Characterizing and circumventing sequence restrictions for synthesis of circular RNA in vitro

Jason W. Rausch 1,*, William F. Heinz 2, Matthew J. Payea 3, Chringma Sherpa 1, Myriam Gorospe 3 and Stuart F. J. Le Grice 1

1 Basic Research Laboratory, National Cancer Institute, NIH, Frederick, MD 21702, USA, 2 Optical Microscopy and Analysis Laboratory, Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA and 3 Laboratory of Genetics and Genomics, National Institute on Aging–Intramural Research Program, National Institutes of Health, Baltimore, MD 21224, USA

ABSTRACT

Just as eukaryotic circular RNA (circRNA) is a product of intracellular backsplicing, custom circRNA can be synthesized in vitro using a transcription template in which transposed halves of a split group I intron flank the sequence of the RNA to be circularized. Such permuted intron–exon (PIE) constructs have been used to produce circRNA versions of ribozymes, mimics of viral RNA motifs, a streptavidin aptamer, and protein expression vectors for genetic engineering and vaccine development. One limitation of this approach is the obligatory incorporation of small RNA segments (E1 and E2) into nascent circRNA at the site of end-joining. This restriction may preclude synthesis of small circRNA therapeutics and RNA nanoparticles that are sensitive to extraneous sequence, as well as larger circRNA mimics whose sequences must precisely match those of the native species on which they are modelled. In this work, we used serial mutagenesis and in vitro selection to determine how varying E1 and E2 sequences in a thymidylate synthase (td) group I intron PIE transcription template construct affects circRNA synthesis yield. Based on our collective findings, we present guidelines for the design of custom-tailored PIE transcription templates from which synthetic circRNAs of almost any sequence may be efficiently synthesized.

INTRODUCTION

In eukaryotes, circRNAs are produced by backsplicing, which occurs when a donor splice site at the 5′ end of an intron and an acceptor splice site 3′ to an upstream intron are incorporated into and processed by the same spliceosome (1). Multiple instances of this inversion of canonical splicing have been demonstrated across numerous cell lines (2), and the abundance of circRNA in human cells relative to polyadenylated mRNA has been estimated to be as high as 1% (3). Expression of circRNAs is often cell-type specific, some are evolutionarily conserved and many exert important biological functions by sequestering microRNA or RNA-binding proteins (RBPs), regulating protein function or themselves being translated (4). CircRNAs have been implicated in, or serve as biomarkers of diseases such as diabetes mellitus, neurological disorders, cardiovascular diseases, and cancer (5).

Because they lack termini, the conformational flexibility of circRNA may be less than that of their linear counterparts, and they are inherently impervious to exoribonucleases. Accordingly, just as endogenous circRNAs are long-lived in eukaryotic cells, synthetic circRNAs have proven to be functionally superior as siRNA mimics, i.e. RNA dumbbells (6), or as components of geometric RNA nanoparticles (7). Construction of small synthetic circRNAs has typically involved chemical synthesis and purification of 5′-phosphorylated RNA, splint ligation of proximal 3′ and 5′-phosphorylated termini using commercially available enzymes, and purification of the ligation product from similarly sized linear precursors. Although this approach may be applied to synthesize simple circRNAs with manageable efficiency, construction of longer and more structurally complex variants in this manner is often complicated by se-
sequence variability at RNA termini, ligation junction imprecision, and low yield.

A simpler and more robust alternative method for producing circRNA in vitro exploits the properties of self-splicing group I introns such as those found in the pre-rRNA_{Leu} and thymidylate synthase (td) genes of the cyanobacteria *Anaabaena* and T4 bacteriophage, respectively (8,9). Natively, these elements are autocatalytically excised from linear RNA precursors and the flanking exons ligated in tandem transesterification reactions. For circularization constructs, DNA sequences encoding these introns are genetically split into 5’ and 3’ halves (5H, 3H) and assembled with the DNA equivalent of the RNA to be circularized (DRC, RC) in an order that promotes backsplicing (i.e. 5’-3H/DRC/5H-3’). Such constructs are typically amplified and an upstream T7 promoter is added by PCR, the products of which are used as templates in run-off in vitro transcription reactions. Segregated intron halves in nascent RNA transcripts spontaneously assemble into functional group I introns by conformational sampling, whereupon circRNA is generated by intron self-excision and ligation of RC termini (10,11).

Such permuted intron–exon (PIE) constructs have been used to produce circRNA versions of ribozymes (hepatitis delta virus, hammerhead, RNase P), mimics of viral RNA motifs (e.g. the 5’ and 3’ terminal regions of human endogenous circRNAs), and to produce RNAs for model circRNAs (i.e. 5’-3H/DRC/5H-3’). Such constructs are typically amplified and an upstream T7 promoter is added by PCR, the products of which are used as templates in run-off in vitro transcription reactions. Segregated intron halves in nascent RNA transcripts spontaneously assemble into functional group I introns by conformational sampling, whereupon circRNA is generated by intron self-excision and ligation of RC termini (10,11).

In this work, we present a strategy for PIE construct design by which the sequence of the RNA to be circularized is permuted so its 5’ and 3’ terminal sequences resemble T4 td group I intron segments E2 and E1, respectively. Though sites of end-joining in synthetic circRNA produced in this manner may differ from their natural counterparts, the sequences of the respective RNAs will exactly match. In addition, by mutational analysis and in vitro selection, we characterize the obligatory E1–E2 incorporation into circRNA in detail and explore the degree to which these requirements may be circumvented or compensated for in T4 td PIE construct design. Based on the results of our analysis, we present a set of guidelines for designing PIE transcription templates for efficient synthesis of almost any circRNA, including models of endogenous human variants, and discuss new potential applications thereof.

