Published online 8 May 2018

Nucleic Acids Research, 2018, Vol. 46, No. 13 ***e79***
doi: 10.1093/nar/gky303

A single-molecule sequencing assay for the comprehensive profiling of T4 DNA ligase fidelity and bias during DNA end-joining

Vladimir Potapov[1], Jennifer L. Ong[1], Bradley W. Langhorst[2], Katharina Bilotti[1], Dan Cahoon[3], Barry Canton[3], Thomas F. Knight[3], Thomas C. Evans, Jr[1] and Gregory J.S. Lohman[1,*]

[1]Research Department, New England Biolabs, Ipswich, MA 01938, USA, [2]Applications and Product Development, New England Biolabs, Ipswich, MA 01938, USA and [3]Ginkgo Bioworks, Boston, MA 02210, USA

Received January 22, 2018; Revised March 13, 2018; Editorial Decision April 09, 2018; Accepted April 12, 2018

ABSTRACT

DNA ligases are key enzymes in molecular and synthetic biology that catalyze the joining of breaks in duplex DNA and the end-joining of DNA fragments. Ligation fidelity (discrimination against the ligation of substrates containing mismatched base pairs) and bias (preferential ligation of particular sequences over others) have been well-studied in the context of nick ligation. However, almost no data exist for fidelity and bias in end-joining ligation contexts. In this study, we applied Pacific Biosciences Single-Molecule Real-Time sequencing technology to directly sequence the products of a highly multiplexed ligation reaction. This method has been used to profile the ligation of all three-base 5′-overhangs by T4 DNA ligase under typical ligation conditions in a single experiment. We report the relative frequency of all ligation products with or without mismatches, the position-dependent frequency of each mismatch, and the surprising observation that 5′-TNA overhangs ligate extremely inefficiently compared to all other Watson–Crick pairings. The method can easily be extended to profile other ligases, end-types (e.g. blunt ends and overhangs of different lengths), and the effect of adjacent sequence on the ligation results. Further, the method has the potential to provide new insights into the thermodynamics of annealing and the kinetics of end-joining reactions.

INTRODUCTION

DNA ligases, present in all domains of life and in many viruses, are critical enzymes for *in vivo* genome replication and maintenance. Ligases, especially the DNA ligase from bacteriophage T4, have also proven essential to molecular biology methodologies, including cloning, next-generation sequencing library preparation, gene synthesis and molecular diagnostics (1–3). DNA ligases canonically join the 3′-hydroxyl of one DNA strand to the 5′-phosphoryl group of another when both are hybridized to a complementary DNA. Flexibility in substrate structure has been documented for a variety of ligases, with many tolerating ribonucleotides, gaps, or base-pair mismatches during the ligation of a break in one strand of a duplex DNA (nick ligation) (4–9). Several ligases, including T4 DNA ligase, can also efficiently join two DNA fragments with short complementary ssDNA overhangs or blunt ends, an activity of critical importance to many modern molecular biology methodologies (10–13).

Ligase fidelity, the ability to discriminate against mismatches at the ligation junction, has been well studied for several ligases in the context of nick ligation (6,8,14–18). DNA ligases are proposed to sterically sense mismatches in nicked DNA through distortion in DNA helix shape, and prefer mismatch pairings that can be accommodated within the normal helix diameter, pairings with multiple hydrogen bonds, or both (19–21). However, the specific mismatches tolerated vary from ligase to ligase, and these preferences must be determined empirically. Past studies profiling ligase mismatch tolerance have typically analyzed individual nick substrates in parallel or in small pools, allowing measurement of relative rates of ligation of nick substrates containing mismatched base pairs (15,22–24). In general, ligases have been found to be more tolerant of mismatches at the side of the junction providing the 5′-phosphate. For example, T4 DNA ligase can ligate all mismatched base pairings at the 5′ side, but prefers C:T, G:T, A:C and T:T mismatches at the 3′-hydroxyl side to the exclusion of others (where the mismatched bases are listed as the ligation junction base:templating strand base) (6,23). The high-fidelity *Thermus thermophilus* (*Tth*) DNA ligase is less tolerant of mis-

*To whom correspondence should be addressed. Tel: +1 978 998 7916; Fax: +1 978 921 1350; Email: lohman@neb.com

matches on the 5′-side of the ligation junction than T4 DNA ligase, readily ligating T:T, T:G, A:C and C:A mismatches, with lesser amounts of G:T, C:C, A:A and G:A ligation (15). Asymmetrical preferences are common in ligation mismatch tolerance; for example, T4 DNA ligase prefers T:C to C:T at the 3′-OH side of the junction, and *Tth* DNA ligase prefers T:G to G:T at the 5′-side of the ligation junction (6,15,23).

While DNA ligase fidelity and bias has been studied for nick ligation, there remains a need to systematically characterize mismatch tolerance in an end-joining context. Here, we report a single-molecule, next-generation sequencing assay to probe the fidelity of DNA ligase end-joining, from a mixed population of ssDNA overhangs. Pacific Biosciences Single-Molecule Real-Time (SMRT) Sequencing allows for true single-molecule sequencing without PCR amplification of the DNA, gaining high accuracy by reading an insert many times via a rolling-circle replication mechanism (25,26). In our assay, hairpin substrates incorporate the SMRTbell adapter and a short three-base 5′-overhang (Figure 1). The overhang region is randomized such that it contains all possible three-base overhangs in approximately equal proportion. Ligation of these randomized pools creates libraries with SMRTbell adapters on both ends and an insert region generated from the ligation of two overhangs. SMRT sequencing of the ligated libraries allows direct read out of both overhangs from single ligation products (27). This method allows the systematic profiling of ligation events for all possible overhangs in a single experiment, with the frequency of each product assumed to be proportional to the efficiency of that particular end-joining reaction. We have applied this method to characterize the ligation of three base overhangs by T4 DNA ligase under typical reaction conditions, allowing the comprehensive evaluation of the fidelity and bias of the reaction.

MATERIALS AND METHODS

All enzymes and buffers were obtained from New England Biolabs (NEB, Ipswich, MA, USA) unless otherwise noted. T4 DNA ligase reaction buffer (1×) is: 50 mM Tris–HCl (pH 7.5), 10 mM $MgCl_2$, 1 mM ATP, 10 mM DTT. NEBuffer 2 (1×) is: 10 mM Tris–HCl (pH 7.9), 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT. CutSmart Buffer (1×) is: 20 mM Tris-acetate (pH 7.9), 50 mM Potassium Acetate, 10 mM Magnesium Acetate, 100 μg/ml BSA. All column cleanup of oligonucleotides and ligated libraries was performed using Monarch® PCR & DNA Cleanup Kit columns (NEB), following the published Oligonucleotide Cleanup Protocol (https://www.neb.com/protocols/2017/04/25/oligonucleotide-cleanup-using-monarch-pcr-dna-cleanup-kit-5-g-protocol-neb-t1030). Oligonucleotide purity and sizing was performed using an Agilent Bioanalyzer 2100, using a DNA 1000 assay, following the standard protocols.

Preparation of substrates for multiplexed ligation fidelity and bias profiling assay

Initial PAGE-purified substrate precursor oligonucleotide was obtained as a lyophilized solid (IDT). The sequence (Table 1) contained a 5′-terminal region, a randomized three-base region, a SapI binding site, a constant region, an internal 6-base randomized region as a control for synthesis bias, and a region corresponding to the SMRTbell sequencing adapter for PacBio SMRT sequencing. The oligonucleotide was designed with a short (7-base) complementary region such that they form a primer–template junction hairpin structure (Figure 1). The precursor oligonucleotide was dissolved in 1× annealing buffer (10 mM Tris pH 8.0, 50 mM NaCl, 0.1 mM EDTA) to a final concentration of 100 μM. In a final reaction volume of 200 μl, the substrate precursor oligonucleotide (40 μl of 100 μM stock) was combined with 200 U Klenow Fragment (3′→5′ exo-), 0.4 U yeast inorganic pyrophosphatase in NEBuffer 2 at 1× final concentration and dNTPs at 1 mM each final concentration. The extension reaction was incubated 1 h at 37°C, 2 μl Proteinase K was added and the reaction incubated 20 min at 37°C. The DNA was purified (Monarch® PCR & DNA Cleanup Kit), and the concentration of the purified DNA (typically 25–30 μM) was determined using an Agilent Bioanalyzer 2100, DNA 1000 kit.

