Self-Documenting Plasmids

Samuel J. Peccoud[1,2†], Sarah I. Hernandez[3†], Diptendu M. Kar[4], Casey-Tyler Berezin[3], Jean Peccoud[2,3,5,6,7*]

1. Department of Electrical Engineering, Colorado State University, Fort Collins, CO, USA
2. GenoFAB, Inc.
3. Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO, USA
4. Department of Computer Sciences, Northeastern University, Boston, MA, USA
5. Department of Computer Sciences, Colorado State University, Fort Collins, CO, USA
6. School of Biomedical Engineering, Colorado State University, Fort Collins, CO, USA
7. Department of Systems Engineering, Colorado State University, Fort Collins, CO, USA

*Correspondence: jean.peccoud@colostate.edu

†These authors should be considered co-first authors.

Summary

Biological products produced by plasmids underscore the bioeconomy, thus ensuring the sequences of physical DNA constructs match their expected designs is critical. Yet, the verification of plasmids remains a difficult and overlooked step in the production process. We developed a web application to generate certificates that embed information about a plasmid and its designer within the sequence itself. Plasmids can be sequenced *de novo* and upon sequencing and assembly upload, plasmid sequences can be verified with a finite number of errors corrected. Users can also encode GenBank or plain text files in certificate sequences to store additional data or documentation associated with a plasmid within the

sequence. Certificate insertion does not adversely affect bacterial DNA yields nor functional protein expression in mammalian cells. This technology accelerates and simplifies plasmid verification and has the potential to transform bioproduction, biosurveillance, and the protection of intellectual property in the life sciences.

Keywords

Plasmid, Digital signature, Documentation, GenBank, Sequencing, Cryptography

Introduction

Biological sequences like DNA are unique because they exist in two states simultaneously: physical samples should be identical to the digital sequences that define them [1]. The increasing accessibility of DNA synthesis and tools like Gibson assembly and CRISPR-Cas-9 for assembling and editing DNA sequences have made it relatively straightforward to design large libraries of genetic constructs [2-5]. However, there is an alarming disconnect between the sequences uploaded to public repositories and the actual sequences of the physical samples received from such repositories, with nearly half of lab-made genetic parts and plasmids deviating from their theoretical sequences [6,7]. There are two major contributors to such deviations: spontaneous mutations can occur over time and in individual clones, but perhaps more importantly, sequence verification is often overlooked in the production process [1,8].

It is ironic that despite the inherent information-encoding capabilities of DNA, documentation about DNA sequences can often be inaccurate, incomplete, or missing [9,10]. Data storage using DNA presents the opportunity for DNA sequences to contain information

about themselves [11]. Self-documentation is common in cyberspace, such as with smartphones that contain their own manuals or software libraries whose manuals can be easily accessed using the '-h' option. In the DNA space, cryptographic approaches like substitution ciphers (e.g., using "AGG" for "D", "CTG" for "N", and "GCA" for "A" to encode "DNA" in 9 nucleotides) enable DNA sequences to theoretically encode anything from simple information like a sequence ID or the name of the sequence designer to complex information like data files, books, operating systems and more [11-14].

The replicability of DNA makes it ideal as a data storage tool. It is trivial to reproduce DNA sequences in the laboratory by performing a PCR, or even more efficiently, by letting cells replicate them [15]. Plasmids can accommodate cargo thousands of base pairs long and be replicated infinitely at little to no cost [16]. Indeed, error-free data retrieval from plasmids passaged for several generations or frozen for several years has been experimentally shown [15,17]. Plasmids can be easily synthesized, assembled, modified and put into bacteria, yeast, mammalian cells, and plants, so they are commonly used as vectors for creating gene libraries, gene editing, and in bioproduction [15]. Similarly, large artificial chromosomes encoding digital data can be synthesized [18], but compared to plasmids in the kilobase range, it is still difficult to sequence megabase-long chromosomes with the single base-pair accuracy needed for practical data sharing and storage, especially at large scales [11,19]. Since the first instance of using plasmids to insert alphanumeric messages into replicating bacteria in 2003 [20], many DNA data storage approaches have cut long nucleotide messages into fragments only a few thousand base pairs long or less and stored them in plasmids [13,14,17,21-23].

Given the ubiquity of plasmids and the availability of tools for editing plasmids and encoding information in DNA sequences, we focused our attention on how to improve the documentation and verification processes associated with plasmids. We recently developed the first iteration of certificate sequences for self-documenting plasmids that use digital signature techniques to securely embed a plasmid's sequence information within the plasmid sequence itself [24]. Here, we present a simple web application that allows any registered user to generate certificates, manage their certified plasmid library, and verify the accuracy of certified plasmids. Certificates include plasmid and author IDs and a compressed version of the sequence but can also include a GenBank file of the plasmid's annotations or a plain text file. The certification of plasmids accelerates the verification process, eliminating the need for reference sequences or alignment steps, ensures plasmid sequencers can assess the proper documentation, and provides security assurances in plasmid distribution.

Results

GSIN certificates link plasmids to their documentation

We developed the GenoFAB Standard Identification Number (GSIN) web application to generate and verify certificate sequences for plasmids (available at https://din.genofab.com/). The 'Generate Certificate' form contains both editable and locked values (Figure 1a). The GSIN delimiter sequences are standardized for all certificates and are used for certificate detection. Registered users are assigned an unchangeable 16-digit Author ID and unique Plasmid IDs are automatically generated by the software. To certify a plasmid, a GenBank file containing the plasmid sequence and its gene annotations must be uploaded. Once uploaded, a plasmid map viewer appears with a prompt to select a location to insert

the certificate (Figure 1b). The plasmid name is populated based on the file name and is editable so that different instances of the same plasmid can be easily identified. Two certificate sizes are available, the 512 base pair (bp) certificate that was originally developed and the smaller 164 bp certificate implemented in this work [24]. Both are generated by digital signature schemes that utilize hash functions to compress the entire plasmid sequence into either 164 or 512 bp depending on the scheme chosen [24]. Lastly, users can optionally supply a URL to documentation about the plasmid, such as sequence and functional information or a material transfer agreement (MTA). Once all the fields are completed, the GSIN sequence is generated, and users receive a new GenBank file to download containing the annotated certificate (Figure 1c). The full certificate for the 164 bp signature is 260 bp and is comprised of the start delimiter (10 bp), author ID (32 bp), plasmid ID (12 bp), the signature containing the hashed (compressed) plasmid sequence (164 bp), the error correction code (32 bp), and the end delimiter (10 bp).

