



Perspective

Tree of life at two levels: from species to cell

Shanjun Deng, Xionglei He*

State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

All species on the Earth can trace their origins back to a common ancestor, and their evolutionary history can be illustrated as a tree-like structure. The root of the tree represents the shared ancestor, while the bifurcations signify past speciation events, and the leaves indicate present-day species (Fig. 1a). Similarly, in multicellular organisms, all cells develop from a zygote, and their developmental lineage can also be depicted using a phylogenetic tree. In this tree, the root represents the zygote, the bifurcations signify mitotic cell divisions, and the leaves indicate terminally differentiated cells. Species tree has long been fundamental to the field of Ecology and Evolutionary Biology (EEB). Numerous sophisticated tools have been developed for reconstructing species trees and using them to draw inferences related to key evolutionary events, such as speciation, extinction, population size expansion or contraction, and so on. In contrast, despite the direct observation of the first complete cell tree in *Caenorhabditis elegans* nearly 40 years ago, the study of cell trees is still in its early stages [1]. The technical challenge of mapping a cell tree for a typical complex organism with millions to trillions of cells is a major contributing factor to the nascent stage of cell tree research. However, the underlying similarities between species and cell trees, which differ only in terms of time scale, have generated a growing interest in expanding the knowledge acquired from studying species trees to the emerging field of cell trees. Accordingly, there is an anticipated convergence between the two biological fields, namely, EEB and Molecular, Cellular, and Developmental Biology (MCDB), with the former focusing on species trees and the latter on cell trees. The integration of the two fields would spark a wave of new ideas to biological research. Notably, there exist unique research languages within each of the fields. For example, researchers in MCDB tend to have a prospective view of biological processes, whereas those in EEB tend to have a retrospective view. MCDB researchers focus on deterministic phenomena and their understanding, while EEB researchers acknowledge and appreciate stochasticity. MCDB specialists excel at creating experimental techniques for direct observation, while EEB experts have mastered the art of developing analytical tools for inference. MCDB researchers aim to characterize the comprehensive picture, while EEB researchers resort to sampling and use a few parameters with mathematical models to describe critical features. Integrating the different viewpoints from both perspectives would benefit both fields of study. This article will concentrate on the potential advances

of utilizing the language of EEB in MCDB studies. We will discuss recent progresses in reconstructing the developmental cell tree of a complex multicellular organism with somatic mutations, and then delve into cell tree-based population genetic modeling for comprehending organ development.

Experimental advances for reconstructing a cell tree. The tracking of cell division and lineage history in complex organisms, where direct observation is often difficult, involves two main methods. The first method utilizes naturally occurring somatic mutations during development to retrospectively reconstruct the cell lineage history. However, this technique is limited by the low natural mutation rate, which necessitates the examination of either numerous sites or the whole genome of individual cells [2,3]. While the application of this method is laborious, particularly when applied to a large number of cells, it yields sparse results, with only a few tens to hundreds of cells examined in comparison to the millions or trillions present in a complex organism. The second method involves the engineering of a barcoding system into model organisms, such as mice, zebrafish, and flies.

Over the past few years, two types of cell barcoding systems have been developed: static and evolving. The former involves viral transduction or transposon integration to integrate a random, invariable DNA barcode that reveals clone-specific hierarchy structures within the cell population. However, it overlooks the phylogenetic relationship among members of the same clone. In contrast, the latter requires heritable barcode sequences with mutations accumulating randomly and continuously, enabling the tracing of the entire evolutionary process of each individual cell over time.

The generation of mutations on evolving DNA barcodes can be achieved using various technologies. The most frequently utilized approach is the clustered regularly interspaced short palindromic repeat/CRISPR-associated nuclease 9 (CRISPR/Cas9) gene-editing system. In brief, the CRISPR/Cas9 gene-editing system utilizes Cas9 protein and guide RNAs to generate double-stranded breaks on specific genomic regions. These breaks are then repaired, inducing unique indels in individual cells, allowing for the generation of mutations on evolving DNA barcodes. Cas9-based lineage tracing systems have been successfully implemented in zebrafish (e.g., genome editing of synthetic target arrays for lineage tracing (GESTALT), lineage tracing by nuclease-activated editing of ubiquitous sequences (LINNAEUS), and ScarTrace) and mouse (e.g., CRISPR array repair lineage tracing (CARLIN), homing CRISPR, and CRISPR Recorder) via transgenic technology (Table 1). Despite the high efficiency of Cas9 in gene-editing, there are several limitations on lineage tracing, such as the difficulty in measuring mutations due to

* Corresponding author.