**MATERIALS AND METHODS**

**T4 td PIE transcription template design and construction**

Transcription templates were constructed to include the permuted T4 td gene group I intron previously reported (11) together with complementary 5’- and 3’-terminal ‘homology arms’ to facilitate co- and post-transcriptional folding of the functional group I intron RNA (13). From 5’ to 3’, template components include the (i) T7 promoter, (ii) 5’ homology arm, (iii) 3’ intron half, (iv) E2, (v) DNA equivalent of the RNA to be circularized, (vi) E1, (vii) 5’ intron half (including IG) and (viii) 3’ homology arm. For this study, DNA encoding RNA to be circularized included a permuted form of *circPVT1* (14) and variants of a circular form of a streptavidin aptamer that also contains native T4 td exon sequence (cSAA; (11)). The full sequence of a generic T4 td PIE vector is provided in Supplementary Materials. Transcription templates were commercially synthesized and cloned into the pUC57 plasmid (GenScript, Piscataway, NJ), and amplified by PCR for run-off in vitro transcription.

**In vitro transcription, RNA circularization, circRNA processing and analysis**

The MEGAscript or MEGashortscript transcription kit (Ambion) was used for *in vitro* transcription of *circPVT1* and cSAA variant precursors, respectively. For *circPVT1* RNA synthesis, 8 μl of transcription template from an unpurified PCR reaction was mixed with 12 μl of other reaction components (2 μl 10× Transcription Buffer, 4 × 2 μl each rNTP, 2 μl T7 RNA polymerase Enzyme Mix) and incubated for 6 h at 37°C. For cSAA variant transcription, 500 fmol of column-purified (Monarch PCR & DNA Cleanup Kit – 5 μg, NEB) transcription template was added to each 20 μl transcription reaction. Transcription templates were removed with TurboDNase as per manufacturer’s instructions, and post-transcriptional circularization was facilitated by supplementing each reaction with 2 mM GTP and incubating at 55°C for 15 min, followed by rapid cooling to 4°C. RNA was purified using a MEGAClear kit (circPVT1; Ambion) or by ammonium acetate/ethanol precipitation (cSAA) according to MEGashortscript instructions. Where indicated, purified RNAs were treated with RNase R (Lucigen) according to manufacturer’s instructions and re-purified as described. Reaction products were fractionated over 2% or 4% non-denaturing agarose gels containing ethidium bromide (Thermo Fisher Scientific), gel images were recorded using an Azure Biosystems c280 imaging system, and circRNA and total RNA quantified for each reaction using AzureSpot software. Fractional cSAA yields relative to total RNA in each reaction were normalized to WT fractional yields (100%) in the same experiment.

**Identifying candidate E1–E2 homologs in the PVT1 exon**

The python script for ranking potential E1-E2 segments in genes encoding RNAs to be circularized is provided in Sup-
Figure 1. Permuted T4 td group 1 intron RNA secondary structure and activity. Secondary structure of RNA transcribed from generic PIIE T4 td group 1 intron construct. 5’ and 3’ intron halves, the 5’ splice site, 3’ splice site, and site of end-joining (black arrows), E1, E2 and the internal guide sequence (IG) are indicated. RNA is cleaved at the 5’ and 3’ splice sites and E1 and E2 termini are joined in sequential transesterification reactions, thereby producing circRNA and two linear byproducts (i.e. the 5’ and 3’ intron halves). Predicted local structures of T4 td mRNA and *Anabaena* tRNA*{\text{Leu}}*, together with respective E1–E2 segments and sites of end-joining, are also depicted (bottom).

**Supplementary Data.** Though this approach was specifically applied to permute the *PVT1* exon for RNA circularization, the script may be easily adapted for use with any RNA to be circularized. In brief, the encoded algorithm scored every 8-nt segment in the *PVT1* exon sequence for homology to the native T4 td E1–E2 sequence, indexed qualifying segments by starting nucleotide number, and outputted a list of index, sequence, and score triples in a comma-separated values (.csv) file readable and sortable using Microsoft Excel or other spreadsheet software. Scoring attributions used in the original script (provided) are as follows: (i) nucleotides at positions E1:6 and E2:1 of the candidate segment must be ‘T’ and ‘C’ respectively to qualify for inclusion in the outputted list, (ii) 3 pts are added to the segment score for each of E1 positions 1–5 and E2:2 that precisely matches the corresponding position in native E1–E2, (iii) 1 pt is added for each of E1:1–5 in which the respective non-WT candidate nucleotide would be expected to maintain basepairing with the WT IG. Because nucleotides at E2:3 and E2:4 were not expected to contribute to intron stability by base-pairing, they were excluded from consideration in this analysis. Once a suitable E1–E2 segment was identified, the sequence en-
coding *circPVT1* was permuted around this segment and the recombinant form embedded into the PIE transcription templates as described above.