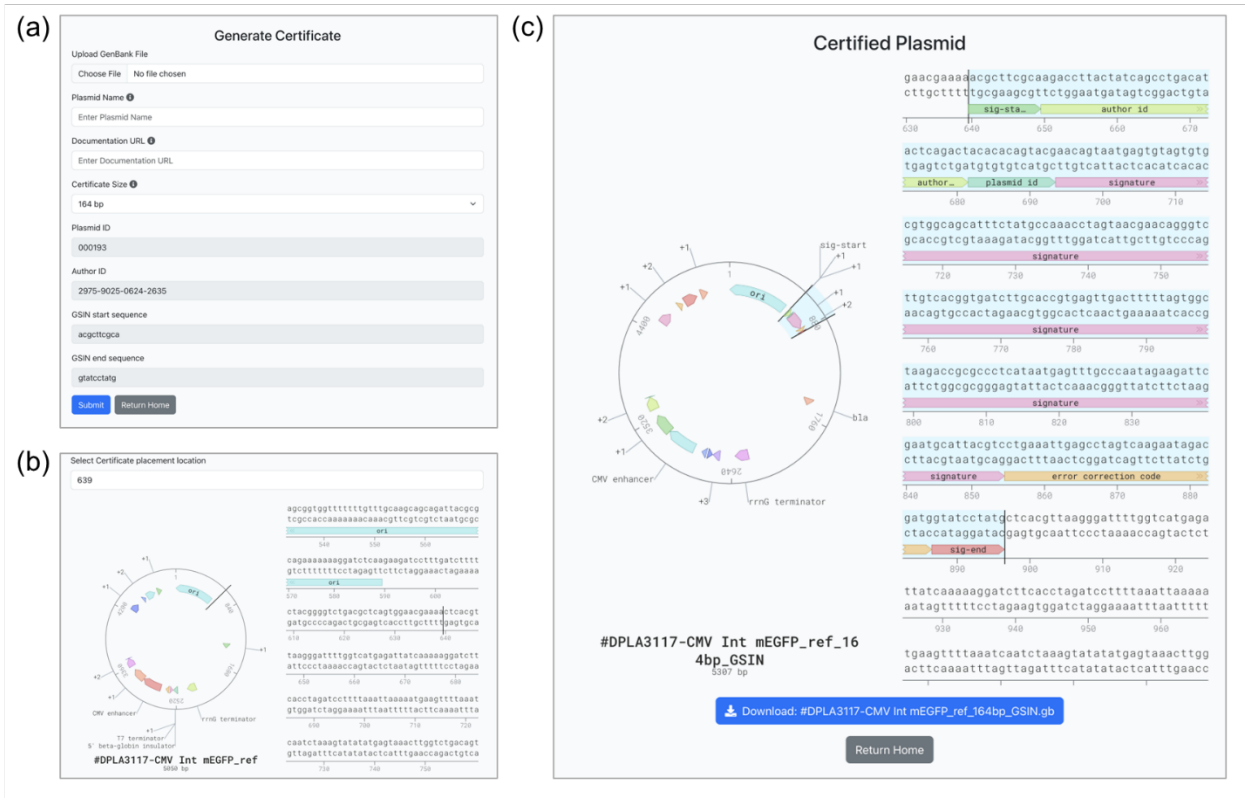


Figure 1. Plasmids are certified through a simple web application. (a) To generate certificate sequences, users upload a GenBank file to the form, then designate a plasmid name, URL, and certificate size. The certificate size determines the length of the compressed plasmid sequence (164 or 512 bp). Plasmid and author IDs are automatically generated by the application, and delimiter (start/end) sequences are standardized. **(b)** Once a GenBank file is uploaded, a plasmid map viewer allows the certifier to select a place to insert the certificate sequence. **(c)** A GenBank file with the certificate features annotated is available to download. Any part of the plasmid sequence can be easily selected (blue highlight) and copied.

We certified 20 plasmids and all had the certificates inserted in the same place in the plasmid backbone between the ampicillin resistance gene and the origin of replication. Each of the URLs point to a FigShare link that hosts the GenBank file associated with the plasmid to illustrate the sharing of plasmid documentation. With the new GenBank files generated by the software, we ordered the certificate sequences as synthesized gene fragments, each with 25-30 bp of overlap to facilitate ligation into the plasmid backbone. We used the Basic Local Alignment Search Tool (BLAST) by NCBI to confirm that none of the certificate sequences had any significant similarity to any sequence in the nucleotide collection [25,26]. We found no

significant difference in the yield of the plasmid DNA extracted between uncertified plasmids and their certified counterparts (Figure 2). Plasmid assemblies were generated using a hybrid assembly pipeline that takes both Illumina short-reads and Nanopore long-reads as input to generate assembly sequences [27].

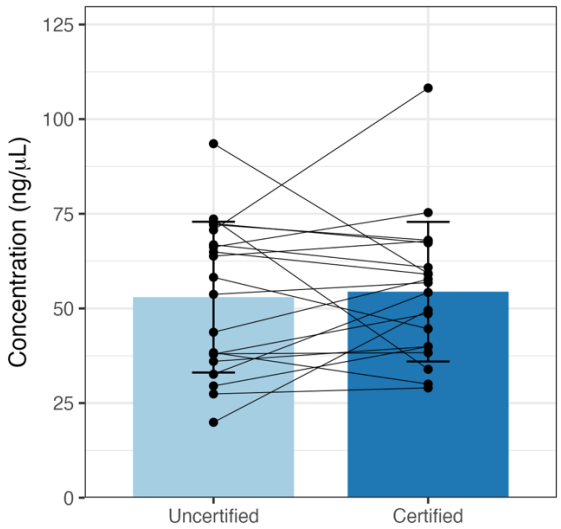


Figure 2. Insertion of certificate sequence does not affect plasmid yield. Yields from miniprep extractions of certified plasmids and their uncertified counterparts (connected by lines) were not significantly different from one another (paired t-test, $t(19) = -0.35$, $p=0.73$). Data presented as mean ± SD.

The FASTA files for plasmid assemblies were uploaded to the 'Verify Certificate' page (Figure 3a). We present both successful and failed verification outcomes (Table 1). FASTA files from plasmids that haven't been certified, or where the certificate was not properly inserted, will fail the verification process (Figure 3b). When the uploaded assembly sequence matches exactly to the original sequence used to generate the certificate, the verifier receives all the information about the plasmid: the signer's author ID, the plasmid ID, the author's name and email address, and the URL associated with that plasmid (Figure 3c). If two or less mismatch mutations have occurred, the error correction code is implemented, the verifier is notified of these errors, and the signature is successfully verified with error correction

(Figure 3d). If more than two mismatches, or any insertions or deletions, have occurred, then verification fails but the software can report the author ID and plasmid ID that were identified in the certificate sequence (Figure 3e). Although this information is revealed, users should verify the accuracy of these IDs and the plasmid sequence by contacting the plasmid distributor, as the mutations in the sequence are not corrected and could have affected sequence accuracy. In our case, reanalysis of the samples that failed verification using a

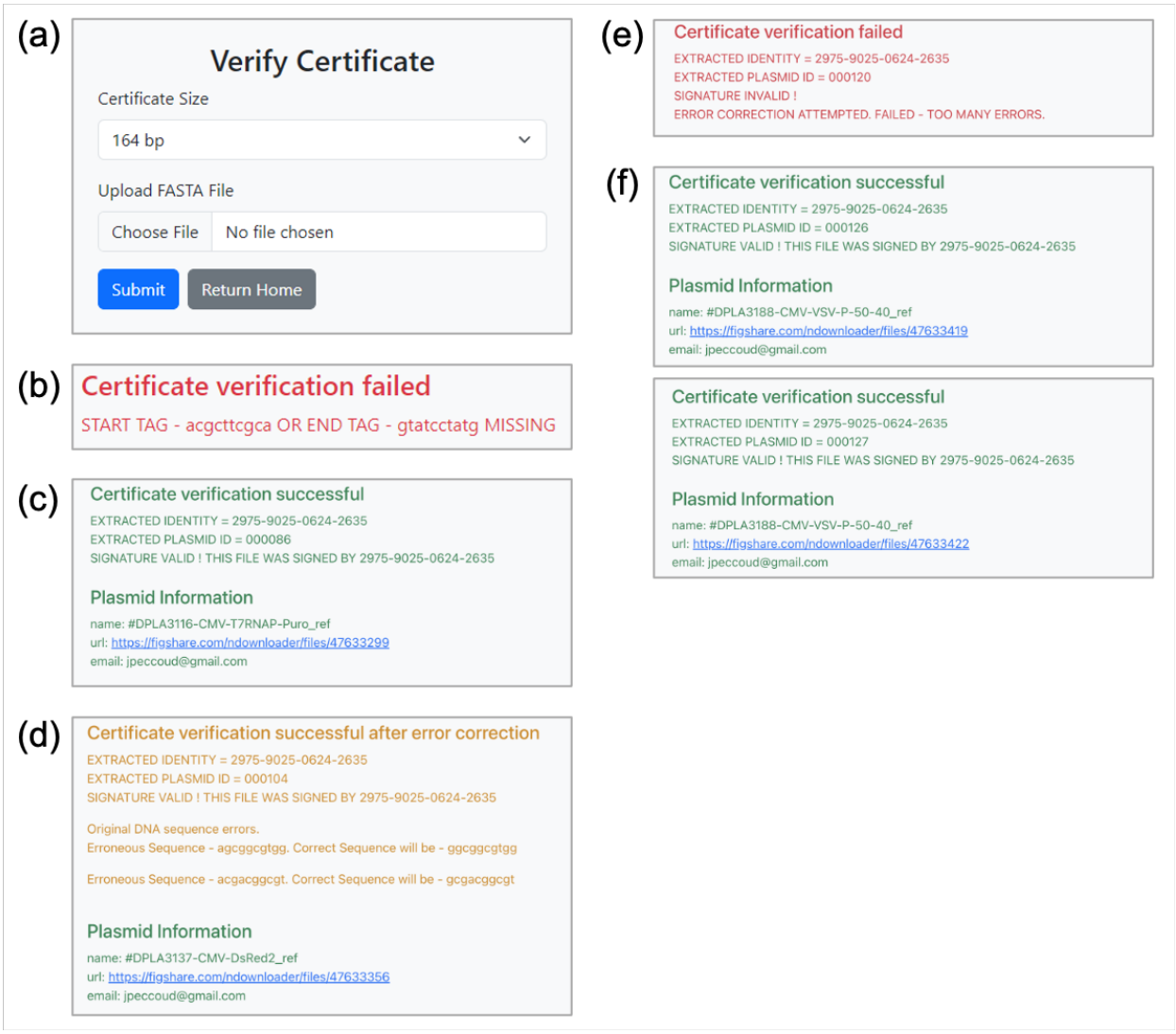


Figure 3. Certificate verification is accurate and secure. (a) The form to verify a certified plasmid only requires the user to choose what type of certificate they expect and upload a FASTA file of a plasmid assembly. **(b)** Plasmids that are not certified or do not have the delimiter sequences will fail

verification. **(c)** Upon successful verification, plasmid and author information is revealed, including a URL to documentation about the plasmid, in this case, a GenBank file hosted on figshare. **(d)** The error correction code can identify and accept up to two mismatches to successfully verify a plasmid, but plasmids with three or more mismatches, or any insertions or deletions, will fail verification **(e)**. **(f)** Identical plasmids can be certified with unique plasmid IDs and URLs.

second bacterial clone showed successful certification, indicating that failure resulted from factors in the ligation and assembly processes rather than an inability to certify certain plasmids.

