Theoretical Computer Science 894 (2021) 203–213

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

RESCOT: Restriction enzyme set and combination optimization tools for rNMP capture techniques

Penghao Xu *, Francesca Storici *

School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA

ARTICLE INFO

Article history:
Received 6 April 2021
Received in revised form 3 August 2021
Accepted 5 August 2021
Available online 12 August 2021

Keywords:
rNMP incorporation
Ribonucleotides in DNA
Restriction enzyme
Optimization
ribose-seq
Next generation sequencing
rNMP capture techniques

ABSTRACT

The incorporation of ribonucleoside monophosphates (rNMPs) in genomic DNA is a frequent phenomenon in many species, often associated with genome instability and disease. The ribose-seq technique is one of a few techniques designed to capture and map rNMPs embedded in genomic DNA. The first step of ribose-seq is restriction enzyme (RE) fragmentation, which cuts the genome into smaller fragments for subsequent rNMP capture. The RE selection chosen for genomic DNA fragmentation in the first step of the rNMP-capture techniques determines the genomic regions in which the rNMPs can be captured. Here, we designed a computational method, Restriction Enzyme Set and Combination Optimization Tools (RESCOT), to calculate the genomic coverage of rNMP-captured regions for a given RE set and to optimize the RE set to significantly increase the rNMP-captured-region coverage. Analyses of ribose-seq libraries for which the RESCOT tools were applied reveal that many rNMPs were captured in the expected genomic regions. Since different rNMP-mapping techniques utilize RE fragmentation and purification steps based on size-selection of the DNA fragments in the protocol, we discuss the possible usage of RESCOT for other rNMP-mapping techniques. In summary, RESCOT generates optimized RE sets for the fragmentation step of many rNMP capture techniques to maximize rNMP capture rate and thus enable researchers to better study characteristics of rNMP incorporation.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

In the DNA replication process, deoxyribonucleoside triphosphates (dNTPs), the building blocks of DNA, are incorporated by DNA polymerases in the form of deoxyribonucleoside monophosphates (dNMPs) after two phosphate groups (PP) are cleaved off in the enzymatic reaction to polymerize DNA. Sometimes the DNA polymerases cannot distinguish ribonucleoside triphosphates (rNTPs), the building blocks of RNA, from the dNTPs because the molecules are similar. In this case, rNTPs are misincorporated into DNA and become ribonucleoside monophosphate (rNMPs). This phenomenon is called rNMP incorporation into DNA [1]. rNMPs incorporated into DNA alter the backbone of double-stranded DNA, resulting in structural and mechanical changes [2] as wells as increased fragility and genome instability [3], which have been linked to diseases [4–6].

* Corresponding authors.
E-mail addresses: pxu64@gatech.edu (P. Xu), storici@gatech.edu (F. Storici).

https://doi.org/10.1016/j.tcs.2021.08.006
0304-3975/© 2021 Elsevier B.V. All rights reserved.

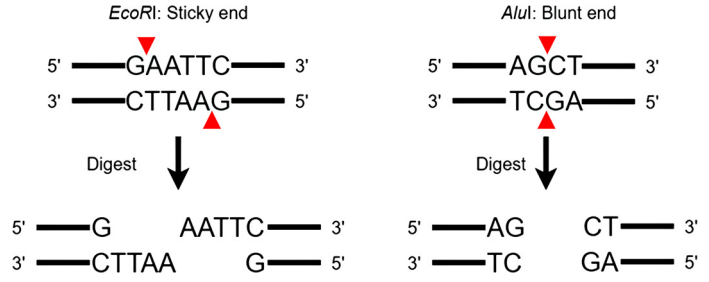

Fig. 1. Digestion of sticky-end and blunt-end restriction enzymes (REs). Diagram of RE digestion. If RE cuts at the middle of the recognition site, blunt end fragments are generated. Otherwise, the digestions lead to sticky end fragments.

A variety of studies on rNMPs embedded in DNA have focused on the composition and distribution of rNMP incorporation, including rNMP incorporation preference in different species [7–11], rNMP removal mechanisms [12], rNMP distribution bias on the leading and the lagging strands of DNA replication [13], and its relation to replication origin and DNA polymerases [11,14,15]. However, we still know very little about the characteristics of rNMP incorporation, particularly in mammalian DNA.

The first step to analyze the characteristics of rNMP incorporation is to locate the genomic coordinates of rNMPs at single-nucleotide resolution in the reference genome. Currently, there are six rNMP capture techniques for rNMP library preparation, including ribose-seq [7], emRiboSeq [13], Alk-HydEn-seq [11], RHII-HydEn-seq [14], Pu-seq [16], and RiSQ-seq [17]. All six techniques can capture rNMPs in different genomes. However, each technique is unique in its approach to capture rNMPs. In particular, the ribose-seq technique uses alkali treatment to cleave at the 3' of rNMPs in DNA. Then, *Arabidopsis thaliana* tRNA ligase (AtRNL) is used to capture the rNMPs. Hence, ribose-seq is the only method that directly captures rNMPs without capturing nicks in DNA or Okazaki fragments [7]. The emRiboSeq technique uses human RNase H2, which cleaves at the 5' of rNMPs in DNA, and T4 Quick ligase to capture the rNMPs. The upstream neighbor deoxyribonucleotide of the rNMPs is captured by emRiboSeq [13,18]. Using *Escherichia coli* RNase HII to cleave 5' to an embedded rNMP and T4 RNA ligase to ligate the rNMP, the RHII-HydEn-seq technique can also capture the DNA sequence terminated with an rNMP. However, the RHII-HydEn-seq technique may also capture Okazaki fragments [14] Alk-HydEn-seq and Pu-seq use alkali to cleave at the 3' side of rNMPs in DNA and use T4 RNA ligase 1 to capture the DNA sequence downstream from the rNMPs. These two methods can also capture Okazaki fragments [11,16]. RiSQ-seq was recently developed, and it captures the upstream neighbor deoxyribonucleotide of the rNMP [17]. Following library preparation, high-throughput DNA sequencing is performed on rNMP libraries. Afterward, the Ribose-Map bioinformatics toolkit can process and analyze the output rNMP sequencing data and locate the genomic coordinates of rNMPs with high performance [19].

