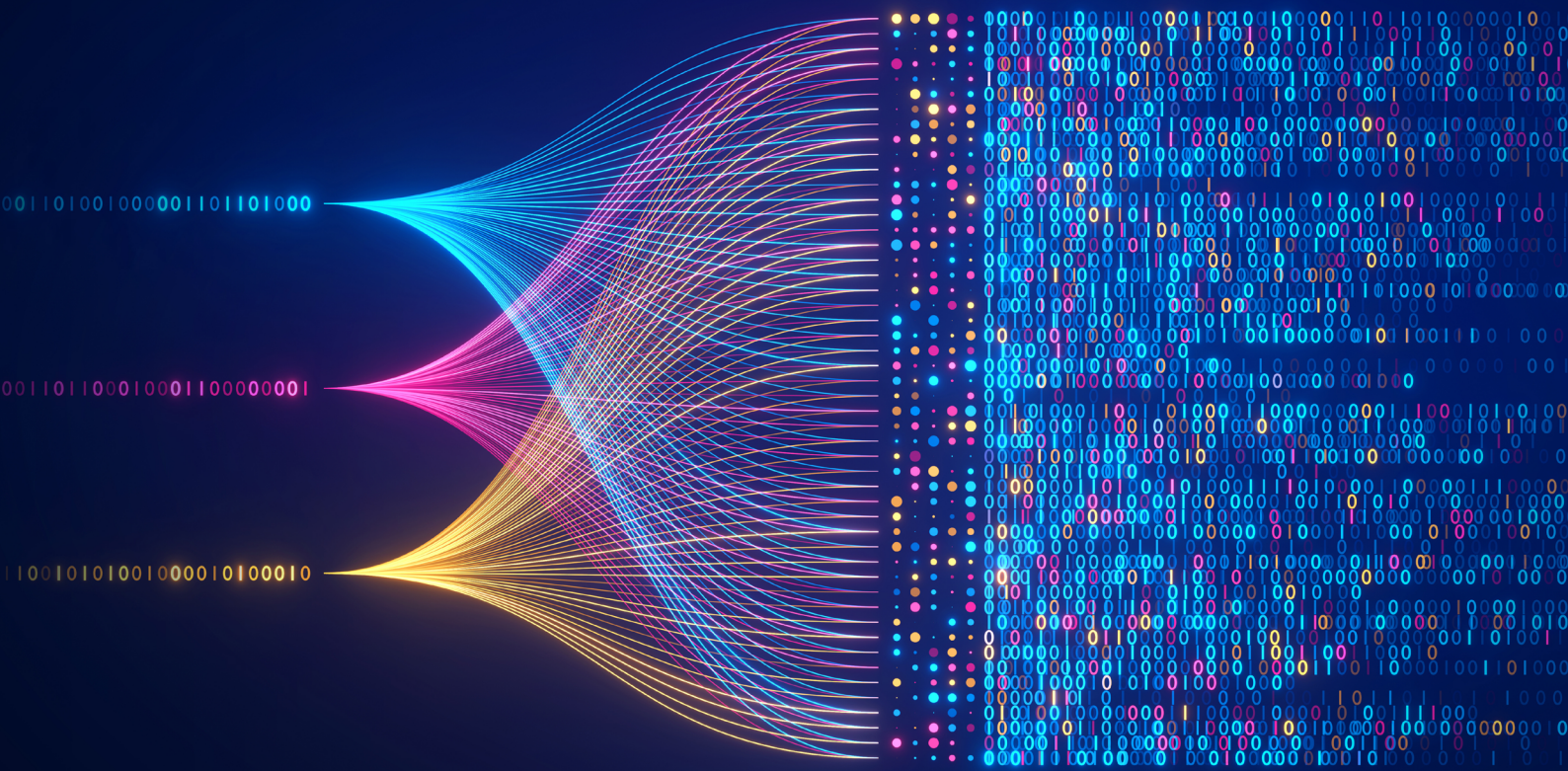


How Pooled Oligo Synthesis is Fostering Innovation in Drug Target Discovery, Protein Engineering, and Synthetic Biology

Expert Insights

Sponsored by

Agilent

WILEY

Contents

Introduction: How Pooled Oligo Synthesis is Fostering Innovation in Drug Target Discovery, Protein Engineering, and Synthetic Biology

Oligos, short for oligonucleotides, are fundamental tools in molecular biology, genetics, and biotechnology that play a key role in research and development. They are generally chemically synthesized, short nucleic acid sequences made up of a variety of nucleotides and can be either DNA or RNA-based. By nature, they are highly specific and have unique compositions suited for precise molecular and genetic applications. Today, they are essential components in various workflows, such as PCR, DNA sequencing, gene expression analysis, and gene synthesis.

Pooled oligo synthesis refers to the process of chemically synthesizing thousands of different oligo sequences in massively parallel fashion on a silicon-chip based substrate using high resolution DNA writing instrumentation. After synthesis is complete, the single stranded oligos are cleaved from the substrate and consolidated as a single library or pool of thousands of custom oligo sequences. As a research tool, oligo pools have emerged as a low-cost source of synthetic DNA for the creation of the enormous sequence diversity required for use in high-throughput genomics screening and next-generation protein and pathway engineering strategies. The significant cost savings relative to other commercial sources of custom sequences also enables lower cost gene assembly and more cost-effective screening when performing massively parallel reporter assays (MPRA) or even loss of function or gain of function research in the form of pooled CRISPR screening. In this context, the oligo pools produced on high resolution DNA writing platforms present researchers with a powerful technology built on high-throughput, parallel synthesis of large numbers of unique oligo sequences on a single chip for custom, high-precision synthetic DNA produced with speed and at scale.

This Expert Insights eBook begins with a study on splicing assays in precision medicine. The research study from Rhine *et al.* [1] represents significant advancements in high throughput splicing assays, specifically the Massively Parallel Splicing Assay (MaPSy), for comprehending and categorizing the effects of genetic variants in precision medicine. The advancement of these assays and predictive modeling of splicing (MMSplice) has the potential to improve the understanding of genetic mutations, especially those that impact pre-mRNA splicing, and their relevance in clinical genetics and disease management.

Our second paper by Kuiper *et al.* [2] is a research review that examines the utility of oligo pools in gene assembly and high-throughput oligonucleotide library design for protein and metabolic pathway engineering. Utilizing custom oligo pools is essential for protein engineering and provides a low-cost source of synthetic DNA, greatly reducing expenses compared to other methods. The authors point out that although oligo pools are an affordable alternative, there are challenges associated with their use in advanced engineering strategies, such as low concentrations and high error rates. Despite the challenges, the review highlights the importance of oligo pools in performing many applications and techniques that should be more accessible to the broader bioengineering community. The researchers summarized currently available methods that use affordable, array-synthesized oligo pools as a source of synthetic DNA for next-generation protein engineering libraries.

Our last research paper is a protocol published by Eroshenko *et al*. [3]. It presents an exciting technique for constructing lengthy double-stranded DNA structures using oligonucleotides that are synthesized on DNA chips with high density. This method overcomes the high costs and technical challenges of traditional synthesis. Long double-stranded DNA synthesis is critical in many biological and bioengineering applications. However, high cost and complexity have kept it from being widely used, especially in academic settings. The researchers in this study developed a protocol for assembling 500- to 800-base pair gene fragments from commercially available DNA chips.

Overall, oligo pools have become indispensable tools in applications as diverse as synthetic biology, pooled CRISPR screening, and high-throughput gene assembly. Furthermore, utilization has driven advances in data driven computational modeling and predictive algorithm development. These developments have had far-reaching consequences in accelerating the research supported by pooled oligo synthesis platforms. This accelerated understanding of functional genomics and protein engineering will undoubtedly improve research outcomes in precision medicine and variant discovery as well as in novel enzyme and therapeutic protein development.

Through this Expert Insights eBook, we hope to educate researchers on oligo pools and potential applications in functional genomics pooled screening and synthetic biology. For more information, we encourage you to visit Agilent's **Pooled Oligo Synthesis** page to gain a deeper understanding of available options for accelerating your research.

Julian Renpenning, Ph.D.
Editor at Wiley Analytical Science

References

[1] Rhine, C.L. *et al*. (2019). Future directions for high-throughput splicing assays in precision medicine. *Human Mutation*. DOI: 10.1002/HUMU.23866.

[2] Kuiper, B.P. *et al*. (2022). Oligo Pools as an Affordable Source of Synthetic DNA for Cost-Effective Library Construction in Protein- and Metabolic Pathway Engineering. *ChemBioChem*. DOI: 10.1002/CBIC.202100507.

[3] Eroshenko, N. *et al*. (2012). Gene Assembly from Chip-Synthesized Oligonucleotides. *Current Protocols in Chemical Biology*. DOI: 10.1002/9780470559277.CH110190.

Exceptional Uniformity and Fidelity

Powering your high-complexity DNA oligo libraries

Custom oligonucleotide libraries made for your unique research goals

Are you looking for reliable and fully customizable oligo libraries for your research? Whether you need short or long oligos, high or low complexity, Agilent has the solution for you. Synthesize your oligo pool on a platform that has been continuously improved and optimized for over 20 years. Our oligo libraries offer:

- Industry-leading fidelity with error rates below one in 2,400 nucleotides
- Exceptional uniformity for fewer false positives and false negatives following pooled screens
- Multifaceted quality control with parallel control synthesis used in every product run
- Continuous production running 24 hours a day and seven days a week for rapid turnaround times

Create your library from the ground up and leverage our pooled oligo synthesis platform to make the start of your experiments more accessible than ever.

Explore the Agilent oligonucleotide library manufacturing advantage today.

Gene Assembly from Chip-Synthesized Oligonucleotides

Nikolai Eroshenko,[1,5] Sriram Kosuri,[2,3,5] Adam H. Marblestone,[3,4] Nicholas Conway,[3] and George M. Church[2,3]

[1]Harvard School of Engineering and Applied Sciences, Cambridge, Massachusetts
[2]Department of Genetics, Harvard Medical School, Boston, Massachusetts
[3]Wyss Institute for Biologically Inspired Engineering, Boston, Massachusetts
[4]Harvard Biophysics Program, Cambridge, Massachusetts
[5]These authors contributed equally to this work

ABSTRACT

De novo synthesis of long double-stranded DNA constructs has a myriad of applications in biology and biological engineering. However, its widespread adoption has been hindered by high costs. Cost can be significantly reduced by using oligonucleotides synthesized on high-density DNA chips. However, most methods for using off-chip DNA for gene synthesis have failed to scale due to the high error rates, low yields, and high chemical complexity of the chip-synthesized oligonucleotides. We have recently demonstrated that some commercial DNA chip manufacturers have improved error rates, and that the issues of chemical complexity and low yields can be solved by using barcoded primers to accurately and efficiently amplify subpools of oligonucleotides. This unit includes protocols for computationally designing the DNA chip, amplifying the oligonucleotide subpools, and assembling 500- to 800-bp constructs. *Curr. Protoc. Chem. Biol.* 4:1-17 © 2012 by John Wiley & Sons, Inc.

Keywords: oligonucleotide • gene synthesis • nucleic acids • synthetic biology

INTRODUCTION

Gene synthesis has been used in applications as diverse as decoding the genetic code (Nirenberg and Matthaei, 1961), screening industrially useful enzymes discovered through environmental metagenomic sequencing (Bayer et al., 2009), and building custom pathways and genomes (Tian et al., 2004; Ro et al., 2006; Hanai et al., 2007; Gibson et al., 2008). Unfortunately, despite the rapid decrease in the cost of synthesizing short single-stranded oligonucleotides, synthesis of double-stranded gene-sized fragments remains too expensive for ubiquitous adoption by academic laboratories (Carr and Church, 2009). To make large-scale synthesis cheaper, we have recently developed a set of techniques, which allow assembly of DNA from commercially available high-density oligonucleotide chips into gene-sized 500- to 800-bp fragments (Kosuri et al., 2010). Specifically, we showed that PCR could be used to separate the DNA synthesized on a chip into subsets (or, as we will refer to them from here on, subpools) consisting of only the oligonucleotide species necessary to build one particular gene-sized fragment. These methods use inexpensive sources of DNA and require no specialized instrumentation or expertise that cannot be found in a typical molecular biology laboratory.

The major stages of our synthesis pipeline are computational design, chip synthesis, serial PCRs that isolate the oligonucleotides necessary to build each construct, and assembly of the constructs. The key principle is that well-designed primers can amplify a desired subset of oligonucleotides and, thereby, dilute the undesired DNA to the point where it does not interfere with the downstream gene assembly reaction. An off-chip pool (see

Table 1 Nomenclature Used to Describe the Synthesis Protocol

Term	Definition	Source
Off-chip pool	The pool of oligonucleotides cleaved from the DNA microchip	Synthesized on a DNA chip
Plate subpool	The subset of the off-chip pool necessary to build 96 assemblies	Amplified from off-chip pool
Plate subpool-specific primers	A pair of PCR primers that bind to sites shared by all members of a plate subpool	Traditional (not chip) synthesis
Assembly subpool	The subset of a plate subpool necessary to build 1 assembly	Amplified from plate subpool
Assembly subpool-specific primers	A pair of PCR primers that bind to sites shared by all members of an assembly subpool	Traditional (not chip) synthesis
Construction primers	A pair of PCR primers used to assemble one or more assembly subpools	Traditional (not chip) synthesis
Assembly	A 500-800 bp dsDNA construct built from an assembly subpool using a pair of construction primers	Built from assembly subpool

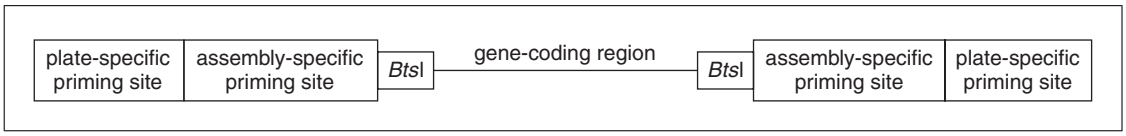


Figure 1 A schematic of the features present on each off-chip oligonucleotide. The gene-coding regions of the oligonucleotides within each assembly subpool partially overlap, allowing them to be assembled into the full-length construct using a high-fidelity polymerase. The gene-coding region is flanked by *Bts*I cut sites that permit enzymatic removal of the subpool-specific priming sites. The gene-coding region is also flanked by a pair of assembly-specific priming sites, which are shared by all the oligonucleotides within a particular assembly subpool. The assembly-specific priming sites are, in turn, flanked by a pair of plate-specific priming sites common to all the oligonucleotides within a particular plate-specific subpool.

Table 1 for nomenclature) consists of oligos with a gene-coding region flanked by nested subpool-specific sequences and a restriction enzyme recognition site for removing the priming sequences (Fig. 1). The subpool-specific sequences act as barcodes for selecting subsets of oligos. Each pair of assembly-specific priming sites is shared by all the oligonucleotides needed to build a particular construct; in turn, each pair of plate-specific subpool priming sites is shared by all the oligos necessary to build 96 constructs. Another pair of primers—the construction primers—are used to amplify the full-length construct.

We have developed a software tool called GASP (Gene Assembly by Subpool PCR) that deconstructs a given set of genes to generate the sequences of oligonucleotides for synthesis on a chip. Using GASP for chip design is described in Basic Protocol 1. Once the oligonucleotides have been designed, synthesized, and cleaved off the chip surface, the off-chip pool is divided into aliquots among plate subpool amplification reactions (Fig. 2). After the PCR, each plate subpool is divided into aliquots into a 96-well plate. Following the addition of assembly-specific primers, another amplification is performed, at which point a restriction enzyme is used to cleave off the priming sites. Basic Protocol 2 provides directions for amplifying the plate and assembly subpools, and Basic Protocol 3 describes the enzymatic removal of priming sites. A polymerase is used to assemble the oligos into the full-length construct, and the assemblies are amplified using the construction primers (which are not necessarily unique for the assembly subpool), as described in Basic Protocol 4.

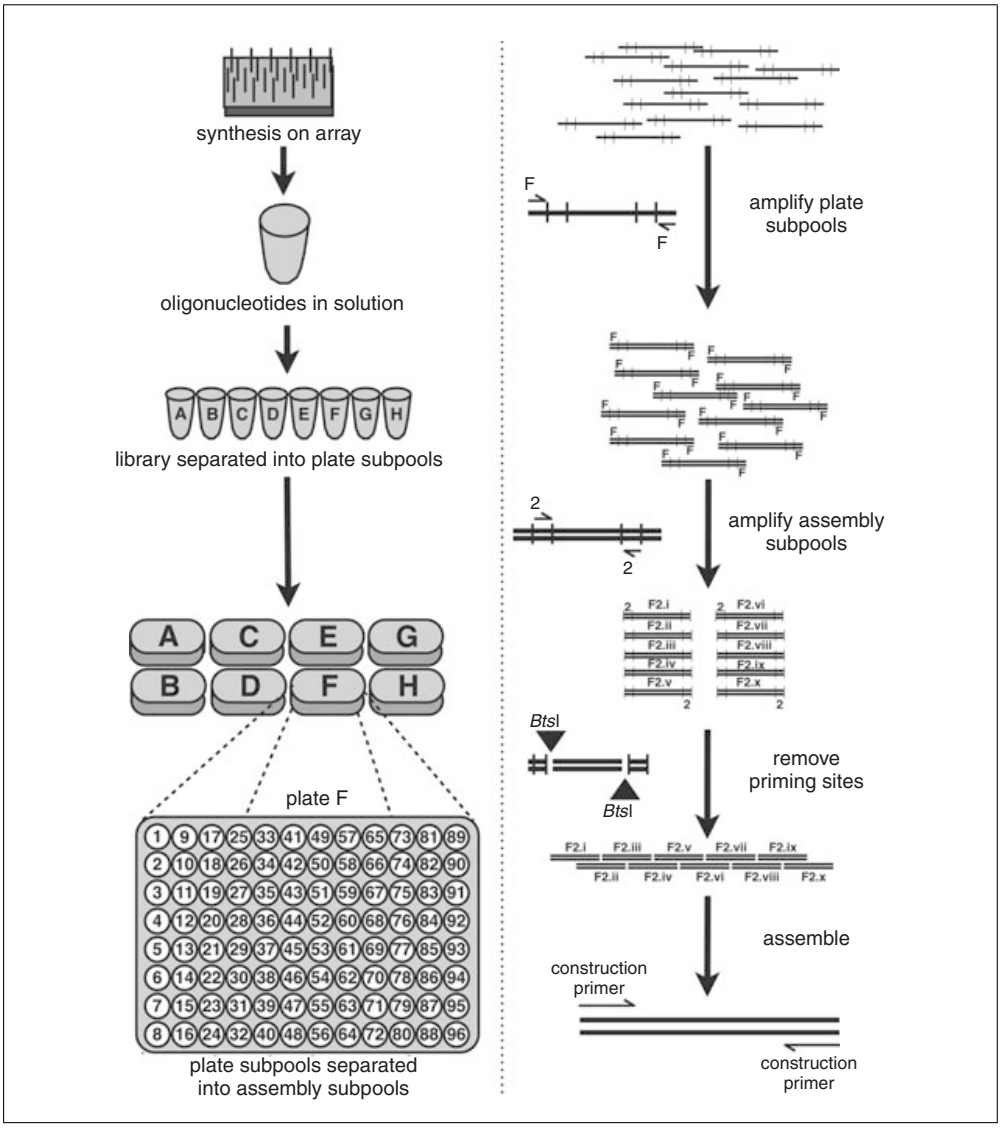


Figure 2 A schematic of the gene synthesis workflow's liquid handling steps (left) and the oligonucleotide processing steps (right). In the left half of the schematic, each off-chip oligonucleotide is drawn as a horizontal line with vertical lines used to indicate the plate and assembly subpool-specific priming sites. After the DNA chip has been synthesized and the oligonucleotides have been cleaved off the array surface, plate-subpool-specific primers are used to amplify only the oligonucleotides needed to build 96 assemblies. In the drawing, the primer pair "F" amplifies all the oligonucleotides that make up plate subpool "F." The plate subpool is then divided into aliquots among the wells of a 96-well plate. Assembly subpools are amplified with assembly-specific primers. In the figure, assembly-specific primers "2" amplify subpool F2, which consists of ten double-stranded fragments (F2.i-F2.x). The assembly subpool-specific priming sites are removed via a restriction digest, and a polymerase assembles the overlapping fragments into full-length constructs. Last, full-length constructs are amplified by a pair of construction primers.

OLIGONUCLEOTIDE DESIGN AND SYNTHESIS

BASIC PROTOCOL 1

Each construct to be built must be split up into short overlapping fragments. Each fragment must, in turn, be flanked by the assembly- and plate-specific subpool priming sequences, as well as restriction sites for removing the priming sequences (Fig. 1). We have automated these design steps with Biopython scripts (Cock et al., 2009). This software can be either run from a server (*http://synbiosis.med.harvard.edu:8080/gaspserver.rpy*) or as a script on any computer that has the Biopython package installed (current and all

future versions of GASP can be found at *https://bitbucket.org/skosuri/gasp/overview*). The protocol below describes using the server version.

Materials

Computer with a Web browser
List of sequences to be built (in FASTA format)

Setting up the parameters

1. Open http*://synbiosis.med.harvard.edu:8080/gaspserver.rpy* with a browser.
2. Enter your name, email address, and the location of your input file (the FASTA file with a list of genes you want built).
3. The parameter configuration text box will contain the following text:

```
"initialPlateNum": 2,
"RESpacing": [
2,
5,
4
],
"REVector": [
  "BtsI",
"BsmBI",
"BspQI"
],
"SearchForRE": "True",
"REToUse": "BtsI",
"avgoverlapsize": 20,
"deltaGThresholdForOverlaps": -3,
"selfDimersThreshold": 3,
"lengthleeway": 10,
"positionleeway": 10,
"oligoSizeMax": 200,
"seqsToAvoidInOverlapRegions":[],
"skip":[]
```

Change the SearchForRE parameter to False.