The same plasmid sequence can be certified multiple times (Figure 3f). Although the initial sequences and the insertion points are the same, by running the certificate generation process multiple times, they can be assigned unique plasmid IDs (leading to different signature sequences) and therefore given different URLs. This feature allows certifiers to, for example, send Plasmid #126 to Alice with an MTA allowing her to replicate the plasmid in bacteria, and send the identical Plasmid #127 to Bob whose URL directs him to a different MTA that forbids replication of the plasmid. The URLs are not embedded into the certificate as sequences, and thus can be changed by the certifier at any time after certificate generation by using the 'Manage Plasmids' feature of the application.

Encoding sequence annotations in a plasmid

In the above case, a user can generate a 164 bp certificate that contains a URL for the verifier to download a GenBank file for the verified plasmid. However, this method is somewhat indirect and carries the risk that the URL could change, making the documentation inaccessible. Thus, we developed a certificate that would encode the same information as above along with a sequence that encodes a GenBank file of the plasmid's annotations, making the plasmid self-documenting.

We illustrate this approach using one of the 20 plasmids that we certified, a mammalian expression vector expressing the GFP gene under the control of the CMV promoter, DPLA3122-CMV-mEGFP, which is 4,966 bp long. DPLA3122-CMV-mEGFP_detailed.gb is 11 kilobytes (KB) and includes 18 features with detailed descriptions, including fields like notes and amino acid coding sequences, as well as some information about the source (Figure 4a). Certificate generation through the GSIN software is completed just as with the above method, but instead the certificate size is 'Self Documenting, size scales with annotations.' The annotations in the GenBank file uploaded by the user are converted into a sequence that can be embedded within the certificate sequence. The software produces a new GenBank file with embedded annotations (e.g., DPLA3122-CMV-mEGFP_detailed_PID171.gb). The sequence annotations are encoded in a feature called 'annotations' that is 6,804 bp long. The resulting self-documenting plasmid is now 12,027 bp long (Figure 4c).

More detailed GenBank files create longer annotations features, and larger plasmids are generally harder to transform or transfect into cells [28-30]. We therefore created DPLA3122-CMV-mEGFP_summary.gb, which is a simplified version of the GenBank file that is only 8 KB. The sequence and features are the same as in the previous file, however, the file preamble has been removed and each feature only has a label and no other field (Figure 4b). Having fewer annotations in the GenBank file reduces the size of the feature encoding the sequence annotations to only 2,064 bp. The resulting self-documenting DPLA3122-CMV-mEGFP_summary_PID172.gb file is only 7,287 bp long.

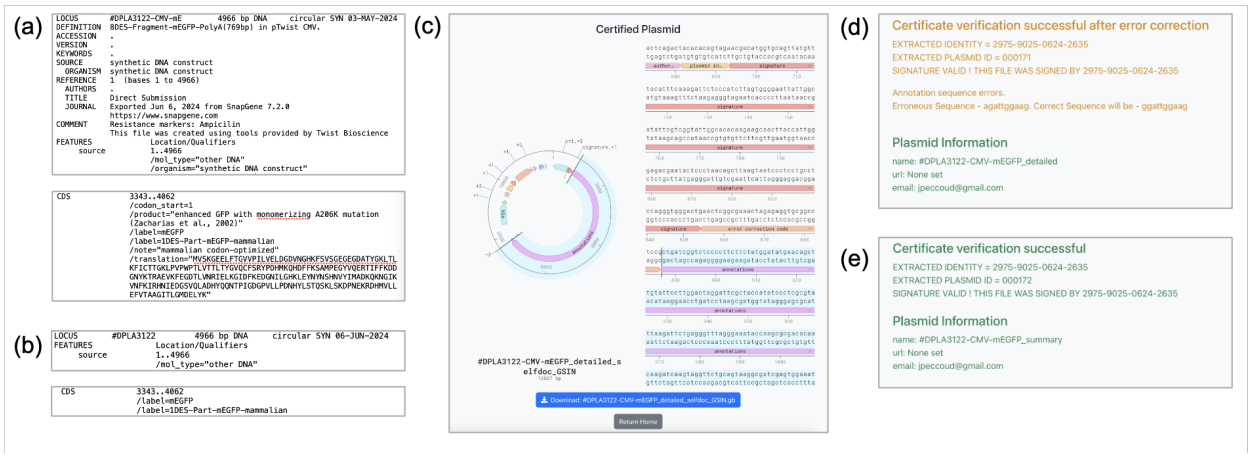


Figure 4. Self-documenting plasmids contain their own GenBank files. To generate a self-documenting plasmid, users upload a GenBank containing information about the plasmid and the gene annotations. We generated one GenBank file with a detailed preamble and gene annotations **(a)** and one simplified GenBank file with minimal annotation data **(b)**. **(c)** The GenBank file is converted into a nucleotide feature called 'annotations' that is embedded within the certificate sequence. The size of the GenBank file determines the size of the 'annotations' feature. **(d)** Self-documenting certificates can be error-corrected. **(e)** Upon plasmid verification, the plasmid information is revealed and the annotated GenBank file is automatically downloaded.

We constructed both self-documenting plasmids by ordering the certificate sequences as sets of gene fragments, which were ligated to produce the full-length plasmids. Both self-documenting plasmids (#171 with the detailed GenBank file and #172 with the summary GenBank file) verified successfully (Figure 4d, e). The error correction code still accommodates up to two mismatches and will alert the user if the errors are in the Annotation sequence (Figure 4d) or in the Original DNA sequence (Figure 3d). Once a certified plasmid is verified, the plasmid information is revealed, including the URL as before, and the annotated GenBank file is automatically downloaded. Thus, while the URL can still be used to share documents, all the pertinent information about a plasmid, its functions, and source are directly embedded into a certified plasmid.

<u>Encoding a text file in a plasmid</u>

GenBank files are flat text files that are specially formatted and amenable to only one type of data. Thus, we next wanted to explore embedding other types of text files into plasmids, again using DPLA3122-CMV-mEGFP to illustrate the approach. One type of documentation that is commonly used in biological resource sharing is the MTA. The Open MTA is a standardized document that can be used to quickly establish global sharing agreements [31]. We copied the Open MTA text (available at https://biobricks.org/open-material-transfer-agreement/) into a text file that is about 8 KB (openmta.txt). Certificates are generated as before, but with the certificate size set to 'File Storage, size scales with file' (Figure 5a). When this option is selected, an entry field appears for the user to upload a file to store in their plasmid. The file is converted into an annotated feature labeled with the file name. The 'openmta.txt' feature encoding the MTA file is 11,218 bp (Figure 5b). We ordered the file storage certificate as synthetic gene fragments and used NEBuilder HiFi DNA assembly to produce the full 16,441 bp plasmid (#192) in one step. The plasmid assembly obtained after sequencing was successfully verified and the text file retrieved with no errors (Figure 5c).

