## Title page:

## Smrt Gate: A solution for the validation of synthetic constructs on the Pacific Biosciences sequencing platform

Rosalinda D’Amore1,2, James Johnson1, Sam Haldenby2, Neil Hall1,2,3, Margaret Hughes2 , Ryan Joynson2,3 , John G. Kenny2, Nicola Patron3, Christiane Hertz-Fowler 1,2, Anthony Hall\*1,2, 3

1GeneMill, Centre for Genomic Research, MerseyBio Building, Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool, L69 7ZB, UK

2Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool, L69 7ZB, UK

3 The Earlham Institute, Norwich Research Park, Norfolk, NR4 7UH, UK

\*corresponding author: Anthony Hall (anthony.hall@Earlham.ac.uk)

Keywords: Golden Gate, synthetic biology, next-generation sequencing, NGS, high throughput, plasmid QC

Abstract word count: 214

Manuscript word count:

### 

### Abstract

Current molecular biology methods of DNA assembly are prone to sequence errors. This requires rigorous quality control (QC) to identify incorrectly assembled or synthesised constructs. These errors can lead to the misinterpretation of phenotypes. Because of this intrinsic problem, routine QC analysis is generally performed on three or more clones using a combination of restriction endonuclease assays, colony PCR, and Sanger sequencing. However, with new automation methods enabling high-throughput assembly, QC has become a major bottleneck. Although Next-Generation Sequencing (NGS) technologies have increased the throughput and decreased the cost of sequencing whole genomes, Sanger sequencing is still routinely used for relatively small 5-50 kb synthetic construct. Despite the low cost per bp of NGS, library preparation and sample-multiplexing are a significant hurdle for projects that require large numbers of clones to be sequenced.

To overcome this issue, we have developed a quick and affordable methodology for the QC of synthetic constructs: a one-pot digestion-ligation DNA assembly based on the Golden Gate methodology coupled with Pacific Bioscience Single Molecule Real Time (PacBio SMRT) sequencing.

In summary, this report describes a new sequencing protocol for a quick and cost-effective screening of multipart synthetic constructs. It uses the Golden Gate assembly method to build barcoded libraries and the long-read sequencing capability of Pacific Bioscience RSII sequencing platform.

**Introduction**

Synthetic biology is an interdisciplinary field that applies engineering principles to biology to build and optimise biological systems for new purposes using design-build-test cycles. Synthetic biology is a relatively new field with the potential to have a major impact on a large number of key application areas such as healthcare (1), agriculture, biofuels (2), novel materials (3), fine chemical production (4), bio-remediation (5) and clean water (6). A key component of synthetic biology is the ability to assemble and re-use biological parts. A powerful approach to part assembly has been Golden Gate Cloning (7, 8), which uses type IIS restriction enzymes and T4 ligase to assemble multiple parts in a single tube. A modular and hierarchical cloning system, MoClo, was developed (9). The system combines the concept of basic functional parts and Golden Gate technology and allows the rapid assembly of complex constructs built up from basic parts. Fundamental to this strategy is that the parts and final constructs are sequenced to ensure that they are correctly ordered and have been synthesised with high fidelity.

Alongside this parts-based approach, another underpinning technology of synthetic biology is the application of high-throughput automation. Automated fabrication and testing allow reaction miniaturisation, removing human error and increasing the speed and scale of experimentation (10). However, the scale of production of DNA assembly now creates another bottleneck: the QC of the final products. Routinely, Sanger Sequencing is used to perform quality control of the assembled constructs. Three or more positive clones are screened using a combination of colony PCR and restriction endonuclease assays to verify the correct size of the assembled constructs and subsequently submitted to Sanger sequencing to confirm that the parts were assembled correctly. The average length of Sanger reads is 800 bp hence, constructs larger than 1.5 kb must go through several rounds of sequencing using bespoke primers to be able to achieve coverage across the full length of the construct. This is an expensive, time-consuming and difficult to automate approach. Here we describe a methodology (SMRT Gate) that takes advantage of the long read technology offered by the Pacific Bioscience (PacBio) RSII sequencing platform (Pacific Bioscience, Menlo Park, California) (11) in combination with a type IIS restriction enzyme-based cloning system to rapidly attach sequencing barcodes and adapters. This generates a barcoded library of SMRTbells suitable for sequencing. We have used this methodology to sequence forty-one synthetic constructs (insert size ranging 1.3-7.7 Kb), built using the MoClo Assembly Standard (9). The method provides a simple, cheap, fast and robust alternative to sanger sequencing for the quality control of constructs (Estimated cost and time required for alternative methods, Supplementary Table S1).

