nature methods

Article

https://doi.org/10.1038/s41592-023-01868-1

Fast and efficient template-mediated synthesis of genetic variants

Received: 18 July 2022

Accepted: 29 March 2023

Published online: 01 May 2023

Check for updates

Liyuan Liu [1,4], Yiming Huang [1,2,4] & Harris H. Wang [1,3]

Efficient methods for the generation of specific mutations enable the study of functional variations in natural populations and lead to advances in genetic engineering applications. Here, we present a new approach, mutagenesis by template-guided amplicon assembly (MEGAA), for the rapid construction of kilobase-sized DNA variants. With this method, many mutations can be generated at a time to a DNA template at more than 90% efficiency per target in a predictable manner. We devised a robust and iterative protocol for an open-source laboratory automation robot that enables desktop production and long-read sequencing validation of variants. Using this system, we demonstrated the construction of 31 natural SARS-CoV2 spike gene variants and 10 recoded *Escherichia coli* genome fragments, with each 4 kb region containing up to 150 mutations. Furthermore, 125 defined combinatorial adeno-associated virus-2 *cap* gene variants were easily built using the system, which exhibited viral packaging enhancements of up to 10-fold compared with wild type. Thus, the MEGAA platform enables generation of multi-site sequence variants quickly, cheaply, and in a scalable manner for diverse applications in biotechnology.

Construction and manipulation of kilobase-sized DNA building blocks is fundamental to synthetic biology and synthetic genomics[1,2]. At the gene and pathway level, synthetic or engineered sequences can be applied to a design-build-test-learn (DBTL) framework to optimize for a desired function[3,4]. At the genome scale, de novo synthesis and genome assembly can be used to explore synthetic genome designs[5–9].

However, despite significant advances in DNA synthesis over the last two decades[10], current methods are still restricted by size limits, synthesis fidelity and long lead times. In addition, the construction of synthetic DNA remains expensive, labor intensive and impractical for gigabase genomes such as those of animals and plants[11]. And although breakthroughs in deep learning and computational design have now enabled the generation of thousands to millions of synthetic variants[12,13], most protein-sized sequences cannot be synthesized and experimentally tested at scale, which underscores an important unmet need in the field.

From a conceptual perspective, de novo synthesis is fundamentally ill-suited for making gene variants in which specific mutations are dispersed across a wild-type core sequence. This is because significant time and resources are wasted in making a common core sequence from scratch. Beyond de novo synthesis, current alternatives have numerous drawbacks. Strategies using cellular machinery such as double-stranded DNA recombineering[14], base editing[15] or prime editing[16] require the assembly of complicated constructs and are not readily multiplexable (that is, able to target more than two distinct sites at a time). Other published or commercial mutagenesis protocols using primers can produce only a few mutations at a time at best and often require an existing cloned vector[17–20]. More multiplexable oligonucleotide-mediated allelic replacement methods[21,22] rely on DNA transformation into cells and the screening of many colonies, which adds significant time, labor and cost burdens.

[1]Department of Systems Biology, Columbia University, New York, NY, USA. [2]Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, NY, USA. [3]Department of Pathology and Cell Biology, Columbia University, New York, NY, USA. [4]These authors contributed equally: Liyuan Liu, Yiming Huang. e-mail: hw2429@columbia.edu

To address these shortcomings, we present here a new in vitro variant synthesis platform, mutagenesis by template-guided amplicon assembly (MEGAA), which can produce kilobase-long sequence variants rapidly, in a scalable manner, and at high fidelity.

Results

Overview of template-mediated variant synthesis

MEGAA uses a seed DNA material to generate an initial template for subsequent annealing, extension and ligation of oligonucleotide pools that carry mutations of interest (Fig. 1a). The generated variant is then specifically amplified against the initial template to yield the final high-fidelity product. MEGAA exploits the specificity for uracil bases conferred by Q5 versus Q5U DNA polymerase. Q5 and similar Archaeal polymerases bind tightly to uracil nucleotides, which stall DNA polymerization[23], and therefore cannot extend off uracil-containing DNA templates. In contrast, Q5U is a modified Q5 DNA polymerase that contains a mutation in the uracil-binding pocket to enable amplification of DNA containing uracil and inosine bases[24]. In the first step of MEGAA, the input seed DNA is amplified by polymerase chain reaction (PCR) using a Q5U hot-start high-fidelity DNA polymerase, in which dTTPs (deoxythymidine triphosphates) are substituted with dUTPs (deoxyuridine triphosphates). This results in MEGAA templates with all thymine bases replaced by uracil bases. In the second step, the uracil-containing template is combined with Taq DNA ligase, Q5U hot-start high-fidelity DNA polymerase, dNTP (deoxyribonucleotide triphosphate), and the desired mutagenic pool of oligonucleotides and a forward extension primer at 500–1,000-fold molar excess of the template as a single-pot reaction in a compatible buffer (Methods). Then the oligonucleotide annealing, extension and ligation reactions proceed. Given that the Q5U polymerase does not exhibit strand displacement activity or 5′→3′ exonuclease activity, once mutagenic oligonucleotides are annealed to the template, the polymerase will fill gaps only between oligonucleotides and enable subsequent ligation by Taq DNA ligase. Rapid oligonucleotide annealing is performed from 95 °C down to 4 °C at a rate of 3 °C s^{-1}. Fast annealing with excess oligonucleotides is crucial to avoid renaturation of the uracil-template DNA. Furthermore, this prevents Taq ligase from unwarranted ligation before the single-stranded variant allele is fully gap-filled. In the last step, the assembled single-stranded variant allele, which has incorporated the mutagenic oligonucleotides, is amplified by PCR using a Q5 hot-start high-fidelity DNA polymerase while avoiding amplification of the initial uracil-containing template. The specifically amplified variant amplicon from the MEGAA reaction can be used directly in downstream applications (for example, cloning, sequencing or transformation) (Fig. 1a).

To analyze the full-length MEGAA products accurately, rapidly and in parallel, we developed a low-cost long-read sequencing pipeline using the Oxford Nanopore MinION platform with a PCR barcoding scheme that enabled multiplexing of up to 96 samples per run (Supplementary Fig. 1). A custom variant-calling pipeline was used to assess MEGAA efficiency across target sites (Methods). With this set-up, MEGAA products can be analyzed cost-effectively (~US$2.90 per sample), with sufficient accuracy for unique variant identification, and with a reduced turnaround time (from 2 days by Sanger sequencing to only ~2 hours).

We first piloted the ability of MEGAA to generate variants using oligonucleotide pools containing 1, 3, 6 or 9 oligonucleotides for a 1,192 bp DNA template (*rsgA* gene from *Escherichia coli* K-12), with each oligonucleotide (20–39 nucleotides) containing a 2–5 base substitution of the template sequence (Supplementary Table 1). The efficiency and completeness of each variant synthesis by MEGAA was assessed using nanopore sequencing. Variants from oligonucleotide pools containing 1 oligonucleotide were generated at >90% efficiency, while >70% were generated completely in a 3-pool reaction, 35% were generated completely in a 6-pool reaction, and 2.5% were generated completely in a 9-pool reaction with ~25% of variants having 8 or 9 mutations. (Supplementary Fig. 2a). In larger oligonucleotide pools (for example, 9-pool), we noted that targets near the 5′ region were generally more efficiently converted than those at the 3′ region (Supplementary Fig. 2b). We included a head-to-head experiment to compare MEGAA efficiency with other commercial directed mutagenesis kits. For single-site mutagenesis, MEGAA efficiency was higher than commercial kits, (that is, 93.5% for MEGAA versus 85.2%, 87.6% and 80.7% for Q5 Site-Directed Mutagenesis Kit, QuikChange II Site-Directed Mutagenesis Kit and QuikChange Lightning Multi Site-Directed Mutagenesis Kit, respectively) (Supplementary Fig. 2a). Importantly, MEGAA was substantially more efficient than the commercial kits in multiplex target reactions. Although only 3.6% of sequences were completely mutated for a 6-oligonucleotide pool reaction in the QuikChange Lighting reaction, 35.4% of sequences in MEGAA were completely mutated.

