Published online 6 November 2013 *Nucleic Acids Research, 2014, Vol. 42, No. 4* ***e22***
doi:10.1093/nar/gkt1088

High-throughput, cost-effective verification of structural DNA assembly

Yandi Dharmadi, Kedar Patel, Elaine Shapland, Daniel Hollis, Todd Slaby, Nicole Klinkner, Jed Dean and Sunil S. Chandran*

Amyris, Inc., 5885 Hollis Street, Suite 100, Emeryville, CA 94608, USA

Received April 23, 2013; Revised September 12, 2013; Accepted October 16, 2013

ABSTRACT

DNA ‘assembly’ from ‘building blocks’ remains a cornerstone in synthetic biology, whether it be for gene synthesis (~1 kb), pathway engineering (~10 kb) or synthetic genomes (>100 kb). Despite numerous advances in the techniques used for DNA assembly, verification of the assembly is still a necessity, which becomes cost-prohibitive and a logistical challenge with increasing scale. Here we describe for the first time a comprehensive, high-throughput solution for structural DNA assembly verification by restriction digest using exhaustive *in silico* enzyme screening, rolling circle amplification of plasmid DNA, capillary electrophoresis and automated digest pattern recognition. This low-cost and robust methodology has been successfully used to screen over 31 000 clones of DNA constructs at <$1 per sample.

INTRODUCTION

As a scientific field, synthetic biology seeks to provide the rigor of an engineering framework to the modification of living organisms through modularization, characterization and standardization (1–3). The central tenet is that well-characterized DNA building blocks could be assembled into larger constructs of predictable biological functions according to standard modular design (4,5). Applications of DNA assembly technologies vary according to final construct size, from ~1 kb for gene synthesis (6,7), ~10 kb for pathway engineering (8–11), to >100 kb for synthetic genomes (12,13). Any assembly method is susceptible to an inherent failure rate and thus multiple clones of the construct need to be screened for verification. Typical methods reported in the literature include analysis of restriction digest or PCR over junctions between individual building blocks by agarose gel electrophoresis, and Sanger sequencing (10,12,14).

We employ synthetic biology for high-throughput rational strain engineering and design and manufacture hundreds to thousands of unique building blocks and assemblies every month. In this context (parallel assembly of DNA constructs at increasing scale), a verification method that is cheap, fast and reliable is absolutely essential. Verification of thousands of DNA constructs by sequencing or junction-PCR quickly becomes cost-prohibitive and/or logistically intractable due to multiplicity of reactions (number of constructs × number of clones × number of junctions).

As a cost-effective yet reliable alternative, we have developed a comprehensive strategy for a high-throughput restriction digest assay, consolidating recent technological advances in the field with custom designed computational framework on our part. We first addressed the selection process of the enzyme used for digesting a huge library of constructs. As will be described in detail, given a set of assembly sequences and a set of restriction enzymes, we can devise an exhaustive, unbiased *in silico* screen to select for the most suitable enzyme based on objective frequency metrics reflecting system constraints. Second, we relied on the use of rolling circle amplification (RCA) (15) to amplify plasmid DNA directly from *Escherichia coli* culture thereby side-stepping the otherwise complex workflow for large-scale plasmid minipreps. Third, identification of the various fragment sizes from the digestions was done by capillary electrophoresis technology, which offers superior resolution, quantitation, speed and automation (16–18) as compared to traditional agarose gel slab fragment analysis. Finally, we provide a rigorous derivation of an algorithm for automated processing of digest data (electropherograms), accounting for fragment sizes and molar abundances. Altogether, this novel and robust methodology has been successfully utilized in pre-screening of >31 000 clones (>100 Mbp) at less than a $1 per sample, and is now an integral part of our industrial scale DNA assembly pipeline.

MATERIALS AND METHODS

Structural design, building blocks and assemblies

Figure 1 illustrates the design and creation of plasmid DNA constructs (‘building blocks’ and ‘assemblies’).

*To whom correspondence should be addressed. Tel: +1 510 597 4765; Fax: +1 510 225 2645; Email: chandran@amyris.com