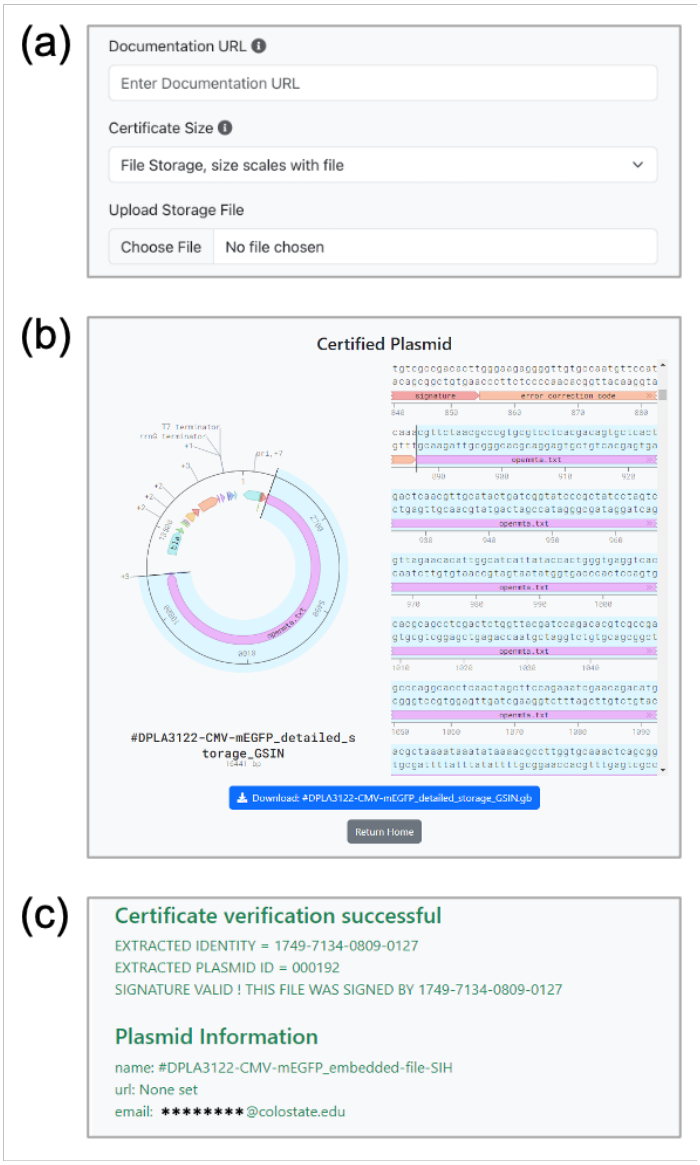


Figure 5. Text files can be embedded in and retrieved from plasmids. (a) Selecting the 'File Storage' certificate size produces a form field to upload the desired file. **(b)** The file is embedded into the plasmid sequence as a feature labeled with the file name (here, 'openmta.txt'). **(c)** Upon sequencing and successful verification, the text file is retrieved alongside the plasmid and author information.

<u>GSIN Certification Does Not Inhibit Functional Protein Expression</u>

We considered that the insertion of several kb of synthetic sequences could impact gene expression. We used spectral flow cytometry to compare EGFP expression in the five different versions of DPLA3122-CMV-mEGFP: uncertified, certified, both self-documenting

plasmids, and the file storage plasmid. We transfected 1 µg of each plasmid solution into HEK293-T17 cells. There was no GFP expression in our unstained negative control group. We performed Dunnett's test to compare each of our sample groups to the unsigned reference group (Figure 6). In terms of the percentage of the cell population that expresses GFP, most certified samples were not significantly different than the unsigned group (~46%). However, samples transfected with Plasmid #192 with the text file had significantly smaller average GFP+ populations around 30% (Figure 6a). Replication of this experiment resulted in the same statistical trends (data not shown).

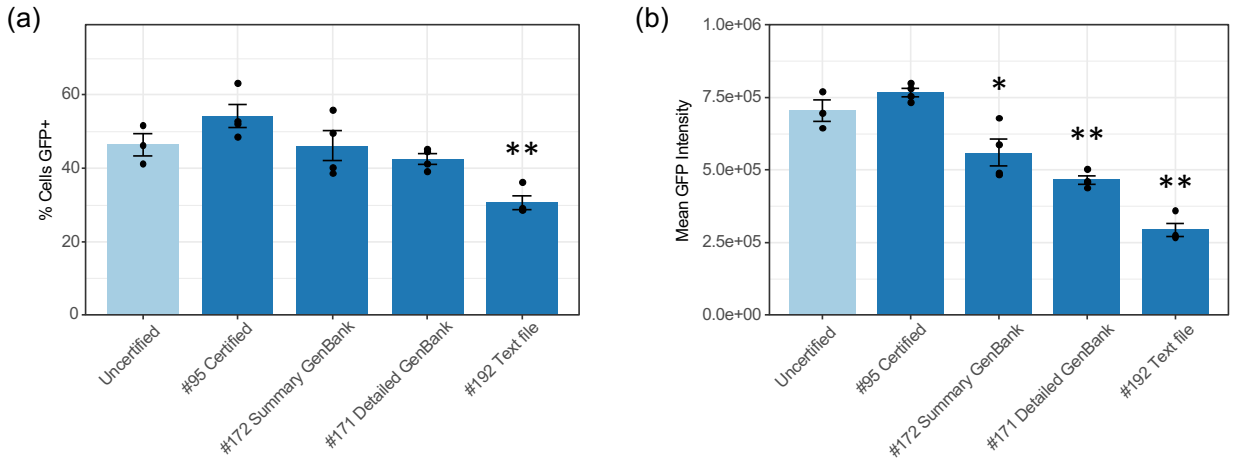


Figure 6. Functional protein expression is maintained after certificate insertion. Each certificate type was inserted into a CMV-mEGFP plasmid backbone. (a) The percentage of the cell population that expresses GFP decreases as plasmid length increases, with the 16 kb text file-encoding plasmid being statistically different than the unsigned reference group. (b) The mean intensity of GFP also decreases as plasmid length increases, with statistically significant differences in the GenBank- and text file-encoding plasmids compared to the unsigned reference group. Data were analyzed using Dunnett's test where * indicates $p<0.05$ and ** indicates $p<0.01$.

The mean GFP intensity in the certified CMV-mEGFP group was not significantly different from the unsigned reference group, whereas the intensity decreased as the size of the plasmid increased (Figure 6b). The sizes of the transfected plasmids ranged from 4,966 to 16,441 bp with decreasing GFP+ cell populations and GFP intensity in larger plasmids.

Performing this experiment with equimolar concentrations of the plasmids provided the same results as the equimass experiments reported here (data not shown). These findings are consistent with studies that show not only that larger plasmids have lower transfection efficiencies, but that fluorescent protein expression (positive population size and intensity) decreases with larger plasmids, even in equimolar experiments [32,33]. It is thought that despite equivalent entry into the cell across plasmid lengths, differences in nuclear translocation and transcription underlie the relationship between plasmid length and gene expression [34]. Nevertheless, functional protein expression is maintained even when an 11 kb certificate is inserted in the plasmid.

Discussion

Sequence verification of plasmids is critical throughout the life science community, but the documentation of DNA sequences is often insufficient to quickly verify the accuracy of a plasmid sequence, especially when assemblies are generated *de novo*. The GSIN application allows individuals to generate and verify certified plasmids to accelerate and simplify the sequence verification of plasmids. We show that insertion of the 164 bp certificate has no effect on DNA yield and that insertion of the larger certificates does not inhibit functional protein expression in mammalian cells, which is consistent with findings from the first iteration of 512 bp certificates that had no effect on bacterial growth or lacZ expression [24].

The advantage of GSIN certificates over simple DNA barcodes is the implementation of a public-key cryptographic approach [11]. It assumes that the verifier is following a *de novo* plasmid assembly protocol with no knowledge of the reference sequence and requires

no specific primers to retrieve the "message" (in this case, whether the sequence is correct). Certified plasmids can be distributed, verified, and used for experimentation without extended communication between the sender and receiver. The documentation provided to the verifier depends on the outcome of the verification process (pass/fail), and the details of the sequence annotations can be limited by the author. This can protect IP and secure information transfer, since the certificate sequence as well as proprietary genetic parts can be left unannotated and thus difficult to parse. On the other hand, since the source and author information are revealed upon verification, it is straightforward for a plasmid verifier to contact the author about any questions they have regarding a sample.

Use of the application by our own laboratory has saved us a significant amount of time by eliminating the need to open each assembly file, align it to the reference sequence, and manually catalogue which errors have occurred. Importantly, the GSIN application links directly to the PlasCAT assembly pipeline we developed for *de novo* plasmid assembly [8,27], providing a simple platform for accurately generating assemblies before verifying them on the GSIN website. This workflow accelerates screening experiments with many variants and replicates to produce accurate samples faster. Here, we validate the GSIN technology and show that it is ready for implementation outside of our laboratory. Users outside our group have been using PlasCAT regularly, and several groups have expressed interest in using the GSIN technology to manage large plasmid libraries. There are no barriers to other groups using the application in its current form, yet we acknowledge there are several avenues to extend the capabilities of the technology.

The primary current limitation to the technology is the inability to correct insertion or deletion (indel) mutations. This is especially important to individuals relying on

nanopore-based DNA sequencing techniques, which are more prone to having indels in their assemblies [8]. Choosing an error correction code involves a tradeoff between the scale of the errors being corrected and the length of the code sequence. For example, the current algorithm requires 16 bp to correct each additional error. Error correction requirements will depend on the downstream applications of the plasmid. Future updates to the application could offer multiple error correction schemes after further experimental validation [35,36].

