Quality and Index Swapping Analysis

Adrian Bubie

9/11/17

Goals: Determine if index swapping occurred in our Hiseq4000 class sequencing samples.

De-multiplex samples to create 48 FASTQ files that contain acceptable index pairs (read1 and read2) and two files of undetermined files that contain unacceptable index pairs, low quality, or undetermined (read1 and read2).
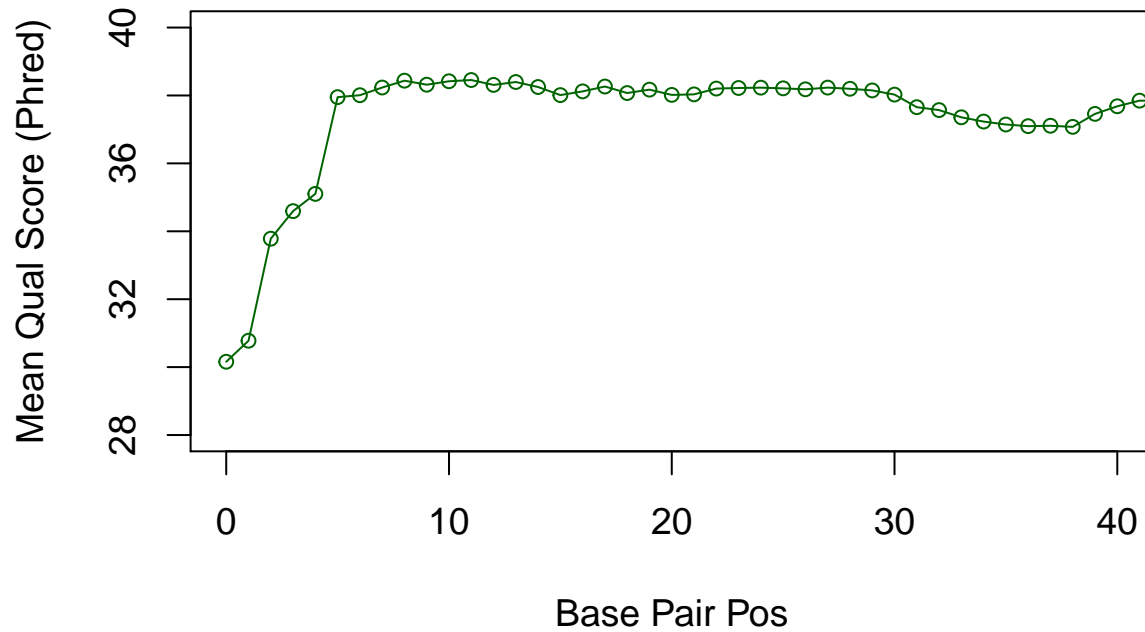
1. (Please see the associated python script, "Qual_Mean_Calc.py" for reference on how the files used to plot here were created).