The extended DNA was cut using SapI to generate a three-base overhang. Typically, 1 μl of DNA from the extension reaction was combined with 900 U of SapI in a 100 μl total volume of NEB CutSmart buffer and incubated for 2 h at 37°C. Reactions were halted by addition of 1 μl Proteinase K followed by 20 min incubation at 37°C, then purified using the Monarch® PCR & DNA Cleanup Kit (NEB). Final concentration and extent of cutting was determined by Agilent Bioanalyzer (DNA 1000) and confirmed to be >95% cut. Remaining uncut starting material (∼5%) was not 5′ phosphorylated, and thus should not interfere with subsequent cohesive-end joining reactions. For use in subsequent steps, DNA substrates were diluted to ∼500 nM in 1× TE buffer, with precise concentration determined by Bioanalyzer. The final substrate sequence can be found in Table 1.

Preparation of ligation fidelity libraries for Pacific Biosciences SMRT sequencing

In a typical reaction, substrate (100 nM) was combined with 2.5 μl high concentration T4 DNA ligase (2000 U, 1.75 μM final concentration) in 1× T4 DNA ligase buffer in a 50 μl total reaction volume and incubated for 1 h or 18 h at 25°C or 37°C. Reactions were quenched with 2.5 μl 500 mM EDTA, and purified using the Monarch® PCR & DNA Cleanup Kit, oligonucleotide cleanup protocol. Each ligation was performed in a minimum of duplicates, and the ligation yield was determined by Agilent Bioanalyzer (DNA 1000) with error reported as one standard deviation. The ligated library was treated with exonuclease III (50 U) and exonuclease VII (5 U) in a 50 μl volume in 1× Standard Taq Polymerase buffer for a 60 min incubation at 37°C. The library was purified using a Monarch® PCR & DNA Cleanup Kit, oligonucleotide cleanup protocol, including a second wash step, then quantified by Agilent Bioanalyzer (DNA 1000). Typical concentrations of final library were between 0.5 and 2 ng/μl.

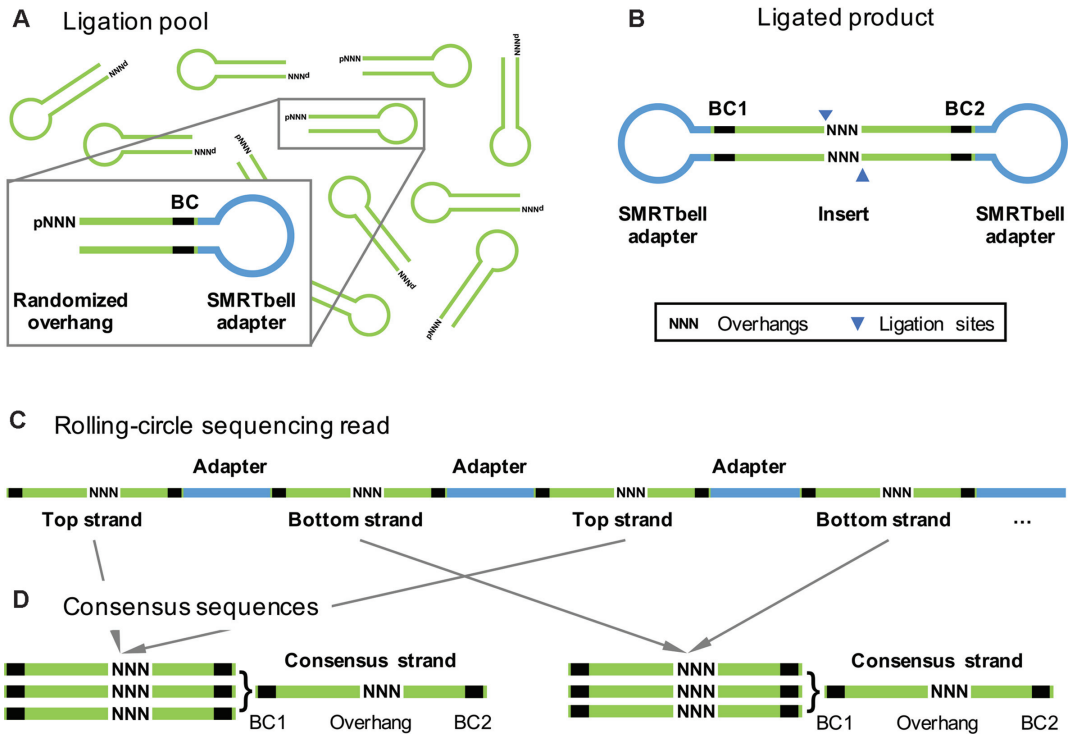

Figure 1. Schematic of multiplexed ligation fidelity and bias profiling assay. (**A**) Libraries containing randomized three-base overhangs were generated and ligated with T4 DNA ligase under various conditions. The hairpin substrates contain the SMRTbell adapter sequence as well as an internal 6-base 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.

Table 1. Precursor and substrate sequences

Substrate	Sequence[a]
Ligation library precursor	TCAGGTNNNC**GAAGAGC**TGCGATCCAGTGCGCCGTGCATTGATCAACGC AANNNNNNATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAG
Ligation library substrate	pNNNC**GAAGAGC**TGCGATCCAGTGCGCCGTGCATTGATCAACGCAANNNN NNATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATNNNN NNTTGCGTTGATCAATGCACGGCGCACTGGATCGCA**GCTCTTC**G
Expected insert[b]	NNNNNNTTGCGTTGATCAATGCACGGCGCACTGGATCGCAGCTCTTCG NNNCGAAGAGCTGCGATCCAGTGCGCCGTGCATTGATCAACGCAANNN NNN

[a]The SapI (type IIS restriction enzyme) binding site is indicated in bold. SMRT adapter region is underlined.
[b]The expected insert length in 99nt. The location of three-base overhang is in position 49..51, 3′-randomized region is in position 1..6, and 5′-randomized region is position 94..99.

Pacific biosciences SMRT sequencing

Ligated overhang substrates form a circular molecule, comprising a double-stranded insert DNA capped with SMRTbell adapters (Figure 1). Libraries were prepared for sequencing according to the Pacific Biosciences Binding Calculator Version 2.3.1.1 and the DNA/Polymerase Binding Kit P6 v2 using the standard protocol, no-DNA control complex, and a custom concentration on plate (0.3375 nM). Libraries were sequenced on a Pacific Biosciences RSII, 1–8 SMRT cells per library, 3 h data collection time per cell, with 'stage start' off.

PacBio SMRT sequencing of each ligated product resulted in a long rolling-circle sequencing read. The sequencing read was comprised of regions corresponding to inserts (summarized in Table 1) from top and bottom strands interspersed by regions corresponding to SMRTbell adapters (Figure 1). A computational workflow was developed (i) to separate insert sequences from opposite strands, (ii) to build accurate consensus reads for each strand and (iii) to extract actual overhang sequences in each strand. For the first step, the longest stretch of inserts was identified in each polymerase read such that each insert is of expected size and the distance between any two inserts is of expected length

corresponding to the length of SMRTbell adapter. A 20% variation in length of either insert or adapter was allowed to account for the high single-pass indel rate of PacBio SMRT sequencing technology. Inserts corresponding to the same strand were grouped together, and a consensus sequence of each strand was built with the Arrow algorithm using the ccs program from SMRT Link software (https://github.com/PacificBiosciences/GenomicConsensus). At least five subreads per strand were required to build a consensus; reads from which fewer than five subreads per strand were found were discarded. Resulting consensus sequences were aligned to the respective insert reference sequences using BLASR aligner, and sequence fragments corresponding to overhangs and barcodes were extracted for each strand. A number of filtering steps were applied to avoid sequencing artifacts and ensure integrity of the derived data. The length of derived overhangs was required to be exactly three nucleotides. Three bases, immediately adjacent to the overhangs on either side, were required to match the reference sequence. Additionally, the length of the derived barcodes was required to be six nucleotides, and respective barcodes in the opposite strands were required to be complementary. One mismatch per barcode was permitted to account for possible replicative errors. Frequencies of all observed overhang pairs in ligation products were tabulated and used to derive results. To avoid bias due to arbitrary definition of top and bottom strands in ligation products, overhang pairs were counted twice: once in the top/bottom direction, and once in the bottom/top direction.