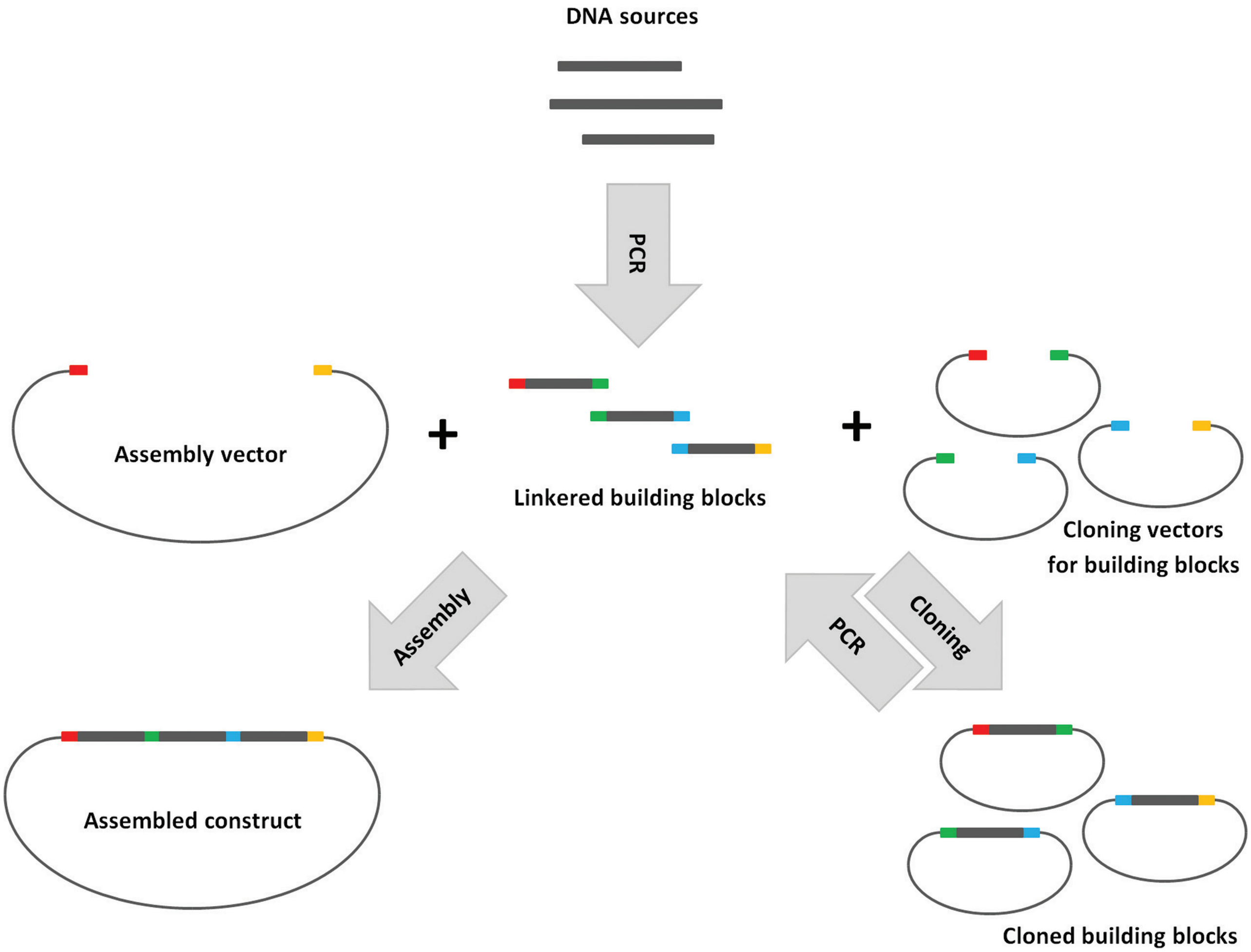

Figure 1. Schematic representation of assembly architecture by overlapping linkers (shown is example of assembly with 3 building blocks). Linear building blocks may be amplified *de novo* using primers with linker overhang, or re-amplified with linker primers from cloned/archived building blocks. Whereas one universal assembly vector with one linker pair accommodates all 5660 assembly designs, 25 unique cloning vectors were required for cloning 2236 building blocks due to combinatorial linkers on 5′ and 3′.

Linear DNA building blocks with 24- to 36-bp linkers at 5′ and 3′ ends were amplified *de novo* from various natural and synthetic sources, or from previously cloned building blocks with specific linkers already in place (linker sequences and usage frequency given in Supplementary Table S1). The PCR products were used in the cloning of 2236 building blocks (0.5–10 kb excluding vector) and construction of 5660 assemblies (1.0–20 kb excluding vector, 2–12 building blocks per assembly) by virtue of overlapping linkers (details on methods included in subsequent sections). Each assembly design corresponds to a specific genotype for yeast pathway engineering, e.g. gene deletion, overexpression or introduction of point mutations, the specifics of which are outside the scope of this discussion. Figure 2 presents the distribution of size and complexity of the constructs: excluding vectors, the average building block is 0.9 kb, while the average assembly is 4.2 kb and has 3.8 building blocks.

Strains, vectors and media

Saccharomyces cerevisiae strain CEN.PK2-1c (19), a tryptophan auxotroph (*MATa; ura3-52; trp1-289; leu2-3,112; his3Δ 1; MAL2-8^{C}; SUC2*), was used as host for DNA assembly by yeast homologous recombination. Yeast growth medium was YPD (1% yeast extract, 2% peptone, 2% dextrose). Complete synthetic medium lacking tryptophan (CSM-W, 2% dextrose) was used as selective medium for yeast outgrowth post-transformation. *Escherichia coli* strain XL1-Blue (Agilent) was used as the host for cloning of DNA building blocks and assemblies. LB (0.5% yeast extract, 1% tryptone, 0.5% NaCl) with 0.1 g/l carbenicillin was used for liquid medium and solid selective medium (with 1.5% agar) for colony formation.

Vectors for assembly and cloning are depicted in Figure 1 (sequences given in Supplementary Sequences S1 and S2). Each building block has 5′- and 3′-linkers specific to the assembly in which it is used, resulting in 25 vectors with unique 5′- and 3′-linker pairs being required for cloning of building blocks. The vectors were derived from *bla*-marked pUC19 (20) with insertion of linkers flanking plac-lacZα ORF in the original sequence. The lacZα is subsequently excluded upon PCR amplification of the vector backbone.

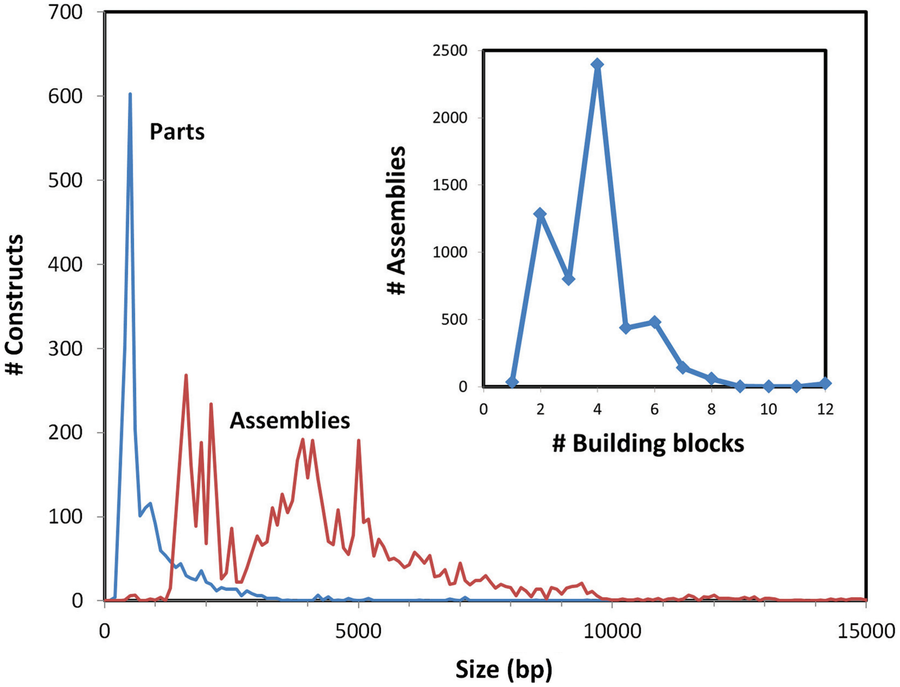

Figure 2. Size distributions of building blocks and assemblies, excluding vectors (bin size = 100 bp). *Inset*: complexity distribution of assemblies based on number of building blocks. Building blocks range from 0.5 to 10 kb (~0.9 kb). Assemblies range from 1 to 20 kb (~4.2 kb). The average assembly has 3.8 building blocks (median = 4 building blocks).

The vector for assembly of building blocks was derived from *TRP1*-marked yeast shuttle vector pRS414 (21) with lacZα ORF in the original sequence disrupted by a pair of linkers, between which assembly of the building blocks would occur.

Automation and instrumentation

All specimens and reagents were handled in 96- or 384-well format. Applied Biosystems thermal cycler model 2720 (Life Technologies) were used for PCR amplifications and high temperature incubations. Liquid transfers (2–200 µl) were done on the Biomek FXP automation workstation (Beckman Coulter). Inoculation of *E. coli* colonies was done using the Qpix 2 automated colony picker (Molecular Devices). Multitron incubation shakers (Infors HT) were used for cultivation of yeast and bacterial culture specimens. Capillary electrophoresis was performed on the Fragment Analyzer^{TM} instrument (Advanced Analytical) equipped with 33-cm, 96-channel capillary array and fluorescence detection.

