Science Advances
AAAS

Supplementary Materials for

Genetic physical unclonable functions in human cells

Yi Li, Mohammad Mahdi Bidmeshki, Taek Kang, Chance M. Nowak,
Yiorgos Makris*, Leonidas Bleris*

*Corresponding author. Email: bleris@utdallas.edu (L.B.); yiorgos.makris@utdallas.edu (Y.M.)

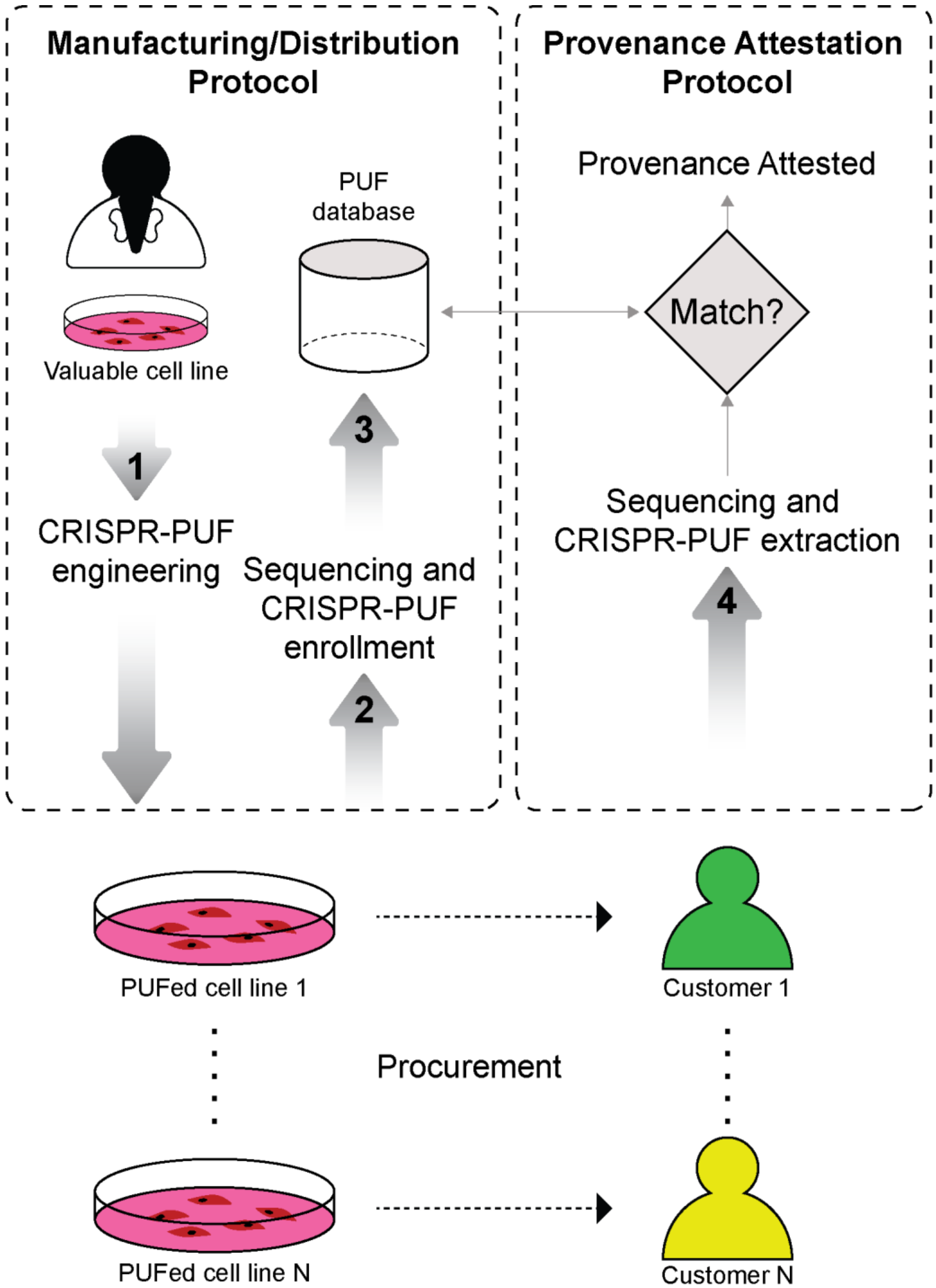
Published 4 May 2022, *Sci. Adv.* **8**, eabm4106 (2022)
DOI: 10.1126/sciadv.abm4106

The PDF file includes:

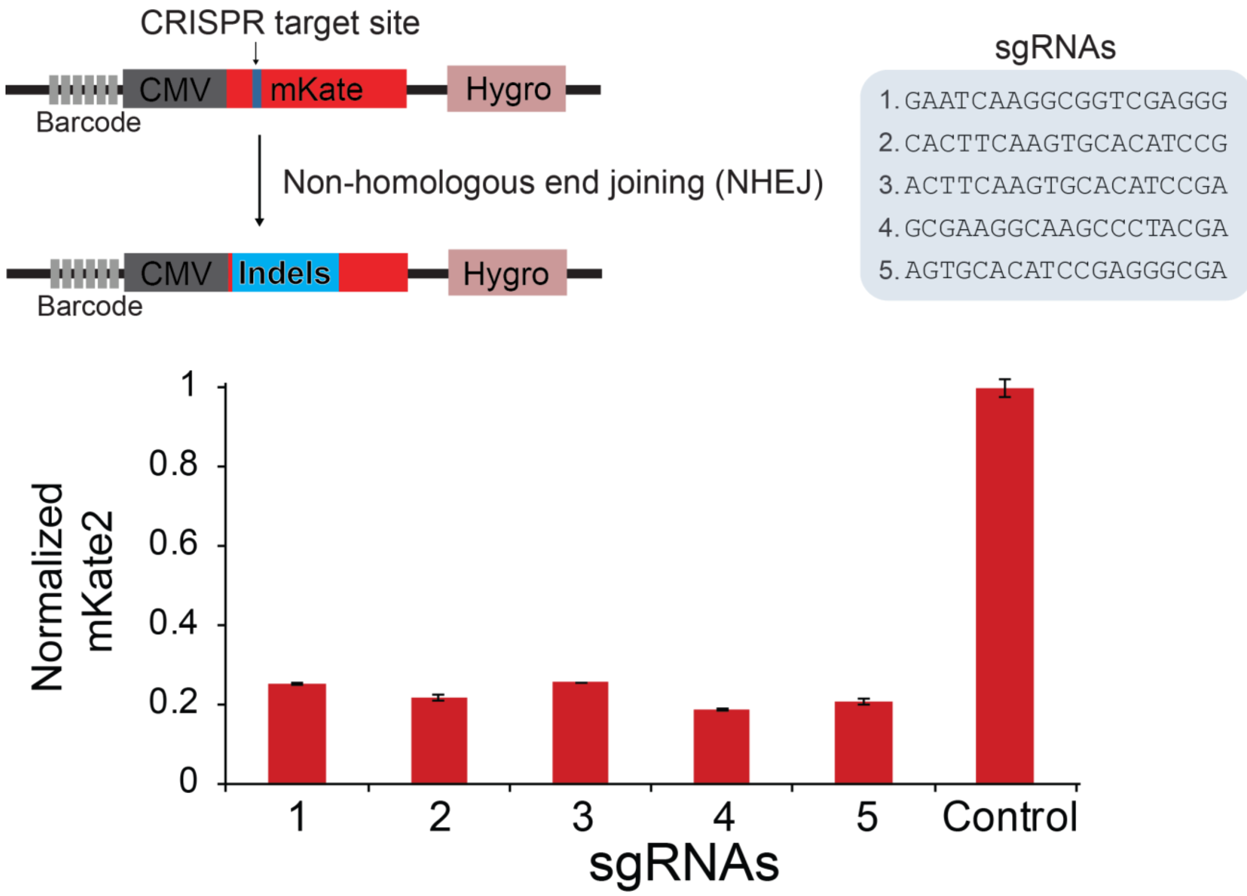
Figs. S1 to S21
Supplementary Text
Legends for tables S1 to S17
Legends for scripts S1 and S2

Other Supplementary Material for this manuscript includes the following:

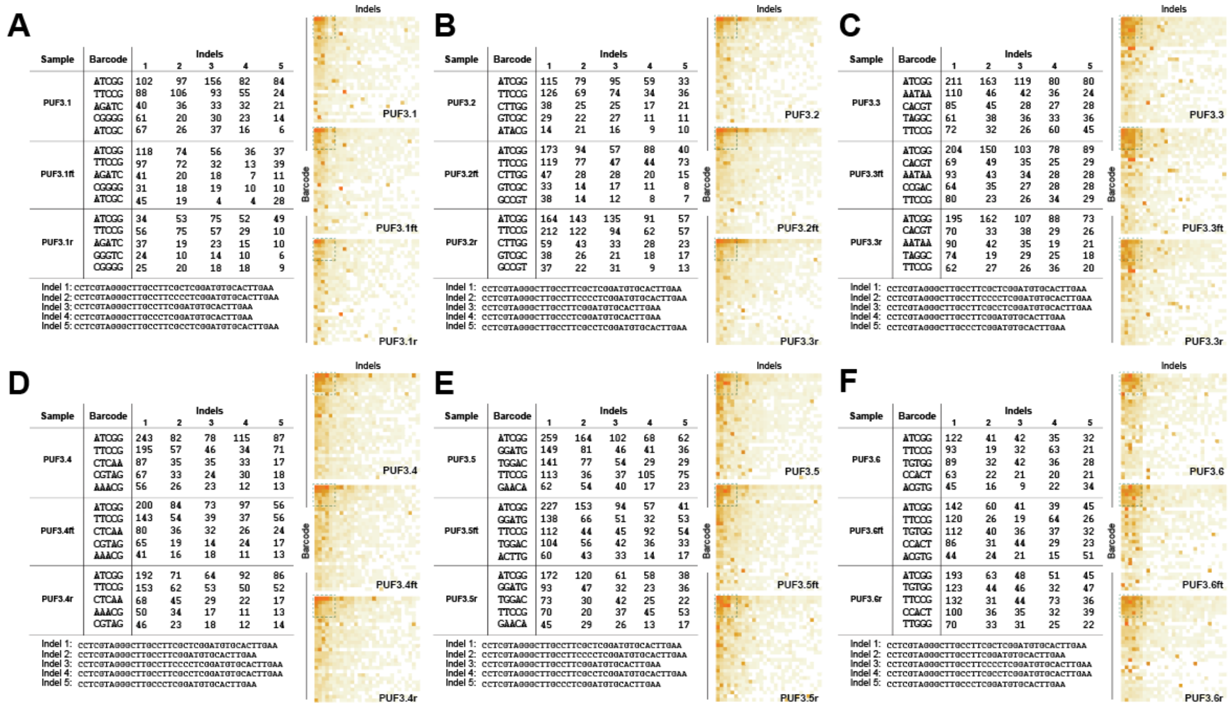
Tables S1 to S17
Scripts S1 and S2



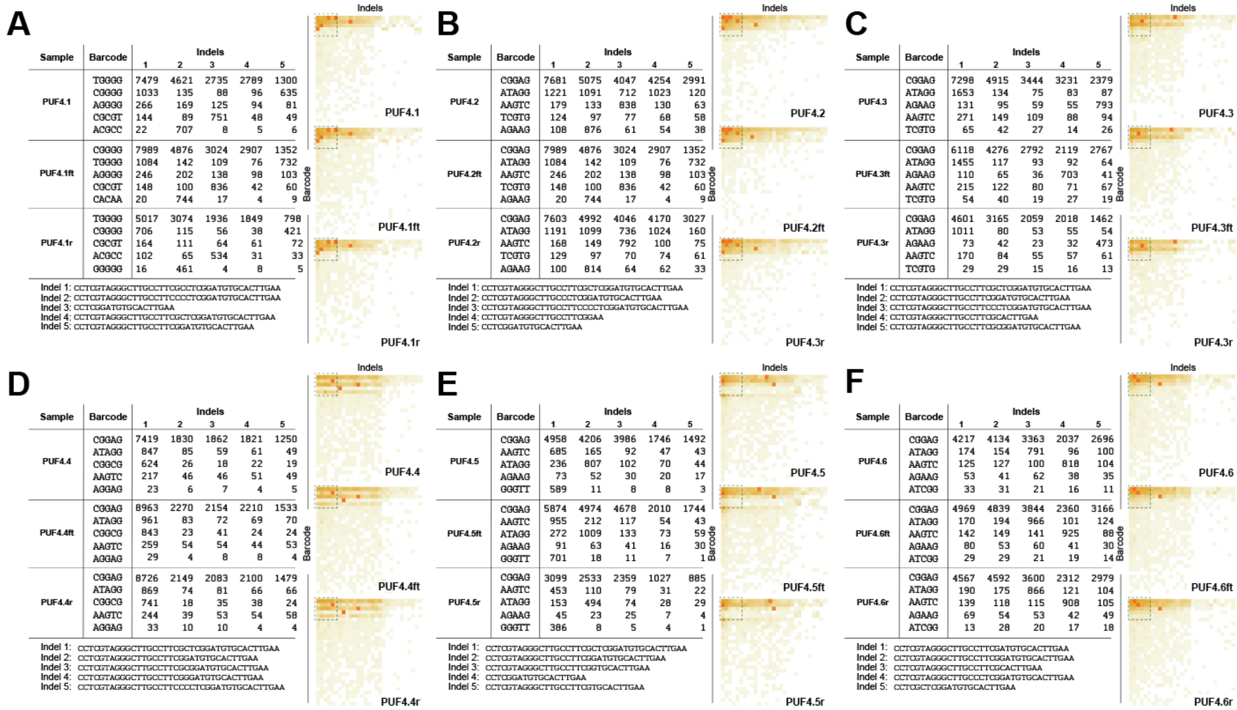
Supplementary Figure 1. Provenance attestation protocols and pilot CRISPR-PUF. (A) The producer of a valuable cell line inserts a unique, robust and unclonable signature in each legitimately produced copy of this cell line. Upon thawing of a frozen sample and prior to its initial use, a customer who purchased a copy of the cell line can obtain this signature and communicate it to the producer who compares it against the signature database of legitimately produced copies of this cell line and, thereby, attests its provenance.



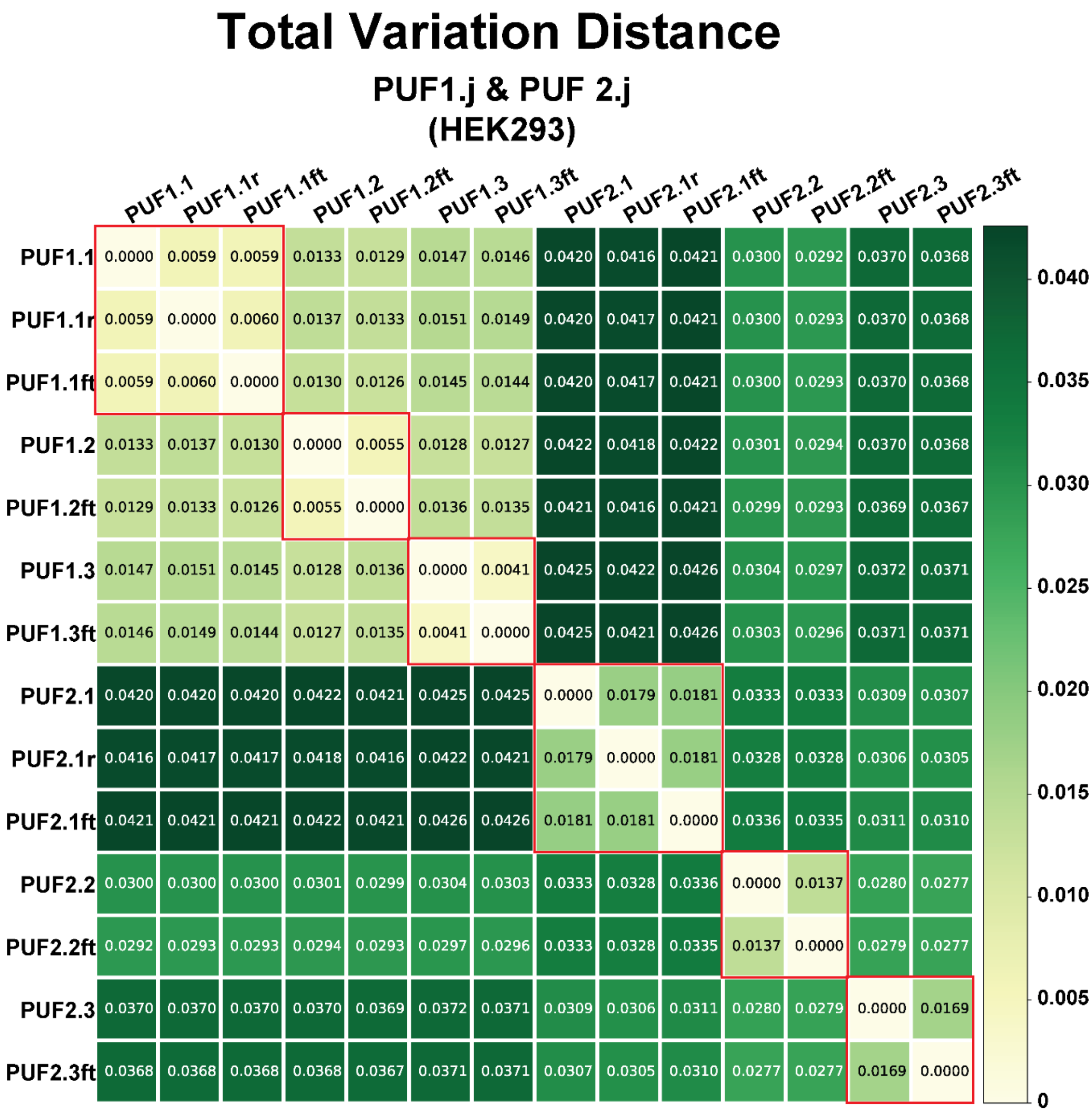
Supplementary Figure 2. Implementation of CRISPR-PUFs in HEK293 cells. Five sgRNAs were designed to target the Open Reading Frames (ORFs) of the mKate2 construct, and demonstrated comparable efficiencies using *in vitro* fluorescence reporter assays.



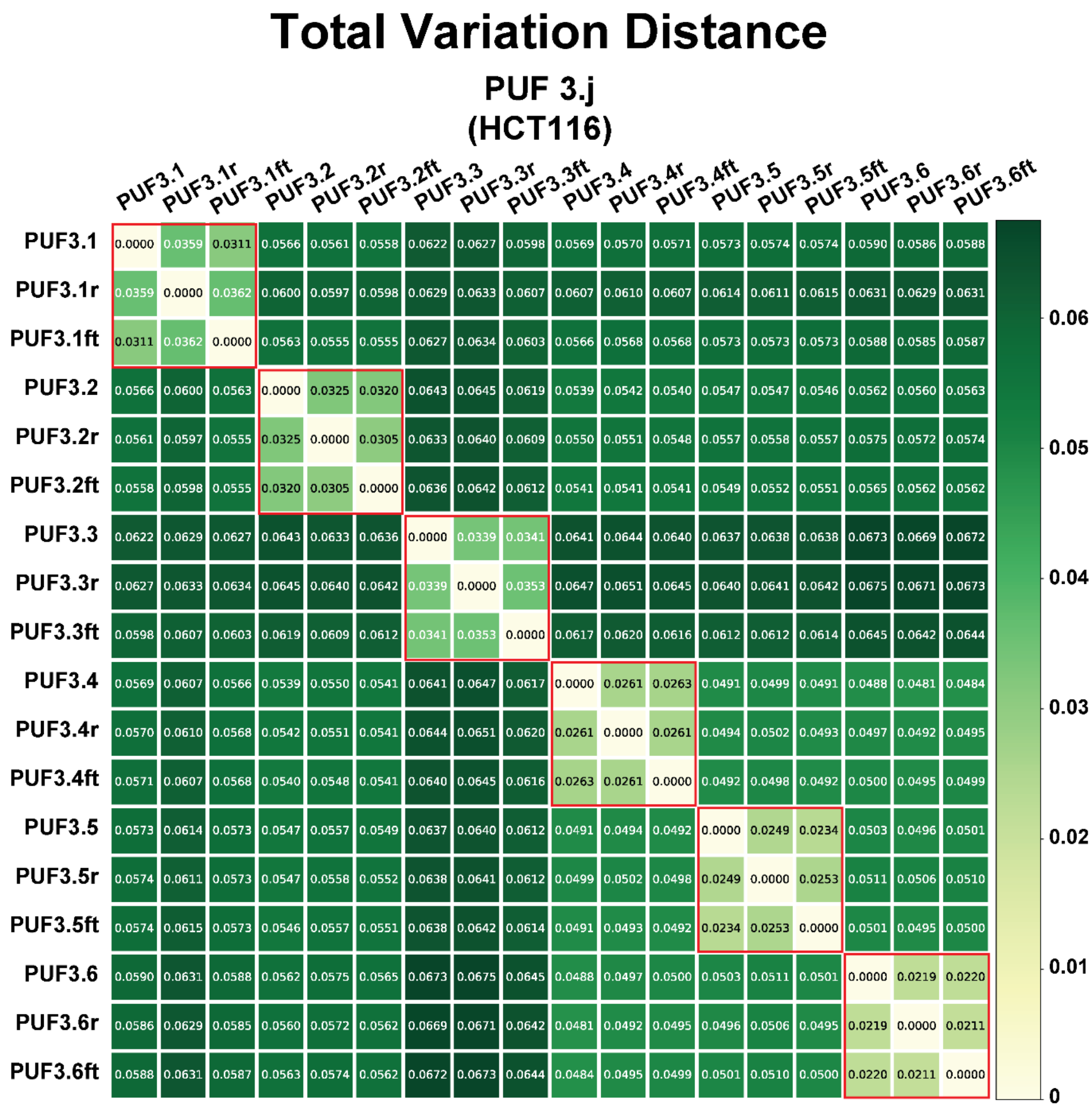
Supplementary Figure 3. Implementation of CRISPR-PUFs in HCT116 cells. Qualitative assessment of CRISPR-PUFs generated using HCT116. (A~E) Frequencies of barcode-indel addresses consisting of the 5 most commonly observed barcodes and indels (Left) and heatmap based on the same data but expanded to the top 30 most commonly observed barcodes and indels (Right) for a given PUF and its freeze-thaw counterparts and technical replicates. The green dashed square on the heatmap represents the data shown on the table. Data shown in (A) are barcode-indel addresses for PUF3.1 with their respective freeze-thaw counterpart and technical replicate. Data shown in (B~E) are for PUFs 3.2 to 3.6, respectively, which are produced identically to PUF3.1 using the same barcoded cell line and same sgRNA to introduce indels.