E-mail address: hexiong1@mail.sysu.edu.cn (X. He).

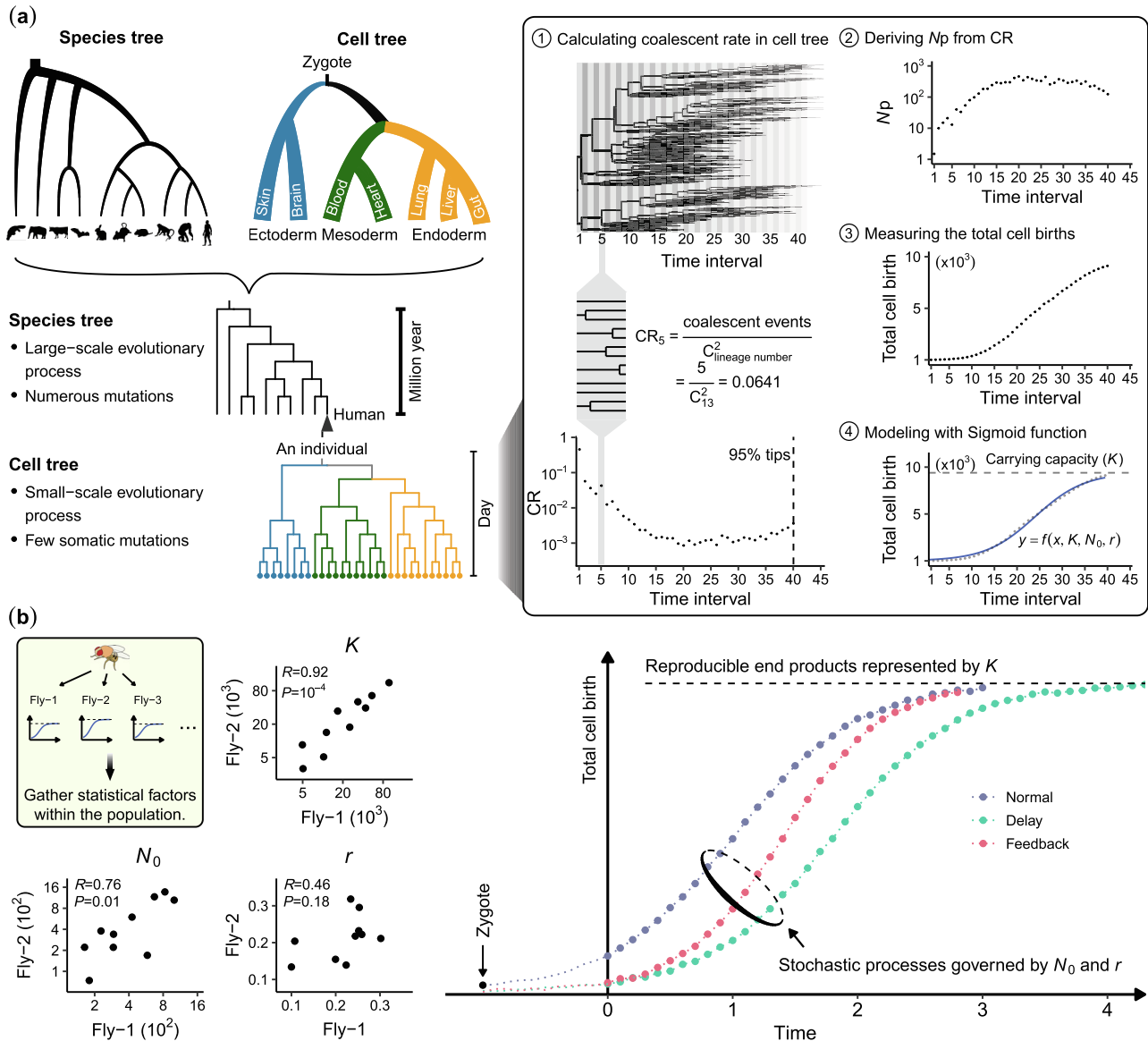


Fig. 1. A population genetic framework for studying developmental stochasticity and reproducibility. (a) Tree of life at two levels. The left side shows the hierarchical relationship and distinction between species tree and cell tree. The box highlighted on the right side shows the step-by-step workflow of applying SMALT and instantaneous coalescent method to calculating the coalescent rate (CR) and modeling development from a cell tree. (b) The left side shows the correlation of the key model parameters (K , N_0 , and r) between two *D. melanogaster* cell trees. Each dot represents a specific organ included in the cell trees. The diagram on the right side shows how the development stochasticity and reproducibility can be understood by examining K , N_0 , and r .

the balance between the number of target sites, the potential loss of historical lineage information caused by large-fragment deletions of adjacent target sites, and the possible cell stress or cytotoxicity resulting from DNA double-stranded breaks [4]. Second, the Cre-loxp recombination system is used in conjunction with the expansion of cellular populations to generate barcodes. For example, *Polylox* utilized a multi-loxp cassette that permits the random excision or deletion of genetic material, creating barcodes with unique arrangements [5]. However, the Cre-loxp method is not suitable for tracking long-term processes due to the irreversible excisions that continually decrease the number of available loxp sites. Third, evolving barcoding systems based on single base substitution have attracted growing interest because of their potential to track individual cells without double-stranded breaks that could cause cytotoxicity and the loss of informative sites in long-term processes. Recent developments in generating evolving barcodes have utilized deaminase enzymes to achieve several advances. For instance, a substitution mutation-aided lineage-tracing system (SMALT) was