Building block and vector preparation

All PCR amplifications were performed with Phusion® Hot Start Flex DNA Polymerase (NEB) using manufacturer-recommended conditions. PCR-amplified building blocks were diluted 1:80 in 10 mM Tris-EDTA, pH 8.0 (1× TE) and analysed by capillary electrophoresis on the Fragment Analyzer instrument for verification of fragment size, concentration and purity. Removal of primer dimer, dNTP and PCR buffer was done using the AxyPrep Mag PCR clean-up magnetic beads (Corning) according to manufacturer's protocols. Vectors were prepared by PCR amplification. Assembly vector DNA was pelleted by precipitation (0.1 volume 3M sodium acetate + 1 volume isopropanol) and centrifugation (10 000*g*, 60 min), followed by rinse with 70% ethanol and air drying. The eluted DNA was then purified by gel filtration using Sephacryl S-500HR matrix (Sigma-Aldrich) and 10 mM Tris-HCl, pH 8 + 50 mM NaCl as mobile phase. Cloning vectors for archiving of building blocks were purified by agarose gel electrophoresis and extraction with the 96-well Zymoclean DNA recovery kit (Zymo Research). All final DNA preparations were eluted in 1× TE.

Assembly and cloning

Linkers flanking the building blocks and assembly vector serve as homologous regions for recombinational DNA repair in yeast (11,22–24). For each assembly design, 16 ng of the assembly vector was combined with required building blocks (typically 100 ng each) through programmed liquid transfers. The DNA mixture was transformed into *S. cerevisiae* strain CEN.PK2-1c following standard LiAc/ssDNA/PEG protocol (25). Upon a 2-day outgrowth period in selective medium (CSM-W + 2% glucose), assembled plasmid DNA was recovered using the Zymoprep yeast plasmid miniprep II kit (Zymo Research). Similarly, homologous recombination in *E. coli* facilitated direct cloning of building blocks into the corresponding vectors with matching linkers (26–28).

Escherichia coli XL1-blue competent cells prepared with the Z-competent *E. coli* transformation kit (Zymo Research) were used for chemical transformation of assembled yeast plasmids as well as co-transformation of individual building blocks (typically 300 ng) and corresponding cloning vectors (100 ng). Upon plating and colony formation on LB agar + 0.1 g/l carbenicillin cast in Q-trays (Molecular Devices), four clones of each building block and assembly constructs were picked and inoculated into liquid cultures (LB + 0.1 g/l carbenicillin) for further verification.

RCA and restriction digest

Amplification of plasmid constructs from *E. coli* cultures was done by multiply-primed RCA (15) using commercial buffers and mastermix (MCLAB). Two microliters of overnight *E. coli* culture was added to 5 µl of lysis buffer, then incubated at 96°C for 5 min. After addition of 5 µl mastermix containing the mesophilic phi29 polymerase and random hexamers (15), RCA reaction was allowed to proceed for 16 h at 30°C. For *Bsr*DI digest of assemblies, the RCA reaction was diluted 2-fold with water. For *Afl*II and *Ava*II digest of cloned DNA building blocks, the RCA reaction was diluted 4-fold with water.

Digest formulations and incubation temperatures for *Bsr*DI, *Afl*II and *Ava*II (NEB) are given in Supplementary Table S2. For any digest, a mastermix of buffer, enzyme and BSA (if applicable) was created, then added to the diluted RCA reactions. After addition of enzyme mastermix to diluted RCA products, digest reaction was allowed to proceed for 4 h to ensure complete digestion, followed by heat inactivation for 20 min and finally 10-fold dilution with water. The diluted sample was analysed by capillary electrophoresis on the Fragment Analyzer instrument for verification of digest pattern.

Capillary electrophoresis

Proprietary fluorescent dye and gel matrix rated for analysis of 75–20 000 bp DNA fragments were purchased from Advanced Analytical and used in capillary electrophoresis runs on the Fragment Analyzer instrument (Advanced Analytical). Upon gel priming, a mixture of 0.1 and 10 kb DNA markers (Thermo Scientific) at 0.2 ng/µl each in 1× TE and the diluted sample (PCR product or restriction digest) were sequentially injected (8 kV, 6 s each), then separation was allowed to proceed for 24 min at 8 kV. With priming and injections, total processing time was <37 min per sample. A separate run for a sample mixture of 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.2, 1.5, 2, 3, 4, 5, 6 and 8 kb DNA markers (Thermo Scientific) at 0.15 ng/µl each in 1× TE forms a complete calibration curve (size versus elution time) with 19 data points ranging from 0.1 to 10 kb. Electropherogram of each sample was processed through the ProSize 2 software (Advanced Analytical), producing tabulated sizes (bp) and abundances (ng/µl) of fragments present in the sample.

In silico enzyme screen

In order to verify DNA assembly by restriction digest, several considerations are taken into account, in order of significance:

(1) Cost and logistical efficiency prescribes single digestion, which means one enzyme is to be selected as 'most suitable' with respect to a set of constructs determined *a priori*.
(2) The enzyme should be available commercially.
(3) At minimum, restriction sites in the vector region should linearize the plasmid so that total construct size can be verified. However, more pertinent are cuts within the assembled region, thereby verifying the assembly sequence. As the assemblies are variable in a given set, this means the recognition sequence should appear at some statistically suitable frequency in the population.
(4) Operational constraints suggest that the cuts should not be too frequent (producing extremely small fragments), and not too rare (leaving extremely large fragments). Based on rating of the gel matrix (75–20 000 bp) and suitable voltage balancing speed and resolution (8 kV), 0.1 and 10 kb were chosen as the required lower and upper electrophoresis markers. Thus ideally all digest fragments should fall within this range.
(5) Finally, the enzyme should be amenable to a robust process (e.g. preferably no star activity, no irreversible binding to DNA).

The set of all commercially available restriction enzymes with defined cleavage was retrieved in January 2013 from REBASE (29), resulting in a list of 236 unique sequence/cut sites (Supplementary Table S3). An exhaustive simulation of >1.8 million restriction digestions was carried out, matching reference sequences of all 5660 assemblies and 2236 building blocks against all 236 enzyme candidates. In order to evaluate the suitability of each enzyme candidate, the following metrics are proposed for a given set of constructs:

(1) *cut1* = fraction of constructs with at least 1 cut outside the vector region.
(2) *min100* = fraction of constructs with smallest fragment > 100 bp.
(3) *max10k* = fraction of constructs with largest fragment < 10 kb.

Thus for a given set of constructs, the ideal enzyme would score highest in all three metrics. A code written on the QB64 platform (http://qb64.net) was used for digest simulations and calculation of metric values.

Automated digest pattern matching and 'best clone' selection

In the ideal case, capillary electrophoresis data of restriction digest should match expected digest pattern according to several criteria. These criteria are formulated below, along with sources of error that create non-idealities:

1 *All fragment peaks are distinct from one another.*
In capillary electrophoresis, axial dispersion (band broadening) is a mechanistic phenomenon arising from chemistry, fluid mechanics, as well as thermal and electrical effects (30). As a result, two fragments of similar size may not be resolved but rather appear as a single peak with combined signal strength. In our

case the resolution threshold appears to be ~4% (e.g. 500- and 519-bp fragments barely resolved).