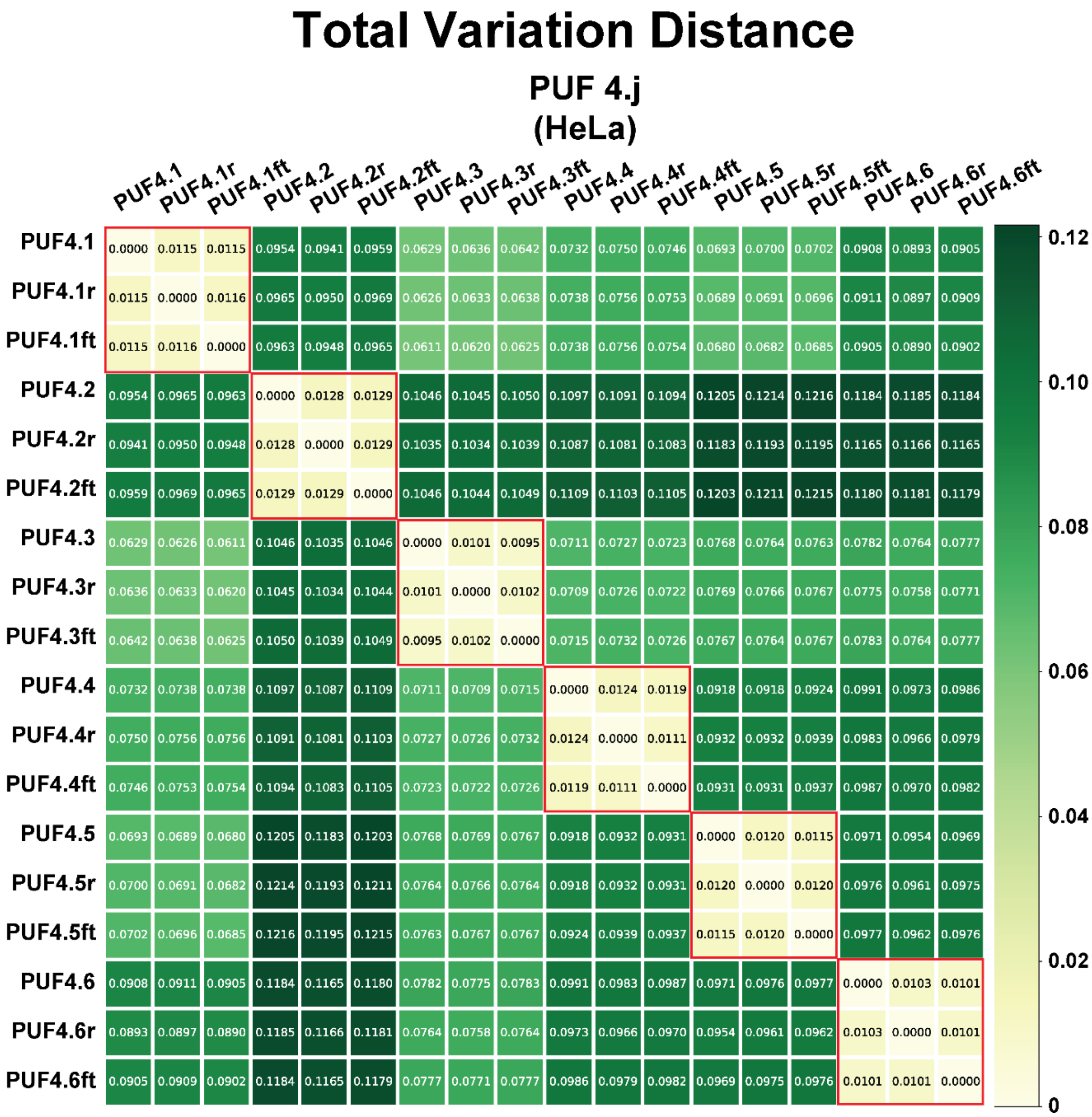
Supplementary Figure 4. Implementation of CRISPR-PUFs in HeLa cells. Qualitative assessment of CRISPR-PUFs generated using HeLa. See Supplementary Figure 2 for detailed description.



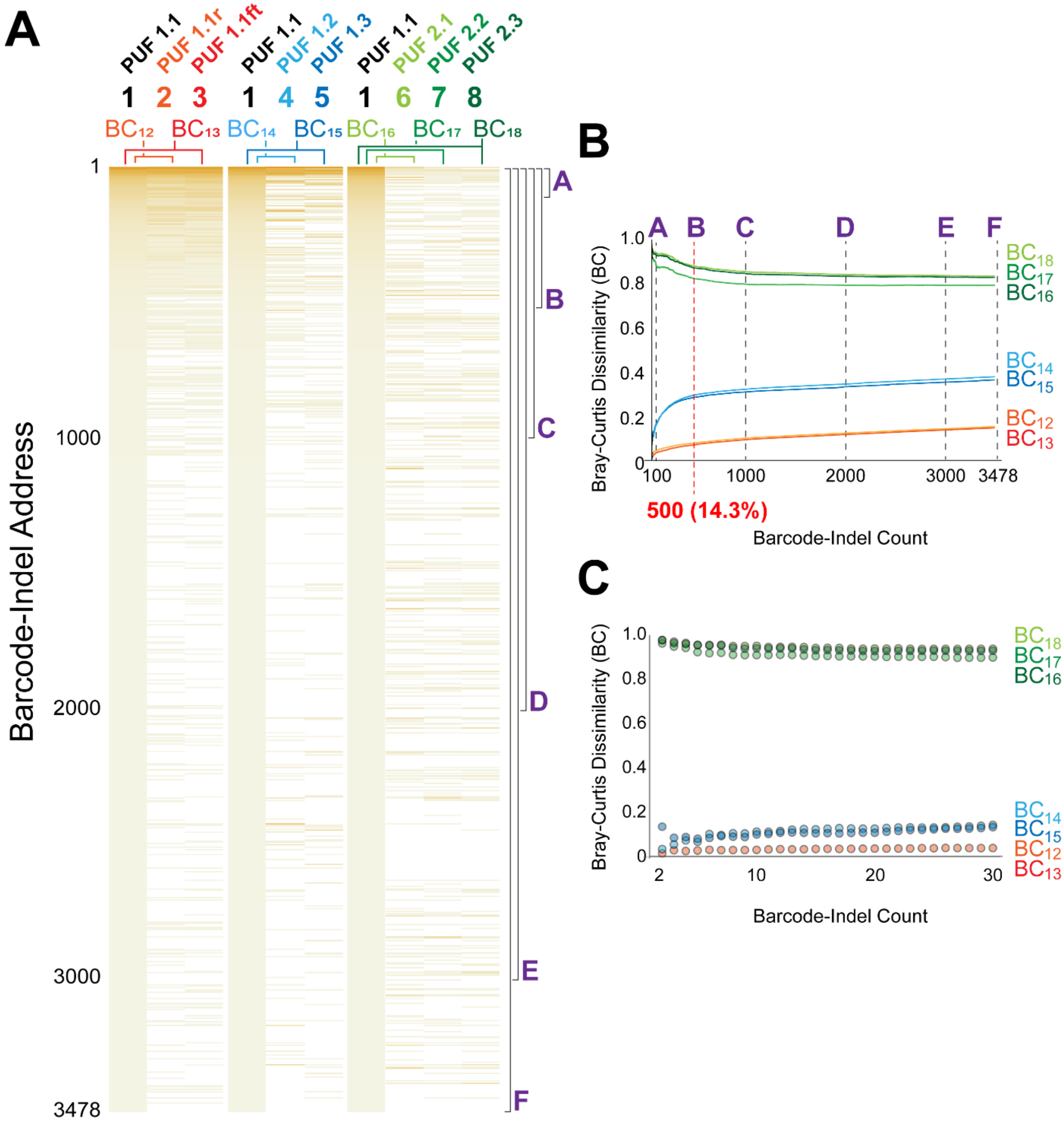
Supplementary Figure 5. Quantitative assessment of CRISPR-PUFs using total variation distance. Pairwise Total Variation Distances between all PUFs were calculated in samples derived from HEK293 cells.



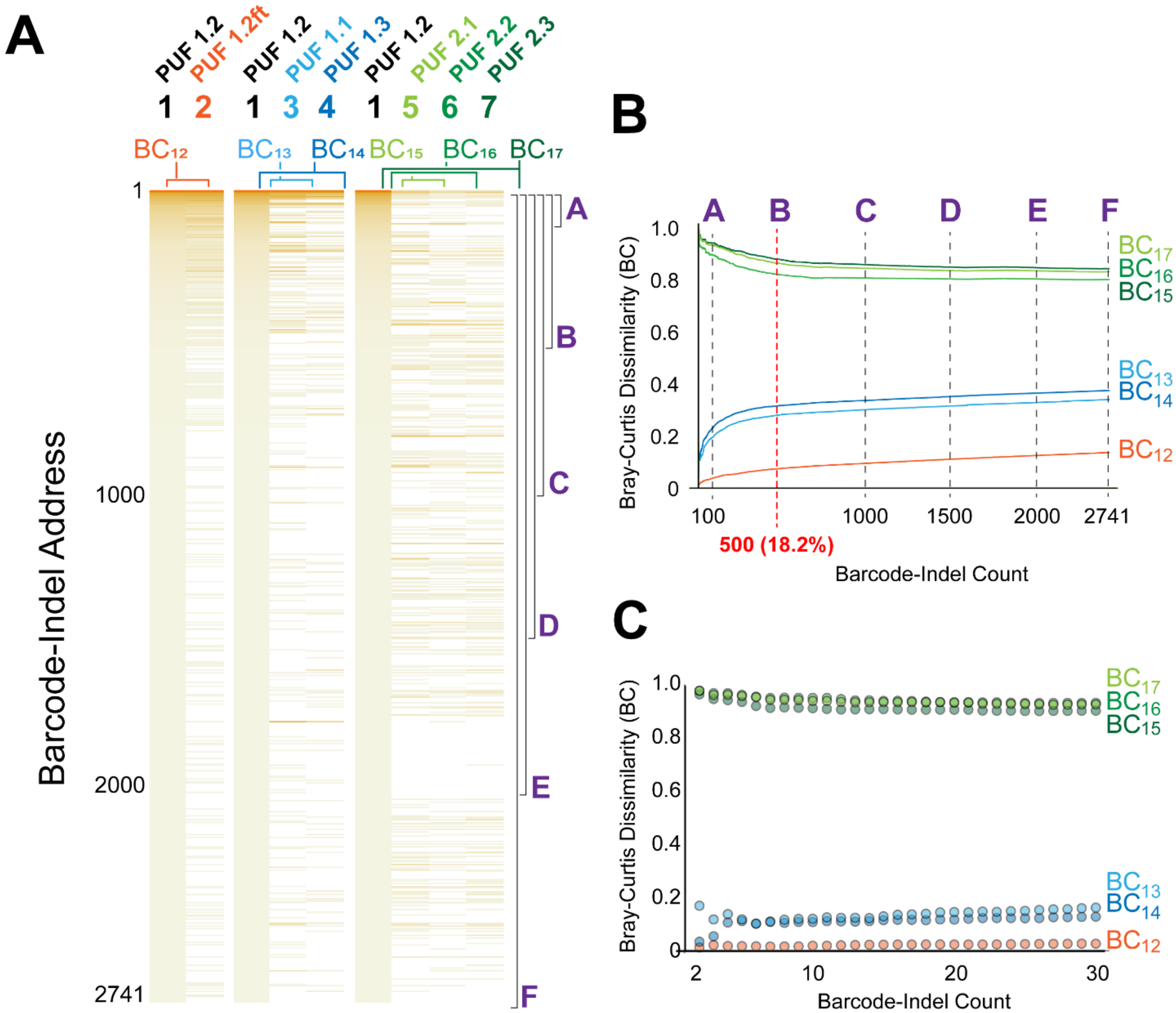
Supplementary Figure 6. Quantitative assessment of CRISPR-PUFs using total variation distance. Pairwise Total Variation Distances between all PUFs were calculated in samples derived from HCT116 cells.



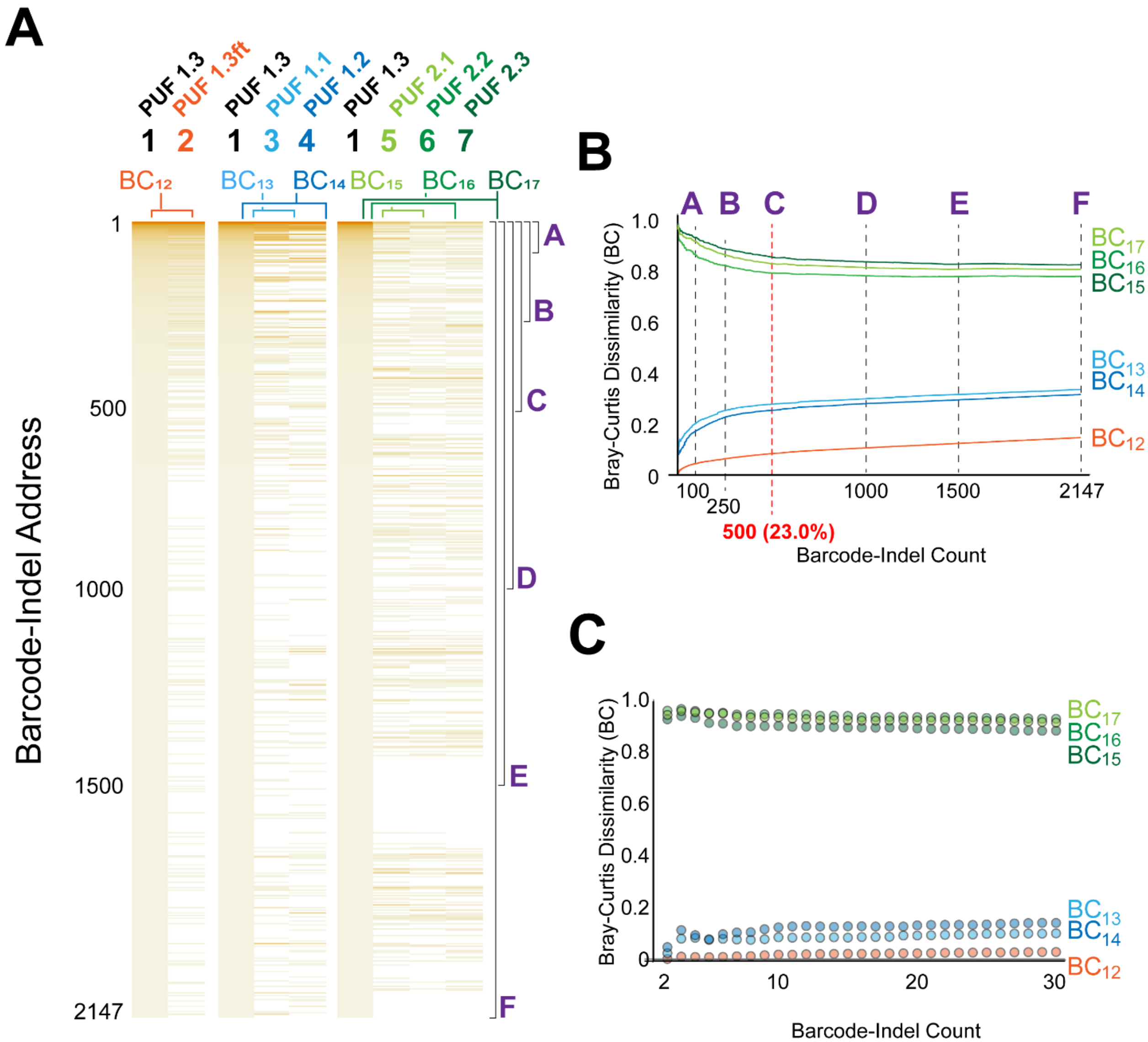
Supplementary Figure 7. Quantitative assessment of CRISPR-PUFs using total variation distance. Pairwise Total Variation Distances between all PUFs were calculated in samples derived from HeLa cells.