Since DNA sequencing typically produces short reads, fragmentation of genomic DNA using restriction enzymes (REs) is an essential step for many molecular biology studies, including rNMP capture studies. REs can recognize specific sequences called RE sites, which are usually between four and eight-base pair palindrome sequences. Then, REs can cut at those sites, generating short DNA fragments. If an RE cuts in the middle of a palindrome RE site, both strands of DNA around the cut will have the same length, which leads to a blunt-end cut. Alternatively, RE can generate sticky-end fragments via a staggered cut at the DNA site (**Fig. 1**). Due to the high efficiency and reliability of REs, both ribose-seq and RHII-HydEn-seq use REs for DNA fragmentation. Ribose-seq uses a set of REs generating blunt ends for different libraries [7], while RHII-HydEn-seq uses *Sbf*I-HF, which recognizes an eight-base pair pattern and generates sticky ends [14]. It is important to optimize the RE usage to reach the maximum rNMP coverage for the genome fragmentation step of the rNMP-capture techniques, especially when the genome of interest is large like the human genome. Considering that rNMP incorporations are abundantly found in genomic DNA from bacteria to human cells and only little is known about rNMP functionality, there is a significant need to understand their characteristics, hotspots, and patterns. Understanding the function and consequences of rNMPs embedded in genomic DNA has a value not only for fundamental but also for clinical research.

The majority of rNMP incorporation studies are performed on species with a well-assembled reference genome, such as the sacCer2 reference genome for *Saccharomyces cerevisiae*, the hg38 reference genome for *Homo sapiens*, and the v5.5 reference genome for *Chlamydomonas reinhardtii*. Since REs usually have high fidelity, we can simulate the fragmentation of DNA by REs *in silico* and generate the predicted fragments using the sequence of the reference genome. There are various tools that determine the RE sites in a particular reference genome, such as NEBcutter [20] and SnapGene software (from Insightful Science; available at snapgene.com). However, generating an optimized RE set or a combination of RE sets for rNMP capture techniques is a more difficult task and has never been conducted before. Here, we propose a method to optimize RE usage for ribose-seq and other rNMP mapping techniques to increase the captured region of rNMPs across the genome. The software, Restriction Enzyme Set and Combination Optimization Tools for rNMP capture techniques (RESCOT), calculates the coverage of the rNMP-captured regions determined by REs in the reference genome of interest and optimizes those regions. While we apply RESCOT for mapping rNMPs in genomic DNA, these computational methods could also be useful for a variety of genomic analyses for which genomic DNA needs to be cut using REs in a manner that maximizes coverage of a particular region on the DNA fragments.

2. Theory and calculation

2.1. rNMP captured region and coverage

The first step to optimize the RE set usage is to locate the rNMP captured region and calculate its coverage. With the ribose-seq technique, an rNMP is captured through several steps including fragmentation, adapter ligation, alkali treatment, self-ligation by AtRNL, degradation of linear single-strand DNA, removal of 2'-phosphate, and PCR (**Fig. 2**). All fragments with an embedded rNMP remain in the concentration, while all fragments without an embedded rNMP are removed [7]. However, some fragments with an embedded rNMP are also filtered out by the size-selection bead purification step depending on the fragment length. Only DNA fragments of suitable length are selected since they bind to the bead surface. The fragment length is directly related to the rNMP location on the fragment. If an rNMP locates near the 5' RE cut, the fragment is too short and is eliminated in the size selection step. Similarly, rNMP incorporation far from 5' RE cut would generate a too-long fragment, which cannot bind to the bead surface, and they are also removed. Only if the rNMP is incorporated in a particular region on fragments, not too close or too far from the 5' RE cut, do the fragments have a suitable length to pass the size-selection filter after ribose-seq, and then to be sequenced. We term the DNA region of rNMP incorporation to generate suitable fragments as the rNMP-captured region. The coverage of the rNMP-captured region in the reference genome determines the performance of an RE set. A high coverage means the majority of rNMP incorporation positions can be captured by the ribose-seq technique. Oppositely, only a small portion of rNMPs can be captured when coverage is low. RE sets with high coverage are what we aim to achieve by RESCOT.

Since the fragment is composed of genome sequence and fixed-length adapters (including barcode and unique molecular identifies, UMI) on both ends, we can subtract the adapter length from the size-selection range to obtain the suitable genome sequence length. In ribose-seq, the 5' end of the genome sequence is an RE cut, and the 3' end of the genome sequence is the incorporated rNMP (**Fig. 2**). Hence, only if rNMP locates in a particular region from the 5'-RE cut, the fragment has a suitable length for genome sequencing. Specifically, the region starts with the minimum suitable genome sequence length (l_{start}) and ends with the maximum suitable genome sequence length (l_{end}). RESCOT calculates the coverage of the captured region using **Algorithm 1**.

Algorithm 1: Calculate the coverage of RE set.