Determining substrate sequence bias resulting from oligonucleotide synthesis

For each multiplexed ligation profile library, the internal barcodes were characterized to assess the degree of sequence bias introduced by the oligonucleotide synthesis. For each read, the barcode sequence was extracted, and the fraction of each base at each position determined. The reported values are for the barcode strand produced by the oligonucleotide synthesis; the complementary strand, inserted during the polymerase extension step, has the complementary ratios. The error is reported as the standard deviation of all six positions from all sequencing runs using a given substrate.

Oligonucleotide ligation assay

Standard ligation assay mixtures were composed of 1× T4 reaction buffer, 350 nM T4 DNA ligase and 100 nM FAM-labeled DNA substrate, in a reaction volume of 100 μl. Reactions were performed at 25°C. Components were gently mixed by pipetting and incubated at reaction temperature for 5 min prior to initiation by the addition of the DNA substrate. Reactions were quenched by a 1:1 (vol/vol) addition of 50 mM EDTA at times as indicated in each figure legend. The ligated products were analyzed by capillary gel electrophoresis as described previously (23,28,29). Reported values are the average of a minimum of three replicates, with the error bars representing the standard deviation of the measurements. Sequences used in this assay can be found in the Supplementary Data, Table S1.

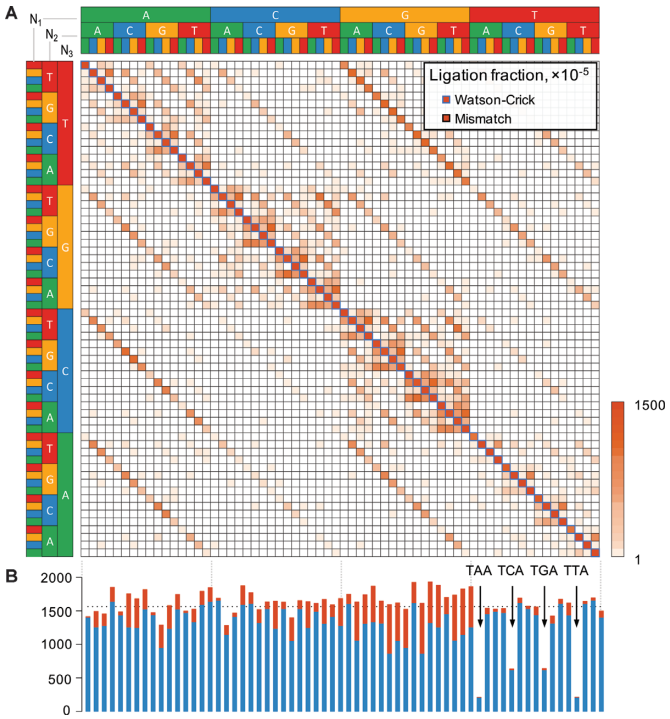

Figure 2. Assay results for the ligation of randomized three-base overhangs by T4 DNA ligase (1 h at 25°C). SMRT sequencing results for ligating 100 nM of the multiplexed three-base overhang substrate 1 h at 25°C, with 1.75 μM T4 DNA ligase in standard ligation buffer. Observations have been normalized to 100 000 ligation events (see Supplementary Data for actual observation totals). (**A**) Frequency heat map of all ligation events (log-scaled). Overhangs are listed alphabetically left to right (AAA, AAC, AAG …TTG, TTT) and bottom to top such that the Watson–Crick pairings are shown on the diagonal. (**B**) Stacked bar plot showing the frequency of ligation products containing each overhang, corresponding to each column in the heat map in (A). Fully Watson–Crick paired ligation results are indicated in blue, and ligation products containing one or more mismatches are in orange.

RESULTS

As a key test case, the method was used to profile the end-joining fidelity of T4 DNA ligase under typical reaction conditions. Ligation libraries were prepared using standard buffer conditions for each overhang pool with ligation temperatures of 25 or 37°C and ligation times of 1 or 18 h. These reactions contain a large excess of ligase (1.75 μM) over ligatable ends (100 nM); thus, the results represent single-turnover ligation conditions. This ratio was chosen as it is similar to the standard ratio recommended in typical cloning protocols, and should be representative of results that would be expected during molecular biology experiments that require end-joining by T4 DNA ligase. Under these conditions, the yield of ligation product was 80 ± 1% for 1 h at 25°C (88 ± 1% at 18 h) and 75 ± 1% for 1 h at 37°C (79 ± 4% at 18 h). Data from replicates were combined before analysis; for an examination of sequencing reproducibility between replicates, see Supplementary Data (Supplementary Text, Figure S1 and Table S2).

The multiplexed ligation profile results for three-base overhangs for 1 h at 25°C are shown in Figure 2, and reported in a tabular format in Table 2. The results for 1 h

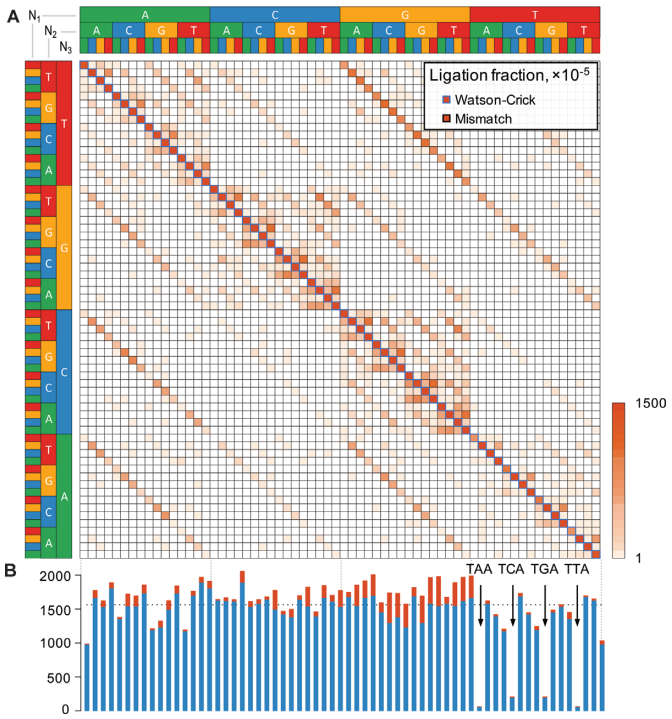

Figure 3. Assay results for the ligation of randomized three-base overhangs by T4 DNA ligase (1 h at 37°C). SMRT sequencing results for ligating 100 nM of the multiplexed three-base overhang substrate 1 h at 37°C, with 1.75 μM T4 DNA ligase in standard ligation buffer. Observations have been normalized to 100 000 ligation events (see Supplementary Data for actual observation totals). (**A**) Frequency heat map of all ligation events (log-scaled). Overhangs are listed alphabetically left to right (AAA, AAC, AAG . . . TTG, TTT) and bottom to top such that the Watson–Crick pairings are shown on the diagonal. (**B**) Stacked bar plot showing the frequency of ligation products containing each overhang, corresponding to each column in the heat map in (A). Fully Watson–Crick paired ligation results are indicated in blue, and ligation products containing one or more mismatches are in orange.

ligation at 37°C are shown in Figure 3 and Supplementary Data, Table S3. In each figure, panel A shows a log-scale frequency heat map of all ligation events, with the overhangs sorted such that the top left to bottom right diagonal represents Watson–Crick paired ligation products (the highest frequency events for any given overhang). This visualization allows for any pair of overhangs leading to significant mismatch ligation to be easily identified. Panel B shows the linear frequency of ligation events for each overhang as a bar plot, with the Watson–Crick ligation frequency shown in blue, and the summed frequency of mismatch products shown in orange. The data in Panel B is sorted in the same order as the heat maps. Results changed little from short to long incubation time at each temperature, despite slightly increased library yields (Supplementary Data, Figures S2 and S3, Tables S4 and S5).

