
8	SRR11102146	SAMN14126930	70	75.24 M	31.55 Mb	L2	SRX7740729	GSM4318946	2020-04-24	2020-02-19 02:56:00Z	GSM4318946	whole c.elegans (L2 st
9	SRR11102147	SAMN14126930	70	106.55 M	44.42 Mb	L2	SRX7740729	GSM4318946	2020-04-24	2020-02-19 03:11:00Z	GSM4318946	whole c.elegans (L2 st

The sra toolkit

```
$ module load sratoolkit/latest
$ fasterq-dump SRR11102139
```

```
coletrap@nexus3:~/tmp/trash/packer_data$ fasterq-dump SRR11102139
spots read      : 1,616,210
reads read      : 3,232,420
reads written   : 3,232,420
coletrap@nexus3:~/tmp/trash/packer_data$ ls -al
total 464897
drwxr-xr-x 2 coletrap trapnelllab      4096 Mar 31 15:45 .
drwxr-xr-x 7 coletrap trapnelllab      4096 Mar 31 15:44 ..
-rw-r--r-- 1 coletrap trapnelllab 183035944 Mar 31 15:45 SRR11102139_1.fastq
-rw-r--r-- 1 coletrap trapnelllab 292938224 Mar 31 15:45 SRR11102139_2.fastq
coletrap@nexus3:~/tmp/trash/packer_data$
```

```
[coletrap@nexus3:~/tmp/trash/packer_data$ head SRR11102139_1.fastq
@SRR11102139.1 1 length=18
TGATTTTGGCCTGATATA
+SRR11102139.1 1 length=18
A/AAAEEEEEEEEEEEE
@SRR11102139.2 2 length=18
GCTCCGCGTGCAGCCTAC
+SRR11102139.2 2 length=18
AA/A/E6EEEEEEE6E/E
@SRR11102139.3 3 length=18
AAAGGATTATCTGCCAAT
[coletrap@nexus3:~/tmp/trash/packer_data$ head SRR11102139_2.fastq
@SRR11102139.1 1 length=52
TTGTACCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR11102139.1 1 length=52
AAAAAEA################################################
@SRR11102139.2 2 length=52
ATTCGTCGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR11102139.2 2 length=52
//AAAE/E################################################
@SRR11102139.3 3 length=52
CTCCTAGTGTATCTCGTTATTGCTGCGGTTAAAAAGCTCGTAGTTGGATCTA
```

There are 966 run files for this paper…

How might we automate the downloading of all of these files?

Clone the starter project repo

Click on this link: https://classroom.github.com/a/GrDP9OaP

`coletrapnell-teaching/packer-workflow-starter-<your_gihub_id>.git`

Now check out the README.md

THE END

For now