**Construction of T4 td PIE transcription templates with randomized E1 or E2**

Transcription templates for generating circular streptavidin aptamer (cSAA) RNA for *in vitro* selection experiments were assembled from overlapping or partially complementary forward and reverse PAGE Ultramer DNA oligo primers (Integrated DNA Technologies, Coralville, Iowa): cSAA-frag1(F), 5′-phosphorylated cSAA-frag2(R), 5′-phosphorylated cSAA-frag3(F) and cSAA-frag4(R) components. Alternative forms of cSAA-frag3(R) were randomized at positions corresponding to the E1 or E2 segments, and a third variant contained both a ‘CC’→‘GG’ substituted IG sequence and a randomized E1. 100 pmol of cSAA-frag1(F) and cSAA-frag3(F) were hybridized to cSAA-frag2(R) and cSAA-frag4(R), respectively, by heating at 90°C for 2 min and slow cooling to 25°C (1° per 5 s) in 1× NEBuffer 2 with 200× dNTPs. Partial duplexes were filled out by incubation with 10 U Klenow Fragment (NEB) for 30 min at 25°C, and these reactions terminated by addition of EDTA to a final concentration of 10 mM. Blunt ended, hemi-phosphorylated duplexes were column-purified (Monarch DNA & PCR Cleanup – 5 μg, NEB), then fractionated by electrophoresis through a 2% non-denaturing agarose gel (Monarch, NEB) and cut out and purified from the gel. Phosphorylated strands of each duplex (cSAA-Klen12 and cSAA-Klen34) were removed by treatment with phage λ exonuclease (NEB) and the remaining strands (cSAA-ss12(F) and cSAA-ss34(R)) fractionated over an 8% denaturing polyacrylamide/urea gel (1× TBE, 7 M urea) and purified by UV shadowing. Partially complementary ssDNAs were hybridized, filled out using Klenow fragment, and the full-length semi-randomized transcription templates purified as described above. Each semi-randomized transcription template (100 fmol) was amplified by 20 cycles of PCR prior to use in an *in vitro* transcription reaction.

**E1 and E2 *in vitro* selection experiments and analysis**

RNA transcribed from E1 or E2 randomized PIE-cSAA transcription templates was fractionated over a 2% non-denaturing agarose gel in lanes adjacent to products of an unmodified PIE cSAA control (WT). Product(s) of the former reaction migrating to the same position as cSAA produced in the latter reaction were excised from the gel, purified (Monarch DNA Gel Extraction Kit, NEB), and subjected to reverse transcription followed by inverse-PCR (RT-iPCR), which should only produce an amplicon from cSAA RNA (not linear precursors or artifacts). Indexed Illumina sequencing adapters were ligated to amplicon libraries using an abbreviated version of NEBNext (NEB) where agarose gel purification was used for size selection, and the resulting libraries sequenced on an Illumina MiniSeq using a 300 cycle mid-output kit. Due to the short length and low diversity of the amplicon library, paired-end sequencing was abbreviated (2 × 80) and 50% PhiX (Illumina) was included in the sample mixture. Custom python scripts were used to identify sequences derived from cSAA containing a 6 nt E1 or a 4 nt E2, bin them by E1 or E2 sequence, and output them into a comma-separated values (.csv) file. Data were further analyzed and the graphs of Figures 4, 5 and 7 generated using Microsoft Excel.

**RESULTS**

**Selecting a PIE model system**

PIE transcription templates commonly utilized for *in vitro* circRNA synthesis employ group I intron sequences from either the *Anabaena* tRNA<sub>Leu</sub> gene or the thymidylate synthase (td) gene of T4 bacteriophage, as both derivatives have been demonstrated to promote efficient RNA circularization (11,13). However, for reasons to be discussed, a framework incorporating the latter intron was used for this study, the previously established secondary structure of which is depicted in Figure 1 (16). Relative positioning and base-pairing interactions among the two intron halves (5H, 3H), E1 and E2, IG nucleotides, and the RNA to be circularized (RC) are shown, as are the terminal homology arms added to facilitate folding and stabilize the intron (13). RNA circularization occurs via tandem transesterification reactions wherein (i) a guanosine nucleophile inserts at the RC–5H junction (5′SS) and displaces the RC 3′ terminus, which itself then (ii) inserts at the 3H-RC junction (3′SS), displacing the intron 3H 3′ terminus as it ligates to the RC 5′ terminus, thereby forming the circRNA.

The E1 and E2 segments are unique in this process in that they are both incorporated into nascent circRNA and involved in base-pairing interactions with intron sequences thought to be essential for catalytic activity. More specifically, the 6 nucleotides of E1 hybridize to the IG segment, and two base pairs are formed between E2 and the IG-proximal loop. The degree to which E1 and E2 sequences can be varied while still optimally supporting catalytic events required for circRNA synthesis is one of the central questions addressed in this work.

Another important consideration in choosing a framework for broadly applicable PIE construct design is the de-
gree to which group 1 intron folding and catalytic activity are affected by the sequence and structure of the RNA to be circularized. As a proxy for this information, we have modeled exonic base-pairing interactions proximal to the respective ligation junctions formed by self-splicing of *Anabaena* or T4 group I introns in their native contexts (Figure 1, bottom). Whereas the td mRNA is predicted to have some local structure—a short A:U-rich stem flanked by an 8-nt internal loop, the structure of the *Anabaena* tRNA*Leu* split product, like all tRNA, is predicted to be extremely stable and probably contributes significantly to intron folding stability. Indeed, ‘tRNA scaffolds’, comprised of tRNA halves placed at opposite ends of RNA transcripts, have previously been engineered into RNA constructs to increase the stability and structural homogeneity of embedded RNA (17). This likely dependence of *Anabaena* tRNA*Leu* intron activity on exon structure ran contrary to our goal of developing a PIE construct that works well regardless of the RNA to be circularized. Consequently, we chose the T4 td group I intron to serve as the foundation for the model system developed in this study.

