



Published in final edited form as:

ACS Synth Biol. 2017 June 16; 6(6): 936–942. doi:10.1021/acssynbio.6b00309.

Quantitative Analysis of Synthetic Cell Lineage Tracing Using Nuclease Barcoding

Stephanie Tzouanas Schmidt^{†,iD}, Stephanie M. Zimmerman[‡], Jianbin Wang[†], Stuart K. Kim^{‡,§}, and Stephen R. Quake^{*,†,||,⊥}

[†]Department of Bioengineering, Stanford University, Stanford, California 94305, United States

[‡]Department of Genetics, Stanford University, Stanford, California 94305, United States

[§]Department of Developmental Biology, Stanford University, Stanford, California 94305, United States

^{||}Department of Applied Physics, Stanford University, Stanford, California 94305, United States

[⊥]Chan Zuckerberg Biohub, San Francisco, California 94518, United States

Abstract

Lineage tracing by the determination and mapping of progeny arising from single cells is an important approach enabling the elucidation of mechanisms underlying diverse biological processes ranging from development to disease. We developed a dynamic sequence-based barcode system for synthetic lineage tracing and have demonstrated its performance in *C. elegans*, a model organism whose lineage tree is well established. The strategy we use creates lineage trees based upon the introduction of synthetically controlled mutations into cells and the propagation of these mutations to daughter cells at each cell division. We analyzed this experimental proof of concept along with a corresponding simulation and analytical model to gain a deeper understanding of the coding capacity of the system. Our results provide specific bounds on the fidelity of lineage tracing using such approaches.

Graphical abstract

*Corresponding Author. quake@stanford.edu.

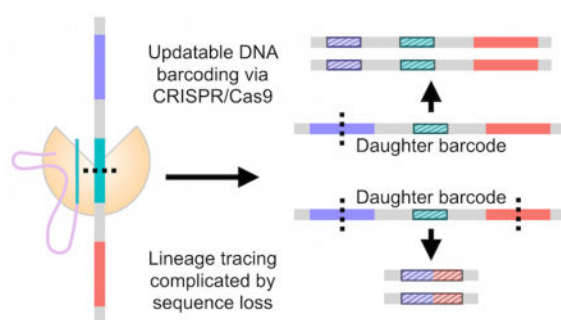
ORCID

Stephanie Tzouanas Schmidt: 0000-0003-3560-5782

Author Contributions

Conceived and designed the experiments: STS, SMZ, JW, SKK, SRQ. Contributed reagents/materials/analysis tools: SKK, SRQ. Performed the experiments: STS, SMZ. Analyzed the data: STS, SMZ, SSK, SRQ. Wrote the paper: STS, SRQ.

The authors declare no competing financial interest.



Keywords

DNA barcoding; lineage tracing; CRISPR/Cas9

Biologists have had a longstanding interest in lineage tracing, or the determination and mapping of progeny arising from single cells.¹ Numerous approaches have been realized and utilized over time, with the earliest attempts involving careful sectioning and staging of embryos² as well as direct observation of developing animals.^{3,4} As direct observation is not implementable for all systems, labeling of cells has been used to study a wide variety of questions in an assortment of experimental systems and organisms, first by dyes^{5,6} and radioactive tracers⁷ and later by fluorescent markers,⁸ genetic recombination,^{9–11} and static genetic barcoding.^{12–17}

While existing methods can identify the descendants of a cell, inferring the relationships between descendants remains to be adequately addressed, and all techniques thus far present drawbacks that render high-throughput and high-resolution lineage tracing a difficult undertaking. Intended for the study of tissues in which genetic lineage tracing is challenging and incomplete, the approach we present involves synthetic lineage tree construction based on the propagation of defined synthetic mutations arising in individual cells to their daughter cells.

Though prior work has explored lineage analysis based upon naturally arising mutations,^{18–22} we have engineered a strategy by which cell lineage trees can be constructed based upon the introduction of specific mutations into cells and the propagation of these mutations to daughter cells at each cell division. With the advent of clustered regularly interspaced short palindromic repeat nucleases (CRISPR) as a genome editing tool, tractable and specific targeting of genomic sites has been made possible.^{23,24} By creating mutations in a predictable manner using CRISPR/Cas9, this technology will enable more complete investigations of cell differentiation in processes ranging from development to cancer emergence. Recent reports also describe CRISPR-enabled barcoding strategies for lineage tracing, which were implemented in bacterial, cell line, mouse, and zebrafish systems.^{25–28} The groups using CRISPR/Cas9 to introduce mutations, as we did, did not analyze the coding capacity of the presented approaches.^{25,27} Our results here provide precise limits on the fidelity of such synthetic lineage tracing approaches.

To engineer dynamic, sequence-based barcodes, we used CRISPR/Cas9 genome editing to target ten specific sites within the gene encoding enhanced green fluorescent protein (EGFP) in the EG6173 strain of *C. elegans*. Our strategy relies upon nonhomologous end joining (NHEJ) to repair the double-stranded breaks caused by CRISPR/Cas9 and to introduce insertion-deletion mutations (indels) at the specified sites contained within the genome of a given cell. After such a site has been mutated, the indel introduced makes it unavailable for further alteration by CRISPR/Cas9 and serves as an identifiable marker of the cell and its daughters (Figure 1).

Given the diversity of indels generated by NHEJ, the accumulation of indels at the targeted sites over the course of several generations of cell division creates a barcode within a cell's DNA that allows for the construction of lineage trees based upon the indels shared between the barcodes of different cells (Figure 1A). That the sites designated for CRISPR/Cas9 editing occur within a region of 500 base pairs allows for the barcode to be easily amplified and read using paired-end sequencing. However, it is important to note that if CRISPR/Cas9 acts upon a second site before a given site can be repaired, the processes of NHEJ can result in a dropout, or the removal of sequence between the two sites, thus potentially omitting useful information (Figure 1B). In spite of this, our simulations show that our proposed dynamic DNA barcode approach could enable the tracing of cell lineages with potentially greater resolution and specificity than existing methods and without the need for continuous monitoring of the system of interest.