Inputs: REs = Used REs;
G = Reference genome;
l_{start} = Start of rNMP-captured region;
l_{end} = End of rNMP-captured region
Output: Ribose-seq coverage

```
M ← 0; # Missing part length
L ← 0; # Genome size
for chromosome C in G do
|   L ← L + len(C);
|   # Get RE sites
|   S ← ∅;
|   for RE in REs do
|   | S_RE ← All RE sites in C;
|   end
|   S = S ∪ S_RE;
|   S.sort();
|   # Count missing parts
|   for i ← 0 to len(S) − 1 do
|   |   M ← M + min(S[i+1] − S[i], l_start); # Missing part before start
|   |   M ← M + max(S[i+1] − S[i] − l_end, 0); # Missing part after end
|   end
end
R ← 1 − M/L; # Calculate coverage
return R
```

Algorithm 1. Calculate ribose-seq coverage of a particular RE set and reference genome. With the RE used and the reference genome, the ribose-seq coverage is calculated. The start and end of the captured region are the two parameters in the algorithm, which should be the minimum and maximum length of suitable genome sequence length for size selection. Instead of counting the length of the captured region, the algorithm counts the length of the uncaptured region, in other words, the missing part. One minus the missing part fraction is the ribose-seq coverage.

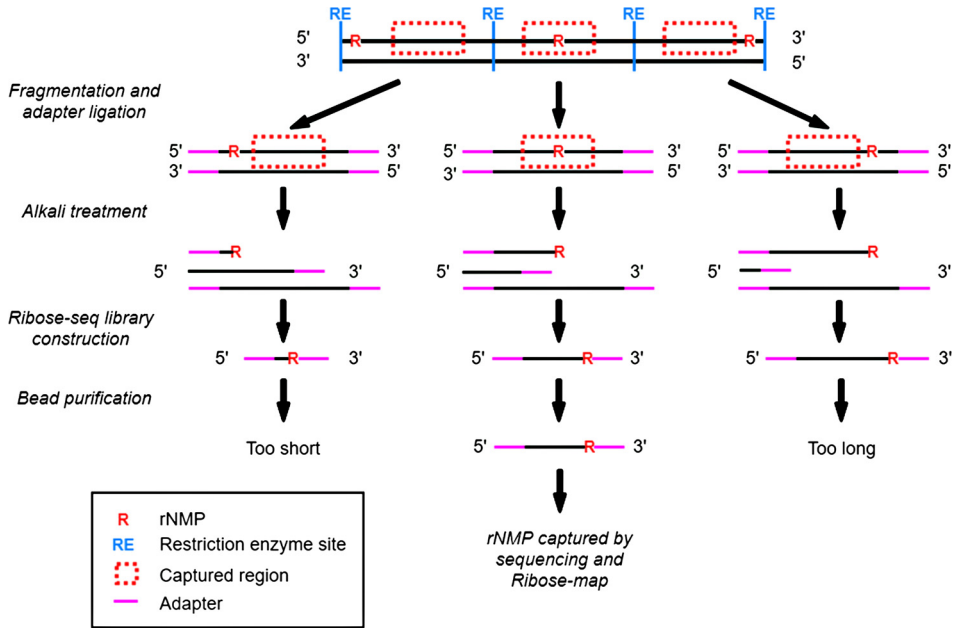

Fig. 2. rNMPs incorporated in the captured region can be captured by the ribose-seq technique. Not all incorporated rNMPs at different locations within the digested fragments can be captured with the ribose-seq technique. After digestion by RE, all fragments with rNMP incorporation are processed through the ribose-seq library construction steps. There are one or two bead-purification steps in the ribose-seq protocol after 2nd PCR as size filters. If an rNMP is incorporated too close to the 5' RE cut, it will be outside of the rNMP-captured region (left panel). The rNMP fragment with adapters is too short to pass the bead purification. If an rNMP is too far from the 5' RE cut, it will be also outside of the rNMP-capture region (right panel). The rNMP fragment with adapters is too long and is also removed by the bead purification step. Only fragments with an rNMP incorporated at a specific distance from the 5'-RE-cut end will be within the rNMP-captured region (central panel), and thus have a suitable length for bead purification to be sequenced. In the ribose-seq technique, usually, 250-650 bp fragments can pass the bead selection. Since the adapter size is 169 bp, the captured region spans 81-481 bp from the 5'-RE cut.

Besides the size selection through beads where the minimum fragment size is 81 bp in our experiments, Ribose-Map has size-selection criteria that filter out fragments less than 50 bp by default [19]. Hence, capturing fragments of 81 bp would be sufficient in this experiment to meet both filtering criteria.

2.2. Optimize RE set used for ribose-seq

If only one of the currently available REs is used in ribose-seq, the fragments tend to be too long, leading to low rNMP coverage. Therefore, ribose-seq uses multiple REs to form an RE sets for fragmentation, which results in shorter fragment length. However, simply using many REs in an RE set cannot achieve high coverage. It is because rNMPs located near the 5' end of all fragments are always out of the captured region and can never be captured. The RE set containing a large number of REs would have too short fragments, which would have a larger portion of the uncaptured region in the whole genome.