**Identifying and exploiting sequences that resemble E1-E2 in circPVT1**

A primary rationale for developing a robust, genetically flexible PIE framework is to synthesize perfect sequence mimics of human circRNAs for structural and functional analysis. Toward this end, we engineered a transcription template for synthesis of circPVT1, a 410-nt human circRNA shown to sequester let-7 miRNA and promote cellular proliferation while inhibiting senescence (14).

Because circPVT1 in its native context is produced by backsplicing across a *PVT1* gene exon, we initially considered simply inserting the unmodified exon sequence directly into our T4 td PIE construct. However, the circRNA produced from such a template would contain extraneous T4-derived E1–E2 sequence and therefore the sequence match with native circPVT1 would be imperfect. Alternatively, we could have removed E1 and E2 from the backbone sequence of our construct, effectively replacing these elements with terminal nucleotides of the inserted *PVT1* exon. This strategy is also flawed, since the terminal nucleotides of the native *PVT1* exon would not be expected to correctly hybridize to the IG sequence of the T4 td group I intron and catalytic activity would probably be impaired.

Our solution to this problem exploits the universal absence of 5′ and 3′ termini in circRNA. Specifically, we engineered a construct for efficient synthesis of a perfect circPVT1 mimic by permuting the linear *PVT1* exon sequence such that the 5′ and 3′ terminal sequences closely match those of T4-derived E2 and E1, respectively. Toward this end, we computationally searched for suitable segments in the *PVT1* exon and ranked them by how closely they resembled the E1–E2 sequence. From this analysis, we determined that the *PVT1* exon sequence segment comprised of nt 324–331 matched E1–E2 almost perfectly, i.e. nt 324–329 are identical to E1, while nt 330–331 match the other 5′ nucleotides of E2.

General features of this construct design strategy, highlighting specific aspects of *PVT1* exon permutation, are illustrated in Figure 2A. The depicted transcription template was commercially synthesized and cloned into a plasmid vector, then amplified by PCR, the products of which were included in *in vitro* transcription and circularization reactions as described in Materials and Methods. RNA from these reactions was purified from other reaction components and either directly fractionated over a 2% non-denaturing agarose gel or first subjected to digestion with RNase R to remove linear RNA (Figure 2B). Collectively, the migration rates of the various RNA products and byproducts and their relative sensitivity to RNase R indicate that we were successful in producing circPVT1 from our genetically permuted template. Moreover, sequence correctness of circPVT1 RNA, including at the artificial site of end-joining, was verified by RT-PCR and next generation sequencing. Collectively, our results closely resemble those previously obtained using similar PIE constructs to produce circRNA as a vector for protein expression (13), though in this case, there was no need to tailor E1–E2 to precisely match the sequence of the RNA being circularized (or vice versa).

To demonstrate the utility of precision, high yield circRNA synthesis *in vitro*, we investigated the structure of our synthetic circPVT1 by atomic force microscopy. Products of *in vitro* transcription and circularization reactions were treated with RNase R and processed for microscopy as described in Materials and Methods. The images generated depict various conformations assumed by this biologically important circRNA (Supplementary Figure S1). Such results could not easily be obtained by isolating and purifying native circPVT1 from human cells, as low yield and representation among total cellular RNA would be prohibitive.

**Influence of mutations in E1 and E2 on T4 td group 1 intron activity**

Though our efforts to produce a perfect sequence-mimic of human circPVT1 from a designer PIE transcription template were successful, we reasoned that because many human circRNAs would not contain a sequence segment closely matching native T4-derived E1–E2 sequence, our permutation-based design strategy could not be universally applied. Therefore, we decided to test the limit to which E1 and E2 sequences in PIE constructs could be altered while still retaining group 1 intron self-splicing activity. To achieve this, we adopted a previously reported PIE model system (11) designed to produce a hybrid circRNA containing the sequence of a streptavidin aptamer (cSAA) flanked by native T4 td exonic sequence, including E1 and E2 (Figure 3A). Given its small size relative to the circPVT1-PIE construct, cSAA-PIE transcription template variants are less expensive and easier to manipulate genetically. Moreover, the secondary structure of the streptavidin aptamer is known, allowing us to make more reliable inferences as to how the structure of the RNA to be circularized affects PIE structure and function.

Our initial approach to this investigation was to conduct a systematic mutational analysis of every nucleotide position in E1 and E2. Transcription templates containing every possible single-nucleotide substitution in these regions...
Figure 2. Design and products of circPVT1 synthesis reactions. (A) Arrangement of DNA sequences encoding transposed T4 td group 1 intron halves, the permuted PVT1 exon, and a T7 promoter to generate a PIE circPVT1 transcription template. (B) Products of in vitro transcription and circularization reaction. M1, NEB 100 bp DNA ladder; CP, transcription/circularization reaction products; RNase R −/+; RNase R-untreated or -treated transcription or circularization reaction products; M2, NEB 50 bp DNA ladder. RNA band assignments are based upon those made for products of similar circRNA synthesis reaction previously described. The correct circPVT1 RNA sequence was confirmed from RNase R-treated reaction products by RT-iPCR and NGS.

were synthesized, amplified by PCR, and used in transcription and circularization reactions as described in Materials and Methods. Individual circRNA quantities were recorded as a fraction of total RNA fluorescence in individual reactions, and these fractions were normalized to the equivalent values calculated for circRNA produced from a WT construct containing native E1 and E2 sequences. Used as a negative control, the NEG construct contains E1 and E2 sequences ('AAUAAA' and 'AUAA', respectively) that cannot base pair with IG and the IG-proximal loop, respectively, in the manner predicted in the secondary structure model.