used to tag two larvae of *Drosophila melanogaster*, enabling more than an average of 20 mutations on a three-kilobase-pair barcode [6]. The exceptional performance of the SMALT system unequivocally demonstrates the immense potential of single base substitutions in cellular phylogeny mapping.

Thoughtful consideration of two critical parameters is necessary when selecting an appropriate barcoding system for mapping a cell tree. The first parameter is the barcoding space, which should be significantly greater than the number of cells examined to avoid unrelated cells receiving the same barcode by chance. The theoretical barcoding space can be calculated by determining the number of mutable sites and the number of possible patterns for each site. The second parameter is the mutation number per barcode, which should be comparable to the number of mitotic generations from zygote to adult to enable precise tracing of the complete cell genealogy. Ideally, a reliable barcoding system should record at least one mutation on the barcode at each cell generation to offer robust bootstrapping support on the internal

tree branches for phylogeny analysis rather than solely clonal analysis.

Here, we have summarized the performance of various current strategies in Table 1. We noticed that most strategies contain around three mutations per barcode providing limited phylogenetic relationships. In contrast, the SMALT strategy contains approximately 1200 mutable sites and can obtain roughly 21 substitution mutations per barcode during the embryonic development of the fruit fly, *D. melanogaster*, *in-vivo*. The effective barcoding space of the SMALT system is theoretically larger than 10^{30} , which enables nearly the complete cell genealogy history of fruit fly larvae to be traced, taking about 26 cell divisions. Consequently, we demonstrated the capacity of a high-quality cell phylogeny in systematically and quantitatively characterizing both deterministic and stochastic processes in the development of a complex organism. Overall, the development of lineage tracing technology provides a valuable tool to investigate the cell tree of an organism throughout embryonic development, tissue homeostasis, or disease progression (e.g., carcinogenesis).

