Indexing and Barcoding for Illumina NextGen Sequencing

(ELK 10/2011)

There are 2 main strategies for indexing/barcoding (which are the same thing for our purposes here)

IN-LINE

The In-Line barcodes are adjacent to the sample DNA and read from the same sequencing primer as part of the sequence read. They are obligatory (must be read) and they need to handled post-image-processing by bioinformatic analysis. They also take up some of the sequencing read length. On the plus side, they are easy to add, cheaper to analyze/run, and usually faster to process.

MULTIPLEX

The Multiplex barcodes are in the stem or arm of the library. Reading them is optional, but requires additional priming and sequencing so it is more expensive and takes longer. On the plus side, the image pipeline generally bins/sorts them for you and you don't give up any space in your sequencing read.

Multiplex barcodes are recommended for short reads where every base is needed and for working with un-annotated or newly acquired genomes. In-Line barcodes are recommended for long reads of well-annotated genomes such as Chip-Seq, SNP, or transcriptome analysis. You can combine both approaches if you need more degrees of variation. In-Line barcodes should be non-redundant and base-balanced as much as possible; the Multiplex barcodes are less affected by base-bias.

```
Below are the 12 barcodes used in the Illumina TruSeq system, they are base-
balanced and work well as In-Line barcodes as well as Multiplex.
```

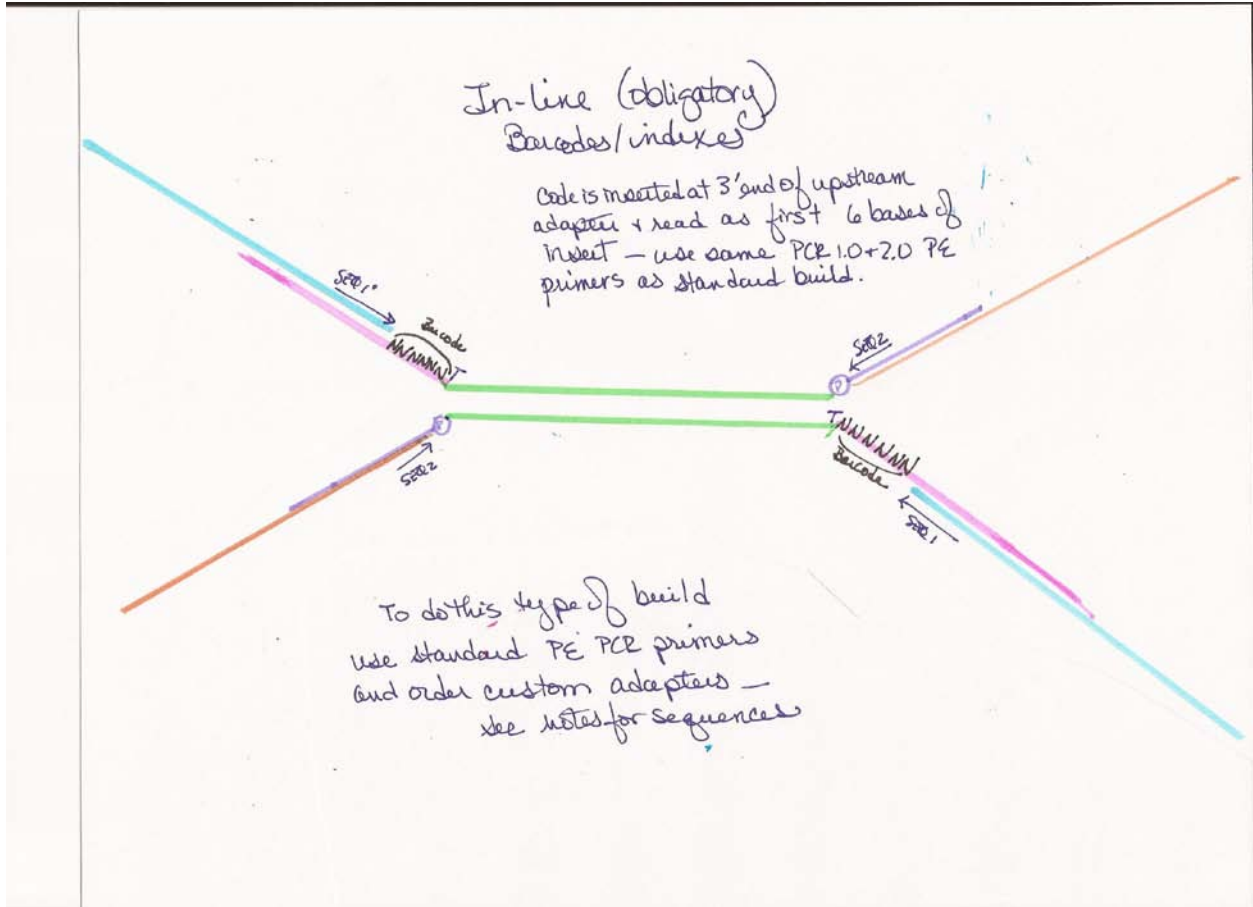|					
ATCACG	CGATGT	TTAGGC	TGACCA	ACATGT	GCCAAT
CAGATC	ACTTGA	GATCAG	TAGCTT	GGCTAG	CTTGTA