To characterize and better understand the CRISPR/Cas9-enabled barcoding technique, we simulated sequence barcodes and the resulting lineage trees resulting from its implementation in *C. elegans*, a nematode and model organism whose cell lineage tree was established in the late 1970s by means of real-time microscopic observation.⁴ Because our proposed method relies upon the stochastic introduction of indel mutations by CRISPR/Cas9 targeting, we represented CRISPR/Cas9 activity as a Poisson process and followed the known cell division pattern from *C. elegans* for the initial four divisions during which major cell lineages are specified. Barcodes were generated to represent two cases, in which NHEJ did and did not result in dropouts, the removal of sequence between targeted sites.

For both cases and for a range of expected values of the Poisson process, 100 sets of simulated barcodes were generated and corresponding lineage trees constructed. Modulating the expected value of the Poisson process enabled us to explore the effects of the CRISPR/Cas9 cutting rate, which we defined as the number of indels introduced per time. To compare the agreement of each simulated tree with the known lineage tree, we calculated the cophenetic correlation between the two. As cophenetic distance describes the intergroup similarity of the cluster in which a pair of observations is placed within a tree structure, the correlation between the resulting cophenetic distance matrices of full tree structures provides a suitable metric for the evaluation of concordance between two trees.²⁹

With this in mind, we computed the cophenetic correlation of each simulated tree with the known tree (Figure 2). We found that the correlation improved with increasing numbers of indels introduced during each generation, but even with approximately one indel introduced per cell division, corresponding to a Poisson expected value of 0.05 cuts per minute, the

lineages constructed agreed well with the reference lineage. As expected, the presence of dropouts reduced the agreement with the known tree, but despite this, for lower Poisson expected values, the difference in performance between barcode sets with and without dropouts was not statistically significant. However, it is important to note that even with the cutting rate approaching 8 cuts per cell division, the best correlations achieved were 0.99 without dropout and only 0.75 with dropouts. This suggests that in any practical situation the stochastic generation of lineage barcodes will not enable perfect lineage reconstruction unless the scheme is modified to eliminate dropouts.

We tested our proposed barcoding strategy by using it to label cells in *C. elegans*. Ten sites within the sequence encoding EGFP, codon optimized for *C. elegans*, were selected for targeting by CRISPR/Cas9, thus collectively serving as a sequence barcode. The stochastic creation of indels by CRISPR/Cas9 and NHEJ over the course of a worm's development was expected to generate a diverse collection of resulting sequence barcodes corresponding to different cells within the organism. Due to the sequential addition of indels during development, the indels observed in the barcodes obtained from the adult worm could then be used to infer the relationships between the cells from which the barcodes came, since once a cell's barcode is modified to include a particular indel, the barcodes of all subsequent daughter cells must also carry that indel.

Ribonucleoprotein complexes (RNPs) of Cas9 and single guide RNA (sgRNA) corresponding to each of the ten sites were injected into the gonad of worms from the EGFP-expressing EG6173 strain so that the RNPs would be encapsulated within the forming eggs. After 48 h post injection we screened the F1 progeny of the parent animals for absence of EGFP expression, since successful targeting by CRISPR/Cas9 of the selected sites within the EGFP gene would lead to the introduction of indels, thus disrupting the fluorescent protein (Figure 3). Progeny displaying the desired phenotype were dissected in order to remove their intestine, since this organ is known to derive from the descendants of a single cell, the E blastomere.^{4,30} The isolated intestine and remainder of the body were lysed and sequenced separately, and the EGFP gene containing the sites targeted was inspected for the presence or absence of indel mutations.

The indels contained within the sequence barcodes enabled the determination of lineage relationships between cells. Since indels are transmitted from parent to daughter cells, barcodes displaying shared indels can be considered to constitute a sublineage. We grouped barcodes by such common indels and used the frequency of indels observed across unique barcodes to infer the order of introduction of indels and thus, to construct lineage trees, a subset of which is shown in Figure 4. We found that barcodes from the intestine displayed an assortment of indels not observed in the rest of the body, suggesting a distinct cell type in agreement with the known lineage, since indels created after the formation of the E blastomere should be restricted to appearing in either the intestine or the body. Moreover, in both the sample derived from the intestine and that from the rest of the body, sets of barcodes contained shared indels, indicating descent from a common progenitor cell, such as the P1 or EMS blastomeres that occur prior to the formation of the E blastomere. The presence of barcodes such as the leftmost two shown that contain no indels or just one indel

in the adult worm reveals that the rate of CRISPR/Cas9 activity needs to be further tuned to record cell divisions over the entirety of development.

To assess the performance of the proposed strategy in correctly capturing the separation of the intestinal lineage from the rest of the body, the correspondence of the generated barcodes to tissues of origin was determined using the k-nearest neighbors algorithm.³¹ By comparing the predicted tissue of origin to the actual tissue of origin for each barcode, we were able to calculate two instructive metrics of information retrieval: recall, the ratio of correctly identified intestinal barcodes to the total number of actual intestinal barcodes, and precision, the ratio of correctly identified intestinal barcodes to the total number of putatively identified intestinal barcodes (Figure 5).³² This analysis was carried out using the full collection of barcodes derived from the sequenced intestine and body of the worm shown in Figure 4, and the results were compared to those generated by randomizing the data set's barcode assignments over 100 trials. The experimental results show that the precision of the actual data is 0.86 compared to 0.32 for the randomized case, while the recall is 0.40 compared to 0.03 for randomization. These results underscore the ability of updateable sequence barcodes to capture lineage relationships using approaches independent of tree construction.

To examine the diversity of indels generated using CRISPR/Cas9 and to calculate the information content of the described barcoding method, the indels observed across the ten targeted sites in the full collection of barcodes resulting from eight sequenced *C. elegans* samples were compared (Figure 6). For these 1039 indels, the probability distribution of positions contained within indels relative to the position of the CRISPR/Cas9-introduced doubled-stranded break (DSB) for cases in which dropouts did not occur can be described using a Gaussian curve (Figure 6a). We note that the distribution of positions included in indels was consistent across the eight samples, as two-factor analysis of variance without replication demonstrated that differences in the likelihood of occupancy of a particular position within an indel were not statistically significant across all worms and positions considered ($P=0.98$). The Shannon entropy calculated from this distribution is 4.42 bits with a 95% confidence interval of 3.24 to 5.48 bits³³ and represents the experimentally observed potential for information encoding at a single cut site.