The parameters, which are described in detail below, may have to be further adjusted if the DNA will be processed using methods that deviate from the workflow described here.

InitialPlaneNum: 96-well plates of assemblies will be numbered sequentially initiating at this value. This should never be set to 1, as plate #1 is reserved for construction primers.

RESpacing: The distance between the end of the recognition site to the cut location for the enzymes listed in REVector setting. The values should be separated by a comma, and be in the same order as the enzymes in the REVector setting. For example, the RESpacing for BtsI should be set to 2 because the enzyme nicks 2 bp away from its binding site (such that the full recognition site is GCAGTGN_2). Similarly, the recognition site for BsmBI is CGTCTCN_5, so its RESpacing should be set to 5. Therefore, if "REVector": ["BtsI","BsmBI"], then "RESpacing": [2,5]. The site can be left blank if SearchForRE is False.

REVector: List of type IIS restriction enzymes to use for removing the subpool-specific amplification priming sites. The software will attempt to find an enzyme from this list

from this list. The enzyme names should be properly capitalized, placed between double quotes, and separated by a comma (e.g., "REVector": ["BtsI","BsmBI"]). The parameter accepts enzymes defined by the Bio.Restriction module of Biopython (for the latest list, see http://www.biopython.org/DIST/docs/api/Bio.Restriction.Restriction-module.html). This can be left blank if SearchForRE is False.

SearchForRE: Set to True if you wish to specify a list of restriction enzymes within the REVector setting. This should be done in applications in which it is impossible to eliminate a single type IIS enzyme cut site from all constructs. Set to False to use the enzyme listed in REToUse to process all the assembly subpools.

REToUse: Restriction enzyme to use if SearchForRE is set to False. This can be left blank if SearchForRE is True.

avgoverlapsize: Each construct will be broken up into assembly oligos that will be fused using a polymerase. The fusion reaction requires priming through overlaps between neighboring oligos. This setting specifies the mean length of the overlap region.

deltaGThresholdForOverlaps: Rejects any overlaps with a secondary structure that has a hybridization free energy less than the value specified (in units of kcal/mol).

selfDimerThreshold: Rejects assembly oligos that have any self-dimerization configurations with a hybridization free energy less than the value specified (arbitrary units).

lengthleeway: Sets allowable variation in the length of the overlap regions.

positionleeway: Sets allowable variation in the assembly oligo junction position. Increasing this value results in a less constrained search space, but increases the computation time and increases variation in synthesized oligonucleotides' lengths.

oligoSizeMax: The maximum oligo size that will be designed. This includes the full-length oligos that include the coding region, the restriction enzyme processing site, and the assembly-specific and plate-specific priming sites. This value should typically be constrained by the commercial synthesis platform used. Note that many of the oligos will be shorter than this maximum value.

seqsToAvoidInOverlapRegion: Specifies positions to be avoided in the overlap between neighboring assembly oligos. This should usually be left blank, but can be used in specialized applications, such as constructing proteins with known repeated regions.

skip: Specifies names of constructs that the algorithm should skip from the design process. This option is used in specialized cases and is normally left blank.

4. Click Submit. An e-mail will be sent to the provided email confirming initiation of the run. Once the run is complete, two more emails should arrive. The first one will contain a report that contains: (1) The sequences to be synthesized on the DNA chip in FASTA format; (2) The plate-specific, position-specific, and construction primers needed to build the set of assemblies; (3) The plate-specific, position-specific, and construction primers that correlate with each individual assembly. The second e-mail will contain a FASTA file that contains the sequences that should be synthesized on the DNA chip.

Ordering DNA

5. Synthesize a DNA chip using the chip oligonucleotide sequences designed by GASP.

   We have validated and extended this protocol using Agilent's Oligo Library Synthesis (OLS) platform (Leproust et al., 2010), though we have also been able to build genes using DNA from LC Sciences (Borovkov et al., 2010) and CustomArray (Liu et al., 2006).

6. Order the plate-specific, position-specific, and construction primers listed in the output email from a commercial vendor (such as Integrated DNA Technologies or Sigma-Aldrich). If the chip has been designed for more than 48 assemblies, it is usually convenient to synthesize the oligos in a 96-well format. If that is the case,

each century of oligos should be synthesized in a separate plate. For instance, skpp-101, skpp-102, skpp-103 . . . , should be synthesized in separate wells of a single plate.

The primers are named with the following format: skpp-#-F/R. The F and R indicate forward and reverse primers, respectively. The numbering scheme assigns the first numbers to plate-specific primers (skpp-1, skpp-2,. . .). If initialPlateNum was set to 2, then the first plate of assembly subpool-specific primers is numbered skpp-201, skpp-202. . ., the second plate is numbered skpp-301, skpp-302 .., and so forth. The oligos numbered in the 100s are the construction primers. Each pair of construction primers assembles one assembly subpool in each plate. Specifically, skpp-101 assembles assembly subpool generated with skpp-201, skpp-301, skpp-401, and so on.

BASIC PROTOCOL 2

PCR AMPLIFICATION OF OLIGONUCLEOTIDE SUBPOOLS

This set of protocols describes going from a pool of off-chip oligonucleotides to 96-well plates containing a different assembly subpool in each well. We have found that while the experimental procedures are technically simple, the logistics can sometimes present a challenge. We highly recommend that each researcher thoroughly understand the logic behind the subpool amplification scheme (as shown in Fig. 2) before starting the procedure. The quantities provided in the protocol are enough to go from an off-chip pool to one plate-subpool, and then from that one subpool to 96 assembly subpools. The volumes used should be scaled in accordance with the actual number of assemblies being built.

Subpool amplification and assembly steps are punctuated with numerous reaction cleanup steps. Their primary purpose is to remove enzymes, unused primers, dNTPs, and salts. At some points of the protocol they have the additional role of concentrating the DNA. Since the protocol necessitates handling 96-well plates of reactions, it practical to use vacuum manifold-driven 96-well column plates. Although not described in this unit, we are currently working integrating the less expensive magnetic bead-based cleanups into our workflow (Hawkins et al., 1994; Rudi et al., 1997; Wiley et al., 2009).

Materials

Array-synthesized library
100× TE buffer (Sigma-Aldirch, cat. no. T9285)
Plate subpool-specific primers (Basic Protocol 1)
Assembly subpool-specific primers (Basic Protocol 1)
UltraPure DNase/RNase-free distilled water (Invitrogen, cat. no. 10977-023)
Phusion high-fidelity DNA polymerase (New England Biolabs, cat. no. M0530L) containing:
5× Phusion HF Reaction buffer
dNTP solution mix (Enzymatics, cat. no. N205L)
QIAquick 96 PCR purification kit (QIAGEN, cat. no. 28183)
EB buffer (10 mM Tris·Cl, pH 8.5)

Galaxy microcentrifuge with 1.5-ml, 2.0-ml, and 0.2-ml tube adapters (VWR, cat. no. 37000-700)
Vortex mixer (VWR International, cat. no. 58816-121)
0.2-ml PCR tubes with flat caps (BioRad, cat. no. TFI-0201)
Microseal 96-well skirted low-profile PCR plates (BioRad, cat. no. MSP-9601)
96-Reaction thermal cycler (BioRad, cat. no. 186-1096)
Microseal ‘F’ Sealing Foil (BioRad, cat. no. MSF-1001)
QIAvac 96 (QIAGEN, cat. no. 19504)

DNA resuspension and dilution

1. DNA synthesized on Agilent OLS arrays is shipped lyophilized in a microcentrifuge tube. Before opening the tube, centrifuge it for 1 min at maximum speed, room temperature, in a microcentrifuge.

   Using qPCR, we have estimated that an OLS chip with 13,000 130-mers yields 1 pmol of DNA (or approximately 0.1 fmol per oligo species).

2. Add 500 µl 1× TE buffer to the lyophilized DNA. Vortex thoroughly for 5 sec, then briefly centrifuge 5 sec at ~2000 × *g*, room temperature, to spin down liquid.

3. Make mixes of primers by diluting the appropriate plate and assembly subpool-specific primer pairs to 5 µM in DNase/RNase-free water.

   We recommend first making primer storage stocks by diluting each primer to 100 µM in 1× TE buffer. Note that high concentrations of EDTA in the working stocks may inhibit PCR and other enzymatic reactions.

Plate subpool amplification

4. Working on ice mix the following reagents in 0.2-ml PCR tube for each plate subpool to be amplified:

   33.1 µl distilled water
   10 µl Phusion HF Reaction buffer (5×)
   0.4 µl dNTPs (25 mM each deoxynucleotide)
   5 µl plate subpool-specific primer mix (5 µM each primer)
   1 µl array-synthesized library (from step 2)
   0.5 µl Phusion polymerase (2 U/µl).
   Vortex thoroughly 5 sec, then briefly centrifuge 5 sec at ~2000 × *g*, room temperature, to spin down liquid.

   If more then one plate subpool will be amplified, then it is convenient to make a common master mix that contains everything except for the subpool-specific primers. The primers should be added once the master mix has been split among the appropriate number of 45-µl aliquots. The volume of the master mix should be in slight excess of how much is needed for the amplifications. For example, if eight plate subpools are to be amplified, an 8.5 reaction 382.5 µl reaction mix should be made. Be sure to vortex the mix thoroughly prior to dividing into aliquots.

5. Place the samples in a thermal cycler and run the following program:

Initial cycle:	30 sec	98°C	(initial denaturation)
25 cycles:	5 sec	98°C	(denaturation)
	10 sec	65°C	(annealing)
	10 sec	72°C	(extension)
1 cycle:	5 min	72°C	(final extension)
Final step:	indefinite	4°C	(hold).

   To minimize sample loss and increase reproducibility either use a thermal cycler with a heated lid or add a few drops of mineral oil (Sigma-Aldrich, cat. no. M8662) to the top of each reaction. Amplified DNA can be stored at 4°C for less than a month, or indefinitely at −20°C.

   At this stage, it is prudent to quantify yield and ensure proper amplification by running 1 to 10 µl of each of the amplified products on an agarose or acrylamide gel. A double-stranded DNA stain such as ethidium bromide should reveal a band for each oligo of a distinct size in the plate subpool.

Assembly subpool amplification

6. Working on ice make the following master mix for each plate subpool amplified:

   11,636 µl distilled water
   4 µl plate subpool amplification reaction products (from step 5)
   4000 µl Phusion HF reaction buffer (5×)
   160 µl dNTPs (25 mM each)
   200 µl Phusion polymerase (2 U/µl).
   Vortex thoroughly 5 sec, then briefly centrifuge 5 sec at ~2000 × *g*, room temperature, to collect liquid.

7. Working on ice, add 160 µl of the mix to each well of a 96-well plate.

   While an ice bucket will suffice for keeping the 96-well plates cold, we have found 96-well coolers (Eppendorf, cat. no. 022510525) to be a convenient alternative. Also, depending on the thermal cycler and plates used, we routinely split the 200-µl reaction into 50 to 100 µl across multiple plates.

8. Add 40 µl of the appropriate primer mix (forward and reverse assembly-specific primers at 5 µM each) to each well of the 96-well plate that contains the master mix aliquots. Cover the wells with foil plate sealer.

9. Place the plate in a thermal cycler and run the following program:

Initial cycle:	30 sec	98°C	(initial denaturation)
30 cycles:	5 sec	98°C	(denaturation)
	10 sec	65°C	(annealing)
	10 sec	72°C	(extension)
1 cycle:	5 min	72°C	(final extension)
Final step:	indefinite	4°C	(hold).

Reaction cleanup

10. Following the manufacturer's instructions, use the QIAquick 96 PCR Purification kit and the QIAvac 96 to bind the DNA from a 96-well assembly amplification to the kit columns. Elute the DNA on each column into 60 µl EB buffer (10 mM Tris·Cl, pH 8.5).

BASIC PROTOCOL 3

ENZYMATIC REMOVAL OF PRIMING SITES

After the second amplification, the subpools still have their subpool-specific priming sites. These must be removed before assemblies can be built. It is for this purpose that each oligo contains a *Bts*I recognition sequence (GCAGTG) between the assembly subpool priming site and the coding region. *Bts*I, like other member of the type IIs family of restriction enzymes, cuts both strands completely outside of its recognition site (Pingoud and Jeltsch, 2001). Consequently, processing with *Bts*I places no sequence restriction on the coding portion of the oligos. Other type IIs restriction enzyme sites may be used by adjusting the design parameters set in Basic Protocol 1.

Materials

UltraPure DNase/RNase-free distilled water (Invitrogen, cat. no. 10977-023)
*Bts*I (New England Biolabs, cat. no. R0614L) containing:
  NEBuffer 4
  BSA
96-well plate containing the assembly subpools (see Basic Protocol 2)
MinElute 96 UF PCR purification kit (QIAGEN, cat. no. 28051)
EB buffer (10 mM Tris·Cl, pH 8.5)

Seal-Rite 1.5-ml microcentrifuge tubes (USA Scientific, cat. no. 1615-5500)
Vortex Mixer (VWR International, cat. no. 58816-121)
Galaxy mini centrifuge with 1.5/2.0 ml and 0.2 ml tube adapters (VWR, cat. no. 37000-700)
QIAvac 96 (QIAGEN, cat. no. 19504)
96-reaction thermal cycler (BioRad, cat. no. 186-1096)

1. For each 96-well plate of cleaned-up assembly subpools to be processed, prepare the following master mix in a 1.5-ml tube:

   145 µl distilled water
   700 µl NEBuffer 4 (10×)
   70 µl BSA (10 µg/µl)
   85 µl *Bts*I (10 U/µl).

   Vortex thoroughly for 5 sec, and then briefly centrifuge 5 sec ~2000 × *g*, room temperature, to spin down liquid.

2. Add 10 µl of the master mix to each well of the 96-well plate containing the assembly subpools. Cover the wells with plate sealer.

3. Heat the plate to 55°C for 2 hr.

4. Following the manufacturer's instructions, use the MinElute 96 UF PCR Purification kit and the QIAvac 96 to clean up each 96-well of *Bts*I-digested assembly subpools. Elute the DNA bound to each column into 15 µl EB buffer

   We highly recommend quantifying the DNA concentration with a spectrophotometer and/or running the samples out on a gel. The total DNA concentration should be 30 to 300 ng/µl. If all of the oligos in a subpool are the same length, then a gel with sufficient resolution should reveal four bands: (1) the properly processed oligos lacking the two priming sites; (2) the oligos cut on only one side; (3) the uncut oligos; and (4) the cut-off pieces containing the priming sites.

ASSEMBLY OF PROCESSED SUBPOOLS INTO FULL-LENGTH CONSTRUCTS

BASIC PROTOCOL 4

Full-length assemblies are built from assembly subpools in two steps—assembly and amplification. The assembly stage consists of 15 thermal cycles with an annealing step consisting of a thermal ramp from 70°C to 50°C. The full-length assemblies are then selected by adding the appropriate construction primers and performing a 25-cycle PCR. In our experience, the more assembly subpool DNA in the assembly reaction, the more likely a full-length product is formed.

Materials

Construction primers
Assembly subpool-specific primers (Basic Protocol 1)
Ice
UltraPure DNase/RNase-free distilled water (Invitrogen, cat. no. 10977-023)
Phusion high-fidelity DNA polymerase (New England Biolabs, cat. no. M0530L) containing:
   HF reaction buffer
dNTP solution mix (Enzymatics, cat. no. N205L)
QIAquick 96 PCR purification kit (QIAGEN, cat. no. 28183)
QIAvac 96 (QIAGEN, cat. no. 19504)

Seal-Rite 1.5-ml microcentrifuge tubes (USA Scientific, cat. no. 1615-5500)
Vortex mixer (VWR International, cat. no. 58816-121)
Galaxy mini centrifuge with 1.5/2.0 ml and 0.2 ml tube adapters (VWR, cat. no. 37000-700)
Microseal 96-well skirted low-profile PCR plates (BioRad, cat. no. MSP-9601)
96-Reaction thermal cycler (BioRad, cat. no. 186-1096)
5-ml tubes

Assembly

1. Mix the forward and reverse construction primer pairs to 5 µM of each primer in water.

   As with the subpool-specific primers, it is a good practice to first make primer storage stocks by diluting each primer to 100 µM in 1× TE buffer.

2. Working on ice, prepare the following master mix for each 96-well plate of assembly subpools in a 1.5-ml tube:

   164 µl distilled water
   400 µl Phusion HF reaction buffer (5×)
   16 µl dNTPs (25 mM each)
   20 µl Phusion polymerase (2 U/µl).

   Vortex thoroughly for 5 sec, then briefly centrifuge 5 sec to spin down liquid.

   We have also successfully used KOD Polymerase (EMD, cat. no. 71085-3) in place of Phusion.

3. Transfer 6 µl of the master mix into each well of a new 96-well PCR plate.

4. Add 14 µl of each assembly subpool to each well of the PCR plate containing the master mix. Cover the wells with plate sealer.

   A total of 14 µl corresponds to 420 ng of assembly subpool DNA at the lowest expected concentration (30 ng/µl). If the concentration has been measured using a spectrophotometer or gel electrophoresis then all subpools can be diluted to 30 ng/µl for maximum reproducibility, although we have not observed higher concentrations inhibiting the assembly reaction.

5. Place the 96-well plate in a thermal cycler and run the following program:

Initial step:	2 min	95°C	(initial denaturation)
15 cycles:	20 sec	95°C	(denaturation)
	1 sec	70°C	
	30 sec	50°C	(cool to 50°C at 0.5°C/sec; annealing)
	20 sec	72°C	(extension)
1 cycle:	5 min	72°C	(final extension)
Final step:	indefinite	4°C	(hold).

   The annealing occurs as the temperature ramps from 70° to 50°C at 0.5°C/sec. The 1-sec step provides the starting point for the ramp.

Assembly amplification

6. Working on ice, prepare the following master mix for each 96-well plate of assembly subpools in a 5-ml tube:

   3.31 ml distilled water
   1 ml Phusion HF reaction buffer (5×)
   40 µl dNTPs (25 mM each)
   50 µl Phusion polymerase (2 U/µl).

Vortex thoroughly for 5 sec, and then briefly centrifuge 5 sec to spin down liquid.

7. Transfer 44 µl of the master mix into each well of a new 96-well plate.
8. For each plate of assemblies, add 198 µl distilled water to a fresh 96-well plate.
9. Dilute the assembly reactions 1:100 by adding 2 µl of the assembly reaction to each well of the PCR plate with water.
10. To each well of the PCR plate containing the amplification master mix, add 5 µl of the appropriate construction primer mix (5 µM each forward and reverse primers) and 1 µl of the 1:100 dilution of the appropriate assembly reaction. Cover the wells with plate sealer.
11. Place the samples in a thermal cycler and run the following program:

1 cycle:	30 sec	98°C	(initial denaturation)
25 cycles:	5 sec	98°C	(denaturation)
	10 sec	62°C	(annealing)
	20 sec	72°C	(extension)
1 cycle:	5 min	72°C	(final extension)
Final step:	indefinite	4°C	(hold).

Reaction cleanup

12. Following the manufacturer's instructions, use the QIAquick 96 PCR Purification kit and the QIAvac 96 to clean up the DNA from the 96-well assembly amplification to QIAquick 96 PCR purification kit column.

GEL-STAB PCR

SUPPORT PROTOCOL

Assembly amplification, particularly that of GC-rich or repetitive constructs, sometimes results in side-products of the wrong size (typically smaller than the construct). If there is substantial yield of the full-length assembly, then it can be purified using agarose gel isolation. However, in rare cases the full-length assembly forms only a faint band on an agarose gel. This protocol describes a simple technique for selectively re-amplifying the correct assembly product from a large background of incorrect assemblies.

Materials

Assembly amplification product (see Basic Protocol 4)
2% E-Gel EX gel (Invitrogen, cat. no. G4020-02)
1-Kb Plus DNA ladder (Invitrogen, cat. no. 10787-018)
UltraPure DNase/RNase-free distilled water (Invitrogen, cat. no. 10977-023)
Appropriate construction primers
Phusion high-fidelity DNA polymerase (New England Biolabs, cat. no. M0530L) containing:
HF reaction buffer
dNTP solution mix (Enzymatics, cat. no. N205L)
QIAquick PCR purification kit (QIAGEN, cat. no. 28104)

E-Gel iBase power system (Invitrogen, cat. no. G6400)
Thin metal edge or E-Gel opener (Invitrogen, cat. no. G5300-01)
Safe Imager 2.0 blue light transilluminator (Invitrogen, cat. no. G6600)
Seal-Rite 1.5-ml microcentrifuge tubes (USA Scientific, cat. no. 1615-5500)
10- or 20-µl pipet tips
0.2-ml PCR tubes with flat caps (BioRad, cat. no. TFI-0201)
Galaxy mini centrifuge with 1.5/2.0-ml and 0.2-ml tube adapters (VWR, cat. no. 37000-700)

Vortex mixer (VWR International, cat. no. 58816-121)
96-Reaction thermal cycler (BioRad, cat. no. 186-1096)

1. Following the manufacturer's instructions, load 1 to 5 µl of the assembly amplification product and 5 µl of the 1-Kb Plus ladder in separate wells of a 2% E-Gel EX and run on the E-Gel base.

   Traditional agarose gels work, as well as the precast gels, but the amount of DNA loaded may have to be adjusted to account for the differences in gel geometry and the dye used.

2. Using a thin metal edge or an E-Gel opener carefully remove the top plastic cover of the gel cassette.

3. Place the gel on a blue light transilluminator. Add 20 µl water into a 1.5-ml tube. Stab the band that corresponds to the desired assembly product with a sterile 10- or 20-µl pipet tip. Look at the tip to make sure a small slice of the gel is stuck in the tip. Pipet up and down in the water using the same pipet tip. Look at the tip again to make sure that the gel that was stuck there is gone.

