Prediction of Golden Gate Assembly (GGA) Using a Comprehensive Analysis of T4 DNA Ligase End-Joining Fidelity and Bias

SEED 2018
Poster 87

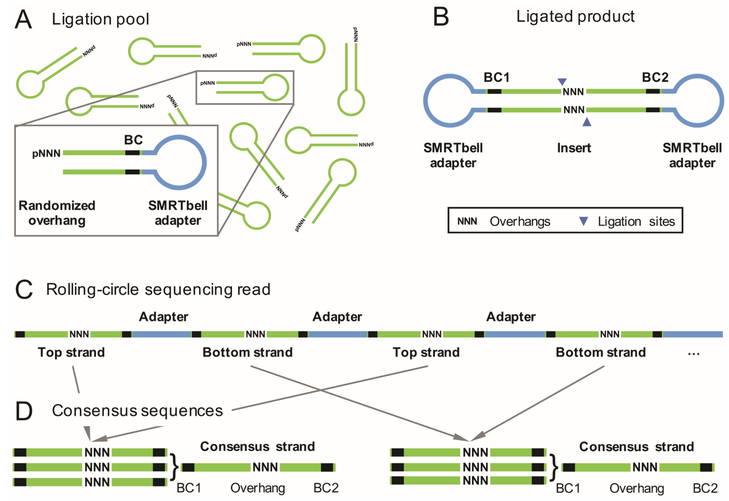
Vladimir Potapov[1], Jennifer L. Ong[1], Rebecca B. Kucera[1], Bradley W. Langhorst[1], Katharina Bilotti[1], John M. Pryor[1], Eric J. Cantor[1], Barry Canton[2], Thomas F. Knight[2], Thomas C. Evans, Jr.[1], and Gregory J. S. Lohman[1]

[1]New England Biolabs, Ipswich, MA 01938, USA [2]Ginkgo Bioworks, Boston, MA, 02210, USA

Abstract

The one-pot assembly of long DNA sequences from multiple component parts is key to the rapid generation of constructs for modern synthetic biology. Methods for the one-pot assembly of multiple fragments linked by short overhangs (e.g. Golden Gate) depend on accurate and unbiased ligation. Design of junctions to date largely depends on the use of rules of thumb and empirical success, rather than detailed data on ligase fidelity and bias. In this study, we have applied Pacific Biosciences Single-Molecule Real-Time sequencing technology to directly measure of the ligation frequency of every possible 5′-four-base overhang pairing in a single experiment. This comprehensive data set has been applied to predict the accuracy of Golden Gate assembly (GGA) using the Type IIS restriction enzyme BsaI. Ten fragment assemblies were designed based on the ligation data with junctions predicted to result in high or low fidelity assembly. Experimental results confirmed not only the overall accuracy, but the specific mismatch ligation errors observed and their relative frequency. The data was further used to design 12- or 24-fragment assemblies of the lac operon, which were shown to assemble with high fidelity and efficiency. Thus, ligase fidelity data allows the prediction of high-accuracy overhang pair sets with greater flexibility in design than the rules of thumb, allowing assembly of >20 fragments at high-accuracy junction points even within defined coding regions without modification of the native DNA sequence.

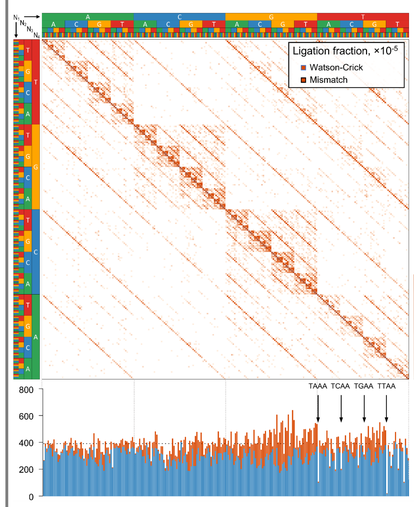
Schematic of Multiplexed Ligation Profiling Assay