Next, we tested the capacity of MEGAA to work on templates of different sizes ranging from 1 kb to 13 kb. We generated 16 uracil-templates of different sizes (*rsgA*1–*rsgA*16) by amplifying the *rsgA* gene region of the *E. coli* K-12 genome. Then, we performed separate MEGAA reactions using the same 9-oligonucleotide pool designed against a shared 1 kb region across the different sized uracil-templates (Fig. 1b). In general, MEGAA products had robust amplicon bands at the expected sizes. Templates larger than 10 kb (for example, *rsgA*16) did not produce a detectable amplicon. From nanopore sequencing of MEGAA products, we observed varying levels of completeness in MEGAA product yield, with more than 40% of all products having at least five of nine targets converted for almost all of the templates (Fig. 1b). The mean MEGAA efficiency across all target sites reached as high as 75% for the *rsgA*6 template (1,192 bp) and as low as 29% for the *rsgA*15 template (9,156 bp) (Fig. 1c). We observed target-specific differences in MEGAA efficiency that were consistent across all template sizes. In general, 5′ targets were more efficiently generated than 3′ targets, suggesting that global factors govern oligonucleotide assembly (Fig. 1c). Targets s5 and s7 were less efficiently generated than expected, which implies that local oligonucleotide annealing factors are also at play. Overall, the size of the template correlated with MEGAA efficiency, with shorter templates being more efficiently converted (Fig. 1d). To verify that these results hold true for another template, we repeated the experiment on 16 additional templates (*pheS*1–*pheS*16) derived from the *E. coli* K-12 genome near the *pheS* gene using a 12-oligonucleotide pool. The same trends were observed: most target sites were generated at high efficiency with some variation in some targets (Supplementary Fig. 3). Together, these findings indicate that MEGAA is efficient and multiplexable across different templates of up to 10 kb in length and has the potential for further improvements.

Optimization of variant synthesis and iterative cycling

We propose that the reduced MEGAA efficiency near the 3′ region of templates was due to extension of the template without having the oligonucleotides annealed in their correct place. Therefore, a strategy was devised in which oligonucleotides were designed to have a gradation of melting temperatures (T_m), with 5′ oligonucleotides having the lowest T_m (47 °C) and 3′ oligonucleotides having the highest T_m (64 °C) (Supplementary Fig. 4). This strategy should support a more ordered assembly process whereby 3′ oligonucleotides first anneal to the uracil-template before 5′ oligonucleotides, which would increase the likelihood of generating a more fully converted variant. We performed a head-to-head comparison of this new oligonucleotide design (Design-2) with the prior oligonucleotide design (Design-1), in which all oligonucleotides had the same T_m. With Design-2 oligonucleotides, the resulting MEGAA variants had a notably improved mean MEGAA efficiency of 86% per target (versus 75% for Design-1), with nearly all of the target positions performing better, especially s7 (Supplementary Fig. 4). We also tested the opposite oligonucleotide design (Design-3), with 5′ oligonucleotides having the highest T_m and 3′ oligonucleotides having the lowest T_m, which yielded even lower oligonucleotide incorporation

https://doi.org/10.1038/s41592-023-01868-1

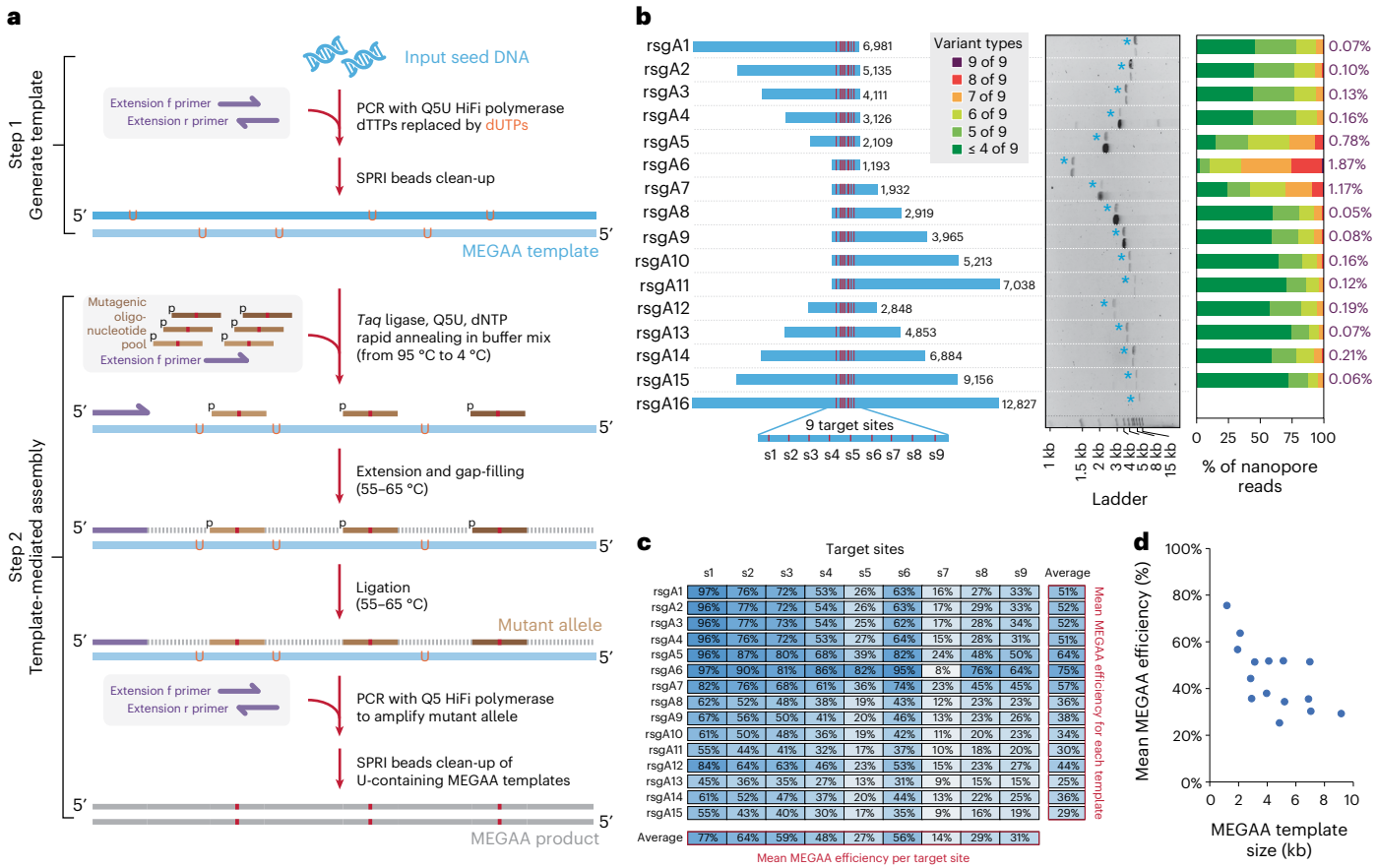


Fig. 1 | The MEGAA method for DNA variant synthesis. a, Overview of the MEGAA protocol. **b**, Testing *rsgA* templates of different lengths and target positions (left panel; the numbers on the right side of the bar show the sizes of the fragments, and the span of the bar represents the start and end positions of the fragments) with their corresponding MEGAA reaction products (middle panel; blue asterisks indicate U-containing templates), and variant generation efficiency (right panel; the numbers in purple on the right side of the bar show the percentage of fully complete variants). **c**, Efficiency of MEGAA per target site across different *rsgA* templates. **d**, Correlation between mean MEGAA efficiency across nine targets versus *rsgA* template size.

at the 3′ region and thus confirmed our oligonucleotide design principle (Supplementary Fig. 5). Furthermore, using the Design-2 strategy we showed that MEGAA could operate on templates with guanine–cytosine content ranging from 29% to 63%, yielding mean conversion rates per target of 91.7% and 81.0%, respectively (Supplementary Fig. 6). To characterize off-target mutations in MEGAA products, we then cloned products into shuttle vectors, transformed them into cells, and isolated selected colonies for Sanger sequencing, which did not show any additional mutations outside of the MEGAA target sites.

Conceptually, MEGAA could be repeatedly cycled such that the output from one round is used as the direct input of the next round, which could further enhance MEGAA product conversion towards the target genotype (Fig. 2a). Using the *rsgA*6 template and Design-1 or Design-2 for 9-oligonucleotide pools, we developed and tested a protocol whereby the MEGAA product from the prior round is reamplified into uracil-containing templates for the next round of MEGAA reactions without any laborious cell transformation or clonal purification steps (Supplementary Methods). For Design-1 oligonucleotides, as more MEGAA rounds are performed, the fraction of fully converted variants increased, reaching near completion after the fifth round (Fig. 2b and Supplementary Fig. 7). For Design-2 oligonucleotides, the desired variant was almost completely generated after only two or three rounds, highlighting the substantially improved performance using the more optimized oligonucleotide design, and we verified select cloned variants by Sanger sequencing (Supplementary Fig. 8). Importantly, the conversion state of the variant product over multiple MEGAA cycles can be modeled using a simple binominal distribution (Methods). For Design-2 oligonucleotides, our experimental data match the model prediction of an overall MEGAA efficiency per site of 80–90% per cycle, while Design-1 oligonucleotides have a more variable MEGAA efficiency of 50–70% (Fig. 2c). Although iterative PCRs during MEGAA cycling could potentially accumulate amplification errors, we can assess the DNA sequence fidelity as a function of PCR cycles, which shows that PCR-associated errors are minimal for five MEGAA cycles (for example, >89% of a 2 kb template maintains the perfect sequence for five cycles) (Supplementary Fig. 9). Therefore, MEGAA can be computationally modeled and experimentally tuned to generate high-fidelity variants of different levels of mutational saturation across a population.