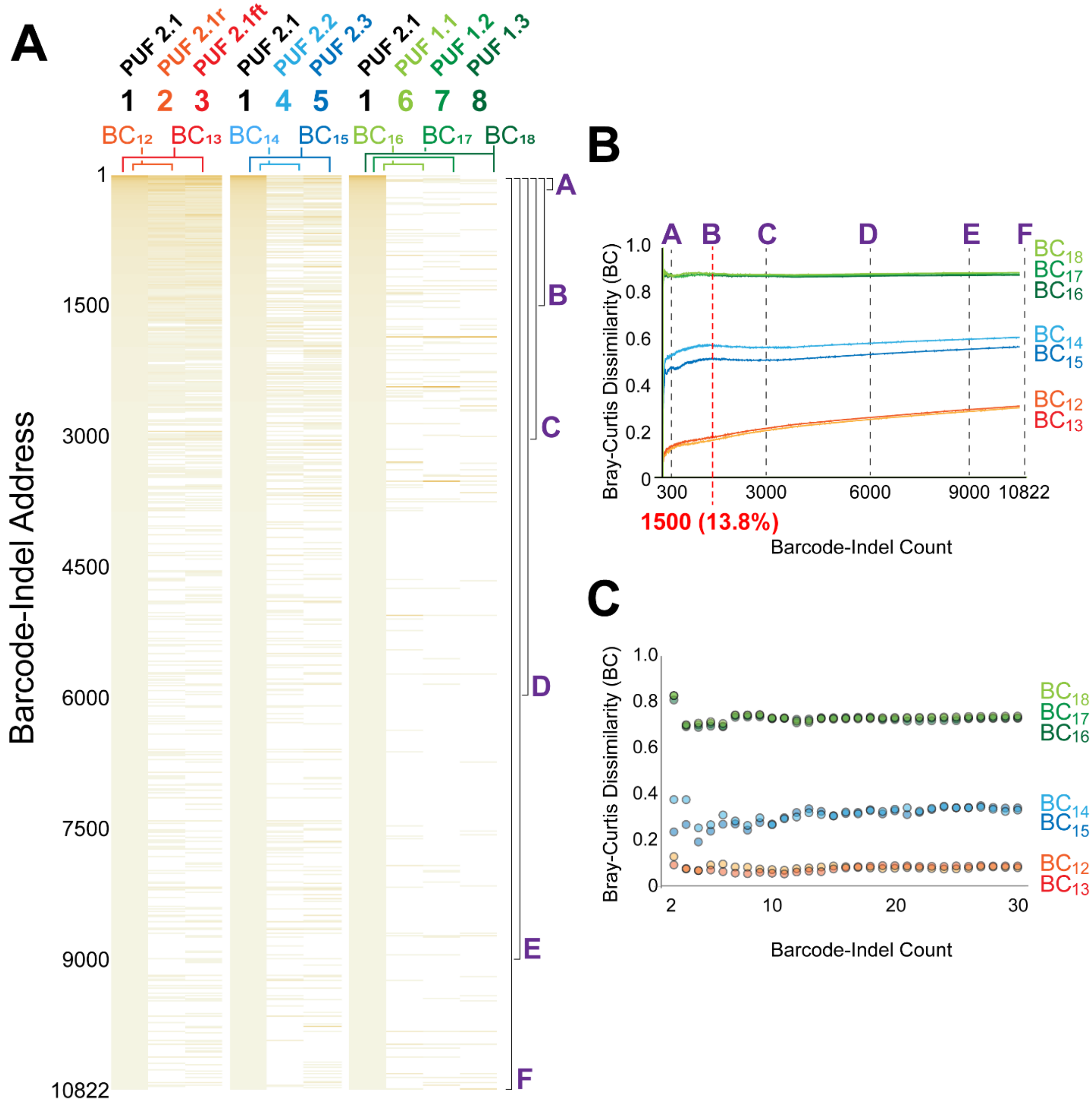
Supplementary Figure 8. Calculation of Bray-Curtis dissimilarities using PUF 1.1 as reference with varying sampling rate. (A) To calculate the Bray-Curtis value between 2 PUFs, the NGS results are first turned into an array of barcode-indel combinations. After sorting the array of the reference PUF based on frequency of occurrence, entries of the other arrays are then sorted to match this order. **(B)** The Bray-Curtis value between the reference and another PUF based on the size of the barcode-indel list used in the calculation, from 2 to the size of the reference sample. Purple letters indicate section of the array shown in **(A)** that corresponds to the visual representation of the list used in the calculation. The barcode-indel count shown in red indicates the list size used for analysis in the main text. **(C)** The Bray-Curtis dissimilarity based on the size of the barcode-indel list used to obtain the distance, from 2 to 30.



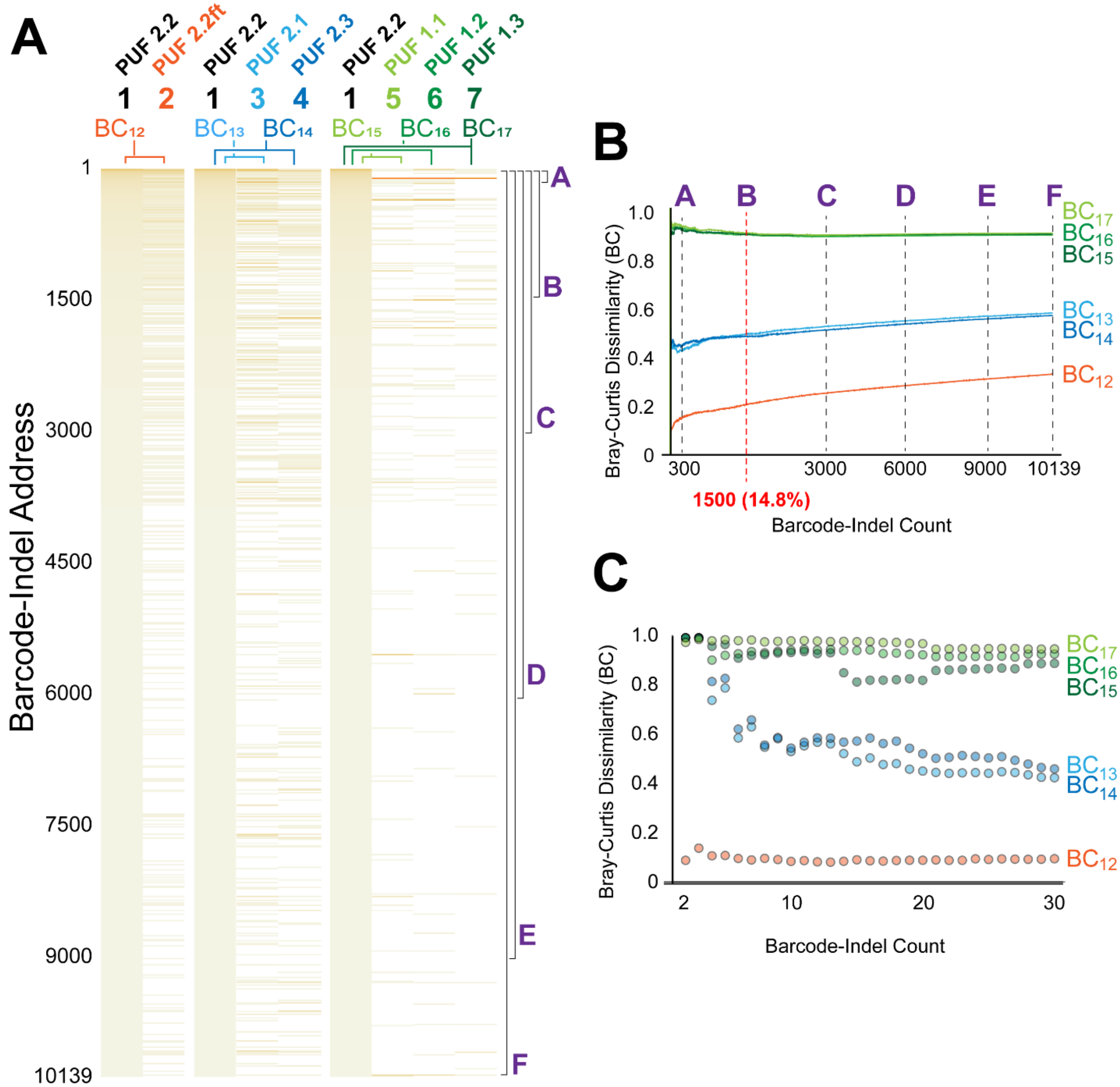
Supplementary Figure 9. Calculation of Bray-Curtis dissimilarities using PUF 1.2 as reference with varying sampling rate. Refer to **Supplementary Figure 4** for a detailed description.