### Results and Discussion

#### SMRT Gate library

PacBio long read sequencing captures sequence information during the replication process of the target DNA molecule. The template, called a SMRTbell, is a closed, single-stranded circular of DNA that is created by ligating hairpin adaptors to both ends of target double-stranded linear DNA (12).

PacBio protocols exist to build SMRTbell libraries with insert sizes ranging from 500bp to greater than 20kb. Although barcoding methods do exist, these are time-consuming and involve multiple steps. The template has to be provided as fragments, repaired and ligated to barcoded SMRTbells to allow multiplexing. The cost of a library preparation using such procedure can be more than $300 per reaction. This represents a limiting factor and a constraint in using PacBio sequencing for synthetic construct quality control.

To streamline the library construction and multiplexing, we designed a sequencing strategy (SMRT Gate) based on Golden Gate MoClo (9). The Golden Gate Cloning method allows basic genetic elements to be made into standardised DNA parts by cloning into a plasmid vector (Level 0). Subsequently, these standard parts are assembled together in a single step building transcriptional units (in Level 1 plasmid vectors), which have defined fusion sites and positions for further assembly. Level 1 transcriptional units can be assembled up to six at a time in the next assembly levels (Level 2 or M). All MoClo assemblies are free of internal recognition sequences for *BbsI* (*BpiI*) and *BsaI*, which flank the parts and are used for iterative assembly.

In our method, we used the Level 1 constructs’ unique fusion sites, generated after the cleavage with *BbsI* (*BpiI*), to design the SMRTbell adapters (Figure 1). Forty-one Level 1 synthetic constructs were pooled into three groups based on their fusion sites (Table 1). The three pools were processed individually using either barcoded SMRT Gate bells (Supplementary material, Table S2), *BbsI (BpiI)* and T4 DNA ligase or the SMRT bells with the standard multi-step library protocol, to generate three barcoded SMRTbell libraries for PacBio sequencing for each method.

The libraries were analysed using a TapeStation to assess the size of the fragments and yield. The fragments profile for both methods was comparable but the SMRT Gate libraries had a lower yield (50 ng) compared to the standard library (80 ng). Prior to going through the SMRT GATE protocol, the libraries showed a single peak (Figure 2, panel 1. See also Figure S1 and S2, panel 1 in supplementary material) comparable with the size of the initial synthetic constructs pool. However, after cleavage with *Bbs*I and SMRT bell ligation, the libraries showed distinct peaks comparable with the size of the fragment released from the construct after cleavage.

(Figure 2, panel 3 . See also Figure S1 and S2, panel 3 in supplementary material) Moreover, a smear between 50 bp and 45 kb was present, which highlight unligated SMRTbell and undigested constructs (Figure 2, panel 2 . See also Figure S1 and S2, panel 2 in supplementary material). To reduce the unligated SMRTbell and undigested synthetic constructs an exonuclease reaction was performed for both libraries.

After performing the reaction, the final SMRT GATE library had a reduced background. The peaks correspond to the average fragment size generates after endonuclease cleavage for each pool (Figure 2, panel 3 . See also Figure S1 and S2, panel 3 in supplementary material).

The better yield of the standard protocol didn’t translate into an increase sequencing throughput. A failure in SMRTbell ligation and exonuclease activities were the likely cause which produced the loading bias. The low sequencing yield was due to under-loading where the majority of wells of the SMRT cell did not contain a template.