Automated desktop construction of DNA sequence variants

To generalize and standardize our variant synthesis platform, we used a low-cost open-source liquid handling and nucleic acid amplification workstation (Opentrons OT-2) to execute MEGAA reactions in an automated end-to-end pipeline dubbed MEGAAtron (Fig. 2d and Supplementary Table 2). First, a MEGAA design tool (MEGAA-dt) was developed to generate sequences of mutagenic oligonucleotides based on input templates and desired changes by automatically optimizing for high MEGAA efficiency and low resource requirements (Supplementary Fig. 10 and Methods). MEGAA oligonucleotides are ordered from commercial vendors individually or as premixed pools. Reagents, templates and oligonucleotides are loaded onto the MEGAAtron robotic system, which can produce 24 different variants in a single run, including all steps

https://doi.org/10.1038/s41592-023-01868-1

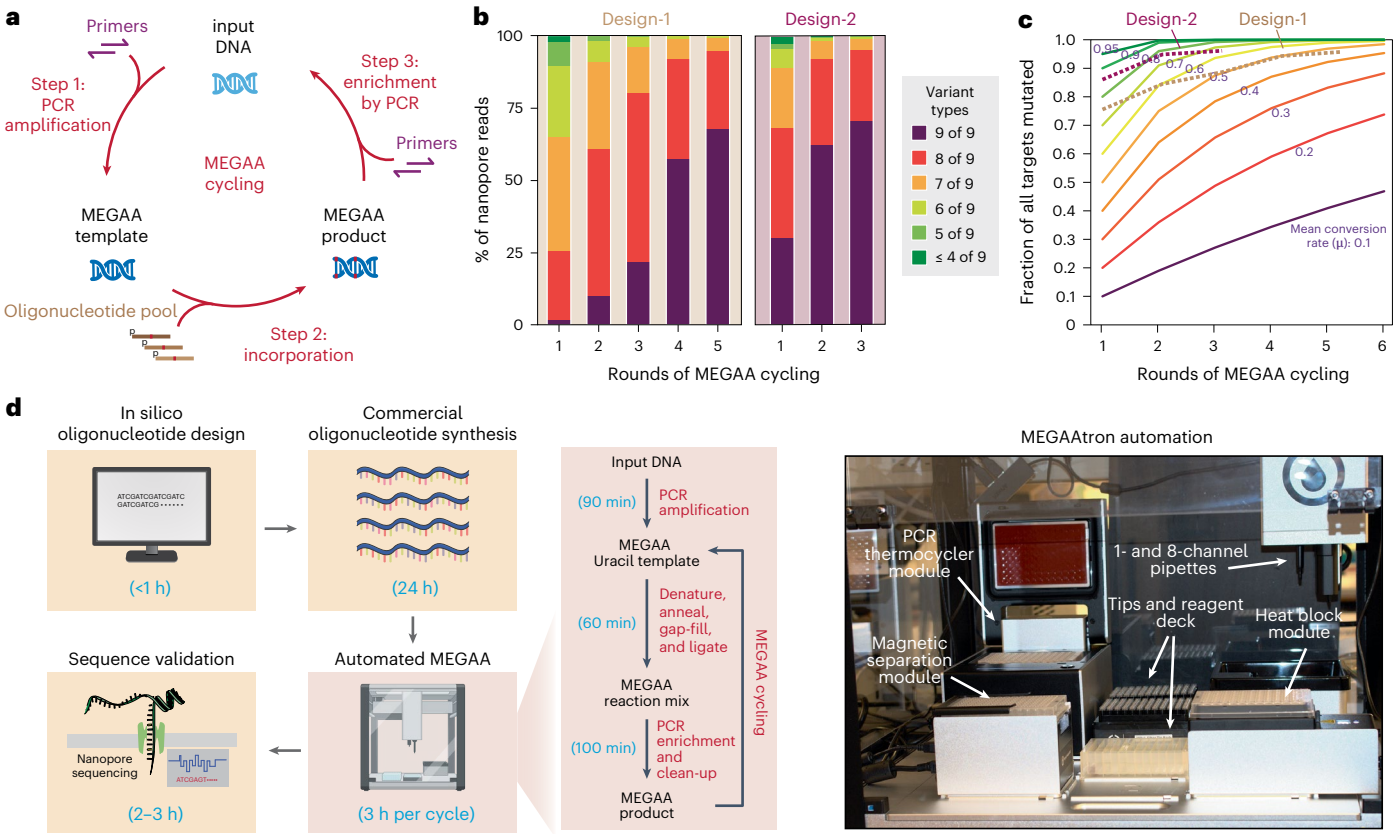


Fig. 2 | MEGAA cycling, optimization, modeling and automation. a, Schematic diagram of MEGAA cycling to regenerate inputs for additional rounds. **b**, Variants generated across an increasing number of MEGAA rounds, with random oligonucleotide annealing design (Design-1) and ordered oligonucleotide annealing design (Design-2). **c**, Modeling of MEGAA cycling efficiency using a binomial process to assess the fraction of all target sites converted at a given mean conversion rate (μ). Solid lines are population distributions at different conversion rates (0.1–0.95) predicted by the model. The dotted lines represent Design-1 data and Design-2 data over multiple MEGAA cycles. **d**, The MEGAAtron platform to automate the design and synthesis of variants and their validation by nanopore sequencing.

of the protocol through a MEGAA round (or multiple rounds) from PCR amplification to product purification (Fig. 2d). The resulting MEGAA products are assessed using nanopore sequencing for quality control and efficiency characterization. The overall turnaround time of the pipeline once all inputs are ready (for example, oligonucleotide pool, initial template) is less than 6 hours, with a cost starting from ~US$20 per variant (depending on variant type) including oligonucleotides, consumables, and sequencing, which is 10-fold cheaper than commercialized gene synthesis (Supplementary Fig. 11). To obtain 100% complete clonal variants, an additional cloning step can be performed, and a minimal amount of colony sequencing is needed to identify the desired variant based on nanopore sequencing analysis (for example, three out of four colonies selected are expected to contain the perfect variant from nanopore reads). To highlight the advantages of MEGAA, we systematically compared the MEGAA platform with other commercial kits and published methods[20] (Supplementary Table 3 and Supplementary Fig. 12).

Gene- and genome-scale templated variant synthesis

Using the MEGAAtron system, we sought to showcase gene- to genome-scale uses of MEGAA for fast and cheap templated variant synthesis. We first explored a new capacity to generate viral variants that would otherwise require total gene synthesis. Fast variant production of key viral components can facilitate the testing of neutralizing antibodies and therapies against variants[25,26] and help establish zoonotic transmission paths for better pandemic preparedness[27]. We chose the 3,822 bp *S* gene that encodes the spike (S) protein from the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus, which has been extensively characterized by surveillance sequencing during the ongoing global pandemic since late 2019. We assimilated a set of 31 representative natural *S* gene variants from different SARS-CoV2 lineages from around the world, encompassing major variants of interest and variants of concern as of September 2021 (Fig. 3a). Across the 31 *S* gene variants, 66 unique mutations are present, with some variants containing up to 13 substitutions and deletions. Oligonucleotides were commercially synthesized for each target site and separately pooled to produce their respective variants on the MEGAAtron system. Within 12 hours and two rounds of MEGAA reactions, we successfully generated all *S* gene variants to a high degree of saturation as assessed using full-length nanopore sequencing (Fig. 3a and Supplementary Table 4). In 27 of 31 variants, we observed the correct complete variant sequence in >50% of single-molecule reads from nanopore sequencing. This means that one of every two molecules in each MEGAA product had the perfect sequence. Furthermore, for 14 variants with residue substitutions, 89.3 ± 8.6% of the nanopore reads were fully mutated. Of the 17 deletion-containing variants, 65.5 ± 16.6% had fully mutated nanopore reads. Notably, variant ID31, which contains a 21 bp deletion along with six separate residue substitutions, shows complete variant generation in 70% of the reads, thus demonstrating the versatility of our method to produce different mutation types (Fig. 3a). Beyond defined variants, we also explored the generation of complex variant populations by MEGAA using oligonucleotides with degenerate bases to target multiple sites. Using a 6-pool or a 9-pool oligonucleotide set