RESCOT aims to generate an optimized RE set with high coverage on the reference genome. All the REs in the optimized RE set should be selected from a candidate pool of REs that can work together in the same buffer. Meanwhile, the maximum and minimum number of REs in a set should be determined. Usually, fewer REs in the set can make the fragmentation preparation easier and reduce the cost. And more REs in the set sometimes can achieve higher coverage. Due to the high time complexity, we cannot find the optimized RE set by traversing all possible RE sets in the candidate pool (Brute force) since it is an NP-complete question. Here, RESCOT uses a local search algorithm to accomplish the optimization in a reasonable time and achieve good coverage. Briefly, at each iteration, RESCOT makes some mutation to the current RE set by adding one RE, removing an RE, or replacing an RE. Then, the coverage of the new RE set and the current RE set is compared, and RESCOT chooses which RE set to keep. This algorithm can significantly reduce the time required to optimize the RE set but are prone to drop in the local maximum instead of the global maximum. Thus, RESCOT uses a stochastic tunneling algorithm, based on simulated annealing to resolve the issue. The detailed steps are shown in **Algorithm 2**. When the new coverage is higher than the current one, the new RE set is accepted. When the new coverage is lower than the current one, RESCOT only accepts the new RE set with a small probability. The lower the coverage is, the less likely the new RE set is accepted. There is a β parameter that decides the likelihood to accept a mutation with a small reduction of coverage. When β is large, a small reduction of coverage is unlikely to be accepted, which speeds up optimization but will easily get into the local maximum. On the other hand, a small β leads to accepting a small reduction of coverage easily, giving us a higher probability to achieve the global maximum but is hard to converge. In our experiment, $\beta = 70$ is used,

which means there is a 10% probability to accept the new RE set when coverage decreases by 1%. This β value gives us good results within a reasonable time in our experiments.

Algorithm 2: Optimize RE set.

Inputs: $REList$ = List of candidate REs;
G = Reference genome;
I = Number of iterations
Output: Optimized RE set

```
# Initialization
REs ← Random REs from REList;
REs_max ← REs;
R_max ← Calc_coverage(REs, G);
R_prev ← R_max;
# Optimization
for i ← 1 to I do
|   REs_cache ← REs;
|   Add, remove, or replace one RE to REs;
|   R_curr ← Calc_coverage(REs, G);
|   # Update max coverage and corresponding RE set
|   if R_curr > R_max, then
|   |   R_max ← R_curr;
|   |   REs_max ← REs;
|   end
|   # Accept or decline current mutation
|   d = (R_curr − R_prev)/R_prev;
|   if rand() < exp(70 × d), then
|   |   R_prev ← R_curr; # Accept
|   else
|   |   REs ← REs_cache; # Decline
|   end
end
return REs_max
```

Algorithm 2. Optimize the RE set with local search and stochastic tunneling method. The RE set for the ribose-seq technique is optimized with an RE candidate pool and the reference genome. Starting with a random RE set, one mutation (adding, removing, replacing) is performed at each iteration. Then the ribose-seq coverage for the new RE set is calculated. If the new coverage is higher, the mutation is accepted. Otherwise, the mutation is only accepted with a small probability. Finally, the RE set with the highest ribose-seq coverage is considered as an optimized RE set.

2.3. Optimization of RE combination

Although RE sets are optimized as described in the section above, the coverage can never reach 100% with a single RE set, which means all rNMPs incorporated to have the chance to be captured. It is because all fragments have a short region at the 5' end out of the captured region. Thereby, DNA fragments with the highest coverage should have a length equal to the end of the rNMP-captured region. If all the DNA fragments have such length, ribose-seq reaches the highest coverage with a single RE set. Let l_{start} denote the start position, l_{end} denote the end position of the rNMP-captured region relative to the 5' end of the fragments. The upper bound of coverage is $(l_{end} - l_{start})/l_{end}$.

To further increase the coverage, we need to combine the RE sets for the ribose-seq technique. In particular, we divide the DNA into two halves. One-half is fragmented by RE set 1. The other half is fragmented by RE set 2. The missing part of combined two halves is the intersection of the two original missing parts, which is usually much shorter. Since the missing part is shortened in the mixture, the coverage is increased with the RE combination. If we use more RE sets to form a combination, higher coverage could be achieved.

Using an RE combination can increase the ribose-seq coverage. However, choosing two or more optimized RE sets may not lead to the best RE combination because their missing parts may be highly overlapping with each other. To generate an optimized RE combination, we generate random RE sets at the beginning. Then, the RE sets are randomly combined to form the RE combinations, and their coverages are calculated. Among all the iterations, the RE combination with the highest ribose-seq coverage represents the optimized RE combination. The detailed steps are shown in **Algorithm 3**.

Algorithm 3: Optimize RE combination.

Inputs: $REList$ = List of candidate REs;
G = Reference genome;
N = Number of RE sets used for a combination;
N_{set} = Number of random RE sets generated;
I = Number of iterations

Output: Optimized RE combination

```
# Initialization
Combination_best ← ∅;
M_min ← +inf; # Missing part length
REsets ← ∅;
Missing ← ∅;
# Generate random RE sets and find missing parts
for i ← 1 to N_set do
|   REs ← Random N REs from REList;
|   REsets.add(REs);
|   Missing[REs] ← Missing part of REs in G;
end
# Combination
for i ← 1 to I do
|   Combination ← N random RE sets from REsets;
|   M ← Length of intersection of all Missing[REs] for REs in| Combination;
|   if M < M_min, then
|   |   M_min ← M;
|   |   Combination_best ← Combination;
|   end
end
return Combination_best
```

Algorithm 3. Generate optimized RE combination with from RE candidate pool and reference genome. The RE combination is composed of several RE sets. To generate an optimized RE combination with high coverage, random RE sets are generated at the beginning. Then, the RE sets are randomly combined, and the ribose-seq coverage is calculated. The optimized RE combination is the one with the highest coverage among all iterations.

3. Material and methods

3.1. Reference genome and rNMP incorporation data

The *S. cerevisiae* sacCer2 reference genome is used for all the calculations in this study. We used three *S. cerevisiae* libraries fragmented by two different RE sets (FS162: RE3, FS164: RE3, and FS166: RE1) in Balachander et al. 2020 [8], which have significantly more rNMP incorporations than other libraries to validate the calculation results,

3.2. RE candidate pool

All REs used for coverage calculation, RE set optimization, and RE combination optimization, with their recognition sites and cutting sites are available on GitHub (https://github.com/xph9876/RESCOT/blob/main/example/res.txt).