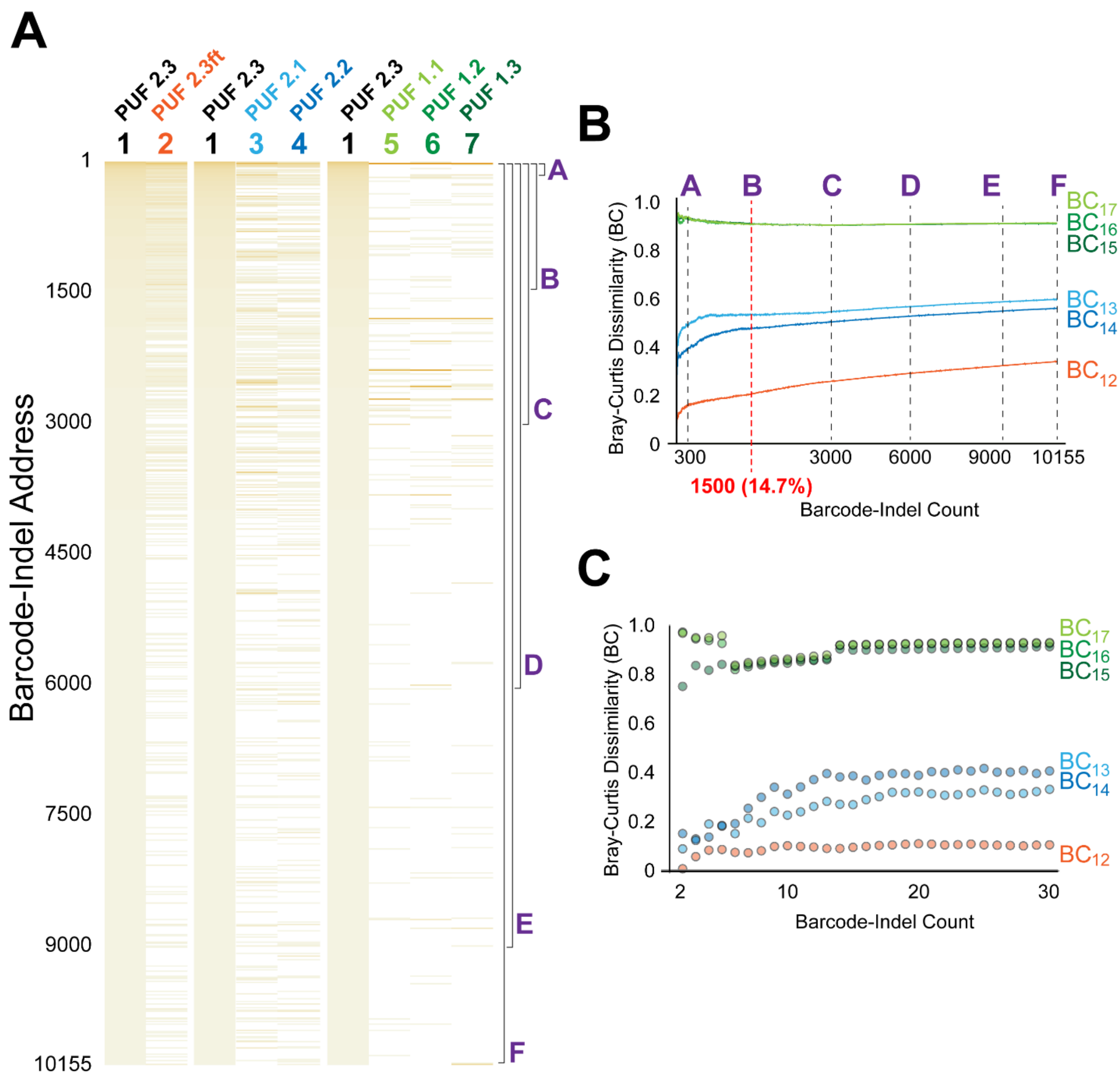
Supplementary Figure 10. Calculation of Bray-Curtis dissimilarities using PUF 1.3 as reference with varying sampling rate. Refer to **Supplementary Figure 4** for a detailed description.



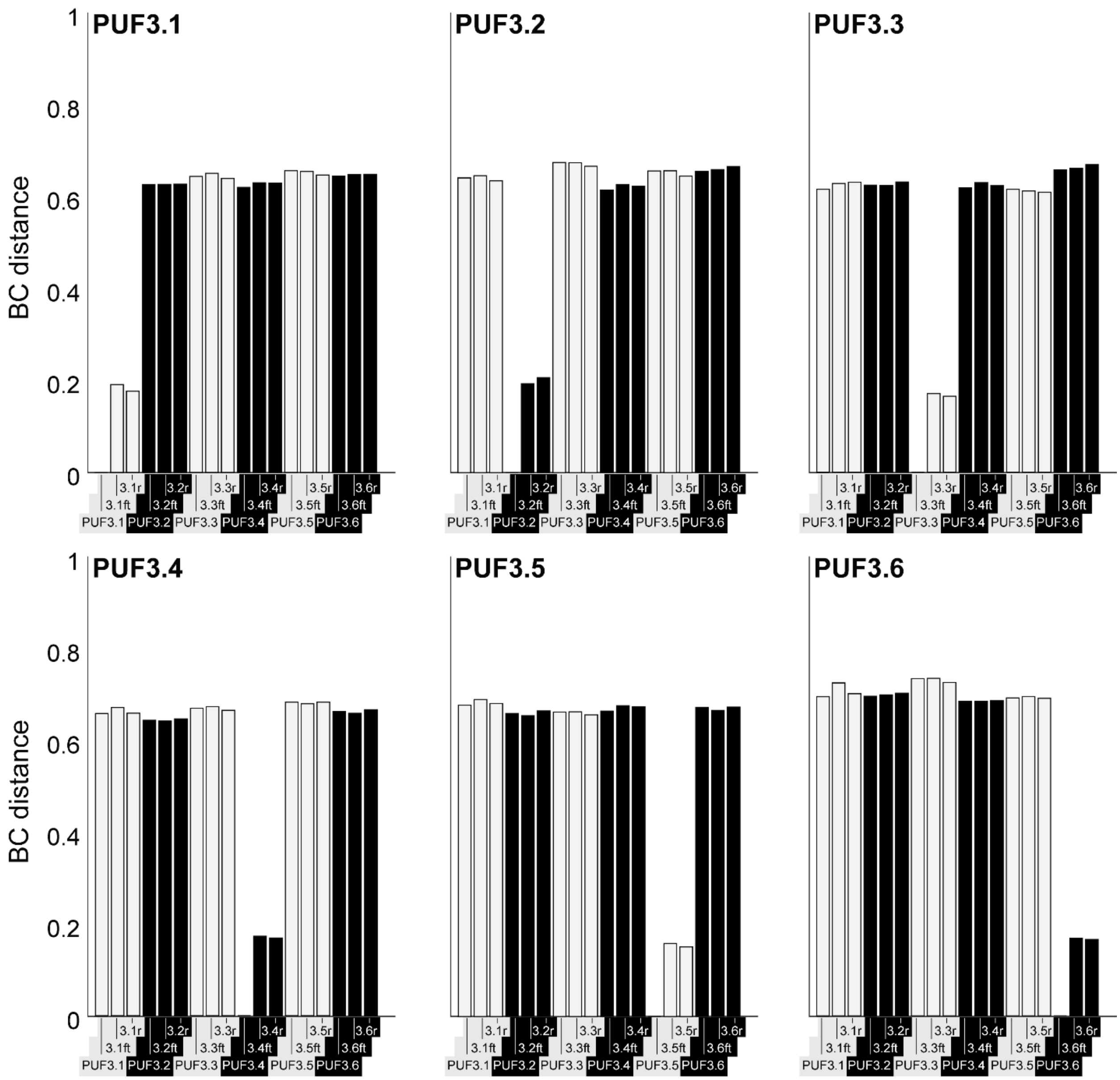
Supplementary Figure 11. Calculation of Bray-Curtis dissimilarities using PUF 2.1 as reference with varying sampling rate. Refer to **Supplementary Figure 4** for a detailed description.



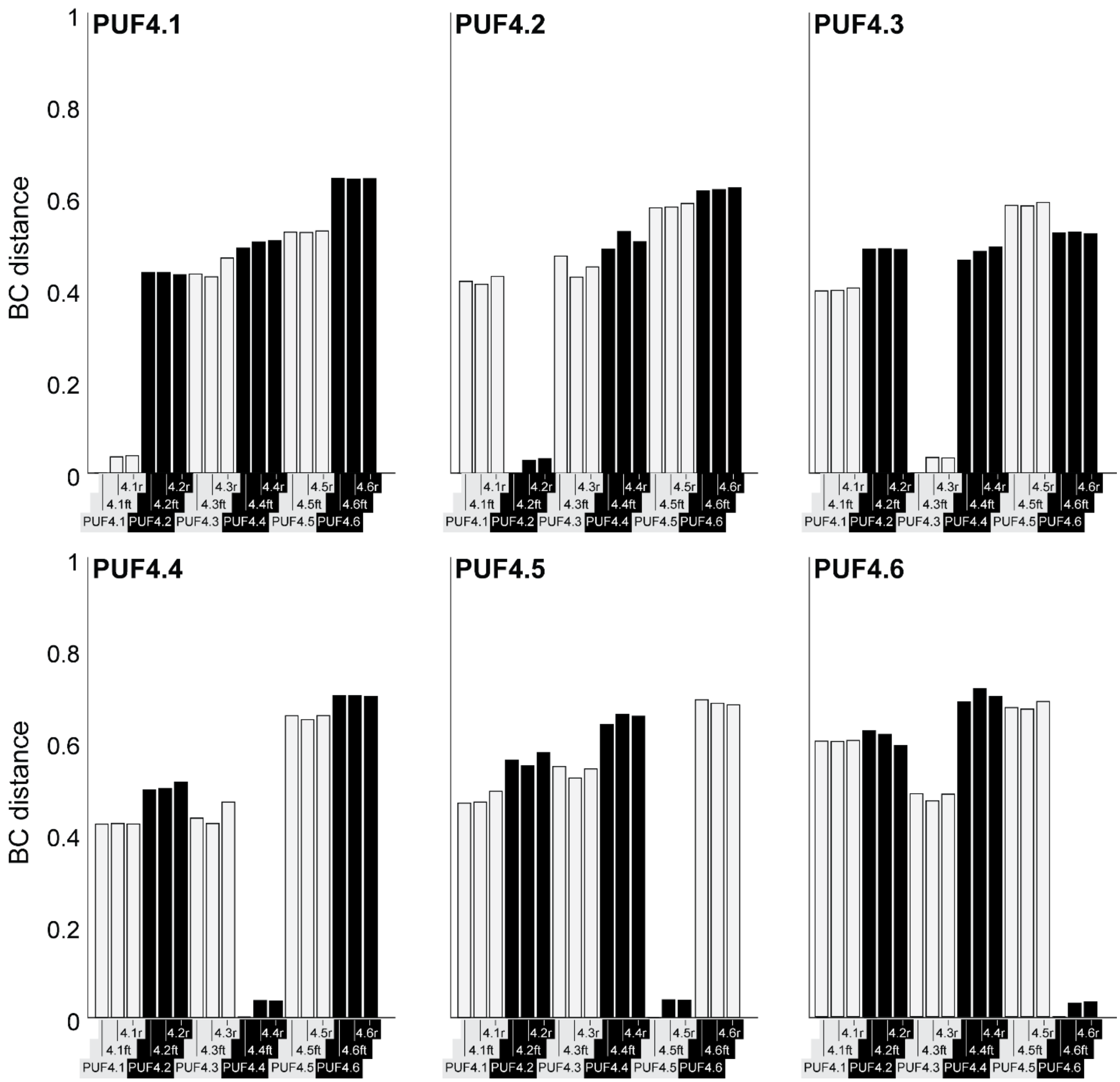
Supplementary Figure 12. Calculation of Bray-Curtis dissimilarities using PUF 2.2 as reference with varying sampling rate. Refer to **Supplementary Figure 4** for a detailed description.



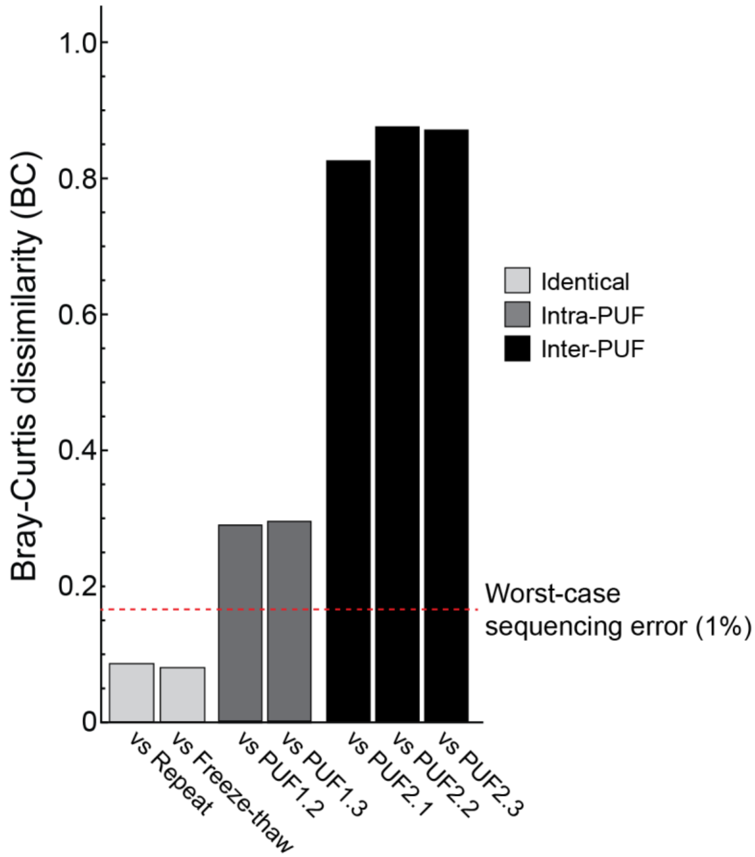
Supplementary Figure 13. Calculation of Bray-Curtis dissimilarities using PUF 2.3 as reference with varying sampling rate. Refer to **Supplementary Figure 4** for a detailed description.