When produced from the WT construct, cSAA migrates as a discrete band and ahead of all circularization reaction precursors and byproducts (Figure 3B, C, D; WT). In contrast, apparent circularization products generated from the NEG template are diffuse (Figure 3B, C, D; NEG), suggesting that transesterification reaction sites and sites of end-joining are highly imprecise. The effects of specific E1 and E2 point mutations are also shown (Figure 3C, D). Although the effects of these mutations did not always correlate with predicted effects on intron structure, some tentative inferences regarding how these sequence changes affect RNA circularization can be drawn. For instance, most point mutations at E1 positions 1 and 2 moderately affect circRNA synthesis (Figure 3C), reducing yield (E1:1C, E1:1G, E1:2A, E1:2G), altering the migration position of the primary product (E1:1G), or giving rise to multiple circRNA products (E1:2C). The latter two effects suggest that the mutations in question may displace one or both transesterification sites, and therefore the site of end-joining, 1–2 nucleotides in either direction. Surprisingly, circRNA yields from E1:3 mutant constructs are all approximately twice that obtained from WT, though the reasons for this increased activity are unclear. Conversely, any nucleotide substitution at E1:4 or a G→A change at E1:5 significantly reduces circRNA yield, suggesting that the native G-C pairs involving these positions are important for maintaining the structure and activity of the group 1 intron. Similarly, E1:6A construct yield is only 28% of that from WT, and the E1:6G product appears heterogeneous. However, E1:6C generates large amounts of homogeneous circRNA, indicating that a U→C substitution at this position is well tolerated. As a potential explanation for this observation, it is worth noting that of the three E1:6 mutations, only E1:6C would be expected to preserve base-pairing with ‘G’ predicted in the secondary structural model.

E2:1U is remarkable among substitutions at the E2:1 position (Figure 3D). Specifically, it is the only one of these variants that would be expected to maintain pairing with ‘G’, and our quantitative analysis suggests that its activity is 164% relative to WT. Conversely, none of the E2:2 substitutions are observed to have any significant effect. CircRNA
Figure 3. Assessment of cSAA products and yield from PIE construct E1 point mutants. (A) Pre-circularization secondary structure of the streptavidin aptamer-T4 td fusion RNA (SAA), with intron segments E1, E2, IG sequence, and splice sites (lightning bolts) also depicted. (B) Transcription/circularization products from WT (native E1, E2, IG) and NEG (mutant E1) constructs fractionated over a 4% agarose gel. The migration position of the circRNA product (cSAA) is indicated. (C and D) Mutant descriptions, including predicted base-pairing with IG, cSAA yield, and synthesis profiles for the indicated WT and E1 and E2 point mutant constructs. Fractional cSAA yields (arrows) relative to total RNA in each reaction are normalized to WT fractional yields (100%) in the same experiment.

yield from the E2:3C and E2:3U constructs are significantly reduced relative to WT, perhaps due to non-native base-pairing with the ‘G’ two nucleotides into the IG-proximal loop. Such an interaction might be expected to alter or stabilize a reaction intermediate, or possibly displace one or both transesterification sites. The latter effect would in turn displace the site of end-joining and alter the lengths of the respective circRNAs, a result consistent with the different migration positions observed for the respective E2:3 reaction products. Finally, the greatest difference in circRNA production between constructs with different mutations at the same position is observed in the E2:4G and E2:4U reactions, with yields of 194% or 19% relative to wild type, respectively. How such seemingly inconsequential mutations can produce such pronounced and opposite effects remains unclear, though respective inhibition or stabilization of unproductive alternative PIE conformations may contribute to these phenotypes.
Using in vitro selection to characterize E1 and E2 sequence restrictions

More in depth characterization of E1 and E2 sequence dependencies required evaluation of PIE transcription template variants containing more than one nucleotide substitution in these regions. Since conducting such an analysis using individually synthesized constructs would be impractical, we instead devised an in vitro selection method to determine the functionality of all 4096 possible E1 and 256 possible E2 sequence permutations simultaneously. In brief, transcription template libraries were assembled from Ultramer oligonucleotides (IDT) in which the sequence segments encoding E1 or E2 were randomized. These libraries were included in in vitro transcription and circularization reactions in the manner described for individual transcription templates. RNA products of these reactions were subjected to RT-iPCR using primers designed to amplify only circRNA containing an E1-E2 junction. Sequencing adapters were ligated to cDNA-derived amplicons by NEBNext (NEB) and the resulting libraries sequenced on an Illumina MiniSeq. Custom Python scripts were used to identify reads containing circRNA sequences flanking E1-E2, collect and bin E1 or E2 sequences recorded in these reads (whichever had been randomized), and output the results in a comma-separated values file (.csv).

The approach and results of our E1-randomization experiment are summarized in Figure 4. From the 25,777 reads containing circRNA sequence with a 6 nt E1, 717 different functional E1 sequences were identified. Not only do these variants comprise only 18% of all possible E1 sequence permutations simultaneously. In brief, transcription template libraries were assembled from Ultramer oligonucleotides (IDT) in which the sequence segments encoding E1 or E2 were randomized. These libraries were included in in vitro transcription and circularization reactions in the manner described for individual transcription templates. RNA products of these reactions were subjected to RT-iPCR using primers designed to amplify only circRNA containing an E1-E2 junction. Sequencing adapters were ligated to cDNA-derived amplicons by NEBNext (NEB) and the resulting libraries sequenced on an Illumina MiniSeq. Custom Python scripts were used to identify reads containing circRNA sequences flanking E1-E2, collect and bin E1 or E2 sequences recorded in these reads (whichever had been randomized), and output the results in a comma-separated values file (.csv).