3.3. Coverage calculation

There are 3 RE sets used in Balachander et al. 2020 [8]. Two of them (RE2 and RE3) are optimized RE sets generated by this method. RE set 1 contains *Dra*I, *Eco*RV, and *Ssp*I. RE set 2 contains *Alu*I, *Dra*I, *EcoR*V, and *Ssp*I. RE set 3 contains *Rsa*I and *Hinc*II. The coverages of all three RE sets are calculated in the study. For the parameters, the fragments of 250-650 bp can pass the bead purification step. Since the total adapter length is 169 bp, the captured region is 81-481 bp from the 5' end of fragments.

3.4. RE set and RE combination optimization

With the same captured region parameters, RE set optimization is performed with a minimum of 2 and a maximum of 5 REs in the set, and 100 iterations. For the RE combination optimization, 100 random RE sets are generated where a

minimum of 1 RE and a maximum of 5 REs are in each RE set. Then 2-4 RE sets are randomly selected to form an RE combination. The best combinations are picked among 1,000 iteration results.

3.5. Visualization of the results

All visualizations were generated using custom Python3 scripts, which are available on GitHub (https://github.com/xph9876/RESCOT).

4. Results

4.1. Validation of ribose-seq coverage calculation

As discussed above, we can locate the captured region and calculate its coverage of REs in the reference genome. Here, we analyzed three libraries prepared by two RE sets (FS162: RE3, FS164: RE3, and FS166: RE1. RE1: *Dra*I, *Eco*RV, and *Ssp*I. RE3: *Rsa*I and *Hinc*II) in Balachander et al. 2020 [8]. These libraries have significantly more rNMPs incorporated, which gives us more convincing data to validate our coverage calculation. We aligned rNMPs and compared their locations with the rNMP-captured regions. The results show that the majority of captured rNMPs are embedded in the captured regions we calculated on both positive and negative strands on mitochondrial DNA (**Fig. 3A, B**). No matter we measure the rNMP incorporation by the total counts of incorporated rNMPs, or the unique positions of incorporated rNMPs, the majority of rNMP incorporation in mitochondrial DNA is in the captured region (**Fig. 3C**). Yeast chromosomal DNA is much longer than mitochondrial DNA. We can still find that the majority of rNMP incorporation counts and positions are in the rNMP-captured region (**Fig. 3D**). Thus, it is reasonable to optimize the RE set or the RE combination used for ribose-seq by increasing rNMP-captured region coverage.

4.2. RE set and combination optimization

With the RE candidate pool and the reference genome, RESCOT can generate the optimized RE set with local search and stochastic tunneling. Here, we used an RE candidate pool of 38 REs generating blunt ends. Most of them have 4-6 bp recognition sites. Some of them have longer sites and degenerated bases. We optimized the RE set and combination in the sacCer2 reference genome. For the RE set optimization, the optimization process is shown in **Fig. 4A**. With the iterations, the coverage gets higher with large variations, which helps the local search method to jump out the local maximum and obtain the global optimized RE set. Since the mutant RE set with reduced coverage can only be accepted by a small possibility, the coverage of the last accepted RE set grows at a similar level with a much smaller variation. That leads to increasing coverage with iterations, which is shown as the maximum line. For the RE combinations, the coverage is also increased with iterations (**Fig. 4B**). We tried to use 2, 3, or 4 RE sets to form an optimized RE combination. The result shows that an RE combination composed of more RE sets can reach higher coverage. However, it also takes more time to optimize the RE combination and prepare the fragmentation step.

To check the performance of the RE set and combination optimization method, we compare the optimized RE set and RE combination containing 2, 3, and 4 RE sets with 1,000 random RE sets on sacCer2 reference genome (**Table 1**). Both random RE sets and optimized RE set contain 2-5 REs. We can find that the optimized RE set has higher coverage (0.5768) than the best one of the random RE sets (0.5766). For the combinations, 100 random RE sets with 1-4 REs in each are generated. Then, 1,000 iterations of the random combination are performed. Using 2 RE sets, the coverage reaches 0.8181, which is 41.83% higher compared to the optimized RE set. Using 3 RE sets, the coverage goes another 11.97% higher than the 2 RE sets. And the coverage of the 4 RE set combination is 4.63% higher than the 3 RE sets. Hence, the RE set and combination generated with this method are optimized to have the highest coverage and are suitable for the ribose-seq technique.

To validate the performance of RESCOT to generate optimized RE sets for the ribose-seq approach to capture rNMPs embedded in genomic DNA, we compared the number of unique rNMPs embedded in yeast DNA captured using ribose-seq with and without the optimization. Among the three RE sets used by the recent Balachander study, RE1 (*Dra*I, *Eco*RV, and *Ssp*I) was used to fragment yeast genomic DNA without any optimization by RESCOT, resulting in 45.61% coverage in the *S. cerevisiae* reference genome. RE2 (*Alu*I, *Dra*I, *Eco*RV, and *Ssp*I) is an RE set optimized by RESCOT that produced 56.50% coverage. RE3 (*Hae*III and *Rsa*I) is also optimized to 56.86% coverage and performs slightly better than RE2 [8] (**Table 2**). We can find the unique rNMPs incorporation coordinates are significantly higher when the optimized RE set is used in libraries of three different genotypes. Also, RE3 always has more rNMPs captured compared to RE2. Therefore, having an optimized RE set using RESCOT increases the coverage of the rNMP-captured region.