Quality plots per basepair position and distribution of average quality score per read, for each of our 4 input files (read1, read2, index1, index2):
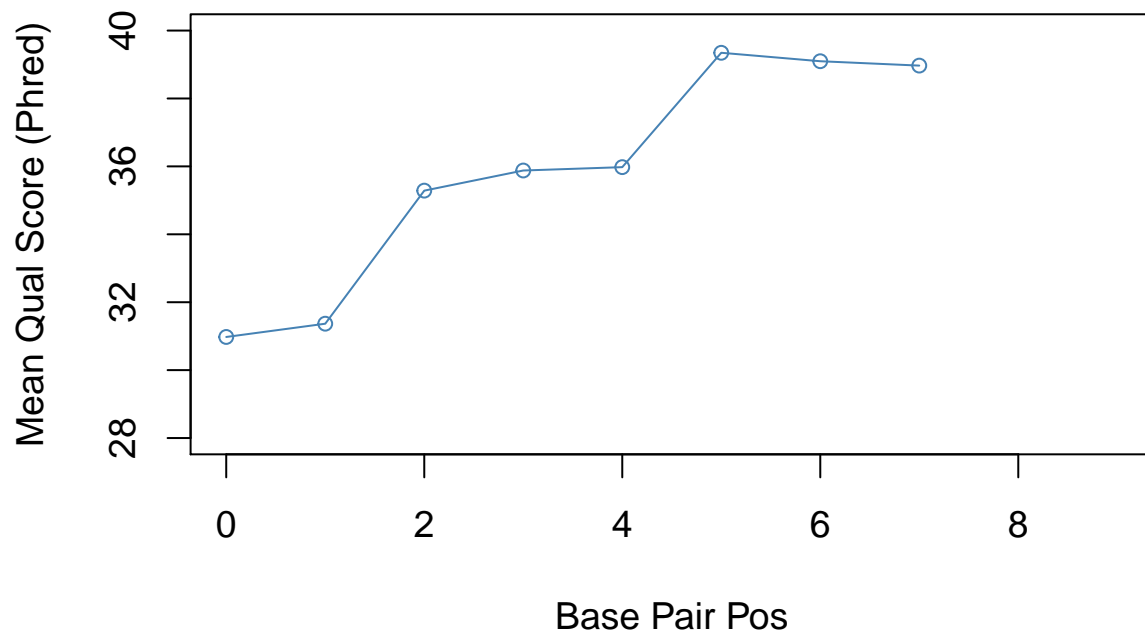