The approach and results of our E1-randomization experiment are summarized in Figure 4. From the 25,777 reads containing circRNA sequence with a 6 nt E1, 717 different functional E1 sequences were identified. Not only do these variants comprise only 18% of all possible E1 sequences, but representation among them is highly skewed, indicating of a high degree of selection (Figure 4B). To this point, the collective representation of the 10 most frequently observed E1 variants exceeds 17% of the total number of reads, while more than half of the selected variants were observed only once (285) or twice (78). Collective analysis of nucleotide identity by position reveals the basis for this selection (Figure 4C, D). Specifically, ‘G’ is highly favored at E1 positions 4 and 5, with collective weighted representations among functional variants of 79% and 99%, respectively. These observations strongly suggest that pairing of these nucleotides with the ‘CC’ motif in IG is critical for maintaining the E1–IG interaction and thus PIE function. Similarly, in an observation consistent with our single-substitution experiments, only ‘U’ or ‘C’ is well-tolerated at E1:6, with respective representations of 67% and 31%, and both nucleotides are capable of pairing with ‘G’ in the IG sequence.

Selection of E2 variants is also skewed, but not to the degree observed with E1 (Figure 5B). Specifically, 249 of the 256 possible E2 variants are identified among the 99 355 reads derived from circRNA. This assessment is further supported by analysis of nucleotide identity by position, where little selection is observed at E2 positions 2–4 (Figure 5C, D). An exception to this overarching observation, however, is that ‘G’ is relatively disfavored at every E2 position, perhaps because its tendency toward promiscuous base-pairing would promote formation of or stabilize unproductive intron conformations or reaction intermediates. Indeed, the only seven E2 variants not represented among the nearly 10^5 sequence reads are 5′-GAGA-3′, 5′-GAGG-3′, 5′-GUAA-3′, 5′-GCCA-3′, 5′-GGCG-3′, 5′-AGGG-3′ and 5′-GUGG-3′. Positive selection pressure for E2 position 1 is also observed, where ‘U’ (46%) and ‘C’ (41%) are dominant. As with E1:6, either ‘U’ or ‘C’ at E2:1 would be expected to support a putative base-pairing interaction with ‘G’ at the intron active site.

Increasing or restoring E1 mutant functionality by making compensatory changes to IG

Of the 10 best represented variants identified in E1 and E2 selection experiments, none exactly matched the respective WT sequences. Moreover, most of the former variants would be expected to incompletely base-pair with IG. To further probe the tolerance for genetic flexibility in E1 together with the importance of E1-IG base-pairing, we measured circRNA yield from transcription templates containing three of the best represented E1 variants in the selection experiment or the same variants plus compensatory changes in IG (Figure 6). The result of this analysis indicates that the activities of the selected variants was at least 47% of WT, indicating that in vitro selection is a reasonably good proxy for activity. Perhaps most notably, the activity of the s7 variant was 148% of WT, despite being predicted to pair with only three of six nucleotides in IG. We also find that making compensatory changes to restore E1-IG base-pairing unexpectedly but consistently decreases the functionality of these variants, indicating that this approach to PIE structural engineering can have unpredictable and detrimental effects on function. Notably, however, of the seven compensatory changes collectively introduced into the three constructs, only one affected the tandem G–C pairs predicted to be important in previous experiments (Figures 3 and 4).

Both single-mutation and in vitro selection experiments have demonstrated that base-pairing between ‘G’s at positions 4 and 5 of E1 and ‘C’s in IG are important for intron function. However, these experiments did not determine whether the polarity of this interaction was also critical, or could be reversed, as might be desirable for synthesis of some precisely tailored circRNAs. To investigate this possibility, we conducted an in vitro selection experiment using a construct containing ‘C’→’G’ substitutions at adjacent sites in IG expected to pair with nucleotides at positions 4 and 5 in E1, together with a randomized E1 sequence.

For reasons that are unclear, only 487 reads generated by Illumina sequencing contained sequence consistent with having been derived from circRNA with a 6 nt E1, among which only 159 of 4096 possible E1 variants are represented. However, even in this relatively small sample, analysis of nucleotide identity at each position reveals a pattern closely resembling that observed in our first E1 experiment except that strong selection at E1:4 and E1:5 is for ‘C’, not ‘G’ (Figure 7A). In further support of this observation, all ten of the best represented variants contain ‘CC’ at E1 positions 4 and 5 (Figure 7B), and the activities of the top two among these are at least 74% of that of a fully WT construct (Figure 7C). Again, however, efforts to augment this activity by introducing compensatory mutations into IG were unsuccessful. Collectively, tolerances revealed in these experiments further expand the options for mandatory sequence incorporation into engineered circRNA.
Figure 4. Selection of cSAA sequences from E1-randomized PIE construct libraries. (A) Local structural depiction of possible RNA precursors from E1-randomized PIE construct libraries. (B) Frequencies with which E1 variants were observed among 25,777 sequencing reads ordered by frequency. The embedded table lists the 10 most frequently identified E1 sequence variants. (C) Fractional nucleotide composition in qualifying sequence reads by E1 position. Qualifying reads contain a 6-nt E1 and flanking sequences consistent with RT-iPCR of cSAA. Consensus nucleotide selections are ordered by the frequencies with which they occur in capital (40% or more) or lower case (20% or more) letters. (D) Bar graph representation of the table in C.