5. Discussion

5.1. Incorporated rNMPs outside of the captured region

In a typical rNMP incorporation library, most of the captured rNMPs are embedded in the rNMP-captured region. However, we find numerous rNMPs incorporated outside of the rNMP-captured region. There are several possible reasons for their occurrence.

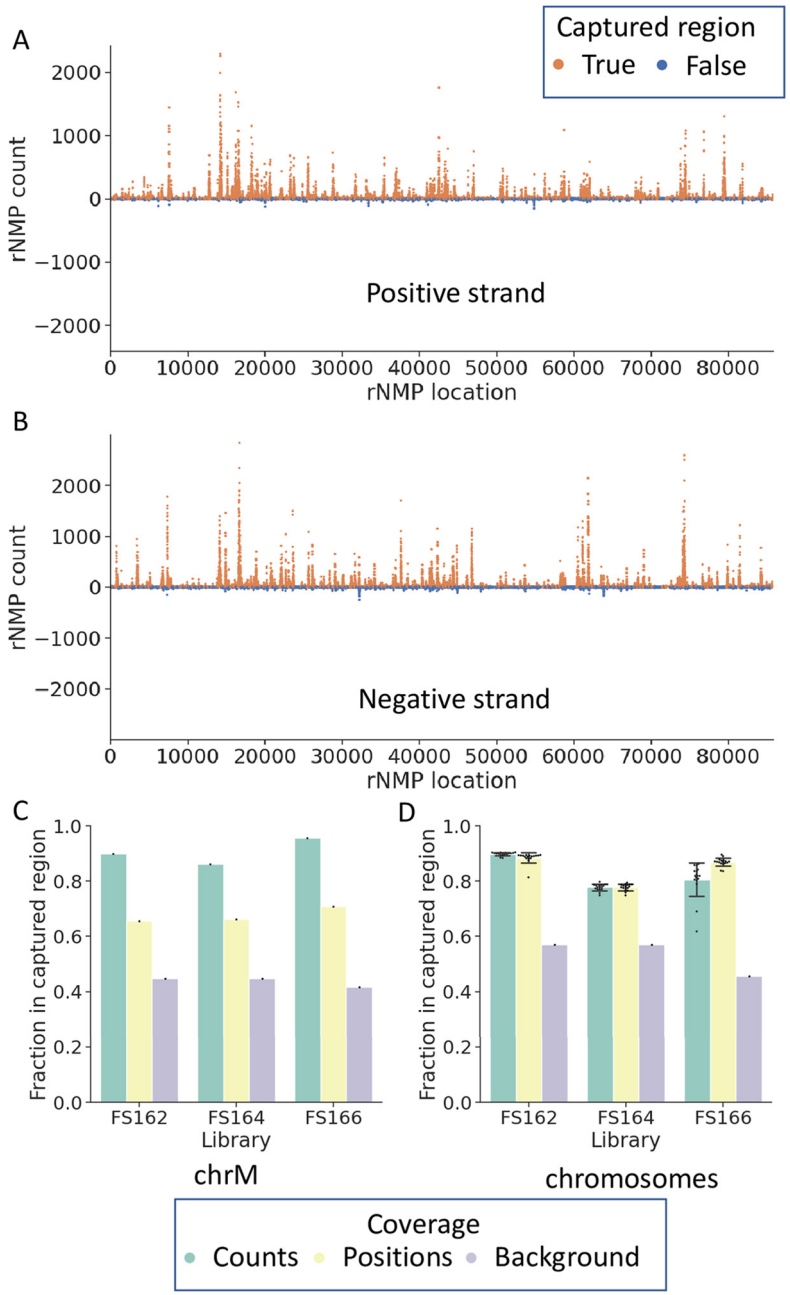

Fig. 3. The majority of captured rNMPs are embedded in the rNMP-captured region. The incorporated rNMPs on each position and whether they are in the captured region or not, are analyzed on mitochondrial DNA positive (A) and negative (B) strands separately in the FS166 ribose-seq library. The X-axis represents the coordinate of the mitochondrial DNA reference sequence. And the Y-axis shows the count of rNMPs captured in that position. The orange points with positive values are in the rNMP-captured region, which is more than 97% of rNMP incorporation. The blue points with negative values are out of the rNMP-captured region. The fraction of count and position of rNMPs incorporated in the rNMP-captured region of mitochondrial DNA in (C) and chromosomes (D) for the three libraries with different RE sets (FS162: RE3, FS164: RE3, FS166: RE1) are also calculated. The green bar measures the count, and the yellow bar measures the number of positions. Each chromosome is count as one data point in D. The standard deviations are shown as the error bar. The purple bar shows the fraction of the rNMP-captured region in the reference sequence. For all three libraries, we can find that the majority of the captured rNMP incorporations are in the rNMP-captured region in both mitochondrial and nuclear DNA. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

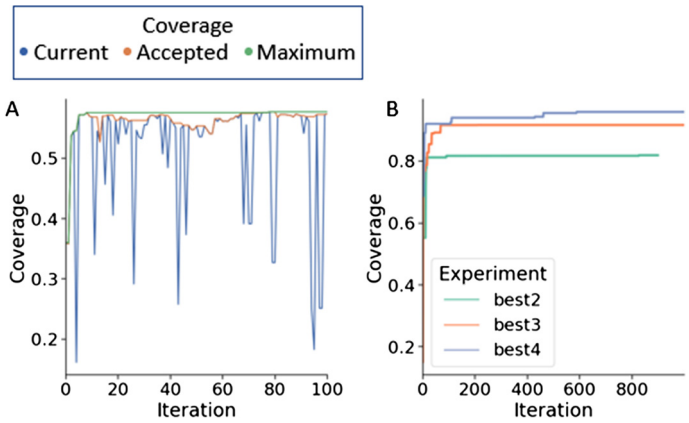

Fig. 4. The optimization process of RE set and combination. The coverage of each iteration during the optimization of the RE set and combination is tracked. There are 100 iterations of mutation as the process of RE set optimization (A). The blue line shows the coverage at each iteration. The orange line shows the coverage of the last accepted RE set. And the green line shows the maximum coverage by the current iteration. For the process of RE combination (B), 100 random RE sets are generated and 1,000 iterations are performed for the combination. The lines represent the maximum coverage by the current iteration. The green, orange, and blue lines show the RE combination composed of 2, 3, and 4 RE sets. We can find both RE set and combination are optimized with the iterations.