4. Heat the water containing the gel slice to 65°C for 15 to 30 min.

5. Combine the following in a 0.2-ml PCR tube:

   4.6 µl distilled water
   1 µl water containing the gel isolate
   2 µl appropriate construction primers (5 µM each primer)
   4 µl Phusion HF Reaction Buffer (5×)
   0.2 µl dNTPs (25 mM each)
   0.2 µl Phusion polymerase (2 U/µl).

   Vortex thoroughly for 5 sec, and then briefly centrifuge for 5 sec at room temperature, to spin down liquid.

6. Place the samples in a thermal cycler and run the following program:

1 cycle:	30 sec	98°C	(initial denaturation)
30 cycles:	5 sec	98°C	(denaturation)
	10 sec	62°C	(annealing)
	20 sec	72°C	(extension)
1 cycle:	5 min	72°C	(final extension)
Final step:	indefinite	4°C	(hold).

7. Cleanup the reaction using a QIAquick kit following the manufacturer's instructions.

COMMENTARY

Background Information

There is a large and fairly mature set of techniques for building genetic constructs without relying on a pre-existing template, including methods for chemically synthesizing short oligonucleotides (Michelson and Todd, 1955; Letsinger and Mahadevan, 1965; Brown, 1993), for using enzymes to fuse the oligonucleotides into long double-stranded fragments (Agarwal et al., 1970; Rossi et al., 1982; Li and Elledge, 2007; Gibson et al., 2009), and for performing in vivo assembly of genome-scale fragments in recombinogenic organisms such as *S. cerevisiae* (Gibson, 2009; Shao et al., 2009). Unfortunately, high costs hinder the widespread adoption of de novo synthesis, primary among them the cost of chemically synthesizing single-stranded oligonucleotides. The maturation of a number of technologies that allow parallel synthesis of thousands or millions of oligonucleotide on solid surfaces (DNA chips) has reduced oligonucleotide cost more than 100-fold over the past decade (Carlson, 2003; Carr and

Church, 2009). Contrary to the expectations of many, and despite a number of proof-of-concept publications (Tian et al., 2004; Zhou et al., 2004; Binkowski et al., 2005), the advent of commercially available DNA chips did not catalyze a widespread drop in the cost of synthesizing double-stranded gene-length fragments.

The three factors that until recently inhibited the adoption of DNA off chips for gene synthesis were the high chemical complexity of the resulting DNA pool, the oligonucleotide error rates, and the low synthesis yields. Chemical complexity refers to the fact that while tens of thousands of oligonucleotide species are synthesized in parallel on the same surface, only 5 to 50 are needed to build any particular gene-sized double-stranded fragment. At typical DNA chip synthesis scales, performing enzymatic assembly of a subset of oligonucletides in the presence of the unrelated species is difficult or impossible (Borovkov et al., 2010). Physical separation of DNA off a chip into subsets of oligonucleotides has recently been achieved using microfluidics (Zhou et al., 2004; Lee et al., 2010; Quan et al., 2011), beads in picotiter plates (Matzas et al., 2010), and barcode PCR (Kosuri et al., 2010). The second problem, high error rates, is important because having more errors raises costs by increasing the amount of screening needed to isolate error-free constructs. Off-chip oligonucleotides typically have an error rate of 1/25 to 1/50 errors/bases, while low-throughput synthesis on controlled pore glass beads results in an error rate of 1/200 to 1/500. Here, too, there has been much recent progress, including improvements in the chemical synthesis process (Leproust et al., 2010), advances in enzyme-mediated error depletion (Carr et al., 2004; Binkowski et al., 2005; Fuhrmann et al., 2005; Bang and Church, 2007; Kosuri et al., 2010; Quan et al., 2011), and the integration of high-throughput sequencing and synthesis for direct error screening (Matzas et al., 2010). The last challenge has been the low yields achieved with on-chip synthesis, which make it difficult to assemble the synthesized oligonucleotides into gene-length fragments. The microfluidic-based methods described above solve this problem by increasing the DNA concentration through low assembly reaction volumes. In contrast, our subpool-based technique achieves high concentrations through selective PCR amplification.

The workflow we presented in this set of protocols relies on the ability to PCR amplify a small subset of oligonucleotides out of a large background of DNA. To prevent hybridization of primers to the wrong priming sites we designed a set of 3000 orthogonal primer pairs (Kosuri et al., 2010). The method, in brief, was as follows: first, we searched the set of 240,000 orthogonal 25-mers designed by Xu et al. (2009) for 20-base windows that lack commonly used restriction enzyme sites and have a melting temperature of 60° to 64°C (SantaLucia, 1998; SantaLucia and Hicks, 2004). The resulting sequences were culled of any pairs that cross-hybridize using a modified version of the AutoDimer tool, a library-on-library BLAT search, and, finally, a BLAST-based network elimination algorithm (Kent, 2002; Vallone and Butler, 2004; Xu et al., 2009). Next, we used UNAFold to select only the sequences with a secondary structure with a $\Delta G \leq -2$ (Markham and Zuker, 2008). The selected sequences were clustered using ClustalW2, and 6000 sequences were selected such that they were maximally spread out on the resulting phylogenetic tree (Larkin et al., 2007). Finally, 3000 primer pairs were generated by matching the sequences on their melting temperature and propensity to form primer-dimers.

Critical Parameters

DNA synthesis

A number of manufacturers sell custom DNA microarrays, and differences between their synthesis methodologies result in differences in the oligo error rates. The error rate affects two key aspects of the gene synthesis process. First, a high error rate means a larger fraction of the synthesized genes will have an error. This directly increases the cost, as constructs must be screened by low-throughput sequencing until a perfect clone is found. A high error rate also makes it more difficult to synthesize longer oligos. Longer oligos are advantageous, as they reduce the complexity of the synthesis reaction by reducing the number of individual DNA pieces that must be put together in order to build a gene of a particular size. Most of our experience with the set of protocols provided in this unit comes from working with Agilent Technologies' Oligo Library Synthesis (OLS) platform, which can synthesize oligonucleotide 100- to 200-bp in length with an error rate on the order of 1/500 errors/bp (Kosuri et al., 2010; Leproust et al., 2010). Because the OLS platform is still under development, OLS pools are not yet widely sold. However, laboratories can gain access to them by signing a material transfer agreement

with Agilent Technologies. The protocols we describe should be applicable to all other array platforms, although if the DNA chip has a high error, then error correction may be necessary. Similarly, if the off-chip oligonucleotides are shorter than 120 bases, then different DNA cleanup steps may have to be used, as the QIAGEN columns used in the protocols do not recover fragments shorter than 40 bp.

It is useful to keep in mind that DNA from microarrays is one of the cheapest reagents in the synthesis process. Therefore, in the interest of rapid prototyping, it often makes sense to waste synthesis capacity rather than trying to design enough constructs to fill up all features on the chip. Unfortunately, subpool-specific and construction primers are expensive because they must be synthesized using low-throughput methods. However, primer synthesis yields are typically in excess of what is needed to process a single set of assemblies, so the primers can be used to process multiple DNA chips.

Construct design

If the constructs being built encode proteins, then there are many possible ways to reverse-translate the amino acid sequence (Welch et al., 2009b). One consideration when choosing among synonymous codons is the natural codon bias within the genome of organism that will express the protein (Gustafsson et al., 2004). Similarly, studies in *Escherichia coli* revealed that using codons that are most highly charged during amino acid starvation leads to significantly higher levels of protein expression (Welch et al., 2009a). The second variable to take into account when designing constructs is the propensity of the resulting DNA to form secondary structures that interfere with polymerase-based assembly methods. Despite our efforts to optimize the robustness of our assembly protocol, the reaction will fail often if the construct being built is highly GC-rich or repetitive. For designing protein-coding DNA sequences, we recommend GeneDesign, a Web-based sequence manipulation environment that allows facile manipulation of codon usage, GC-bias, restriction site placement, and other design parameters (Richardson et al., 2010).

Downstream processing

An assembly reaction produces a heterogeneous population of molecules, some of which contain point mutations or structural rearrangements, such as large deletions or insertions. If error-free constructs are needed, then the assemblies must be cloned and screened by sequencing. If the constructs are to be used in a screen or selection, then it may be possible to use the heterogeneous assembly products directly. Error rates can be reduced by creating mismatches by melting and reannealing the assemblies and depleting heteroduplexes by a MutS pulldown (Carr et al., 2004; Binkowski et al., 2005) or cleavage with resolvases (Fuhrmann et al., 2005; Bang and Church, 2007). Assemblies with frame-shift mutations can be screened for by cloning them in-frame to a downstream *lacZ* fragment or fluorescent protein (Cronan et al., 1988; Kim et al., 2010). Error rates can be estimated by building control fluorescent protein assemblies, which allow the error rate to be estimated by counting colonies or performing flow cytometry. Such protein function assays give an accurate error rate estimate if most errors are deletions and insertions. However, due to the degeneracy of the genetic code and the stability of protein function in the face of some residue substitutions, function-based screens underestimate the error rate if there are many transitions and transversions.

Troubleshooting

No assembly products or products that are too short

Assembly of GC-rich or repetitive sequences sometimes proves challenging. We have found that most of the time this problem can be solved by increasing the amount of processed assembly subpool DNA added to the assembly reaction. At 5 to 20 ng/μl, the highest assembly subpool concentrations tested, we were able to get a correctly sized band for ~95% of the difficult templates tested (Kosuri et al., 2010). Assembly reactions can be further optimized by introducing PCR additives that increase specificity and reduce DNA secondary structure, such as betaine and dimethyl sulfoxide (DMSO; Winship, 1989; Henke et al., 1997). Similarly, varying annealing temperature or the concentrations of magnesium ions and deoxynucleotides may alter PCR performance (Cobb and Clarkson, 1994; Roux, 1995).

Assembly results in short side-products

Gel-isolate the band of the correct size and re-amplify using the appropriate construction primers (see Support Protocol).

Assembly results in high-molecular-weight smears

This is typically caused by "over-amplifying" the template, which happens

when PCR primers run out, causing the amplicon molecules to generate long fusions by mispriming off each other (Bell and DeMarini, 1991). This can be avoided by decreasing the number of amplification cycles, increasing the concentration of construction primers, decreasing the amount of post-assembly DNA added to the assembly amplification reaction, shortening extension times, or optimizing PCR specificity as described in the section on troubleshooting a lack of assembly products.

Synthesized genes have high error rates

Errors can be introduced during the chemical synthesis of DNA on a chip or at any of the subsequent enzymatic steps. Sequencing at various points of the synthesis workflow may help detect the processes that introduce the most error. Error detection or depletion techniques can be introduced after the assembly amplification step, some of which are described in greater detail within the *Downstream processing* section of Critical Parameters.

Loss of DNA during reaction cleanups

Our protocols rely on using QIAGEN's column-based PCR cleanups that have reduced recovery of double-stranded DNA shorter than 100 bp, and no recovery of DNA shorter than 40 bp. It is important to keep track of how the DNA length changes at each step of the protocol, as it decreases after both the assembly subpool amplification and the *Bts*I digest. If DNA is lost during cleanups in a size-dependent manner, then it is best to switch to either a different column-based cleanup kit or to use alternative methods such as magnetic beads (Hawkins et al., 1994; Rudi et al., 1997; Wiley et al., 2009).

Anticipated Results

Basic Protocol 1 will generate a FASTA file with the sequences of the oligonucleotides that need to be synthesized on the DNA chip. After the oligonucleotides have been synthesized and cleaved off the array surface, they will be amplified first into plate subpool and then into assembly subpools, as described in Basic Protocol 2. Running each sample out on an agarose or a polyacrylamide gel should result only bands that correspond to the expected lengths of the DNA species in each subpool. Note assembly subpools will be shorter than the plate subpools because plate-specific priming sites flank the assembly-specific priming sites. Basic Protocol 3 will result in the removal of the assembly-specific priming sites. Gel electrophoresis will reveal bands corresponding to the cleaved-off priming sites, the members of the assembly subpool with the priming sites removed, and partially digested DNA cut on just one of the two sides. Basic Protocol 4 should result in full-length assemblies, some of which may have to be gel-isolated or, if the assembly amplification yields were low, re-amplified from a gel-isolated sample. The constructs are ready to be used for downstream applications once they produce a single band of the expected size when resolved by gel electrophoresis.

Time Considerations

Going from a list of sequences to be built to a chip design should take a couple of hours. Once both the DNA chip and the amplification primers have been synthesized, the synthesis process should take 1 to 2 days, though assembling a large number of constructs will increase the amount of time spent setting up reactions (and, conversely, automation will reduce it significantly). Post-assembly gel-isolation and re-amplification typically add an additional day of work.

Acknowledgements

This work was supported by the U.S. Office of Naval Research (N000141010144), National Human Genome Research Institute Center for Excellence in Genomics Science (P50 HG003170), and the Department of Energy Genomes to Life grant (DE-FG02-02ER63445) (all to G.M.C.). Also, we thank Jeff Sampson from Agilent Technologies for helpful discussions related to process development and oligonucleotide libraries.

Literature Cited

Agarwal, K.L., Büchi, H., Caruthers, M.H., Gupta, N., Khorana, H.G., Kleppe, K., Kumar, A., Ohtsuka, E., Rajbhandary, U.L., Van de Sande, J.H., Sgarawella, H., and Yamada, T. 1970. Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature* 227:27-34.

Bang, D. and Church, G.M. 2007. Gene synthesis by circular assembly amplification. *Nat. Methods* 5:37-39.

Bayer, T.S., Widmaier, D.M., Temme, K., Mirsky, E.A., Santi, D.V., and Voigt, C.A. 2009. Synthesis of methyl halides from biomass using engineered microbes. *J. Am. Chem. Soc.* 131:6508-6515.

Bell, D.A. and DeMarini, D.M. 1991. Excessive cycling converts PCR products to random-length higher molecular weight fragments. *Nucleic Acids Res.* 19:5079.

Binkowski, B.F., Richmond, K.E., Kaysen, J., Sussman., M.R., and Belshaw, P.J. 2005.

Correcting errors in synthetic DNA through consensus shuffling. *Nucleic Acids Res.* 33:e55.

Borovkov, A.Y., Loskutov, A.V., Robida, M.D., Day, K.M., Cano, J.A., Le Olson, T., Patel, H., Brown, K., Hunter, P.D., and Sykes, K.F. 2010. High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. *Nucleic Acids Res.* 38:e180.

Brown, D.M. 1993. A brief history of oligonucleotide synthesis. *In* Methods in Molecular Biology, Vol. 20: Protocols for Oligonucleotides and Analogs (S. Agrawal, ed.) pp. 1-20. Humana Press, Totowa, N.J.

Carlson, R. 2003. The pace and proliferation of biological technologies. *Biosecur. Bioterror.* 1:203-214.

Carr, P. A. and Church, G.M. 2009. Genome engineering. *Nat. Biotechnol.* 27:1151-1162.

Carr, P.A., Park, J.S., Lee, Y.-J., Yu, T., Zhang, S., and Jacobson, J.M. 2004. Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res.* 32:e162.

Cobb, B.D. and Clarkson, J.M. 1994. A simple procedure for optimizing the polymerase chain reaction (PCR) using modified Taguchi methods. *Nucleic Acids Res.* 22:3801-3805.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hood, M.J.L2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422-1423.

Cronan, J.E. Jr., Narasimhan, M.L., and Rawlings, M. 1988. Insertional restoration of β-galactosidase α-complementation (white-to-blue colony screening) facilitates assembly of synthetic genes. *Gene* 15:161-170.

Fuhrmann, M., Oertel, W., Berthold, P., and Hegemann, P. 2005. Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage. *Nucleic Acids Res.* 33:e58.

Gibson, D.G. 2009. Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res.* 37:6984-6990.

Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A., Merryman, C., Young, L., Noskov, V.N., Glass, J.I., Venter, J.C, Hutchison, C.A. III, and Smith, H.O. 2008. Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319:1215-1220.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* 6:343-345.

Gustafsson, C., Govindarajan, S., and Minshull, J. 2004. Codon bias and heterologous protein expression. *Trends Biotechnol.* 22:346-353.

Hanai, T., Atsumi, S., and Liao, J.C. 2007. Engineered synthetic pathway for isopropanol production in *Escherichia coli*. *Appl. Environ. Microb.* 73:7814-7818.

Hawkins, T.L., O'Connor-Morin, T., Roy, A., and Santillan, C. 1994. DNA purification and isolation using a solid-phase. *Nucleic Acids Res.* 22:4543-4544.

Henke, W., Herdel, K., Jung, K., Schnorr, D., and Loening, S.A. 1997. Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Res.* 25:3957-3958.

Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* 12:656-664.

Kim, H., Han, H., Shin, D., and Bang, D. 2010. A fluorescence selection method for accurate large-gene synthesis. *Chembiochem* 11:2448-2452.

Kosuri, S., Eroshenko, N., Leproust, E.M., Super, M., Way, J., Li, J.B., and Church, G.M. 2010. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.* 28:1295-1299.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948.

Lee, C.C., Snyder, T.M., and Quake, S.R. 2010. A microfluidic oligonucleotide synthesizer. *Nucleic Acids Res.* 38:2514-2521.

Leproust, E.M., Peck, B.J., Spirin, K., McCuen, H.B., Moore, B., Namsaraev, E., and Caruthers, M.H. 2010. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* 38:2522-2540.

Letsinger, R.L. and Mahadevan, V. 1965. Oligonucleotide synthesis on a polymer support. *J. Am. Chem. Soc.* 87:3526-3527.

Li, M.Z. and Elledge, S.J. 2007. Harnessing homologous recombination *in vitro* to generate recombinant DNA via SLIC. *Nat. Methods* 4:251-256.

Liu, R.H., Munro, S.B., Nguyen, T., Siuda, T., Suciu, D., Bizak, M., Slota, M., Fuji, H.S., Danley, D., and McShea, A. 2006. Integrated microfluidic CustomArray device for bacterial genotyping and identification. *J. Assoc. Lab. Autom.* 11:360-367.

Markham, N.R. and Zuker, M. 2008. UNAFold: Software for nucleic acid folding and hybridization. *In* Bioinformatics, Vol. 2: Structure, Function and Applications, Vol. 453 (J.M. Keith, ed.) pp. 3-31. Humana Press, Totowa, N.J.

Matzas, M., Stähler, P.F., Kefer, N., Siebelt, N., Boisguérin, V., Leonard, J.T., Keller, A., Stähler, C.F., Häberle, P., Gharizadeh, B., Babrzadeh, F., and Church, G.M. 2010. High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat. Biotechnol.* 28:1291-1294.

Michelson, A.M. and Todd, A.R. 1955. Nucleotides part XXXII. Synthesis of a dithymidine dinucleotide containing a 3′: 5′-internucleotidic linkage. *J. Chem. Soc.* 2632-2638.

Nirenberg, M.W. and Matthaei, J.H. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U.S.A.* 47:1588-1602.

Pingoud, A. and Jeltsch, A. 2001. Structure and function of type II restriction endonucleases. *Nucleic Acids Res.* 29:3705-3727.

Quan, J., Saaem, I., Tang, N., Ma, S., Negre, N., Gong, H., White, K.P., and Tian, J. 2011. Parallel on-chip gene synthesis and application to optimization of protein expression. *Nat. Biotechnol.* 29:449-452.

Richardson, S.M., Nunley, P.W., Yarrington, R.M., Boeke, J.D., and Bader, J.S. 2010. GeneDesign 3.0 is an updated synthetic biology toolkit. *Nucleic Acids Res.* 38:2603-2606.

Ro, D.-K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., Chang, M.C.Y, Withers, S.T., Shiba, Y., Sarpong, R., and Keasling, J.D. 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440:940-943.

Rossi, J.J., Kierzek, R., Huang, T., Walker, P.A., and Itakura, K. 1982. An alternate method for synthesis of double-stranded DNA segments. *J. Biol.Chem.* 257:9226-9229.

Roux, K.H. 1995. Optimization and troubleshooting in PCR. *Genome Res.* 4:S185-S194.

Rudi, K., Kroken, M., Dahlberg, O.J., Deggerdal, A., Jakobsen, K.S., and Larsen, F. 1997. Rapid, universal method to isolate PCR-ready DNA using magnetic beads. *BioTechniques* 22:506-511.

SantaLucia, J. Jr. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U.S.A.* 95:1460-1465.

SantaLucia, J. Jr. and Hicks, D. 2004. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* 33:415-440.

Shao, Z., Zhao, H., and Zhao, H. 2009. DNA assembler, an *in vivo* genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.* 37:e16.

Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X., and Church, G. 2004. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 432:1050-1054.

Vallone, P.M. and Butler, J.M. 2004. AutoDimer: A screening tool for primer-dimer and hairpin structures. *BioTechniques* 37:226-231.

Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J., and Gustafsson, C. 2009a. Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One* 4:e7002.

Welch, M., Villabos, A., Gustafsson, C., and Minshull, J. 2009b. You're one in a googol: Optimizing genes for protein expression. *J. R. Soc. Interface* 6:S467-S476.

Wiley, G., Macmil, S., Qu, C., Wang, P., Xing, Y., White, D., Li, J., White, J.D., Domingo, A., and Roe, B.A. 2009. Methods for generating shotgun and mixed shotgun/paired-end libraries for the 454 DNA sequencer. *Curr Protoc. Hum. Genet.* 61:18.1.1-18.1.21.

Winship, P.R. 1989. An improved method for directly sequencing PCR amplified material using dimethyl sulfoxide. *Nucleic Acids Res.* 17:1266.

Xu, Q., Schlabach, M.R., Hannon, G.J., and Elledge, S.J. 2009. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci. U.S.A.* 106:2289-2294.

Zhou, X., Cai, S., Hong, A., You, Q., Yu, P., Sheng, N., Srivannavit, O., Maranjan, S., Rouillard, J.M., Xia, Y., Zhang, X., Xiang, Q., Ganesh, R., Zhu, Q., Matejko, A., Gulari, E., and Gao, X. 2004. Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Res.* 32:5409-5417.