2 *Every fragment of expected size is present.*
Errors in measured versus expected fragment sizes can occur due to slight changes in calibration curves, sample differences and capillary variations. The observed coefficient of variations (CV) of size ratios is <3% across the whole 0.1–10 kb spectrum.

3 *No extra peaks are present.*
Incomplete digestion, minor impurities and suboptimal peak integration threshold in electropherogram processing may result in extra peaks being observed/reported.

4 *Expected fragments are equimolar with respect to one another.*
For each sample, noise in signal detection and resolution-dependent baseline peak integration contribute to error in fragment quantification. As molarity values (nmol/μl) are calculated from estimates of abundance (ng/μl) and size (bp), any measurement errors in these variables will contribute to total error in molarity estimates.

In order to reconcile the non-idealities, it is necessary to reformulate the four ideal match criteria above into a single criterion: *Each expected fragment is represented by an 'assigned normalized molarity' no less than an abundance threshold and of size divergence no more than a sizing threshold.* The following are derivations and algorithm used in generating the new match criterion.

For a restriction digest with n expected fragments and m observed peaks, the unit molarity u (nmol/μl) is defined as

$$u = \frac{\sum_{i=1}^{m} c_i}{n}, \tag{1}$$

where c_i (nmol/μl) is the molarity of the i-th peak. Thus the normalized molarity r (unitless) is defined as

$$r_i = \frac{c_i}{u} = \frac{nc_i}{\sum_{i=1}^{m} c_i}. \tag{2}$$

Through mathematical characterization (see Supplementary Note S1), the significance of r can be illustrated in three general scenarios:

(1) *All expected peaks resolved, no extra peaks ($m = n$)*: in the ideal case each peak should have an r_i-value close to 1, subject to measurement error relative to unit molarity u.
(2) *Some peaks unresolved, no extra peaks ($m < n$)*: subject to measurement error, r-values should be close to round numbers with the multiple representing number of unresolved peaks in the observed peak.
(3) *All expected peaks resolved, with extra peaks ($m > n$)*: relative to total molarity of expected peaks, extra peaks decrease r-values across the board (e.g. at 1:1 total molarity of extra and expected peaks, all r-values would drop by 50%).

With the normalized molarity r defined and characterized, below is a sketch of the digest pattern matching algorithm:

(1) Initialization: prepare list of expected fragments (i = 1–n) and list of observed peaks (j = 1–m), both sorted ascending by size. To avoid complications with peaks close to or smaller than the lower marker (100 bp), a size cutoff may be adopted, below which both expected fragments and observed peaks are ignored.
(2) Starting with the first expected fragment (i = 1), scan the list of observed peaks ascending, to find the first peak that satisfies two criteria:
   (a) The divergence f_i between expected size (s_i) and estimated size (s_j) is within sizing threshold (ϵ):

$$\max\left(\frac{s_j}{s_i}, \frac{s_i}{s_j}\right) = f_i < 1+\epsilon. \tag{3}$$

   (b) The normalized molarity is no less than an abundance threshold ($r_j \geq \delta$).
(3) If such a peak is found, 1 is subtracted from the normalized molarity r_j and *assigned* to r_i (alternatively if $r_j < 1$, r_j is subtracted from itself and assigned to r_i): fragment i finds a match.
(4) Repeat steps 2 and 3 for all n expected fragments.

The digest data passes verification if each expected fragment finds a match, i.e. represented by an assigned normalized molarity $r_i \geq \delta$ and estimated size s_j within ϵ of expected size s_i. In general, threshold values should not be too stringent (to avoid false negatives), and not too permissive (to avoid false positives). In this work ϵ is set at 0.1 according to electrophoresis performance specifications endorsed by the instrument manufacturer, and δ is set at 0.5.

Screening of multiple clones of the same construct could indicate that more than one clone was correct for any given assembly, in which case the 'best clone' can be selected based on digest data. Given a list of clones that already pass the match algorithm (k = 1–p), excess normalized molarity x for each clone can be calculated post-run:

$$x_k = \sum_{j=1}^{m} r_j. \tag{4}$$

Also for each clone, the maximum size divergence d can be calculated over all matched fragments:

$$d_k = \max\{f_i\}_{i=1}^{n}. \tag{5}$$

Thus the best clone may be selected according to these criteria:

(1) Extra peaks can now be defined more rigorously as ones whose r-values never got subtracted in match algorithm step 3 above.
(2) Clones without extra peaks are preferred to ones with extra peaks.

(3) Among clones with extra peaks, the best clone is one with lowest *x*-value (lowest excess normalized molarity).
(4) Among clones without extra peaks, the best clone is one with lowest *d*-value (lowest maximum size divergence).

A code written on the QB64 platform was used to execute pattern matching and best clone selection.

Sanger sequencing

Whereas verification of DNA assembly at the structural level was achieved by matching of restriction digest pattern as described above, further verification was provided by Sanger sequencing. For building blocks, the best clones were sequenced by primers annealing on the cloning vector backbone, from 5′ (GGAGCAGACAAGC CCGTCAGGG) and 3′ (GCTGATACCGCTCGCCGC AG). For assemblies, the best clones were sequenced from the 5′ end by a primer annealing on the assembly vector backbone (GCGGATAACAATTTCACACAGG AAACAGC) and by primers annealing on each 5′ linker introduced by the building blocks (linker sequences in Supplementary Table S1 in Supplementary Information). The number of sequencing reactions for each assembly is thus the same as the number of building blocks. Primer binding sites on vector backbones are highlighted in vector sequences (Supplementary Sequences S1 and S2).

RESULTS

This work proposes a methodology for high-throughput verification of structural DNA assembly by restriction digest. To establish the reliability of the assay, we choose as a test set the 2236 cloned building blocks and 5660 assemblies manufactured over 30 weeks in our DNA assembly pipeline (total number of constructs = 7896). With four clones screened per construct, the total number of clones assayed by restriction digest was 4 × 7896 = 31 584. For each construct, the best clone selected by restriction digest analysis was further verified by Sanger sequencing, thus providing two sets of data (2 × 7896 constructs) for comparison of the two methods.

Selection of most suitable enzyme for building block and assembly sets

From the size distribution of building blocks (Figure 2) it is apparent that total size of building block + 2.2 kb vector is <10 kb for essentially the entire population, making the *max10k* metric superfluous (this is more evident in Figure 3a, where *max10k* for building blocks cluster around value 1.0). Figure 3b shows the *in silico* enzyme screen results for 2236 cloned building blocks. Plotting only *cut1* versus *min100*, two regimes may be highlighted:

(1) Low-frequency cutters: as these cut rarely, they score low in *cut1* (stringent) but have no problem scoring high in *min100* (permissive) due to large fragments being left.
(2) High-frequency cutters: as these create small fragments, they score low in *min100* (stringent) but have no problem scoring high in *cut1* (permissive).

