Storage-D: a user-friendly platform that enables practical and personalized DNA data storage

Xiaoluo Huang[1#*], Junting Cui[1#], Wei Qiang[1#], Jianwen Ye[3#], Yu Wang[1], Xinying Xie[3], Yuanzhen Li[1], Junbiao Dai[1,2*]

[1]Shenzhen Key Laboratory of Synthetic Genomics, Guangdong Provincial Key Laboratory of Synthetic Genomics, Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

[2]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518000, China
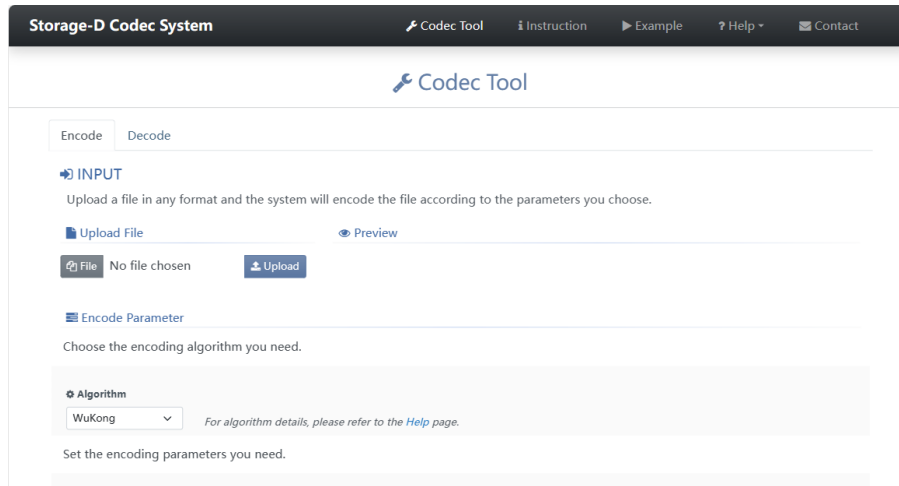
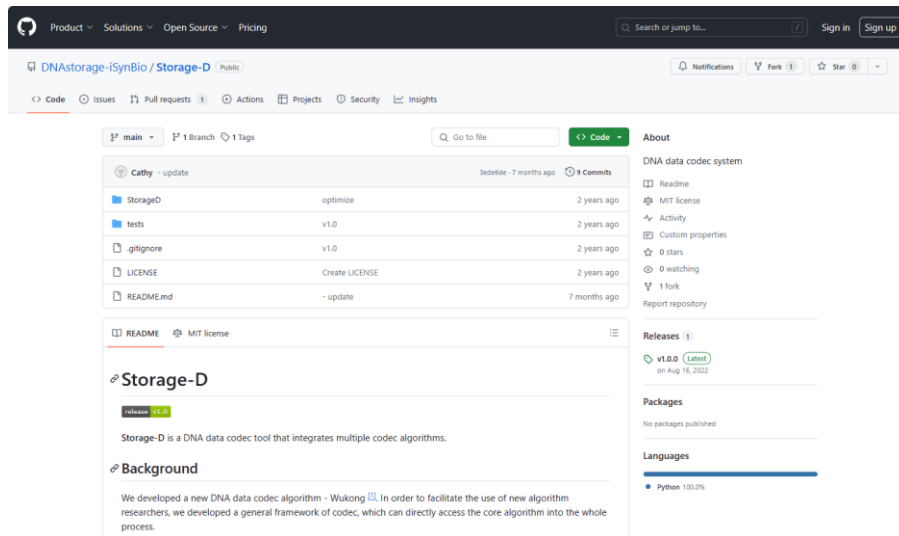[3]School of Biology and Biological Engineering, South China University of Technology, Guangzhou, 510006, China;

website：http://storage.dailab.xyz:16666/


code：https://github.com/DNAstorage-iSynBio/Storage-D/

Deoxyribonucleic acid (DNA) is suggested as a promising data storage medium, but its practical use lacks a user-friendly platform. We developed the "Storage-D" web-based codec platform, which modularized key codec functions and provided a customizable choice for practical data storage use. A novel algorithm called "Wukong" was developed and integrated into the platform, which employed a flexible encoding logic and was able to generate a large number of encoding rules that can be employed for various DNA data storage demands. We verified the algorithm through in vitro and in vivo experiments, and provided the website service and source code.