https://doi.org/10.1038/s41592-023-01868-1

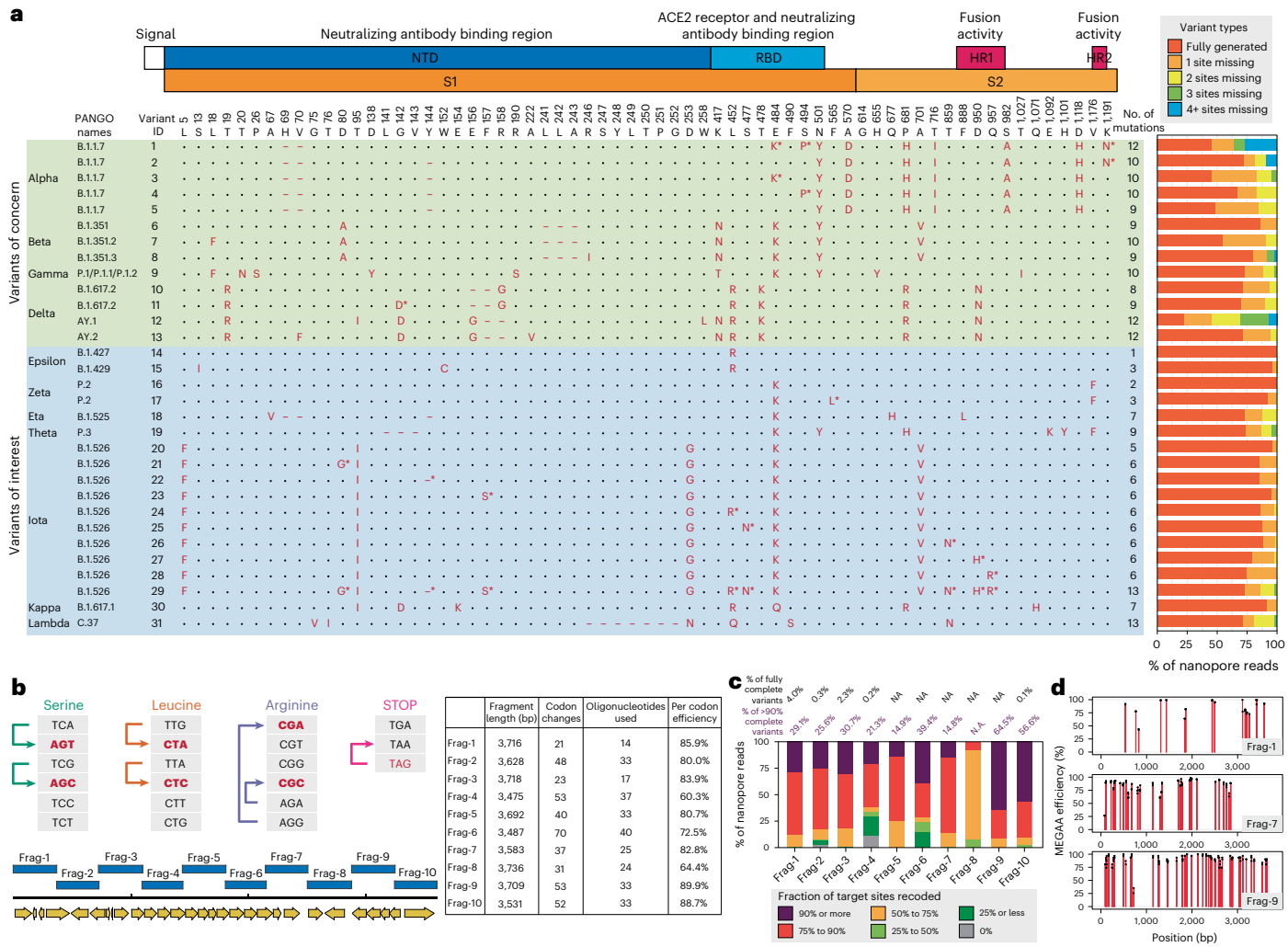


Fig. 3 | Generation of SARS-CoV2 spike gene variants and *E. coli* codon compressed recoded fragments using MEGAA. a, Thirty-one natural spike gene variant sequences are individually made with MEGAA. The MEGAA yield after two cycles measured by nanopore sequencing is shown. Asterisks indicate additional mutations in spike gene variants according to the World Health Organization. **b**, Generation of recoded genomes by systematically removing codons in the *E. coli* genome. **c**, MEGAA reaction results for 10 fragments showing the fraction of recoded target sites in each fragment. The MEGAA yield after 1 cycle measured by nanopore sequencing is shown. NA, not applicable (the value is less than 0.1%). **d**, Recoding efficiency across all target sites in three representative fragments. ACE2, angiotensin-converting enzyme 2; HR1, heptad repeat 1; HR2, heptad repeat 2; NTD, N-terminal domain; RBD, receptor-binding domain.

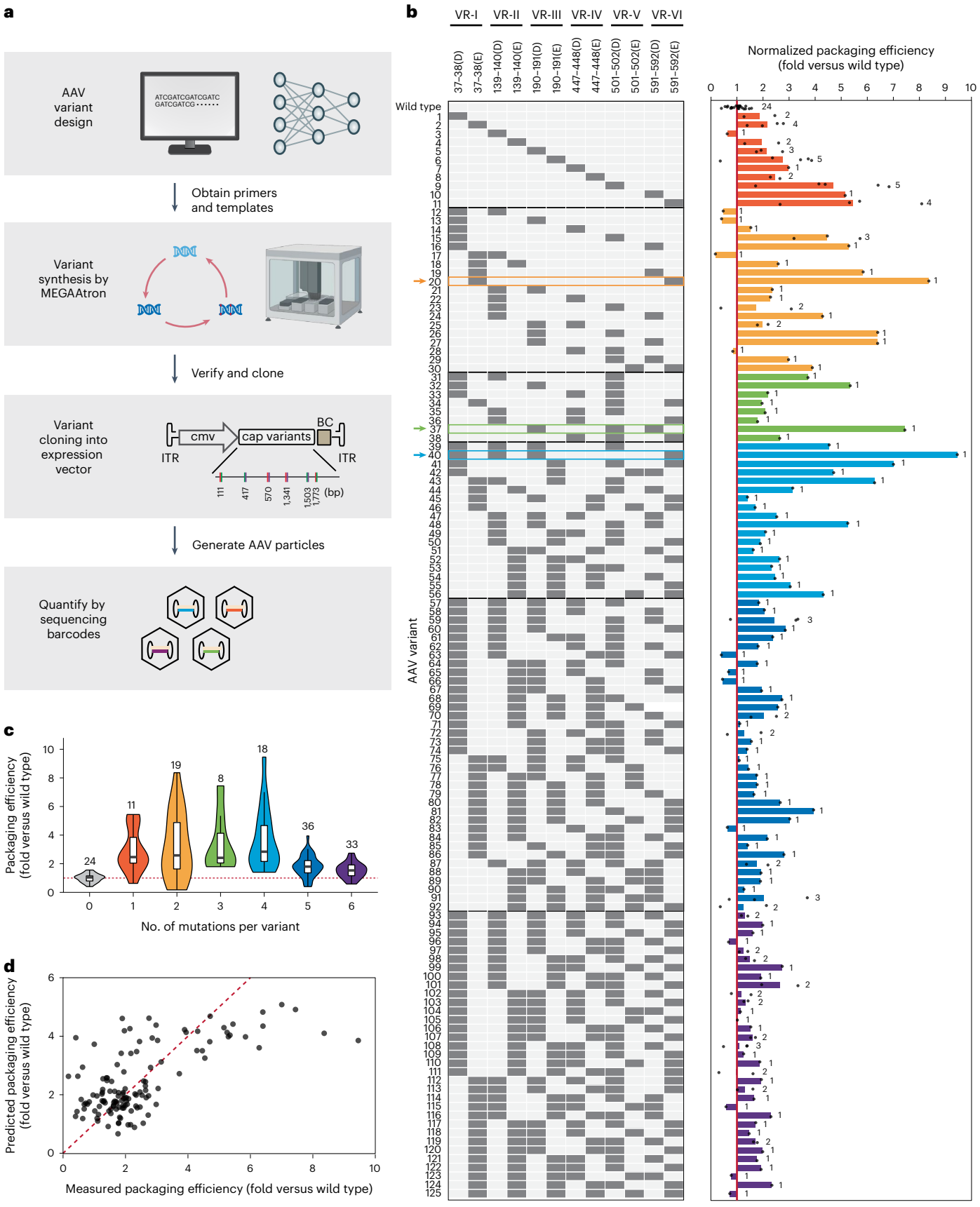
with NNS base degeneracies on the B.1.617.2 (Delta) *S* gene template, we produced complex yet even variant pools estimated to contain >1×10^9 and >3.5×10^{13} unique sequences for the 6-pool and 9-pool MEGAA reactions, respectively (Supplementary Fig. 13).