To determine bounds on the performance of the evaluated barcoding technique and to contextualize the empirical estimate provided above, the theoretical upper and lower bounds of information provided by CRISPR/Cas9 barcoding was determined using simple analytical models. The probability of position occupancy by an indel conditional upon the indel's length was used to calculate an upper estimate of the Shannon entropy of such a system. Making use of the parameters describing the distribution of experimental indel lengths and assuming a uniform distribution for the utilization of positions themselves, we found that the Shannon entropy of an individual cut site within such a sequence barcode could encode 6.86 bits of information at most. To provide a more realistic estimate, we weighted the occupancy of particular positions by their likelihood of inclusion in the indel itself, giving a Shannon entropy of 6.61 bits.

The difference between the theoretical values obtained and the empirical result can be explained in part by the bias observed in the creation of indels (Figure 6b). By describing the

location of the DSB with respect to the center of the subsequent indel normalized to the length of the indel, it is evident that the indels resulting from CRISPR/Cas9 activity tend to lie to the left of the specified DSB, reducing the number of possible states and amount of information able to be encoded through barcoding. This point can be further illustrated by employing a modified version of the aforementioned analytical model. By using a Heaviside step function instead to describe the occupancy of positions around the DSB to determine a lower bound on performance, we calculated a Shannon entropy of 3.80 bits for a single cut site, which is still a considerable improvement over the effectively binary information supplied through current recombination- and color-based lineage tracing techniques. We assume that the greater the number of possible states offered by a lineage tracing system, the higher the amount of information it can possibly provide, translating into improved performance in capturing the lineage relationships between cells.

Hence, we report a novel technology enabling dynamic sequence barcoding for lineage tracing and definitively determine the information content of the proposed method. Existing approaches are limited in the resolution provided, the ability to combine lineage information with other system readouts, and the need for real-time observation.³⁴ We were motivated by the great potential for data storage and manipulation within DNA and thus developed our dynamic, sequence-based barcoding technology to address the aforementioned shortcomings. Previous work has exploited the natural occurrence of mutations in microsatellite loci^{18–22} or mitochondrial DNA³⁵ to trace lineages. Controllably introducing mutations in a desired genomic region and at rates higher than background as we propose offers the ability to track parent-daughter relationships with even greater clarity and without the need for whole-genome amplification. Though earlier approaches, *e.g.*, zinc finger nucleases³⁶ and TALENS,³⁷ have been used to edit genomes, the flexibility and specificity afforded by CRISPR/Cas9 made it attractive for the strategy we have described.

Our work demonstrates the potential of such CRISPR-based barcoding methods by illustrating the agreement between the fully known lineage of *C. elegans* and that obtained using dynamic sequence barcodes, delving into the challenge presented by sequence dropouts through a simulation-based approach, and underscoring the technique's promise by establishing its information content. Certainly, future work to develop control of CRISPR/Cas9 activity or to explore other methods of introducing mutations into DNA could improve the performance of such a system toward achieving the predicted upper limit of information content of 6.61 bits. With that much information in each site and just ten sites as we used in our experiments, one could create approximately 7.9×10^{19} unique barcodes, far more than all of the 3.7×10^{13} cells in the human body or the total number of cells in virtually any model organism of interest. Taken together, we present proof of concept and limits on the fidelity of cell tracking by means of dynamic sequence-based barcodes, which we are confident will enable unprecedented and more complete studies of cell lineage.

METHODS

Mathematical Modeling

Due to the stochastic nature of the proposed approach, the activity of CRISPR/Cas9 was represented as a Poisson process, which can be represented by the following:

$$P(n \text{ indels in a given time}) = \frac{\lambda^n e^{-\lambda}}{n!}$$

where λ is the average number of indels created per minute and n is the number of indels created in that time during the development of *C. elegans*. A Python script was written and implemented to update theoretical sequence barcodes corresponding to those expected in *C. elegans*³ based upon the introduction of indels whose occurrence followed a Poisson distribution within cells following the known division times and patterns. Across a range of values used for λ , similar to an indel introduced every other generation and to eight indels introduced each generation, barcodes were generated with and without dropouts, the removal of genetic material between cut sites. Cophenetic correlation was then used to compare the theoretical lineage trees determined from the modeled barcodes to the known lineage tree.²⁹

***C. elegans* Maintenance**

C. elegans expressing enhanced green fluorescent protein (EGFP) were obtained from the University of Minnesota's Caenorhabditis Genetics Center and served as the wild-type strain for these experiments. The nematodes were maintained per the established protocols.³⁸

Ribonucleoprotein Complex Preparation and Injection into *C. elegans*

Ten EGFP-specific sgRNAs were designed and transcribed using the MEGAshortscript T7 Transcription Kit from Ambion; purified Cas9 was purchased from PNA Bio. The pooled sgRNAs and Cas9 protein were mixed at a 1:2 mass ratio and incubated for 10 min at 37 °C.³⁹ The sgRNA-Cas9 complexes were injected into the gonads of wild-type young-adult P0 animals, and the resulting F1 progeny from the following 48 h were screened for knockdown of EGFP.

Sequencing and Analysis

F1 progeny were screened for EGFP knockdown. Selected worms were lysed and sequenced whole, while others were dissected to remove the intestine. Each intestine, corresponding body, or whole body was lysed separately per previously described methods.⁴⁰ The region encoding EGFP in each lysate was PCR-amplified and underwent paired-end sequencing using the Illumina MiSeq system. Low-quality and low-abundance reads were removed, and only reads joined by FLASH were used.⁴¹ A custom Python pipeline was used to align the obtained reads to the reference sequence over each selected cut site *via* Smith–Waterman alignment⁴² and identify indels introduced by the activity of the sgRNA-Cas9 complexes.

Only indels, all of which included or were within 3 base pairs of the DSB, were considered in order to mitigate potential amplification and sequencing errors within the barcodes. Each sequence was then expressed as a vector in which the presence or absence of each indel observed across the collection of indels was recorded using ones and zeros, respectively, to create binary barcodes. Maximum parsimony was implemented in R to reconstruct the relationships between unique binary barcodes and create lineage tree representations.^{43,44} To evaluate the accuracy and precision of the barcodes in correctly identifying different tissue

types, the k -nearest neighbors algorithm was executed in R for $k = 3$.³¹ Randomized assignments of the experimental data set were generated and the results compared to those from the actual assignments.