Mean Quality Score by Base Pair Pos. (Read 1)
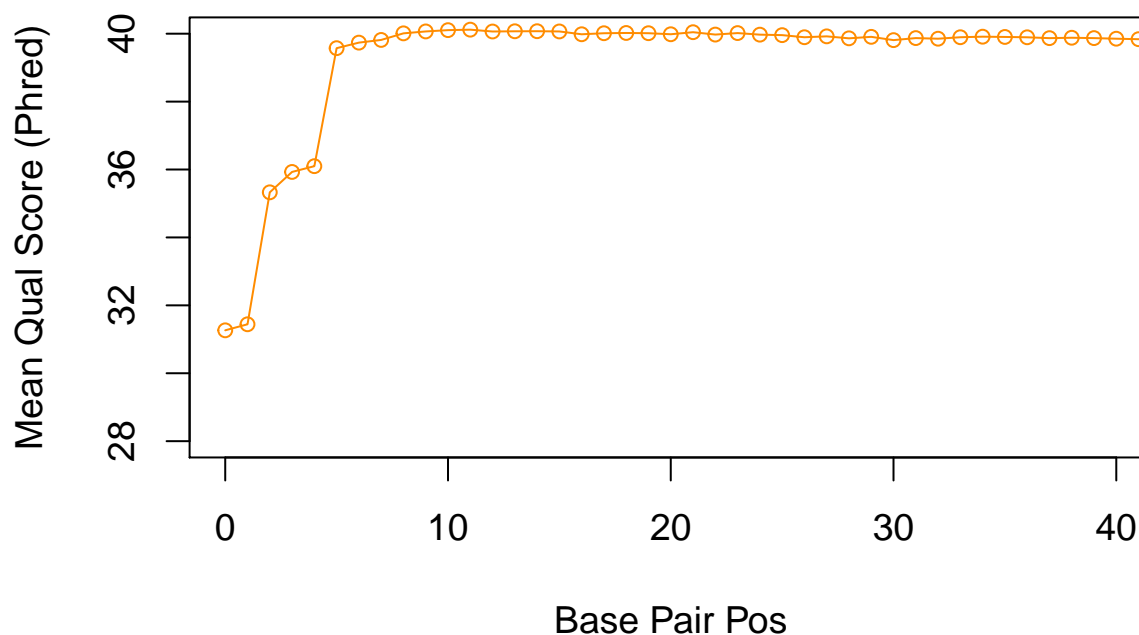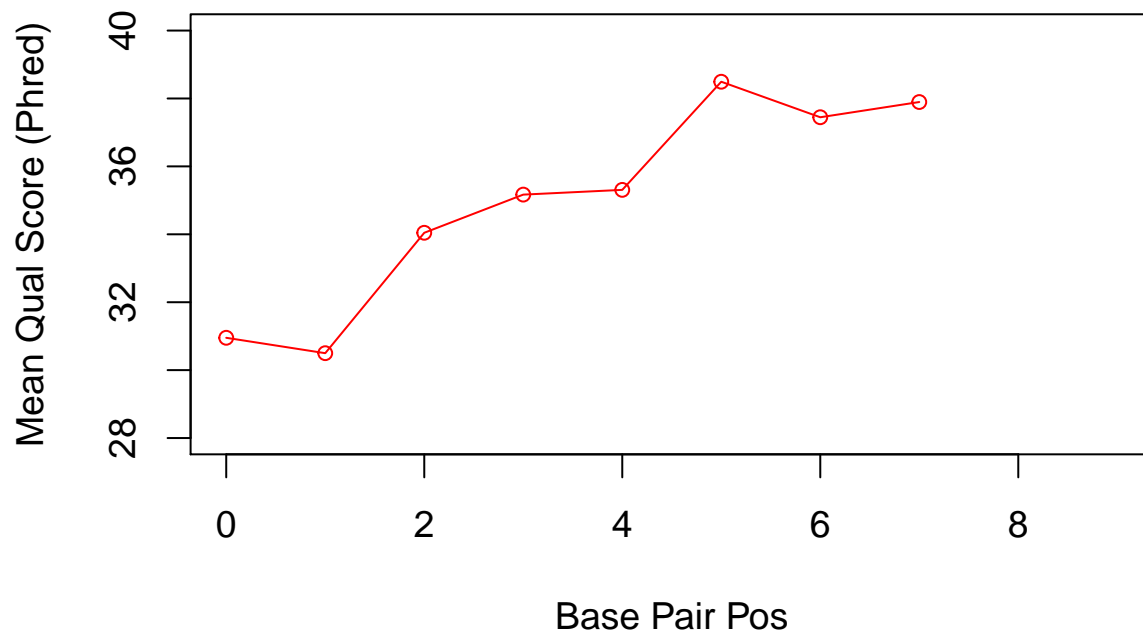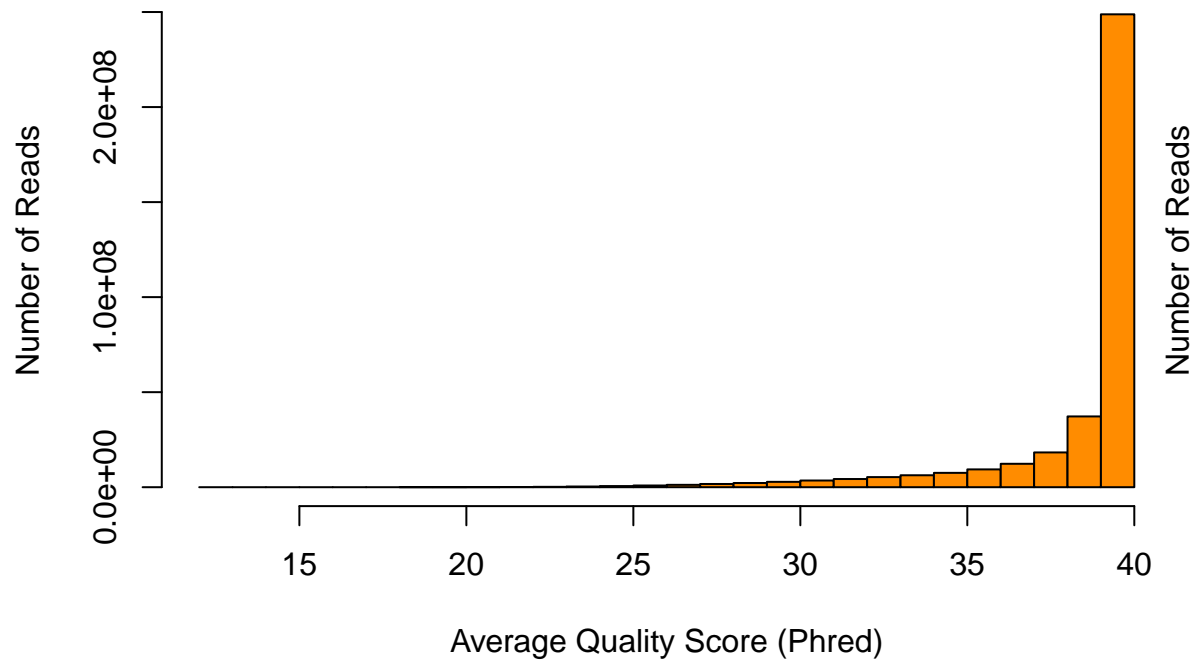
Mean Quality Score by Base Pair Pos. (Read 2)