Effects of circRNA structure on PIE function

Because the 5′ and 3′ intron halves are located at opposite termini in PIE constructs, assembly of a functional ribozyme is likely to be at least somewhat dependent upon folding of the RNA to be circularized. To control for this potential dependence, we utilized a previously described model construct designed to produce a compact circRNA of known structure together with a mostly unpaired segment comprised of native T4 td sequence, including E1 and E2 (11). Despite these efforts, we consistently found that our circRNA mass yield from this construct (by ethidium bromide staining) was only 5–10%, equivalent to circularization efficiencies of 25–50%. Upon further analysis, we considered that the lack of structure predicted for the T4 td segment in this construct did not recapitulate base-pairing interactions normally assumed by the native sequence, probably due to non-native base-pairing with ‘CGG’ in the engineered segment (Figure 8). To determine whether imposing a structure more like that of wild-type in this region might increase PIE activity, we designed and tested the variants shown in Figure 8. Indeed, whereas the activity of v2 was comparable to that of the original PIE construct, the relative activity of v3, the variant designed to fold precisely like T4 td mRNA, was more than 4.5-fold greater (468%) and indicative of a circularization efficiency approaching 100%. Together, these results indicate that the structure of the RNA to be circularized can be an important determinant of circRNA yield.

DISCUSSION

Base-pairing among E1, E2 and IG is a highly conserved feature of group 1 intron structure that helps establish and maintain the catalytic active center while facilitating the transition between transesterification reactions (18). Moreover, because E1 and E2 are incorporated into PIE construct reaction products, the limits of genetic flexibility in circRNA synthesis are determined by the length and composition of these elements. E1 and E2 segments associated with the T4 td intron are relatively small, with only 6 and 2 base pairs predicted to form between these elements and IG sequences, respectively (Figure 1). For comparison, the *Tetrahymena* group 1 intron forms 6 E1-IG and 7 E2-IG base-pairs in this region (19), suggesting that PIE derivatives thereof would impose a much heavier burden of sequence restriction. Hence, as a foundation for designing tailored PIE constructs for synthesis of circRNAs of virtually any sequence, the T4 td group I intron is close to ideal.

Our detailed development of this model system demonstrates that circRNA sequence restrictions imposed by E1 and E2 can be reduced even further, though which specific
Figure 5. Selection of cSAA sequences from E2-randomized PIE construct libraries. (A–D) Format of in vitro selection results obtained from E2-randomized PIE construct library is equivalent to that described in Figure 4.

In general, tolerated variations can be summarized succinctly as 5′-NNNGGY-3′ and 5′-YNNN-3′, respectively, and more precisely according to the results presented in Figures 4 and 5. It is perhaps predictable that the identities of nucleotides adjacent to the site of end joining are restricted, and that the tandem G-C pairs central to the E1-IG base-pairing interaction must be preserved for optimal intron function. Despite these restrictions, we were surprised by the relative flexibility of nucleotide composition in more distal portions of E1 and E2, given the seeming importance of native nucleotides at these positions in maintaining intron structure. That ‘G’ was disfavored at all E2 positions is another noteworthy observation, perhaps indicating that this nucleotide tends to promote alternative intron conformations that either slightly shift the active center or are not at all compatible with catalysis. Our inability to increase the functionality of several E1 mutants by introducing compen-
Figure 7. Selection and activity of E1 variants in an IG mutant background. (A) Fractional nucleotide composition by E1 position among qualifying sequence reads. Reads were generated in an E1 in vitro selection experiment conducted using an E1-randomized library with a ‘GG’ IG background. Results are presented in both tabular and graphical forms. (B) List of the 10 most frequently represented E1 variants among qualifying sequence. (C) Sequence composition and results obtained from synthetic PIE constructs containing native E1 and IG sequences (WT, WT), native E1 and ‘GG’ IG (WT, GG), the two best-represented variants isolated in the in vitro selection experiment (s1a, GG; s1b, GG), or these variants plus compensatory mutations in IG (s1a, GGc; s1b, GGc). Constructs were processed and analyzed as described for Figures 3 and 6.

Compensatory changes designed to restore E1-IG base-pairing likewise demonstrates that the structural effects of even individual RNA nucleotide changes can be complex and challenging to predict. The consistent selection of ‘CC’ at E1 positions 4 and 5 in a ‘GG’ IG mutant background represents a notable exception to this finding and shows that base-pairing composition but not polarity in this interaction is important for intron function. In more practical terms, this determination reveals yet another option for circumventing E1 sequence restrictions in PIE construct design.

Although we have interpreted many of our observations in the context of predicted interactions among E1, E2, and IG, some findings are difficult to reconcile on this basis. For instance, several constructs mutated at individual sites in E1 or E2 exhibit phenotypes inconsistent with the roles predicted for these nucleotides in structural models (Figure 3). More significantly, we discovered from experimental results shown in Figure 8 that the structure of the RNA to be circularized can have a marked effect on circRNA synthesis yield. We have also found this to be the case in ongoing circRNA synthesis projects, though when necessary, we have so far successfully implemented the principles and strategies outlined here to redesign PIE constructs and improve synthesis yield. More broadly, we believe RNA sequences that naturally bring the sites of end joining into proximity during co-transcriptional folding are also likely to facilitate proper PIE folding and thus increase circularization efficiency. This tendency can either be a product of structural engineering or a natural feature of the RNA being studied. Indeed, with respect to endogenous human circRNAs, it has been postulated that RNA structure plays an important role in determining which exonic RNA segments are prone to form circles (20). Conversely, the circularization efficiency of highly structured PIE-derived circRNA precursors that do not bring the sites of end-joining into proximity will likely be extremely low.