Calculation of Information Content

The distributions of lengths, l , and positions, d , occupied by indels were determined, and from the experimental position distribution, D , the Shannon entropy of the system in bits was calculated as follows, where p represents the likelihood of a particular position's inclusion in an indel per the observed distribution:

$$H = - \sum_{d \in D} p(d) \times \log_2 p(d)$$

The 95% confidence interval of this estimate was determined using empirical bootstrapping with ten thousand replicates.

To determine the upper bound of information able to be encoded within a single cut site, the following was considered:

1. We assume that the distribution of indel lengths, l , is Gaussian with mean, μ_l , and standard deviation, σ_l , where both parameters have been taken from the empirical distribution. We approximate the likelihood of a given indel length, l , of size up to 25 bp, by using the Gaussian cumulative distribution function:

$$p(l) = \Phi \left(\frac{l - \mu_l}{\sigma_l} \right)_{l-0.5}^{l+0.5}$$

2. We can express the probability of a particular position, d , being occupied within an indel in terms of the corresponding indel's length per the following equations.
 - a. Uniform Distribution

$$p(d|l) = \frac{1}{2l}$$

- b. Weighted Distribution

$$p(d|l) = \frac{l - |d| + 1}{l + 1}$$

- c. Heaviside Step Function

$$p(d|l) = \begin{cases} 1, & d \leq l \\ 0, & d > l \end{cases}$$

3. Combining the above two equations and normalizing the probability distribution obtained so that it sums to one,

$$p(d) = \sum_{l=d}^{\infty} p(d|l) \times p(l)$$

4. We can write the Shannon entropy as

$$H = - \sum_{d=0}^{\infty} p(d) \times \log p(d)$$

These calculations were carried out by numerical integration in Python using the parameters describing the Gaussian curve fit to the empirical distribution of indel lengths.

Acknowledgments

We thank Robert Phillips, John Beausang, and Winston Koh for helpful discussions; Norma Neff and Gary Mantalas for assistance with sequencing; and Ben Passarelli, Derek Croote, and Mark Kowarsky for computing support. *C. elegans* strains were provided by the University of Minnesota's Caenorhabditis Genetics Center, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). STS is supported by the Fannie and John Hertz Foundation Fellowship, National Science Foundation Graduate Research Fellowship, and Gabilan Stanford Graduate Fellowship. This work was supported by a National Heart, Lung, and Blood Institute (NHLBI) U01HL099995 Progenitor Cell Biology Consortium Grant (STS, JW, SRQ).

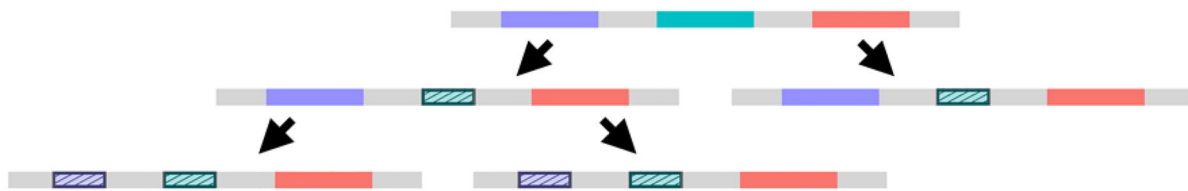
References

1. Kretzschmar K, Watt FM. Lineage Tracing. *Cell*. 2012; 148:33–45. [PubMed: 22265400]
2. Wilson, HV. Embryology of the Sea Bass. Government Printing Office; Washington, D.C.: 1891.
3. Sulston JE, Horvitz HR. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* 1977; 56:110–156. [PubMed: 838129]
4. Sulston JE, Schierenberg E, White JG, Thomson JN. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* 1983; 100:64–119. [PubMed: 6684600]
5. Vogt W. Teil Gastrulation und Mesodermbildung bei Urodelen und Anuren. Gestaltungsanalyse am Amphibienkeim mit örtlicher Vitalfärbung. *Dev. Genes Evol.* 1929; 120:384–706.
6. Weisblat DA, Sawyer RT, Stent GS. Cell lineage analysis by intracellular injection of a tracer enzyme. *Science*. 1978; 202:1295–1298. [PubMed: 725606]
7. Axelrod D. Carbocyanine dye orientation in red cell membrane studied by microscopic fluorescence polarization. *Biophys. J.* 1979; 26:557–573. [PubMed: 263688]
8. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. Green fluorescent protein as a marker for gene expression. *Science*. 1994; 263:802–805. [PubMed: 8303295]
9. Harrison DA, Perrimon N. Simple and efficient generation of marked clones in *Drosophila*. *Curr. Biol.* 1993; 3:424–433. [PubMed: 15335709]
10. Nowak JA, Polak L, Pasolli HA, Fuchs E. Hair follicle stem cells are specified and function in early skin morphogenesis. *Cell Stem Cell*. 2008; 3:33–43. [PubMed: 18593557]
11. Sun J, Ramos A, Chapman B, Johnnidis JB, Le L, Ho Y-J, Klein A, Hofmann O, Camargo FD. Clonal dynamics of native haematopoiesis. *Nature*. 2014; 514:322–327. [PubMed: 25296256]
12. Lu R, Neff NF, Quake SR, Weissman IL. Tracking single hematopoietic stem cells *in vivo* using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* 2011; 29:928–933. [PubMed: 21964413]
13. Naik SH, Perie L, Swart E, Gerlach C, van Rooij N, de Boer RJ, Schumacher TN. Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature*. 2013; 496:229–232. [PubMed: 23552896]
14. Kim S, Kim N, Presson AP, Metzger ME, Bonifacino AC, Sehl M, Chow SA, Crooks GM, Dunbar CE, An DS, Donahue RE, Chen ISY. Dynamics of HSPC Repopulation In Non-human Primates