In discussing the results in the following sections, individual overhangs are written in the 5′ to 3′ direction with the phosphate omitted, and ligation products are written as overhang pairs with the top overhang written in the 5′ to 3′ direction and the bottom overhang in the 3′ to 5′ direction. For example, $\frac{ATT}{TAA}$ represents the fully Watson–Crick paired ligation product between a substrate with a 5′-pATT overhang and a substrate with a 5′-pAAT overhang.

TNA overhangs show greatly reduced ligation efficiency

While the majority of correctly base-paired ligation partners were observed in very similar overall frequency (Figures 2B and 3B, blue bars), it was noted at both 25°C and 37°C that the four TNA overhangs had notably reduced incidence compared to the average. The corresponding ANT overhangs, despite being expected to be present in the exact same proportion of the initial substrate pool, did not show a reduced incidence compared to the other overhangs in the set.

This discrepancy suggested that the ligation of TNA overhangs may be fundamentally inefficient. To confirm this hypothesis, defined, fluorescently-labeled dsDNA substrates with 5′-three-base overhangs were ligated and the degree of ligation monitored over time. Indeed, ligation of defined substrate pools of TNA and ANT substrates (Figure 4A) showed that substrates with a TNA overhang ligated ∼100× slower than the same substrate with an ANT overhang. This result indicated that the low incidence of ligation products resulting from TNA overhangs was indeed a result of dramatically lower ligation rates compared to other sequences.

Ligation fidelity of overhangs varies dramatically with sequence

The range of observed ligation fidelity as a function of overhang identity was quite broad, from overhangs with very few observed mismatch ligation events (e.g. AAA, ATA, CAA) to those where a very large fraction of observations (>40%) found a ligation partner with one or more base-pair mismatches (e.g. GCC and GGC). Overall, there was a weak trend towards lower fidelity with higher G/C content. More specifically, for three-base overhangs, 5′-GNN sequences were highly represented in the lowest-fidelity overhangs, with 5′-TNN over-represented in the highest fidelity group. Increasing temperature (37°C, Figure 3) resulted in an overall suppression of mismatch ligation, with no significant change to the overall patterns: the Kendall rank correlation coefficient equals 0.87 when comparing individual ligation fractions observed at 25°C and 37°C (only ligations with more than 10 counts per ligation event were considered). Additionally, there was minimal effect on the identity of the mismatches observed (Figure 5). Mismatch rates at 37°C were 2-fold lower than at 25°C as measured by the fraction of mismatch ligations (orange bars in Figures 2B and 3B; mean mismatch ligation fraction per overhang was 2.53×10^{-3} and 1.29×10^{-3}, respectively).

For most overhangs, even those of the lowest fidelity, it should be noted that the bulk of the mismatch ligation events were derived from pairing with only a few (typically 2–3) other overhangs (Table 2). For example, GGC was one of the lowest fidelity overhangs, with only 54% of observations showing it paired with its Watson–Crick partner, GCC. However, its mispairing events are dominated by GCT (43% of mismatch ligation events) and ACC (38% of mismatch ligation events). For CGT, 76% of ligations

Table 2. Ligation fidelity for three-base overhangs (1 h at 25°C)

Overhang	Correct, $\times 10^{-5}$		Mismatch, $\times 10^{-5}$		Fidelity, %[a]		Mismatch overhangs[b]
	Value	S.D.	Value	S.D.	Value	S.D.	
AAA	1403.8	14.2	11.1	4.1	99.2	0.3	TTC (43%); ATT (12%); TGT (10%)
TAA	209.0	10.8	2.9	1.9	98.6	0.9	GTA (36%); TTT (27%); TTG (18%)
ATA	1472.5	11.9	24.9	3.1	98.3	0.2	AAT (44%); GAT (26%); TTT (13%)
CAA	1658.4	0.2	30.0	8.0	98.2	0.5	GTG (54%); ATG (34%); TGG (2%)
TTC	1608.3	29.6	34.2	2.5	97.9	0.2	GAG (42%); GAT (17%); AAA (14%)
TTG	1658.4	0.2	36.3	7.4	97.9	0.4	CAG (46%); CAT (22%); CTA (9%)
AGA	1440.7	7.9	32.4	18.9	97.8	1.3	GCT (24%); ACT (23%); TCA (16%)
TAG	1491.6	12.2	35.3	0.3	97.7	0.0	CTG (55%); CTT (35%); CTC (5%)
TTA	209.0	10.8	5.0	2.0	97.6	1.0	GAA (21%); AAA (21%); TAG (11%)
TCA	617.9	1.1	16.2	4.2	97.4	0.6	AGA (31%); GGA (28%); TGG (11%)
TCG	1531.7	10.1	41.9	9.0	97.3	0.6	CGG (68%); CGT (9%); AGA (6%)
TGA	617.9	1.1	21.5	1.2	96.6	0.2	GCA (43%); ACA (31%); TCT (10%)
ACA	1437.3	19.3	50.4	15.3	96.6	1.0	TGC (54%); TGG (16%); TGA (13%)
CAG	1415.5	35.6	51.2	0.8	96.5	0.1	TTG (33%); ATG (19%); GTG (19%)
TGG	1609.6	39.8	66.0	18.0	96.1	1.1	CCG (41%); CCT (20%); ACA (12%)
TAT	1472.5	11.9	63.9	3.6	95.8	0.2	GTA (71%); ATG (17%); ATT (7%)
TCC	1619.4	43.6	71.6	18.6	95.8	1.2	GGG (68%); GGT (14%); AGA (5%)
TAC	1453.7	30.9	87.0	3.7	94.4	0.3	GTG (69%); GTT (26%); GTC (3%)
CGA	1531.7	10.1	98.1	1.3	94.0	0.1	ACG (51%); GCG (41%); CCG (2%)
CCG	1531.1	69.8	101.0	24.2	93.8	1.6	AGG (39%); TGG (27%); CTG (14%)
TTT	1403.8	14.2	94.7	8.5	93.7	0.5	GAA (82%); AAG (4%); AGA (4%)
CGG	1531.1	69.8	120.9	29.1	92.7	1.9	ACG (41%); TCG (24%); GCG (11%)
CTA	1491.6	12.2	120.4	23.9	92.5	1.4	AAG (56%); GAG (32%); TTG (3%)
TGC	1316.9	18.0	107.4	20.8	92.5	1.3	GCG (47%); ACA (25%); GCT (20%)
TCT	1440.7	7.9	119.6	7.8	92.3	0.4	GGA (81%); AGG (13%); TGA (2%)
GAA	1608.3	29.6	139.5	15.5	92.0	0.9	TTT (56%); GTC (27%); ATC (12%)
CCA	1609.6	39.8	162.3	8.0	90.8	0.6	AGG (73%); GGG (23%); CGG (2%)
CAC	1147.7	67.3	133.6	8.9	89.6	1.1	ATG (54%); GCG (17%); GTT (13%)
ATG	1593.4	13.4	198.9	20.5	88.9	1.1	CAC (36%); CTT (33%); TAT (5%)
CTG	1415.5	35.6	177.7	12.4	88.8	0.9	CTG (50%); AAG (12%); TAG (11%)
TGT	1437.3	19.3	181.6	6.7	88.8	0.2	GCA (83%); ACG (11%); ACT (3%)
ATT	1638.5	19.7	207.1	11.2	88.8	0.4	GAT (84%); AGT (4%); ATT (3%)
AAT	1638.5	19.7	211.6	21.3	88.6	0.9	GTT (74%); ATC (16%); ATA (5%)
AAG	1279.7	15.5	175.8	12.3	87.9	0.9	CTC (40%); CTA (38%); CTG (13%)
ATC	1336.5	8.9	188.5	19.1	87.6	1.2	GAC (33%); AAT (17%); GGT (14%)
AGT	1526.9	0.5	217.2	27.7	87.5	1.4	GCT (83%); ATT (4%); ACC (3%)
CCC	1322.4	107.4	197.6	68.8	87.0	4.8	AGG (61%); GTG (20%); GAG (6%)
GTA	1453.7	30.9	248.5	7.0	85.4	0.6	AAC (39%); GAC (31%); TAT (18%)
CAT	1593.4	13.4	281.4	38.2	85.0	1.9	GTG (87%); ACG (3%); TTG (3%)
AAC	1259.1	35.1	234.7	18.4	84.3	0.7	GTC (42%); GTA (41%); GTG (12%)
GGA	1619.4	43.6	305.0	27.6	84.2	1.6	GCC (33%); TCT (32%); ACC (31%)
ACT	1526.9	0.5	289.8	20.6	84.0	1.0	GGT (64%); AGC (24%); AGG (5%)
GCA	1316.9	18.0	330.4	2.2	79.9	0.3	TGT (45%); AGC (33%); GGC (15%)
CTC	1313.9	25.9	358.8	10.9	78.6	0.8	GTG (57%); AAG (19%); GCG (12%)
AGG	1236.8	51.8	342.9	52.8	78.3	3.3	CCC (35%); CCA (35%); CCG (11%)
CGT	1250.0	0.8	386.9	35.2	76.4	1.7	GCG (90%); GAG (2%); AGG (2%)
CTT	1279.7	15.5	405.7	18.1	75.9	0.6	GAG (71%); ATG (16%); AGG (4%)
GAG	1313.9	25.9	425.9	4.2	75.5	0.5	CTT (67%); CTA (9%); CTG (4%)
CGC	1055.4	20.0	342.1	33.9	75.5	2.2	ACG (82%); GTG (5%); GCT (4%)
CCT	1236.8	51.8	408.9	48.8	75.2	3.0	GGG (90%); TGG (3%); ATG (2%)
ACG	1250.0	0.8	432.2	44.7	74.3	2.0	CGC (65%); CGA (12%); CGG (12%)
AGC	950.7	21.1	337.8	13.5	73.8	1.2	GCA (32%); GCC (31%); ACT (20%)
ACC	1258.3	110.3	498.5	58.2	71.6	4.1	GGC (57%); GGA (19%); GGG (14%)
GAT	1336.5	8.9	533.5	16.6	71.5	0.8	GTC (59%); ATT (33%); ATG (1%)
GGG	1322.4	107.4	608.8	90.4	68.5	4.9	CCT (61%); ACC (11%); TCC (8%)
GTT	1259.1	35.1	603.3	20.1	67.6	1.3	GAC (55%); AAT (26%); ACC (4%)
GGT	1258.3	110.3	623.4	97.8	66.9	5.4	GCC (60%); ACT (30%); ATC (4%)
GCG	1055.4	20.0	564.8	33.5	65.1	1.8	CGT (61%); TGC (9%); CTC (7%)
GAC	1061.5	2.1	570.1	36.7	65.1	1.4	GTT (58%); GTA (13%); ATC (11%)
GTG	1147.7	67.3	679.6	48.1	62.8	3.0	CAT (36%); CTC (30%); TAC (9%)
GCT	950.7	21.1	572.5	46.7	62.4	2.5	GGC (56%); AGT (32%); TGC (4%)
GTC	1061.5	2.1	673.5	26.0	61.2	0.9	GAT (47%); AAC (15%); GTC (12%)
GCC	864.7	68.8	730.8	98.5	54.2	5.4	GGT (51%); AGC (14%); GGA (14%)
GGC	864.7	68.8	748.9	48.2	53.6	3.6	GCT (43%); ACC (38%); GCA (7%)

