nature portfolio

Peer Review File

Random access and semantic search in DNA data storage enabled by Cas9 and machine-guided design

Corresponding Author: Professor Jeff Nivala

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Communications.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The reviewer thanks the authors for their responses. In general the additions to the manuscript have addressed most of the comments I had from the previous review. I have one set of related concerns remaining.

I remained concerned about the reference to “1.74 million images”. First, while I appreciate the authors’ edits to the text to address this concern, the syntax used will unlikely be understood by the broader scientific community outside of computer scientists. Second, from a molecular/physical standpoint, I respectfully disagree with how the authors are presenting the work in this regard and believe it is overclaiming what is being done.

Related, as brought up by R2.4 and R2.5, a physical system comprised of <500 oligos that then point to 1.74 million in silico images is quite different from actually executing a search/retrieval with the full data represented physically in the system. As the authors note, “addressing this issue requires the careful design of primers that are orthogonal to both the encoded data and each other, as well as the addition of physical redundancies in the data.” The same type of careful design of the encoded data (which does not exist physically in the authors’ system) so that they account for gRNA sequences would be required to be a true comparator to PCR (and to claim 1.74 million images) but was not demonstrated by the authors.

I would suggest at least removing the mention of 1.74 million images from the abstract, or being very clear using broadly understandable syntax that the 1.74 million images are in silico.

And also one additional modification that avoids additional experiments would be to calculate the relative benefit of omitting PAMs from payloads for Cas9 approach versus needing to omit entire primer sequences for PCR from payloads.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have made systematic revisions to their manuscript, particularly in the introduction and discussion sections. The current version has been greatly improved in clarity and credibility. Due to its sequence-specific double-strand cutting ability, the CRISPR system has the potential to provide new tools for data searching and manipulation in DNA storage. The demonstrated similarity-based searching capability is attractive, although the experimental results are somewhat poor.

However, I still believe that relying solely on the enrichment score is not convincing enough to support the claim that CRISPR can be used for DNA data access. This is because there is no direct evidence to prove that the data has indeed been retrieved at a high success rate. For example, suppose a file contains 100 DNA strands, of which 50 are amplified 1000 times, and the other 50 are not amplified. According to the authors' definition, a relatively high enrichment score could still be obtained, but in reality, a portion of the data has not been retrieved. Since Cas9 is the new data access tool

introduced here, the authors need to provide concrete experimental evidence. The requirement is not for 100% data recovery, but rather for quantification and discussion of the DNA data.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The manuscript is a resubmission of an earlier version.

The authors have done a great job in addressing the Reviewers' comments. The current form of the manuscript conveys the results of the paper clearly.

The code used in the paper is made available for reproducibility of the results. The reviewer did not check the code in detail. The provided "Readme" fille gives clear instructions on how to run the code.

The paper presents a novel contribution allowing fast random access in DNA-based storage. The solution consists of encoding the data using Neural Networks into DNA sequences and using the CRISPR-Cas9 technology to perform the similarity search on the desired data. Numerical results show the accuracy of the method in the number of times the "trageted" files are accessed compared to the "untrageted" ones.

The work meets the standards in this research field. To the extent of the Reviewer's understanding, the contributions made by this work are significant.

It would be great if the Authors could comment on the following question:
Does the time to retrieval include the time spent by the nanopore sequencing as well or is it only the time spent by the CRISPR-Cas9 cleavage? If so, when comparing to previous papers, is the sequencing method used ion the literature also nanopore sequencing? If not, the sequencing time of the previous should be accounted to showcase the savings brought by the introduced cleavage method.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Point-by-Point response to reviewer comments

Random access and semantic search in DNA data storage enabled by Cas9 and machine-guided design

We would like to thank the reviewers: **R1**, **R2**, **R3** for their continued review of our manuscript.

Responses to specific comments:

R1.1 I remained concerned about the reference to "1.74 million images". First, while I appreciate the authors' edits to the text to address this concern, the syntax used will unlikely be understood by the broader scientific community outside of computer scientists. Second, from a molecular/physical standpoint, I respectfully disagree with how the authors are presenting the work in this regard and believe it is overclaiming what is being done.

Response: We appreciate the reviewers concerns and we have now edited the abstract to clarify that the molecular payloads are addresses to groups of images stored in silico.