- Revealed by A Decade-Long Clonal-Tracking Study. *Cell Stem Cell*. 2014; 14:473–485. [PubMed: 24702996]
15. Wu C, Li B, Lu R, Koelle SJ, Yang Y, Jares A, Krouse AE, Metzger M, Liang F, Lore K, Wu CO, Donahue RE, Chen ISY, Weissman IL, Dunbar CE. Clonal Tracking of Rhesus Macaque Hematopoiesis Highlights A Distinct Lineage Origin for Natural Killer Cells. *Cell Stem Cell*. 2014; 14:486–499. [PubMed: 24702997]
 16. Perie L, Duffy KR, Kok L, de Boer RJ, Schumacher TN. The branching point in erythro-myeloid differentiation. *Cell*. 2015; 163:1655–1662. [PubMed: 26687356]
 17. Biasco L, Pellin D, Scala S, Dionisio F, Basso-Ricci L, Leonardelli L, Scaramuzza S, Baricordi C, Ferrua F, Cicalese MP, Giannelli S, Neduva V, Dow DJ, Schmidt M, Von Kalle C, Roncarolo MG, Ciceri F, Vicard P, Wit E, Di Serio C, Naldini L, Aiuti A. *In vivo* tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell Stem Cell*. 2016; 19:107–119. [PubMed: 27237736]
 18. Wasserstrom A, Adar R, Shefer G, Frumkin D, Itzkovitz S, Stern T, Shur I, Zangi L, Kaplan S, Harmelin A, Reisner Y, Benayahu D, Tzahor E, Segal E, Shapiro E. Reconstruction of Cell Lineage Trees in Mice. *PLoS One*. 2008; 3:e1939. [PubMed: 18398465]
 19. Frumkin D, Wasserstrom A, Itzkovitz S, Stern T, Harmelin A, Eilam R, Rechavi G, Shapiro E. Cell Lineage Analysis of a Mouse Tumor. *Cancer Res*. 2008; 68:5924–31. [PubMed: 18632647]
 20. Reizel Y, Chapal-Ilani N, Adar R, Itzkovitz S, Elbaz J, Maruvka YE, Segev E, Shlush LI, Dekel N, Shapiro E. Colon Stem Cell and Crypt Dynamics Exposed by Cell Lineage Reconstruction. *PLoS Genet*. 2011; 7:e1002192. [PubMed: 21829376]
 21. Reizel Y, Itzkovitz S, Adar R, Elbaz J, Jinich A, Chapal-Ilani N, Maruvka YE, Nevo N, Marx Z, Horovitz I, Wasserstrom A, Mayo A, Shur I, Benayahu D, Skorecki K, Segal E, Dekel N, Shapiro E. Cell Lineage Analysis of the Mammalian Female Germline. *PLoS Genet*. 2012; 8:e1002477. [PubMed: 22383887]
 22. Shlush LK, Chapal-Ilani N, Adar R, Pery N, Maruvka Y, Spiro A, Shouval R, Rowe JM, Tzukerman M, Bercovich D, Izraeli S, Marcucci G, Bloomfield CD, Zuckerman T, Skorecki K, Shapiro E. Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability. *Blood*. 2012; 120:603–612. [PubMed: 22645183]
 23. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc*. 2013; 8:2281–2308. [PubMed: 24157548]
 24. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*. 2014; 346:1258096–1258096. [PubMed: 25430774]
 25. McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J. Whole organism lineage tracing by combinatorial and cumulative genome editing. *Science*. 2016; 353:aaf7907. [PubMed: 27229144]
 26. Shipman SL, Nivala J, Macklis JD, Church GM. Molecular recordings by directed CRISPR spacer acquisition. *Science*. 2016; 353:aaf1175. [PubMed: 27284167]
 27. Perli SD, Cui CH, Lu TK. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science*. 2016; 353:aag0511. [PubMed: 27540006]
 28. Frieda KL, Linton JM, Hormoz S, Choi J, Chow KK, Singer ZS, Budde MW, Elowitz MB, Cai L. Synthetic recording and *in situ* readout of lineage information in single cells. *Nature*. 2017; 541:107–111. [PubMed: 27869821]
 29. Sneath, PHA., Sokal, RR. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W.H. Freeman; San Francisco: 1973.
 30. Deppe U, Schierenberg E, Cole T, Krieg C, Schmitt D, Yoder B, von Ehrenstein G. Cell lineages of the embryo of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U. S. A.* 1978; 75:376–380. [PubMed: 272653]
 31. Shakhnarovich, D., Indyk, P., Darell, T. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press; Cambridge: 2006.
 32. van Rijsbergen, CJ. *Information Retrieval*. Butterworths; London: 1979.
 33. Cover, TM., Thomas, JA. *Elements of Information Theory. 2*. Wiley-Interscience; Hoboken: 2006.

34. Chen C, Puliafito A, Cox BD, Primo L, Fang Y, Di Talia S, Poss KD. Multicolor cell barcoding Technology for Long-Term Surveillance of Epithelial Regeneration in Zebrafish. *Dev. Cell.* 2016; 36:668–680. [PubMed: 27003938]
35. Walther V, Alison MR. Cell lineage tracing in human epithelial tissues using mitochondrial DNA mutations as clonal markers. *Wiley Interdiscip Rev. Dev Biol.* 2016; 5:103–117. [PubMed: 26302049]
36. Urnov FD, Miller JC, Lee Y, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature.* 2005; 435:646–651. [PubMed: 15806097]
37. Joung JK, Sander JD. TALENs: a widely applicable technology for targeted genome editing. *Nat. Rev. Mol. Cell Biol.* 2013; 14:49–55. [PubMed: 23169466]
38. Brenner S. The genetics of *Caenorhabditis elegans*. *Genetics.* 1974; 77:71–94. [PubMed: 4366476]
39. Cho SW, Lee J, Carroll D, Kim J, Lee J. Heritable gene knockout in *Caenorhabditis elegans* by direct injection of Cas9–sgRNA ribonucleoproteins. *Genetics.* 2013; 195:1177–1180. [PubMed: 23979576]
40. Ahringer J. Reverse Genetics WormBook. 2006; doi: 10.1895/wormbook.1.47.1
41. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics.* 2011; 27:2957–2963. [PubMed: 21903629]
42. Durbin, R., Eddy, S., Krogh, A., Mitchison, G. *Biological Sequence Analysis*. Cambridge University Press; Cambridge: 2005.
43. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics.* 2011; 27:592–593. [PubMed: 21169378]
44. Bolker, B., Butler, M., Cowan, P., Vienne, DD., Eddelbuettel, D. *Phylobase: Base Package for Phylogenetic Structures and Comparative Data*. R package version 0.8.2. 2010. <http://r-forge.r-project.org/projects/phylobase/>

A) NHEJ without dropouts



B) NHEJ with dropouts

**Figure 1.**

Inheritance of introduced mutations enables inference of cell lineages. Making CRISPR/Cas9 available during development allows for the introduction and transmission of indels in the sequence barcode of cut sites in individual cells. Sample sequence barcodes are represented by gray line segments, and the contained cut sites are represented by each differently colored subsection. Upon targeting by CRISPR/Cas9, the resulting indels are denoted by the shortening of the colored subsections and by the dashed fill pattern. (A) Schematic of representative barcodes in the case where no sequence between cut sites is lost (*i.e.*, no dropouts). (B) Schematic of representative barcodes in the case where dropouts occur, resulting in concomitant loss of information provided by previously created indels.