The regime of importance is one with highest scores in both *cut1* and *min100*, from which a short list of best enzymes may be drawn—at this point further considerations may be applied. Despite being statistically suitable (high scores), some enzymes were disqualified for the sake of process robustness, e.g. *Bsr*FI (irreversibly binds DNA, altering electrophoretic mobility), *Mme*I (requires 1:1 stoichiometry to substrate DNA), *Nae*I (requires another cut site as *cis*-acting effector). *Ava*II and *Afl*II passed all statistical and robustness requirements and thus were used in for restriction digest assay. It should be noted that although the 2236 building blocks are treated as a unified set in this work, they were actually created in 10 waves over 30 weeks. Thus of the 10 subsets, some were verified using *Ava*II and others using *Afl*II depending on their metric scores on particular subsets.

Figure 3c shows the enzyme screen results for 5660 assemblies. Analysis of the assembly set is analogous to that of cloned building blocks, but due to the larger size distribution (Figure 2), *max10k* is discriminating (values range from 0.75 to 1.0, as shown in Figure 3a). The frequency regimes now gain an added dimension: high-frequency cutters have no problem scoring high on *max10k* (permissive) due to small fragments, but low-frequency cutters in general score low on *max10k* (stringent) due to large fragments being left. *Bsr*DI (*cut1* = 0.97, *min100* = 0.94, *max10k* = 1.0) passed all statistical and robustness requirements and thus was used universally for restriction digest assay of assembled constructs. It should be noted that once applied, the frequency metrics (*cut1*, *min100*, *max10k*) constitute an unbiased screen, with selection of high-scoring enzymes predestined by construct reference sequences (i.e. distribution of restriction sites in the sequences). In our case, we find that that *Bsr*DI is still a suitable enzyme for assemblies beyond the 5660 described here, presumably because the sequences still share a degree of similarity (data not shown).

Quality verification of linear building blocks

DNA building blocks were amplified by PCR as described in Materials and Methods section. To ensure successful assembly and cloning, linear DNA building blocks were analysed on the Fragment Analyzer instrument to verify size, concentration and molar purity. Across >5600 samples, the CV of relative size (estimate/expected) was 2.8%, the average concentration was 98.5 ng/µl, and the average molar purity was 0.98 (distribution plots of this data can be found in Figure 4). The instrument sizing CV is particularly significant as a statistical parameter to determine suitable deviation threshold between expected and observed fragment sizes in restriction digest assay (more details discussed in subsequent section).

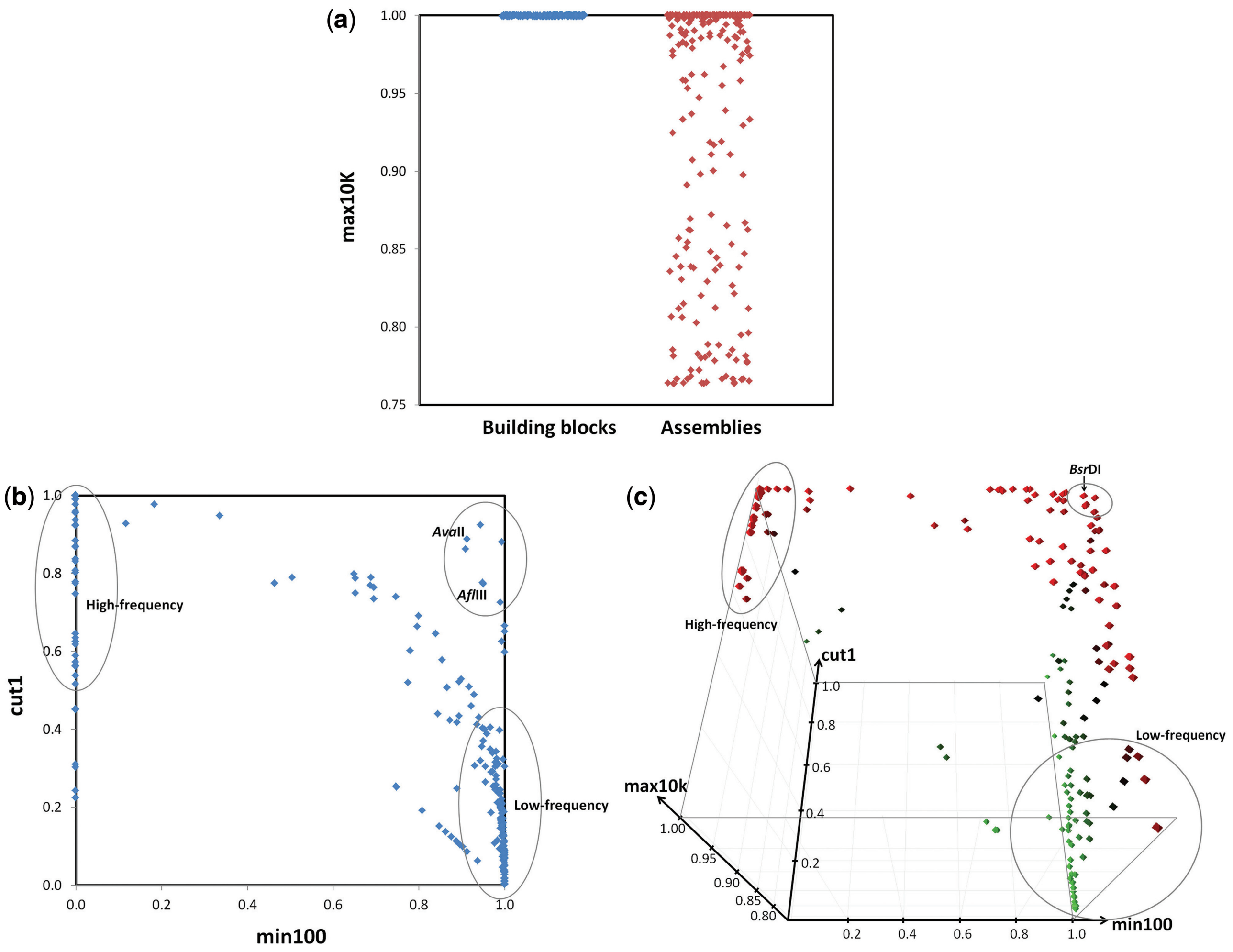

Figure 3. *In silico* evaluation of 236 unique restriction enzyme sequence/cut sites according to frequency metrics *cut1*, *min100* and *max10k*. Regimes of low-frequency and high-frequency cutters are highlighted. Best enzymes are found where metrics values are highest. (**a**) *max10k* values for building blocks cluster ~1.0 as essentially all have total size (including vector) below 10 kb. In contrast, *max10k* is discriminating for assemblies (values range from 0.75 to 10) due to larger size distribution. (**b**) Screening against 2236 cloned building blocks using metrics *cut1* and *min100*. Most suitable enzymes are *Afl*III and *Ava*II. (**c**) Screening against 5660 assemblies using metrics *cut1*, *min100* and *max10k*. Most suitable enzyme is *Bsr*DI.