#### Analysing Data using the SMRT portal

The raw reads generated by the RSII PacBio sequencing platform were analysed using the “on-board” tools on the SMRT portal, a cloud-based, software platform (<http://www.pacb.com/support/software-downloads/>) The datasets were analysed using the RS\_Resequencing pipeline. Briefly, length and quality filtered reads were demultiplexed using the Resequencing Barcoding module and subsequently locally aligned to the reference sequence based on the construct designs using BLASR. Following mapping, the Quiver algorithm was employed to detect variants and generate consensus sequences. An overview of variants calling and coverage for all the constructs for both methodologies is provided in Table 1 and Supplementary material, Table S3. We sequenced 41 synthetic constructs with an average depth of coverage of 13,162 fold for the SMRT gate constructs and 878 fold for the standard library. The depth of sequencing for each construct was unbalanced, ranging from 12- to 57,120-fold for the SMRT Gate constructs and 2- to 4,565-fold for the Standard protocol. Although the constructs were prepared with consideration given to fragment sizes in the pooling calculation, consistent with our previous experience with the Pacific Biosciences RSII, small fragments were still preferentially loaded over the larger ones. Loading of SMRTbell template is size dependent, this is particularly important when sequencing libraries with different insert size. To overcome this issue and to achieve a more even representation we would highly recommend to select libraries of a similar size (+/- 10%) to minimise loading biases or increase the concentration of larger fragment in the pools.

(http://www.pacb.com/wp-content/uploads/2015/09/Guide-Pacific-Biosciences-Template-Preparation-and-Sequencing.pdf)

#### Analysis pipeline and variants calling

Due to the nature of PacBio sequencing, with average Polymerase Reads of 50 kb and considering a max. fragment length in the pool of 7.7 kb, we expected the reads to map across the entire length of the synthetic construct.

After mapping and variant calling we filtered out the four constructs (11021, 11030, 11045,11046, indicated as FAIL in Table 1) which had a drop in local coverage across the reference sequence (within a 200 bp window) in spite of their overall coverage being more than 20-fold. For these constructs we used a combination of PCR and Sanger Sequencing. For the PCR we designed primers to target the vector sequence and amplifying across the inserts, and we submit the amplicon to Sanger sequencing to verify the sequence. We confirmed that the constructs provided were different from the reference sequence and hence looked as if they had failed because we had not included a reference sequence for these constructs in our reference mapping database.

Of the remaining thirty-seven constructs, the PacBio sequence data agreed 100% with the reference sequence in thirty-one constructs. In six constructs, the RS\_resequencing pipeline identified five SNPs and four indels in the constructs generated via SMRT Gate protocol while four SNPs and three indels were identified in the Standard method (Table 1 and Supplementary material, Table S3). All the variants were confirmed by Sanger sequencing and highlighted that they have occurred during the assembly of the original Level O parts.

Part of the screening process is selecting two to five colonies and sequence them to assess point mutation. To confirm the ability of the PacBio to discriminate single point mutations, we included 2 constructs in the pool, which were identical with the only exception of a base in position 3332 (11065 T/11066 C). The reference sequences for 11065 and 11066 were aligned to the consensus sequences generated by Quiver (Fig.3A left panel) and the software assigned correctly the SNP in position 3332. Moreover, a new SNP in position 2774 was found (Fig 3A right panel) and subsequently validated.

### Conclusion

SMRT Gate can be used to generate libraries suitable for Single Molecule Sequencing technology without the need of individual barcoding and time-consuming procedures. Although we have yet to test the methodology on constructs made using other Type IIS assembly methods, it could also be applied to those made using BASIC (13), Golden Braid (14) or TNT (15). The methodology can be adapted to any construct flanked by recognition sequences for Type IIS restriction endonucleases that generate a 4 bp overhang. The RSII PacBio platform can generate up to 100,000 reads while the Sequel can produce 5 times more reads. In our experiment the uneven distribution was probably due to the loading bias which can be mitigated providing libraries with the same size range (+/- 10%)

If consideration is given to library size, the reads distribution should be more even and therefore at least 10x coverage can be achieved when multiplexing 384 samples per SMRT. That would provide enough coverage for calling variants with confidence and a cheap alternative to Sanger sequencing.