An additional consideration is that, when encoding a file, the certificate length scales with file size. While very large certificates can be generated in the software, they may be difficult to construct experimentally. Many gene synthesis companies currently impose limitations on the length and complexity of the sequence, yet here we show that ordering the certificates as multiple fragments and ligating them into the final product is a feasible approach to generating large plasmids. To simplify this workflow, the software could implement a screening algorithm to detect and correct problematic sequences within the certificate sequence. Furthermore, a screening algorithm could be developed to detect any sequences of concern or novel genetic parts (e.g., start or stop codons) present in the certificate sequence. As is, the onus is on the plasmid designer to screen their sequences and adhere to not only biological constraints, but federal regulations surrounding DNA sequences before synthesizing the novel construct [37].

An additional feature that would improve the security of distribution of certified plasmids is the ability to limit which users can verify certain plasmids, or limit which users can download the embedded file. We also foresee this technology being expanded to allow for certifying smaller fragments within a plasmid, or for certifying gene networks composed of multiple plasmids. Lastly, the current plasmid IDs are 6 digits, imposing a maximum of 10^6

plasmids with the current architecture, however the plasmid ID could be easily expanded to 8 or 10 digits.

This work focused on developing self-documenting plasmids to improve the sequence verification process, yet it sets the stage for more complex information encoding in plasmids. On one hand, the ability to embed more complex data files within plasmid sequences further improves verification efforts and promotes openness in science. The ease and fidelity of plasmid replication and maintenance in bacteria gives the GSIN scheme an advantage over information-encoding linear DNA fragments and facilitates long-term, verifiable data storage in DNA [11]. Yet, it also opens the door to more nefarious acts. It has already been shown that malicious software ("malware") can be encoded in DNA and unleashed onto the computer that is sequencing the strand of DNA [38]. Files encoding malware or computer viruses could theoretically be embedded in a similar scheme as we used for embedding GenBank and flat text files. This possibility emphasizes the need for improved screening algorithms for synthetic DNA orders, not only to filter out sequences of concern but also to determine whether unknown sequences are innocuous or potentially harmful on their own or in tandem with other constructs.

Concluding Remarks

GSIN certificates simplify sequence verification efforts, provide thorough documentation to plasmid verifiers, and allow plasmid designers to protect their IP while encouraging plasmid distribution. They eliminate the need to track reference sequences. By embedding the plasmid identity and source within the sequence, plasmid verifiers can identify any unknown plasmid, determine who the designer was, and whether any mutations have occurred. We

show that certificates do not adversely affect DNA yield or mammalian gene expression, although further testing is needed to evaluate the practicality of inserting certificates into higher organisms (see Outstanding Questions). Widescale adoption of the GSIN will accelerate plasmid production pipelines, allow the validity of biological products to be frequently confirmed, and improve biosurveillance, for example by simplifying the identification of GMOs.

Acknowledgements

This work was supported by the National Science Foundation [#MCB-2123367 to J.P.]; the National Institutes of Health [R01GM147816 and R21AI168482 to J.P.]; and a generous gift from the Suzanne and Walter Scott Foundation.

Author Contributions

Conceptualization, J.P., D.K., S.P., C.-T.B.; Software, S.P.; Investigation, S.H.; Formal Analysis, S.H., C.-T.B.; Validation, S.H., C.-T.B., J.P.; Writing – Original Draft, C.-T.B.; Writing – Review & Editing, C.-T.B., S.H., D.K., S.P., J.P.; Supervision, J.P., C.-T.B.

Declaration of Interests

J.P. and S.P. have a financial interest in GenoFAB, Inc., a company that may benefit or be perceived to benefit from this publication.

Technology Readiness Box

We developed the GenoFAB Standard Identification Number (GSIN) web application to generate certificate sequences and to verify certified plasmids. Certificates embed identifying information about a plasmid and its author directly in the sequence itself to make self-documenting plasmids that are easy to sequence verify. The GSIN application is publicly accessible to users that register with an email address and password. Users can generate, manage, and verify certified plasmids with optionally embedded GenBank or plain text files. All certificate formats have been experimentally validated and all components of the application function as expected. Beyond the introduction of features that offer enhanced flexibility (i.e., extended error correction capability, built-in screening algorithms) or security at larger scales (i.e., user-specific limitations on verification, increased maximum number of certificates), the application has met its goal of providing a simple process for protecting and verifying plasmid sequences. The technology has moved into the deployment phase corresponding to Technology Readiness Level (TRL) 7. GSIN certificates not only improve the documentation process for plasmid sharing but provide an avenue for protecting intellectual in basic research, clinical development, and bioproduction.

Tables

Table 1. Verification outcomes for certified plasmids.

Plasmid Name	Length (bp)	Plasmid ID	Verified	Notes
DPLA3116	9,013	86	Yes	
DPLA3117	5,307	88	Yes	
DPLA3118	5,298	91	Yes	

DPLA3119	3,119	92	Yes	
DPLA3122	5,223	95	Yes	
DPLA3125	4,376	97	Yes	
DPLA3135	13,724	101	Yes, with ECC	1 mismatch
DPLA3136	5,174	102	Yes	
DPLA3137	5,132	104	Yes, with ECC	2 mismatches
DPLA3140	5,165	107	Yes	
DPLA3142	5,174	108	Yes	
DPLA3146	5,153	112	Yes	
DPLA3148	5,204	115	Yes	
DPLA3152	14,448	116	Yes	
DPLA3182	7,087	120	No, ECC fails	2 bp deletion
DPLA3183	7,076	123	No, ECC fails	3 mismatches
DPLA3187	15,301	124	Yes	
DPLA3188	5,117	126	Yes	
DPLA3190	5,009	130	Yes	
DPLA3192	5,855	135	Yes	

Methods

Certification Software Development

The GSIN web application was developed with Java using Apache Tomcat on the server-side and React.js on the client-side. Firebase Firestore, a non-relational database service from Google Cloud, is used for backend storage. Firebase is also used for user authentication using email and password credentials. The app is deployed on an Amazon Web Services Elastic Beanstalk instance. Whenever a new plasmid is generated in the app, an entry in the database is created. This entry contains the name of the new plasmid, a backup of the GenBank File as text, the ID of the user who generated the file, and the URL.

The algorithm used for certificate generation and verification is as previously published [24], with modifications to how user IDs are generated and used as private keys. The previously published method signed each plasmid using a private key derived from the signer's unique ORCID, necessitating the user to have an ORCID account. Instead, this new

software assigns a unique user identifier (uuid) to each account upon signing up. The uuid is auto generated by the Firebase authentication service, but it does not conform to same 16-digit number format as the ORCID. The Firebase uuid is converted to the ORCID format using a truncated SHA-256 hash. Each users' private key is now derived from the truncated hash instead of the ORCID.

Experimental Validation of Certified Plasmids

Reagents. The Zyppy-96 Plasmid MagBead Miniprep Kit was purchased from Zymo Research (Irving, CA, USA, #D4102). The Oxford Nanopore Rapid Barcoding Kit and R10.4 flow cell were purchased from VWR (USA, #SQK-RBK114.96 and #FLO-MIN114). For the short-reads, the MiSeq Nano V2 cartridges were purchased from Illumina (San Diego, CA, USA. #20060059 and #MS-103-1003). Library preparation kits (ExpressPlex™ Formula M) for the short-reads were sourced from SeqWell (Beverly, MA, USA).

Biological Resources. The plasmids used for library preparation, sequencing, and analysis were obtained from two vendors or were designed and manufactured in house. Eleven plasmids were synthesized and sequence-verified by Twist Biosciences (San Francisco, CA), and five were procured from Addgene (Watertown, MA, USA). The remaining three plasmids were made using fragments and vectors sourced from Twist Biosciences (San Francisco, CA). The array of plasmids was given unique identifiers (e.g., DPLA1234). Primers for plasmid linearization were sourced through Integrated DNA Technologies (Coralville, IA, USA).

Plasmid Construction. To certify the plasmids used in this paper, uncertified plasmids were first linearized via PCR amplification using the NEB Q5 Hot Start High Fidelity 2X Master Mix (Ipswich, MA, USA. #M0494S) according to the manufacturer's instructions with annealing

temperatures determined using the NEB Tm calculator tool and 25 cycles. Post linearization, plasmids were quantified using both a Synergy HTX plate reader for A260/280 measurements and a TapeStation 4200 using the genomic screen tapes and reagents to verify fragment size, purity, and quantity (Santa Clara, CA, #5067-5365 and #5067-5366). The fragment assembly consisted either entirely of fragments sourced from Twist Biosciences (San Francisco, CA) or from fragments and the linear plasmids, using the NEBuilder HiFi DNA Assembly Master Mix (Ipswich, MA, USA., #E2621L) according to the NEBuilder design tool available on the manufacturer's website (with the exception of an increased incubation time to one hour). In brief, equal molar equivalents of fragments were added to a tube with the master mix. The ligation solution was transformed into chemically competent *E. coli* cells and grown overnight.