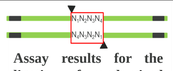


(A) Libraries containing randomized three-base overhangs were synthesized and ligated with T4 DNA ligase under various conditions. The hairpin substrates contain the SMRTbell adapter sequence as well as an internal 6-base random barcode used to confirm strand identity and monitor the substrate sequence bias derived from oligonucleotide synthesis. (B) Ligated substrates form circular molecules, in which a double-stranded insert DNA is capped with SMRTbell adapters. (C) Ligated products were sequenced utilizing PacBio SMRT sequencing, which produced long rolling-circle sequencing reads. Sequencing reads are comprised of regions corresponding to top and bottom strands separated by regions corresponding to SMRTbell adapters. (D) Consensus sequences were built for the top and bottom strands independently, allowing extraction of the overhang identity and barcode sequence.

Methods/Results

Fidelity and Bias: Four-Base Overhangs

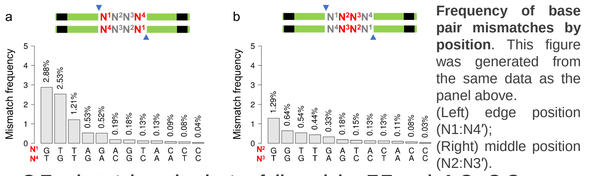




Assay results for the ligation of randomized four-base overhangs by T4 DNA Ligase 25°C. SMRT sequencing results for ligating 100 nM of the multiplexed four-base overhang substrate 18h at 25°C, with 1.75 μM T4 DNA ligase in standard ligation buffer. (Top) Frequency heat map of all ligation events (log-scaled). Watson-Crick (WC) pairings are shown on the diagonal. (Bottom) Stacked bar plot showing the frequency of ligation products containing each overhang. Fully Watson-Crick paired ligation results are indicated in blue, and ligation products containing one or more mismatches are in orange.

- **Method allows comprehensive profiling of fidelity for a given ligase/end-structure pairing in one experiment.**
- **Overhangs range in fidelity from <50% WC to >99%.**
- **Most ligate with similar efficiency, but TNNA overhangs are notably inefficient.**
- **For even very low-fidelity overhangs, mismatch ligation products result from only a few (2 – 8) other ligation partners.**
- **Increasing temperature increases fidelity, but also increases bias against AT-rich overhangs.**
- **PEG drastically reduces fidelity.**

Position-Dependence of Mismatches

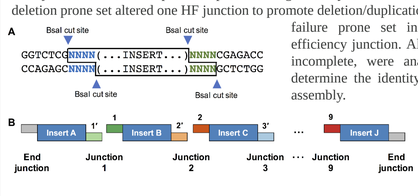


Frequency of base pair mismatches by position. This figure was generated from the same data as the panel above. (Left) edge position (N1:N4′); (Right) middle position (N2:N3′).

- **G:T mismatches dominate, followed by T:T and A:G. G:G appear internally but rarely at the ligation junction.**
- **Mismatches are 2x more likely at the edge than the middle.**
- **At higher temperature, same mismatches but reduced frequency.**

Ligase Fidelity is Predictive in GGA

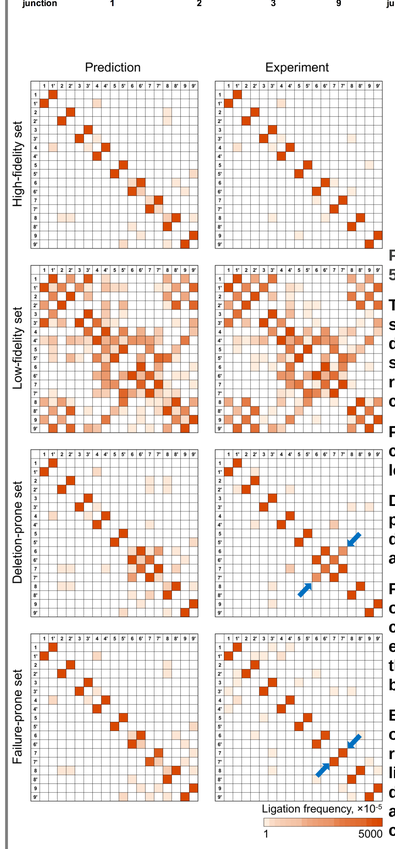
Overview of Golden Gate assembly design. Ten fragments of arbitrary sequence were designed, with junctions chosen to provide predicted high-fidelity or low-fidelity per the below table. A deletion prone set altered one HF junction to promote deletion/duplication of Fragment G, and a failure prone set including a predicted low efficiency junction. All products, full length and incomplete, were analyzed by sequencing to determine the identity of every junction in the assembly.