Standard deviations were derived from two experimental replicates, while the values themselves were derived from the combined data.
[a]Fidelity is calculated as the fraction of correct ligations divided by the total fraction of ligations for a given overhang.
[b]Top three mismatch overhangs are given for each overhang. All overhangs are written in the 5′-to-3′ direction. The numbers in parentheses give the percentage for the given mismatch ligation relative to the total number of mismatch ligations for the overhang.
Fidelity for other reaction conditions can be found in the Supplementary Data, Tables S3–S5.

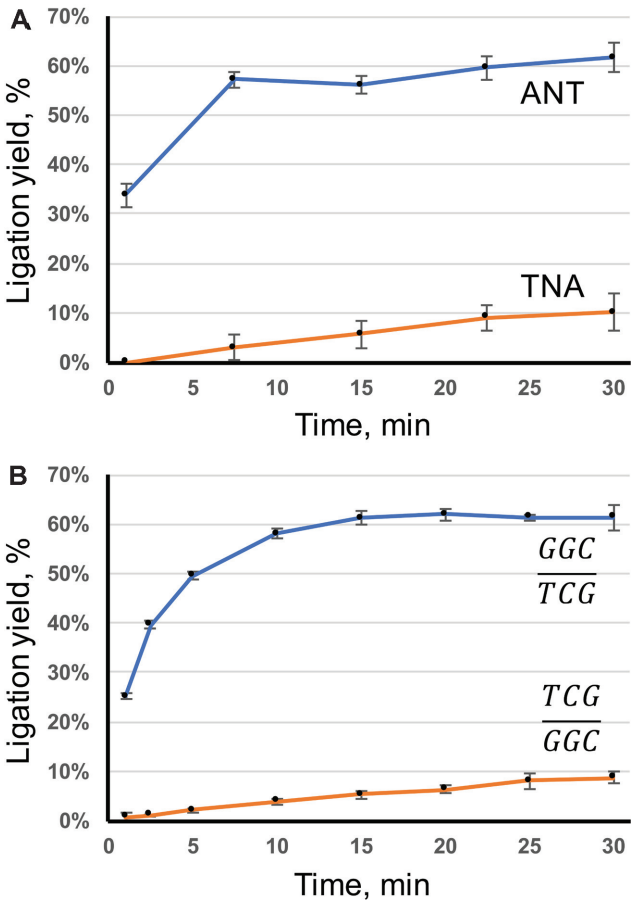

Figure 4. Ligation time courses for defined substrates. Ligation of three-base overhang substrates, FAM-labeled on the 3′-end of the phosphorylated strand. Ligation reactions were composed of 1× T4 reaction buffer, 350 nM ligase and 100 nM FAM-labeled DNA substrate. Reactions were performed at 25°C, removing time points over a 2 h incubation. Reported values are the average of a minimum of three replicates, with the error bars the standard deviation of the measurements. (**A**) Substrate with an ANT overhang versus a TNA overhang. (**B**) Ligation of $\frac{GGC}{TCG}$ pair, 100 nM each overhang, versus ligation of a $\frac{TCG}{GGC}$ pair.

were with its Watson–Crick partner, with nearly all of the observed mismatch ligations coming from pairing with the overhang GCG (90% of mismatch ligation events). Importantly, the overwhelming majority (98%) of mismatch overhangs at 1 h at 25°C had a single-base mismatch; two- and three-base mismatches were 1.8% and <0.1%, respectively. Very similar results were observed for 1 h at 37°C (one, two and three-base mismatches were 97.6%, 2.2% and <0.2%, respectively). Thus, individual overhangs that exhibit low fidelity are not promiscuous, rather they tend to pair with only a few specific mismatched sequences.

Ligation preference for mismatches varies by position and strand sense and is context-sensitive

Analysis of the observed mismatch ligation events further allowed identification of the types of mismatches tolerated by the ligase, and the effect of position and structural context on ligation preferences. Figure 5 shows the observed frequency of ligated mispairs at the 'edge' position versus the 'middle' position for three-base overhangs at 25°C and 37°C. As with the overall fidelity, increasing temperature had little effect on the specific mismatch pairings observed or their relative frequency to each other, simply reducing overall mismatch frequency relative to Watson–Crick paired products (Supplementary Data, Figure S4).

