**Supplementary Material**

**S1. Experimental information**

Table S1. Sequence patterns used in DNA data storage experiments

|  |  |
| --- | --- |
| **Type** | **Sequence pattern** |
| Adapter  (Forward, 26nt) | GTTCAGAGTTCTACAGTCCGACGATC |
| Adapter  (Reverse, 21nt) | GGAATTCTCGGGTGCCAAGG |
| Primer (RP1) | 5' AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA 3' |
| Primer (RP1I) | 5' CAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA TTGGCACCCGAGAATTCCA 3' |
| Index adapter | ATCACG |

**∙ Synthesis**

18,000 oligo sequences are produced with a length of 199nt in the encoding process.

Oligo synthesis was requested from Twist Bioscience and a 300ng of oligo pool was synthesized.

**∙ PCR**

The lyophilized PCR sample was diluted to 25 ng/µl by nuclease-free water (Invitrogen).

PCR was performed with the below conditions and thermocycling process.

After amplification, the PCR sample was purified using 50 µl (1X) of AMPure XP (Beckman Coulter) and eluted with 20 µl of nuclease-free water (Invitrogen).

Table S2. PCR experiment conditions

|  |  |
| --- | --- |
| Oligo pool sample (25ng/ul) | 1µl (25ng) |
| RP1 primer(10uM) | 2.5µl |
| RPI1 primer(10uM) | 2.5µl |
| Q5 master mix\* | 25µl |
| Nuclease free water | 19µl |
| Total | 50µl |

\* Q5® Hot Start High-Fidelity 2X Master Mix

Table S3. Thermocycling cycles of PCR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 98°C | 98°C | 60°C | 72°C | 72°C |
| 30sec | 10sec | 30sec | 30sec | 5min |

\* The highlighted area is repeated for 10 cycles

**∙ Sequencing**

The PCR sample and Phi-X control v3 were denatured and diluted according to the Illumina guide to prepare a 10 pM library.

MiSeq Reagent Kit v3 (600-cycle) was used in sequencing with a 30% Phi-X control sample spike-in.

**S2. Calculate errors of a read**

We calculated the error rates to provide the characteristics of the DNA data storage channel. We computed the edit distances between all the oligo sequences and each read to estimate the errors using the edit distance algorithm. Among them, an oligo sequence with the smallest edit distance was regarded as the true source of the read. We randomly selected one if multiple oligo sequences have the same edit distance as one read. This matching process was executed for each of forward/reverse direction, and the merged reads. It sometimes happens that the forward/reverse direction and the merged read are not matched to the same oligo sequence. In this case, the edit distances were recalculated by selecting the oligo sequence with the smallest edit distance from the three types.

The base error rate is calculated by dividing the total number of errors by the product of the number of all reads and the original length. In both types of reads, all the error rates are lowest around the original length. The NPF reads show higher error rates than the PF reads and could be seen useless but the decoding experiment will show that the extra use of NPF reads helps generate more error-free sequences.