Junction[2]	High-fidelity set	Low-fidelity set
1 / 1′	AAGG / TTCC	GCCC / GGGG
2 / 2′	ACTC / TGAG	GCCA / CGGT
3 / 3′	AGGA / TCCT	ACCC / TGGG
4 / 4′	AGTC / TCAG	AGCC / TCGG
5 / 5′	ATCA / TAGT	CGCC / GCGG
6 / 6′	CCGC / CGGC	AGCA / TCCT
7 / 7′	CTGA / GACT	AGCG / TCGC
8 / 8′	GCGA / CGCT	CGGC / GCCG
9 / 9′	GGAA / CCTT	AGGC / TCCG

[1] A notation of —— indicates the junction pair used is identical to the HF set.

[2] In the Deletion-prone set, Junction 7 is replaced with $\frac{CGGC}{GCCG}$, predicted to result in frequent deletion of Fragment G. In the Failure-prone set, Junction 7 is replaced with $\frac{CAAA}{ATTT}$, a predicted inefficient junction.



Protocol: 37°C 5 min/16°C 5 min, 30 cycles.

The WC connections are shown in the double diagonal visible in the HF set. All other squares represent mismatched connections.

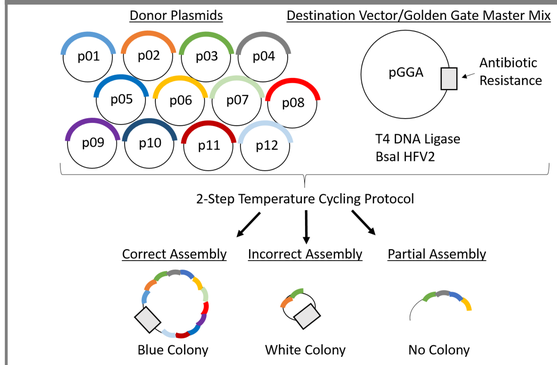
Predictions match results closely, excepting only lowest frequency events.

Deletion prone set shows predicted deletion and duplications (indicated by arrows).

Failure prone set shows only a modest drop In connections at the low efficiency junction, less than predicted (indicated by arrows).

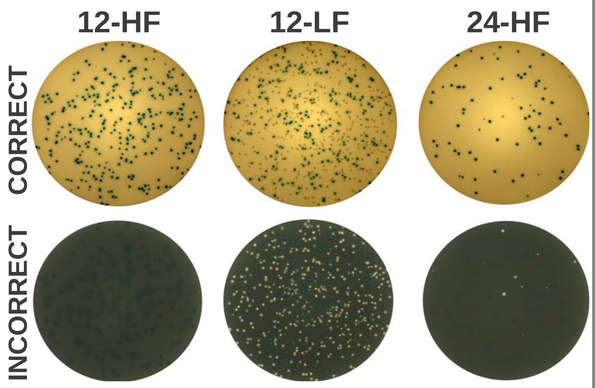
Bias against 100% GC overhangs observed. This result was not predicted by ligation assay, and is likely due to ligation profiling assay not accounting for cutting and melting steps.