At the edge position of three-base overhangs (N1:N3′, Figure 5A and C), mismatch ligation was dominated by G:T mispairs, which make up 52% of all observed N1:N3′ mismatches (mismatches at this position total 7.1% of all ligation events, with G:T mismatches at this position comprising 3.7% of all ligation events). Interestingly, this preference was only for 5′-G mispairs, $\frac{GNN}{TNN}$; the reciprocal $\frac{TNN}{GNN}$ mismatch was not especially prevalent (3.7% and 0.3% of all observed ligations, respectively). The ligation of a $\frac{GNN}{TNN}$ mispair was independently assayed and compared to the ligation of the reciprocal $\frac{TNN}{GNN}$ mispair to confirm this observation. Indeed, the 5′-G mismatch ligated ∼80-fold faster than the 5′-T mismatch (Figure 4B). The preference for edge G:T mismatches accounted for the over-representation of GNN overhangs displaying low overall fidelity in Figures 2 and 3; a G:T simply ligated much faster than any other mispairing in this position. After G:T, several purine:purine (A:A, A:G, and G:A) and the other purine:pyrimidine (A:C, C:A, T:G) mismatches were observed, with pyrimidine:pyrimidine mismatches disfavored (Figure 5A). Adenine displayed the same 5′ mismatch preference; the $\frac{ANN}{CNN}$ mispairing was ∼40-fold more prevalent than the reciprocal $\frac{CNN}{ANN}$ mispairing (1.3% and 0.03% of ligation events, respectively).

At the middle position of three-base overhangs (N2:N2′, Figure 5B and D), a lower overall frequency of mismatches was observed, ∼3-fold less common compared to the total frequency for mismatches at the edge position (7.1% and 2.3% for edge and middle mismatches, respectively). At the middle position, T:T mismatches were modestly favored, ∼2-3 times more prevalent than any other single mismatch, with C:T, T:C, G:T and T:G all present at similar frequencies (Figure 5B). A:C, C:A, C:C and purine:purine mismatches appeared to be strongly disfavored at these positions.

DISCUSSION

Our results revealed the mismatch tolerance of T4 DNA ligase when joining short overhangs. In the end-joining reactions described here, the substrate did not have randomized bases at the 3′-OH side of the ligation junction, as the sequence adjacent to the overhangs was kept constant; thus, only mismatches at the 5′-phosphate side of the junction could be observed. T4 DNA ligase is very tolerant of mismatches on 5′-phosphate of nicks, making it difficult to predict from previous studies what mismatches to expect in end-joining (6,23). In this study, we nevertheless observed specific mismatch preferences that line up with the general preferences of ligases in nick ligation (15,23). Specifically, mismatches observed were those that could form multiple hydrogen bonds (A:G, C:A, A:A), those which can be accommodated within the normal helix diameter upon enzyme binding (T:T), and those that meet both criteria (G:T) (20,21). As with nick ligation, G:T mismatches were broadly

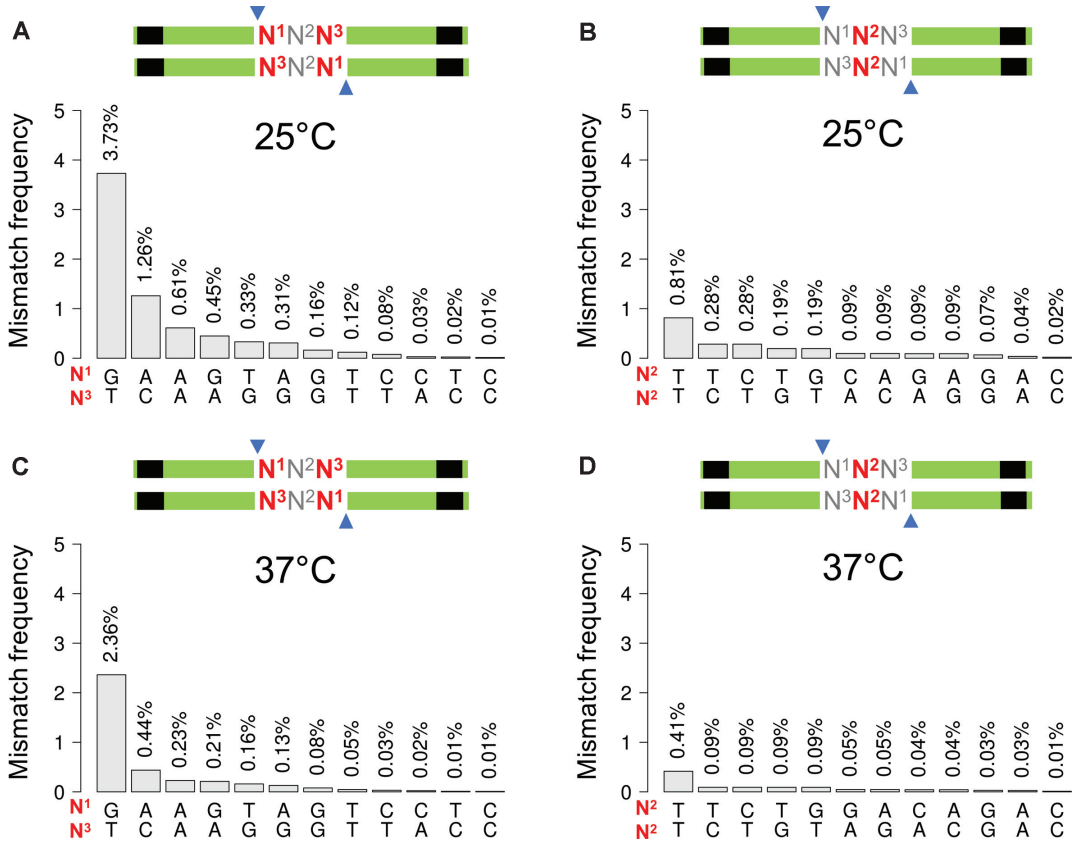

Figure 5. Frequency of specific base pair mismatches by position. Incidence of each possible mismatched base pair observed for ligation of three-base overhangs. The results shown are for SMRT sequencing of ligation reactions with 100 nM of the multiplexed three-base overhang substrate, 1 h at 25°C (**A** and **B**) or 37°C (**C** and **D**), with 1.75 μM T4 DNA ligase in standard ligation buffer. This figure was generated from the same data as shown in Figures 2 and 3. A and C show the results for the edge position (N1:N3′); B and D for the middle position (N2:N2′).

tolerated in cohesive end ligation, as are T:T, and to a lesser extent, A:C, A:G and A:A mismatches. Interesting asymmetries were also observed; i.e. G:T mismatches are much more (∼10-fold) prone to ligation than T:G, and A:C mismatches much more prone to ligation than C:A (∼40-fold). This result suggests that the ligase active site prefers a phosphorylated purine over a pyrimidine, a result that is reciprocal to the preference previously observed at this position for Taq DNA ligase (15).

Mismatches in the middle position were very poorly tolerated (2.3% of all ligation events), with overall incidence much lower than at the edge positions and only T:T mismatches present in relative high frequency (36% of all mismatches at this position; Figure 5C). This result suggests that two correct base pairs in a row are required at the overlap for efficient ligation. It is unclear, however, if the low incidence of mismatches at this position is an enzyme effect, with helix distortions at this position disrupting the active site, or an annealing effect, with the mismatch in the middle position disrupting stable end annealing. A survey of predicted ΔG of annealing of mismatched overhangs (Supplementary Data, Figure S5) shows no correlation between ΔG and observation frequency; the overwhelming determinant of ligation frequency is the type of mismatch, not the stability of annealing (30). Likely, annealing thermodynamics do play some role in mismatch ligation frequency. For a given type of mismatch (Supplementary Data, Figures S6 and S7), there is a weak correlation between increasing ΔG and observations, though there is still quite a bit of scatter in these plots indicating that the sequence context is also highly important, not just the annealing strength/number of hydrogen bonds. Further, increasing ligation temperature greatly reduces mismatch ligation; this result is almost certainly an annealing effect, destabilizing the pool of available annealed mismatched substrates for the ligase to act on. High-GC overhangs display a higher frequency of mismatch ligation, especially when two of three bases form G:C Watson–Crick pairs. Among overhangs with single-base mismatches, the two correct base pairings were both G:Cs in 50% of cases, G:Cs and A:Ts in 42% of cases, and both A:Ts in 8% of cases at 25°C (with similar ratios at 37°C). Interestingly, it is not simply the GC content, but the location of G:Cs, that makes mismatch ligation more favorable. For example, overhangs with single-base edge mismatches containing middle A:T pairs and edge G:C base pairs ($\frac{XAG}{YTC}$, where X:Y indicates mismatch) outnumber the reverse ($\frac{XGA}{YCT}$) by a factor of 2.5 (32% versus 13%, respectively), despite having similar predicted ΔG of annealing (e.g. the G:T panel, Supplementary Data, Figure S6).