Table 1
Comparison of the optimized coverage of different methods and random RE sets. The coverage of optimized RE sets and RE combinations are compared with random RE sets. 1,000 Random RE sets are generated with 2-5 REs in each RE set. For RE set optimization, 2-5 REs are used to compose an RE set, and 100 iterations of optimization are done. For RE combination, random 100 RE sets are generated with 1-4 REs in each set. Afterward, 1,000 iterations of random combinations are performed.

Method	Coverage
Random RE sets	
Mean $\pm$ Std	0.4041 ± 0.1478
Maximum	0.5766
Optimized RE set (2-5 REs)	0.5768
Optimized RE combinations	
2 RE sets (1-4 REs per set)	0.8181
3 RE sets (1-4 REs per set)	0.9160
4 RE sets (1-4 REs per set)	0.9584

Table 2
The number of unique rNMPs coordinates increases when an optimized RE set is used. The numbers of unique rNMP captured of each selected library in Balachander et al. are shown [8]. The genotypes are indicated in bold font. The RE set usage and its coverage are marked after the library name. Significant more rNMPs are captured by using optimized RE set RE2 and RE3.

Libraries	Unique rNMP coordinates
WT	
FS166 (RE1, 46.51%)	81,019
FS162 (RE2, 56.50%)	134,716
rnh201	
FS117 (RE2, 56.50%)	299,237
FS141 (RE2, 56.50%)	338,052
FS115 (RE3, 56.86%)	435,978
FS116 (RE3, 56.86%)	407,515
RED	
FS103 (RE1, 46.51%)	8,590
FS150 (RE2, 56.50%)	178,400

The first possible explanation is linked to the size-selection bead purification step of the ribose-seq protocol. The DNA fragments containing an rNMP that have an appropriate size can bind to the bead surface. The DNA fragments containing an rNMP that have a size outside the range remain in the concentration and are removed in the washing step. In the computational calculation, we used a fixed maximum and minimum size threshold for bead purification. However, in the experimental procedure, the threshold of maximum and minimum size of the DNA fragments that the beads can bind is not precise. If DNA fragments are slightly longer than the maximum, or slightly shorter than the minimum size for the specific type of beads, there is still a probability that such DNA fragments are bound to the beads. For these fragments, the actual rNMP-captured region will be longer than what is indicated in the calculations.

The activity of the REs also contributes to capturing rNMPs outside of the rNMP-captured region. An RE may leave some RE sites uncut due to inefficient activity in the buffer or interaction with other REs in the set. For example, if there is only one RE site in the middle of a short DNA sequence, when the site is cut by the RE, it generates two extremely short DNA fragments. Hence, there should be no rNMP-captured region in both of these DNA fragments because they are shorter than the start of the rNMP-captured region (l_{start}). However, if the RE does not cut at that site all the time in the genomic DNA preparation, there is one longer DNA fragment, which may be longer than the start of the rNMP captured region (l_{start}) and have an rNMP-captured region at the end.

Another possible explanation for capturing rNMPs that are outside the rNMP-capture regions is the presence of genomic breaks that are not caused by the chosen REs. Such breakages can convert a long fragment into two smaller fragments, which may extend the rNMP-captured regions.

5.2. Usage in other rNMP capturing techniques

The RESCOT method is designed for the ribose-seq technique. However, it can also serve to optimize other rNMP-capture techniques. Among these, the emRiboSeq, Pu-seq, and RiSQ-seq methods use sonication for fragmentation, and a bead purification process is used before sequencing [13,17,21]. RESCOT could be applied to these methods if the sonication step is replaced by the RE digestion to generate fragments of defined length. The Alk-HydEn-seq method uses alkaline hydrolysis for fragmentation [11]. Alkaline hydrolysis may also be replaced by using an optimized RE set to obtain fragments with defined lengths. For the RHII-HydEn-seq technique, the RE SbfI-HF is used, which is an RE with an 8-bp recognition site generating a sticky end [14]. RESCOT could serve to find another RE set to increase the coverage of RHII-HydEn-seq.

6. Conclusions

We designed a method termed RESCOT to calculate the coverage of the rNMP-captured region of the reference genome. We analyzed a subset of previously published results obtained using ribose-seq and found most of the rNMPs in the rNMP-captured region calculated by RESCOT. Furthermore, as part of RESCOT, we designed an algorithm to optimize the RE set and combination used in ribose-seq with the RE candidate pool and the reference genome, which significantly increases the coverage of the rNMP-captured region and the number of unique rNMPs captured by ribose-seq. RESCOT has been designed for ribose-seq, but the tools can also be expanded to accommodate other rNMP capture techniques and potentially other types of genomic analyses.

Funding

This work was supported by the National Institutes of Health (NIEHS R01 ES026243 to F.S.), and the Howard Hughes Medical Institute Faculty Scholars Award (HHMI 55108574 to F.S.).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology for their research cyberinfrastructure resources and services, T. Yang for helping to prepare the restriction enzyme, A.L. Gombolay, and D.L. Kundnani for critically reading the manuscript and all members of the Storici laboratory for assistance and feedback on this study.