At this time, Illumina TruSeq adapters and primers are only available in the TruSeq kits from Illumina. We can order those for you through the core lab, but there is little or no discount. The oligos-only kits are not available for the TruSeq system yet, however we do have the Paired-End Oligo Set available and that works well to build In-Line barcodes as shown below. When you're designing the In-Line adapters remember to add a T at the end, as the first ligation step attachs the adapter to the product of an A-Tailing reaction.

This cartoon is intended to show how the regular Paired-End library is constructed and read. (fancy graphics will come later).
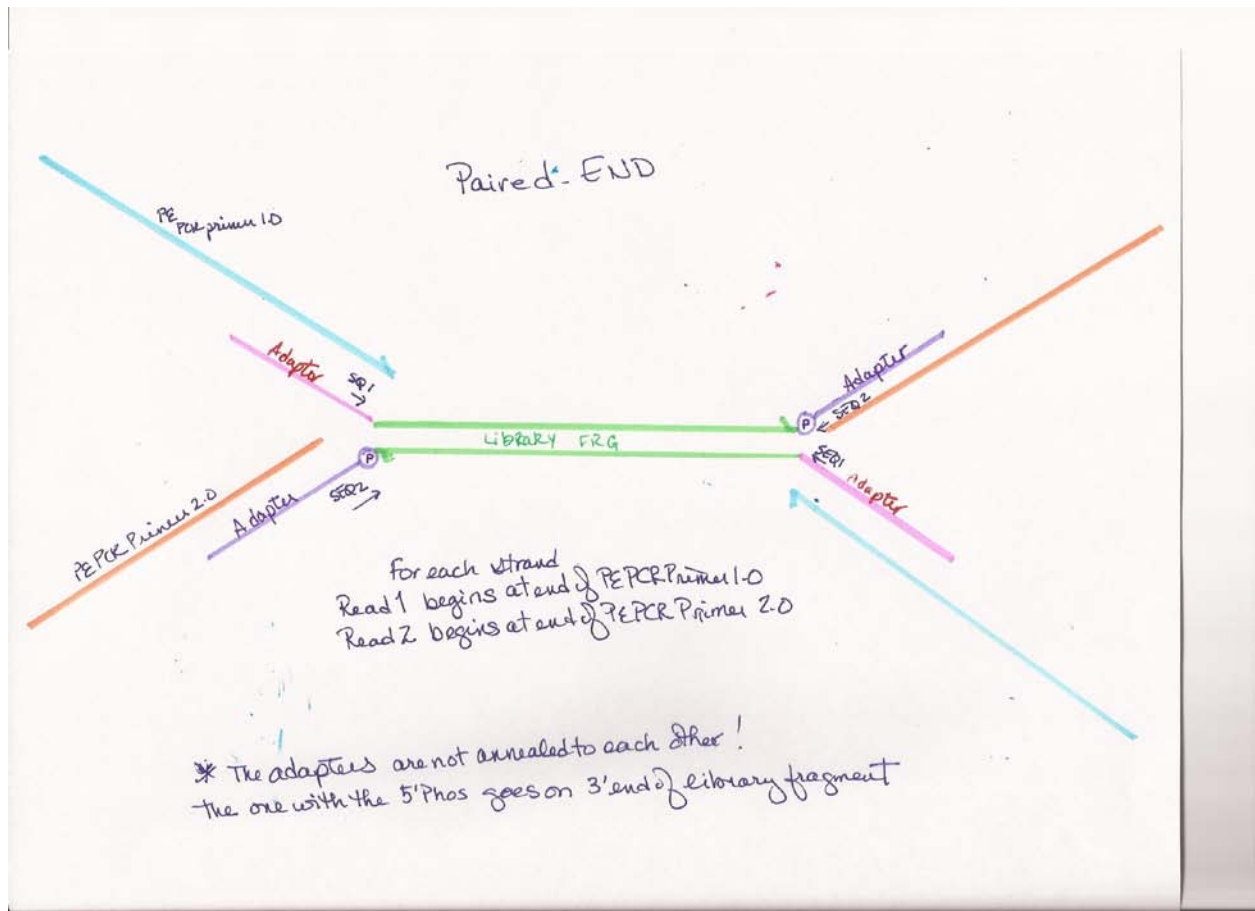
The adapters are ligated onto the prepared insert fragment. The 5' adapter ends in T and the 3' adapter begins with a phosphorylated base (they are not complimentary sequences). The first read sequence priming site is in the upstream adapter. If a side 2 read is ordered, then the clusters are reversed and the second read is done from the priming site in the downstream adapter.