The web server and codes of the platform are available at http://storage.dailab.xyz:16666/ and https://github.com/DNAstorage-iSynBio/Storage-D/, respectively.
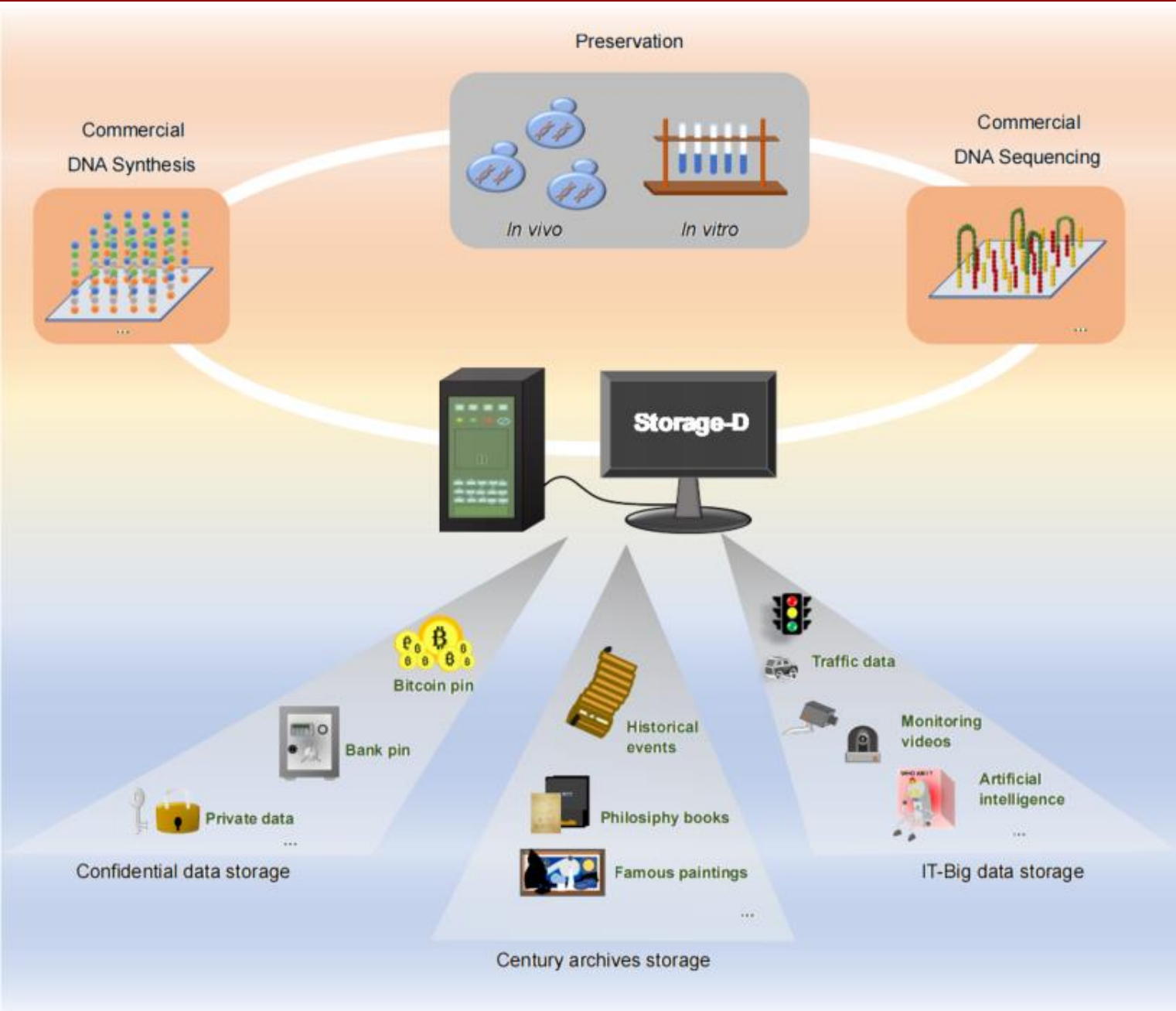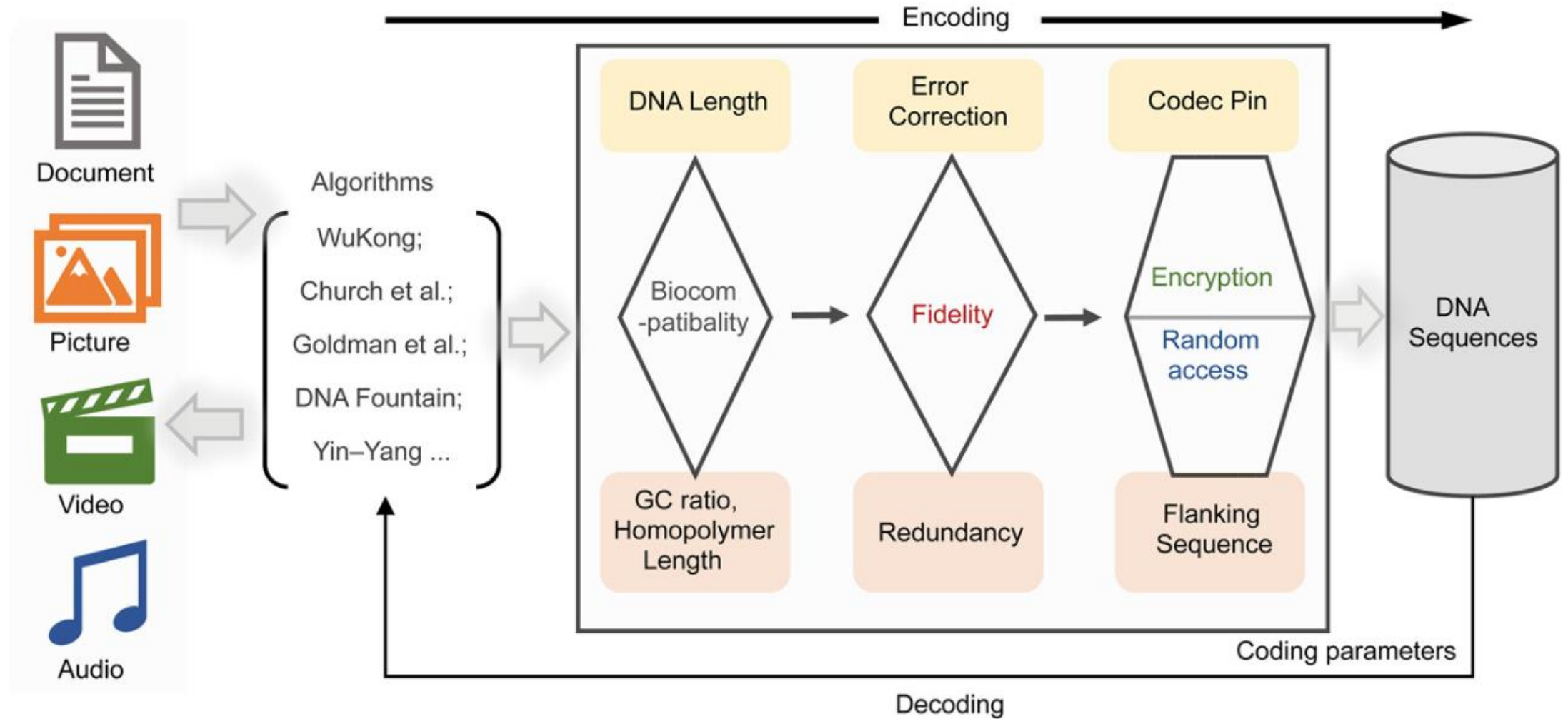
➢ Deoxyribonucleic acid (DNA) data storage exhibits remarkable advantages, such as high density and long lifespan, and is suggested to be one of the most promising media for coping with future data storage crises.