> *We then developed a molecular similarity-search approach combining machine learning with Cas9-based retrieval. Using a deep neural network, we mapped a database of 1.74 million images into a reduced-dimensional embedding, encoding each embedding as a Cas9 target sequence. These target sequences act as molecular addresses, capturing clusters of semantically related images. By leveraging Cas9's off-target cleavage activity, query sequences cleave both exact and closely related targets, enabling high-fidelity retrieval of molecular addresses corresponding to in silico image clusters similar to the query.*

Furthermore, we have edited the discussion to include language that clarifies this point:

> *Our C9SS similarity search architecture enables the retrieval of file addresses that point to content stored in silico that is similar to a query*

R1.2 Related, as brought up by R2.4 and R2.5, a physical system comprised of <500 oligos that then point to 1.74 million in silico images is quite different from actually executing a search/retrieval with the full data represented physically in the system. As the authors note, "addressing this issue requires the careful design of primers that are orthogonal to both the encoded data and each other, as well as the addition of physical redundancies in the data." The same type of careful design of the encoded data (which does not exist physically in the authors' system) so that they account for gRNA sequences would be required to be a true comparator to PCR (and to claim 1.74 million images) but was not demonstrated by the authors.

Response: We agree with the reviewer here that the design of a system where 1.74 million images are encoded as payloads into this architecture would present a challenge. However, the comment the reviewer linked from our previous responses is related to multiplexing our random access method, which we demonstrated in a system containing twenty five unique files. This we directly compare to PCR, which is the current state of the art for random ac-

cess in DNA data storage. We fully encoded all 25 files in this system and retrieved varying numbers of files in multiplex to make these comparisons.

We make no direct comparisons of our similarity search system to PCR based access because the currently published method instead relies on hybridization. The system is intentionally designed to store file addresses in a database pool, which are then used to index into a secondary database that contains the file content.
We hope that the text changes we made in addressing the previous concern also apply to this concern.

R1.3 I would suggest at least removing the mention of 1.74 million images from the abstract, or being very clear using broadly understandable syntax that the 1.74 million images are in silico.

Response: We have included clarity to our abstract that considers a broad range of backgrounds when desicribing our contributions. The full scope of our additions are detailed in **R1.1** and the specific sentence outlined here:

> *By leveraging Cas9's off-target cleavage activity, query sequences cleave both exact and closely related targets, en- abling high-fidelity retrieval of molecular addresses corresponding to in silico image clusters similar to the query.*

R1.4 And also one additional modification that avoids additional experiments would be to calculate the relative benefit of omitting PAMs from payloads for Cas9 approach versus needing to omit entire primer sequences for PCR from payloads.

Response: We appreciate the reviewer's point and would like to clarify that we did not omit PAM sequences from our payload designs. Given Cas9's stringent specificity and low frequency of off-target cleavage, the likelihood of unintentionally generating functional PAM-adjacent off-target sites within payload sequences is extremely low, and therefore, we did not consider additional mitigation necessary.

R2.1 However, I still believe that relying solely on the enrichment score is not convincing enough to support the claim that CRISPR can be used for DNA data access. This is because there is no direct evidence to prove that the data has indeed been retrieved at a high success rate. For example, suppose a file contains 100 DNA strands, of which 50 are amplified 1000 times, and the other 50 are not amplified. According to the authors' definition, a relatively high enrichment score could still be obtained, but in reality, a portion of the data has not been retrieved. Since Cas9 is the new data access tool introduced here, the authors need to provide concrete experimental evidence. The requirement is not for 100% data recovery, but rather for quantification and discussion of the DNA data.

Response: We appreciate the reviewer's consideration of this point and would like to clarify an important distinction between PCR amplification and Cas9-based targeting. Unlike PCR, the Cas9 targeting step in our experiments does not involve amplification and therefore does not introduce amplification-related biases. To provide concrete experimental evidence of

our retrieval efficiency, we refer to the single-plex file g10 random-access experiments: from approximately 1.8 million reads aligned to the g10 file address, we successfully recovered 97% of the 65,000 unique oligo strands encoding the file. This level of recovery represents only a modest dropout rate, consistent with typical expectations for sequencing-based retrieval methods.

R3.1 Does the time to retrieval include the time spent by the nanopore sequencing as well or is it only the time spent by the CRISPR-Cas9 cleavage? If so, when comparing to previous papers, is the sequencing method used ion the literature also nanopore sequencing? If not, the sequencing time of the previous should be accounted to showcase the savings brought by the introduced cleavage method.

Response: The sequencing time is not considered in when we report metrics on the time for retrieval and the reviewer is correct that the number just reflects the Cas9 reaction time. Sequencing is also not included in the time reported for the DNA hybridization protocol metric (approximately 24 hour reaction time).