

Supplementary Information for

PacBio sequencing output increased through uniform and directional 5-fold concatenation

Nisha Kanwar^{1,2#}, Celia Blanco^{3#}, Irene A. Chen³, & Burckhard Seelig^{1,2*}

¹Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, Minnesota, 55455, USA. ²BioTechnology Institute, University of Minnesota, St. Paul, Minnesota, 55108, USA. ³Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA 90095

Co-first authors, * Correspondence to nisha.kanwaruk@gmail.com or seelig@umn.edu.

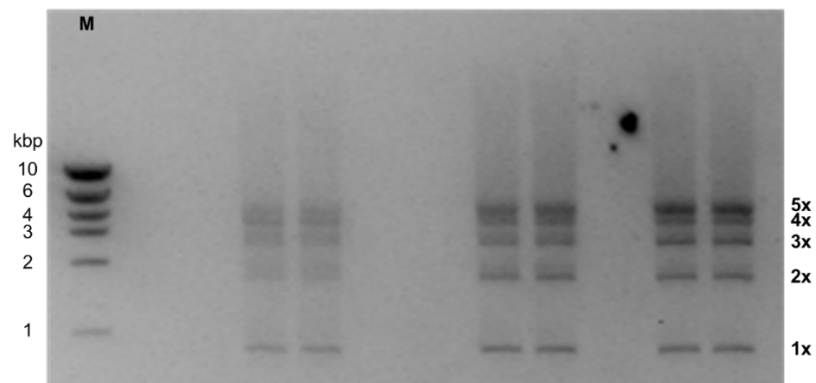
Variant	Scar		Oligonucleotide sequence (5'-3')
1	n/a A	Fwd Rev	See Barcodes GCGGCG <u>GGTCTCT</u> GCTG <i>GGCCAGATCCAGACATTCC</i>
2	A B	Fwd Rev	GCGGCG <u>GGTCTC</u> CAGC <i>CGACTCACTATAGGGACA</i> GCGGCG <u>GGTCTC</u> GACAT <i>GGCCAGATCCAGACATTCC</i>
3	B C	Fwd Rev	GCGGCG <u>GGTCTC</u> CATGT <i>CGACTCACTATAGGGACA</i> GCGGCG <u>GGTCTC</u> GACA <i>GGCCAGATCCAGACATTCC</i>
4	C D	Fwd Rev	GCGGCG <u>GGTCTC</u> TGTG <i>CGACTCACTATAGGGACA</i> GCGGCG <u>GGTCTC</u> ATTGC <i>GGCCAGATCCAGACATTCC</i>
5	D n/a	Fwd Rev	GCGGCG <u>GGTCTC</u> TGCAA <i>CGACTCACTATAGGGACA</i> See Barcodes

Supplementary Table S1. Oligonucleotides used for the five PCR reactions per sample to prepare the library variants for the concatenation procedure. Bsa I restriction sites are in bold and underlined, the unique 4 bp scar regions are highlighted in colour, and gene-overlapping regions are shown in italics.

Variant	Barcode Number		Oligonucleotides used to attach barcodes (5'-3')
1	001	Fwd	CACATATCAGAGTGCG <i>CGACTCACTATAGGGACA</i>
	004		CACGCACACACGCGCG <i>CGACTCACTATAGGGACA</i>
	009		ACACACGCGAGACAGA <i>CGACTCACTATAGGGACA</i>
	012		ACACTAGATCGCGTGT <i>CGACTCACTATAGGGACA</i>
	015		CGCATGACACGTGTGT <i>CGACTCACTATAGGGACA</i>
5	002	Rev	ACACACAGACTGTGAG <i>GGCCAGATCCAGACATTCC</i>
	008		ACAGTCGAGCGCTGCG <i>GGCCAGATCCAGACATTCC</i>
	010		ACGCGCTATCTCAGAG <i>GGCCAGATCCAGACATTCC</i>
	014		CTCACTACGCGCGCGT <i>GGCCAGATCCAGACATTCC</i>
	016		CATAGAGAGATAGTAT <i>GGCCAGATCCAGACATTCC</i>

Supplementary Table S2. Oligonucleotides used to attach the sample-specific barcodes. The barcode-encoding regions are shown in bold and colour. The regions overlapping with the termini of the concatenated amplicon are shown in italics. The oligonucleotides attach the barcodes either to the 5'-end of library variant 1 (Fwd primer, orange) or to the 3'-end of library variant 5 (Rev primer, blue).