➢ A user-friendly platform, "Storage-D" was developed, which enables users to store any format of practical data with personalized choice.

➢ Specifically, a new algorithm, named "Wukong" was developed, which contains a sizeable collection of codec pins that enables encoding data into any DNA sequence with considerable privacy.

➢ The "Wukong" algorithm shows overall better performance than earlier algorithms in encoding a suitable length of DNA sequence matching downstream biochemical working flow for in vitro and in vivo storage.

➢ The tool provides an open-frame for integrating other codec algorithms and can be easily connected to commercial DNA synthesis and sequencing platform for building a complete pipeline for practical data storage into DNA. The web server and codes of the platform are available at http://storage.dailab.xyz:16666/ and https://github.com/DNAstorage-iSynBio/Storage-D/, respectively.

The working flow of Storage-D

The interface of Storage-D

The "Wukong" codec algorithm implemented in Storage-D

➤ Encoding with different algorithms
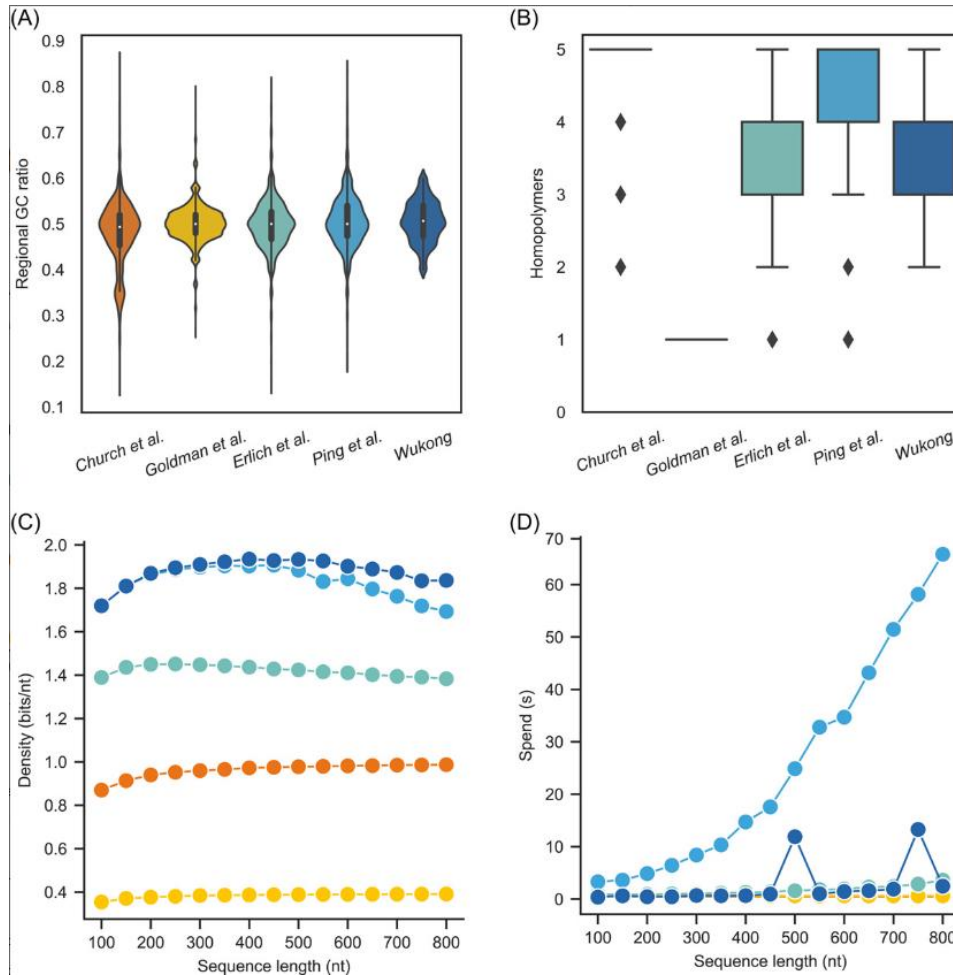
➤ Comparison between "Wukong" and other codec algorithms.

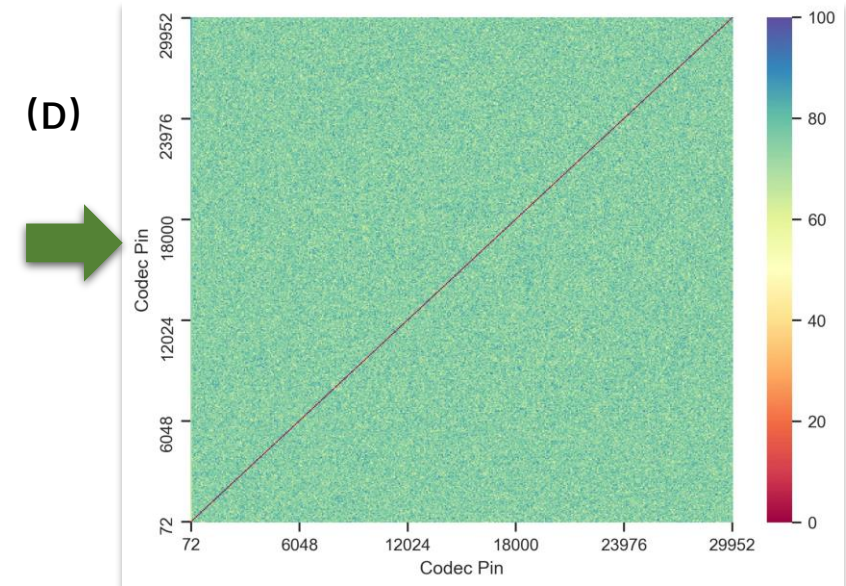| General features	Details	Church et al.	Goldman et al.	Erlich et al.	Ping et al.	Wukong
Density	Coding potential*	1	1.58	1.98	1.95	1.98
Biochemical Compatibility	Regional GC (%)	15−85	26−79	15−80	20−84	40−60 or defined
	Homopolymer	4	1	4	4	4 or defined
Data Fidelity	Error correction	NO	NO	Yes	Yes	Yes or defined
	Redundancy	NO	Yes	Yes	Yes	Yes or defined
Encryption	Codec Pin	--	--	--	1536[#]	> 20 Trillion
Random-Access	Primer Design	--	--	--	--	Yes or defined