When interpreting the quantitative results, it should be noted that overhang pools were not perfectly random and contained bias introduced during the initial oligonucleotide

synthesis. The distribution of nucleobases in the random synthesized region was 25.4±0.3%, 27.0±0.4%, 25.6±0.6%, and 22.0±0.3%, for A, T, G and C, respectively. This base bias did not vary significantly by position. In the particular substrate used, the most abundant overhang (TTT) is estimated to be about 2-fold more prevalent than the least abundant (CCC). While it would be desirable to normalize the data based on the bias in the substrate library, the ligation reaction is a combination of dynamic annealing of overhangs and ligase kinetic preferences for each annealed overhang pair; thus, it was unclear how to normalize in a way that is consistent with this complex biochemistry. Consequently, mismatch products resulting from particularly over- or under-represented overhangs may have their frequency over- or under-represented. The strongest effects are expected to be on the substrates with the weakest annealing and/or the weakest K_M, which will be the most responsive to changes in substrate concentration. Thus, the lowest-frequency observed (poor annealing and/or weak binding) substrate pairs are likely to be those most strongly influenced by the sequence synthesis bias. Future work will focus on deconvoluting the effects of annealing versus enzyme kinetic preferences, potentially allowing for data normalization and, thus, more quantitatively accurate predictions.

The most surprising result uncovered in this study was the remarkably sluggish ligation rate of TNA overhangs as compared to all other Watson–Crick pairings, including ANT and other high AT overhangs. We attempted to determine if this result might be due to unusual secondary structure, either of the unligated substrate, the paired hairpins, or the nicked intermediate formed after the first intermolecular ligation step. To this end, Vienna was used to predict the structure of the full hairpin, the ligated product with one nick, and the first 18 bases of the hairpin substrate with or without its complement strand (30). In no case did we find predictions of unusual secondary structure that would explain the slow ligation rate, or any other significant predicted differences from substrates with ANT in this position. ΔG of annealing was also predicted to be in line with other high AT overhangs (Supplementary Data, Figure S5). Thus, we do not expect the poor reactivity of the TNA substrates to be a result of DNA structure alone, and must result from interaction with the ligase. There is unfortunately no reported crystal structure of T4 DNA ligase; a co-structure with this substrate sequence might be able to show if the TNA sequence somehow formed a complex with the ligase active site that pulled the bases out of alignment for adenylyl transfer and/or phosphodiester bond formation. Mechanistically, we observed very little adenylylated intermediate in the ligation of the defined substrates (Figure 4A), suggesting slow adenylyl group transfer. This is in contrast to the reaction of T4 ligase with other inefficient end-joining substrates, such as blunt ends and single base overhangs, which show substantial adenylylated substrate accumulation under comparable single-turnover conditions used in this study (31). However, additional work, including with analogous nicked substrates, will be required to definitively determine the mechanism of the slow reaction of TNA overhangs.

The current method has proven effective in rapidly profiling the ligation fidelity of T4 DNA ligase in a single experiment. Further application of the method will allow for the profiling of any ligase that can carry out the ligation of short, cohesive ends, and measuring the influence of reaction conditions that are likely to influence annealing and/or ligase kinetics (e.g. ionic strength, buffer pH, divalent cation concentration and identity) on ligation fidelity and bias. The effect of additives, such as the crowding agent polyethylene glycol, commonly used to enhance ligation rates, could also be measured; this class of additives accelerates reaction rate by increasing effective concentration of substrates, which may have a significant effect on ligation fidelity and bias (32,33).The method could be used to find not only high-fidelity ligases and conditions, but potentially to find ligases with very different fidelity profiles and mismatch preferences. The latter could be substituted for T4 DNA ligase in cases where the use of low-fidelity ligation pairs is desired, or allow for identification of ligases optimized for particular applications.

Additionally, there is potential for use of this data to explore the kinetics and thermodynamics of end-joining ligation. Varying ligase identity, concentration, reaction time, and temperature could allow the deconvolution of the contributions of annealing and ligase substrate preference on each reaction, potentially allowing the extraction of substrate-dependent kinetic parameters. The current study is performed entirely under single-turnover conditions, with a large excess of enzyme over ligatable ends. At minimum, varying the enzyme:substrate ratio and including multiple turnover studies will be necessary to deconvolute substrate-dependent kinetics. Deeper analysis will also require the use of a nicked version of the substrate pool to separate out the differences between the first, intermolecular ligation event and the subsequent intramolecular nick ligation. Another potential approach, such as our previously published multiplexed nick ligation profiling method, would be better suited to complement SMRT sequencing analysis of the end-joining substrates (23).

A comprehensive understanding of the fidelity and bias of cohesive end-joining may facilitate the optimization of methods requiring high-fidelity ligation; e.g. the ligase chain reaction and related ligation-dependent methods for detecting specific DNA sequences (34–37). This method could be further adapted to explore other end structures; for example, extending the length of the overhang to study how length of the annealing region affects specificity, or modifying the substrate to allow for profiling the sequence-dependent ligation bias of blunt ends and single-base overhangs. Knowledge of the sequence-dependent blunt and T/A ligation bias would be of great potential interest to the generation of DNA libraries for NGS sequencing, including potential effects from using adapters with different sequences. Application and extension of this methodology thus promises to generate helpful foundational data for the optimization of many modern molecular biology protocols.

DATA AVAILABILITY

Sequencing data pertaining to this study has been deposited into the Sequencing Read Archive under accession number SRP130363. Custom software tools are available in

the GitHub repository at: https://github.com/potapovneb/ligase-fidelity.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Laurence Ettwiller, Laurie Mazzola, Rick Morgan, Yvette Luyten (NEB) and Pacific Biosciences for assistance with sequencing reactions. We are grateful to Bill Jack, Andy Gardner, Eric Cantor and Karen Lohman for critical feedback on this manuscript.

FUNDING

Internal funding from NEB and Ginkgo Bioworks. Funding for open access charge: New England Biolabs.

Conflict of interest statement. Vladimir Potapov, Jennifer L. Ong, Bradley W. Langhorst, Katharina Bilotti, Thomas C. Evans, Jr, Gregory J.S. Lohman are employees of New England Biolabs, a manufacturer and vendor of molecular biology reagents, including DNA ligases. This affiliation does not affect the authors' impartiality, adherence to journal standards and policies, or availability of data.

Dan Cahoon, Barry Canton, and Thomas F. Knight are employees of Ginkgo Bioworks, Inc., a corporation that uses enzymes and reagents for gene synthesis in the course of developing engineered microbes. This affiliation does not affect the authors' impartiality, adherence to journal standards and policies, or availability of data.