Base Pair Pos

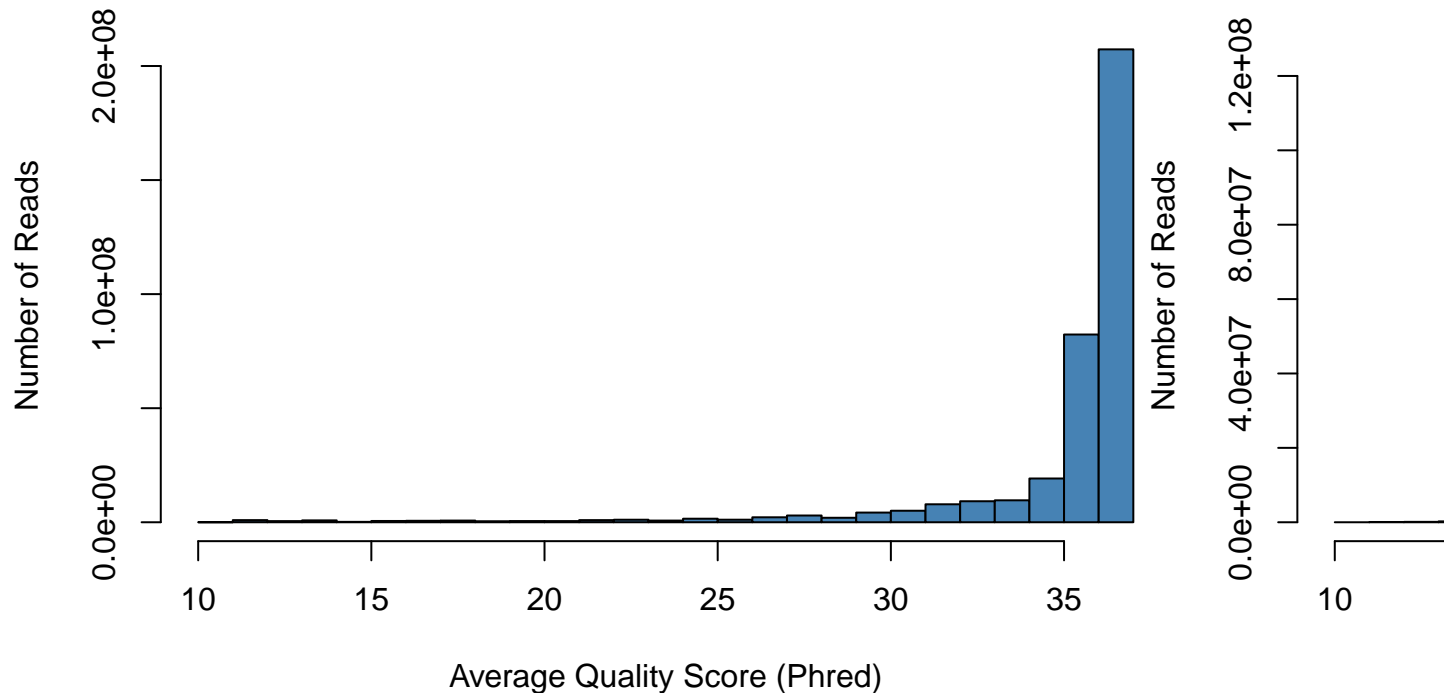# Mean Quality Score by Base Pair Pos. (Index 1)

Base Pair Pos

Mean Quality Score by Base Pair Pos. (Index 2)

Distribution of Reads by Average QS for R1

Distribution of Reads by Average QS for I1

Average Quality Score (Phred)

From these distributions, we can most likely set a quality cutoff of 35 for the index pairs and quality cutoff of 37 for our paired reads, and retain about 3/4 (about 250 million) of our reads, while improving the quality of our reads overall.