Future directions for high-throughput splicing assays in precision medicine

Christy L. Rhine[1*] | Christopher Neil[1*] | David T. Glidden[2*] | Kamil J. Cygan[1,2*] | Alger M. Fredericks[1] | Jing Wang[1] | Nephi A. Walton[4] | William G. Fairbrother[1,2,3]

[1]Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, Rhode Island

[2]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island

[3]Hassenfeld Child Health Innovation Institute of Brown University, Providence, Rhode Island

[4]Genomic Medicine Institute, Geisinger, Danville, Pennsylvania

Correspondence
William G. Fairbrother, Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, RI.
Email: william_fairbrother@brown.edu

Funding information
National Human Genome Research Institute, Grant/Award Numbers: U41 HG007346, R13 HG006650; National Institute of General Medical Sciences, Grant/Award Number: R01 GM127472

Abstract

Classification of variants of unknown significance is a challenging technical problem in clinical genetics. As up to one-third of disease-causing mutations are thought to affect pre-mRNA splicing, it is important to accurately classify splicing mutations in patient sequencing data. Several consortia and healthcare systems have conducted large-scale patient sequencing studies, which discover novel variants faster than they can be classified. Here, we compare the advantages and limitations of several high-throughput splicing assays aimed at mitigating this bottleneck, and describe a data set of ~5,000 variants that we analyzed using our Massively Parallel Splicing Assay (MaPSy). The Critical Assessment of Genome Interpretation group (CAGI) organized a challenge, in which participants submitted machine learning models to predict the splicing effects of variants in this data set. We discuss the winning submission of the challenge (MMSplice) which outperformed existing software. Finally, we highlight methods to overcome the limitations of MaPSy and similar assays, such as tissue-specific splicing, the effect of surrounding sequence context, classifying intronic variants, synthesizing large exons, and amplifying complex libraries of minigene species. Further development of these assays will greatly benefit the field of clinical genetics, which lack high-throughput methods for variant interpretation.

KEYWORDS
assay, disease, high-throughput, precision medicine, splicing, variant

1 | INTRODUCTION

The cost of next generation sequencing (NGS) has fallen several thousand-fold in the last 10 years, which has allowed for whole-genome sequencing and whole-exome sequencing to become common approaches in personal genomics and clinical medicine. A typical exome sequencing study will reveal thousands of variants of unknown significance (Telenti et al., 2016). The effects of these coding variants on protein function are particularly difficult to interpret, as individual functional assays do not exist for most proteins. However, variants in splice-regulatory elements typically result in deleterious phenotypes. Splicing mutations are not only harmful, but they are also prevalent. In fact, it has been predicted that one-third of all disease-causing variants confer some degree of aberrant splicing (Lim, Ferraris, Filloux, Raphael, & Fairbrother, 2011). The effect of variants on splicing is measurable through the use of minigene assays (Cooper, 2005). Because splicing mutations are deleterious, prevalent, and measurable, splicing minigene assays are a valuable method for interpreting the pathogenicity of variants discovered.

As thousands of variants are discovered in sequencing studies, the challenge for precision medicine lies in the ability to classify variants at the same rate they are discovered. Recently, our group

*Christy L. Rhine, Christopher Neil, David T. Glidden, and Kamil J. Cygan contributed equally to this work.

developed a Massively Parallel Splicing Assay (MaPSy) to screen ~5,000 disease-causing exonic mutations for splicing defects. Using highly stringent criteria, this study showed that 10% of exonic mutations altered splicing (Soemedi et al., 2017b). The ability to evaluate variants for defective splicing is beginning to emerge as an achievable goal with the advent of massively parallel reporter assays (MPRAs) and high-throughput screens (Adamson, Zhan, & Graveley, 2018; Ke et al., 2018; Soemedi et al., 2017b). Computational methods aimed at leveraging MPRAs and high-throughput assay data have led to improved predictive models for classifying splicing variants that have not been empirically verified (Bretschneider, Gandhi, Deshwar, Zuberi, & Frey, 2018; Desmet et al., 2009; Fairbrother, Yeh, Sharp, & Burge, 2002; Mort et al., 2014). The Critical Assessment of Genome Interpretation (CAGI) recognized the need for a community effort in advancing the computational methods in predicting the impacts of genomic variation and devised a prediction challenge. In the challenge, participants were asked to identify variants causing splicing defects and estimate the severity of each defect. A variety of different machine learning approaches were submitted, and the top performer, a program called MMSplice, was recently described in a publication (Cheng et al., 2019). Here, we outline methods, challenges, and future directions for this hybrid experimental/computational approach. We focus in particular on the role of MPRAs in functional genomics and clinical medicine. In addition to applications of this technology to consortia sequencing science and discussing drug screening technologies, the effect of sequence context on splicing in MPRAs and technical issues relating to oligonucleotide synthesis are discussed.

2 | CAGI AND THE MAPSY DATA SET CHALLENGE

Increasingly sophisticated predictive models have been developed to estimate the effect of variants on splicing. Many of these models are trained on data from various implementations of MPRAs (FAS-ESS, ESRseq scores, and HAL; Ke et al., 2011; Rosenberg, Patwardhan, Shendure, & Seelig, 2015; Wang et al., 2004). However, these models lack training on large data sets describing the effect of single nucleotide variants on the process of splicing, hindering their ability to produce reliable splicing predictions for variant interpretation. The development of the MaSPy has now offered the splicing and machine learning fields with a rich training data set describing the effect of ~5,000 single nucleotide variants on splicing. Recognizing the need for improved prediction models for variant interpretation, CAGI devised a competition where multiple machine learning teams were challenged to construct splicing variant predictive models to aid in the variant interpretation demands facing precision medicine. The following section describes the MaPSy training and test sets provided to CAGI, the challenge posed to the machine learning teams, and the resulting machine learning splicing model which outperformed the competing CAGI teams.

2.1 | Massively Parallel Splicing Assay (MaPSy) experiment

The challenge was prepared from a splicing analysis of publicly available disease-causing variants. Nonsynonymous mutations classified as disease-causing (DM) were downloaded from Human Genome Mutation Database (Stenson et al., 2009). Mutations were mapped to internal exons of 100 nucleotides or less in length and selected for those that fit into 170 nucleotide genomic windows. The genomic window included 15 nucleotides of downstream intronic sequence and at least 55 nucleotides of upstream intronic sequence (n = 4,964). The mutant and wild-type versions of the 170-mer genomic fragments were flanked with 15-mer common primers and synthesized as a 200-mer oligo library (Figure 1a).

An additional 797 mutations were mapped to exons greater than 100 nucleotides. Each of these longer variant exons were "cut" in a way to (a) preserve the 5′ and 3′ splice site signals and (b) a middle portion of the exon was removed to decrease the size of the exon to 100 nucleotides or less to meet oligonucleotide synthesis size restrictions.

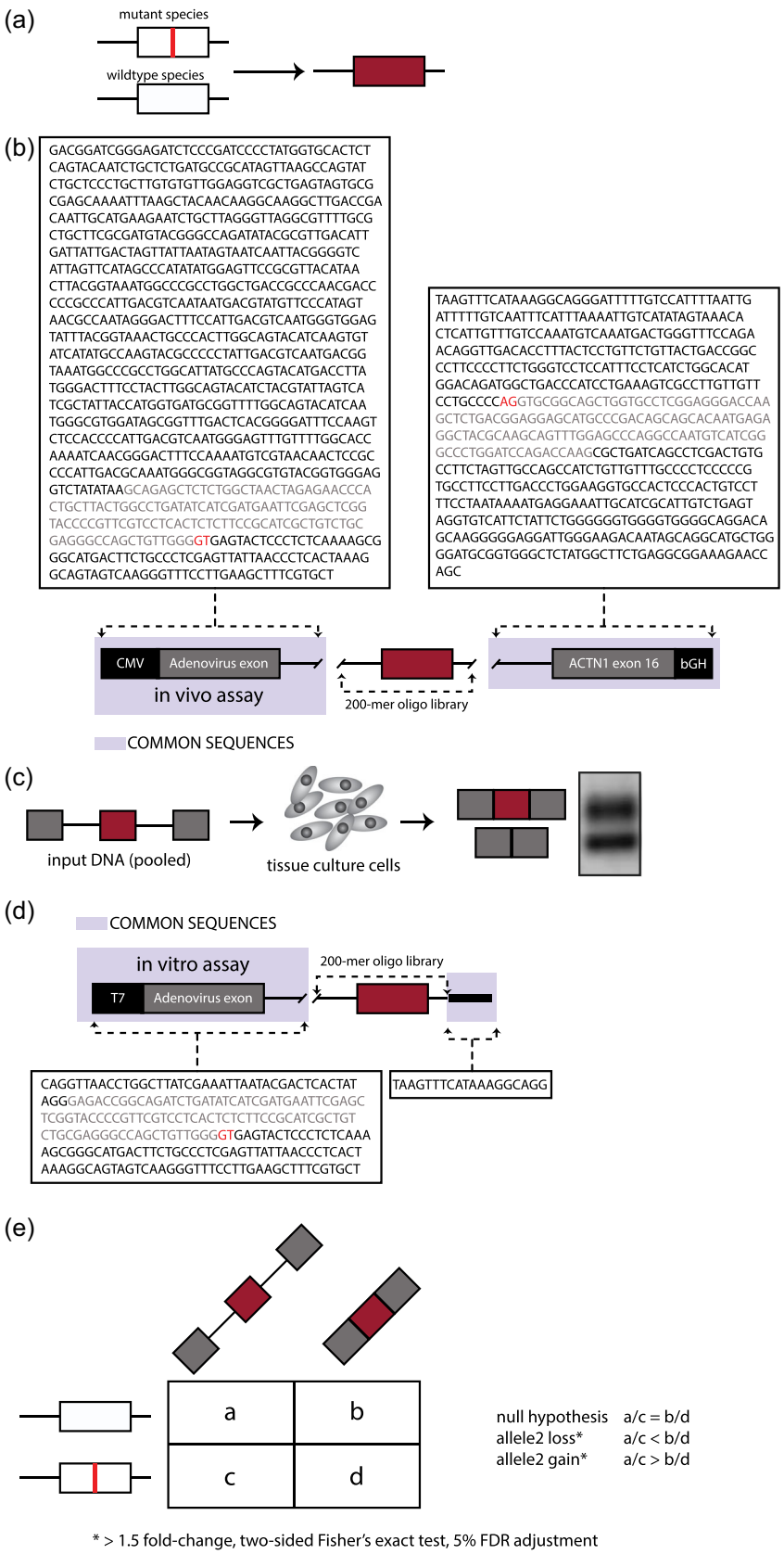
A three exon in vivo splicing reporter was constructed to include a Cytomegalovirus (CMV) promoter and a common first exon, followed by the 200-mer oligo library, and a common downstream exon (Figure 1b). The resulting in vivo reporters were transfected to human embryonic kidney HEK293T cells. RNA was extracted 24 hr post transfection (Figure 1c). Input reporters and spliced species were sequenced by Illumina HiSeq. 2500.

The in vitro splicing reporter includes a T7 promoter and a common first exon, followed by the oligo library (Figure 1d). In vitro reporters were obtained via in vitro transcription using T7 RNA Polymerase. The resulting RNA was gel purified and used for splicing reactions in 40% HeLa-S3 nuclear extract for 80 min at 30°C. Pools of input and spliced RNAs were converted to complemenatry DNA (cDNA) and prepped into an Illumina library for deep sequencing.

A contingency table was created for each mutant/wild-type pair and includes the counts obtained from deep sequencing of the input pool as well as the output-spliced fractions (Figure 1e). To determine pairs with significant allelic skew we required at least 1.5-fold change and a two-sided Fisher's exact test adjusted with 5% false discovery rate (FDR). The following formula was used to calculate allelic skew: $log_2\left(\frac{mut_s / mut_i}{wt_s / wt_i}\right)$, where mut_s is the count of reads in the spliced fraction for the mutant, mut_i is the count of reads in the input for the mutant, wt_s is the count of reads in the spliced fraction for the wild-type, and wt_i is the count of reads in the input for the wild-type.

2.2 | Prediction challenge

Two sets of variants that were tested by MaPSy were provided to CAGI, the training set and the test set. The training set included all 4,964 published variants (Soemedi et al., 2017b). The test set includes all 797 mutant/wild-type pairs of variants that fall within exons that needed to by "cut" to fit in the oligonucleotide library. The



* > 1.5 fold-change, two-sided Fisher's exact test, 5% FDR adjustment

FIGURE 1 MaPSy challenge (a) Mutant and wild-type versions of 170-mer genomic fragments flanked by 15-mer common. (b) The in vivo splicing reporter consists of the Cytomegalovirus (CMV) promoter and Adenovirus (pHMS81) exon with part of its downstream intron at the 5′ end, followed by the 200-mer oligo library, and exon16 of ACTN1 with part of intron15 and bGH PolyA signal sequence at the 3′ end. (c) The in vivo reporters were transfected in hek293 cells. (d) The in vitro reporter includes a T7 promoter and Adenovirus (pHMS81) exon. (e) Contingency tables were created for each mutant/wild-type pair and include the counts obtained from deep sequencing of the input pool as well as the output-spliced fractions to assess defects in splicing

sequence for both constructs including exon/intron boundaries needed to evaluate allelic ratio, as well as the counts for the input for both mutant and wild-type species for both panels were provided.

CAGI participants were asked to submit predictions on the subset of variants in the test set that passed our threshold as disruptors of splicing both in vitro and in vivo and therefore were categorized as exonic splicing mutations. Participants provided the probability that each variant is an exonic splicing mutation and which allele from each pair spliced better. In addition, given the input read counts, the participants predicted the log2 allelic skew ratio for in vivo and in vitro panels for each pair in the test set.

2.3 | Prediction winner: Modular modeling of splicing

The winning prediction model, modular modeling of splicing (MMSplice), trained a set of neural network modules separately for exons, 3′ and 5′ splice-sites, and intronic sequences. This method of building modules for individual splicing-relevant sequence regions allowed the group to leverage multiple data sets to predict percent spliced-in (psi) values, splicing efficiency, and pathogenicity. The resulting program, MMSplice, was shown to outperform previous highly predictive models on predicting the effect genetic variants have on splicing (Cheng et al., 2019).

3 | HIGH-THROUGHPUT METHODS IN SPLICING

The success of the challenge prompted an examination of the potential for this technology in precision medicine. Building from advancements in solid-phase oligonucleotide synthesis technologies, massively parallel reporter assays (MPRAs) have become an increasingly attractive approach for the study of alternative splicing (Park, Pan, Zhang, Lin, & Xing, 2018). Extensive libraries of sequence variants constructed into minigene reporters can be screened in parallel for functional impacts on splicing. The study of wild-type and variant exons in minigene cassettes allows for direct assessment of sequence contributions to splicing outcomes (Singh & Cooper, 2006). Such MPRAs have been used to analyze the ability of sequence variants in degenerate or mutationally saturated libraries to influence 5′ and 3′ splice site selection (Rosenberg et al., 2015; Wong, Kinney, & Krainer, 2018) and exon definition (Ke et al., 2018). Recently, an MPRA that measured mutations within the primate lineage helped identify a mathematical equation to calculate the magnitude of splicing disruption caused by a novel exonic mutation. In this equation, exonic mutations have a maximal impact in exons with an intermediate degree of splicing (Baeza-Centurion, Minana, Schmiedel, Valcarcel, & Lehner, 2019). In other words, less efficient splicing substrates are more prone to splicing defects. MaPSy, another MPRA, provides a direct measure of splicing disruption caused by exonic mutations. Mutations from thousands of different exons can be assayed in parallel, offering both insights into the determinants of splicing aberrations and a potential high-throughput technology for the classification of disease variants (Soemedi et al., 2017b).

MPRAs are not the only approach for testing the effects of variation on splicing. The CRISPR-Cas9 system has also been used to screen thousands of mutations in parallel within endogenous genomic loci (CRISPR-arrays). In CRISPR-arrays, pools of guide RNAs are used to introduce numerous mutations at one locus. For example, a 6-bp region of BRCA1 exon 18 was replaced with all possible hexamers. The utility of CRISPR-arrays is widely applicable in functional genomics. They have identified novel regulatory elements, pathogenic variants, and quantified effects such as nonsense-mediated decay (Canver et al., 2017; Findlay, Boyle, Hause, Klein, & Shendure, 2014; Sanjana, 2017).

CRISPR-arrays have some unique advantages. By editing endogenous genes, they capture the physiologic context of the cell. All relevant cis-elements or secondary structural components are preserved. They are also unconstrained by size limitations of solid-state oligonucleotide synthesis. Therefore, any full-length exon may be screened by this method. It is also more technically straightforward to construct a pool of guide RNAs than a pool of minigene species, which may require polymerase chain reaction (PCR) and other molecular biology techniques to assemble. Despite these advantages, there are some important drawbacks to consider. Haplotype cells lines have been required to achieve efficient multiplex gene editing with CRISPR-Cas9. CRISPR-array throughput can test variants to saturation, but only within a small window. In other words, CRISPR screening is limited to one exon per experiment, and also requires sufficient gene expression for downstream analysis. Lastly, genes considered essential for cell survival may pose additional limitations (Findlay et al., 2014). Splicing mutations in essential genes may have lethal effects, because the only copy of the gene is mutated in these assays.

MPRAs have several advantages over CRISPR-arrays. Because they utilize minigenes, MPRAs are not dependent on endogenous gene expression. MPRAs tend to represent a pure measure of splicing effects. Many clinical whole-exome sequencing data sets are being generated, which return large numbers of variants for interpretation. MPRAs that leverage minigenes are better suited for studying the functional consequences of these variants at the scale and widespread genomic distribution of variants returned by exome sequencing. For example, MPRAs can assay many or potentially all variants of interest from a whole-exome sequencing study instead of being restricted to one gene or exon (Adamson et al., 2018; Soemedi et al., 2017b). Therefore, MPRAs are the method of choice to analyze variants from consortia sequencing because of these unique advantages over other methods, like CRISPR-arrays.

There are still several challenges that limit the potential of MPRAs. First, splicing is a tissue-specific process. For example, the brain has the highest degree of exon skipping, and splicing in the liver is almost entirely limited to cryptic alternative splicing events (alternative 5′ and 3′ splice sites; Yeo, Holste, Kreiman, & Burge, 2004). MPRAs only report splicing outcomes in one tissue type, and

cannot be extrapolated to other tissue types. However, we can identify similar splicing events across tissues from RNA-seq studies in multiple tissues, such as the GTEx consortium, in to determine the potential effect of a variant across tissue types (Consortium et al., 2017). In addition, MPRAs rely on artificial minigene constructs that lack valuable surrounding endogenous sequence context that may impact splicing. Exons that are tested in these assays are typically flanked by common exons to all species in a minigene library as opposed to the exons from the endogenous transcripts. Sequence context from the whole pre-mRNA transcript affects the order of intron removal and can lead to alternative splicing events not captured in MinGenes (Kim et al., 2017). Moreover, as splicing is a cotranscriptional process, chromatin binding state, absent in minigenes, has also been shown to affect splicing (Jaganathan et al., 2019). Lastly, large data sets containing genetic variants for screening by MPRAs often lack corresponding phenotypic or other relevant patient information and thus limit the use of MPRAs in returning informative variant discoveries.

4 | NEW SCIENTIFIC AND HEALTHCARE INITIATIVES: GEISINGER HEALTH SYSTEM AND SIMONS FOUNDATION OF AUTISM RESEARCH INITIATIVE

Many data sets reporting disease-causing or disease-associated variants, such as the Human Gene Mutation Database (Stenson et al., 2009) and ClinVar (Landrum et al., 2016), have limited information on clinical phenotypes and lack methods in contacting and/or requesting biospecimens from patients. Fortunately, new scientific and healthcare initiatives have recognized the need in accurately identifying and interpreting genomic findings that will prove relevant to clinical efforts and precision medicine. Such relationships provide a direct means for validation of functional genomic approaches, return incidental findings to patients, and further analyze the relationship between variants and patient phenotypic characteristics.

A prominent example of this type of integrated data set can be found in the DiscovEHR cohort (Dewey et al., 2016). Through a partnership with the Regeneron Genetics Center, Geisinger has created the DiscovEHR cohort (Dewey et al., 2016). The DiscovEHR cohort is a large population of patients from the Geisinger healthcare system who have had exome sequencing added to their electronic healthcare records to pair genotype with phenotype in a single data set. This patient cohort currently includes 92,805 participants drawn entirely from participants in the MyCode Community Health Initiative (Carey et al., 2016). MyCode participants are consented for collection of biospecimens to be used in conjunction with all the data from their electronic health record (EHR). The participants in MyCode have an average of 14 years of medical records that can be linked with their exome sequencing results, including; clinical notes, lab values, ICD10 Codes, medications, and imaging studies. This combination of genotypic and detailed phenotypic information provides for an extremely rich data set for genomic discovery. MyCode participants are also consented for recontact for additional research which allows for additional clinical evaluation with more targeted phenotyping to supplement the rich data set that already exists in the EHR. DiscovEHR has already been proven to be a tremendous resource for genomic discovery (Abul-Husn et al., 2018; Gusarova et al., 2018; Verma et al., 2019).

An integrated data set such as the DiscovEHR cohort is a suitable platform for the aggregation of data from additional functional genomic experiments like high-throughput splicing assays. By comparing the comprehensive profiles of patients with splicing variants to matched controls, overrepresented phenotypes can be discovered that are representative of known gene effects and perhaps even discover phenotypes related to these variants that have not previously been described. For example, splicing defects could be a tissue-specific phenomenon, which could alter the presentation of particular genetic disorders. The interactive nature of the healthcare system allows researchers to recontact and assess patients for phenotypic features that may not be in their medical record through additional clinical evaluation, laboratory testing, or imaging studies. The sheer size of the DiscovEHR cohort which accounts for a large amount of rare variation allows for a more complete analysis of the phenotypic consequences of rare variants and the power of this resource increases as it continues to grow (Mirshahi et al., 2018).