### Materials and Methods

**SMRT Bell**

The oligonucleotides (5’ Phosporylation, PAGE purification) which are used to generate the SMRTbell adapter were ordered from IDT (Supplementary Table S2; Leuven, Belgium). The SMRTbell adapter must be annealed to form hairpin prior to ligation. To anneal the SMRT bell adaptor, the oligonucleotides were resuspended in 1x annealing buffer (10 x annealing buffer100 mM Tris-HCl pH7.5, 5MNaCl) at a final concentration of 20 μM in 20 μl. They annealing reactions were incubated at 80°C for 2 minutes, then ramped down to 25°C at a rate of 0.1°C/second, before transferring to 4°C for immediate use or store at -20°C.

#### Generation of SMRT Gate library

Forty-one plasmids were (see Table 1 Experimental Design) assembled in three pools using an equimolar amount of each plasmid. Each pool (500-1000 ng) was subjected to a Golden Gate type reaction to digest and ligate the barcoded SMART Gate SmtBell as following : 1 μl of T4 Ligase (1U/μl ,ThermoFisher Scientific, Runcorn, UK) , 1 μl of ATP (10nM ,NEB, Hitchin, UK), 0.5 μl of BbsI (10,000 U/ul ,NEB, Hitchin, UK) 1x Buffer 2.1 (NEB, Hitchin, UK) 1μl SmartGateBell (Supplementary material, Table S2) a and b (20 μM, IDT, Leuven, Belgium) in a final reaction volume of 50 μl or scale down to 10 ul when automation were used. The reaction was incubated on a thermocycler (Agilent SureCycler 8800, Stockport, UK) as following: First step at 37°C for 20 minutes, 2nd Step at 37°C for 1:50 and 16°C for 5 minutes (15 Cycles) and then 50°C for 5 minutes and finally and incubation at 80° for 10 minutes. After ligation, the reaction was submitted to an exonuclease reaction using 1μl Exo III (100 U/μl NEB, Hitchin, UK) and Exo VII (10 U/μl NEB, Hitchin, UK) to degrade products which failed ligation and templates containing internal nicks. The libraries were cleaned using 1x of AMPureXP Beads (Beckman, High Wycombe, UK) and resuspended in 10 μl of TE buffer. One μl of each library was run on Agilent Genomic DNA ScreenTape (Agilent, Stockport, UK) to ensure the final product was the correct size and quantified using a Qubit® (ThermoFisher Scientific, Runcorn, UK) dsDNA HS assay. The three libraries were pooled and the final library was 5 ng/μl.

#### Standard Library

The same pools of constructs were prepared as described above. Each pool was subjected to a digestion reaction to cleave the DNA fragments using 1 μl of BbsiI (10,000 U/ul, NEB, Hitchin, UK) 1x Buffer 2.1 (NEB, Hitchin, UK) at 37°C for 60 minutes in a final volume of 20 μl. Following 2 μl of rSAP (NEB, Hitchin, UK) was added to the reaction and incubated at 37°C for 60 minutes and subsequently at 65°C for 5 minutes. A dephosphorylation was performed to remove a phosphate group from a DNA fragment to inhibit self-ligation of the fragments. A ligation reaction was performed adding 1 μl of T4 Ligase (1U/μl , ThermoFisher Scientific, Runcorn, UK), 1 μl of ATP (10nM ,NEB Hitchin, UK), 1x Buffer 2.1 (NEB, Hitchin, UK) 1μl barcoded SMRTBell a and b (20 μM, IDT Leuven, Belgium) and incubated at 25 °C for 60 minutes . A clean-up step was performed using 1x of AMPureXP Beads (Beckman, High Wycombe, UK) to remove enzymes and salts and resuspended in 10 μl of TE buffer. The repair reaction was performed using the PreCR Repair Mix (NEB, Hitchin, UK) following manufacturer recommendation. After the repairing step the library was submitted to an exonuclease reaction and a cleanup as above. Final library concentration was 8 ng/μl.