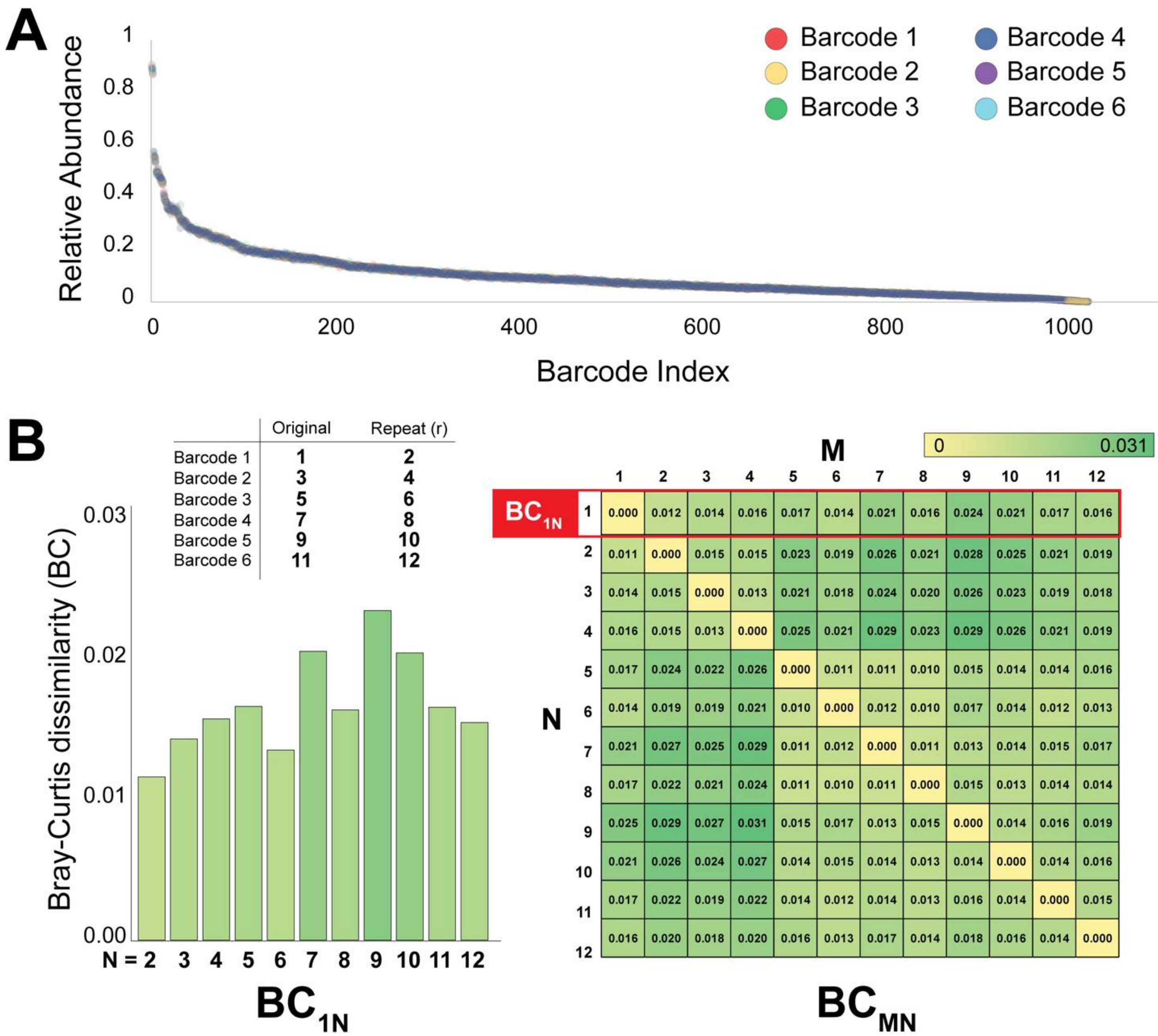
Supplementary Figure 14. Quantitative assessment of HCT116-derived CRISPR-PUFs using Bray-Curtis dissimilarity. Comparison of Bray-Curtis dissimilarities for a single PUF3.i (i={1,2,3,4,5,6}) generated in HCT116 against 17 other PUFs generated in the same cell line.



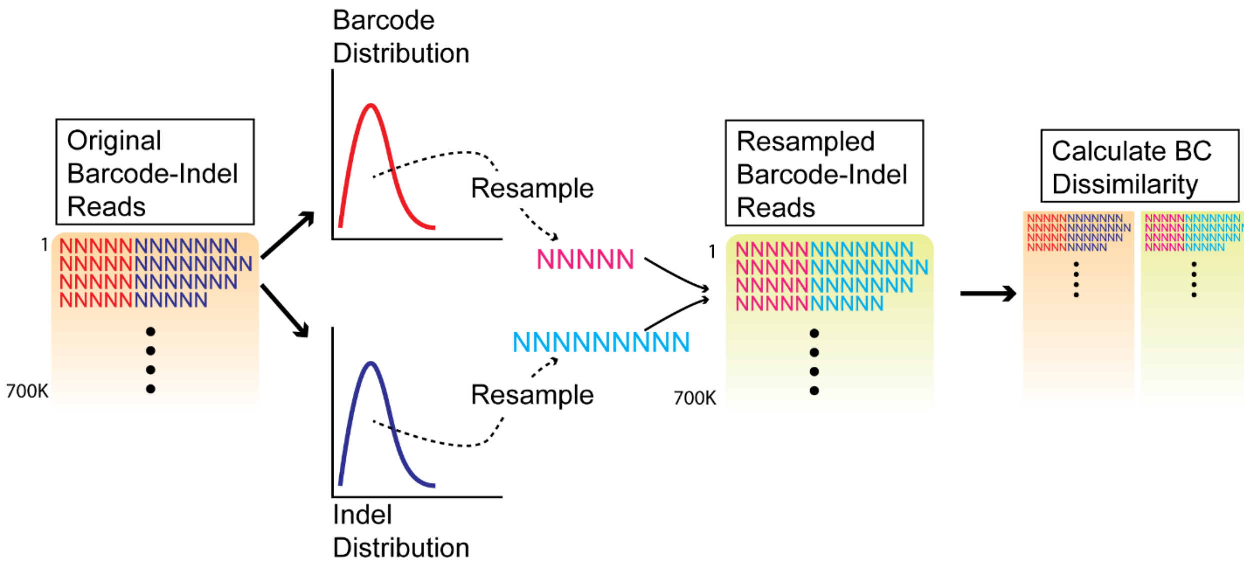
Supplementary Figure 15. Quantitative assessment of HeLa-derived CRISPR-PUFs using Bray-Curtis dissimilarity. Comparison of Bray-Curtis dissimilarities for a single PUF4.i (i={1,2,3,4,5,6}) generated in HeLa against 17 other PUFs generated in the same cell line.



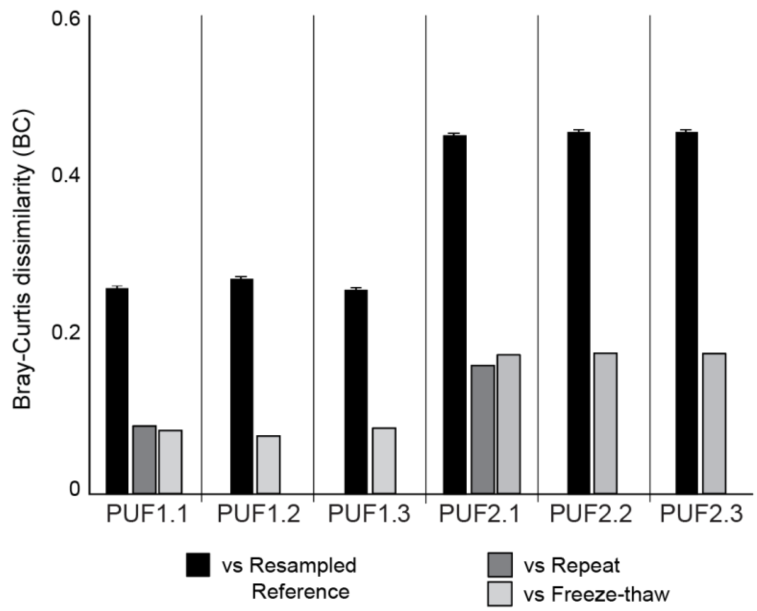
Supplementary Figure 16. Simulated maximum Bray-Curtis dissimilarity from sequencing error for PUFs. To obtain the worst-case Bray-Curtis values from sequencing error, each PUF barcode-indel sequencing data were mutated *in silico* using an error rate of 1% per base. The resulting dataset was then used to calculate the Bray-Curtis value against the original sequence and the technical replicates of the original sequence (repeat and freeze-thaw). The value shown for worst-case sequencing error is an average of 100 different simulations.



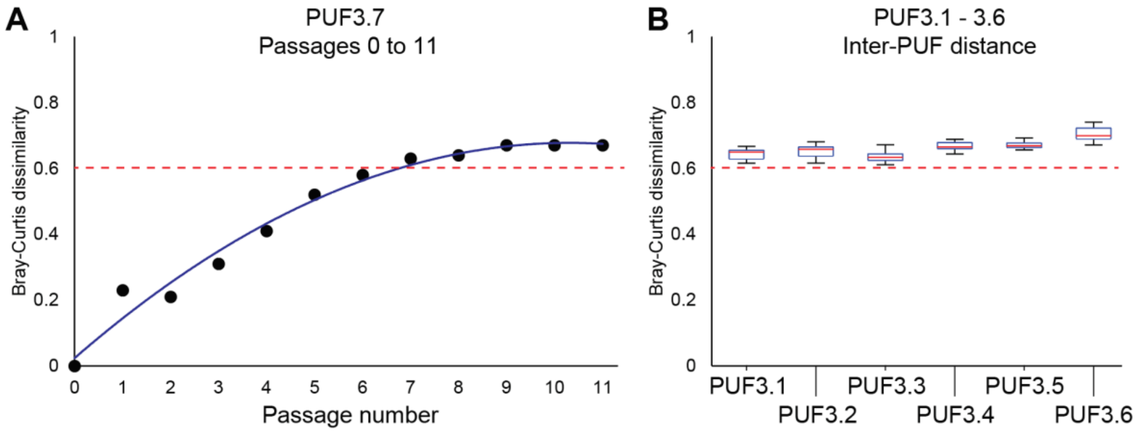
Supplementary Figure 17. Barcode library alone does not satisfy the uniqueness requirement of PUFs. A 5-nucleotide barcode library was stably integrated into the AAVS1 locus of HEK293 cells in 6 parallel trials. **(A)** The relative abundances of stably integrated barcodes in 6 replicates. **(B)** The Bray-Curtis dissimilarity values between barcode 1 and all other 6 samples and their NGS sequencing replicates (left) and of any given pair of all barcodes (right). Note the *intra*-sample dissimilarities generally overlapped with those of *inter*-samples, thus violating the uniqueness requirement of PUFs.



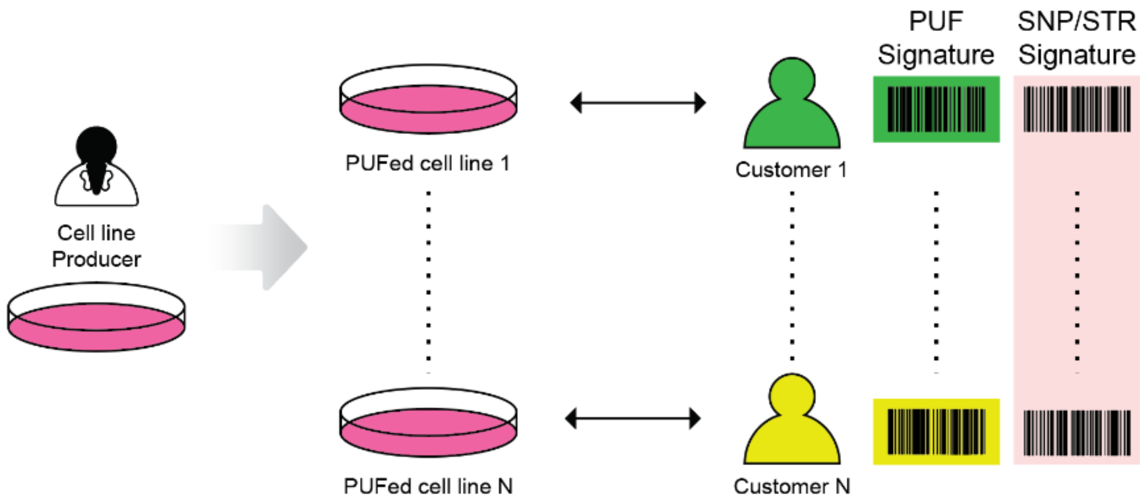
Supplementary Figure 18. Procedure for generating resampled Barcode-Indel reads and corresponding BC dissimilarity



Supplementary Figure 19. Bray-Curtis dissimilarities for intra-PUFs and simulated inter-PUFs.