In contrast to initiatives identifying variants across individuals sampled from a population, additional initiatives are taking a disease-centric approach to identify variants relating to a single disease. For example, the Simons Foundation of Autism Research Initiative (SFARI) was launched in 2003 to fund innovative research to understand the etiology of autism spectrum disorders (ASD). SFARI Simons Simplex Collection (SSC) has performed whole-exome sequencing on families with one ASD affected child, and unaffected parents and siblings (quad families) to identify inherited and de novo variants. In combination with genomic data, SSC has collected extensive phenotypic data (i.e., IQ, cognitive, developmental, behavioral, etc.) and biospecimens (i.e., blood samples/cell lines) for each participant. This wealth of data offers a unique advantage to researchers attempting to decipher the phenotypic and genetic heterogeneity that characterizes ASD. More specifically, we can leverage this data by identifying potentially deleterious variants and ASD risk gene through the use of MaPSy, validate splicing defects using the relevant biospecimens, and even analyze phenotypic attributes that may have arisen due to defective splicing.

5 | TECHNICAL CHALLENGES: DESIGNING AND UTILIZING COMPLEX LIBRARIES TO ASSAY SPLICING

The design of libraries for use in MaPSy assays is constrained by several technical challenges. Most notably, only mutations in exons of fewer than 100 nucleotides can be included as a consequence of

current limitations in oligonucleotide synthesis technology. As the median length of internal exons is approximately 130 nucleotides, more than half of all human exons are excluded from MaPSy splicing characterization. Advances in solid-phase oligonucleotide synthesis will continue to expand the window size for sequence design moving forward. Currently, exons greater than 100 nucleotides can be truncated to preserve 5′ and 3′ splice site signals and proximal regions, to approximate the effect of potential splice variants.

5.1 | MPRAs and intronic variants

An additional technical challenge with MaPSy, and MPRAs using minigene approaches, lies in the analysis of intronic variants. As introns are excised, identifying variants in introns are lost during splicing and cDNA generated from mutant and wild-type alleles become indistinguishable. Recently, Adamson et al. (2018) devised a barcoding strategy using an eight-nucleotide barcode that, after subcloning into a reporter plasmid, designated a particular variant at the end of a transcript. These extra steps can potentially limit library complexity and random octamers may themselves affect gene expression. To circumvent this issue in MaPSy, a one-step barcoding strategy has been developed to allow for the identification of the mutant and wild-type exons from the spliced product. In our method, a barcode was added to every intronic mutant species as a unique marker. Each oligonucleotide library species consisted of a mutant or wild-type intronic sequence and 26 nucleotides of the endogenous exon. The last six nucleotides of the endogenous exons were used to design each barcode (Figure 2a). All possible variants within the barcode window were submitted to the Spliceman prediction software, and the three variants least likely to disrupt splicing were selected as barcodes for each mutant species (Lim & Fairbrother, 2012). Therefore, each barcode consisted of unique point variants for intronic mutant identification, and each intronic mutant species was tested in triplicate using three unique barcodes.

To evaluate the effect of the barcodes on splicing, the counts for each intronic mutant barcoded triplicate in the unspliced input versus the spliced output were plotted to test for a correlation. Presumably if the representation of each barcoded species in the unspliced input and spliced output are similar, we can be confident the barcodes are likely not affecting splicing. Of the 208 triplicates which had at least 10 reads each, 175 (84%) of these were highly correlated ($r^2 > 0.9$), suggesting that the barcoding strategy was effective (Figure 2b,c). This result suggests the observed allelic imbalances in splicing are more likely to be caused by the mutant being tested, and not a result of the barcode. This new approach will allow for the expansion of analysis into intronic variants.

5.2 | Challenges with complex library amplifications

A related technical issue arises from difficulties in maintaining initial oligonucleotide library complexity during amplification. Library synthesis provides a highly complex pool of oligonucleotides, each at sub pmol quantities. To apply library contents to MPRAs, amplification through PCR is necessary to acquire experimentally tractable quantities of DNA. However, amplification may alter the overall composition of the library, changing both the overall content and the ratio of constituents possessed there within. Such changes can be the result of either PCR drift or PCR selection (Polz & Cavanaugh, 1998). PCR drift is a bias that is assumed to be the result of stochastic variation in early cycles of amplification, and is not reproducible in replicate PCR amplifications. Alternatively, PCR selection operates on mechanisms which inherently favor amplification of particular templates relative to others. Factors such as the GC content and relative structure of oligonucleotides may dictate their representation in an amplified library. PCR of complex libraries also holds the heightened potential to generate artifacts in the form of chimeras and heteroduplexes, which can quickly change the compositional landscape of a library (Qiu et al., 2001). To circumvent these issues, amplification can most effectively be achieved through multiple rounds of fewer amplification cycles (5–10) followed by size selected purification of PCR products. However, even with optimized protocols, population dynamics are observed to shift between rounds of amplification (Figure 3a) suggesting that PCR selection continues to restrict the maximum complexity that can be achieved in an applied library. Fortunately, within the context of paired oligonucleotides (wild-type and variant), both species seem to behave similarly during amplification (Figure 3b). Overall representation, including observed dropout, is typically conserved in final data sets between compared oligonucleotides, minimizing the number of unproductive reads. In moving forward, the use of low pass sequencing allows for optimization of amplification protocols that place an emphasis on

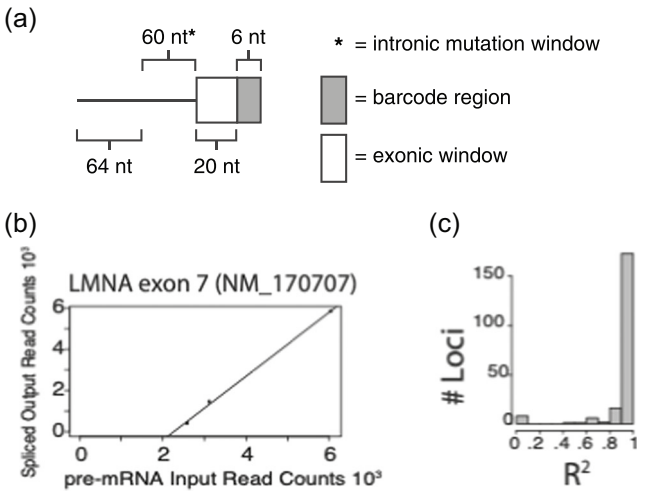


FIGURE 2 New MaPSy barcoding strategy(a) The schematic of the barcoding strategy shows that the last six nucleotides (nt) of the region containing endogenous sequence (150 nt) was used to design barcodes. Mutations tested by the assay fell within 60 nt upstream of the 3′ss. (b) Preliminary Data demonstrates correlation between three barcoded triplicates for LMNA exon 7. (c) Distribution of R squared values across barcoded triplicates of 208 other exons demonstrates selected barcodes do not alter splicing

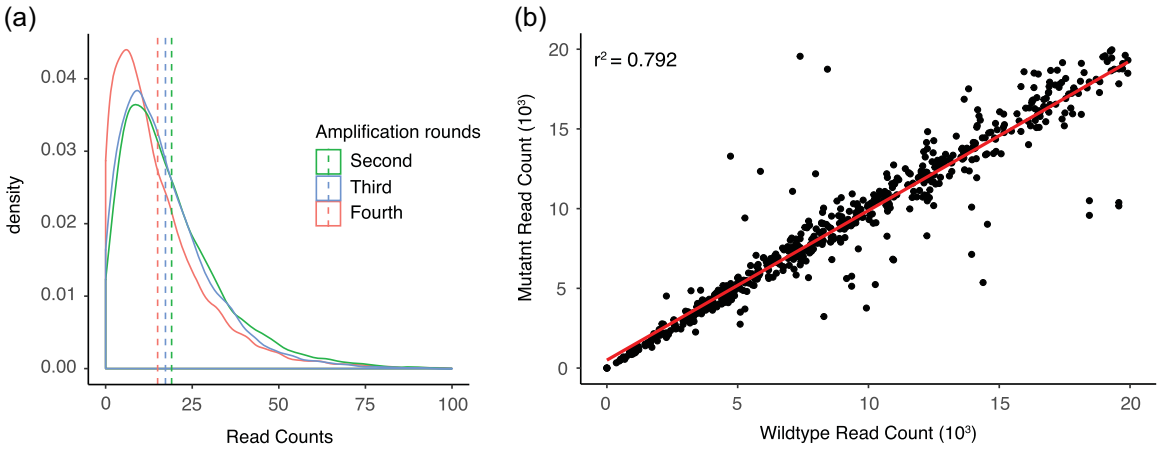


FIGURE 3 Amplification of complex oligonucleotide libraries (a) Density of oligonucleotide library substrate read counts between successive rounds of amplification (second, third, and fourth). Initial library contained 7,520 oligonucleotide species generated using Agilent solid-phase oligonucleotide synthesis technologies. Next generation sequencing performed using Illumina MiSeq. Dashed lines represent mean substrate read counts. (b) Relationship of paired mutant and wild-type oligonucleotide substrate read counts after initial amplification of an oligonucleotide library containing 1,504 substrates. Deep sequencing was performed using Illumina HiSeq. 3,000 (2 × 150)

retaining both complexity of oligonucleotide content and uniformity of representation there within.

5.3 | The effect of flanking sequence context on variant perturbation in MaPSy

Another limitation lies in the accuracy of MRPAs in representing physiological outcomes. Previous reports have suggested that surrounding sequence context is an important determinant in splicing outcome (Kim et al., 2017). Our original MaPSy tested ~5,000 disease-causing exonic variants in a three exon minigene where each variant was flanked by an upstream adenovirus exon and a downstream ACTN1 exon (Figure 1). To determine the potential contribution of sequence context to splicing outcome, we re-implemented MaPSy to assess potential splicing defects in 748 alleles caused by de novo variants reported in the SSC using three slightly different three exon minigene reporters. Instead of using the adenovirus upstream exon as described in (Soemedi et al., 2017b), three exons representing a range of 5′ss strengths, determined by MaxEntScan (Yeo & Burge, 2004), were synthesized into three separate in vivo minigene reporters and assayed in parallel. Each reporter contained either the VCP exon 15, EMC7 exon 3, or VCP exon 10, a 230-mer genomic fragment containing either the mutant or wild-type (reference) sequence, and a downstream ACTN4 exon (Figure S1). This resulted in each de novo variant being assayed in triplicate under three separate upstream exonic sequence contexts. Deep sequencing of input libraries and output-spliced fractions were used to determine the allelic ratio of mutant/wild-type pairs (M/W splice ratio) as described previously (cite Nat gen paper) (Figure 4a, Table S1). Despite the differences in the sensitivity of different reporter constructs, general agreements were observed between the relative allelic imbalances (i.e., M/W splice ratios) in all three assay runs (Figure 4b). Although sequence context does impose an effect on splicing outcome, as described previously (Kim et al., 2017), the validation rate of the original MaPSy (~83%; Soemedi et al., 2017b) and the general agreement between the variants allelic imbalance given the three new minigene constructs, suggests that the MaPSy assay offers a reliable method for prioritizing variants by their ability to affect splicing.

6 | FUTURE POTENTIAL

In summary, MPRAs show great promise for future efforts in precision medicine and drug discovery. Variants identified in consortia sequencing and integrated genetic data sets such as the Geisinger MyCode program, are well suited for MPRAs. MPRAs can help interpret incidental findings in clinical sequencing studies and guide clinical decisions. Moreover, the relatively low cost of deep sequencing has and will continue to produce large data sets containing novel variants. MPRAs are currently the ideal method for interpreting the functional consequences of novel variants, as they keep pace with discovery, and in the case of MaPSy, accurately assess a variant's effect on splicing with a ~83% validation rate (Soemedi et al., 2017b). In addition to the functional interpretation of variants, the data generated from MPRAs have also been used to train predictive models for the effects of novel variants on splicing (Cheng et al., 2019; Ke et al., 2011; Rosenberg et al., 2015; Wang et al., 2004), offering additional tools for variant interpretation. A recent analysis evaluated the predictive power of three splicing variant prediction programs (SPANR [Xiong et al., 2015], ESRseq scores [Ke et al., 2011], and Hexplorer [Erkelenz et al., 2014]) and found that the model trained on a minigene screening of all possible hexamers, ESRseq, was the most predictive in nature (Soukarieh et al., 2016). Even more recently, the predictive ability of MMSplice, the splicing prediction model trained on the single nucleotide variant MaPSy data and additional splicing MPRA implementations, was shown to outperform multiple splicing prediction programs (Cheng

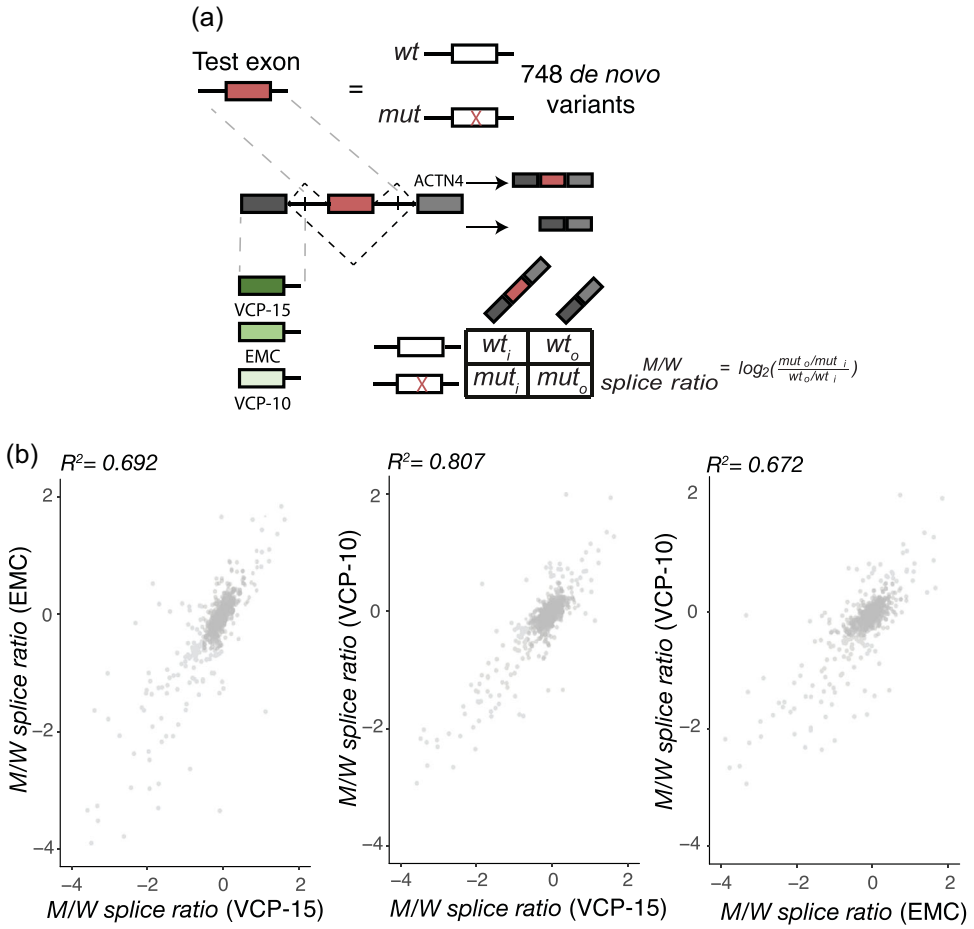


FIGURE 4 Context dependence in MaPSy (a) 748 de novo mutant exons and their corresponding wild-type counterparts were incorporated into three different three exon in vivo constructs. Both the unspliced input and spliced output library were deep sequenced to establish allelic imbalance between mutant and wild-type species. (b) Comparison of individual allelic ratios of variants in the reporter constructs

et al., 2019). These analyses further highlight the utility of MRPAs in not only assessing functionally a variant's effect on splicing but also in the construction of predictive models. In addition to MPRA-trained prediction models, a recent splicing model was trained on primary sequence alone and produced reliable splicing predictions (Jaganathan et al., 2019). The advantage of this model is that it can identify long-range sequence features that are not captured in the minigenes used in MPRAs. In time, it is likely that multiple algorithms will be combined when making classifications to improve the interpretation of variants. In general, the advantage of computational methods is their ability to assess more variants than can be assayed in a single MPRA and will help improve the novel variant classification problem facing clinical sequencing studies.

In addition to variant interpretation, MPRAs can identify the effects of drugs on splicing. Many drugs, such as amiloride, demonstrate widespread, but tolerable effects on splicing (Chang et al., 2011; Soemedi et al., 2017a). These drugs can be screened against a library of variants to determine their personalized effects on patients with rare diseases (Soemedi, Vega, Belmont, Ramachandran, & Fairbrother, 2014). For example, a drug may be found to exacerbate a splicing defect in a patient. The patient's physician could be informed of the adverse event, and a safer drug may be prescribed instead. Conversely, some splicing defects may be rescued by a drug. In this case, follow-up studies may be indicated, which might lead to the discovery of novel therapeutics for diseases that are too rare to justify the expense of other methods such as high-throughput screening.

Although there are some limitations to the scope and scale of MPRAs, viable strategies are being developed to circumvent them. Solid-phase oligonucleotide synthesis technology currently limits the length of the species to be tested in parallel to a few hundred base pairs. For splicing, this limits the number of full-length exons that can be tested to less than half of all human exons. This challenge can be addressed by designing chimeric exons that contain only one of the splice sites of larger exons. Intronic variants are also more challenging to test, because the fully-spliced species of the mutant and wild-type are degenerate. We have discussed barcoding methods that help identify degenerate species after splicing. The Vex-seq library design utilized one of these methods to test intronic variants (Adamson et al., 2018).

The winners of this CAGI challenge, who developed the MMSplice prediction software, can accurately predict the splicing outcomes of novel variants. The CAGI data set, we have generated

represents the importance and future promise of variant interpretation algorithms. Similar data sets are likely to be generated from future clinical sequencing studies. Platforms such as MMSplice will help classify novel variants from these studies. Such classifications will help both for returning incidental findings to patients, and for determining the safety and efficacy of drugs for patients with rare variants.

ACKNOWLEDGMENTS

The work was funded by NIH R01 GM127472.

The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650.

We would like to acknowledge Dr. Julien Gagneur and Jun Cheng for their text recommendations. We would also like to acknowledge CAGI for providing a platform to critically assess predictive models and their application to larger data sets.

ORCID

David T. Glidden http://orcid.org/0000-0002-2742-9531

REFERENCES

Abul-Husn, N. S., Cheng, X., Li, A. H., Xin, Y., Schurmann, C., Stevis, P., & Dewey, F. E. (2018). A protein-truncating HSD17B13 variant and protection from chronic liver disease. *New England Journal of Medicine*, *378*(12), 1096–1106. https://doi.org/10.1056/NEJMoa1712191

Adamson, S. I., Zhan, L., & Graveley, B. R. (2018). Vex-seq: High-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biology*, *19*(1), 71. https://doi.org/10.1186/s13059-018-1437-x

Baeza-Centurion, P., Minana, B., Schmiedel, J. M., Valcarcel, J., & Lehner, B. (2019). Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell*, *176*(3), 549–563. https://doi.org/10.1016/j.cell.2018.12.010. e523.