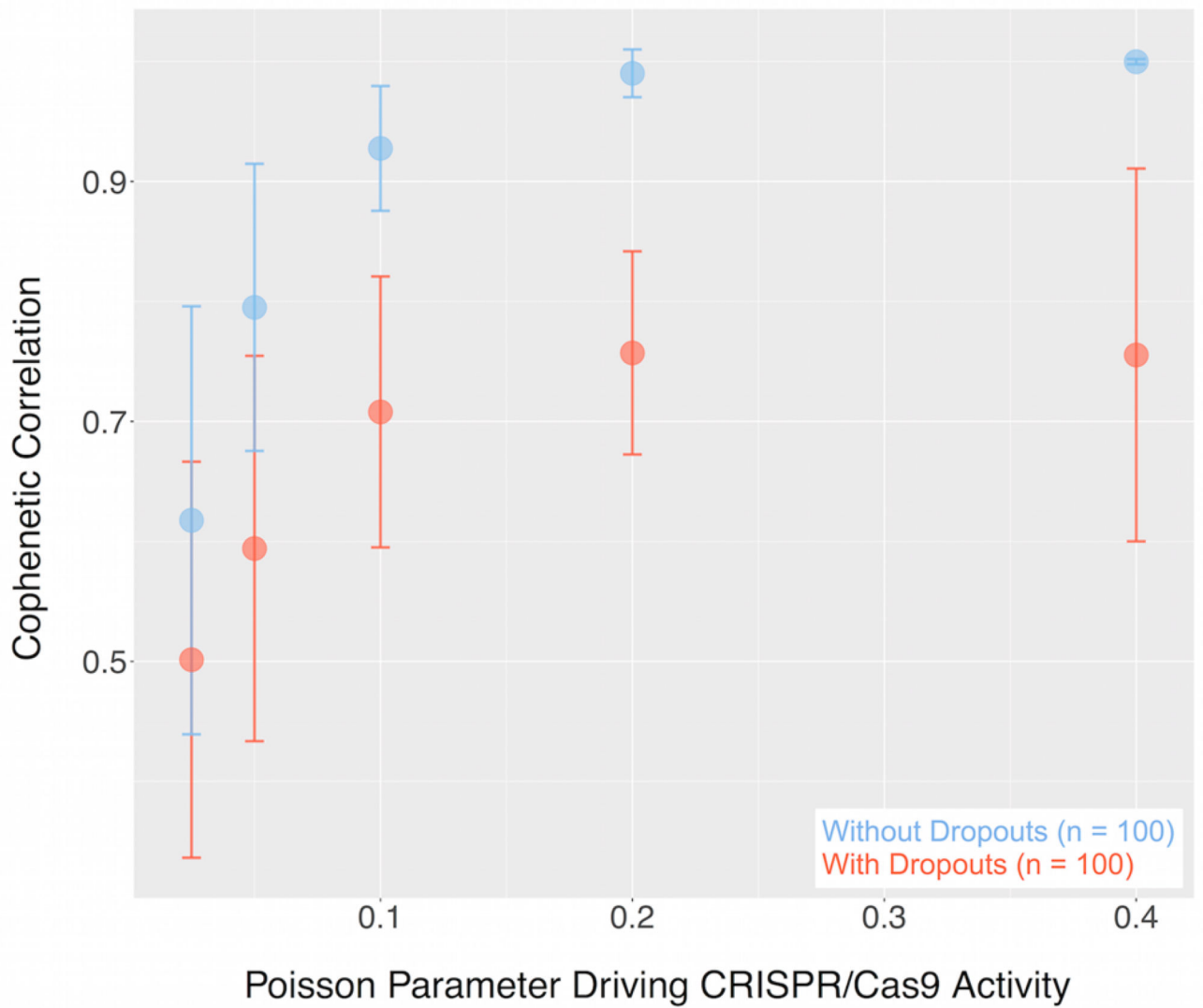


Figure 2. Performance overlaps predicted for cases with and without dropouts. Simulation of sequence barcodes and resulting lineage trees by modeling indel generation as a Poisson process without (blue) and with (red) dropouts enabled comparison with the known *C. elegans* lineage using cophenetic correlation for a range of Poisson expected values ($n = 100$).