Table S4. Length distribution and average base error rate of merged reads

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Length (nt)** | **PF** | | | | **NPF** | | | |
| **Ratio** | **Error rate** | | | **Ratio** | **Error rate** | | |
| **Sub** | **Del** | **Ins** | **Sub** | **Del** | **Ins** |
| **< 140** | 2.87% | 0.0069 | 0.3863 | 0.0009 | 2.65% | 0.0281 | 0.3798 | 0.0023 |
| **140-144** | 0.27% | 0.0111 | 0.0680 | 0.0025 | 0.21% | 0.0539 | 0.0710 | 0.0081 |
| **145** | 0.07% | 0.0116 | 0.0488 | 0.0029 | 0.05% | 0.0443 | 0.0513 | 0.0074 |
| **146** | 0.08% | 0.0099 | 0.0418 | 0.0024 | 0.05% | 0.0461 | 0.0454 | 0.0075 |
| **147** | 0.10% | 0.0090 | 0.0350 | 0.0022 | 0.06% | 0.0463 | 0.0386 | 0.0065 |
| **148** | 0.13% | 0.0066 | 0.0277 | 0.0014 | 0.09% | 0.0339 | 0.0296 | 0.0038 |
| **149** | 0.21% | 0.0040 | 0.0206 | 0.0009 | 0.15% | 0.0390 | 0.0235 | 0.0041 |
| **150** | 0.49% | 0.0023 | 0.0136 | 0.0004 | 0.33% | 0.0269 | 0.0143 | 0.0013 |
| **151** | 2.80% | 0.0011 | 0.0067 | 0.0001 | 1.89% | 0.0238 | 0.0070 | 0.0004 |
| **152\*** | 82.40% | 0.0009 | 0.0000 | 0.0000 | 56.37% | 0.0390 | 0.0003 | 0.0003 |
| **153** | 1.01% | 0.0018 | 0.0003 | 0.0069 | 0.65% | 0.0265 | 0.0008 | 0.0075 |
| **> 153** | 0.62% | 0.1444 | 0.0249 | 0.2075 | 1.50% | 0.1722 | 0.0311 | 0.2075 |
| **Total** | 91.06% | 0.0175 | 0.0561 | 0.0190 | 64.02% | 0.0483 | 0.0577 | 0.0214 |

\* Original length

**S3. Erlich’s sequence analysis[R1]**

As mentioned in the paper, we utilized Erlich's sequence analysis method as Step 1 of our proposed method. Erlich and his team performed length filtering and sequence clustering. First, R1 and R2 reads are selected and merged. If the length of the merged read is not equal to standard length *l*, it is discarded. Only reads with standard length are used in clustering. Erlich’s clustering is the task of making only reads having the same bases into one cluster. After sorting the representative sequences in descending order of cluster size, the sequences are input one by one into the error detector and decoder in this order. Entering the order according to the cluster size is for the reliability of the sequence.

**S4. Simulation results**

**∙ Dataset**

We used 3,071,898 paired-end reads for our simulation. Among them, 2,746,375 were PF reads, and 325,523 were NPF reads. The read data are available as FASTQ files at <https://github.com/PParkJy/SAD-DNAstorage>.

**∙ Edit distance threshold**

In the paper, the edit distance-based clustering methods that we used and proposed allow the user to select the edit distance threshold. We calculated the number of error-free sequences when using reads with lengths ranging from 145 to 153. Sequence analysis was attempted 100 times with a random sampling number of 100,000, and the edit distance that yielded the highest average number of error-free sequences was selected. In both cases—when using only PF reads and when also including NPF reads in the Extra category—the most error-free sequences were obtained when was set to 5 and ​ was set to 4 (shown in Table S5).

Table S5. Comparison of the number of error-free sequences based on the choice of edit distance

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Edit  distance | |  | | | | | | | |
| PF | | | | Extra | | | |
| 2 | 3 | **4** | 5 | 2 | 3 | **4** | 5 |
|  | 2 | 17331.21 | 17331.75 | 17331.95 | 17332.07 | 17409.33 | 17410.92 | 17411.96 | 17412.68 |
| 3 | 17333.34 | 17331.93 | 17334.07 | 17334.19 | 17413.52 | 17413.41 | 17416.14 | 17416.86 |
| 4 | 17333.78 | 17334.32 | 17332.91 | 17334.63 | 17415.35 | 17416.93 | 17416.07 | 17418.68 |
| **5** | 17334.02 | 17334.56 | **17334.75** | 17333.27 | 17416.84 | 17418.4 | **17419.43** | 17418.45 |

**∙ Decoding result**

The number of successful decodings based on performance gain for all scenarios is shown in Tables S5-S7. The performance gain is expressed as a percentage, calculated by subtracting the ratio of the random sampling number that achieved perfect decoding in each scenario from the random sampling number that achieved perfect decoding in the Erlich-PF scenario from 1. The number of perfect decoding trials is 100.

