SeqEMA, a seq implementation of EMA preprocessing

EMA

SeqEMA Functions

Raw FASTQ

Bucket 1 → EMA → BAM 1
Bucket 2 → EMA → BAM 2
Bucket 3 → EMA → BAM 3
Bucket 4 → EMA → BAM 4
Bucket 5 → BWA → BAM 5

BAM merge

Final BAM

Load Library Barcodes from whitelist into a Dictonary, B

Look for Barcodes from each input FASTQ file and update the count if it matches a barcode in B

Take each read and its respective barcode from the FASTQ file and check all barcodes within a Hamming distance of 1. If the barcode is not in the library, pick the most likely barcode.

Each read is assigned to a bucket. For N reads and B buckets, you get ~N/B reads per bucket. Calculate which barcode goes to each read and put the corrected read in each bucket.

── SeqEMA Test 1   ── SeqEMA Test 2   ── SeqEMA Test 3   ── EMA Test 1   ── EMA Test 2   ── EMA Test 3

Cumulative Time (s)

70
58.333
46.667
35
23.333
11.667
0

Loading whitelist   Counting reads   Read correction   Sorting   Sorting buckets by read