Encode data with different codec pins
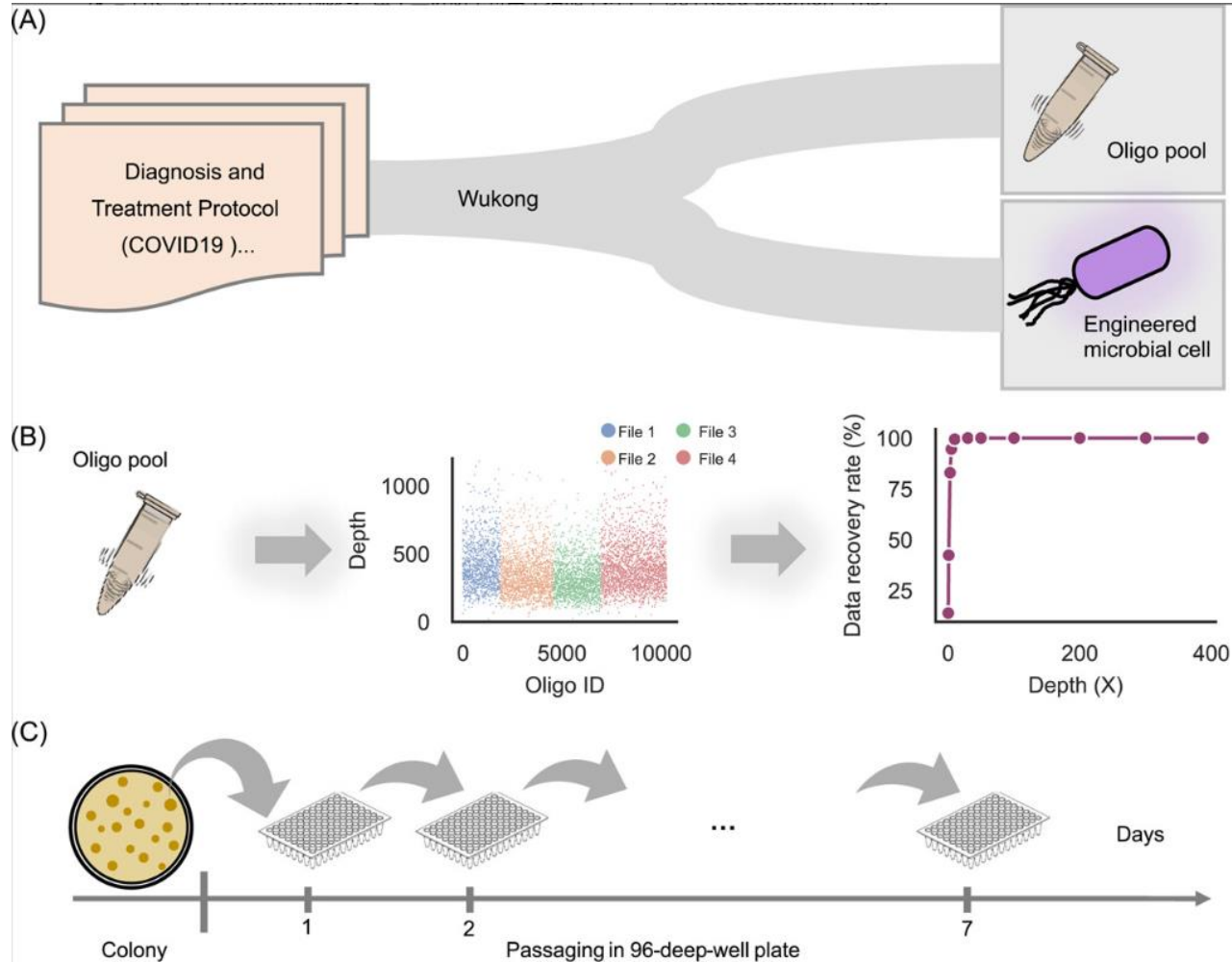
(A)