Table S6. Comparison of the number of successful decodings when the performance gain is at its highest

|  |  |  |  |
| --- | --- | --- | --- |
| **Best case** | | | |
| **Random sampling**  **number** | **Erlich-PF** | **Erlich-ExtraNPF** | **Prop-ExtraNPF** |
| 87000 | 0 | 6 | 46 |
| 88000 | 1 | 15 | 54 |
| 89000 | 2 | 29 | 66 |
| 90000 | 9 | 46 | 76 |
| 91000 | 20 | 58 | 82 |
| 92000 | 25 | 67 | 85 |
| 93000 | 38 | 70 | 91 |
| 94000 | 49 | 79 | 93 |
| 95000 | 62 | 86 | 99 |
| **96000** | 72 | 93 | **100** |
| 97000 | 82 | 96 | 100 |
| 98000 | 88 | 97 | 100 |
| 99000 | 92 | 99 | 100 |
| **100000** | 92 | **100** | 100 |
| 101000 | 96 | 100 | 100 |
| 102000 | 98 | 100 | 100 |
| 103000 | 98 | 100 | 100 |
| 104000 | 98 | 100 | 100 |
| 105000 | 99 | 100 | 100 |
| 106000 | 99 | 100 | 100 |
| **107000** | **100** | 100 | 100 |

Table S7. Comparison of the number of successful decodings when the performance gain is average

|  |  |  |  |
| --- | --- | --- | --- |
| **Average case** | | | |
| **Random**  **Sampling**  **Number** | **Erlich-PF** | **Erlich-ExtraNPF** | **Prop-ExtraNPF** |
| 87000 | 0 | 3 | 32 |
| 88000 | 0 | 6 | 48 |
| 89000 | 1 | 23 | 66 |
| 90000 | 8 | 38 | 81 |
| 91000 | 19 | 48 | 89 |
| 92000 | 26 | 62 | 90 |
| 93000 | 45 | 74 | 95 |
| 94000 | 61 | 85 | 96 |
| 95000 | 70 | 90 | 97 |
| 96000 | 83 | 94 | 98 |
| 97000 | 90 | 95 | 98 |
| 98000 | 91 | 95 | 98 |
| 99000 | 92 | 95 | 98 |
| 100000 | 92 | 95 | 98 |
| 101000 | 93 | 95 | 98 |
| 102000 | 95 | 96 | 98 |
| 103000 | 96 | 96 | 98 |
| **104000** | 98 | 98 | **100** |
| 105000 | 99 | 99 | 100 |
| **106000** | 99 | **100** | 100 |
| 107000 | 99 | 100 | 100 |
| 108000 | 99 | 100 | 100 |
| 109000 | 99 | 100 | 100 |
| **110000** | **100** | 100 | 100 |

Table S8. Comparison of the number of successful decodings when the performance gain is low

|  |  |  |  |
| --- | --- | --- | --- |
| **Worst case** | | | |
| **Random**  **Sampling**  **Number** | **Erlich-PF** | **Erlich-ExtraNPF** | **Prop-ExtraNPF** |
| 87000 | 0 | 7 | 40 |
| 88000 | 1 | 10 | 53 |
| 89000 | 4 | 24 | 69 |
| 90000 | 8 | 39 | 79 |
| 91000 | 20 | 56 | 84 |
| 92000 | 27 | 67 | 87 |
| 93000 | 40 | 76 | 91 |
| 94000 | 58 | 82 | 93 |
| 95000 | 69 | 88 | 96 |
| 96000 | 72 | 92 | 97 |
| 97000 | 82 | 94 | 97 |
| 98000 | 91 | 96 | 98 |
| 99000 | 95 | 97 | 98 |
| 100000 | 97 | 97 | 98 |
| 101000 | 98 | 98 | 99 |
| **102000** | 99 | 99 | **100** |
| 103000 | 99 | 99 | 100 |
| 104000 | 99 | 99 | 100 |
| 105000 | 99 | 99 | 100 |
| 106000 | 99 | 99 | 100 |
| 107000 | 99 | 99 | 100 |
| 108000 | 99 | 99 | 100 |
| **109000** | **100** | **100** | 100 |

Tables S4 and S6 are included as Figures 6 and 7 in the paper, respectively, along with the performance gain. The figure for the average case in Table S7 corresponds to Figure S1 below.