REFERENCES

1. Tomkinson,A.E., Vijayakumar,S., Pascal,J.M. and Ellenberger,T. (2006) DNA ligases: structure, reaction mechanism, and function. *Chem. Rev.*, **106**, 687–699.
2. Pascal,J.M. (2008) DNA and RNA ligases: structural variations and shared mechanisms. *Curr. Opin. Struct. Biol.*, **18**, 96–105.
3. Shuman,S. (2009) DNA ligases: progress and prospects. *J. Biol. Chem.*, **284**, 17365–17369.
4. Nilsson,S.V. and Magnusson,G. (1982) Sealing of gaps in duplex DNA by T4 DNA ligase. *Nucleic Acids Res.*, **10**, 1425–1437.
5. Goffin,C., Bailly,V. and Verly,W.G. (1987) Nicks 3′ or 5′ to AP sites or to mispaired bases, and one-nucleotide gaps can be sealed by T4 DNA ligase. *Nucleic Acids Res.*, **15**, 8755–8771.
6. Wu,D.Y. and Wallace,R.B. (1989) Specificity of the nick-closing activity of bacteriophage T4 DNA ligase. *Gene*, **76**, 245–254.
7. Harada,K. and Orgel,L.E. (1993) Unexpected substrate specificity of T4 DNA ligase revealed by in vitro selection. *Nucleic Acids Res.*, **21**, 2287–2291.
8. Sriskanda,V. and Shuman,S. (1998) Specificity and fidelity of strand joining by Chlorella virus DNA ligase. *Nucleic Acids Res.*, **26**, 3536–3541.
9. Showalter,A.K., Lamarche,B.J., Bakhtina,M., Su,M.I., Tang,K.H. and Tsai,M.D. (2006) Mechanistic comparison of high-fidelity and error-prone DNA polymerases and ligases involved in DNA repair. *Chem. Rev.*, **106**, 340–360.
10. Deugau,K.V. and van de Sande,J.H. (1978) T4 polynucleotide ligase catalyzed joining of short synthetic DNA duplexes at base-paired ends. *Biochemistry*, **17**, 723–729.
11. Sgaramella,V. and Ehrlich,S.D. (1978) Use of the T4 polynucleotide ligase in the joining of flush-ended DNA segments generated by restriction endonucleases. *Eur. J. Biochem.*, **86**, 531–537.
12. Pheiffer,B.H. and Zimmerman,S.B. (1983) Polymer-stimulated ligation: enhanced blunt- or cohesive-end ligation of DNA or deoxyribooligonucleotides by T4 DNA ligase in polymer solutions. *Nucleic Acids Res.*, **11**, 7853–7871.
13. Kukshal,V., Kim,I.K., Hura,G.L., Tomkinson,A.E., Tainer,J.A. and Ellenberger,T. (2015) Human DNA ligase III bridges two DNA ends to promote specific intermolecular DNA end joining. *Nucleic Acids Res.*, **43**, 7021–7031.
14. Shuman,S. (1995) Vaccinia virus DNA ligase: specificity, fidelity, and inhibition. *Biochemistry*, **34**, 16138–16147.
15. Luo,J., Bergstrom,D.E. and Barany,F. (1996) Improving the fidelity of Thermus thermophilus DNA ligase. *Nucleic Acids Res.*, **24**, 3071–3078.
16. Nakatani,M., Ezaki,S., Atomi,H. and Imanaka,T. (2002) Substrate recognition and fidelity of strand joining by an archaeal DNA ligase. *Eur. J. Biochem.*, **269**, 650–656.
17. Wang,Y., Lamarche,B.J. and Tsai,M.D. (2007) Human DNA ligase IV and the ligase IV/XRCC4 complex: analysis of nick ligation fidelity. *Biochemistry*, **46**, 4962–4976.
18. Schmier,B.J. and Shuman,S. (2014) Effects of 3′-OH and 5′'-PO4 base mispairs and damaged base lesions on the fidelity of nick sealing by Deinococcus radiodurans RNA ligase. *J. Bacteriol.*, **196**, 1704–1712.
19. Liu,P., Burdzy,A. and Sowers,L.C. (2004) DNA ligases ensure fidelity by interrogating minor groove contacts. *Nucleic Acids Res.*, **32**, 4503–4511.
20. Werntges,H., Steger,G., Riesner,D. and Fritz,H.J. (1986) Mismatches in DNA double strands: thermodynamic parameters and their correlation to repair efficiencies. *Nucleic Acids Res.*, **14**, 3773–3790.
21. Aboul-ela,F., Koh,D., Tinoco,I. Jr. and Martin,F.H. (1985) Base-base mismatches. Thermodynamics of double helix formation for dCA3XA3G + dCT3YT3G (X, Y = A,C,G,T). *Nucleic Acids Res.*, **13**, 4811–4824.
22. Tong,J., Cao,W. and Barany,F. (1999) Biochemical properties of a high fidelity DNA ligase from Thermus species AK16D. *Nucleic Acids Res.*, **27**, 788–794.
23. Lohman,G.J., Bauer,R.J., Nichols,N.M., Mazzola,L., Bybee,J., Rivizzigno,D., Cantin,E. and Evans,T.C. Jr (2016) A high-throughput assay for the comprehensive profiling of DNA ligase fidelity. *Nucleic Acids Res.*, **44**, e14.
24. Chauleau,M. and Shuman,S. (2016) Kinetic mechanism and fidelity of nick sealing by Escherichia coli NAD+-dependent DNA ligase (LigA). *Nucleic Acids Res.*, **44**, 2298–2309.
25. Roberts,R.J., Carneiro,M.O. and Schatz,M.C. (2013) The advantages of SMRT sequencing. *Genome Biol.*, **14**, 405.
26. D'Amore,R., Johnson,J., Haldenby,S., Hall,N., Hughes,M., Joynson,R., Kenny,J.G., Patron,N., Hertz-Fowler,C. and Hall,A. (2017) SMRT Gate: A method for validation of synthetic constructs on Pacific Biosciences sequencing platforms. *Biotechniques*, **63**, 13–20.
27. Potapov,V. and Ong,J.L. (2017) Examining sources of error in PCR by single-molecule sequencing. *PLoS One*, **12**, e0169774.
28. Lohman,G.J., Chen,L. and Evans,T.C. Jr (2011) Kinetic characterization of single strand break ligation in duplex DNA by T4 DNA ligase. *J. Biol. Chem.*, **286**, 44187–44196.
29. Greenough,L., Schermerhorn,K.M., Mazzola,L., Bybee,J., Rivizzigno,D., Cantin,E., Slatko,B.E. and Gardner,A.F. (2016) Adapting capillary gel electrophoresis as a sensitive, high-throughput method to accelerate characterization of nucleic acid metabolic enzymes. *Nucleic Acids Res.*, **44**, e15.
30. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
31. Bauer,R.J., Zhelkovsky,A., Bilotti,K., Crowell,L.E., Evans,T.C. Jr., McReynolds,L.A. and Lohman,G.J.S. (2017) Comparative analysis of the end-joining activity of several DNA ligases. *PLoS One*, **12**, e0190062.
32. Hayashi,K., Nakazawa,M., Ishizaki,Y. and Obayashi,A. (1985) Influence of monovalent cations on the activity of T4 DNA ligase in the presence of polyethylene glycol. *Nucleic Acids Res.*, **13**, 3261–3271.
33. Hayashi,K., Nakazawa,M., Ishizaki,Y., Hiraoka,N. and Obayashi,A. (1986) Regulation of inter- and intramolecular ligation with T4 DNA ligase in the presence of polyethylene glycol. *Nucleic Acids Res.*, **14**, 7617–7631.

34. Laffler,T.G., Carrino,J.J. and Marshall,R.L. (1993) The ligase chain reaction in DNA-based diagnosis. *Ann. Biol. Clin. (Paris)*, **51**, 821–826.
35. Wiedmann,M., Wilson,W.J., Czajka,J., Luo,J., Barany,F. and Batt,C.A. (1994) Ligase chain reaction (LCR)–overview and applications. *PCR Methods Appl.*, **3**, S51–S64.
36. Kim,J. and Lee,H.J. (2000) Rapid discriminatory detection of genes coding for SHV beta-lactamases by ligase chain reaction. *Antimicrob. Agents Chemother.*, **44**, 1860–1864.
37. Cheng,Y., Zhao,J., Jia,H., Yuan,Z. and Li,Z. (2013) Ligase chain reaction coupled with rolling circle amplification for high sensitivity detection of single nucleotide polymorphisms. *Analyst*, **138**, 2958–2963.