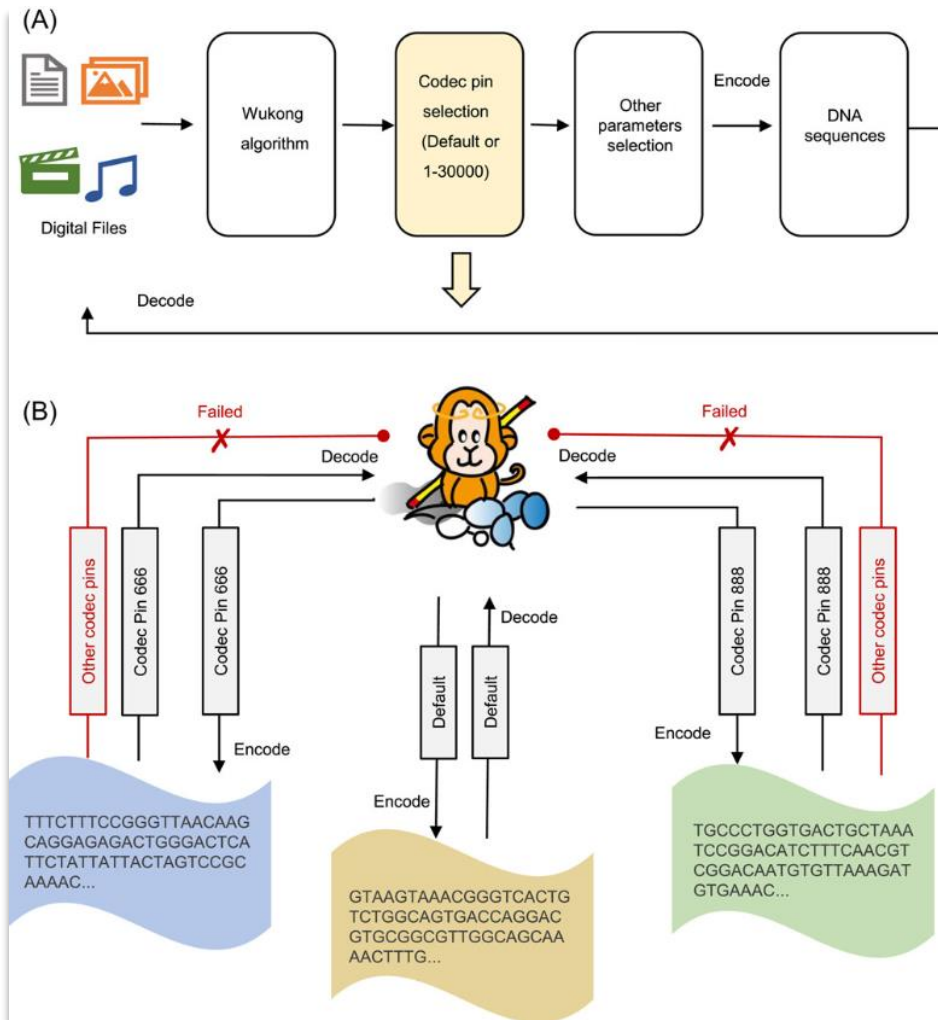
Encode data with different codec pins