We also know that the illumina system will sometimes record bases as 'N' for 'any' when it is not certain what base to call for a particular read. We can count the number of indexes that contain an 'N' in the sequence using the following UNIX command:

```
>$ for FILE in /projects/bgmp/2017_sequencing/*.fastq; do counts=$(awk 'NR%4==2' $FILE | grep "N" | wc -
/projects/bgmp/2017_sequencing/1294_S1_L008_R1_001.fastq 2602560
/projects/bgmp/2017_sequencing/1294_S1_L008_R2_001.fastq 3976613
/projects/bgmp/2017_sequencing/1294_S1_L008_R3_001.fastq 3328051
/projects/bgmp/2017_sequencing/1294_S1_L008_R4_001.fastq 3591851
```

Our Indexes are contained in the 'R2' and 'R3' files; so we have 3976613 and 3328051 reads with 'N's for our index 1 and index 2, respectively (a total of 7304664 reads across both indexes)

Overall, the read distributions show that most of our data clustered pretty high on the quality scale. Illumina defines Phred 30 as 99.9% accuracy for the basecall, and <90% of our reads have an average score of 30 or above. However, there is significant disparity between our Index read scores and our insert read qualities. Index scores appear to be overall of lower quality on average than our insert reads' qscore; additionally, we see in the graph of quality by basepair that the bases at the beginning of each index/read almost always have a average quality score lower than those of the bases further into the read. This indicates that the bases read at the beginning of the sequence (near the oligo linkage on the flowcell) are more difficult for the sequencer to resolve properly. Notice also that vast majority of the Indexes that contain an 'N' base (for unresolved or 'any' base) have that 'N' at the starting position of the index. This provides further evidence that the sequencer's resolution for bases starts poor but gets much better as synthesis continues into the insert.

2. (Please see the associated python script, "Qual_Index_Swp.py" for reference and indication of how reads were binned by index pairs). Based on analysis of the index reads above, the reads were filtered by an index average quality cut-off Phred score of 35. This means that reads that had an *both* indexes with an average quality score of 35 or above were retained and all other reads were removed.