Examples illustrating digest data and successful pattern matching

Figure 5a shows an electropherogram with baseline resolution of 19 peaks between 0.1 and 10 kb, forming the calibration ladder (separation time <23 min). To demonstrate robustness of the restriction digest assay, *Ava*II digest of cloned building block #30655 (5.4 kb plasmid) is showcased in Figure 5b. Of 9 expected fragments (152, 175, 222, 318, 522, 859, 898, 1107, 1172 bp), the last two are not resolved but rather appear as a single 1124-bp peak, and there are 3 extra peaks as a result of minor impurities (986, 1380, 2229 bp). Despite unresolved peaks and extra peaks, the algorithm was able to match the data to the expected DNA construct, thus verifying the structural integrity of this clone.

Figure 5c showcases the *Bsr*DI digest of a 12-part assembly #54520 (16.6 kb plasmid). Of 9 expected fragments (174, 234, 1058, 1205, 1943, 2194, 2430, 3617, 3754 bp) peaks # 6 and 7 are only partially resolved and peaks # 8 and 9 are unresolved, appearing as a single 3760-bp peak. Structural integrity of this clone is nevertheless verified since it passes the match criteria as vetted by the algorithm. Comparison of the overall fluorescent signal levels between Figure 5b and c indicates that smaller DNA substrates are preferentially amplified during the 16 h RCA reaction (compare 5.4 kb versus 16.6 kb). Consistent performance across this dynamic range further demonstrates robustness of the restriction digest assay in its entirety (physical process and matching algorithm).

Comparison of restriction digest assay and Sanger sequencing

The restriction digest assay was used to pre-screen four clones of each DNA assembly/cloning and the best clone for each was further verified by Sanger sequencing (Table 1). While the percentage of constructs verified by both methods indicate our overall success rates from an operational standpoint (>95% for both building blocks cloning and DNA assembly), more pertinent to this

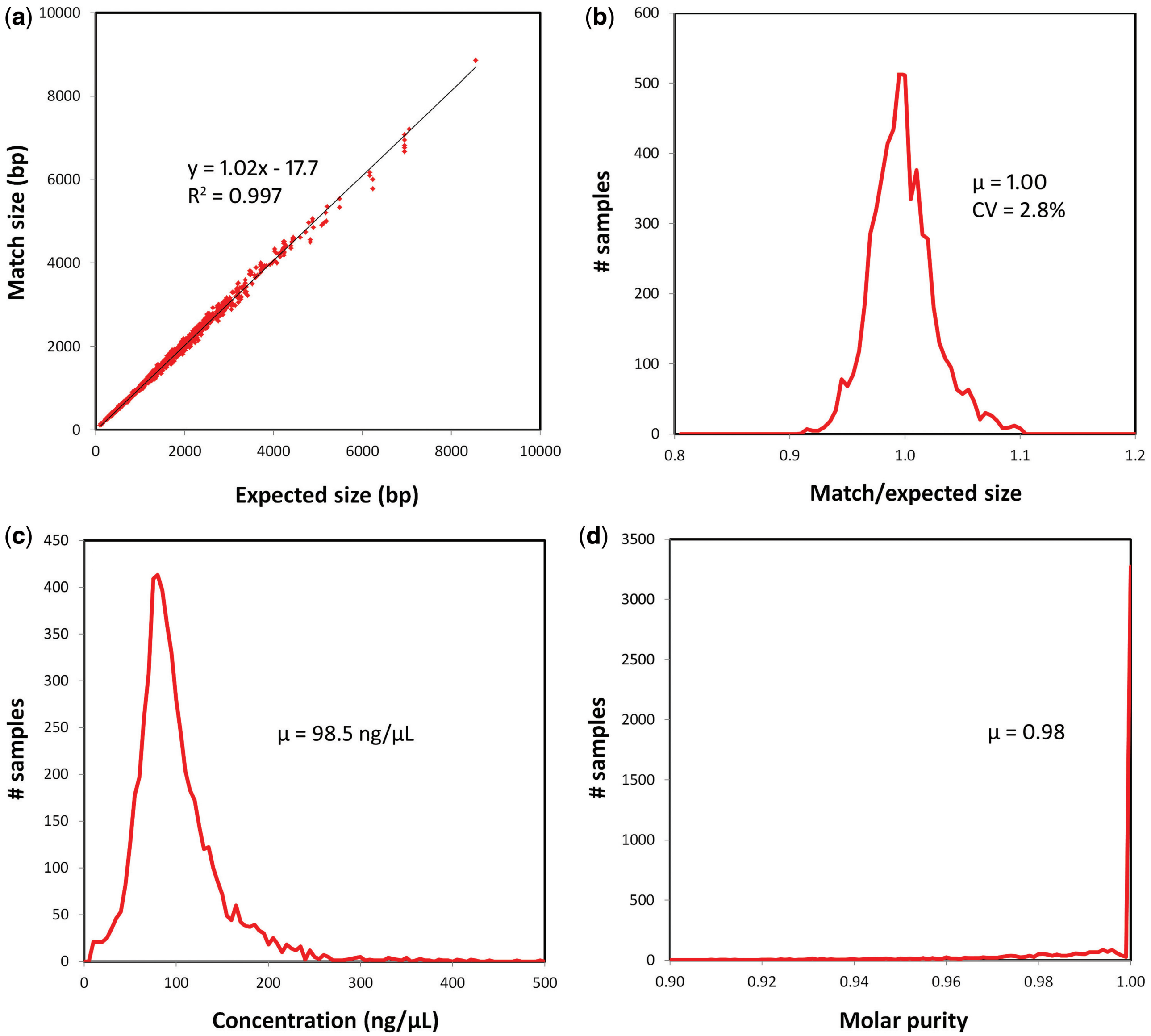

Figure 4. Quality verification of linear DNA building blocks used in this work. Of 5608 data points, the CV of match/expected size ratio = 2.8%, the average concentration = 98.5 ng/μl (median = 88 ng/μl), and average molar purity = 0.98 (median = 1).

work is the agreement between our proposed method and sequencing. The results indicate a disagreement rate below 2%, i.e. 36/2236 = 1.6% for building blocks and 107/5660 = 1.9% for assemblies. Due to lack of a perfect assay, this is not truly a false positive rate because it cannot be distinguished from the false negative rate of Sanger sequencing. In fact, as much as 0.9% of disagreement can be attributed to poor data quality, mostly observed in templates with repeat sequences (possibly due to mispriming) and homonucleotide stretches (possibly due to polymerase slippage).