Plasmid Minipreparation. The plasmid DNA pre- and post-assembly was extracted from each of the 21 isolates on an epMotion 5075 TC liquid handler (Eppendorf, Hamburg, DE) using the Zyppy-96 Plasmid MagBead Miniprep Kit (Zymo Research, Irving CA, USA), according to manufacturer's instructions with the following modifications: pipet mixing during the lysis and neutralization steps, increased Zyppy wash buffer to 800 uL, and an extended elution time of 10 minutes. Samples were all quantified on a Synergy HTX plate reader to determine the quality and quantity of the solutions after purification. Statistical analysis of the DNA yield was performed using R.

Plasmid Isolation and Sequencing. The plasmids were sequenced with the Oxford Nanopore Rapid Barcoding Kit V14.96 according to the manufacturer's specifications using 50 ng per plasmid. The samples were set to run for 24 hours on the MinION with the following parameters: super accurate basecalling, barcode trimming, and basecalling every

hour. The Express Plex Formula M kit was used to prepare the plasmids for short read sequencing according to the manufacturer specifications using 2.5 ng per plasmid. The samples were loaded on to the Illumina MiSeq and demultiplexed to reflect the indexes supplied within the kit. All data was processed into final assemblies using PlasCAT [27].

Transfections and Flow Cytometry. Plasmids were transfected into HEK 293T/17 cells using Lipofectamine 2000 (Invitrogen, ThermoFisher, MA, USA) according to the manufacturer's instructions in a 24-well plate. Cells were transfected at around 70% confluency and the growth media (DMEM supplemented with 10% v/v fetal bovine serum) was changed 4 hours post-transfection. After 24 hours of growth, cells were harvested using TrypLE Express and phosphate buffered saline according to the manufacturer's instructions adjusting the volume for the 24-well plate (ThermoFisher, MA, USA). EGFP expression was quantified using the Aurora spectral flow cytometer (Cytek, Fremont, CA) to measure 10,000 events per sample. Unmixing was performed with autofluorescence extraction, using unstained cells as a control. Statistical analysis of the spectral data collected was performed using R.

Data Availability. GenBank files of all plasmid designs and FASTA sequencing assembly files generated for this study are available at: https://figshare.com/s/fbc52c8931ad07a3c83a.

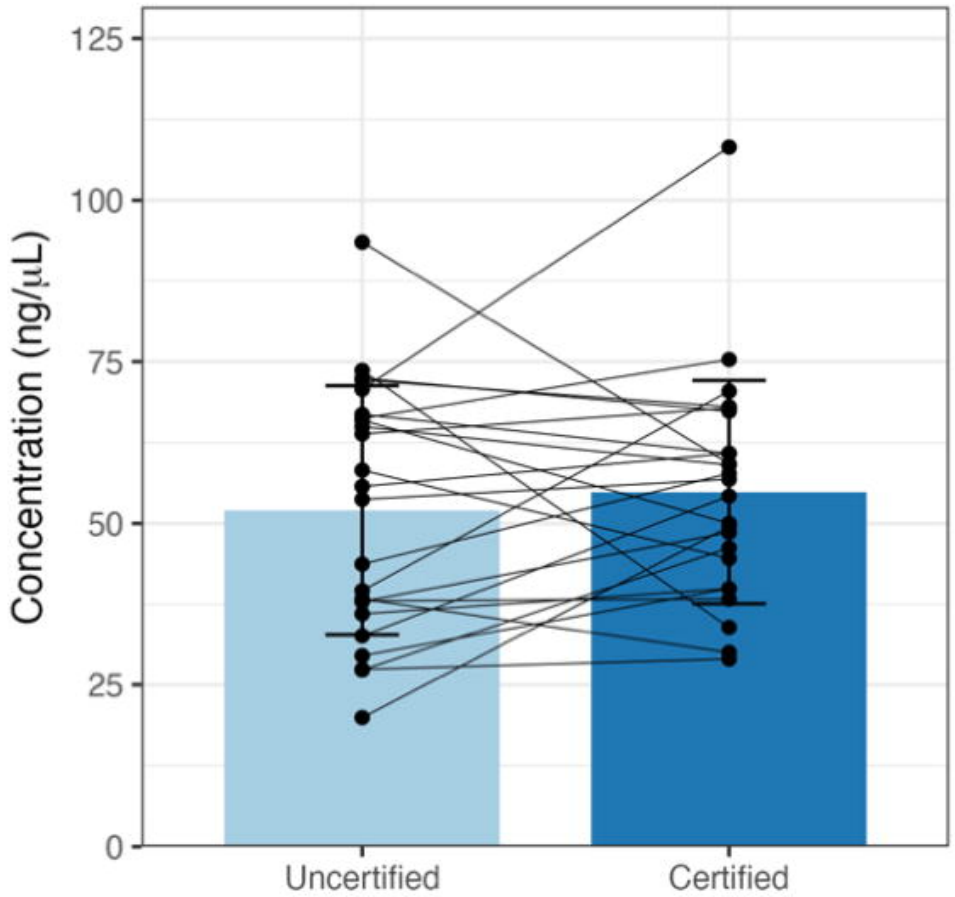
References

1. Peccoud, J. *et al.* (2018) Cyberbiosecurity: from naive trust to risk awareness. *Trends in biotechnology* 36, 4-7
2. Ellis, T. *et al.* (2011) DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr Biol (Camb)* 3, 109-118. 10.1039/c0ib00070a
3. Liu, G. *et al.* (2022) The CRISPR-Cas toolbox and gene editing technologies. *Molecular cell* 82, 333-347

4. Prather, K.J. *et al.* (2003) Industrial scale production of plasmid DNA for vaccine and gene therapy: plasmid design, production, and purification. *Enzyme and Microbial Technology* 33, 865-883. https://doi.org/10.1016/S0141-0229(03)00205-9
5. Liu, Y. *et al.* (2022) In vivo processing of digital information molecularly with targeted specificity and robust reliability. *Science Advances* 8, eabo7415
6. Bai, X. *et al.* (2024) Prevalence of errors in lab-made plasmids across the globe. *bioRxiv*, 2024.2006.2017.596931. 10.1101/2024.06.17.596931
7. Peccoud, J. *et al.* (2008) Targeted Development of Registries of Biological Parts. *PLOS ONE* 3, e2671. 10.1371/journal.pone.0002671
8. Hernandez, S.I. *et al.* (2024) Sequencing Strategy to Ensure Accurate Plasmid Assembly. *bioRxiv*, 2024.2003.2025.586694. 10.1101/2024.03.25.586694
9. Peccoud, J. *et al.* (2011) Essential information for synthetic DNA sequences. *Nature Biotechnology* 29, 22-22. 10.1038/nbt.1753
10. Thuronyi, B.W. *et al.* (2023) No assembly required: Time for stronger, simpler publishing standards for DNA sequences. *Plos Biology* 21, e3002376
11. Berezin, C.-T. *et al.* (2024) Cryptographic approaches to authenticating synthetic DNA sequences. *Trends in Biotechnology*. https://doi.org/10.1016/j.tibtech.2024.02.002
12. Clelland, C.T. *et al.* (1999) Hiding messages in DNA microdots. *Nature* 399, 533-534. 10.1038/21092
13. Church, G.M. *et al.* (2012) Next-generation digital information storage in DNA. *Science* 337, 1628-1628
14. Erlich, Y. and Zielinski, D. (2017) DNA Fountain enables a robust and efficient storage architecture. *science* 355, 950-954
15. Sun, J. *et al.* (2019) Digital information storage on DNA in living organisms. *Medical Research Archives* 7. 10.18103/mra.v7i6.1930
16. Hao, Y. *et al.* (2021) Data storage based on DNA. *Small Structures* 2, 2000046
17. Nguyen, H.H. *et al.* (2018) Long-term stability and integrity of plasmid-based DNA data storage. *Polymers* 10, 28
18. Chen, W. *et al.* (2021) An artificial chromosome for data storage. *National Science Review* 8, nwab028
19. Hu, T. *et al.* (2021) Next-generation sequencing technologies: An overview. *Human Immunology* 82, 801-811. 10.1016/j.humimm.2021.02.012
20. Wong, P.C. *et al.* (2003) Organic data memory using the DNA approach. *Communications of the ACM* 46, 95-98
21. Koch, J. *et al.* (2020) A DNA-of-things storage architecture to create materials with embedded memory. *Nature Biotechnology* 38, 39-43. 10.1038/s41587-019-0356-z
22. Ailenberg, M. and Rotstein, O.D. (2009) An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques* 47, 747-754
23. Hao, M. *et al.* (2020) A mixed culture of bacterial cells enables an economic DNA storage on a large scale. *Communications biology* 3, 416
24. Gallegos, J.E. *et al.* (2020) Securing the Exchange of Synthetic Genetic Constructs Using Digital Signatures. *ACS Synthetic Biology* 9, 2656-2664. 10.1021/acssynbio.0c00401
25. National Library of Medicine (US) - National Center for Biotechnology Information (2004). Standard Nucleotide BLAST.