Next, we used MEGAAtron in a synthetic biology application involving genome-scale codon replacement (Fig. 3b). Several recent studies have explored the generation of synthetic genomes with recoded and reduced codon assignments that could provide biocontainment for viral infections and expansion of the genetic code with non-natural amino acids[9,28,29]. Thus far, these efforts required either multiplex oligonucleotide recombineering[21] or de novo synthesis of kilobase-sized fragments and subsequent hierarchical assemblies into full genomes[6], which are highly resource-intensive approaches. We sought to showcase MEGAA as a new 'templated synthetic genome synthesis' framework that is simpler, less expensive and more scalable. A codon replacement scheme (TTA→CTC, TTG→CTA, AGA→AGA, AGG→CGA, TCG→AGC, TCA→AGT) was adopted for the *E. coli* K-12 genome based on prior recoding strategies[5,6] (Fig. 3b). MEGAAtron was used in a proof-of-concept study to generate 10 recoded fragments each of ~3.6 kb in length using *E. coli* K-12 genomic DNA as the seed sequence (Supplementary Table 5). A total of 428 codon changes were made across this 36 kb genomic region. The resulting MEGAA products were pooled and analyzed using nanopore sequencing. Impressively, many fragments with >50 codon changes (for example, Frag-9 and Frag-10) had >70% of products with >75% of targets recoded from a single cycle of MEGAA (Fig. 3c). In general, most sites were efficiently targeted (78.8% efficiency) although some outliers were observed (Fig. 3d and Supplementary Fig. 14), which may require further oligonucleotide design optimizations. Importantly, these fragments were generated in less than 3 days at a 20-fold lower cost than commercial de novo gene synthesis. Once generated, these 3.6 kb fragments could then be combined into larger blocks by established genome assembly methods[6]. We anticipate that this approach will be useful for recoding bacterial and eukaryotic genomes[11].

Generating gene therapy carrier variants using MEGAA

Adeno-associated viruses (AAVs) have emerged as a safe and promising viral vector for DNA-based gene therapy, with more than 149 past or ongoing clinical trials[30]. The AAV capsid consists of 60 molecules of viral proteins encoded by the *cap* gene in the 4.8 kb single-stranded

https://doi.org/10.1038/s41592-023-01868-1



DNA genome of AAV. Mutations in the *cap* gene can lead to a variety of altered viral properties including changes in tissue tropism, packaging efficiency, thermal stability and neutralization escape by canonical antibodies[31]. An elegant recent study[32] generated a comprehensive single-residue saturation mutagenesis library of the *cap* gene in AAV2 and found many variable regions of the *cap* gene that individually modified AAV properties. However, the combinatorial effects of multiple distant mutations were not explored.

Fig. 4 | AAV2 *cap* gene engineering using MEGAAtron. a, Overview of AAV2 variant generation workflow. **b**, Packaging efficiency of 125 AAV2 variants normalized to wild-type levels. Six mutation sites VR-I to VR-VI are noted on the left. Dots in the plot represent individual barcoded replicates, and numbers next to the bars represent numbers of barcoded replicates for AAV variants used in this study. **c**, Violin plot showing packaging efficiency of variants based on number of mutations per variant. Numbers above the violin plot represent the number of barcoded replicates (for wild type) or the number of variants (for AAV2 variants) shown in the plot. In the box plot the center line represents the median; the box limits represent the upper and lower 25th quartiles; and the whiskers represent 1.5-fold the interquartile range. **d**, Plot showing measured versus predicted packaging efficiency compared with a linear regression model (the red dashed line). BC, barcode; CMV, cytomegalovirus promoter; ITR, inverted terminal repeat.

We used MEGAAtron to build AAV variants each containing up to six insertions at selected sites along the capsid protein that individually showed enhanced packaging efficiency based on saturation insertion data[32] (Fig. 4a and Supplementary Fig. 15). In particular, we chose negatively charged residues of aspartic acid (D) or glutamic acid (E) to insert into the capsid protein at residue positions 37–38, 139–140, 190–191, 447–448, 501–502 or 591–592, which in general are surface facing on the capsid. Altering the AAV surface charge can impact various viral particle properties and enhance purification by ion exchange during manufacturing[33,34]. Each variant also contained a unique 24 bp barcode that enables rapid identification and quantification by short-read Illumina sequencing. Variants were produced using MEGAAtron in an arrayed format with 1–12 defined mutant oligonucleotide combinations and cloned into a pAAV-CMV plasmid. Isolates were verified using nanopore sequencing (Supplementary Table 6 and Supplementary Fig. 16). In total, 192 barcoded clones were verified, corresponding to 125 unique variants, including 24 wild-type barcoded variants (Supplementary Table 6). Plasmids carrying each variant were equally pooled and transfected together into HEK293T cells (a human embryonic kidney cell line) to assess viral packaging efficiency by Illumina barcode sequencing (Methods). Packaging efficiency was quantified as the abundance of variants in the virus pool relative to the plasmid pool.

We first confirmed that in general, single-residue D or E insertions at each of the six chosen sites showed improvements in AAV2 packaging efficiency compared with the wild type, which correlated well with previous data[32] and thus verified our quantitative assay (Supplementary Fig. 17). Next, we explored the combinatorial variants in the library and identified several that had substantially improved packaging efficiency (Fig. 4b). Notably, Var20 (37–38E, 591–592E), Var37 (190–191D, 501–502D, 591–592D) and Var40 (37–38D, 139–140D, 190–191D, 591–592E) had an 8.4-fold, 7.4-fold and 9.5-fold improvement over wild type, respectively. Interestingly, we observed that variants containing five or six insertions had a much poorer packaging efficiency overall (mean of 1.9-fold and 1.5-fold, respectively) than variants with 1–4 mutations (mean of 2.9–3.7-fold, respectively) (Fig. 4c). This suggests that an excess of negative surface charge residues substantially reduced improvements in AAV packaging and indicates an upper limit to guiding combinatorial optimizations of this set of variant designs. Nevertheless, even with a limited combinatorial survey of 4-site or fewer variants, high-performing mutants were identified in our AAV2 library.

Finally, we applied our library data to a linear regression model to investigate mutation determinants of AAV2 packaging (Methods). In general, the linear model was able to predict improved packaging efficiency to a reasonable level (adjusted R^2 of 0.383, $P = 5.7 \times 10^{-10}$) (Fig. 4d and Supplementary Fig. 18). Interestingly, we found that in general 591–592(E) and 190–191(E) mutations had statistically significant positive coefficients in the linear model ($P < 0.05$). By contrast, 447–448(D/E) mutations had large negative coefficients and 139–140(D/E) mutations had small negative coefficients and these were both statistically significant ($P < 0.05$). These results show that a linear model can capture meaningful information of multi-site D/E mutational effects on AAV2 packaging efficiency, while also suggesting non-linear combinatorial effects that will warrant more sophisticated models[35,36].

Discussion

Genetic variants are crucial for understanding biological function and evolution. The capacity to build variants quickly and cheaply from an existing template can accelerate new biological discoveries and biotechnology. MEGAA offers an unprecedented capability to generate tens to hundreds of multi-site variations across kilobases of DNA at high efficiency in a matter of hours with automation. We have shown that MEGAA can be cycled to drive the mutagenesis reaction to near completion or be used to generate degenerate variations in DNA of up to 10 kb in length. Moreover, the distribution of variants can be reliably modeled to offer increased control of the in vitro mutagenesis reaction process. This variant synthesis platform can be more economical than de novo gene synthesis for long sequences. In this study, our 125-member AAV2 variant library costs ~US$33,000 to build through a commercial de novo gene synthesis vendor (2.2 kb at US$0.12 per bp) and can take up to 3 weeks to obtain, compared with ~US$2,900 (at US$0.01 per bp, reagents cost) using the MEGAAtron platform in a few days by a single person.

MEGAA has a few current limitations that could be addressed in the future. Given that the method is based on template-guided synthesis, MEGAA requires an available pre-existing template DNA. MEGAA is also sensitive to the fidelity of the oligonucleotides, and any unintended mutations in the oligonucleotides from DNA synthesis can become incorporated into MEGAA products. Furthermore, clonal generation of a perfect sequence is a costly and time-consuming step in gene synthesis that also applies to MEGAA, but it could be side-stepped if MEGAA efficiency is sufficiently high. The platform could be further improved by exploring other DNA polymerases to increase amplicon length and combined with pathway-scale DNA assembly methods to yield 100-kb-sized fragments[37,38]. Use of more sophisticated DNA folding and annealing models[39] could facilitate sequence-based MEGAA efficiency predictions and oligonucleotide design. Desktop DNA synthesizers can further improve MEGAA turnaround time, while better long-read sequencing technologies will further enhance accurate analysis and validation of MEGAA products. Furthermore, implementation of droplet-based DNA synthesis strategies[40] could increase the throughput of MEGAA. We envision that template-mediated synthesis, which can rapidly iterate many genetic designs, will become a crucial part of the synthetic biology arsenal to tackle pressing challenges facing the world.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-023-01868-1.