Sequence data nevertheless uncovered several failure modes compromising the structural integrity of DNA constructs, which are invisible with the restriction digest assay. A common failure observed in assembly is omission of a building block due to non-specific recombination between mismatched linkers. This mode of failure can escape detection when the omitted block is small enough that the associated fragments are still within sizing threshold. Within individual blocks, structural defects can be traced to anomalies in PCR amplification: deletions/insertions as large as 75 bp were observed, which could have arisen from repeat sequences in the original PCR template (i.e. mispriming events). These false positives can be minimized with more stringent sizing threshold (ϵ) with respect to sizing CV, but at the expense of increased false negative rate of the assay. For a normally distributed sizing performance with CV = 2.8% (Figure 4b), the 1-tailed statistical type I error α would grow > 200 times from 0.018 to 3.7% of all size matching instances if ϵ is made twice as stringent from current level (0.1 to 0.05). With our current workflow, it is appropriate that the pre-screen assay is not too stringent as false positives would eventually be screened out by sequencing.

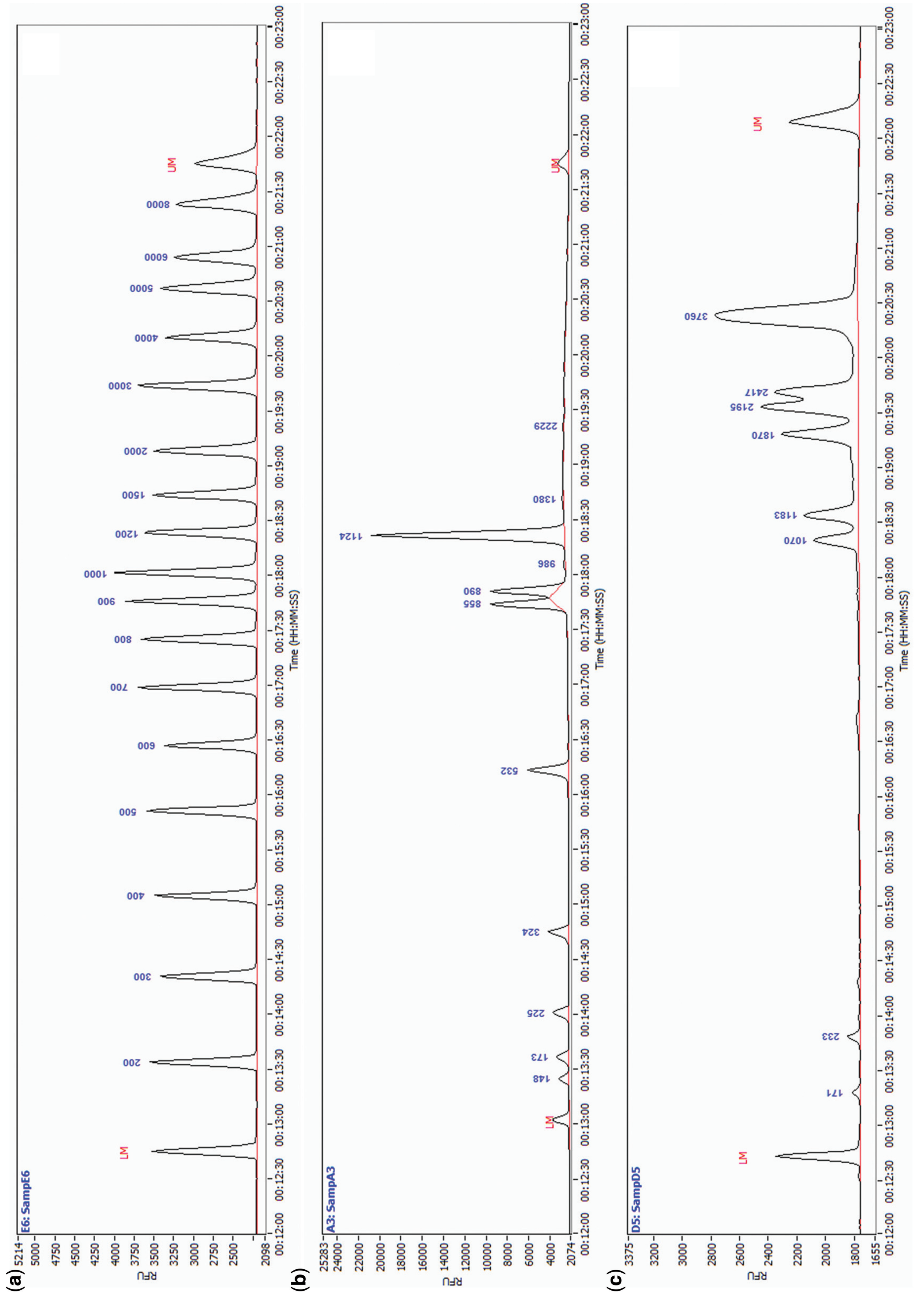

Figure 5. (**a**) Baseline resolution of 19 peaks between 0.1 and 10 kb, forming the calibration ladder. Analysis time is <23 min. (**b**) *Ava*II digest of part #30655 (5.4 kb plasmid), exemplifying unresolved (1124 bp) and extra peaks (986, 1380, 2229 bp). (**c**) *Bsr*DI digest of 12-part assembly # 54520 (16.6 kb plasmid), exemplifying partially resolved peaks (2195 and 2417 bp) and unresolved peaks (3760 bp).

Table 1. Restriction digest assay versus Sanger sequencing of building blocks and assemblies

Category	Specimen	Restriction digest assay	Sanger sequencing	No. of building blocks	%	No. of assemblies	%
1	Four clones	Fail	N/A	65	2.9	170	3.0
2	Best clone	Pass	Pass	2135	95.5	5383	95.1
3	Best clone	Pass	Fail	36	1.6	107	1.9
TOTAL				2236	100.0	5660	100.0

Category 1 is the subset in which all four clones of each construct failed structural verification by restriction digest assay, and hence were not further verified by sequencing. The rest of the set have at least one in four clones pass restriction digest assay, which upon further sequencing of best clone may pass (category 2) or fail (category 3).

DISCUSSION

Assembly of well-characterized DNA building blocks into larger constructs of predictable biological functions is a process widely used in the field of synthetic biology. In this context, ensuring accurate mapping of genotypes to phenotypes relies on verifying DNA constructs, a task considered to be elementary in rational strain engineering. However, at a large scale we learn that the logistics and expenses become a mounting challenge, thus making a cheap, fast and reliable assay absolutely essential. To this end, we have described a comprehensive methodology for high-throughput structural verification of DNA constructs by restriction digest and capillary electrophoresis, including laboratory protocols, exhaustive bioinformatics screening for the most suitable enzyme, and algorithms for digest pattern matching and best clone selection. We demonstrated robustness of the assay across large datasets of DNA building blocks and assemblies covering the full spectrum of construct size and multiple enzyme/buffer systems, as well as excellent agreement with sequencing results owing to suitable assay thresholds chosen based on empirical system performance.