Sample	S4		S7		S2
Total DNA after PCR (μg)	0.62		1.04		2.24
Original template (%)	28		17		8

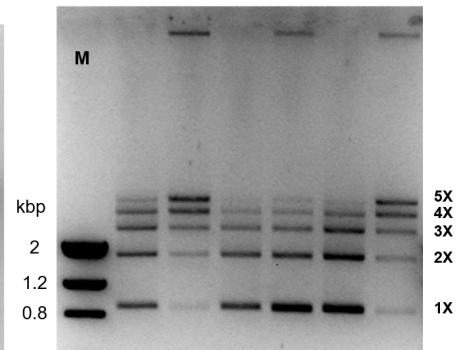
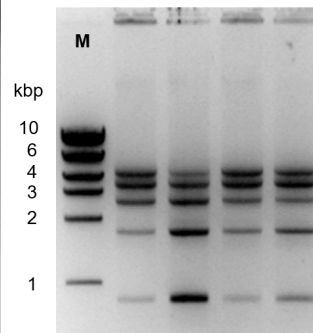
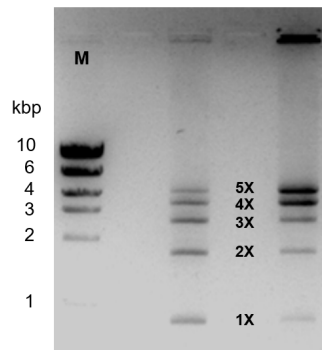


Supplementary Figure S1. Initial attempts at the concatenation showed different efficiencies of assembly for three different samples. Improved efficiency correlated with increased PCR yield of the samples. The total DNA after PCR pertains to the DNA recovered after PCR amplification of the 5x amplicons 1-5 for each sample calculated as an average of the two identical repeats. As PCR yield increased, the assembly efficiency increased because the percentage of original template DNA was reduced. A DNA marker as a size reference was loaded in the leftmost lane (M).

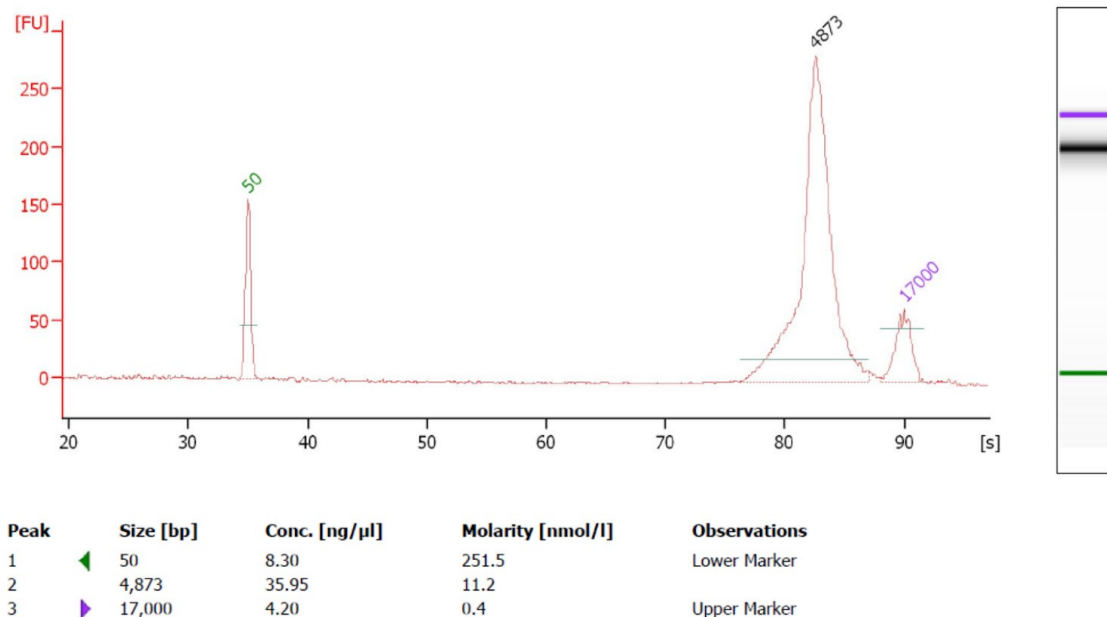
Bsal (NEB)	+		
Bsal HF v2 (NEB)			
Eco31I (Promega)			
T4 DNA Ligase (Promega)	+		
T7 Ligase (Enzymatics)			
Golden Gate kit (NEB)			+
Enzymes 3-fold increased			
Fully assembled amplicon (%)	7		40