텍스트, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

Figure S1. Comparison of decoding results for random sampling numbers (Average case)

**∙ Sequence analysis result**

Tables 2 and 3 in the paper present the sequence analysis results for the best case. These results are represented in detail in Tables S9 and S10 below. The results for the average and worst cases are summarized in Tables S11-S14. The random sampling numbers used as the basis for the result analysis represent the number of times the proposed method achieved perfect decoding.

Table S9. Result of Erlich’s sequence analysis at random sampling number 100,000 (Best case)

|  |  |  |  |
| --- | --- | --- | --- |
| **RSN = 100,000** | | **PF** | **PF + NPF** |
| **Step 1** | Number of merged reads | 81424.74 | 88218.10 |
| Number of error-free sequences (S1) | 17238.25 | 17299.99 |
| Number of error clusters | 7140.83 | 11023.53 |
| Ratio of error-free sequences (S1) to all clusters | 70.71% | 61.08% |
| Ratio of error clusters to all clusters | 29.29% | 38.92% |
| Error-free sequence gain compared to using only PF reads |  | 67.24 |
|  | Ratio of error-free sequence gain compared to using only PF reads | 0.36% |

Table 10. Result of proposed sequence analysis at random sampling number 96,000 (Best case)

|  |  |  |  |
| --- | --- | --- | --- |
| **RSN = 96,000** | | **PF** | **PF + NPF** |
| **Step 1** | Number of merged reads | 78169.33 | 84690.04 |
| Number of error-free sequences (S1) | 17150.5 | 17217.74 |
| Number of error clusters | 6854.41 | 10580.04 |
| Ratio of error-free sequences (S1) to all clusters | 71.45% | 61.94% |
| Ratio of error clusters to all clusters | 28.55% | 38.06% |
| Error-free sequence gain compared to using only PF reads |  | 41.16 |
| Ratio of error-free sequence gain compared to using only PF reads | 0.39% |
| **Step 2** | Number of EDOL reads before clustering |  | 10580.04 |
| Number of all clusters | 8694.24 |
| Number of clusters of size 2 or larger | 1563.62 |
| Number of corrected clusters by CAPMB consensus | 1544.97 |
| Number of uncorrected clusters by CAPMB consensus | 18.65 |
| Number of reads in uncorrected clusters | 37.47 |
| Number of duplicated clusters with S1 | 1503.81 |
| Number of error-free clusters (S2) (duplications were removed) | 41.16 |
| Ratio of clusters of size 2 or larger to all clusters | 17.99% |
| Ratio of corrected clusters to size 2 or larger clusters | 98.81% |
| Ratio of error-free clusters in corrected clusters | 2.66% |
| Ratio of uncorrected clusters to size 2 or larger clusters | 1.19% |
| Ratio of error-free sequence gain compared to using only PF reads | 0.24% |
| **Step 3** | Number of REDOL reads (RREDOL) before clustering |  | 7168.53 |
| Number of AL reads (RAL) before clustering | 4537.6 |
| Number of all clusters | 10375.59 |
| Number of clusters of size 2 or larger | 1162.8 |
| Number of corrected clusters by CAPMB consensus | 1143.51 |
| Number of uncorrected clusters by CAPMB consensus | 19.29 |
| Number of reads in uncorrected clusters | 38.69 |
| Number of duplicated clusters with S1 and S2 | 1111.18 |
| Number of error-free clusters (S3) (duplications were removed) | 32.33 |
| Ratio of clusters of size 2 or larger to all clusters | 11.21% |
| Ratio of corrected clusters to size 2 or larger clusters | 98.34% |
| Ratio of error-free clusters in corrected clusters | 2.83% |
| Ratio of uncorrected clusters to size 2 or larger clusters | 1.66% |
| Ratio of error-free sequence gain compared to using only PF reads | 0.19% |

