Barcode Deconvolution:

Results from a barcoding experiment are returned as a set of FASTQ files from next-generation sequencing, one FASTQ file per sample. A single PERL script is used to parse the set of FASTQ files, enumerate the barcodes for each sample, and merge the barcode counts into a matrix of (M barcodes x N samples). The output is written to a file called *mergedBarcodeMatrix.txt*.

The algorithm is as follows. The list of FASTQ files is concatenated into a comma-separated list, which is piped into the PERL script. For each gzipped FASTQ file, the file is opened, and read in in groups of four lines (Definition line, sequence, spacer, quality string). The barcode sequence is assumed to start at the first base of the sequence.

1. The library code substring is extracted from bases 15..18
2. Vector sequence that follows the barcode is extracted as bases 19..26
3. The library code is checked to make sure it conforms to one of three expected library codes (CCAA, ACGT or TGGA), allowing a single mismatch (ie: a Hamming distance ≤ 1)
4. The vector sequence is checked to make sure it conforms to the expected sequence (ie: ATCGATAC) allowing a single mismatch (ie: a Hamming distance ≤ 1)
5. If both the library code and vector sequence are found, the barcode is extracted as the first 18 bases of the sequence string
6. The barcode is stored in a hashtable of hashtables, indexed by the sample name and barcode sequence, with the hashtable value incremented by one each time a barcode is found

After all FASTQ files are processed, a subroutine writes a 2-dimensional matrix of barcodes-by-samples to a tab-delimited text file.

In a second step, one a sample-by-sample basis, highly similar barcodes arising through sequencing errors are merged into a single barcode record in a high-performance computing environment. Briefly, the list of barcodes for each sample is ordered from the most abundant to the least. Next, the ‘stringdist’ R package is used to compute the Hamming distance between the most abundant barcode and all other barcodes, using the parameters method=”hamming”, useBytes=T, and nthread=16. Barcodes with a Hamming distance of one and an abundance ratio < 0.125, or a Hamming distance of two and a ratio < 0.025, are merged into a single record, summing the read counts and retaining the sequence of the most abundant barcode. The process is then repeated for all the remaining barcodes, and the collapsed matrix is then saved.