#### PCR verification

Five ng each of constructs 11016, 11030, 11046, 11027 and 11045 were used to PCR amplify the inserts using 0.3 μM Level1TDNAL\_for; 0.3 μ M Level1 RB\_rev (IDT, Leuven, Belgium ,Supplementary material, Table S2); 1x OneTaq® Hot Start DNA Polymerase ready mix (NEB, Hitchin, UK). The PCR for each variable region was carried out in triplicate in a 25 μl reaction in the thermal cycler with the following parameters: initial denaturation at 94 °C for 5 min, followed by 25 cycles of 98 °C for 20 s, 60 °C for 15 s, and 72 °C for 3min with a final extension at 72 °C for 3 min.

#### Sanger sequencing

Five µl of Purified plasmid DNA (80 - 100 ng/µl) was mixed with 5 µl of primer leverl1TDNAL\_for and Level1 RB\_rev (5 µM) and submitted to a sequencing provider GATC (Konstanz, Germany) for Sanger sequencing.

#### PacBio Sequencing

The Pacific Biosciences calculator was used to determine the amount of primers and polymerase needed for the binding reactions, based on an insert size of 4500 bp. The primers and Pacific Biosciences proprietary p6 SA DNA polymerase v2 was bound to the library and the MagBead Kit (Pacific Biosciences, Menlo Park, CA, USA) was used to bind the library complex with MagBeads before sequencing to reduce adapter dimers. The Mag-Beads SMRT bell-polymerase complexes were loaded into a 96 well plate. The plate, along with a DNA sequencing kit 6.0, was loaded onto the instrument. Each SMRT cell was loaded with a single binding complex and 360 min movies were collected on a Pacific Biosciences RSII sequencer (Pacific Biosciences, Menlo Park, CA, USA).

#### Sequencing Data Analysis

The sequencing data generated by the Pacific Biosceince RSII sequencer was analysed using RS\_resequencing pipeline on the onboard software provided in the SMRT portal (software v2.2, https://github.com/PacificBiosciences/SMRT-Analysis/wiki/SMRT-Analysis-Software-Installation-v2.2.0) using the construct design sequences as a reference file. The depth of coverage across the constructs was determined and plotted using the BED format file produced by the PacBio portal, facilitating identification of regions of construct coverage dropout. SNPs and INDELs were called as part of the SMRT portal RS\_Resequencing workflow after consensus sequence identification using quiver (https://github.com/PacificBiosciences/GenomicConsensus).

**Table 1** Experimental Design: 5’ and 3’ Overhangs generated after digestion with BbsI enzyme. The final size of each fragment, the status and the variant calling is showed.



**Figure 1** Schematic SMRT Gate library preparation and sequencing reaction: Double-stranded DNA fragments are generated using *BbsI (BpiI*) restriction sites (purple) and release from the backbone. The barcoded (red) SMRT Gate hairpin adapters (blue) are ligated onto double-stranded DNA fragments using the overhang generated by the cleavage. The DNA polymerases (grey) was bound to the SMRT library to generate polymerase reads (red). Dashed lines represent single-stranded DNA.