Bretschneider, H., Gandhi, S., Deshwar, A. G., Zuberi, K., & Frey, B. J. (2018). COSSMO: Predicting competitive alternative splice site selection using deep learning. *Bioinformatics*, *34*(13), i429–i437. https://doi.org/10.1093/bioinformatics/bty244

Canver, M. C., Lessard, S., Pinello, L., Wu, Y., Ilboudo, Y., Stern, E. N., & Orkin, S. H. (2017). Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nature Genetics*, *49*(4), 625–634. https://doi.org/10.1038/ng.3793

Carey, D. J., Fetterolf, S. N., Davis, F. D., Faucett, W. A., Kirchner, H. L., Mirshahi, U., & Ledbetter, D. H. (2016). The Geisinger MyCode community health initiative: An electronic health record-linked biobank for precision medicine research. *Genetics in Medicine*, *18*(9), 906–913. https://doi.org/10.1038/gim.2015.187

Chang, J. G., Yang, D. M., Chang, W. H., Chow, L. P., Chan, W. L., Lin, H. H., & Yang, W. K. (2011). Small molecule amiloride modulates oncogenic RNA alternative splicing to devitalize human cancer cells. *PLoS One*, *6*(6), e18643. https://doi.org/10.1371/journal.pone.0018643

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Celik, M. H., Fairbrother, W. G., Avsec, Z., & Gagneur, J. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, *20*(1), 48. https://doi.org/10.1186/s13059-019-1653-z

Consortium, G. T., Laboratory, D. A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G. g, Fund, N. I. H. C., & Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. https://doi.org/10.1038/nature24277

Cooper, T. A. (2005). Use of minigene systems to dissect alternative splicing elements. *Methods*, *37*(4), 331–340. https://doi.org/10.1016/j.ymeth.2005.07.015

Desmet, F. O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M., & Beroud, C. (2009). Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*, *37*(9), e67–e67. https://doi.org/10.1093/nar/gkp215

Dewey, F. E., Murray, M. F., Overton, J. D., Habegger, L., Leader, J. B., Fetterolf, S. N., & Carey, D. J. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*, *354*(6319), aaf6814. https://doi.org/10.1126/science.aaf6814

Erkelenz, S., Theiss, S., Otte, M., Widera, M., Peter, J. O., & Schaal, H. (2014). Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Research*, *42*(16), 10681–10697. https://doi.org/10.1093/nar/gku736

Fairbrother, W. G., Yeh, R. F., Sharp, P. A., & Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. *Science*, *297*(5583), 1007–1013. https://doi.org/10.1126/science.1073774

Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C., & Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, *513*(7516), 120–123. https://doi.org/10.1038/nature13695

Gusarova, V., O'Dushlaine, C., Teslovich, T. M., Benotti, P. N., Mirshahi, T., Gottesman, O., & Gromada, J. (2018). Genetic inactivation of ANGPTL4 improves glucose homeostasis and is associated with reduced risk of diabetes. *Nature Communications*, *9*(1), 2252. https://doi.org/10.1038/s41467-018-04611-z

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., & Farh, K. K. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, *176*(3), 535–548. https://doi.org/10.1016/j.cell.2018.12.015. e524

Ke, S., Anquetil, V., Zamalloa, J. R., Maity, A., Yang, A., Arias, M. A., & Chasin, L. A. (2018). Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Research*, *28*(1), 11–24. https://doi.org/10.1101/gr.219683.116

Ke, S., Shang, S., Kalachikov, S. M., Morozova, I., Yu, L., Russo, J. J., & Chasin, L. A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Research*, *21*(8), 1360–1374. https://doi.org/10.1101/gr.119628.110

Kim, S. W., Taggart, A. J., Heintzelman, C., Cygan, K. J., Hull, C. G., Wang, J., & Fairbrother, W. G. (2017). Widespread intra-dependencies in the removal of introns from human transcripts. *Nucleic Acids Research*, *45*(16), 9503–9513. https://doi.org/10.1093/nar/gkx661

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., & Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Lim, K. H., & Fairbrother, W. G. (2012). Spliceman--a computational web server that predicts sequence variations in pre-mRNA splicing. *Bioinformatics*, *28*(7), 1031–1032. https://doi.org/10.1093/bioinformatics/bts074

Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J., & Fairbrother, W. G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11093–11098. https://doi.org/10.1073/pnas.1101135108

Mirshahi, U. L., Luo, J. Z., Manickam, K., Wardeh, A. H., Mirshahi, T., Murray, M. F., & Carey, D. J. (2018). Trajectory of exonic variant discovery in a large clinical population: Implications for variant curation. *Genetics in Medicine*, *21*, 1417–1424. https://doi.org/10.1038/s41436-018-0353-5

Mort, M., Sterne-Weiler, T., Li, B., Ball, E. V., Cooper, D. N., Radivojac, P., & Mooney, S. D. (2014). MutPred Splice: Machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biology*, *15*(1), R19. https://doi.org/10.1186/gb-2014-15-1-r19

Park, E., Pan, Z., Zhang, Z., Lin, L., & Xing, Y. (2018). The expanding landscape of alternative splicing variation in human populations. *American Journal of Human Genetics*, *102*(1), 11–26. https://doi.org/10.1016/j.ajhg.2017.11.002

Polz, M. F., & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, *64*(10), 3724–3730

Qiu, X., Wu, L., Huang, H., McDonel, P. E., Palumbo, A. V., Tiedje, J. M., & Zhou, J. (2001). Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Applied and Environmental Microbiology*, *67*(2), 880–887. https://doi.org/10.1128/AEM.67.2.880-887.2001

Rosenberg, A. B., Patwardhan, R. P., Shendure, J., & Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, *163*(3), 698–711. https://doi.org/10.1016/j.cell.2015.09.054

Sanjana, N. E. (2017). Genome-scale CRISPR pooled screens. *Analytical Biochemistry*, *532*, 95–99. https://doi.org/10.1016/j.ab.2016.05.014

Singh, G., & Cooper, T. A. (2006). Minigene reporter for identification and analysis of cis elements and trans factors affecting pre-mRNA splicing. *Biotechniques*, *41*(2), 177–181. https://doi.org/10.2144/000112208

Soemedi, R., Cygan, K. J., Rhine, C. L., Glidden, D. T., Taggart, A. J., Lin, C. L., & Fairbrother, W. G. (2017a). The effects of structure on pre-mRNA processing and stability. *Methods*, *125*, 36–44. https://doi.org/10.1016/j.ymeth.2017.06.001

Soemedi, R., Cygan, K. J., Rhine, C. L., Wang, J., Bulacan, C., Yang, J., & Fairbrother, W. G. (2017b). Pathogenic variants that alter protein code often disrupt splicing. *Nature Genetics*, *49*(6), 848–855. https://doi.org/10.1038/ng.3837

Soemedi, R., Vega, H., Belmont, J. M., Ramachandran, S., & Fairbrother, W. G. (2014). Genetic variation and RNA binding proteins: Tools and techniques to detect functional polymorphisms. *Advances in Experimental Medicine and Biology*, *825*, 227–266. https://doi.org/10.1007/978-1-4939-1221-6_7

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frebourg, T., & Martins, A. (2016). Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools. *PLoS Genetics*, *12*(1), e1005756. https://doi.org/10.1371/journal.pgen.1005756

Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S., & Cooper, D. N. (2009). The Human Gene Mutation Database: 2008 update. *Genome Medicine*, *1*(1), 13. https://doi.org/10.1186/gm13

Telenti, A., Pierce, L. C., Biggs, W. H., di Iulio, J., Wong, E. H., Fabani, M. M., & Venter, J. C. (2016). Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(42), 11901–11906. https://doi.org/10.1073/pnas.1613365113

Verma, A., Bang, L., Miller, J. E., Zhang, Y., Lee, M. T. M., Zhang, Y., & Discov, E. H. R. C. (2019). Human-disease phenotype map derived from PheWAS across 38,682 individuals. *American Journal of Human Genetics*, *104*(1), 55–64. https://doi.org/10.1016/j.ajhg.2018.11.006

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., & Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. *Cell*, *119*(6), 831–845. https://doi.org/10.1016/j.cell.2004.11.010

Wong, M. S., Kinney, J. B., & Krainer, A. R. (2018). Quantitative activity profile and context dependence of all human 5′ splice sites. *Molecular Cell*, *71*(6), 1012–1026. https://doi.org/10.1016/j.molcel.2018.07.033. e1013

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., & Frey, B. J. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, *347*(6218), 1254806–1254806. https://doi.org/10.1126/science.1254806

Yeo, G., & Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, *11*(2-3), 377–394. https://doi.org/10.1089/1066527041410418

Yeo, G., Holste, D., Kreiman, G., & Burge, C. B. (2004). Variation in alternative splicing across human tissues. *Genome Biology*, *5*(10), R74. https://doi.org/10.1186/gb-2004-5-10-r74

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Rhine CL, Neil C, Glidden DT, et al. Future directions for high-throughput splicing assays in precision medicine. *Human Mutation*. 2019;40:1225–1234. https://doi.org/10.1002/humu.23866

Oligo Pools as an Affordable Source of Synthetic DNA for Cost-Effective Library Construction in Protein- and Metabolic Pathway Engineering

Bastiaan P. Kuiper,[a] Rianne C. Prins,[a] and Sonja Billerbeck*[a]

The construction of custom libraries is critical for rational protein engineering and directed evolution. Array-synthesized oligo pools of thousands of user-defined sequences (up to ~350 bases in length) have emerged as a low-cost commercially available source of DNA. These pools cost ≤10% (depending on error rate and length) of other commercial sources of custom DNA, and this significant cost difference can determine whether an enzyme engineering project can be realized on a given research budget. However, while being cheap, oligo pools do suffer from a low concentration of individual oligos and relatively high error rates. Several powerful techniques that specifically make use of oligo pools have been developed and proven valuable or even essential for next-generation protein and pathway engineering strategies, such as sequence-function mapping, enzyme minimization, or *de-novo* design. Here we consolidate the knowledge on these techniques and their applications to facilitate the use of oligo pools within the protein engineering community.

1. Introduction

Our capacity to enhance and change the function of proteins or design entirely new ones is essential to unlocking their potential for medicine, green catalysis, and the food and textile industry.

The field of protein engineering has rapidly evolved, and while early techniques are mostly built upon random mutagenesis, currently more targeted approaches are available. The rise of structural data, generation of detailed sequence-function maps[1] and computational tools,[2–5] combined with next-generation sequencing technologies,[6] enable scientists to generate smart libraries for directed evolution[7–9] and to perform rational engineering or to design new proteins from scratch.[4,5]

For example, in deep-mutational scanning (DMS) all residues of a protein are saturation-mutagenized and then functionally characterized, allowing a protein engineer to create detailed sequence-function maps of a protein.[1] These fitness maps can identify structurally or functionally important residues that can then infuse further library design. In the context of enzyme engineering, information-rich sequence-function maps obtained from such methods allowed researchers to probe the relationship between enzyme fitness and solubility,[10] folding[11] and heterologous expression levels,[12] it has been used to identify sequence determinants of enzymatic substrate specificity,[13] and it has been used to create smart libraries for directed evolution.[9] A collection of available DMS data sets is currently consolidated here: https://www.mavedb.org.[14]

Further, computational tools such as artificial intelligence (AI)[8] and *de novo* protein design strategies[5,15] currently revolutionize the way we do protein science. For example, AI has been used to guide and accelerate the pace of directed evolution[2] and has recently been used to predict from mostly database-available sequences which combination of mutations likely yields a functionally optimized protein or enzyme (in respect to a specific user-set function).[3]

While these developments will continue to expand the functionality of proteins in general and enzymes specifically, they also demand the creation of massive variant libraries. For instance, to create a DMS library of a 300 amino acid protein, 5701 gene variants (the wildtype sequence plus 19 amino acid exchanges for all 300 positions) need to be generated. For the functional optimization of a protein or the *de novo* design of a protein – even when using state-of-the-art AI algorithms or the best *de novo* protein design protocols to reduce library size – between ten and a few thousand full genes need to be synthesized and tested to find the desired function.[3,4] While massive parallel sequencing of DNA has become very affordable over time, the synthesis of DNA is however still prohibitively expensive when considering the scale needed to build complex libraries.[16] As such, DNA synthesis becomes a new bottleneck for next-generation protein engineering.

The currently cheapest available source of synthetic DNA is micro-array-synthesized oligonucleotides, commercially available as 'oligo pools'.[16] Oligo pools are mixes of thousands of individually designed polynucleotides of up to 350 bases in length. Traditionally, oligonucleotides have been synthesized by

[a] *B. P. Kuiper, R. C. Prins, S. Billerbeck*
Molecular Microbiology
Groningen Biomolecular Sciences and Biotechnology Institute
University of Groningen
Groningen (The Netherlands)
E-mail: s.k.billerbeck@rug.nl

This article is part of a Special Collection dedicated to the NextGenBiocat 2021 virtual symposium. To view the complete collection, visit our homepage.

solid-phase phosphoramidite chemistry.[17] This column-based synthesis generates up to 200 mers with error rates of 1 in 200, yields of 10 to 100 nmol per product for a cost of 0.05–1 USD per base dependent on the length and concentration yield.[16,18,19] These individually synthesized oligonucleotides are then routinely further used for the synthesis of gene-length DNA fragments using different PCR-based methods.[20,21] To increase throughput and decrease the cost of oligonucleotide synthesis, several technologies have been developed over the last three decades to synthesize oligonucleotides in spatially decoupled microarrays,[22–25] lowering costs by several orders of magnitude (0.00001 to 0.001 USD per base).[16] Microarray-based technologies allow synthesizing thousands of individual user-defined sequences, eventually delivered as a pool of molecules.

Even though these microarray-based oligo pools are cheap, there are several challenges in using them for gene synthesis and library creation.[16] First, while the number of individual user-defined oligo sequences in a pool is large, their individual concentration is quite low. This challenges their use in traditional gene synthesis protocols, which still mostly rely on oligos from column-based synthesis. Second, the longer the oligos the higher the percentage of truncated molecules, further lowering the expected concentration of full-length molecules. Third, the error rates for oligo pools are usually higher than those for column-synthesized oligos.

Noteworthy, a recently published oligo pool purification method (multiplex oligonucleotide library purification by synthesis and selection (MOPSS))[26] that distinguishes full-length oligos from oligos with insertions and deletions, will partly overcome this issue if a user is willing to add an extra purification step before library creation.

Early on, oligo pools have been adopted for the creation of guide RNA libraries for functional genomics studies,[27,28] for barcoding within screening platforms,[29] and the creation of libraries of short regulatory elements such as promoters,[30–32] enhancers[33] or silencers.[34] These applications require only short DNA (20 to 100 bases) stretches and can thus manage the shortcomings of oligo pools in limited sequence length, high error rate, and incomplete synthesis.

Libraries for enzyme engineering ideally require the synthesis of error-free DNA. Each off-target mutation increases the library size and consequently increases the number of variants that need to be functionally tested to reach full library coverage.

To still leverage oligo pools for library creation – despite their low yield, their short length, and high synthesis errors – several powerful techniques have been developed over the last five years. These techniques managed to use oligo pools for creating DMS libraries,[35–37] insertion libraries, or for direct in-lab assembly of gene fragments,[38,39] full genes[39] and pathways[40] and, as such, have and will make many next-generation protein sciences approaches economically feasible for many laboratories.

Here we give an overview of these methods to consolidate the knowledge for the protein and enzyme engineering community (Figure 1A).

2. Changing Residues: Oligo Pool-Based Multiple Site-Saturation Mutagenesis and Deep Mutational Scanning Libraries

Site-directed mutagenesis has been foundational to protein and enzyme engineering. While commercially available methods like QuikChange™ have long been available to create targeted libraries that mutate single and multiple residues, next-generation protein engineering technologies require massive parallel changes to a protein sequence. Towards this end, a handful of scalable mutagenesis methods that achieve (near) comprehensive mutagenesis of open reading frames have been developed: PFunkel,[48] Nicking Mutagenesis (NM)[35,36] and programmed allelic series (PALS)[37] are *in vitro* methods and plasmid recombineering (PR)[41] and CRISPR-enabled trackable

Bastiaan Kuiper received his BSc and MSc degree in Biomolecular Sciences from the University of Groningen in 2021 where he worked on his research project in the molecular genetics group. He conducted a research internship at the Fukushima Medical University in the department of cell science, Japan. His research interests include structural biology, protein-protein interactions, and fluorescent spectroscopy.

Rianne C. Prins obtained her MSc in molecular cell biology from the University of Groningen in 2018 and is currently working on her PhD project at the Molecular Microbiology group at the Groningen Biomolecular Sciences and Biotechnology Institute, Faculty of Science and Engineering in Groningen, the Netherlands. Her PhD research is focused on sequence-function relationships and protein engineering of yeast killer toxins.

Sonja Billerbeck holds a Master in Microbiology and Biochemistry from the University of Tübingen and the Max Planck Institute for Developmental Biology, and a PhD in Bioengineering from ETH Zürich. After postdoctoral work in yeast Synthetic Biology at Columbia University in New York City, she joined the University of Groningen as an Assistant Professor in 2019. Her laboratory uses a combination of synthetic biology, genome engineering, protein engineering, and environmental microbiology to access, understand and engineer the functional diversity of nature's yeast-based mycobiome for applications in human health, industrial biotechnology and to answer fundamental questions on yeast (pathogen) biology.

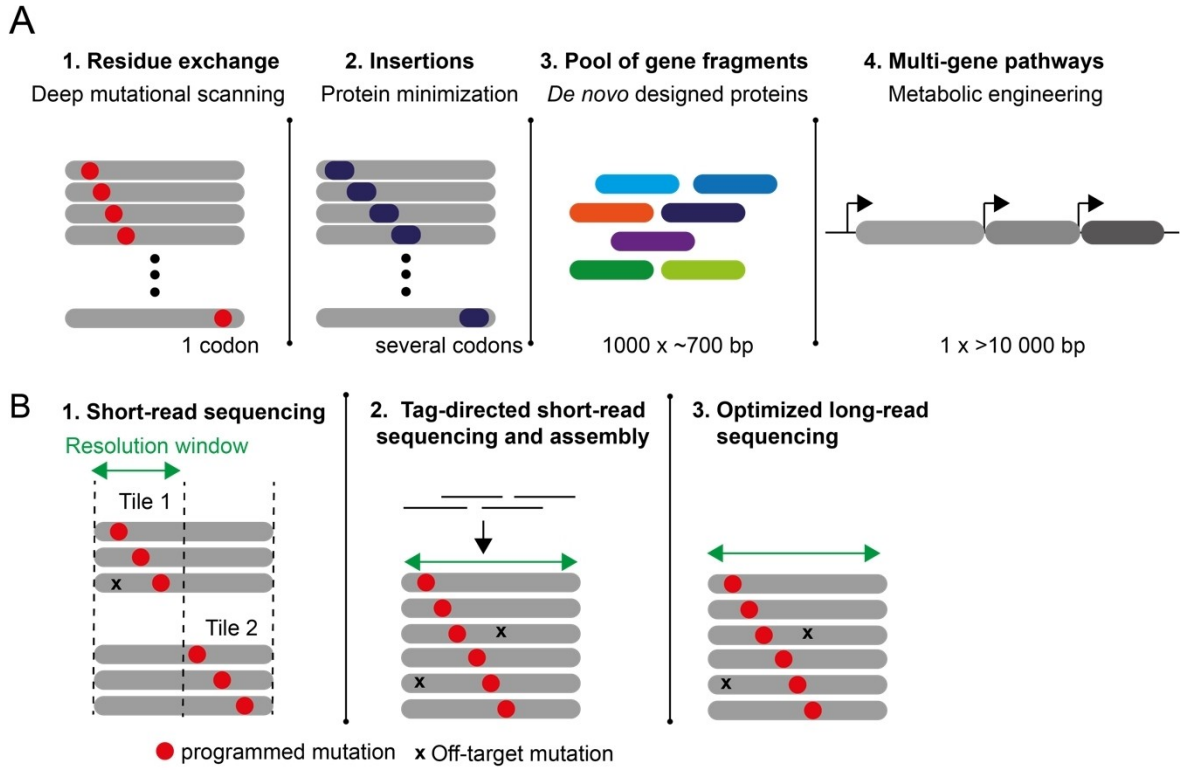


Figure 1. Overview of the discussed methods for mutagenic library creation, gene- and pathway assembly, and library analysis. A) Oligo pool-based methods to generate protein libraries, gene fragment libraries, or pathways. 1) Single residue saturation mutagenic libraries, such as those used for deep-mutational scanning, can be created from oligo pools by Nicking Mutagenesis (NM),[35,36] programmed allelic series (PALS),[37] plasmid recombineering (PR)[41] or CRISPR-enabled trackable Genome engineering (CREATE).[42] 2) Libraries where short DNA stretches are comprehensively inserted, such as required for protein minimization, have been created by PR and can likely be created by NM, PALS, and CREATE. Single-residue deletion libraries have been created by PALS. 3) Complex pools of many different proteins fragments can be created by multiple pairwise assembly[38] and DropSynth.[39,43] 4) Larger genes and pathways can be assembled as shown by Wan *et al.*[40] B) Library sequence analysis: The quality of libraries needs to be analyzed for the frequency of programmed mutations (red dot) and the frequency of off-target mutations (cross). 1) Short-read sequencing – as offered via Illumina-based services – is highly accurate, high in throughput, and widely accessible but the short read-length leads to a narrow resolution window. Libraries need to be tiled to be accurately quality controlled, as off-target mutations outside the resolution window are otherwise invisible.[6] 2) Molecular barcoding and computational assembly can overcome the limited read length in NGS as full-length sequences of mutagenized clones can be obtained from short NGS reads.[44] 3) Single-molecule real-time (SMRT) long-read sequencing, followed by computational error correction[45] or combined with variant-concatenation[46,47] starts to allow long-read sequencing for library quality control in protein engineering.

Genome engineering (CREATE)[42] are *in vivo* methods for use in *E. coli* and the yeast *S. cerevisiae*.

In this review, we highlight those methods (NM, PALS, PR and CREATE), that have been tested to be compatible with the use of array-synthesized oligo pools rather than individually column-synthesized and hand-mixed oligonucleotide pools. Thus, these methods allow for the cost-effective creation of custom libraries. Noteworthy, even though not explicitly tested in a published format, likely, PFunkel[48] would also work with array-synthesized oligo pools as it is the precursor technique for the development of NM.

2.1. *In vitro* techniques for saturation mutagenesis

Nicking mutagenesis (NM) allows for the one-pot generation of targeted or comprehensive single-site or multi-site saturation mutagenesis libraries in a single day, using regular molecular biology techniques: It requires routinely prepped plasmid DNA, a pool of oligonucleotides, two nicking restriction endonucleases (Nt.BbvCI and Nb.BbvCI), an exonuclease (Exo III), a high-fidelity DNA polymerase (Phusion) and a ligase (Taq). The only non-routine requirement is that the plasmid DNA needs to encode the 7 bp BbvCI recognition site, which can be introduced into any plasmid by regular cloning techniques.

The workflow of NM starts by selectively degrading one plasmid DNA strand via the nicking restriction endonuclease Nt.BbvCI and exonuclease treatment. The obtained circular single-stranded DNA subsequently serves as the template for primers that encode for the desired mutations. The ratio of primer to template allows tuning the number of mutations per gene. If the number of primer molecules is much lower than the number of template molecules, then effectively only one primer can bind per template. Each primer is then extended by a high-fidelity polymerase and the created strand is ligated with Taq DNA ligase resulting in covalently closed circular double-stranded DNA. One strand contains the mutations, the other the wild-type sequence. Subsequent treatment with a second

nicking restriction endonuclease (Nb.BbvCI, binds and nicks the reverse complementary Nt.BbvCI recognition site) and exonuclease leads to digestion of the still-wild-type strand of the plasmid DNA, yielding a single-stranded mutated plasmid. Amplification with a universal primer that binds outside the target area for mutagenesis eventually creates a double-stranded plasmid with the manifested mutations in both strands.

In the original paper describing NM,[35] the authors used a hand-made mix of column-synthesized oligonucleotides to establish the method. Later, it became clear that very little primer was needed for the introduction of point mutations and that primers encoded as micro-array synthesized oligo pools could be used.[36] The authors test the oligo-pool based NM with three short protein segments of <100 residues (short enough to enable full sequencing via paired-end reads, see section 5) and show that dependent on the oligo-concentration, 97.4 to 100% mutational coverage could be reached after a single round of NM (Table 1). Noteworthy, all 7118 oligos required for mutagenizing all three proteins were ordered together as one single pool. They further show that the use of oligo pools leads to a more even representation of each amino acid when compared to using NNN codon randomized and hand-mixed column-synthesized primers.