With an electrophoresis runtime of <37 min per 96-well plate, the assay accommodates analysis of up to $12 \times 96 = 1152$ clones every 8 h per instrument, a throughput and automation unmatched by manual pouring and loading of agarose gels, not to mention the quantitative data (fragment size, abundance) amenable to automated processing that is simply not available in qualitative gel images. Most importantly, the assay is quite inexpensive: accounting for all reagents and labware, the cost is only $0.84 per clone. This cost is independent of the number of building blocks per clone as opposed to Sanger sequencing, which requires a sequencing reaction for every building block at a cost of $3 per reaction. Assuming an average of 4 building blocks per assembly, the cost of Sanger sequencing would be $12 per clone, one order of magnitude higher than that for restriction digest assay. Overall, the use of the restriction digest assay as a pre-screen followed by sequencing of the best clone achieves the same outcome as verification of DNA constructs by sequencing exclusively, but with substantial savings.

Beyond its original purpose, the assay will be immediately applicable to verification of any plasmid collection, for example EST libraries (http://clones.invitrogen.com), and in particular re-arrayed clone sets (http://bacpac.chori.org). Given scant literature on DNA construct verification beyond traditional methods (i.e. Sanger sequencing, digest or PCR analysis by agarose gel), we are convinced that the proposed method would be a valuable contribution to the research community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge the support provided by the Lab Services team at Amyris. We also thank Advanced Analytical for the technical assistance provided during the setting up of the Fragment Analyzer^{TM} instruments.

FUNDING

Funding for open access charge: Company funds.

Conflict of interest statement. All the authors possess stock or stock options in Amyris Inc.

REFERENCES

1. Kitney,R. and Freemont,P. (2012) Synthetic biology – the state of play. *FEBS Lett.*, **586**, 2029–2036.
2. Agapakis,C.M. and Silver,P.A. (2009) Synthetic biology: exploring and exploiting genetic modularity through the design of novel biological networks. *Mol. Biosyst.*, **5**, 704–713.
3. Pasotti,L., Politi,N., Zucca,S., Cusella De Angelis,M.G. and Magni,P. (2012) Bottom-up engineering of biological systems through standard bricks: a modularity study on basic parts and devices. *PLoS ONE*, **7**, e39407.
4. Anderson,J.C., Dueber,J.E., Leguia,M., Wu,G.C., Goler,J.A., Arkin,A.P. and Keasling,J.D. (2010) BglBricks: a flexible standard for biological part assembly. *J. Biol. Eng.*, **4**, 1–1.
5. Galdzicki,M., Rodriguez,C., Chandran,D., Sauro,H.M. and Gennari,J.H. (2011) Standard biological parts knowledgebase. *PLoS ONE*, **6**, e17005.
6. Gibson,D.G. (2009) Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res.*, **37**, 6984–6990.
7. Ma,S., Tang,N. and Tian,J. (2012) DNA synthesis, assembly and applications in synthetic biology. *Curr. Opin. Chem. Biol.*, **16**, 260–267.
8. Jiang,X., Yang,J., Zhang,H., Zou,H., Wang,C. and Xian,M. (2012) In vitro assembly of multiple DNA fragments using successive hybridization. *PLoS ONE*, **7**, e30267.
9. Quan,J. and Tian,J. (2009) Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS ONE*, **4**, e6441.

10. Schmid-Burgk,J.L., Xie,Z., Frank,S., Virreira Winter,S., Mitschka,S., Kolanus,W., Murray,A. and Benenson,Y. (2012) Rapid hierarchical assembly of medium-size DNA cassettes. *Nucleic Acids Res.*, **40**, e92.
11. Shao,Z. and Zhao,H. (2009) DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.*, **37**, e16.
12. Gibson,D.G., Benders,G.A., Axelrod,K.C., Zaveri,J., Algire,M.A., Moodie,M., Montague,M.G., Venter,J.C., Smith,H.O. and Hutchison,C.A. (2008) One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. *Proc. Natl Acad. Sci. USA*, **105**, 20404–20409.
13. 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.
14. Constante,M., Grunberg,R. and Isalan,M. (2011) A biobrick library for cloning custom eukaryotic plasmids. *PLoS ONE*, **6**, e23685.
15. Dean,F.B., Nelson,J.R., Giesler,T.L. and Lasken,R.S. (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.*, **11**, 1095–1099.
16. Maschke,H.E., Frenz,J., Belenkii,A., Karger,B.L. and Hancock,W.S. (1993) Ultrasensitive plasmid mapping by high performance capillary electrophoresis. *Electrophoresis*, **14**, 509–514.
17. Peck,K., Wung,S.L., Chang,G.S., Yen,J.J. and Hsieh,Y.Z. (1997) Restriction mapping of genes by capillary electrophoresis with laser-induced fluorescence detection. *Anal. Chem.*, **69**, 1380–1384.
18. Strege,M. and Lagu,A. (1991) Separation of DNA restriction fragments by capillary electrophoresis using coated fused silica capillaries. *Anal. Chem.*, **63**, 1233–1236.
19. Entian,K.-D. and Kötter,P. (2007) Yeast genetic strain and plasmid collections. *Method. Microbiol.*, **36**, 629–666.
20. Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene*, **33**, 103–119.
21. Sikorski,R.S. and Hieter,P. (1989) A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics*, **122**, 19–27.
22. Ma,H., Kunes,S., Schatz,P.J. and Botstein,D. (1987) Plasmid construction by homologous recombination in yeast. *Gene*, **58**, 201–216.
23. Manivasakam,P., Weber,S.C., McElver,J. and Schiestl,R.H. (1995) Micro-homology mediated PCR targeting in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **23**, 2799–2800.
24. Oldenburg,K.R., Vo,K.T., Michaelis,S. and Paddon,C. (1997) Recombination-mediated PCR-directed plasmid construction in vivo in yeast. *Nucleic Acids Res.*, **25**, 451–452.
25. Gietz,R.D. and Woods,R.A. (2001) Genetic transformation of yeast. *Biotechniques*, **30**, 816–820.
26. Bubeck,P., Winkler,M. and Bautsch,W. (1993) Rapid cloning by homologous recombination in vivo. *Nucleic Acids Res.*, **21**, 3601–3602.
27. Oliner,J.D., Kinzler,K.W. and Vogelstein,B. (1993) In vivo cloning of PCR products in *E. coli*. *Nucleic Acids Res.*, **21**, 5192–5197.
28. Zhu,D., Zhong,X., Tan,R., Chen,L., Huang,G., Li,J., Sun,X., Xu,L., Chen,J., Ou,Y. *et al.* (2010) High-throughput cloning of human liver complete open reading frames using homologous recombination in Escherichia coli. *Anal. Biochem.*, **397**, 162–167.
29. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2010) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, 234–236.
30. Ghosal,S. (2006) Electrokinetic flow and dispersion in capillary electrophoresis. *Annu. Rev. Fluid Mech.*, **38**, 309–338.