When adding In-Line adapters, place the sequence at the end of the 5' adapter after the sequence priming site but before the final T as shown here. Barcode will be read at the beginning of sequence on Side 1.

In-line (obligatory) Barcodes/indexes

Code is inserted at 3' end of upstream adapter & read as first 6 bases of insert — use same PCR 1.0+2.0 PE primers as standard build.

To do this type of build use standard PE PCR primers and order custom adapters — see notes for sequences

Figure 4: Adapter Ligation Results in Sequence-Ready Constructs without PCR

Ai. Fragment genomic DNA

Aii. Double-stranded cDNA (from figure 2B)

B. End repair and phosphorylate

C. A-tailing

D. Ligate index adapter

E. Denature and amplify for final product

Library construction begins with either fragmented genomic DNA or double-stranded cDNA produced from total RNA (Figure 4A). Blunt-end fragments are created (Figure 4B) and an A-base is then added (Figure 4C) to prepare for indexed adapter ligation (Figure 4D). Final product is created (Figure 4E), which is ready for amplification on either the cBot or the Cluster Station.

The Multiplex barcodes are in the "stem" (actually the middle of the downstream adapter) and are read by a separate sequencing primer and reactions. This is illustrated in this adiagram from Illumina's "TruSeq RNA and DNA Sample Preparation Kits v2".

There are many more barcodes available and there are kits for sale from vendors such as NEB, Nextera, and NuGen to name a few. This short note is intended to give you an idea of how these work and help you decide which type you might want to use. Illumina will soon have many more in their TruSeq set. A

The SOLiD_4 platform available here in the core lab also supports Multiplex barcodes. Currently there are 96 available from ABI and soon they will offer 384. The kits with the oligos for those constructions are available through the core lab with no shipping or handling fees from ABI. The ABI barcodes are sorted by the image analysis pipeline, but they do require an additional read and more time as do the Illumina Multiplex barcodes. In-Line barcoding is not recommended for sequencing in Color-Space.