Computational integration with single-cell expression data. One of the fundamental goals of developmental biology is to precisely understand the mechanisms that result in a complex multicellular organism's development from a single zygote, including when and where cell fate determinations occur. Since cell fate determinations are often accompanied by cell state transitions, which entail changes in cellular gene expression, morphology, and function, the similarity between cells could potentially reflect a cell population's development history. To infer the temporal ordering of cells along a developmental or differentiation trajectory with single-cell transcriptomics, a popular computational approach called cellular trajectory inference has been developed. Despite its remarkable achievements, it's worth noting that cellular trajectory inference has three major limitations. First, in materializing cell state manifolds on a population level, it depends primarily on cell state snapshots and averages many individual cells, leading to the loss of elaborate information on cell individual dynamics. Second, the true histories of cell individuals should strictly follow a bifurcation structure corresponding to cell division events rather than a graph-like process with multiple in- or out-degrees. Third, researchers have come to realize that cellular trajectory inference cannot resolve complex lineage dynamics, especially when cell state transitions are not consistent with cell division history [7], especially when cell state transitions are incongruent with cell division history.

To overcome these limitations, several new approaches have emerged for integrating cell lineage and cell state information. For example, KP-Tracer [8], an evolving lineage-tracing system, has been applied to investigate the evolution and dynamics of lung adenocarcinoma and metastatic tumors and to map the phylogenetic relationships between cell states, revealing the common paths of tumor evolution *in-vivo*. Another integrative algorithm,

CoSpar [9], is capable of modeling the transition map in ancestral populations. PhyloVelo [10], a recent computational framework, reconstructs a novel transcriptomic velocity field based on monotonically expressed genes along with cell divisions on the phylogeny. Altogether, since lineage tracing can provide direct information about cell division history, like parent–child relationships and branching events, involving cell lineage data alongside cellular trajectory inference is crucial for a comprehensive understanding of cellular dynamics.

Despite of the promising application of lineage tracing technologies, analyzing the tree of cells generated from lineage tracing data remains challenging due to the large number of cells involved. For an individual of a complex organism, like mouse, the body is estimated to be composed with one to ten million of cells. Traditional phylogeny reconstruction algorithms, unfortunately, have high computational costs and time-consuming features that make it difficult to handle big data. Recently, new approaches have emerged to reconstruct cell phylogenies from big data obtained through various lineage tracing technologies. These include a scalable maximum parsimony approach, Cassiopeia [11], a maximum-likelihood approach called LinTIMaT [12], and a time-scaled method for phylogeny reconstruction known as PhyloTime [13]. Moreover, the FRACTAL framework enhances the scalability of classical lineage reconstruction algorithms by incrementally performing top-down reconstructions of a large lineage target through a hierarchy of distributed computing [14], enabling the reconstruction of a phylogeny of over 235 million cells.

Tree-based population genetic modeling of organ development. In principle, a complete cell tree is capable of providing comprehensive details of an organism's ontogeny. However, it's often unrealistic to obtain the complete cell tree of a complex organism due to the huge number of cells. In addition, cell lineage history is frequently non-deterministic in non-nematode, like higher metazoans, which causes the developmental cell tree of a species to vary among individuals. Hence, inferring the whole picture through sampling and developing mathematical models to extract stable rules from seemingly stochastic individual cell trees is necessary. Notably, by working on organismal populations in the past decades, the community of population genetics is now well-equipped with such tools. Specifically, there are sophisticated population genetic models available for inferring the population dynamics from the phylogenetic tree of sampled individuals [15,16]. One frequently employed theory in this context is known as coalescent theory, which rests on the premise that the rate of coalescence of the individuals in their phylogenetic tree is inversely proportional to the size of the examined population.

By adapting conventional population genetic models used for analyzing species trees, several inspiring studies have endeavored to uncover the latent cellular history from cell phylogeny [3,13,17,18]. Due to space limitations, we will focus on one of the representative algorithms, namely, the instantaneous coalescent

Table 1
Statistics of available evolving barcoding methods developed since 2016.