References

[1] J.S. Williams, S.A. Lujan, T.A. Kunkel, Processing ribonucleotides incorporated during eukaryotic DNA replication, Nat. Rev. Mol. Cell Biol. 17 (2016) 350–363, https://doi.org/10.1038/nrm.2016.37.

[2] H.C. Chiu, K.D. Koh, M. Evich, et al., RNA intrusions change DNA elastic properties and structure, Nanoscale 6 (2014) 10009–10017, https://doi.org/10.1039/c4nr01794c.

[3] H.L. Klein, Genome instabilities arising from ribonucleotides in DNA, DNA Repair 56 (2017) 26–32.
[4] C.F. Moss, I.D. Rosa, L.E. Hunt, et al., Aberrant ribonucleotide incorporation and multiple deletions in mitochondrial DNA of the murine MPV17 disease model, Nucleic Acids Res. (2017), https://doi.org/10.1093/nar/gkx1009.
[5] B. Hiller, A. Hoppe, C. Haase, et al., Ribonucleotide excision repair is essential to prevent squamous cell carcinoma of the skin, Cancer Res. (2018), https://doi.org/10.1158/0008-5472.CAN-18-1099.
[6] K. Aden, K. Bartsch, J. Dahl, et al., Epithelial RNase H2 maintains genome integrity and prevents intestinal tumorigenesis in mice, Gastroenterology (2019), https://doi.org/10.1053/j.gastro.2018.09.047.
[7] K.D. Koh, S. Balachander, J.R. Hesselberth, F. Storici, Ribose-seq: global mapping of ribonucleotides embedded in genomic DNA, Nat. Methods 12 (2015) 251–257, https://doi.org/10.1038/nmeth.3259.
[8] S. Balachander, A.L. Gombolay, T. Yang, et al., Ribonucleotide incorporation in yeast genomic DNA shows preference for cytosine and guanosine preceded by deoxyadenosine, Nat. Commun. 11 (2020), https://doi.org/10.1038/s41467-020-16152-5.
[9] W.M.M. El-Sayed, A.L. Gombolay, P. Xu, et al., Disproportionate presence of adenosine in mitochondrial and chloroplast DNA of Chlamydomonas reinhardtii, GISci. Remote Sens. 24 (2021) 102005, https://doi.org/10.1016/j.isci.2020.102005.
[10] S. Jinks-Robertson, H.L. Klein, Ribonucleotides in DNA: hidden in plain sight, Nat. Struct. Mol. Biol. 22 (2015) 176–178, https://doi.org/10.1038/nsmb.2981.
[11] A.R. Clausen, S.A. Lujan, A.B. Burkholder, et al., Tracking replication enzymology in vivo by genome-wide mapping of ribonucleotide incorporation, Nat. Struct. Mol. Biol. 22 (2015) 185–191, https://doi.org/10.1038/nsmb.2957.
[12] J.L. Sparks, H. Chon, S.M. Cerritelli, et al., RNase H2-initiated ribonucleotide excision repair, Mol. Cell (2012), https://doi.org/10.1016/j.molcel.2012.06.035.
[13] M.A.M. Reijns, H. Kemp, J. Ding, et al., Lagging-strand replication shapes the mutational landscape of the genome, Nature 518 (2015) 502–506, https://doi.org/10.1038/nature14183.
[14] Z.X. Zhou, S.A. Lujan, A.B. Burkholder, et al., Roles for DNA polymerase δ in initiating and terminating leading strand DNA replication, Nat. Commun. 10 (2019) 1–10, https://doi.org/10.1038/s41467-019-11995-z.
[15] P. Xu, F. Storici, Ribonucleotide incorporation characteristics around yeast autonomously replicating sequences reveal the labor division of replicative DNA polymerases, bioRxiv 2020.08.27.270728, https://doi.org/10.1101/2020.08.27.270728, 2020.
[16] A. Keszthelyi, Y. Daigaku, K. Ptasińska, et al., Mapping ribonucleotides in genomic DNA and exploring replication dynamics by polymerase usage sequencing (Pu-seq), Nat. Protoc. 10 (2015) 1786–1801, https://doi.org/10.1038/nprot.2015.116.
[17] T. Iida, N. Iida, J. Sese, T. Kobayashi, Evaluation of repair activity by quantification of ribonucleotides in the genome, Genes Cells (2021), https://doi.org/10.1111/gtc.12871.
[18] J. Ding, M.S. Taylor, A.P. Jackson, M.A.M. Reijns, Genome-wide mapping of embedded ribonucleotides and other noncanonical nucleotides using emRiboSeq and EndoSeq, Nat. Protoc. 10 (2015) 1433–1444, https://doi.org/10.1038/nprot.2015.099.
[19] A.L. Gombolay, F. Storici, Mapping ribonucleotides embedded in genomic DNA to single-nucleotide resolution using Ribose-Map, bioRxiv 2020.08.27.267153, https://doi.org/10.1101/2020.08.27.267153, 2020.
[20] T. Vincze, J. Posfai, R.J. Roberts, NEBcutter, in: NEBcutter: a Program to Cleave DNA with Restriction Enzymes, in: Nucleic Acids Re., vol. 31, 2003.
[21] A. Keszthelyi, Y. Daigaku, K. Ptasińska, et al., Mapping ribonucleotides in genomic DNA and exploring replication dynamics by polymerase usage sequencing (Pu-seq), Nat. Protoc. 10 (2015) 1786–1801, https://doi.org/10.1038/nprot.2015.116.