Table 11. Result of Erlich’s sequence analysis at random sampling number 106,000 (Average case)

|  |  |  |  |
| --- | --- | --- | --- |
| **RSN = 106,000** | | **PF** | **PF + NPF** |
| **Step 1** | Number of merged reads | 86324.68 | 93530.88 |
| Number of error-free sequences (S1) | 17349.73 | 17404.22 |
| Number of error clusters | 7557.01 | 11674.83 |
| Ratio of error-free sequences (S1) to all clusters | 69.66% | 59.85% |
| Ratio of error clusters to all clusters | 30.34% | 40.15% |
| Error-free sequence gain compared to using only PF reads |  | 54.49 |
|  | Ratio of error-free sequence gain compared to using only PF reads | 0.31% |

Table 12. Result of proposed sequence analysis at random sampling number 104,000 (Average case)

|  |  |  |  |
| --- | --- | --- | --- |
| **RSN = 104,000** | | **PF** | **PF + NPF** |
| **Step 1** | Number of merged reads | 86324.68 | 93530.88 |
| Number of error-free sequences (S1) | 17349.73 | 17404.22 |
| Number of error clusters | 7557.01 | 11674.83 |
| Ratio of error-free sequences (S1) to all clusters | 69.66% | 59.85% |
| Ratio of error clusters to all clusters | 30.34% | 40.15% |
| Error-free sequence gain compared to using only PF reads |  | 56.67 |
| Ratio of error-free sequence gain compared to using only PF reads | 0.33% |
| **Step 2** | Number of EDOL reads before clustering |  | 11455.19 |
| Number of all clusters | 9271.66 |
| Number of clusters of size 2 or larger | 1782.47 |
| Number of corrected clusters by CAPMB consensus | 1761.86 |
| Number of uncorrected clusters by CAPMB consensus | 20.61 |
| Number of reads in uncorrected clusters | 41.37 |
| Number of duplicated clusters with S1 | 1727.08 |
| Number of error-free clusters (S2) (duplications were removed) | 34.78 |
| Ratio of clusters of size 2 or larger to all clusters | 19.23% |
| Ratio of corrected clusters to size 2 or larger clusters | 98.84% |
| Ratio of error-free clusters in corrected clusters | 1.97% |
| Ratio of uncorrected clusters to size 2 or larger clusters | 1.16% |
| Ratio of error-free sequence gain compared to using only PF reads | 0.20% |
| **Step 3** | Number of REDOL reads (RREDOL) before clustering |  | 7530.92 |
| Number of AL reads (RAL) before clustering | 4921.52 |
| Number of all clusters | 10958.76 |
| Number of clusters of size 2 or larger | 1293.53 |
| Number of corrected clusters by CAPMB consensus | 1270.74 |
| Number of uncorrected clusters by CAPMB consensus | 22.79 |
| Number of reads in uncorrected clusters | 45.78 |
| Number of duplicated clusters with S1 and S2 | 1243.1 |
| Number of error-free clusters (S3) (duplications were removed) | 27.64 |
| Ratio of clusters of size 2 or larger to all clusters | 11.80% |
| Ratio of corrected clusters to size 2 or larger clusters | 98.24% |
| Ratio of error-free clusters in corrected clusters | 2.18% |
| Ratio of uncorrected clusters to size 2 or larger clusters | 1.76% |
| Ratio of error-free sequence gain compared to using only PF reads | 0.16% |