In this work, we provide guidelines for design of PIE constructs to synthesize circRNA of virtually any sequence, together with a permutation strategy by which the synthesis yields from precursors engineered to have different end-joining junctions may be tested and optimized. A generalized workflow for application of these guidelines is provided in Supplementary Data. Using these strategies, it is now possible to synthesize circRNA nanoparticle precursors easily and with high efficiency, as well as perfect sequence mimics of naturally occurring forms of circRNA. Moreover, given the high yields and purity achievable through in vitro circRNA synthesis, structural features of these RNAs can now be investigated in ways not previously possible, including atomic force microscopy, SAXS, NMR, and even X-ray...
Figure 8. Dependence of cSAA yield on the structure of the RNA to be circularized. Secondary structural models of local T4 td mRNA, the original cSAA-PIE construct, and two mutant variants (v2, v3) are depicted. SAA and T4 td sequences are shown in blue and red, respectively. Sequences introduced solely for the purpose of altering RNA structure are shown in gray. Each of the structures represent the forms of RNA that exist after group 1 intron excision and reflect the lowest energy conformation produced by using RNAstructure software. Construct synthesis, in vitro transcription and circularization, RNA fractionation, and circRNA quantitation are as described in Materials and Methods and for Figure 3.

crystallography. Synthesis yields should also permit transfection of circRNA into living cells using established RNA delivery vehicles (21), thereby providing an opportunity to analyze the cellular functions of their native counterparts. This approach can potentially be expanded by randomly incorporating modified nucleotides with different functionalities (e.g. biotin, fluorescein) into nascent RNA circularization products, or alternatively, vectors may be designed for direct PIE-based synthesis of circRNA in eukaryotic cells in a manner not dependent upon cellular splicing machinery (22). In summary, we present here a means of increasing the flexibility and utility of circRNA synthesis for incorporation into nanoparticles, probing the structure and function of native variants, and other applications that have yet to be fully explored.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

J.W.R., C.S., W.F.H. and S.F.J.L.G. were supported by the Intramural Research Program of the National Cancer Institute, NIH; M.G. and M.J.P. were supported by the NIA IRP, NIH; Federal funds from the National Cancer Institute, National Institutes of Health [75N91019D00024]; Intramural Program of the NIH, NCI, Center for Cancer Research. Funding for open access charge: Frederick National Laboratory for Cancer Research.

Conflict of interest statement. The contents of this publication do not necessarily reflect the views of the Department of Health and Human Services and mention of trade names, commercial products or organizations does not imply endorsement by the US government.

REFERENCES

1. Kristensen, L.S., Andersen, M.S., Stagsted, L.V.W., Ebbesen, K.K., Hansen, T.B. and Kjems, J. (2019) The biogenesis, biology and characterization of circular RNAs. Nat. Rev. Genet., 20, 675–691.
2. Salzman, J., Gawad, C., Wang, P.L., Lacayo, N. and Brown, P.O. (2012) Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One*, 7, e30733.
3. Salzman, J., Chen, R.E., Olsen, M.N., Wang, P.L. and Brown, P.O. (2013) Cell-type specific features of circular RNA expression. *PLoS Genet.*, 9, e1003777.
4. Yu, C.Y. and Kuo, H.C. (2019) The emerging roles and functions of circular RNAs and their generation. *J. Biomed. Sci.*, 26, 29.
5. Ren, S., Lin, P., Wang, J., Yu, H., Lv, T., Sun, L. and Du, G. (2020) Circular RNAs: promising molecular biomarkers of human aging-related diseases via functioning as an miRNA sponge. *Mol. Ther. Methods Clin. Dev.*, 18, 215–229.
6. Abe, N., Abe, H., Nagai, C., Harada, M., Hatakeyama, H., Harashima, H., Ohshiro, T., Nishihara, M., Furukawa, K., Maeda, M. *et al.* (2011) Synthesis, structure, and biological activity of dumbbell-shaped nanocircular RNAs for RNA interference. *Bioconjug. Chem.*, 22, 2082–2092.
7. Jasinski, D., Haque, F., Binzel, D.W. and Guo, P. (2017) Advancement of the emerging field of RNA nanotechnology. *ACS Nano*, 11, 1142–1164.
8. Chu, F.K., Maley, G.F., Maley, F. and Belfort, M. (1984) Intervening sequence in the thymidylate synthase gene of bacteriophage T4. *Proc. Natl. Acad. Sci. U.S.A.*, 81, 3049–3053.
9. Xu, M.Q., Kuczyk, S.D., Goodrich-Blair, H., Nierzwicki-Bauer, S.A. and Shub, D.A. (1990) Bacterial origin of a chloroplast intron: conserved self-splicing group I introns in cyanobacteria. *Science*, 250, 1566–1570.
10. Pervouchine, D., and Been, M.D. (1992) Group I permuted intron-exon (PIE) sequences self-splice to produce circular exons. *Nucleic Acids Res.*, 20, 3537–3564.
11. Litke, J.L. and Jaffrey, S.R. (2019) Highly efficient expression of circular RNA aptamers in cells using autocatalytic transcripts. *Nat. Biotechnol.*, 37, 667–675.