Supplementary Figure 20. Bray-Curtis dissimilarity values for PUF samples collected from each of the 11 passages. (A) Bray-Curtis dissimilarity values for PUF samples collected from each of the 11 passages. (B) Observed inter-PUF distances of HCT166-based PUF3.1 – 3.6.



Supplementary Figure 21. Attestation using PUFs. Integrating CRISPR-PUF into an engineered cell line product can enable attestation of multiple transactions from identical cell lines.

Supplementary Text

General cloning protocols

Q5 High-Fidelity 2X Master Mix (New England Biolabs) was used for all polymerase chain reactions (PCR) according to the manufacturer's protocol. All oligonucleotides were ordered from Sigma-Aldrich and were listed in **Supplementary Table 1**. The plasmids were constructed using PCR amplification, restriction digest (all restriction enzymes were ordered from New England Biolabs), and ligation with T4 DNA ligase (New England Biolabs). Gel purification and PCR purification were performed with QIAquick Gel Extraction and PCR Purification kits (Qiagen). Transformations were performed using NEB 5-alpha electrocompetent *Escherichia Coli* (New England Biolabs). The minipreps were performed using QIAprep Spin Miniprep kit (Qiagen). The final plasmids were confirmed by both restriction enzyme digestions and direct Sanger sequencings.

DNA constructs

Barcode-Truncated CMV-mKate-PGK1-hygromycin resistance gene: CMV-mKate-PGK1-hygromycin resistance gene was used as the PCR template with primers P3 and P4. The purified PCR product was then cloned into CMV-mKate-PGK1-hygromycin resistance gene vector using AscI and SbfI sites.

CMV-SpCas9-U6-sgRNA1: CMV-SpCas9-U6-BRIP1-sgRNA was used as the PCR template with primers P5 and P6. Next, the purified PCR product was used as the PCR template with primers P5 and P7. The purified PCR product was then cloned into CMV-SpCas9 vector using KpnI and XbaI sites.

CMV-SpCas9-U6-sgRNA2: CMV-SpCas9-U6-BRIP1-sgRNA was used as the PCR template with primers P5 and P8. Next, the purified PCR product was used as the PCR template with primers P5 and P7. The purified PCR product was then cloned into CMV-SpCas9 vector using KpnI and XbaI sites.

CMV-SpCas9-U6-sgRNA3: CMV-SpCas9-U6-BRIP1-sgRNA was used as the PCR template with primers P5 and P9. Next, the purified PCR product was used as the PCR template with primers P5 and P7. The purified PCR product was then cloned into CMV-SpCas9 vector using KpnI and XbaI sites.

CMV-SpCas9-U6-sgRNA4: CMV-SpCas9-U6-BRIP1-sgRNA was used as the PCR template with primers P5 and P10. Next, the purified PCR product was used as the PCR template with primers P5 and P7. The purified PCR product was then cloned into CMV-SpCas9 vector using KpnI and XbaI sites.

CMV-SpCas9-U6-sgRNA5: CMV-SpCas9-U6-BRIP1-sgRNA was used as the PCR template with primers P5 and P11. Next, the purified PCR product was used as the PCR template with primers P5 and P7. The purified PCR product was then cloned into CMV-SpCas9 vector using KpnI and XbaI sites.

NGS (next generation sequencing)-based amplicon sequencing data analysis pipeline with sample commands

Step 1: extracting the 100-bp reads

```
awk 'NR%4 ==2' < f1.fastq | cat > f2.fastq
awk 'NR%4 ==2' < r1.fastq | cat > r2.fastq
```

Step 2: joining the paired-end reads

```
paste -d '\0' f2.fastq r2.fastq | cat > fr1.fastq
```

Step3: filtering out corrupted reads

```
grep “^CTTATATTCCCAGGGCCGGTTCGCGATCGCCCTGCAGG[A-Z][A-Z][A-Z][A-Z][A-Z]TAGTTATTAATGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGACCGCCCTCGACCGCCTTGATTCTCATGGTCTGGGTGC[A-Z]*GTGGTGGTTGTTCACGGTGCCCT” < fr1.fastq | cat > fr2.fastq
```

Step 4: extracting the barcode and indel sequences

```
sed -e "s/CTTATATTCCCAGGGCCGGTTCGCGATCGCCCTGCAGG \(.*\) TAGTTATTAATGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGACCGCCCTCGACCGCCTTGATTCTCATGGTCTGGGTGC[A-Z]*GTGGTGGTTGTTCACGGTGCCCT [A-Z]*/\1/"  < fr2.fastq | cat > barcode1.fastq
sed -e "s/CTTATATTCCCAGGGCCGGTTCGCGATCGCCCTGCAGG[A-Z][A-Z][A-Z][A-Z][A-Z]TAGTTATTAATGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGACCGCCCTCGACCGCCTTGATTCTCATGGTCTGGGTGC \(.*\) GTGGTGGTTGTTCACGGTGCCCT [A-Z]* /\1/" <  fr2.fastq | cat > indel1.fastq
```

Step 5: joining the paired barcode and indel sequences

```
paste -d '\0' barcode1.fastq indel1.fastq | cat > fr3.fastq
```

Step 6: isolating indels containing insertions/deletions

```
grep -v –x '.\{45\}' fr3.fastq  | cat > fr4.fastq
```

Reverse Engineering a CRISPR-PUF

The effort needed to reverse engineer a CRISPR-PUF, i.e., to synthesize a population that produces an identical barcode-indel matrix, requires an insurmountable amount of time, effort, and cost. Indeed, doing so would necessitate that each individual barcode/indel sequence pair be individually integrated into the required cell line, followed by monoclonal verification and, ultimately, mixing of the individual cells in the right proportions to reproduce the same barcode/indel frequencies observed from the CRISPR-PUF. Simply installing the barcode/indel sequence can, on average, take a single researcher up to seven attempts over 19 weeks with 472 hours of hands-on time and approximately $18,000 to complete a single CRISPR editing workflow, i.e., generation of the desired monoclonal cell line. Furthermore, outsourcing a CRISPR-mediated genetic knock-in, such as a barcode/indel sequence described in our CRISPR-PUFs, has an equivalent cost with a similar time of completion. This process would simply produce cells with the same barcode/indel sequences contained in an individual CRISPR-PUF. For example, to replicate PUF1.1, one would need to create 500 cell lines, which would cost several million USD. Moreover, to dial in the right frequency of engineered cells to reproduce the CRISPR-PUF, would largely be trial and error with no guarantee that it is even possible.

Barcode-Truncated CMV-mKate-PGK1-hygromycin resistance gene Sequence

TAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGGCCAGCTCCCAT
AGCTCAGTCTGGTCTATCTGCCTGGCCCTGGCCATTGTCACTTTGCGCTGC
CCTCCTCTCGCCCCCGAGTGCCCTTGCTGTGCCGCCGGAACTCTGCCCTCT
AACGCTGCCGTCTCTCTCCTGAGTCCGGACCACTTTGAGCTCTACTGGCTT
CTGCGCCGCCTCTGGCCCACTGTTTCCCCTTCCCAGGCAGGTCCTGCTTTC
TCTGACCTGCATTCTCTCCCCTGGGCCTGTGCCGCTTTCTGTCTGCAGCTT
GTGGCCTGGGTCACCTCTACGGCTGGCCCAGATCCTTCCCTGCCGCCTCCT
TCAGGTTCCGTCTTCCTCCACTCCCTCTTCCCCTTGCTCTCTGCTGTGTTGC
TGCCCAAGGATGCTCTTTCCGGAGCACTTCCTTCTCGGCGCTGCACCACGT
GATGTCCTCTGAGCGGATCCTCCCCGTGTCTGGGTCCTCTCCGGGCATCTC
TCCTCCCTCACCCAACCCCATGCCGTCTTCACTCGCTGGGTTCCCTTTTCCT
TCTCCTTCTGGGGCCTGTGCCATCTCTCGTTTCTTAGGATGGCCTTCTCCGA
CGGATGTCTCCCTTGCGTCCCGCCTCCCCTTCTTGTAGGCCTGCATCATCA
CCGTTTTTCTGGACAACCCCAAAGTACCCCGTCTCCCTGGCTTTAGCCACCT
CTCCATCCTCTTGCTTTCTTTGCCTGGACACCCCGTTCTCCTGTGGATTCGG
GTCACCTCTCACTCCTTTCATTTGGGCAGCTCCCCTACCCCCCTTACCTCTC
TAGTCTGTGCTAGCTCTTCCAGCCCCCTGTCATGGCATCTTCCAGGGGTCC
GAGAGCTCAGCTAGTCTTCTTCCTCCAACCCGGGCCCCTATGTCCACTTCA
GGACAGCATGTTTGCTGCCTCCAGGGATCCTGTGTCCCCGAGCTGGGACC
ACCTTATATTCCCAGGGCCGGTTCGCGATCGCCCTGCAGGNNNNNTAGTTA
TTAATGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGA
GTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTC
CGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATA
TAAGCAGAGCTGGTTTAGTGAACCGACCAGCTAAGACACTGCCACGGTCAG
ATCCGCTAGCGCTACCGGTCGCCACCATGGTGAGCGAGCTGATTAAGGAG
AACATGCACATGAAGCTGTACATGGAGGGCACCGTGAACAACCACCACTTC
AAGTGCACATCCGAGGGCGAAGGCAAGCCCTACGAGGGCACCCAGACCAT
GAGAATCAAGGCGGTCGAGGGCGGCCCTCTCCCCTTCGCCTTCGACATCC
TGGCTACCAGCTTCATGTACGGCAGCAAAACCTTCATCAACCACACCCAGG
GCATCCCCGACTTCTTTAAGCAGTCCTTCCCCGAGGGCTTCACATGGGAGA
GAGTCACCACATACGAAGACGGGGGCGTGCTGACCGCTACCCAGGACACC
AGCCTCCAGGACGGCTGCCTCATCTACAACGTCAAGATCAGAGGGGTGAAC
TTCCCATCCAACGGCCCTGTGATGCAGAAGAAAACACTCGGCTGGGAGGC
CTCCACCGAGACCCTGTACCCCGCTGACGGCGGCCTGGAAGGCAGAGCCG
ACATGGCCCTGAAGCTCGTGGGCGGGGGCCACCTGATCTGCAACTTGAAG
ACCACATACAGATCCAAGAAACCCGCTAAGAACCTCAAGATGCCCGGCGTC
TACTATGTGGACAGAAGACTGGAAAGAATCAAGGAGGCCGACAAAGAGACC
TACGTCGAGCAGCACGAGGTGGCTGTGGCCAGATACTGCGACCTCCCTAG
CAAACTGGGGCACAGAGGTGGAGGAGGTTCCGGATCTCACGGCTTCCCTC
CCGAGGTGGAGGAGCAGGCCGCCGGCACCCTGCCCATGAGCTGCGCCCA
GGAGAGCGGCATGGATAGACACCCTGCTGCTTGCGCCAGCGCCAGGATCA
ACGTCTCTAGATAACTGATCATAATCAGCCATACCACATTTGTAGAGGTTTTA
CTTGCTTTAAAAAACCTCCCACACCTCCCCCTGAACCTGAAACATAAAATGA
ATGCAATTGTTGTTGTTAACTTGTTTATTGCAGCTTATAATGGTTACAATAAA

GCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGT
TGTGGTTTGTCCAAACTCATCAATGTATCTTAACGCGTAAATTGGGCGCGCC
CTTAAGCTGGGACGGAGGCTTGTTTGCGAGGCCGCGGCCGGCCGAAGTTC
CTATTCTCTAGAAAGTATAGGAACTTCTACCGGGTAGGGGAGGCGCTTTTCC
CAAGGCAGTCTGGAGCATGCGCTTTAGCAGCCCCGCTGGGCACTTGGCGC
TACACAAGTGGCCTCTGGCCTCGCACACATTCCACATCCACCGGTAGGCGC
CAACCGGCTCCGTTCTTTGGTGGCCCCTTCGCGCCACCTTCTACTCCTCCC
CTAGTCAGGAAGTTCCCCCCCGCCCCGCAGCTCGCGTCGTGCAGGACGTG
ACAAATGGAAGTAGCACGTCTCACTAGTCTCGTGCAGATGGACAGCACCGC
TGAGCAATGGAAGCGGGTAGGCCTTTGGGGCAGCGGCCAATAGCAGCTTT
GCTCCTTCGCTTTCTGGGCTCAGAGGCTGGGAAGGGGTGGGTCCGGGGGC
GGGCTCAGGGGCGGGCTCAGGGGCGGGGCGGGCGCCCGAAGGTCCTCC
GGAGGCCCGGCATTCTGCACGCTTCAAAAGCGCACGTCTGCCGCGCTGTT
CTCCTCTTCCTCATCTCCGGGCCTTTCGACCTGCATCCATCTAGATCTCGAT
CGAGCAGCTGAAGCTTACCGCAGGCTATGAAAAAGCCTGAACTCACCGCGA
CGTCTGTCGAGAAGTTTCTGATCGAAAAGTTCGACAGCGTCTCCGACCTGA
TGCAGCTCTCGGAGGGCGAAGAATCTCGTGCTTTCAGCTTCGATGTAGGAG
GGCGTGGATATGTCCTGCGGGTAAATAGCTGCGCCGATGGTTTCTACAAAG
ATCGTTATGTTTATCGGCACTTTGCATCGGCCGCGCTCCCGATTCCGGAAG
TGCTTGACATTGGGGAATTCAGCGAGAGCCTGACCTATTGCATCTCCCGCC
GTGCACAGGGTGTCACGTTGCAAGACCTGCCTGAAACCGAACTGCCCGCT
GTTCTGCAGCCGGTCGCGGAGGCCATGGATGCGATCGCTGCGGCCGATCT
TAGCCAGACGAGCGGGTTCGGCCCATTCGGACCGCAAGGAATCGGTCAAT
ACACTACATGGCGTGATTTCATATGCGCGATTGCTGATCCCCATGTGTATCA
CTGGCAAACTGTGATGGACGACACCGTCAGTGCGTCCGTCGCGCAGGCTC
TCGATGAGCTGATGCTTTGGGCCGAGGACTGCCCCGAAGTCCGGCACCTC
GTGCACGCGGATTTCGGCTCCAACAATGTCCTGACGGACAATGGCCGCATA
ACAGCGGTCATTGACTGGAGCGAGGCGATGTTCGGGGATTCCCAATACGA
GGTCGCCAACATCTTCTTCTGGAGGCCGTGGTTGGCTTGTATGGAGCAGCA
GACGCGCTACTTCGAGCGGAGGCATCCGGAGCTTGCAGGATCGCCGCGGC
TCCGGGCGTATATGCTCCGCATTGGTCTTGACCAACTCTATCAGAGCTTGGT
TGACGGCAATTTCGATGATGCAGCTTGGGCGCAGGGTCGATGCGACGCAA
TCGTCCGATCCGGAGCCGGGACTGTCGGGCGTACACAAATCGCCCGCAGA
AGCGCGGCCGTCTGGACCGATGGCTGTGTAGAAGTACTCGCCGATAGTGG
AAACCGACGCCCCAGCACTCGTCCGAGGGCAAAGGAATAGGGGAGGCTAA
CTGAAGCTTCCCGGGGGTACCAAATTCGTCGACAGATCTAACTTGTTTATTG
CAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAA
GCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATC
TTATGATGTCTGCATATGGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCG
CGGCCGCTCCCACCCGCTCGTCCCCCCGCGCACCTTTGCTAGGAGCGGGT
CGCCCATGTGGCTCTCAGGTTCTGGGTACTTTTATCTGTCCCCTCCACCCCA
CAGTGGGGCCACTAGGGACAGGATTGGTGACAGAAAAGCCCCATCCTTAG
GCCTCCTCCTTCCTAGTCTCCTGATATTGGGTCTAACCCCCACCTCCTGTTA
GGCAGATTCCTTATCTGGTGACACACCCCCATTTCCTGGAGCCATCTCTCTC
CTTGCCAGAACCTCTAAGGTTTGCTTACGATGGAGCCAGAGAGGATCCTGG
GAGGGAGAGCTTGGCAGGGGGTGGGAGGGAAGGGGGGGATGCGTGACCT

GCCCGGTTCTCAGTGGCCACCCTGCGCTACCCTCTCCCAGAACCTGAGCTG
CTCTGACGCGGCCGTCTGGTGCGTTTCACTGATCCTGGTGCTGCAGCTTCC
TTACACTTCCCAAGAGGAGAAGCAGTTTGGAAAAACAAAATCAGAATAAGTT
GGTCCTGAGTTCTAACTTTGGCTCTTCACCTTTCTAGTCCCCAATTTATATTG
TTCCTCCGTGCGTCAGTTTTACCTGTGAGATAAGGCCAGTAGCCAGCCCCG
TCCTGGCAGGGCTGTGGTGAGGAGGGGGGTGTCCGTGTGGAAAACTCCCT
TTGTGAGAATGGTGCGTCCTAGGTGTTCACCAGGTCGTGGCCGCCTCTACT
CCCTTTCTCTTTCTCCATCCTTCTTTCCTTAAAGAGTCCCCAGTGCTATCTGG
GACATATTCCTCCGCCCAGAGCAGGGTCCCGCTTCCCTAAGGCCCTGCTCT
GGGCTTCTGGGTTTGAGTCCTTGGCAAGCCCAGGAGAGGCGCTCAGGCTT
CCCTGTCCCCCTTCCTCGTCCACCATCTCATGCCCCTGGCTCTCCTGCCCC
TTCCCTACAGGGGTTCCTGGCTCTGCTCTTCAGACTGAGCCCCGTTCCCCT
GCATCCCCGTTCCCCTGCATCCCCCTTCCCCTGCATCCCCCAGAGGCCCCA
GGCCACCTACTTGGCCTGGACCCCACGAGAGGCCACCCCAGCCCTGTCTA
CCAGGCTGCCTTTTGGGTGGATTCTCCTCCAACTGTGGGGTGACTGCTTGG
CAAACTCACCGGTACCCGGCCGCGACTCTAGATCATAATCAGCTCGAGCCT
TAACAAGCTTCGAAACGATATGGGCTGAATACAAAAACGATATGGGCTGAAT
ACAAAAACGATATGGGCTGAATACAAACCGCTTGAAGTCTTTAATTAAACCG
CTTGAAGTCTTTAATTAAACCGCTTGAAGTCTTTAATTAAAGGATCCACCGGA
TCTAGATAACTGATCATAATCGCGGCCGCACTCCTCAGGTGCAGGCTGCCT
ATCAGAAGGTGGTGGCTGGTGTGGCCAATGCCCTGGCTCACAAATACCACT
GAGATCTTTTTCCCTCTGCCAAAAATTATGGGGACATCATGAAGCCCCTTGA
GCATCTGACTTCTGGCTAATAAAGGAAATTTATTTTCATTGCAATAGTGTGTT
GGAATTTTTTGTGTCTCTCACTCGGAAGGACATATGGGAGGGCAAATCATTT
AAAACATCAGAATGAGTATTTGGTTTAGAGTTTGGCAACATATGCCATATGC
TGGCTGCCATGAACAAAGGTGGCTATAAAGAGGTCATCAGTATATGAAACA
GCCCCCTGCTGTCCATTCCTTATTCCATAGAAAAGCCTTGACTTGAGGTTAG
ATTTTTTTTATATTTTGTTTTGTGTTATTTTTTTCTTTAACATCCCTAAAATTTTC
CTTACATGTTTTACTAGCCAGATTTTTCCTCCTCTCCTGACTACTCCCAGTCA
TAGCTGTCCCTCTTCTCTTATGAAGATCCCTCGACCTGCAGCCCAAGCTTGG
CGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATT
CCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAA
TGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGT
CGGGAAACCTGTCGTGCCAGCGGATCCGCATCTCAATTAGTCAGCAACCAT
AGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGC
CCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAG
GCCGCCTCGGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGA
GGCCTAGGCTTTTGCAAAAAGCTAACTTGTTTATTGCAGCTTATAATGGTTAC
AAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCA
TTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGGATCC
GCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTG
GGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGC
TGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAG
AATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAG
GCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGC
CCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAAC

CCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTG
CGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTC
CCTTCGGGAAGCGTGGCGCTTTCTCAATGCTCACGCTGTAGGTATCTCAGT
TCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTT
CAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCG
GTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGC
AGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAAC
TACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCA
GTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACC
GCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAA
AAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGT
GGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGAT
CTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTAT
ATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCT
ATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCG
TGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAA
TGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACC
AGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCC
TCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAG
TTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACG
CTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCG
AGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCT
CCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGG
CAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGT
GACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACC
GAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAG
AACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCA
AGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCC
AACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAA
CAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTT
GAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATT
GTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAA

Green: left homology arm
Red: 5-nucleotide barcode
Dark Red: truncated CMV promoter
Light Blue: mKate open reading frame
Purple: PGK1 promoter
Blue: hygromycin resistance gene open reading frame
Orange: right homology arm

Bray-Curtis and sequencing reads

Assume that a PUF sample contains N barcode-indel reads, the average length of each read is L, and the error rate per base is e. Thus, the total number of mutations is N * L * e.

When N * L * e << N, each mutation most likely will occur within a different read. We further assume that the mutation does not result in a sequence identical to one of the original reads. Thus, for the (N – N * L * e) non-mutated reads, they will appear in both the original and in the mutated samples. In contrast, for the (N * L * e) mutated reads, they will only appear in the original sample.

Therefore, the Bray-Curtis value will be: (N * L * e) / (N + N – N * L * e) = (L * e) / (2 – L * e).

Since L * e << 1, the Bray-Curtis value is (L * e) / 2, therefore the BC values are directly related to the read size L.

Supplementary Tables

Supplementary Table 1. Primers used in this study
Supplementary Table 2. The list of individual barcodes and their frequencies for the pilot PUF (Figure 3).
Supplementary Table 3. The list of individual indels and their frequencies for the pilot PUF (Figure 3).
Supplementary Table 4. The PUF matrix for the pilot PUF (Figure 3).
Supplementary Table 5. The list of individual barcodes/indels and their frequencies for PUF1 samples (Figure 5).
Supplementary Table 6. The list of individual barcodes/indels and their frequencies for PUF2 samples (Figure 5).
Supplementary Table 7. The PUF matrices for PUF1 samples (Figure 5).
Supplementary Table 8. The PUF matrices for PUF2 samples (Figure 5).
Supplementary Table 9. The list of individual barcodes/indels and their frequencies for PUF3 samples (Figure 6).
Supplementary Table 10. The list of individual barcodes/indels and their frequencies for PUF4 samples (Figure 6).
Supplementary Table 11. The list of individual barcode-indel addresses and their frequencies for all PUF samples (Figures 8 and 9).
Supplementary Table 12. Total variation distances between PUF samples (Supplementary Figures 5 – 7).
Supplementary Table 13. The Bray-Curtis dissimilarities between PUFs and their corresponding mutated samples (Supplementary Figure 16).
Supplementary Table 14. The relative abundances of stably integrated barcodes in 6 replicates (Supplementary Figure 17).
Supplementary Table 15. The Bray-Curtis dissimilarities between barcode replicates and their NGS sequencing replicates (denoted as r) (Supplementary Figure 17).
Supplementary Table 16. The Bray-Curtis dissimilarities between PUFs and their corresponding reshuffled samples (Supplementary Figure 19).
Supplementary Table 17. Bray-Curtis dissimilarity values for PUF samples collected from each of the 11 passages (Supplementary Figure 20).

Supplementary Scripts

BC_value_calculations.py. The Python script for Bray-Curtis dissimilarities.

Total_variation_distance_calculations.py. The python script for total variation distances.