	+		
		+	+
	+	+	+
+			
25	4	27	18

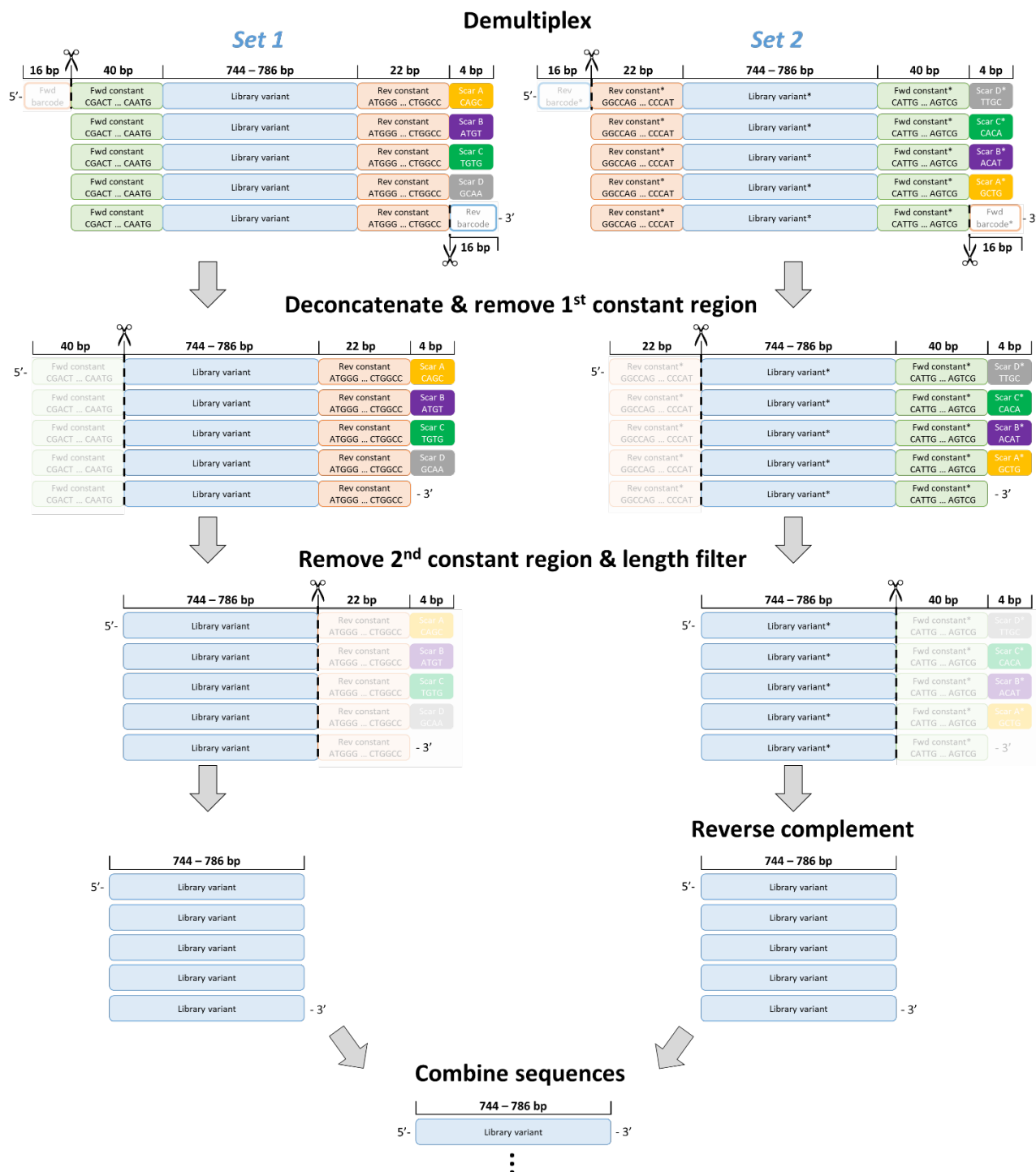
+		+		+	
+		+		+	
	+		+		+
+	+				
4	36	1	2	1	31



Supplementary Figure S2. Optimizing the efficiency of the assembly process. The efficiency of concatenating the five library variants was compared for assemblies using different commercially available enzymes and following protocols 1-3 (black, green and orange, respectively; see methods section). Optimal performance was determined by the percentage of the fully assembled 5x product (between 1% and 40%). The most efficient assembly was achieved with the NEB Golden Gate assembly kit and protocol 1 (increased reaction time and number of cycles). Further details can be found in the methods section. The lane containing a molecular mass ladder is labelled M in each gel.



Supplementary Figure S3. Analysis of size distribution by Agilent Bioanalyzer of the pooled concatenated amplicons S1-S9 directly before submission for sequencing.



Supplementary Figure S4. Detailed schematic representation of the data processing workflow to extract individual library variants (gene sequences) from the PacBio sequencing run. PacBio can read a sequence from either direction and therefore produces two sets of amplicons: one as the sense strand and the other as the reverse complement (Set 1, Set 2). Our sequencing output represents a pool of different samples that each contain a unique pair of 16 bp sequencing barcodes flanking the fully assembled 5x amplicon. Therefore, amplicons were

first demultiplexed by identifying their pair of barcodes and assigning them into sample specific sub-pools accordingly. Each sub-pool was then deconcatenated to extract the five individual library variants contained within each 5x amplicon. As each sequence can be sequenced in the forward or reverse direction, sequences were scanned for either of the two constant regions that flank each variant and the first identified constant region was removed. Each sequence was trimmed by identifying and removing the opposing flanking region (including the second constant region), and filtered by size using a 5% error margin of the expected size range (707-825 bp). Finally, the library variants from reads in the reverse direction were reverse-complemented and merged with the forward reads. The library variants from each sample were subsequently clustered into families of similar sequences, translated to amino acid sequences, and dereplicated within each family (not shown in figure).