**Figure 2** Smrt Gate Library (pool 1) QC Tapestation Electropherograms. Electropherogram before SMRT Gate reaction (1) before Exonuclease treatment (2) and final SMRT Gate library (3)

**Figure 3** Example of Variant Calling. Alignment of 11019 Quiver (Q11019) and Sanger (S11019) Reads mapped to the Reference (R11019) showed a Single nucleotide Polymorphism (SNP) at position 208 (A). Alignment of Quiver consensus (Q11065, Q11066) and reference sequences (R11065, R11066) for 11065 and 11066 showed  a new variant (B).and concordance in SNP calling (C)

### Acknowledgement

Sequencing data was performed at the Centre for Genomic Research at the University of Liverpool. All plasmids were provided by Nicola Patron (NP). LD and AH were funded by a BBSRC grant (BB/M00094X/1) and the Pacific Bioscience RS II was bought with a capital grant from the BBSRC (BB/L014777/1). We would like to thank Chris Lounds and Adam Peltan from NEB for advice and recommendation on enzymatic assays, John Harting from Pacific Bioscience for the advice on bioinformatics tools. Hannah McCue form the GeneMill team for her support.

### Authors’ contributions

LD and AH conceptualized the study, LD performed the experiments and interpretation of the data. JK and JJ reviewed the experiment design. LD and MH performed the sequencing experiments. SH and RJ assisted with bioinformatic analyses. NP, AH, CHF and NH assisted in the data interpretation. LD and AH wrote the manuscript. All authors read and approved the paper.

1. **Slomovic, S., K. Pardee, and J.J. Collins**. 2015. Synthetic biology devices for in vitro and in vivo diagnostics. Proc Natl Acad Sci USA *112*:14429–14435.

2. **Georgianna, D.R., and S.P. Mayfield**. 2012. Exploiting diversity and synthetic biology for the production of algal biofuels. Nature *488*:329–335.

3. **Bryksin, A.V., A.C. Brown, M.M. Baksh, M.G. Finn, and T.H. Barker**. 2014. Learning from nature – Novel synthetic biology approaches for biomaterial design. Acta Biomaterialia *10*:1761–1769.

4. **Keasling, J.D.** 2012. Synthetic biology and the development of tools for metabolic engineering. Metabolic Engineering *14*:189–195.

5. **Zhang, W., and D. Nielsen**. 2014. Synthetic biology applications in industrial microbiology. Front. Microbiol. *5*.

6. **Molyneux-Hodgson, S., and A.S. Balmer**. 2014. Synthetic biology, water industry and the performance of an innovation barrier. Sci Public Policy *41*:507–519.

7. **Engler, C., R. Kandzia, and S. Marillonnet**. 2008. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. PLoS ONE *3*:e3647.

8. **Engler, C., R. Gruetzner, R. Kandzia, and S. Marillonnet**. 2009. Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. PLoS ONE *4*:e5553.

9. **Weber, E., C. Engler, R. Gruetzner, S. Werner, and S. Marillonnet**. 2011. A Modular Cloning System for Standardized Assembly of Multigene Constructs. PLoS ONE *6*:e16765.

10. **Kanigowska, P., Y. Shen, Y. Zheng, S. Rosser, and Y. Cai**. 2015. Smart DNA Fabrication Using Sound Waves: Applying Acoustic Dispensing Technologies to Synthetic Biology. Journal of Laboratory Automation.

11. **Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, et al.** 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science *323*:133–138.

12. **Roberts, R.J., M.O. Carneiro, and M.C. Schatz**. 2013. The advantages of SMRT sequencing. Genome Biol *14*:405.

13. **Storch, M., A. Casini, Ben Mackrow, T. Fleming, H. Trewhitt, T. Ellis, and G.S. Baldwin**. 2015. BASIC: A New Biopart Assembly Standard for Idempotent Cloning Provides Accurate, Single-Tier DNA Assembly for Synthetic Biology. ACS Synth. Biol. *4*:781–787.

14. **Sarrion-Perdigones, A., M. Vazquez-Vilar, J. Palací, B. Castelijns, J. Forment, P. Ziarsolo, J. Blanca, A. Granell, and D. Orzaez**. 2013. GoldenBraid 2.0: a comprehensive DNA assembly framework for plant synthetic biology. Plant Physiol *162*:1618–1631.

15. **De Paoli, H.C., G.A. Tuskan, and X. Yang**. 2016. An innovative platform for quick and flexible joining of assorted DNA fragments. Scientific Reports *6*:19278.