References

1. Bartley, B. A., Beal, J., Karr, J. R. & Strychalski, E. A. Organizing genome engineering for the gigabase scale. *Nat. Commun.* **11**, 689 (2020).
2. Esvelt, K. M. & Wang, H. H. Genome-scale engineering for systems and synthetic biology. *Mol. Syst. Biol.* **9**, 641 (2013).
3. Brophy, J. A. & Voigt, C. A. Principles of genetic circuit design. *Nat. Methods* **11**, 508–520 (2014).

4. Di Blasi, R., Zouein, A., Ellis, T. & Ceroni, F. Genetic toolkits to design and build mammalian synthetic systems. *Trends Biotechnol.* **39**, 1004–1018 (2021).
5. Ostrov, N. et al. Design, synthesis, and testing toward a 57-codon genome. *Science* **353**, 819–822 (2016).
6. Fredens, J. et al. Total synthesis of *Escherichia coli* with a recoded genome. *Nature* **569**, 514–518 (2019).
7. Mitchell, L. A. et al. Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond. *Science* **355**, eaaf4831 (2017).
8. Hutchison, C. A. 3rd et al. Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
9. Lajoie, M. J. et al. Genomically recoded organisms expand biological functions. *Science* **342**, 357–360 (2013).
10. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S. & Ryadnov, M. G. DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.* **7**, 144–161 (2023).
11. Boeke, J. D. et al. The genome project – write. *Science* **353**, 126–127 (2016).
12. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
13. Blazejewski, T., Ho, H. I. & Wang, H. H. Synthetic sequence entanglement augments stability and containment of genetic information in cells. *Science* **365**, 595–598 (2019).
14. Sharan, S. K., Thomason, L. C., Kuznetsov, S. G. & Court, D. L. Recombineering: a homologous recombination-based method of genetic engineering. *Nat. Protoc.* **4**, 206–223 (2009).
15. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
16. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
17. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **8**, 91 (2008).
18. Tseng, W. C., Lin, J. W., Hung, X. G. & Fang, T. Y. Simultaneous mutations up to six distal sites using a phosphorylation-free and ligase-free polymerase chain reaction-based mutagenesis. *Anal. Biochem.* **401**, 315–317 (2010).
19. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–206 (2015).
20. Cozens, C. & Pinheiro, V. B. Darwin Assembly: fast, efficient, multi-site bespoke mutagenesis. *Nucleic Acids Res.* **46**, e51 (2018).
21. Wang, H. H. et al. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).
22. DiCarlo, J. E. et al. Yeast oligo-mediated genome engineering (YOGE). *ACS Synth. Biol.* **2**, 741–749 (2013).
23. Lasken, R. S., Schuster, D. M. & Rashtchian, A. Archaebacterial DNA polymerases tightly bind uracil-containing DNA. *J. Biol. Chem.* **271**, 17692–17696 (1996).
24. Abellan-Schneyder, I., Schusser, A. J. & Neuhaus, K. ddPCR allows 16S rRNA gene amplicon sequencing of very small DNA amounts from low-biomass samples. *BMC Microbiol.* **21**, 349 (2021).
25. Liu, L. et al. Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2022).
26. Iketani, S. et al. Antibody evasion properties of SARS-CoV-2 Omicron sublineages. *Nature* **604**, 553–556 (2022).
27. Harvey, W. T. et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).
28. Robertson, W. E. et al. Sense codon reassignment enables viral resistance and encoded polymer synthesis. *Science* **372**, 1057–1062 (2021).
29. Rovner, A. J. et al. Recoded organisms engineered to depend on synthetic amino acids. *Nature* **518**, 89–93 (2015).
30. Kuzmin, D. A. et al. The clinical landscape for AAV gene therapies. *Nat. Rev. Drug Discov.* **20**, 173–174 (2021).
31. Bartel, M. A., Weinstein, J. R. & Schaffer, D. V. Directed evolution of novel adeno-associated viruses for therapeutic gene delivery. *Gene Ther.* **19**, 694–700 (2012).
32. Ogden, P. J., Kelsic, E. D., Sinai, S. & Church, G. M. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science* **366**, 1139–1143 (2019).
33. Qu, G. et al. Separation of adeno-associated virus type 2 empty particles from genome containing vectors by anion-exchange column chromatography. *J. Virol. Methods* **140**, 183–192 (2007).
34. Hsu, H. L. et al. Structural characterization of a novel human adeno-associated virus capsid with neurotropic properties. *Nat. Commun.* **11**, 3279 (2020).
35. Bryant, D. H. et al. Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
36. Zhu, D. et al. Machine learning-based library design improves packaging and diversity of adeno-associated virus (AAV) libraries. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.02.467003 (2021).
37. Jia, H., Guo, Y., Zhao, W. & Wang, K. Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. *Sci. Rep.* **4**, 5737 (2014).
38. Ellis, T., Adie, T. & Baldwin, G. S. DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol. (Camb.)* **3**, 109–118 (2011).
39. McDevitt, S., Rusanov, T., Kent, T., Chandramouly, G. & Pomerantz, R. T. How RNA transcripts coordinate DNA recombination and repair. *Nat. Commun.* **9**, 1091 (2018).
40. Plesa, C., Sidore, A. M., Lubock, N. B., Zhang, D. & Kosuri, S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* **359**, 343–347 (2018).

Methods

Chemical, oligonucleotide and enzyme reagents

All chemicals were purchased from Sigma-Aldrich unless otherwise noted. The CloneJET PCR cloning system was purchased from Thermo Fisher Scientific. Mutagenic oligonucleotides and sequencing primers were purchased from Integrated DNA Technologies. All enzymes were purchased from New England BioLabs.

Strains, viruses and culture conditions

Genomic DNA from *Escherichia coli* strain K-12 MG1655 was used as the uracil (U)-containing DNA template for MEGAA studies and codon replacement experiments. Plasmid pcDNA3.1 SARS-CoV-2 S D614G was obtained from Addgene to produce the SARS-CoV2 *S* gene variants. Plasmids pAAV-CMV vector, pRC2-mi342 vector and pHelper vector were purchased from Takara Bio Inc. (cat. no. 6230) to produce the AAV2 capsid variants and for AAV packaging. NEB Turbo Competent *E. coli* was used for cloning reactions using standard protocols.

In silico design of MEGAA oligonucleotides

MEGAA oligonucleotides were designed to target a template sequence using a custom python script (MEGAA-dt) available at https://github.com/wanglabcumc/MEGAAdt. The MEGAA-dt script takes reference template sequences and information on the desired mutations as input and generates designs of mutagenic oligonucleotides and sequences of final variants. In brief, desired mutations of each variant are evaluated based on their proximity, given that mutations that are too close to each other will be covered by the same oligonucleotide. Next, the number of perfectly matched bases at the 5′ end and 3′ end of the oligonucleotide is determined sequentially based on melting temperature to ensure that oligonucleotides are assembled in order (lower melting temperatures for upstream oligonucleotides). Mutagenic oligonucleotides are then evaluated for their length and distance to adjacent oligonucleotides, and the sequences of final oligonucleotide designs and variants are generated.

MEGAA protocol and MEGAAtron automation system

In the first step, a MEGAA template is generated using an input DNA source (for example, wild-type genomic DNA) by PCR amplification with a Q5U hot-start high-fidelity DNA polymerase (New England Biolabs, cat. no. M0515L) with a buffer mix in which dTTPs are replaced with dUTPs. Q5U hot-start high-fidelity DNA polymerase is able to use dUTPs at the same fidelity as dTTPs, and as a result the MEGAA template contains uracil (U) bases instead of thymine (T) bases. In the second step, a mix of the MEGAA template, Q5U hot-start high-fidelity DNA polymerase, Taq DNA ligase (New England Biolabs, cat. no. M0208L) and dNTP is made. Phosphorylated mutagenic oligonucleotides (~30–40 nucleotides) containing the desired mutations (that is, substitutions, insertion and deletions) and a forward extension primer are also added to the mix at 500–1,000-fold excess of the template. Oligonucleotide annealing, extension and ligation reactions then proceed in the single-pot reaction. Rapid oligonucleotide annealing (95 °C→4 °C at a rate of 3 °C s^{-1}) is performed on a standard thermal cycler. To increase MEGAA reaction throughput, a liquid handling robot (OT-2; Opentrons) equipped with magnetic, temperature and PCR modules is used to automate the MEGAA reaction. A detailed step-by-step protocol for the method is given in the Supplementary Methods and a detailed description and set-up breakdown of the MEGAAtron system can be found in Supplementary Table 2.