26. Zhang, Z. *et al.* (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational biology* 7, 203-214
27. Peccoud, S. *et al.* (2024) PlasCAT: Plasmid Cloud Assembly Tool. *Bioinformatics* 40. 10.1093/bioinformatics/btae299
28. Kreiss, P. *et al.* (1999) Plasmid DNA size does not affect the physicochemical properties of lipoplexes but modulates gene transfer efficiency. *Nucleic Acids Research* 27, 3792-3798. 10.1093/nar/27.19.3792
29. Villa-Diaz, L.G. *et al.* (2010) Enhanced transfection efficiency of human embryonic stem cells by the incorporation of DNA liposomes in extracellular matrix. *Stem Cells Dev* 19, 1949-1957. 10.1089/scd.2009.0505
30. Walsh, D.I. *et al.* (2019) Standardizing Automated DNA Assembly: Best Practices, Metrics, and Protocols Using Robots. *SLAS TECHNOLOGY: Translating Life Sciences Innovation* 24, 282-290. 10.1177/2472630318825335
31. Kahl, L. *et al.* (2018) Opening options for material transfer. *Nature Biotechnology* 36, 923-927. 10.1038/nbt.4263
32. Boye, C. *et al.* (2022) Reduction of plasmid vector backbone length enhances reporter gene expression. *Bioelectrochemistry* 144, 107981. https://doi.org/10.1016/j.bioelechem.2021.107981
33. Walker, W.E. *et al.* (2004) The effects of plasmid copy number and sequence context upon transfection efficiency. *Journal of Controlled Release* 94, 245-252. https://doi.org/10.1016/j.jconrel.2003.10.003
34. Hornstein, B.D. *et al.* (2016) Effects of Circular DNA Length on Transfection Efficiency by Electroporation into HeLa Cells. *PLOS ONE* 11, e0167537. 10.1371/journal.pone.0167537
35. Hawkins, J.A. *et al.* (2018) Indel-correcting DNA barcodes for high-throughput sequencing. *Proceedings of the National Academy of Sciences* 115, E6217-E6226
36. Press, W.H. *et al.* (2020) HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints. *Proceedings of the National Academy of Sciences* 117, 18489-18496. 10.1073/pnas.2004821117
37. Fast Track Action Committee for Synthetic Nucleic Acid Procurement Screening of the National Science and Technology Council (2024). Framework for Nucleic Acid Synthesis Screening. In Executive Office of the President of the United States, ed.
38. Greenberg, A. (2017) Biohackers Encoded Malware in a Strand of DNA. *WIRED* Published online 8/10/27

(a)
Generate Certificate
Upload GenBank File
Choose File
No file chosen
Plasmid Name
Enter Plasmid Name
Documentation URL
Enter Documentation URL
Certificate Size
164 bp
Plasmid ID
000193
Author ID
2975-9025-0624-2635
GSIN start sequence
acgcttcgca
GSIN end sequence
gtatcctatg
Submit
Return Home
(b)
Select Certificate placement location
639
#DPLA3117-CMV Int mEGFP_ref
(c)
Certified Plasmid
sig-start
ori
bla
CMV enhancer
rrnG terminator
#DPLA3117-CMV Int mEGFP_ref_164bp_GSIN
5307 bp
sig-sta…
author id
author…
plasmid id
signature
error correction code
sig-end
Download: #DPLA3117-CMV Int mEGFP_ref_164bp_GSIN.gb
Return Home


125
100
75
50
25
0
Concentration (ng/μL)
Uncertified
Certified

(a)

Verify Certificate

Certificate Size

164 bp

Upload FASTA File

Choose File | No file chosen

Submit | Return Home

(b)

Certificate verification failed

START TAG - acgcttcgca OR END TAG - gtatcctatg MISSING

(c)

Certificate verification successful

EXTRACTED IDENTITY = 2975-9025-0624-2635
EXTRACTED PLASMID ID = 000086
SIGNATURE VALID ! THIS FILE WAS SIGNED BY 2975-9025-0624-2635

Plasmid Information

name: #DPLA3116-CMV-T7RNAP-Puro_ref
url: https://figshare.com/ndownloader/files/47633299
email: jpeccoud@gmail.com

(d)

Certificate verification successful after error correction

EXTRACTED IDENTITY = 2975-9025-0624-2635
EXTRACTED PLASMID ID = 000104
SIGNATURE VALID ! THIS FILE WAS SIGNED BY 2975-9025-0624-2635

Original DNA sequence errors.
Erroneous Sequence - agcggcgtgg. Correct Sequence will be - ggcggcgtgg

Erroneous Sequence - acgacggcgt. Correct Sequence will be - gcgacggcgt

Plasmid Information

name: #DPLA3137-CMV-DsRed2_ref
url: https://figshare.com/ndownloader/files/47633356
email: jpeccoud@gmail.com

(e)

Certificate verification failed

EXTRACTED IDENTITY = 2975-9025-0624-2635
EXTRACTED PLASMID ID = 000120
SIGNATURE INVALID !
ERROR CORRECTION ATTEMPTED. FAILED - TOO MANY ERRORS.

(f)

Certificate verification successful

EXTRACTED IDENTITY = 2975-9025-0624-2635
EXTRACTED PLASMID ID = 000126
SIGNATURE VALID ! THIS FILE WAS SIGNED BY 2975-9025-0624-2635

Plasmid Information

name: #DPLA3188-CMV-VSV-P-50-40_ref
url: https://figshare.com/ndownloader/files/47633419
email: jpeccoud@gmail.com

Certificate verification successful

EXTRACTED IDENTITY = 2975-9025-0624-2635
EXTRACTED PLASMID ID = 000127
SIGNATURE VALID ! THIS FILE WAS SIGNED BY 2975-9025-0624-2635

Plasmid Information

name: #DPLA3188-CMV-VSV-P-50-40_ref
url: https://figshare.com/ndownloader/files/47633422
email: jpeccoud@gmail.com

(a)

```
LOCUS       #DPLA3122-CMV-mE        4966 bp DNA     circular SYN 03-MAY-2024
DEFINITION  8DES-Fragment-mEGFP-PolyA(769bp) in pTwist CMV.
ACCESSION   .
VERSION     .
KEYWORDS    .
SOURCE      synthetic DNA construct
  ORGANISM  synthetic DNA construct
REFERENCE   1  (bases 1 to 4966)
  AUTHORS   .
  TITLE     Direct Submission
  JOURNAL   Exported Jun 6, 2024 from SnapGene 7.2.0
            https://www.snapgene.com
COMMENT     Resistance markers: Ampicilin
            This file was created using tools provided by Twist Bioscience
FEATURES             Location/Qualifiers
     source          1..4966
                     /mol_type="other DNA"
                     /organism="synthetic DNA construct"
```

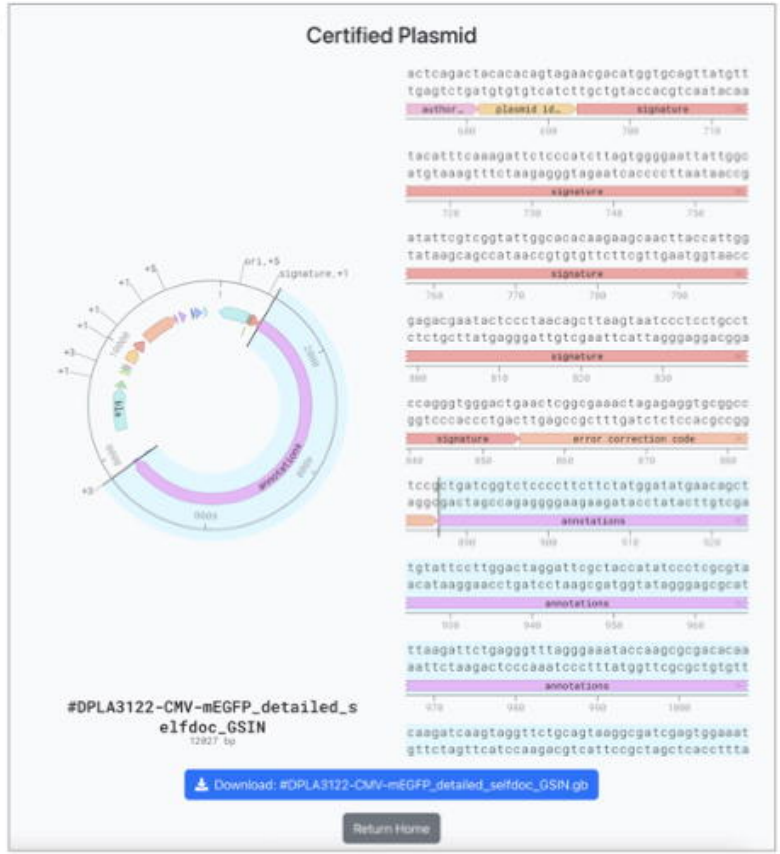
```
CDS             3343..4062
                /codon_start=1
                /product="enhanced GFP with monomerizing A206K mutation
                (Zacharias et al., 2002)"
                /label=mEGFP
                /label=1DES-Part-mEGFP-mammalian
                /note="mammalian codon-optimized"
                /translation="MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTL
                KFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDD
                GNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIK
                VNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSKLSKDPNEKRDHMVLL
                EFVTAAGITLGMDELYK"
```