Full sequencing analysis of the three test libraries showed that they consisted of 14–37% plasmids with the designed one non-synonymous mutation, 52–71% plasmids with wild-type sequence (likely due to template molecules that did not participate in or complete the mutagenesis cycle), and 11–15% plasmids with more than one non-synonymous mutation (Table 1). The authors observed that a higher melting temperature of a given primer correlated with higher mutational frequency, giving room for optimization of the protocol. The authors further tested if NM could be used for subsequent multi-site mutagenesis by subjecting the single-site saturation library to a second cycle of NM. After this second cycle performed with one of the test proteins, the authors show coverage of 79% of all possible positions. Double mutations were depleted in near adjacent positions, likely because a second oligonucleotide would either not anneal properly or overwrite the mutation from the first oligonucleotide. An optimized protocol for performing multi-site NM has recently been developed.[49] This protocol uses model-driven oligo design to achieve $>99\%$ coverage of a multi-site mutagenized 32.768 membered antibody library (Table 2). The authors further provide an automated workflow for oligo pool design using a protein-coding sequence as the sole input.

Scalability of the single-site NM method[36] was later shown by saturation mutagenizing the bacteriophage ΦX174,[50] specifically its two main proteins the F capsid protein (421 amino acids scanned) and G spike protein (172 amino acids scanned). The libraries possessed greater than 99% of all 11.860 programmed mutations (Table 1). All required oligos could be ordered in one pool. To enable the replication of the plasmids in *E. coli*, the genome was divided into 14 non-toxic and non-replicative fragments. The F capsid protein and G spike protein were each tiled into two fragments. Golden Gate cloning was then used to assemble the complete ΦX174 mutant genome

Table 1. Performance overview of single-site saturation mutagenesis methods (e.g. used for deep mutational scanning).

Method	Protein (# of mutated codons)	% Coverage[a]	% 1 NSM[b]	% 0 NSM (wild-type)[c]	% >1 NSM (off-target)[d]	Ref.
in vitro						
Nicking mutagenesis (NM)	*E. coli* AmiE (70)[e]	100	36.4	52.7	10.8	[36]
	A. thaliana PYR1 (86)[e]	100	26.5	59.4	13.6	[36]
	Anti-influenza human antibody variable heavy gene UCA9 (99)[e]	97.4	14.1	71.3	15.0	[36]
	Phage ΦX174 F capsid protein (421)	100	41.8	28.6	29.5	[50]
	Phage ΦX174 G spike protein (172)	99.9	49.3	29.3	21.4	[50]
PALS	*S. cerevisiae* Gal-DBD (64)	99.9	47	24	21	[37]
	Human p53 (393)	93.4	33	30	35	[37]
in vivo						
Plasmid recombineering (PR)[f]	*A. thaliana*-derived iLOV (110)	99.8	28.8	60.7	10.5	[41]
CREATE (genomic mutagenesis)	*E. coli* GalK (1)	100	56.8[g]	22.4[g]	n/a	[42]
	S. cerevisiae ADE2 (2)	100	95	5	n/a	[42]
	E. coli lysine metabolism, 19 genes, 16,300 designed edits[h]	22.7 to 61.6	n/a	n/a	n/a	[54]

n/a: not available; [a] Number of actually observed mutations per 100 designed mutations. Note: differences in sequencing depth used for quality control in the different studies influences the number of observed mutations, thus the apparent coverage. [b] Percent of mutants that carry exactly one desired non-synonymous mutation (NSM). [c] Percent of mutants that do not carry any NSM, thus being wild-type (non-edited) variants. Note that CREATE is the only method that counter-selects for wild-type via CRISPR-Cas9 mediated double-strand breaks. [d] Percent of mutants that carry more than one NSM, e.g. off-target mutations. [e] The average of two reported independent runs of NM is given (Table 1 in Ref. [36]). [f] Here, hand-mixed pools of column-synthesized oligonucleotides were used. It was shown later that PR also works with array synthesized oligo pools.[53] [g] Calculated based on 80% of clones being edited (based on colorimetric screen) and 71% of those 80% being correctly edited (based on sequencing, 0.8×0.71); and 20% of clones being wild-type (based on colorimetric screen) plus 3% of the 80% phenotypically edited clones (0.8×0.03) being still unedited at the programmed locus. [h] CREATE was developed and applied for multiplexed pathway mutagenesis. The percent coverage refers to the observed coverage range of five test loci that were analyzed in-depth (Table 1 in Ref. [54]).

Table 2. Performance overview of double- and multi-site saturation mutagenesis methods (e.g. for targeted protein engineering).

Method	Library size	% Coverage	% 1 NSM	% 0 NSM (wild type)	% 2 NSM	% > 2 NSM	Ref.
Double-site mutagenesis							
Nicking mutagenesis (NM)	n/a	79.2	n/a	60.0	n/a	n/a	[36]
Plasmid recombineering	5940	98.0	26.3	32.6	24.5	16.4	[41]
Multi-site mutagenesis							
Optimized multi-site nicking mutagenesis (NM)[a]	16384	99.9	n/a	2.58	n/a	n/a	[49]
	32768	99.4	n/a	0.25	n/a	n/a	[49]

n/a: not available; [a] The average of the two reported independent runs of optimized multi-site NM is given (Table 1 in Ref. [49]).

and generate libraries of infective viruses that could be used in the future to study viral evolution and to engineer bacteriophages for therapeutic applications.

A second method that achieves saturation mutagenesis in one volume is PALS (programmed allelic series).[37] Similar to NM, PALS allows for the generation of targeted or (near)-comprehensive single-site or multi-site saturation mutagenesis libraries using regular molecular biology techniques and no special equipment. It has a few more steps than NM, takes rather two instead of one day to prepare the library and it leads overall to slightly lower coverage than NM (as such, libraries are called near-comprehensive, reported 98% coverage after a single round, Table 1). Instead of amplifying the full plasmid that encodes the target region for mutagenesis as done in NM, PALS only uses a PCR amplified target fragment as input. Therefore, PALS might become the method of choice in case the used plasmids are large and full amplification becomes inconvenient (e.g. as usually the case for mammalian vectors) or in case the mutagenic library is supposed to be encoded in the genome of an organism. In that case, the PALS PCR library can directly be used as a homology-directed repair fragment for genome engineering.

PALS is based on the annealing and extension of mutagenic oligos along a deoxyuracil (dU)- marked wild-type sense strand. The dU-marking allows the experimenter to specifically digest the wild-type sense strand after the extension is completed using uracil-DNA-glycosylase and exonuclease VIII (also known as USER enzyme mix). Each mutagenic oligo encodes for a unique 3'-priming site and a common 5'-priming site which are subsequently used to amplify the extended oligos by PCR creating a single-mutagenized double-stranded library (the unique priming sites get cleaved). At this point each library member has a different 3'end, depending on where the mutagenic oligo annealed. To create the full template length, the 5'–3' strand of the mutagenized amplicons are used as mega primers to extend their 3' ends along a dU-marked wild-type anti-sense strand.

The required dU-marked (single-stranded) sense and anti-sense strands can be created via PCR using dUs instead of dTs as nucleotides and by using a phosphorylated reverse or forward primer. The phosphorylation of the 5'end is harnessed to digest the antisense or sense strand via lambda exonuclease, leaving a dU-marked (single-stranded) sense and antisense strand.

After one round of PALS on two test proteins (the DNA-binding domain of the yeast transcription factor Gal4, and the full human transcription factor p53) the authors report 93–98% coverage, with 33–47% of the molecules showing the correct single mutations, 24–30% of the molecules being wild type and 21–35% of molecules having additional mutations on top of the programmed single mutation (Table 1).

2.2. *In vivo* techniques for saturation mutagenesis

Here we highlight two techniques, PR[41] and CREATE,[42] that achieve near-comprehensive saturation mutagenesis directly in the living microbial host *E. coli* (PR) and both, *E. coli* and *S. cerevisiae* (CREATE). *In vivo* techniques are useful, as the libraries can be directly functionally screened without the need for a transformation step.

PR relies on a method called "recombineering", which, historically has been widely used for genomic engineering in *E. coli*,[51] for example in a process called multiplexed automated genome engineering (MAGE) that mutagenized the genome via oligo recombineering for functional genomics and metabolic engineering.[52] Higgins and coworkers demonstrated that recombineering can also be used to mutagenize regions of plasmids, e.g. regions encoding for proteins.

PR is based on incorporating synthetic oligonucleotides directly into a gene of interest via the help of lambda phage protein ß-mediated recombination.[51,52] Mechanistically this (likely) involves that the ß protein binds to an oligonucleotide and directs it to the lagging strand at the replication fork of replicating DNA. The oligo is subsequently incorporated into the growing strand, thus allowing the experimenter to edit the new DNA molecule in a programmable fashion.

The procedure is cheap, as it does not involve enzymes or specialized equipment, and is simple, just involving a transformation and overnight growth selection for plasmid transformants: In brief, electro-competent *E. coli* cells are co-transformed with the plasmid of interest and a pool of oligos with user-defined sequences (60 bp, targeting the lagging strand and encoding the mutagenic codon in its middle). After the

electroporation cells are recovered for several hours without antibiotic selection, followed by overnight growth under plasmid-selective conditions. Plasmids can then be miniprepped to recover the library or one can directly continue with functional selection for the desired protein phenotype. Higgins *et al.* show that a nearly comprehensive mutational-scanning library can be achieved in a single transformation. They found 99.8% of all possible single amino acid conversions represented, for their 110-residue model protein iLOV (Table 1). Similar to the *in vitro* methods, no selection against wild-type plasmid is included in the method and they find that about 29% of their plasmids have a single mutation while over 61% of all plasmids are the wild type (Table 1). Five subsequent rounds of PR lowered the frequency of finding wild-type plasmids to 26%, while 33% of colonies had a single mutation and 41% of plasmids had two or more mutations. As such, the authors also show that multiple rounds of PR (five rounds in the reported case) can be used to create targeted double-site saturation libraries with high coverage (98%, Table 2).

In their first proof-of-concept of PR, Higgins *et al.* use a mix of 110 column-synthesized oligonucleotides rather than array-synthesized oligo pools, but in a later study, they show that amplified oligo pools can be used for PR mutagenesis.[53]

As a technical remark, PR should theoretically work with any template that replicates in *E. coli*, but it needs to be performed in a suitable strain background that encodes and expresses the Lambda Red system (including the ß protein) and carries a deletion for MutS. The MutS deletion is required to inactivate the mismatch repair system in *E. coli* and allow for effective recombination.[52] The *E. coli* strain EcNR2 which is widely used for recombineering is available via Addgene (#26931).[52]

A second method that uses array-synthesized oligo pools as a starting point for *in vivo* saturation mutagenesis is CREATE.[42] In comparison to PR, which acts upon plasmid-encoded targets, CREATE is designed to act upon chromosomally encoded proteins and full pathways and allows for pathway- or even genome-wide deep mutational analysis. Mutagenesis is based on inducing a genomic double-strand break in a protein of interest via CRISPR/Cas9 and allowing the cell to repair the break via homologous recombination with a short homology arm (up to 120 bp) that encodes the programmed mutation and a synonymous mutation in the protospacer adjacent motif (PAM) to prevent future cleavage. Both, the guide RNA (gRNA), which programs the cut site, and the homology arm are encoded together on a plasmid as a so-called CREATE cassette. The CREATE cassette library (library of all gRNAs and repair arms, typically 10^4 to 10^6 members) is thereby designed computationally and ordered as an array-synthesized oligo pool, amplified by PCR, and cloned into a cassette vector. The covalent linkage of the gRNA with the programmed edit on the homology arm allows the experimenter to use the CREATE cassette as a plasmid-encoded barcode during future selections, assuming that frequency of changes in a CREATE plasmid and the edited genome frequency stay coupled during growth.

The authors first test CREATE by inactivating the *galK* gene of *E. coli* by introducing a single nonsense mutation. Depending on the length of the used homology arm (80 to 120 bp) and the distance of the mutagenized site from the PAM (17 to 59 bp), CREATE achieves editing efficiencies of 75 to 90%, as judged by a *galK*-based colorimetric assay. Sequencing revealed that 71% of those edited clones carried the designed nonsense mutation plus the designed synonymous mutation in the PAM sequence, 26% contained only the PAM edit, and 3% showed neither of the designed edits at this locus (Table 1).

Further, CREATE could be functionally transferred into the yeast *S. cerevisiae*. Editing of the *S. cerevisiae ade2* gene to introduce a tandem stop codon showed editing efficiencies of 95% based on a colorimetric read-out, with 100% of those phenotypic hits being correctly programmed (Table 1).

In a follow-up study, the authors show the scalability of CREATE by deep mutagenesis of an entire metabolic network.[54] Specifically, the authors designed 16,300 mutations within the binding-pockets of 19 enzymes or transporters, that comprise four primary routes that guide lysine flux in *E. coli*. Using NGS analysis of four test loci, the authors show the library coverage to range between 22.7 to 61.6%, indicating high editing efficiencies given the size of the library (Table 1). Eventually, by mutationally perturbing *E. coli*'s lysin metabolic network and subsequently challenging this library with the antimetabolite AEC, the authors could evaluate in parallel the contribution of these 16,300 targeted mutations toward antimetabolite resistance and thus overall pathway flux. As such, the authors could (near) comprehensively map sequence-function relations that alter the pathway's function, setting a framework for investigating complex multigenic phenotypes.

2.3. Further considerations for saturation mutagenesis

In summary, NM, PALS, PR and CREATE are effective ways to create (near-)comprehensive single-site or multi-site mutagenic libraries with the use of oligo pools. For library screening and associated library coverage calculations, it needs to be kept in mind that all methods lead to libraries that show only ~50% or less of the total molecules to be desired single mutations, and contain a significant amount of wild-type plasmid or plasmids with non-desired mutations (Table 1 and 2). Mutation efficiency might not matter in case a high-throughput selection is available, but might be limiting for protein engineering efforts that require expensive or laborious screens as the number of screened mutants needs to be doubled to cover such a library (compared to a 100% mutagenized library).

3. Encoding Short DNA Stretches: Oligo Pool-Based Insertion and Deletion Scanning Libraries

Besides changing single amino acids, several protein engineering approaches require the insertion of short stretches of DNA that either encode for functional peptide tags[55] – such as affinity tags for purification or detection,[56] protease cleavage sites for on-demand inactivation,[57,58] tags for post-translational

modification via click chemistry,[59] or labeling sites for protein visualization[60] – or encode for molecular recognition sites required for further library processing; for example, restriction enzymes sites that can subsequently be used for creating systematic insertion or deletion libraries.

While tagging with functional peptides is usually done at the N- or C-terminus of the protein, in certain instances this is impossible – in case the termini are functionally relevant or the functional insert needs to be inserted in the middle, as is the case for targeted inaction via protease cleavage[58] – and proteins need to be screened for permissive sites that accept additional amino acids.

In the context of enzyme engineering, random insertion and deletion (indel) libraries have recently been shown to enhance the evolvability of proteins.[61] Therefore, indel evolution could be a promising yet underexplored route towards new enzymatic function.

Indel libraries are traditionally created via transposon mutagenesis,[61–63] but those protocols are often time-consuming and transposon insertion site bias[64] compromises the creation of uniform and comprehensive indel libraries. Here, the above-outlined array-oligo pool-based mutational protocols could become effective alternatives.

3.1. *In vitro* techniques for creating insertion or deletion libraries

PALS was used to create a comprehensive one amino acid deletion library of the yeast transcription factor Gal4.[37] It is imaginable that PALS could also be used to systematically delete more than one residue or insert a short peptide tag, by encoding a tag or deletion in the middle of a mutagenic oligo. Similarly, it should also be feasible to create indel libraries using NM. An oligo-encoded deletion or insertion approach (at one or a few sites) has shown feasible by the QuikChange™ protocol,[65] but to the best of our knowledge, it has not yet been experimentally used to create comprehensive indel libraries using NM or PALS.

3.2. *In vivo* techniques for creating insertion or deletion libraries

PR was shown capable of generating a comprehensive insertion library (insertion of a given tag after each amino acid) for the entire open reading frame of the CRISPR protein Cas9 from *Streptococcus pyogenes* (SpCas9, 1368 amino acids).[53] The insert in this case was a 6 bp restriction site encoding for either one of the two restriction sites NheI or SpeI (both libraries need to be created for their MISER method). Eventually, those sites were used to generate a comprehensive deletion library of SpCas9 with the goal to systematically size-minimize this large protein to make it better suitable for medical and bioengineering applications; a general method that can systematically explore deletion-landscapes of any given protein and which they call genetic **m**inimization by **i**terative **s**ize-**e**xclusion and **r**ecombination (MISER).

To address the low mutagenic efficiency of PR (the frequency of plasmids containing an insert after one round of PR, versus the number of wild-type plasmid), the authors use the restriction sites in the tag to sub-clone an antibiotic resistance marker (Cm), allowing the experimenter to select for plasmids with an insert. In a second step this marker is cut out and the plasmid re-ligated, yielding a 100% mutagenized library. Of course, this strategy is only viable for tags encoding one or several restriction sites, but those could be either encoded or added to a given tag.

Further, CREATE was shown to enable the introduction of deletions (100 bp) in the genomically encoded *E. coli galK* gene with 70% efficiency.[42] Based on this efficiency and the overall scalability of CREATE,[54] it is likely that customized (shorter or longer) (near)-comprehensive deletion or insertion libraries can be created using this method.

4. Encoding Complex Libraries of Short Genes (up to 700 bp) via Oligo Pools

Once the number of required changes per open reading frame increases, site-directed mutagenesis can become inefficient, and complete *de novo* synthesis of a gene is required. In enzyme engineering, this can be the case when a heterologous enzyme or a full pathway of multiple enzymes needs to be recoded for optimized codon usage or if various genes of a pathway need to be engineered together. One field that historically depended on the massive synthesis of new-to-nature DNA is the field of *de novo* protein design. Here, often >7000 computational designs need to be tested for sequence-function relations. Not surprisingly, this field has early on developed techniques that use oligo pools for the synthesis of these designs.

In this section, we will introduce gene synthesis methods that are optimized to produce large libraries of different gene fragments in one pool. Those libraries are typically required during large-scale functional testing of *de novo* designed proteins, and for targeted protein engineering at multiple sites. Those methods are less suitable for synthesizing long genes (>700 bp) or a pathway. Methods that are suitable to assemble a single gene or a multi-gene pathway from oligo pools, rather than a library of variants, will be discussed in section 4.

4.1. Gene assembly in one pot without compartmentalization

In an effort to assemble 2271 designs of a short, 64 to 84 amino acid long protein domain, Klein *et al.* developed a method called multiplexed pairwise assembly (herein abbreviated as MPA) that allows assembling gene fragment libraries of ~250 bp in length in joint-pools of 250 to 2271 targets.[38] Similar to the saturation-mutagenesis methods reported above, MPA relies on relatively standard enzymatic molecular biology

reactions for assembly. In pools of relatively low complexity (250 targets), MPA achieved the error-free synthesis of on average 90.5% of these 250 targets (% coverage, Table 3). Increasing the target pool complexity to the synthesis of the complete desired set of 2271 protein designs simultaneously in one-pot, MPA still allowed the error-free synthesis of 70.6% of the 2271 targets (Table 3); and 11.8 to 31.3% of all assembled molecules are error-free (% accuracy, Table 3) Although not shown, one can imagine that the 250 bp fragments assembled via MPA can be assembled to longer genes using, for example, Golden Gate assembly.

MPA starts with designing the 250 bp target. Each target is then computationally divided into two partially overlapping fragments A and B (each fragment is 160 bp including the adaptors required for priming later in the protocol). Oligo encoded fragments are ordered containing one 3′ pool-specific and one 5′ general adaptor for priming. This allows encoding all fragments in one oligo pool but to specifically amplify those that go into an assembly reaction together. The primers binding to the pool-specific adaptors are uracil-containing, such that these priming sites (as they would interfere with hybridization of A and B fragments) can be removed prior to assembly with Uracil Specific Excision Reagent (USER enzymes). The assembly is based on overlap extension PCR using the same outer primers as used for oligo amplification. Although deep sequencing revealed that between 70% and 90.5% (depending on the number of targets per pool) of the assembled targets had correct error-free molecules represented in the pool of all molecules (Table 3), those error-free molecules needed to be purified from the pool of all error-containing molecules for downstream applications. Here the authors use dial-out PCR, a method that allows identifying and amplifying error-free assemblies to retrieve them for downstream applications.[66,67] Specifically, in dial-out PCR, single molecules get tagged with unique barcodes before sequencing. Error-free molecules can then be identified via NGS and amplified via PCR using primers that bind to the unique barcodes of error-free assemblies.

4.2. Compartmentalized assembly in water-in-oil droplets

A second method that can assemble thousands of genes from microarray-derived oligos in a single reaction is DropSynth[39] and its advanced version DropSynth 2.0.[43] Here, fragments of ~450–675 bp in length are assembled within water-in-oil droplets, allowing the experimenter to multiplex many assemblies. The concept is that all oligos required for one assembly get hybridized to a microbead using oligo-barcodes. The oligo-coated microbeads then get encapsulated in a water-in-oil droplet, creating spatially separate reactions.

DropSynth follows several steps: The genes to be synthesized are first bioinformatically split into several fragments, such that each fragment can fit onto one oligo. Restriction sites and a microbead barcode are also added to each oligo. Oligos are then amplified by PCR using a biotinylated primer, followed by digestion at high temperature to expose the microbead barcode as a single-stranded DNA overhang. Processed oligos are mixed with a pool of either 384 or 1536 barcoded microbeads (limited by the number of unique barcodes), with each microbead containing only one complementary barcode sequence. Complementary oligos (all that are required for a specific assembly) hybridize and are ligated to the microbeads. The loaded beads are then mixed with PCR reagents, a restriction enzyme, and some fluorinated oil and vortexed to form a water-in-oil emulsion, which is placed into a thermocycler where the restriction enzyme displaces the oligos from the bead and the gene assembly reaction takes place inside the droplets. Upon completion, the aqueous solution containing the assembled genes is recovered from the emulsion and PCR-amplified for downstream applications. The authors show that the optimized DropSynth 2.0 protocol[43] can be used to build thousands of gene-length fragments at >20% accuracy, meaning that >20% of the assemblies are error-free, with a coverage of 80–92% depending on the number of targets that were attempted to be assembled in one pool (% coverage, Table 3).