12 and 24-Fragment One-Pot GGA



Overview of *Lac* cassette assembly Assay. A cassette for β-galactosidase (β-gal) expression was split into 12 or 24-fragments at junctions chosen based on the fidelity data and placed in holding vectors, with each fragment flanked by BsaI sites. In each assembly reaction, the fragments were combined with a pGGA destination vector containing the chloramphenicol resistance gene. A high-fidelity (HF) 12 and 24 fragment system was designed, as well as an intentionally deletion-prone low-fidelity (LF)12 fragment system.

Assemblies containing all fragments in the proper order result in β-gal expression and a blue colony; incorrect assembly leads to a white colony; partial assembly leads to no colony.



Protocol: 30 cycles (16°C 5 min, 37°C 5min), 5 min 60C end hold.

Results match predictions closely:
12-HF predicted 99% correct, observed 99.2 ± 0.6 %
12-LF predicted 31% blue colonies, observed 45 ± 5%.
24-HF predicted 91% blue colonies, observed 84 ± 5%.

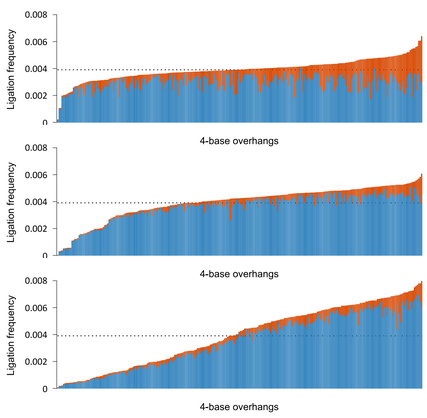
Recommended Overhangs for GGA

Set	Number of overhangs	Estimated fidelity	Overhang sequences [1]
1 (MoClo Compatible)	15	98.5%	TGCC, GCAA, ACTA, TTAC, CAGA, TGTG, GAGC, AGGA, ATTC, CGAA, ATAG, AAGG, AACT, AAAA, ACCG
2	20	98.1%	AGTG, CAGG, ACTC, AAAA, AGAC, CGAA, ATAG, AACC, TACA, TAGA, ATGC, GATA, CTCC, GTAA, CTGA, ACAA, AGGA, ATTA, ACCG, GCGA
3	25	95.8%	CCTC, CTAA, GACA, GCAC, AATC, GTAA, TGAA, ATTA, CCAG, AGGA, ACAA, TAGA, CGGA, CATA, CAGC, AACG, AAGT, CTCC, AGAT, ACCA, AGTG, GGTA, GCGA, AAAA, ATGA
4	30	91.7%	TACA, CTAA, GGAA, GCCA, CACG, ACTC, CTTC, TCAA, GATA, ACTG, AACT, AAGC, CATA, GACC, AGGA, ATCG, AGAG, ATTA, CGGA, TAGA, AGCA, TGAA, ACAT, CCAG, GTGA, ACGA, ATAC, AAAA, AAGG, CAAC

[1] Only one member of each complementary overhang pair is shown.

Set 1 is an extended MoClo set (TGCC, GCAA, ACTA, TTAC, CAGA, TGTG, GAGC) with additional 8 overhangs.
All sets are predicted to assemble with the specified overall fidelity if every overhang and its complement is used; subsets of these sets are predicted to have even higher fidelity.

Why not T7 Ligase?



T7 provides a modest fidelity boost, at the cost of dramatic bias and lower overall efficiency.

Much greater improvements are possible through careful junction choice guided by fidelity and bias data.

Summary/Conclusions

- Our new sequencing assay permits the rapid profiling of fidelity and bias in end-joining. Many sequences can be tested in a single pot.
- Fidelity and bias data for four-base overhangs predicts GGA accuracy.
- Specific erroneous junctions can be predicted in kind and degree
- Current data underestimates bias against 100% GC WC pairings.
- Use of data to guide junction overhang choice permits assembly of up to 24 fragments in one pot with high accuracy and efficiency.
- We have provided predicted best junction sets for GGA taking into account fidelity and efficiency.
- Future work will refine the data sets and examine bias of RE cutting.

REFERENCES: NAR Methods Online: https://doi.org/10.1093/nar/gky303 BioRxiv: https://doi.org/10.1101/322297