(C)

(B)

(D)
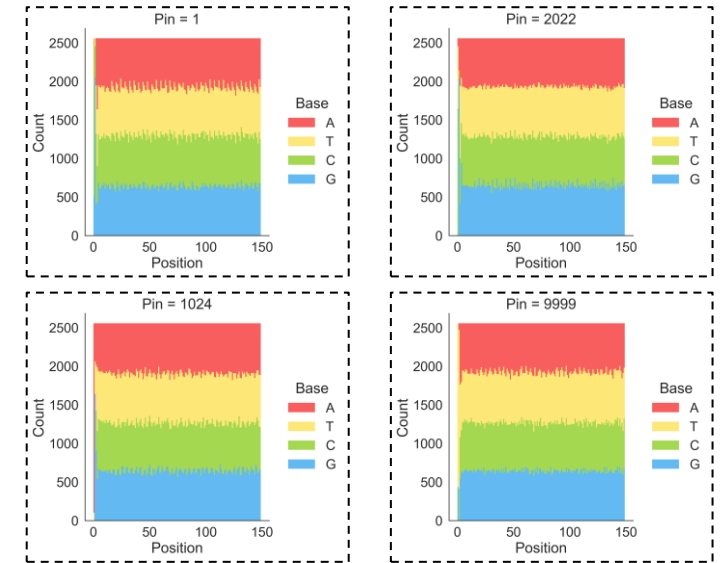
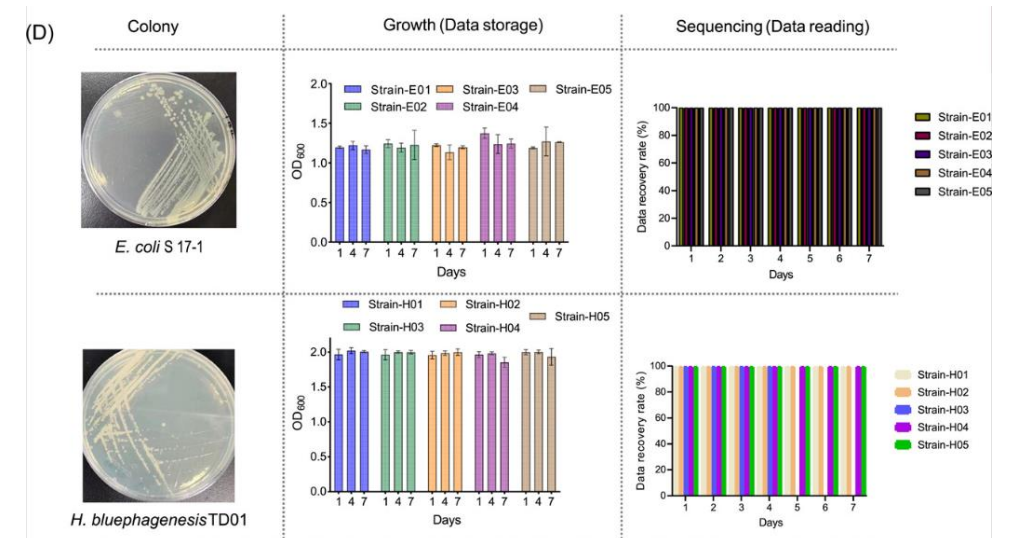Heatmap of hamming distance between encoded DNA sequences by different codec Pins

Validation of "Storage-D" by practical data storage

(A) Illustration of data storage in vitro and in vivo by "Wukong" algorithm

(B) The oligo-pool is read by sequencing

(C) Illustration of data passage in vivo.

(D) The growth state and data recovery rate obtained through sequencing

Summary

❑ "Storage-D" modularizes the basic functions of DNA data storage, such as encoding and decoding, error correction, and random access, and integrates four existing algorithms.

❑ The "Wukong" algorithm is capable of encoding DNA into diverse lengths with controlled homopolymer runs and regional even GC content while maintaining a high coding potential and offering a unique codec pin library. The feasibility of the "Wukong" algorithm was verified through computational simulations and in vitro and in vivo experiments.

❑ A user-friendly online web service and source code available for local download and integration are provided.

iMeta: Integrated meta-omics to change the understanding of the biology and environment

<WILEY>WILEY</WILEY>

"*iMeta*" is an open-access Wiley partner journal launched by scientists of the Chinese Academy of Sciences. iMeta aims to promote metagenomics, microbiome, and bioinformatics research by publishing original research, methods, or protocols, and reviews. The goal is to publish high-quality papers (Top 10%, IF > 15) targeting a broad audience. Unique features include video submission, reproducible analysis, figure polishing, APC waiver, and promotion by social media with 500,000 followers. Three issues were released in March, June , and September 2022.

Society: http://www.imeta.science
Publisher: https://wileyonlinelibrary.com/journal/imeta
Submission: https://mc.manuscriptcentral.com/imeta

office@imeta.science

iMeta

iMetaScience

iMetaScience