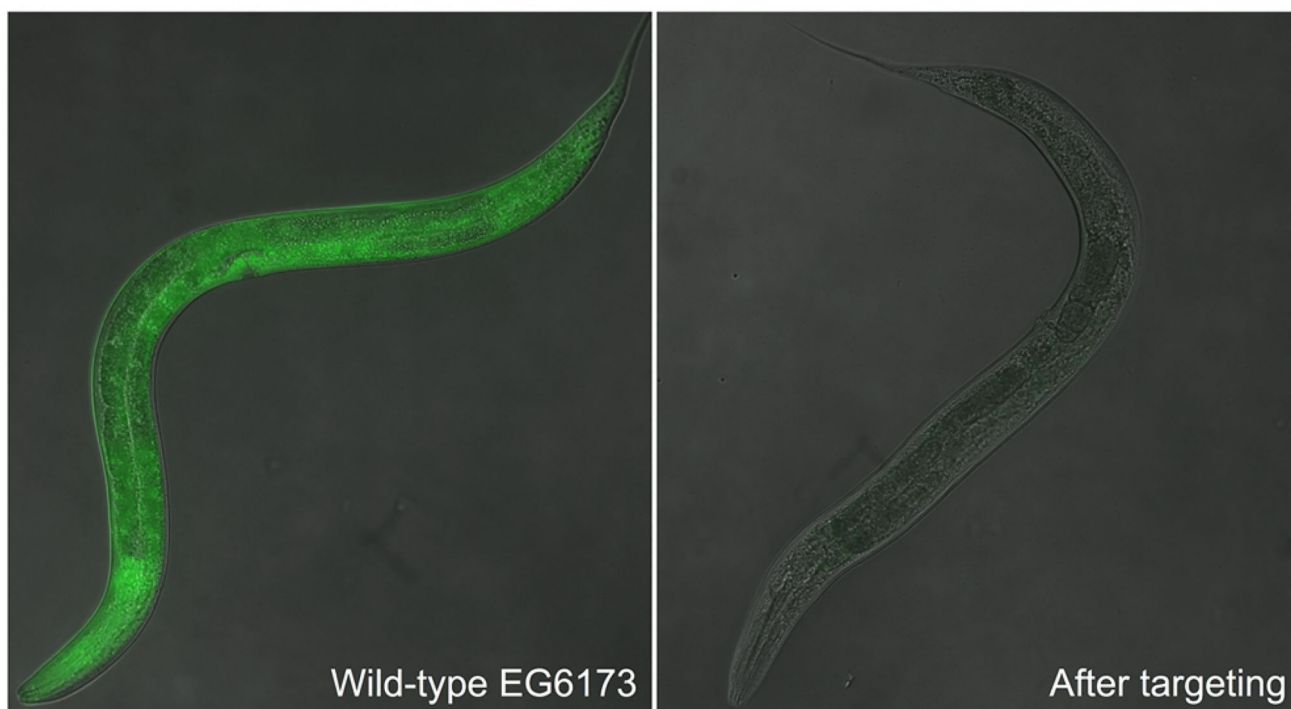


Figure 3. Effective CRISPR/Cas9 targeting screenable by phenotype. Introduction of indels into barcode encoded in EGFP sequence allows for identification of barcoded organisms through disruption of EGFP expression.

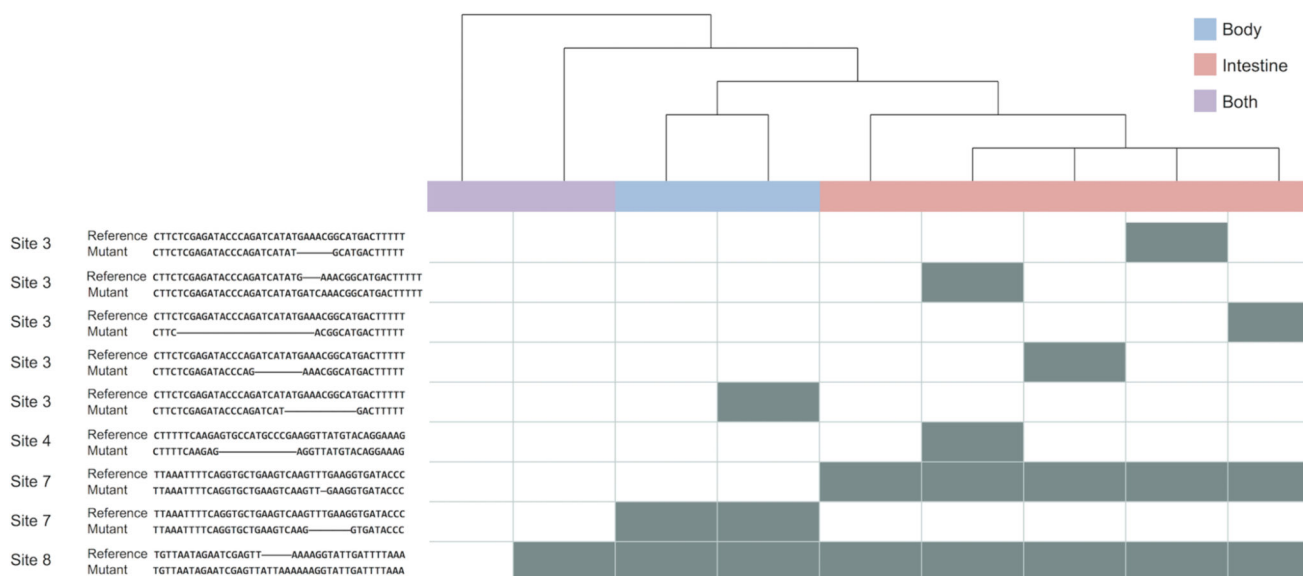


Figure 4.

Dynamic sequence barcoding identifies distinct cell populations. The relationships between a subset of unique sequence barcodes derived from the body (blue), intestine (red), and both samples (purple) of a single worm was used to create the following lineage tree. The barcodes corresponding to each member of the tree are presented in the columns of the heatmap, the individual indels are denoted by its rows, and each gray-shaded box indicates the presence of a given indel. The indels are presented in the table to the left of the heatmap as alignments between the reference (top) and observed (bottom) sequences, and the cut site with which each signature is associated is listed in the leftmost column of the table.