Table 13. Result of Erlich’s sequence analysis at random sampling number 109,000 (Worst case)

|  |  |  |  |
| --- | --- | --- | --- |
| **RSN = 109,000** | | **PF** | **PF + NPF** |
| **Step 1** | Number of merged reads | 88749.71 | 96154.48 |
| Number of error-free sequences (S1) | 17399.71 | 17449.89 |
| Number of error clusters | 7765.19 | 11989.31 |
| Ratio of error-free sequences (S1) to all clusters | 69.14% | 59.27% |
| Ratio of error clusters to all clusters | 30.86% | 40.73% |
| Error-free sequence gain compared to using only PF reads |  | 50.18 |
|  | Ratio of error-free sequence gain compared to using only PF reads | 0.29% |

Table 14. Result of proposed sequence analysis at random sampling number 102,000 (Worst case)

|  |  |  |  |
| --- | --- | --- | --- |
| **RSN = 102,000** | | **PF** | **PF + NPF** |
| **Step 1** | Number of merged reads | 83052.14 | 89981.94 |
| Number of error-free sequences (S1) | 17278.28 | 17336.41 |
| Number of error clusters | 7266.24 | 11221.46 |
| Ratio of error-free sequences (S1) to all clusters | 70.40% | 60.71% |
| Ratio of error clusters to all clusters | 29.60% | 39.29% |
| Error-free sequence gain compared to using only PF reads |  | 58.13 |
| Ratio of error-free sequence gain compared to using only PF reads | 0.34% |
| **Step 2** | Number of EDOL reads before clustering |  | 11221.46 |
| Number of all clusters | 9119.07 |
| Number of clusters of size 2 or larger | 1722.96 |
| Number of corrected clusters by CAPMB consensus | 1704.04 |
| Number of uncorrected clusters by CAPMB consensus | 18.92 |
| Number of reads in uncorrected clusters | 37.99 |
| Number of duplicated clusters with S1 | 1666.65 |
| Number of error-free clusters (S2) (duplications were removed) | 37.39 |
| Ratio of clusters of size 2 or larger to all clusters | 18.89% |
| Ratio of corrected clusters to size 2 or larger clusters | 98.90% |
| Ratio of error-free clusters in corrected clusters | 2.19% |
| Ratio of uncorrected clusters to size 2 or larger clusters | 1.10% |
| Ratio of error-free sequence gain compared to using only PF reads | 0.22% |
| **Step 3** | Number of REDOL reads (RREDOL) before clustering |  | 7434.34 |
| Number of AL reads (RAL) before clustering | 4813.24 |
| Number of all clusters | 10800.55 |
| Number of clusters of size 2 or larger | 1257.92 |
| Number of corrected clusters by CAPMB consensus | 1234.52 |
| Number of uncorrected clusters by CAPMB consensus | 23.4 |
| Number of reads in uncorrected clusters | 46.95 |
| Number of duplicated clusters with S1 and S2 | 1206.02 |
| Number of error-free clusters (S3) (duplications were removed) | 28.5 |
| Ratio of clusters of size 2 or larger to all clusters | 11.65% |
| Ratio of corrected clusters to size 2 or larger clusters | 98.14% |
| Ratio of error-free clusters in corrected clusters | 2.31% |
| Ratio of uncorrected clusters to size 2 or larger clusters | 1.86% |
| Ratio of error-free sequence gain compared to using only PF reads | 0.16% |

**S5. DNA data storage characteristics**

To show the characteristics of DNA data storage, we calculated the distribution of coverage as shown in Figure S2. First, we performed a random sampling of 107,000 reads and conducted the preprocessing described in Section 2.3.1 of the paper. The random sampling number is the number that achieved perfect decoding in the best case of Erlich-PF scenario (Section S4). Afterward, we matched the merged reads to the oligo sequences following the same method outlined in the S2 section. Coverage is calculated by determining how many times a given oligo sequence was sequenced. In the case of 107,000 reads and 18,000 oligo sequences, the ideal coverage distribution forms an ideal Poisson distribution with a mean of approximately 5.94. As shown in Figure S2, both PF and Extra exhibit a leftward shift, meaning lower coverage compared to the ideal distribution. The primary reason is that some reads fail to merge during the paired-end read merging process in preprocessing. (Approximately 9% of PF reads and 36% of NPF reads fail to merge.) Nevertheless, Extra which uses both PF and NPF reads was closer to the ideal coverage distribution compared to PF alone.