Barcoding method	Publication ^a	Applied system	Mutation type	No. of mutable sites (× barcode number)	No. of mutations per barcode	No. of mutations per barcode set	Phylogeny type
GESTALT	2016, <i>Science</i>	Zebrafish	Indel	10	2.86	–	Cell phylogeny
Polylox	2017, <i>Nature</i>	Mouse	Recombination	9	2.67	–	Cell phylogeny
ScarTrace	2018, <i>Nature</i>	Zebrafish	Indel	1 × 8	–	2.95	Tissue phylogeny
LINNEAUS	2018, <i>Nat. Biotechnol.</i>	Zebrafish	Indel	1 × 32	–	3.1	Tissue phylogeny
Homing CRISPR	2018, <i>Science</i>	Mouse	Indel	1 × 60	–	5.93	Organ phylogeny
Molecular Recorder	2019, <i>Nature</i>	Mouse	Indel	3 × 15	2.61	6.56	Cell phylogeny
CRISPR Recorder	2019, <i>eLife</i>	Fruit fly	Indel	1 × 32	<1	–	Cell phylogeny
Cas9-deaminase Barcoding	2019, <i>Nat. Commun.</i>	Cell line	Substitution	1 × 10 ⁵	<1	–	Cell phylogeny
CARLIN	2020, <i>Cell</i>	Mouse	Indel	10	~3	–	Cell phylogeny
TRACE	2020, <i>Nat. Biotechnol.</i>	Cell line	Substitution	~2000	~3	–	Cell phylogeny
SMALT	2021, <i>Nat. Methods</i>	Fruit fly	Substitution	1200	21.3	–	Cell phylogeny
DNA Typewriter	2022, <i>Nature</i>	Cell line	Insertion	5 × 13	39.4	–	Cell phylogeny

^a The detailed publication information is shown in Table S1 (online).

method developed by Yao et al. [17], which has been successfully adapted for high-resolution cell trees in fruit flies. The method allowed the estimation of the number of actively dividing parental cells (N_p) throughout development. Here, parental cells are those that had divided before the moment of sampling, with no any other functional meaning implied. Because each parental cell divides into two, N_p is equivalent to the net gain of cells. Liu et al. [6] applied the instantaneous coalescent method to the cell tree of two *D. melanogaster* samples that were barcoded with SMALT (Fig. 1a). They successfully obtained the trajectory of cell number growth throughout the fly development in each of 10 organs.

Obtaining the entire trajectory of cell number growth by sampling a single developmental time point would illuminate how developmental reproducibility emerges from stochasticity. In the study of Liu et al. [6], the time-dependent cumulative cell number in an organ often fits a logistic function in S-shape, which is widely used to model the population growth of organisms ranging from bacteria to humans. The function has three parameters, all of which have straight-forward biological interpretations: the initial population size N_0 , the growth rate r , and the carrying capacity K . It typically represents an exponential growth constrained by the carrying capacity of environment. The K values of the examined fly organs are highly consistent between the two individuals, suggesting a constraint on the final cell number in each organ (Fig. 1b). Contrarily, there is only moderate consistency for N_0 , and little for r between the two individuals. It is not surprising that the initial size N_0 is more variable than the final size K , because the former represents a snapshot of the stochastic process while the latter shows the sum of all previous moments. The poor consistency for r , however, suggests that the cell population growth in organs is subject primarily to K -selection, rather than r -selection. To understand how the carrying capacity K imposes constraints to ensure reproducible final organ size, one may examine the interaction between N_0 and r . There are two likely scenarios: N does not affect r , resulting in early developmental delay that isn't compensated; alternatively, N negatively affects r , resulting in delayed growth at an early stage compensated for by faster growth at later stages (Fig. 1b). If more fly individuals' cell trees were available, the correlation between the N_0 and r of each organ could determine which scenario is true. This shows how a cell tree informs the developmental reproducibility at an organ level, despite the unavoidable stochastic behaviors of individual cells.

Given the recent boom in tracking and reconstructing cell trees, an increasing number of high-resolution cell trees would be anticipated in years ahead. Therefore, it becomes imperative to contemplate how to amalgamate cell trees from diverse individuals characterized by sparse sampling and considerable stochasticity, in order to acquire shared insights within the population. In this context, leveraging tools and theories from population genetics can facilitate the quantitative depiction of cell trees, elucidating the underlying mechanisms of embryogenesis and tissue homeostasis.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China (2021YFA1302500 and 2021YFA1302501), the National Natural Science Foundation of China (32293190, 32293191, 31970570, and 32200492) and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (33000-31620269).