MEGAA characterization experiments

A uracil-containing DNA template (*rsgA*6, 1,192 bp) was generated by PCR using a Q5U hot-start high-fidelity DNA polymerase with the primers *rsgA*-F0 and *rsgA*-R0. In the meantime, 15 uracil-templates of different sizes (1.2–13 kb) were constructed by amplifying the *rsgA* gene region of the *E. coli* genome with primers from *rsgA*-F1 to *rsgA*-F5 with *rsgA*-R0, from *rsgA*-R1 to *rsgA*-R5 with *rsgA*-F0, and from *rsgA*-F1 to *rsgA*-F5 with *rsgA*-R1 to *rsgA*-R5, respectively. Another set of 16 uracil-templates (1.8–12 kb) was constructed by amplifying the *pheS* gene region using a similar approach (Supplementary Table 1). For the *rsgA*6 fragment, four individual MEGAA reactions are performed with 1, 3, 6 and 9 phosphorylated mutagenic oligonucleotides. For each *rsgA/pheS* gene region mutagenesis, a 9-target and a 12-target phosphorylated mutagenic oligonucleotide pool were added to the MEGAA reactions. Three *rsgA*6 variants (1-, 3- and 6-target), *rsgA1*–*rsgA15* and *pheS1*–*pheS15* amplicons were prepared for nanopore sequencing.

MEGAA optimization experiments

The oligonucleotide pools OP1 (OP1.1–OP1.9) and OP2 (OP2.1–OP2.9) were designed to target nine sites in the *rsgA* gene. Oligonucleotides in OP1 were designed with a similar melting temperature. However, oligonucleotides in OP2 were designed to have a gradation of melting temperature from 47 °C to 64 °C. The 1,192 bp DNA uracil-containing template was constructed by amplifying the *rsgA* gene region of the *E. coli* genome with primers *rsgA*-F0 and *rsgA*-R0. After the separate MEGAA reactions were performed, PCR amplicons (*rsgA*6) were cleaned up and processed via iteratively cycled MEGAA. The 5- and 3-cycle MEGAA reactions were performed with OP1 and OP2, respectively. After each MEGAA reaction, the *rsgA*6r1–r5 and *rsgA*6r1–r3 PCR amplicons from MEGAA OP1 and OP2 were prepared for nanopore sequencing, respectively.

Comparison of the mutagenesis efficiency of MEGAA and commercial kit experiments

MEGAA does not require the template to first be cloned into a circular plasmid DNA. Given that circular plasmids are required for all of these other methods as input, we first generated target plasmids (pJET1.2-*rsgA*6) by cloning the linear DNA fragments (*rsgA*6 template in Fig. 1b) into pJET1.2/blunt (Thermo Scientific cat. no. K1231). Mutagenesis was then performed according to the manufacturer's instructions. Subsequently, the mutagenesis efficiency of different methods was assessed using nanopore sequencing. In brief, transformation was performed for mutated products and all colonies were scraped from plates and pooled together. On average, more than 500 colonies were obtained for commercial kits. Targeted 1.2 kb fragments were then amplified from the pooled colonies using uniquely barcoded primers targeting the plasmid backbone region to avoid amplification of the endogenous *rsgA* gene in the *E. coli* genome. Finally, the amplified products underwent gel examination and the desired bands were excised for nanopore sequencing.

SARS-CoV-2 S mutagenesis experiment

Uracil-containing *S* gene templates were PCR amplified using the pcDNA3.1 SARS-CoV-2 *S* D614G plasmid[41] as the DNA template with the primers SARS-CoV-2 *S*_tempF and SARS-CoV-2 *S*_tempR. Mutagenesis of the SARS-CoV-2 *S* was performed via a MEGAA reaction with the modification that primer lengths were adjusted to ensure ordered oligonucleotide annealing. We designed mutagenic oligonucleotides containing target codons to generate all representative variants from alpha to lambda variants. Meanwhile, oligonucleotides containing degenerate bases (NNS) were designed to generate all combinations based on B.1.617.2 and AY.2 variants. Finally, 33 MEGAA reactions were carried out with 64 defined oligonucleotides and 10 degenerated oligonucleotides (Supplementary Table 1). All variants were prepared for nanopore sequencing after a clean up with SPRI (solid-phase reversible immobilization) beads.

Genome recoding mutagenesis experiment

Approximately 36 kb DNA sections of the *E. coli* K-12 genome were randomly chosen to be recoded with synonymous mutation DNA by compressed redundant codons (TTA → CTC, TTG → CTA, AGA → AGA,

AGG → CGA, TCG → AGC, TCA → AGT). The DNA sections were split into 10 fragments with 17–54 bp overlaps. Ten paired primers, 36K-F1/R1 to 36K-F10/R10, were designed and applied to amplify 10 uracil-containing DNA templates, respectively. Meanwhile, 289 mutagenic oligonucleotides were designed with an ordered oligonucleotide annealing strategy to cover 1,015 mutated bases in the 36 kb DNA. Then 10 oligonucleotide pools, which contained 14–40 mutagenic oligonucleotides per pool rather than individual oligonucleotides, were synthesized for MEGAA reactions. Following the MEGAA reaction steps, nanopore sequencing was applied to verify the recoding products.

AAV cap mutagenesis experiment

Uracil-containing DNA of the wild-type barcoded AAV2 *cap* gene were generated by two rounds of PCR amplification of the pRC2-mi342 vector (Takara Bio Inc., cat. no. 6230) with primers AAV2-tempF/AAV2-tempR1 and AAV2-tempF/AAV2-tempR2 (Supplementary Fig. 16). AAV2-tempR1 and AAV2-tempR2 included 12 bp of random barcoding DNA. Twelve oligonucleotides were designed to carry D or E insertions in the six variable regions (VR), VR-I to VR-VI, the positions of which were 35–40, 132–152, 188–192, 445–460, 490–500 and 576–596 in capsid protein, respectively. After oligonucleotide phosphorylation, MEGAA reactions were carried out with the oligonucleotide pool covering six variable regions. All *cap* variants were assessed using nanopore sequencing. The pAAV-CMV vector was linearized using the *EcoRI* and *BamHI* restriction enzymes. Unique barcoded wild-type *cap* gene and MEGAA variants were digested using *EcoRI* and *BamHI*. The two products were purified using SPRI beads, ligated using T4 DNA ligase (New England Biolabs, cat. no. M0202M), and incubated at 16 °C overnight. A pooled plasmid was transformed into competent NEB Turbo cells and grown for 10 h at 37 °C. Individual colonies were selected for colony PCR with Oxford Nanopore sequencing barcoded primers. Indexed amplicons were pooled in an equimolar fashion and sequenced on the Oxford Nanopore platform to identify mutation sites as well as barcode sequences.

AAV virus was produced using AAVpro Helper Free System (Takara Bio Inc., cat. no. 6230), with minor adjustments. In brief, a 150 mm cell culture dish (Thermo Scientific 150468) was inoculated with 6.0×10^6 293T cells in DMEM culture medium supplemented with 1× GlutaMAX, 1× Pen–Strep antibiotic, and 5% FBS according to standard cell culture protocols. The 293T cells were split into ten 150 mm cell culture dishes for the experiment when cells were approximately 90% confluent. Two days after splitting the cells, polyethylenimine (PEI) transfection was performed with a PEI : DNA mass ratio of 3:1 with 36 µg pR2-mi342, 70 µg pHelper and 0.25 µg pool variants, which included 125 unique barcoded pAAV-CMV-aav2cap variants along with 24 barcoded *cap* wild-type plasmids. The culture medium was completely replaced with fresh DMEM containing 1× GlutaMAX, 1× Pen–Strep antibiotic, and 5% FBS at 12 h after transfection. Fifty percent media volume (200 ml) was added after 72 h. After 5 days, isolation of AAV2 particles from AAV-producing cells was performed according to the AAVpro Helper Free System instructions. Nuclease treatment was performed by adding a 1:100 volume of 1 M $MgCl_2$ solution to a mixture consisting of supernatant obtained from the AAVpro Helper Free System and TURBO DNase (Thermo Scientific, cat. no. AM1907), to a final concentration of 0.4 U μl^{-1}. After centrifugation at 5,000 ×*g* for 10 min at 4 °C, the supernatant was harvested and the AAV2 particles were purified and concentrated following the AAVpro Purification Kit protocol (Takara Bio Inc., cat. no. 6232).

To evaluate the packaging efficiency of the variants, barcode regions of variants for input plasmids and virus particles were quantitatively amplified and sequenced on the NextSeq platform. In brief, 1 µl purified AAV2 particles (1×10^8 GC μl^{-1}) and input plasmid pools underwent 12-cycle PCR amplification using AAV_bcRead primer pairs and cleaning up with SPRI beads to generate amplicons of barcode regions. Next, quantitative PCR was performed to add indexed Illumina TruSeq adapters to the amplicon and advanced to the final extension step during exponential amplification. The libraries that were produced were then purified by gel electrophoresis and sequenced on the Illumina NextSeq platform (2 × 75 paired-end mode; Control Software v4.0) with a 20% PhiX spike-in (Illumina FC-110-3001) according to the manufacturer's instructions. Sequences of primers used for library preparation for barcode sequencing are listed in Supplementary Table 7.