(b)

```
LOCUS       #DPLA3122        4966 bp DNA     circular SYN 06-JUN-2024
FEATURES             Location/Qualifiers
     source          1..4966
                     /mol_type="other DNA"
```

```
 CDS            3343..4062
                /label=mEGFP
                /label=1DES-Part-mEGFP-mammalian
```

(c)



(d)

Certificate verification successful after error correction

EXTRACTED IDENTITY = 2975-9025-0624-2635
EXTRACTED PLASMID ID = 000171
SIGNATURE VALID ! THIS FILE WAS SIGNED BY 2975-9025-0624-2635

Annotation sequence errors.
Erroneous Sequence - agattggaag. Correct Sequence will be - ggattggaag

Plasmid Information

name: #DPLA3122-CMV-mEGFP_detailed
url: None set
email: jpeccoud@gmail.com

(e)

Certificate verification successful

EXTRACTED IDENTITY = 2975-9025-0624-2635
EXTRACTED PLASMID ID = 000172
SIGNATURE VALID ! THIS FILE WAS SIGNED BY 2975-9025-0624-2635

Plasmid Information

name: #DPLA3122-CMV-mEGFP_summary
url: None set
email: jpeccoud@gmail.com

(a)

Documentation URL

Enter Documentation URL

Certificate Size

File Storage, size scales with file

Upload Storage File

Choose File | No file chosen

(b)

Certified Plasmid

#DPLA3122-CMV-mEGFP_detailed_storage_GSIN

Download #DPLA3122-CMV-mEGFP_detailed_storage_GSIN.gb

Return Home

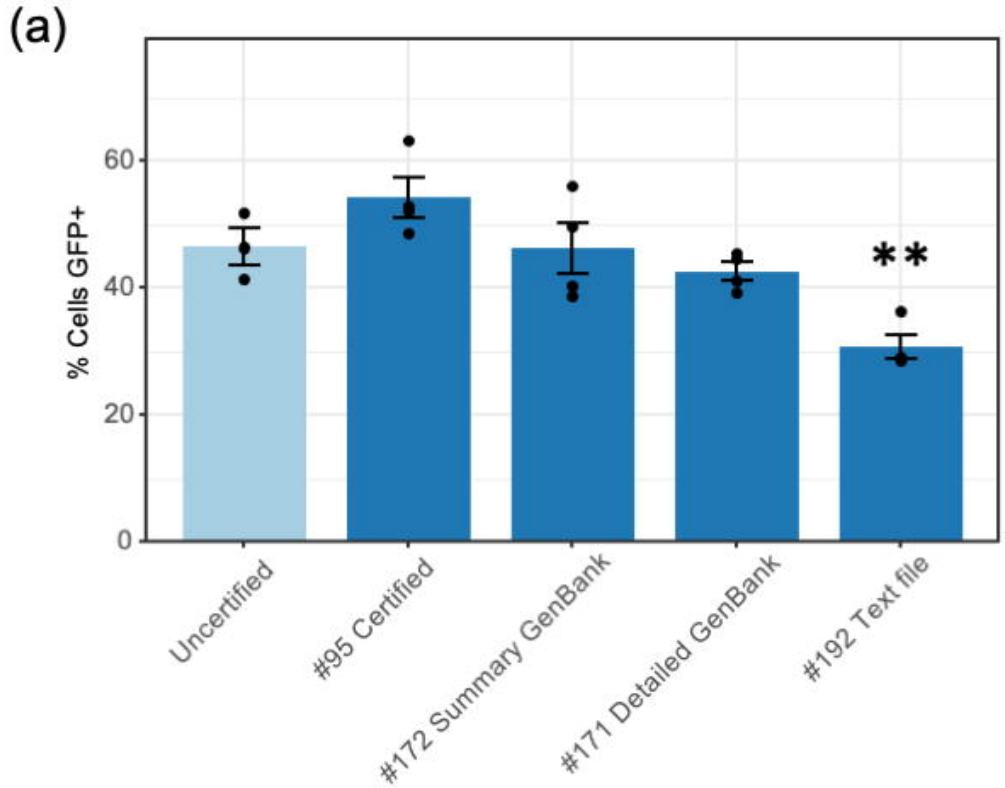
(c)

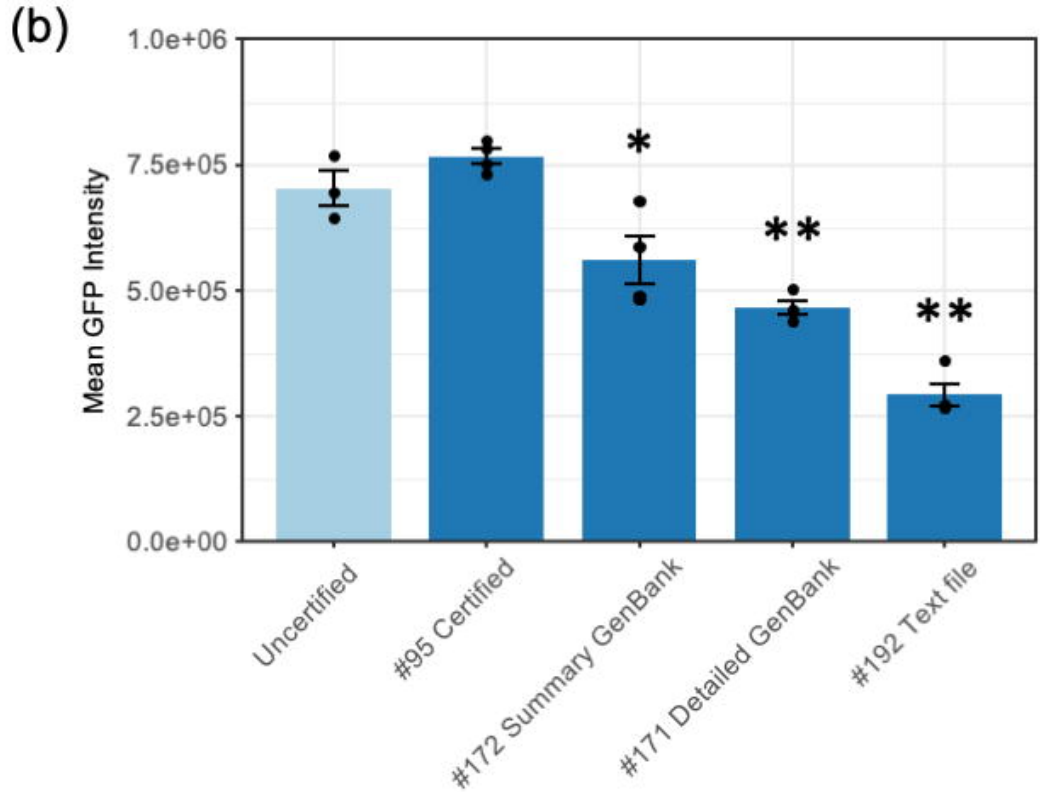
Certificate verification successful

EXTRACTED IDENTITY = 1749-7134-0809-0127
EXTRACTED PLASMID ID = 000192
SIGNATURE VALID ! THIS FILE WAS SIGNED BY 1749-7134-0809-0127

Plasmid Information

name: #DPLA3122-CMV-mEGFP_embedded-file-SIH
url: None set
email: ********@colostate.edu


(a)
% Cells GFP+
0
20
40
60
**
Uncertified
#95 Certified
#172 Summary GenBank
#171 Detailed GenBank
#192 Text file


(b)
Mean GFP Intensity
0.0e+00
2.5e+05
5.0e+05
7.5e+05
1.0e+06
*
**
**
Uncertified
#95 Certified
#172 Summary GenBank
#171 Detailed GenBank
#192 Text file

* p<.05
** p<0.01