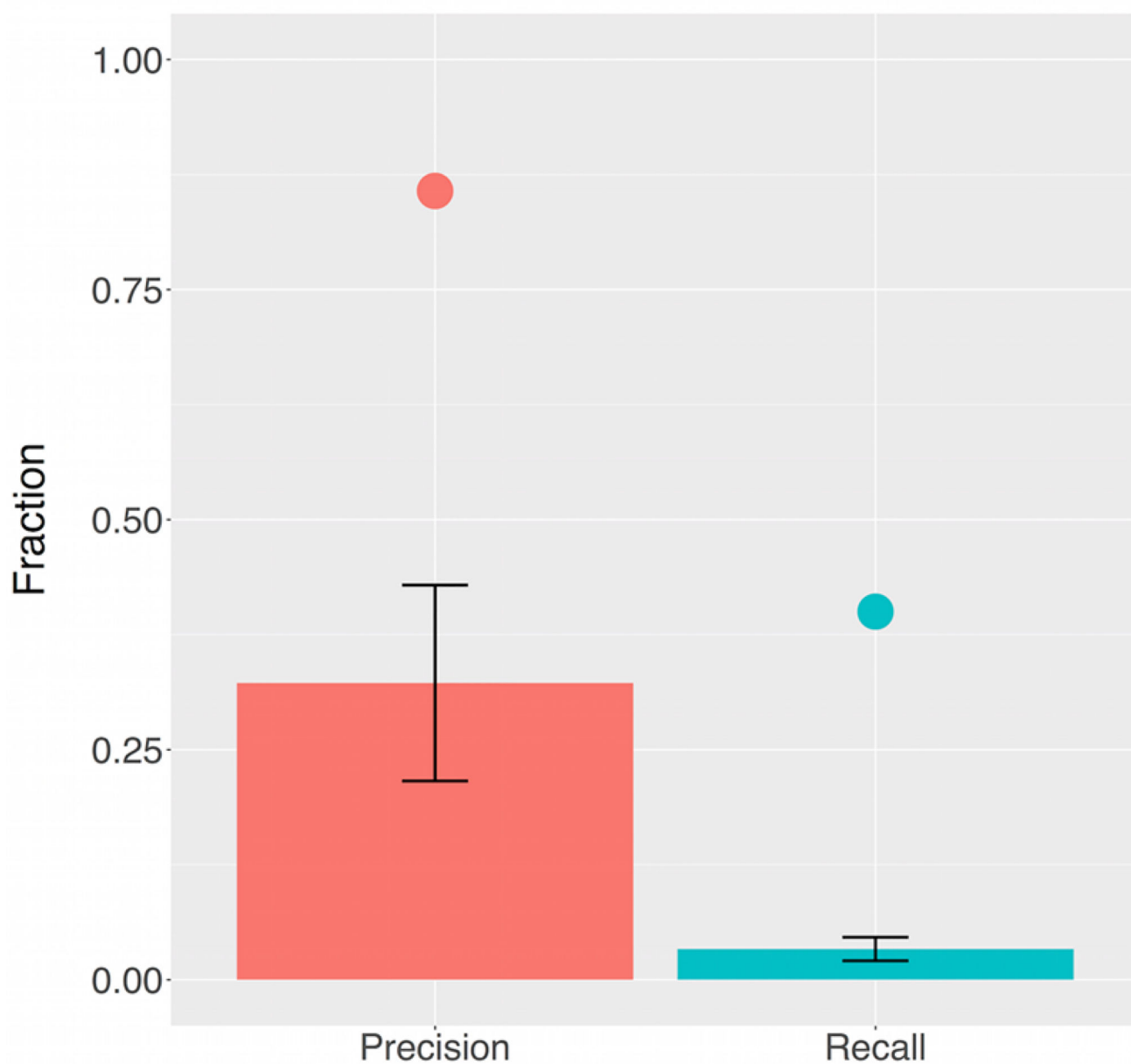


Figure 5. Barcoding permits determination of correct tissue of origin. The precision and recall of tissue identification based upon the full set of barcodes derived from the worm shown in Figure 4 for which the intestine and body were sequenced separately were calculated using the k-nearest neighbors algorithm. The experimental results, shown by the individual points over the columns, compare favorably with the results obtained by randomizing the data set's labels ($n = 100$).

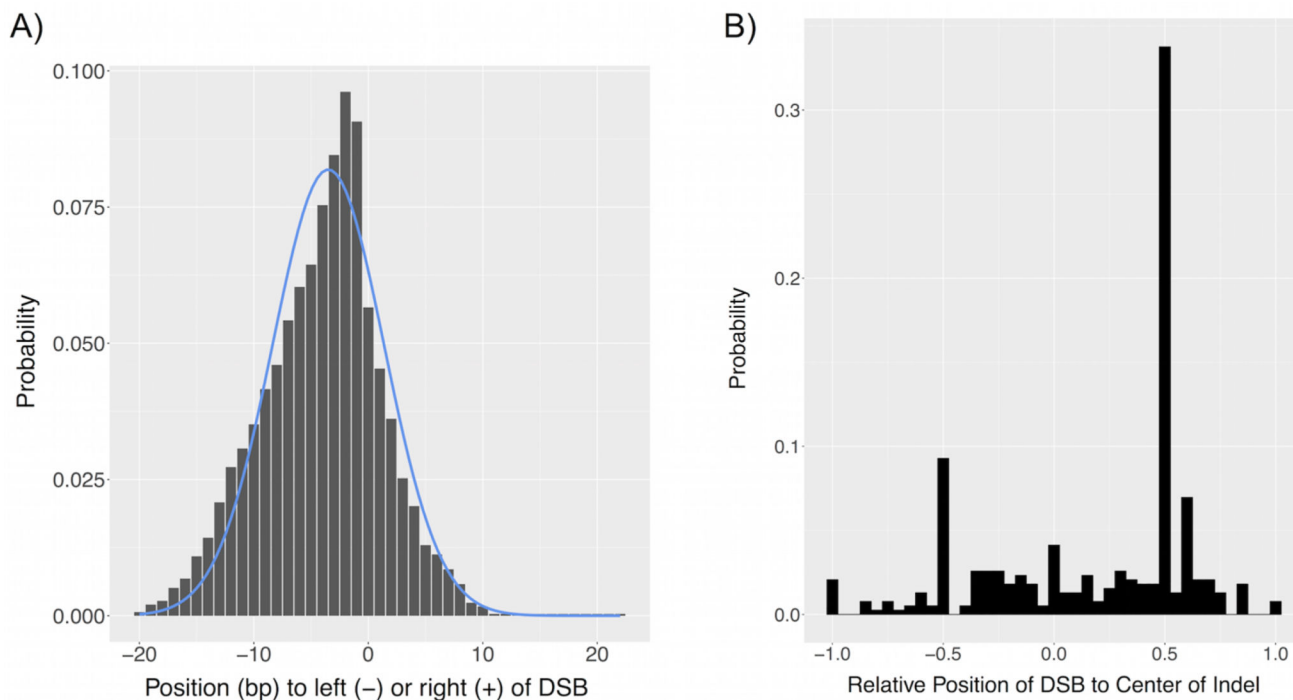


Figure 6.

Profile of unique indels generated in *C. elegans* by CRISPR/Cas9 provides insight into information content of proposed barcoding technique. The unique indels observed in sequenced animals ($n = 8$) across all ten targeted sites were compared. (A) The probability distribution of which positions flanking the CRISPR/Cas9-created double-strand break (DSB) were contained in the resulting indels has a Gaussian form. (B) The length-normalized position of the DSB as compared to the center of the indel tends to occur to its right (+) rather than left (-).