Appendix A. Supplementary materials

Supplementary materials to this perspective can be found online at <https://doi.org/10.1016/j.scib.2023.09.018>.

References

- [1] Sulston JE, Schierenberg E, White JG, et al. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev Biol* 1983;100:64–119.
- [2] Luo T, He X, Xing K. Lineage analysis by microsatellite loci deep sequencing in mice. *Mol Reprod Dev* 2016;83:387–91.
- [3] Lee-Six H, Obro NF, Shepherd MS, et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* 2018;561:473–8.
- [4] Geisinger JM, Stearns T. CRISPR/Cas9 treatment causes extended TP53-dependent cell cycle arrest in human cells. *Nucleic Acids Res* 2020;48:9067–81.
- [5] Pei W, Feyerabend TB, Rossler J, et al. Polylox barcoding reveals haematopoietic stem cell fates realized *in vivo*. *Nature* 2017;548:456–60.
- [6] Liu K, Deng S, Ye C, et al. Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. *Nat Methods* 2021;18:1506–14.
- [7] Wagner DE, Klein AM. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat Rev Genet* 2020;21:410–27.
- [8] Yang D, Jones MG, Naranjo S, et al. Lineage tracing reveals the phylogenetics, plasticity, and paths of tumor evolution. *Cell* 2022;185:1905–23.
- [9] Wang SW, Herriges MJ, Hurley K, et al. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat Biotechnol* 2022;40:1066–74.
- [10] Wang K, Hou L, Lu Z, et al. Cell division history encodes directional information of fate transitions. *bioRxiv* 2022:511094.
- [11] Jones MG, Khodaverdian A, Quinn JJ, et al. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. *Genome Biol* 2020;21:92.
- [12] Zafar H, Lin C, Bar-Joseph Z. Single-cell lineage tracing by integrating CRISPR-Cas9 mutations with transcriptomic data. *Nat Commun* 2020;11:3055.
- [13] Fang W, Bell CM, Sapirstein A, et al. Quantitative fate mapping: a general framework for analyzing progenitor state dynamics via retrospective lineage barcoding. *Cell* 2022;185:4604–20.
- [14] Konno N, Kijima Y, Watano K, et al. Deep distributed computing to reconstruct extremely large lineage trees. *Nat Biotechnol* 2022;40:566–75.
- [15] Wu C-I, Wang H-Y, Ling S, et al. The ecology and evolution of cancer: the ultra-microevolutionary process. *Annu Rev Genet* 2016;50:347–69.
- [16] Karcher MD, Palacios JA, Lan SW, et al. PHYLODYN: an R package for phylogenetic simulation and inference. *Mol Ecol Resour* 2017;17:96–100.
- [17] Yao Z, Liu K, Deng S, et al. An instantaneous coalescent method insensitive to population structure. *J Genet Genomics* 2021;48:219–24.
- [18] He X, Deng S, Gong H, et al. A statistical method for quantifying progenitor population reveals incipient cell fate commitments. *Research Square*; 2023. <https://doi.org/10.21203/rs.3.rs-2654427/v1>.



Shanjun Deng is a postdoctoral researcher at Sun Yat-sen University. He got his Ph.D. from the School of Life Science, Sun Yat-sen University in 2021. His research is centered on the study of evolutionary processes using mutation information, with a focus on understanding the dynamics of progenitors during the embryogenesis of multicellular organism. He is currently working on the development of lineage tracing technologies for higher vertebrates.



Xionglei is a full professor at Sun Yat-sen University. He obtained his Ph.D. from the University of Michigan–Ann Arbor in 2007 and subsequently returned to China through the "Hundred Talents Program". He has made notable contributions to the fields of genetics and systems biology. His recent research is centered on the development of a cell lineage barcoding system to explore microevolution processes in individual development.