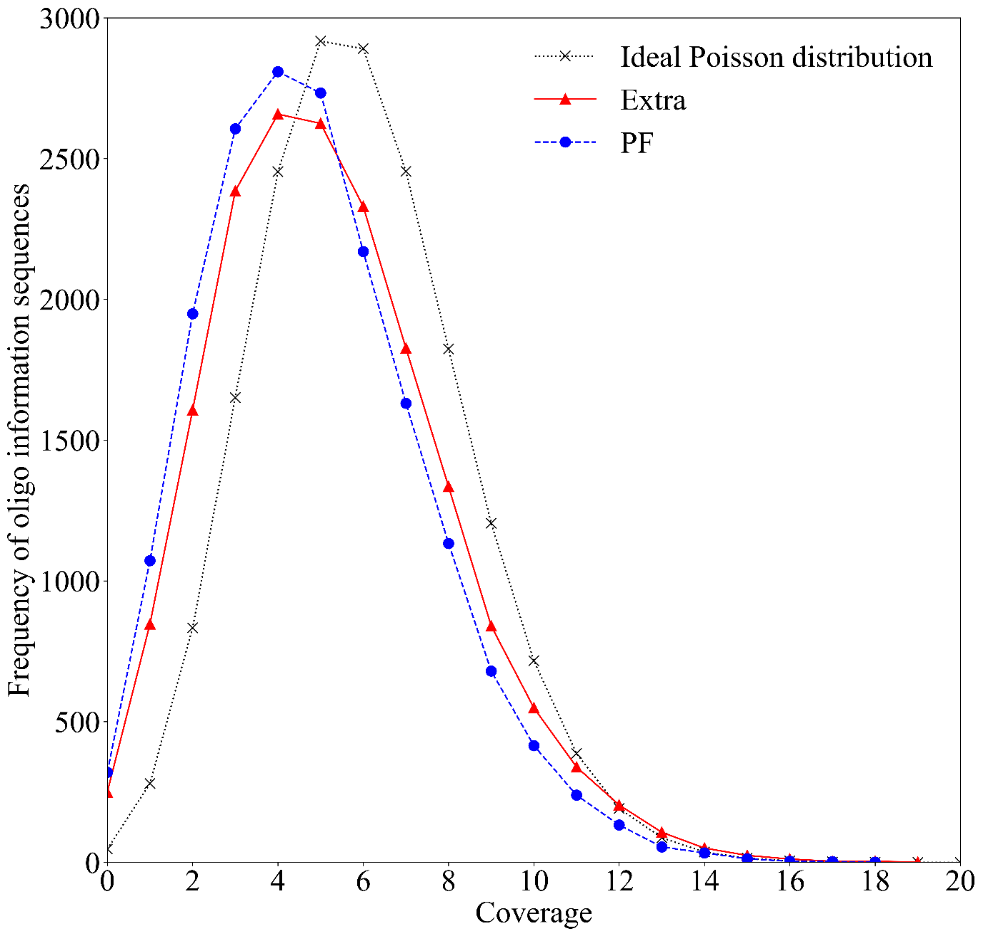
****

Figure S2. Comparison of the sequence coverage distribution for 18,000 oligo sequences from 107,000 randomly sampled reads to the ideal Poisson distribution

**Reference**

[R1] Yaniv Erlich, and Dina Zielinski. "DNA Fountain enables a robust and efficient storage architecture." *science* 355.6328 (2017): 950-954.