Raw sequencing reads of variant barcode amplicons were analyzed using an in-house script to calculate the packaging efficiency of the variants. In brief, barcode sequences of the variants were extracted from reads and matched to variant identity based on references from nanopore sequencing. Reads mapped to each variant were then counted and the relative abundance (RA) was calculated as: RA (variant-X) = reads count mapped to variant-X divided by the total mapped reads count. Next, the ratio of the relative abundance of the variant in the generated virus particles to that in the input plasmid was determined to quantify packaging efficiency: efficiency (variant-X) = RA (variant-X) in virus pool divided by RA (variant-X) in the plasmid pool. The efficiency was then normalized by wild-type variants to generate a final variant packaging efficiency used in downstream analysis: normalized efficiency(variant-X) = efficiency (variant-X) divided by the average of efficiency (wild-type variants). To explore the determinants of AAV2 packaging efficiency, a linear regression model was constructed in R v4.1.2 to predict packaging efficiency based on the binarized mutation profile of all 125 variants, and the predicted efficiency as well as the coefficients of mutation sites were extracted from the linear model to evaluate the overall performance and combinatorial effect of each site.

Nanopore sequencing and data analysis

To determine the overall variant generation efficiency, we implemented a barcoded Oxford Nanopore strategy to sequence the full length of generated variants. In brief, unique dual 12-bp barcodes were added to both ends of the variants by PCR amplification, and the resulting barcoded variants were pooled together and purified using gel electrophoresis. After clean up, ~300 fmol pooled variants were processed using Oxford Nanopore library preparation and sequencing following the manufacturer's instructions. Variants underwent Nanopore sequencing using the protocol 'Amplicons by Ligation (SQK-LSK110)' (Oxford Nanopore Technologies). Both MinION Flow Cell R9.4.1(FLO-MIN106D) and R10.4 (FLO-MIN112) were used for sequencing on a MinION with the MinKNOW v21.11.8 (Oxford Nanopore Technologies). Base-calling was performed with Guppy v3.6.0 (Oxford Nanopore Technologies) in GPU mode. Full-length reads were first demultiplexed based on the barcodes of both ends using an in-house Python script and quality filtered to retain only high-quality reads (no more than 3 bp mismatches and 1 bp gap in a 20 bp region of both 5′ and 3′ ends). Demultiplexed reads were then aligned to the reference sequence by MUSCLE[42] v3.8.31 using default settings. Variant generation efficiency was then calculated based on reads alignment using an in-house Python script. In-house scripts used for Oxford Nanopore sequencing data analysis can be accessed at https://github.com/wanglabcumc/MEGAAdt.

Analytical model of MEGAA cycling process

Using a binomial distribution, we can predict the completeness of MEGAA reactions (C_N) at MEGAA cycle N with the average oligonucleotide incorporation efficiency per locus (μ) through the formula $C_N = 1 - (1 - \mu)^N$. The completeness C_N metric indicates the fraction (or % completeness) of all target sites mutated in the end-product mix, with 1.0 indicating that 100% of products have 100% complete mutations at all target sites. In this model, a C_N of 0.5 could mean that either 50% of products have all sites mutated or that 100% of products have half of their sites mutated. Mapping our experimental data to this simple model gives us an estimated average oligonucleotide incorporation efficiency μ of 0.8–0.9 for Design-2 oligonucleotides (that is, 80–90% mutagenesis efficiency per site per MEGAA round), compared with 0.5–0.7 for Design-1 oligonucleotides (50–70% mutagenesis efficiency).

Statistics and reproducibility

The gel electrophoresis in Fig. 1b and Supplementary Fig. 3b were repeated more than three times independently and similar results were obtained. No statistical method was used to predetermine sample size because the number analyzed depends on the yield from experiments. Data exclusion was based on sequencing coverage or amplicon quality to remove technical artifacts. The experiments were not randomized, and the researchers were blinded to samples given that different designs were processed together in the experiment. Blinding in analysis was not possible and all analyses were performed with the same parameters.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Processed packaging efficiency data in a previous study were obtained from GitHub (https://github.com/churchlab/AAV_fitness_landscape) to identify insertion sites with potential enhanced packaging efficiency for AAV variants design and correlate with packaging efficiency obtained in this study. The sequencing data generated in this study have been submitted to the NCBI BioProject database under accession number PRJNA834093. Source data are provided with this paper.

Code availability

Scripts used for Oxford Nanopore sequencing data analysis can be accessed at https://github.com/wanglabcumc/MEGAAdt.

References

41. Yurkovetskiy, L. et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **183**, 739–751 (2020).
42. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

Acknowledgements

The authors thank G. Urtecho, J. Qian and L. Huang for technical support, and K. Beiswenger and other members of the Wang laboratory for advice and comments on the manuscript. H.H.W. acknowledges funding support from the NSF (MCB-2032259), DOE (47879/SCW1710), NIH (1R01DK118044, 1R01EB031935, 2R01AI132403, 75N93021C00014), ONR (N00014-17-1-2353), Burroughs Wellcome Fund (1016691), Irma T. Hirschl Trust and Schaefer Research Award.

Author contributions

L.L. and H.H.W. developed the initial concept; L.L. and Y.H. performed the experiments and analyzed the data with input from H.H.W.; Y.H. developed the software and genomic data analysis pipeline. L.L. and H.H.W. wrote the manuscript. All other authors discussed the results and approved the manuscript.

Competing interests

H.H.W. is a scientific advisor of SNIPR Biome, Kingdom Supercultures, Fitbiomics, Arranta Bio, VecX Biomedicines, Genus PLC, and a scientific co-founder of Aclid, all of which are not involved in the study. A patent application on methods described in this paper has been filed by Columbia University. All other authors have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-01868-1.

Correspondence and requests for materials should be addressed to Harris H. Wang.

Peer review information *Nature Methods* thanks Kaihang Wang, Hongzhou Gu, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Lei Tang and Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team.

nature portfolio

Corresponding author(s): Harris Wang

Last updated by author(s): Mar 25, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Illumina sequencing data was collected with NextSeq Control Software v4.0. Nanopore sequencing data was collected with Control Software MinKNOW v21.11.8 provided by Oxford Nanopore.
Data analysis	Both MinION Flow Cell R9.4.1(FLO-MIN106D) and R10.4 (FLO-MIN112) were used for sequencing on a MinION flow cell. Base-calling was performed with Guppy v3.6.0 (ONT) in GPU mode. Full-length reads were firstly demultiplexed based on barcodes of both ends using in-house Python script and subjected to quality filtering to only keep high-quality reads (no more than 3-bp mismatches and 1-bp gap in 20-bp region of both 5′ and 3′ ends). Demultiplexed reads were then aligned to reference sequence by MUSCLE v3.8.31 using default setting. Variant generation efficiency was then calculated based on reads alignment using in-house Python script. In-house scripts used for Oxford Nanopore sequencing data analysis can be accessed at https://github.com/wanglabcumc/MEGAAdt. Linear regression model of binarized mutation profile and packaging efficiency was constructed in R v4.1.2 with default setting.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Processed packaging efficiency data in previous study was obtained from GitHub (https://github.com/churchlab/AAV_fitness_landscape) to identify insertion sites with potential enhanced packaging efficiency for AAV variants design and correlate with packaging efficiency obtained in this study. The sequencing data generated in this study have been submitted to the NCBI BioProject database under accession number PRJNA834093.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed as the number analyzed depends on the yield from the experiment; sample sizes are listed in Methods section where applicable. The # of variants generated and sequenced in this study are based on experimental design and the specific # of variants was designed in this study to cover (1) difference length/characteristics of variants to evaluate the performance of MEGAA in the range of most common gene length/characteristics or (2) difference distributions of mutation sites shown in most cases of variants generation.
Data exclusions	Data exclusion was based on sequencing coverage or amplicons quality to remove technical artifacts as described in the Methods section.
Replication	In proof of concept, MEGAA was performed more that 100 biological replicates. In MEGAA characteristics parts: 54 biological replicates (Supplementary Fig. 2, Fig. 1, Supplementary Fig. 3, Fig. 2, Supplementary Fig. 4, Supplementary Fig. 5). In the applications part: 56 biological replicates (Fig. 3 and Fig. 4). The reactions of these replicates details were as described in the Methods section.
Randomization	The experiments were not randomized.
Blinding	The researchers were blinded to different types of samples as different designs are mixed in the same round of experiments. Blinding in analysis was not possible during experiments. All analyses of associated data were performed with the same parameters and criteria described in Methods section.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	293T cells (CRL-3216) were obtained from ATCC
Authentication	Cell lines used were not authenticated
Mycoplasma contamination	Cell lines tested negative for mycoplasma
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used