SUPPORTING INFORMATION

Sequencing depth in Ligase Fidelity Profiling experiments

To get reproducible experiments in Ligase Fidelity Profiling experiments, it is necessary to take into account the frequency of ligation products and required sequencing depth. Table S1 summarizes ligation products for the four-base substrate experiment. There are 32,896 possible unique ligation products. Of these, there are 136 unique fully complementary Watson-Crick ligation products (0 mismatches). They account for 74% of all observed ligation events with an average frequency of 5.5×10^{-3} per ligation product type. There are 1,536 possible unique single-mismatch ligation products, and most of them (1,460) are observed at least once in our experiment. They account for 24% of all observed ligation events and are, on average, ~30 times less frequent compared to Watson-Crick ligations (1.7×10^{-4}). For the typical PacBio Sequel II sequencing run, other ligation products are much less common and cannot be reliably quantified.

Table S1. Summary of ligation products for the four-base substrate

Number of mismatches	Number of possible ligation products	Number of detected ligation products	Total number of ligation events	Total fraction of ligation events	Average frequency
0	136	136	331,780	74.4435%	5.5×10^{-3}
1	1,536	1,460	108,378	24.3174%	1.7×10^{-4}
2	6,960	1,937	4,826	1.0828%	5.6×10^{-6}
3	13,824	408	575	0.1290%	3.2×10^{-6}
4	10,440	93	121	0.0271%	2.9×10^{-6}
	32,896	4,034	445,680	100.0000%	

To estimate how many sequencing reads (post filtering) are required get reproducible results, we draw two random samples of equal size from the total pool of observed ligation events and compute correlation coefficient between ligation frequencies in these two samples. These random samplings are repeated 1,000 times for each sample size, and the average correlation coefficient is reported (Figure S1). As can be seen in Figure S1, the sample size of ~10,000 reads results in $R = 0.9$ correlation for random replicates when considering only Watson-Crick pairs. To achieve the same level of reproducibility for single-mismatch ligations, it is necessary to obtain ~100,000 sequencing reads. Quantifying 2-, 3-, and 4-base mismatch ligations requires at least an order of magnitude more sequencing data, impractical on current generation Pacific Biosciences instruments

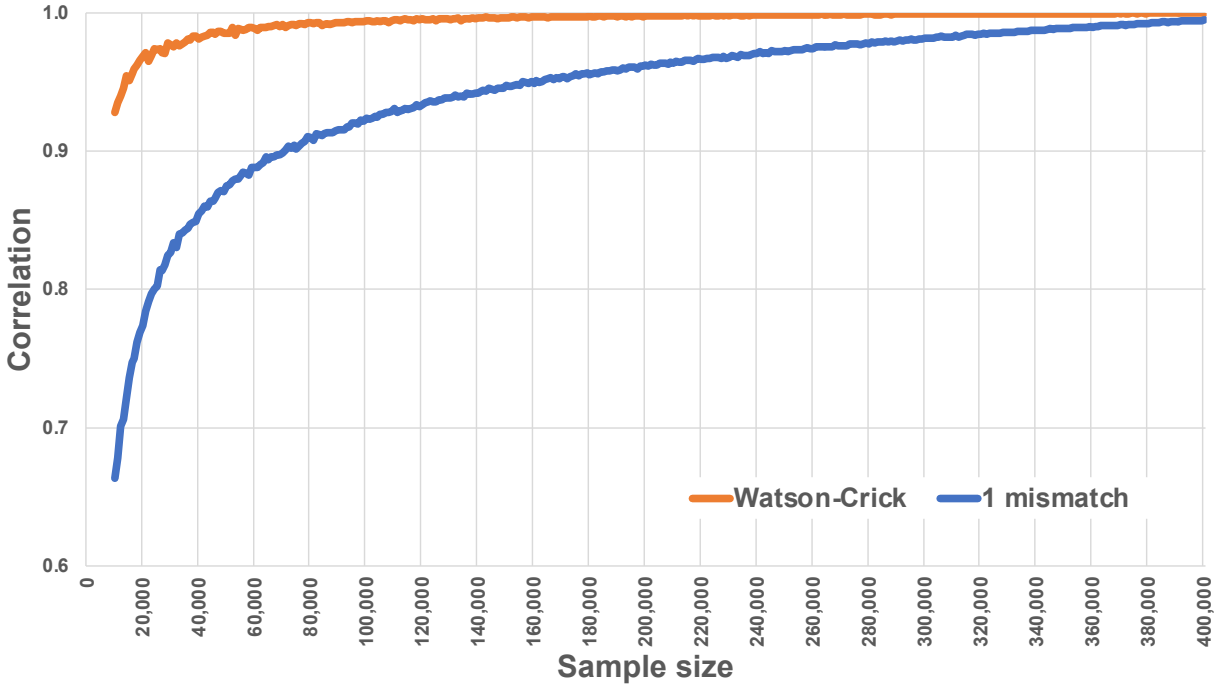

Figure S1. Estimating sequencing depth for Ligase Fidelity Profiling.

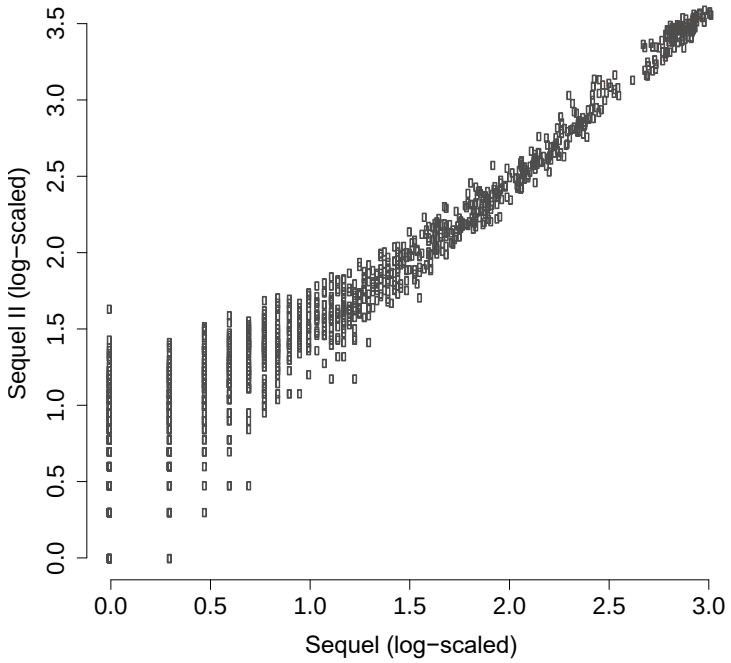

Figure S2. Agreement between observed overhang frequencies for Sequel and Sequel II instruments. Ligase Fidelity Profiling experiments were conducted for the same condition on PacBio Sequel and Sequel II instruments. The number ligation events were determined and plotted against each other (log-scaled). The Watson-Crick ligations (fully complementary ligations) are in orange, and the single-mismatch ligations are in grey. The total number of observed ligation events for Sequel and Sequel II instruments is 124,805 and 450,819, respectively. The Pearson correlation coefficient for the observed Watson-Crick ligations events is 0.97 and for the single-mismatch ligations is 0.94. There were 364 two-base mismatches with at least one observation in both Sequel and Sequel II experiments (not shown in the figure). The Pearson correlation coefficient for the observed ligation frequencies is 0.61.