5. Synthesizing Individual Genes and Pathways

5.1. *In vitro*-based methods for gene and pathway assembly

Over the last decade, several protocols have been developed that allow the assembly of genes and multi-gene pathways from error-prone array-based oligo pools.[18,68–70] They all involve three core steps: First, amplifying sub-pools of oligos via pool-specific primers; this is necessary to get enough oligos for the assembly. Second, assembly of sub-pools by overlap PCR or ligase chain reaction, and third, an error removal step, that removes erroneous molecules. Error removal either involves sequencing-based methods such as dial-out PCR[66] or it is based on depleting erroneous molecules via enzymatic cleavage or affinity-based capture. The latter is based on the fact that

Table 3. Performance overview of methods for gene library assembly (e.g. for testing various protein designs).

Method	Gene size (bp)	Pool size[a]	% coverage[b]	% accurate assemblies[c]	Ref.
MPA	192–252	131 to 250	72.7 to 96.4	11.8 to 31.3	[38]
	192–252	1212	84.2	11.8 to 31.3	[38]
	192–252	2271	70.6	11.8 to 31.3	[38]
DropSynth 2.0	675	384	92.0	23.5	[43]
	675	1536	80.0	22.6 to 27.6	[43]

[a] Number of independent assemblies performed in one pool. [b] Number of assemblies (per 100 designed assemblies) that show at least one error-free molecule. [c] Number of error-free assembled molecules per 100 assembled molecules.

during gene assembly a pool of perfect and imperfect double-stranded sequences is produced. Melting and reannealing pairs perfect and imperfect strands and mis-hybridized bases can be recognized by mismatch binding proteins or mismatch cleaving proteins. The efficiency of various enzymes and binding proteins has been systematically compared.[71]

Here we will highlight one gene- and pathway- assembly protocol that builds on available methods for each of the three above outlined steps, but is optimized to achieve a very low error rate (0.53 per kb) and most importantly can be implemented with regular molecular biology expertise and equipment:[40] The authors use the full *de novo* synthesis of the 10-gene lycopene biosynthetic pathway (11.9 kb) from 479 oligos (64 to 124 bp) as an example. The protocol uses PCR to amplify the entire oligo pool followed by a first error removal step via multiple consecutive annealing and MutS-immobilized column purifications (so-called MutS-immobilized cellulose column, MICC).[72] Error-depleted oligos are then assembled into 500 bp fragments using ligase chain reaction. Residual errors are removed by a second round of MICC. The 500 bp parts are then assembled into the full pathway (encoded as three operons) using Gibson assembly.

5.2. Potential *in vivo* approaches for gene and pathway assembly

Besides the above outlined enzymatic gene assembly, it is well known that simple yeast assembly – based on *S. cerevisiae's* intrinsic capacity to perform homology-repair-based assembly of overlapping fragments - can be used to assemble genes of >1000 bp length from short overlapping oligonucleotides.[73] While in the original method the authors use column-synthesized oligonucleotides at a scale of >10 nmol, it can be imagined that this method is viable using amplified next-generation oligo pools, although this has not been shown yet.

6. Library Analysis by Second- and Third-Generation Sequencing

Next-generation protein engineering relies on the creation of large libraries often created in one pot and often created by methods that are not error-free. As such, the created libraries must be properly evaluated for coverage, target mutation frequency, and off-target mutation frequency, to calculate a sufficient screening sample size to ensure library coverage. Evaluation has mostly been done by short-read NGS (also called second-generation sequencing), such as Illumina-based sequencing, because of the high throughput, low cost, and wide accessibility of the method. One limitation of second-generation sequencing is the inherent short read length (75 to 500 nucleotides for Illumina sequencing platform). Therefore, a mutation that is located outside of the read-window would be invisible in a library analysis. Because of these limitations, DMS or *de novo* protein-coding is usually performed on small genes or subsets of genes (tiling) (Figure 1B). Several second-generation sequencing approaches designed for protein engineering have been reviewed elsewhere.[6]

Here we only want to highlight methods that overcome the resolution window in second-generation sequencing, and recent methods that improve so-called long-read (third-generation) sequencing approaches and make them suitable for library analysis in protein engineering.

6.1. Increasing the resolution window of short-read (second-generation) sequencing-based library analysis methods

A method called "subassembly" is powerful to overcome the limited read length in NGS as full-length sequences of mutagenized clones can be obtained from short NGS reads.[44] In subassembly, each mutant clone in a complex library is individually coupled to a random molecular tag. Paired-end reads are obtained with one fixed end reporting the tag sequence, and one shotgun end derived randomly from the insert. Shotgun reads are then grouped by tag to yield an accurate full-length consensus haplotype that is longer than the constituent reads and can detect random sequencing errors.

6.2. Increasing accuracy and throughput of long-read (third-generation) sequencing for library analysis

Read lengths longer than 500 bp have been possible for a while using single-molecule real-time (SMRT) sequencing or single-molecule Nanopore sequencing (offered by PacBio Oxford Nanopore respectively), but both technologies show reduced throughput and reduced accuracy when compared to short-read sequencing methods,[74] which so far made them less suitable for library analysis.[6]

To enhance the accuracy of SMRT-based sequencing, Waltenspül *et al.* recently developed a computational error correction workflow that eventually allowed them to use SMRT-based sequencing for the full-length sequencing of a G-protein coupled receptor library (gene length >1000 bp).[45] Like this, information of mutational linkage was maintained due to the fact that each read covers the full protein gene.

In addition to the above-outlined accuracy improvements, Schlecht *et al.*[46] and Kanwar *et al.*[47] developed methods to increase the throughput of SMRT-based sequencing by up to 5-fold. They use protein-encoding libraries of up to 870 bp in length as an example. The throughput enhancement was achieved by concatenating individual library members – either using Gibson assembly or Golden gate assembly – into longer fragments of up to five library members, which are then sequenced together, taking full advantage of the long reads that can be achieved by SMRT-based sequencing.

These advancements in accuracy and throughput will likely make SMRT-based sequencing a valid method of choice for the validation of protein engineering libraries.

7. Conclusions and Outlook

Here we summarized currently available methods that use cheap but error-prone array-synthesized oligo pools as a source of synthetic DNA for protein engineering libraries. These methods allow the creation of targeted or comprehensive mutagenic libraries required for deep-mutational scanning, directed evolution, or rational protein engineering. Further, they allow the assembly of *de novo* designed gene libraries of several thousands of genes of up to 700 bp in length as well as the assembly of defined longer genes and pathways (Figure 1A).

All methods only require standard molecular biology expertise and equipment to create libraries and relatively standard bioinformatics expertise to analyze the NGS data. As such, they provide affordable access to next-generation protein engineering libraries for many laboratories.

Combining the herein outlined methods with newly developed methods for oligo pool purification,[26] should further enhance the capacity of oligo pools for protein and pathway engineering by enhancing the sequence accuracy of assemblies, a major bottle-neck for scale-up.

In addition to library creation methods, we highlight how these libraries can be evaluated for quality and coverage using next-generation sequencing and bioinformatics (Figure 1B); specifically, we point to newly developed long-read (second-generation) sequencing workflows that overcome accuracy and throughput limitations formerly associated with those second-generation methods in the context of library quality control. Being able to sequence longer (gene length) fragments with high accuracy and throughput facilitates the quality control and sequence analysis of gene- and pathway length libraries without the need for tiling it into shorter fragments.

In summary, in-house array-synthesized oligo pool-based protein library creation and analysis comes of age and can enable many exciting next-generation protein engineering endeavors in many laboratories.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: array-based oligonucleotides · gene synthesis · mutagenesis · protein engineering

[1] D. M. Fowler, S. Fields, *Nat. Methods* **2014**, *11*, 801–807.
[2] K. K. Yang, Z. Wu, F. H. Arnold, *Nat. Methods* **2019**, *16*, 687–694.
[3] S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, G. M. Church, *Nat. Methods* **2021**, *18*, 389–396.
[4] P. S. Huang, S. E. Boyken, D. Baker, *Nature* **2016**, *537*, 320–327.
[5] X. Pan, T. Kortemme, *J. Biol. Chem.* **2021**, *296*, 100558.
[6] E. E. Wrenbeck, M. S. Faber, T. A. Whitehead, *Curr. Opin. Struct. Biol.* **2017**, *45*, 36–44.
[7] Z. Wu, S. B. Jennifer Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 8852–8858.
[8] K. K. Yang, Z. Wu, F. H. Arnold, *Nat. Methods* **2019**, *16*, 687–694.
[9] C. G. Acevedo-Rocha, M. Ferla, M. T. Reetz, in *Methods in Molecular Biology*, Humana, **2018**, pp. 87–128.
[10] J. R. Klesmith, J. P. Bacik, E. E. Wrenbeck, R. Michalczyk, T. A. Whitehead, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2265–2270.
[11] M. S. Faber, E. E. Wrenbeck, L. R. Azouz, P. J. Steiner, T. A. Whitehead, *Mol. Biol. Evol.* **2019**, *36*, 2764–2777.
[12] E. E. Wrenbeck, M. A. Bedewitz, J. R. Klesmith, S. Noshin, C. S. Barry, T. A. Whitehead, *ACS Synth. Biol.* **2019**, *8*, 474–481.
[13] E. E. Wrenbeck, L. R. Azouz, T. A. Whitehead, *Nat. Commun.* **2017**, *8*, 15695.
[14] D. Esposito, J. Weile, J. Shendure, L. M. Starita, A. T. Papenfuss, F. P. Roth, D. M. Fowler, A. F. Rubin, *Genome Biol.* **2019**, *20*, 1–11.
[15] B. Basanta, M. J. Bick, A. K. Bera, C. Norn, C. M. Chow, L. P. Carter, I. Goreshnik, F. Dimaio, D. Baker, *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 22135–22145.
[16] S. Kosuri, G. M. Church, *Nat. Methods* **2014**, *11*, 499–507.
[17] S. L. Beaucage, M. H. Caruthers, *Tetrahedron Lett.* **1981**, *22*, 1859–1862.
[18] S. Kosuri, N. Eroshenko, E. M. Leproust, M. Super, J. Way, J. B. Li, G. M. Church, *Nat. Biotechnol.* **2010**, *28*, 1295–1299.
[19] D. S. Kong, P. A. Carr, L. Chen, S. Zhang, J. M. Jacobson, *Nucleic Acids Res.* **2007**, *35*, e61.
[20] K. L. Agarwal, H. Büchi, C. H. Caruthers, N. Gupta, H. G. Khorana, K. Kleppe, A. Kumar, E. Ohtsuka, U. L. Rajbhandary, J. H. Van de Sande, V. Sgaramella, H. Weber, T. Yamada, *Nature* **1970**, *227*, 27–34.
[21] M. J. Czar, J. C. Anderson, J. S. Bader, J. Peccoud, *Trends Biotechnol.* **2009**, *27*, 63–72.
[22] A. P. Blanchard, R. J. Kaiser, L. E. Hood, in *Biosens. Bioelectron.*, Elsevier Science Ltd, **1996**, 687–690.
[23] T. R. Hughes, M. Mao, A. R. Jones, J. Burchard, M. J. Marton, K. W. Shannon, S. M. Lefkowitz, M. Ziman, J. M. Schelter, M. R. Meyer, S. Kobayashi, C. Davis, H. Dai, Y. D. He, S. B. Stephaniants, G. Cavet, W. L. Walker, A. West, E. Coffey, D. D. Shoemaker, R. Stoughton, A. P. Blanchard, S. H. Friend, P. S. Linsley, *Nat. Biotechnol.* **2001**, *19*, 342–347.
[24] I. Saaem, K. S. Ma, A. N. Marchi, T. H. LaBean, J. Tian, *ACS Appl. Mater. Interfaces* **2010**, *2*, 491–497.
[25] A. L. Ghindilis, M. W. Smith, K. R. Schwarzkopf, K. M. Roth, K. Peyvan, S. B. Munro, M. J. Lodes, A. G. Stöver, K. Bernards, K. Dill, A. McShea, *Biosens. Bioelectron.* **2007**, *22*, 1853–1860.
[26] H. Choi, Y. Choi, J. Choi, A. C. Lee, H. Yeom, J. Hyun, T. Ryu, S. Kwon, *Nat. Biotechnol.* **2021**, 1–7.
[27] O. Shalem, N. Sanjana, F. Zhang, *Nat. Rev. Genet.* **2015**, *16*, 299–311.
[28] A. Read, S. Gao, E. Batchelor, J. Luo, *Nucleic Acids Res.* **2017**, *45*, e101.
[29] J. D. Smith, U. Schlecht, W. Xu, S. Suresh, J. Horecka, M. J. Proctor, R. S. Aiyar, R. A. O. Bennett, A. Chu, Y. F. Li, K. Roy, R. W. Davis, L. M. Steinmetz, R. W. Hyman, S. F. Levy, R. P. S. Onge, *Mol. Syst. Biol.* **2017**, *13*, 913.
[30] R. P. Patwardhan, C. Lee, O. Litvin, D. L. Young, D. Pe'Er, J. Shendure, *Nat. Biotechnol.* **2009**, *27*, 1173–1175.
[31] E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, E. Segal, *Nat. Biotechnol.* **2012**, *30*, 521–530.
[32] M. R. Schlabach, J. K. Hu, M. Li, S. J. Elledge, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 2538–2543.
[33] A. Melnikov, A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C. G. Callan, J. B. Kinney, M. Kellis, E. S. Lander, T. S. Mikkelsen, *Nat. Biotechnol.* **2012**, *30*, 271–277.
[34] S. Rong, L. Buerer, C. L. Rhine, J. Wang, K. J. Cygan, W. G. Fairbrother, *Nat. Commun.* **2020**, *11*, 1–10.
[35] E. E. Wrenbeck, J. R. Klesmith, J. A. Stapleton, A. Adeniran, K. E. J. Tyo, T. A. Whitehead, *Nat. Methods* **2016**, *13*, 928–930.
[36] A. V. Medina-Cucurella, P. J. Steiner, M. S. Faber, J. Beltrán, A. N. Borelli, M. B. Kirby, S. R. Cutler, T. A. Whitehead, *Protein Eng. Des. Sel.* **2019**, *32*, 41–45.
[37] J. O. Kitzman, L. M. Starita, R. S. Lo, S. Fields, J. Shendure, *Nat. Methods* **2015**, *12*, 203–206.
[38] J. C. Klein, M. J. Lajoie, J. J. Schwartz, E. M. Strauch, J. Nelson, D. Baker, J. Shendure, *Nucleic Acids Res.* **2015**, *44*, e43.
[39] C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, S. Kosuri, *Science* **2018**, *359*, 343–347.
[40] W. Wan, M. Lu, D. Wang, X. Gao, J. Hong, *Sci. Rep.* **2017**, *7*, 1–11.
[41] S. A. Higgins, S. V. Y. Ouonkap, D. F. Savage, *ACS Synth. Biol.* **2017**, *6*, 1825–1833.
[42] A. D. Garst, M. C. Bassalo, G. Pines, S. A. Lynch, A. L. Halweg-Edwards, R. Liu, L. Liang, Z. Wang, R. Zeitoun, W. G. Alexander, R. T. Gill, *Nat. Biotechnol.* **2017**, *35*, 48–55.

[43] A. M. Sidore, C. Plesa, J. A. Samson, N. B. Lubock, S. Kosuri, *Nucleic Acids Res.* **2020**, *48*, E95.
[44] J. Hiatt, R. Patwardhan, E. H. Turner, C. Lee, J. Shendure, *Nat. Methods* **2010**, *7*, 119–122.
[45] Y. Waltenspühl, J. R. Jeliazkov, L. Kummer, A. Plückthun, *Sci. Rep.* **2021**, *11*, 1–12.
[46] U. Schlecht, J. Mok, C. Dallett, J. Berka, *Sci. Rep.* **2017**, *7*, 1–10.
[47] N. Kanwar, C. Blanco, I. A. Chen, B. Seelig, *Sci. Rep.* **2021**, *11*, 1–13.
[48] E. Firnberg, M. Ostermeier, *PLoS One* **2012**, *7*, e0052031.
[49] M. B. Kirby, A. V. Medina-Cucurella, Z. T. Baumer, T. A. Whitehead, *Protein Eng. Des. Sel.* **2021**, *34*, gzab017.
[50] M. S. Faber, J. T. Van Leuven, M. M. Ederer, Y. Sapozhnikov, Z. L. Wilson, H. A. Wichman, T. A. Whitehead, C. R. Miller, *ACS Synth. Biol.* **2020**, *9*, 125–131.
[51] H. M. Ellis, D. Yu, T. DiTizio, D. L. Court, *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 6742–6746.
[52] H. H. Wang, F. J. Isaacs, P. A. Carr, Z. Z. Sun, G. Xu, C. R. Forest, G. M. Church, *Nature* **2009**, *460*, 894–898.
[53] A. Shams, S. A. Higgins, C. Fellmann, T. G. Laughlin, B. L. Oakes, R. Lew, S. Kim, M. Lukarska, M. Arnold, B. T. Staahl, J. A. Doudna, D. F. Savage, *Nat. Commun.* **2021**, *12*, 1–11.
[54] M. C. Bassalo, A. D. Garst, A. Choudhury, W. C. Grau, E. J. Oh, E. Spindler, T. Lipscomb, R. T. Gill, *Mol. Syst. Biol.* **2018**, *14*, e00232-e00320.
[55] S. Billerbeck, in *Synthetic Biology: Parts, Devices and Applications* (Eds.: C. Smolke, S. Y. Lee, J. Nielsen, G. Stephanopoulos), Wiley-VCH, Weinheim, **2018**, pp. 217–235.
[56] A. S. Pina, Í. L. Batalha, A. C. A. Roque, *Methods Mol. Biol.* **2014**, *1129*, 147–168.
[57] R. B. Kapust, D. S. Waugh, *Protein Expression Purif.* **2000**, *19*, 312–318.
[58] S. Oesterle, T. M. Roberts, L. A. Widmer, H. Mustafa, S. Panke, S. Billerbeck, *BMC Biol.* **2017**, *15*, 100.
[59] I. S. Carrico, B. L. Carlson, C. R. Bertozzi, *Nat. Chem. Biol.* **2007**, *3*, 321–322.
[60] S. R. Adams, R. E. Campbell, L. A. Gross, B. R. Martin, G. K. Walkup, Y. Yao, J. Llopis, R. Y. Tsien, *J. Am. Chem. Soc.* **2002**, *124*, 6063–6076.
[61] S. Emond, M. Petek, E. J. Kay, B. Heames, S. R. A. Devenish, N. Tokuriki, F. Hollfelder, *Nat. Commun.* **2020**, *11*, 1–14.
[62] V. De Lorenzo, M. Herrero, U. Jakubzik, K. N. Timmis, *J. Bacteriol.* **1990**, *172*, 6568–6572.
[63] S. Billerbeck, B. Calles, C. L. Müller, V. De Lorenzo, S. Panke, *ChemBioChem* **2013**, *14*, 2310–2321.
[64] B. Green, C. Bouchier, C. Fairhead, N. L. Craig, B. P. Cormack, *Mob. DNA* **2012**, *3*, 1–6.
[65] H. Liu, J. H. Naismith, *BMC Biotechnol.* **2008**, *8*, 91.
[66] J. J. Schwartz, C. Lee, J. Shendure, *Nat. Methods* **2012**, *9*, 913–915.
[67] H. Lim, N. Cho, J. Ahn, S. Park, H. Jang, H. Kim, H. Han, J. H. Lee, D. Bang, *Nucleic Acids Res.* **2018**, *46*, e40.
[68] A. Y. Borovkov, A. V. Loskutov, M. D. Robida, K. M. Day, J. A. Cano, T. Le Olson, H. Patel, K. Brown, P. D. Hunter, K. F. Sykes, *Nucleic Acids Res.* **2010**, *38*, e180-e180.
[69] H. Kim, H. Han, J. Ahn, J. Lee, N. Cho, H. Jang, H. Kim, S. Kwon, D. Bang, *Nucleic Acids Res.* **2012**, *40*, e140.
[70] N. Cho, H. N. Seo, T. Ryu, E. Kwon, S. Huh, J. Noh, H. Yeom, B. Hwang, H. Ha, J. H. Lee, S. Kwon, D. Bang, *Nucleic Acids Res.* **2018**, *46*, e55.
[71] L. N. B., Z. D., S. A. M., C. G. M., K. S, *Nucleic Acids Res.* **2017**, *45*, 9206–9217.
[72] W. Wan, L. Li, Q. Xu, Z. Wang, Y. Yao, R. Wang, J. Zhang, H. Liu, X. Gao, J. Hong, *Nucleic Acids Res.* **2014**, *42*, e102.
[73] D. G. Gibson, *Nucleic Acids Res.* **2009**, *37*, 6984.
[74] E. J. Fox, K. S. Reid-Bayliss, M. J. Emond, L. A. Loeb, *Next Gener. Seq. Appl.* **2014**, *1*.

Manuscript received: September 23, 2021
Revised manuscript received: November 23, 2021
Accepted manuscript online: November 24, 2021
Version of record online: December 7, 2021

Title TBD

Further reading and resources

Precision Genome Editing Puts Cancer Mutations Under the Spotlight
Customer Spotlight

University of Washington's David Baker Lab Develops a New Way to Design Proteins
Case study

Research Citation Collection
Agilent Custom SurePrint Oligonucleotide Libraries

SurePrint HiFi Oligo Pools
Product details

SurePrint Oligo Pools
Product details

WILEY

Imprint

Email: info@wiley-vch.de

Editor: Róisín Murtagh

RAXXXXXXXXXXX