a. There are a total of 363246735 reads in the original Sequencing output. After filtering by index quality, a total of 245598505 reads were retained, representing a retention rate of just over 67.6%. The read counts can be broken down by index pairs (libraries), as seen below:

```
##        Index1    Index2 Read_Count
## 1   TACCGGAT TACCGGAT   51581889
## 2   TCTTCGAC TCTTCGAC   31203016
## 3   CTCTGGAT CTCTGGAT   25933763
## 4   CTAGCTCA CTAGCTCA   13661286
## 5   TGTTCCGT TGTTCCGT   12247310
## 6   AGAGTCCA AGAGTCCA    8234276
## 7   TAGCCATG TAGCCATG    7951357
## 8   TATGGCAC TATGGCAC    7880872
## 9   TCGAGAGT TCGAGAGT    7791551
## 10  ATCATGCG ATCATGCG    7448228
## 11  GTCCTAAG GTCCTAAG    6658709
## 12  AACAGCGA AACAGCGA    6606498
## 13  AGGATAGC AGGATAGC    6571536
## 14  ACGATCAG ACGATCAG    6234505
## 15  GTAGCGTA GTAGCGTA    6033486
## 16  ATCGTGGT ATCGTGGT    5031584
## 17  GATCAAGG GATCAAGG    4900440
## 18  GCTACTCT GCTACTCT    4720342
## 19  CGATCGAT CGATCGAT    4398491
## 20  TCGGATTC TCGGATTC    3149983
## 21  CGGTAATC CGGTAATC    3097661
## 22  GATCTTGC GATCTTGC    2861868
## 23  TCGACAAG TCGACAAG    2786134
## 24  CACTTCAC CACTTCAC    2765861
## 25  TATGGCAC TGTTCCGT      63153
## 26  TGTTCCGT TATGGCAC      61181
## 27  CTAGCTCA TCGACAAG       9926
## 28  GATCAAGG TCTTCGAC       9552
## 29  TCGACAAG ATCATGCG       5896
## 30  CTCTGGAT TACCGGAT       5821
## 31  TACCGGAT CTCTGGAT       5796
## 32  TACCGGAT TCTTCGAC       5626
## 33  GTCCTAAG TATGGCAC       4869
## 34  TCTTCGAC TACCGGAT       4550
## 35  CGGTAATC TACCGGAT       3861
## 36  TCTTCGAC ATCGTGGT       3083
## 37  CTCTGGAT TCTTCGAC       2776
## 38  TCTTCGAC CTCTGGAT       2747
## 39  TACCGGAT TGTTCCGT       2339
## 40  CACTTCAC TAGCCATG       2198
## 41  TATGGCAC TACCGGAT       2105
## 42  TGTTCCGT TACCGGAT       2021
## 43  CTAGCTCA CTCTGGAT       2004
## 44  TACCGGAT CTAGCTCA       1975
## 45  TATGGCAC TCTTCGAC       1885
## 46  CTAGCTCA TACCGGAT       1846
## 47  TACCGGAT TAGCCATG       1826
## 48  TACCGGAT TATGGCAC       1813
## 49  TACCGGAT TCGAGAGT       1632
```

```
## 50  TACCGGAT CGGTAATC        1622
## 51  CTCTGGAT GCTACTCT        1471
## 52  TCTTCGAC TGTTCCGT        1413
## 53  TCGAGAGT TACCGGAT        1410
## 54  CTCTGGAT CTAGCTCA        1381
## 55  TGTTCCGT TCTTCGAC        1260
## 56  TCTTCGAC TATGGCAC        1250
## 57  TCTTCGAC TCGAGAGT        1170
## 58  TCTTCGAC CTAGCTCA        1144
## 59  CTCTGGAT TCGAGAGT        1134
## 60  TACCGGAT AGAGTCCA        1123
## 61  CTCTGGAT TGTTCCGT        1116
## 62  CTAGCTCA ATCATGCG        1103
## 63  TCTTCGAC GATCAAGG        1079
## 64  TACCGGAT AGGATAGC        1063
## 65  TACCGGAT ACGATCAG        1041
## 66  CTAGCTCA TCTTCGAC        1014
## 67  AACAGCGA TACCGGAT        1011
## 68  CTAGCTCA GCTACTCT        1010
## 69  TAGCCATG TACCGGAT        1002
## 70  AACAGCGA TCTTCGAC         998
## 71  TACCGGAT GTCCTAAG         992
## 72  TACCGGAT ATCATGCG         982
## 73  TACCGGAT ATCGTGGT         973
## 74  CGATCGAT TACCGGAT         972
## 75  CTCTGGAT AGAGTCCA         969
## 76  GTAGCGTA GCTACTCT         953
## 77  AGAGTCCA TACCGGAT         949
## 78  ATCATGCG TACCGGAT         932
## 79  TACCGGAT CGATCGAT         920
## 80  CTCTGGAT CGATCGAT         913
## 81  GTCCTAAG TACCGGAT         898
## 82  ACGATCAG TACCGGAT         888
## 83  TACCGGAT AACAGCGA         878
## 84  TCTTCGAC AGGATAGC         873
## 85  TGTTCCGT CTCTGGAT         871
## 86  TACCGGAT GTAGCGTA         857
## 87  TACCGGAT GATCAAGG         826
## 88  ATCGTGGT TACCGGAT         814
## 89  TCGAGAGT CTCTGGAT         796
## 90  GTAGCGTA TACCGGAT         791
## 91  CTCTGGAT TATGGCAC         760
## 92  TACCGGAT GCTACTCT         741
## 93  ATCATGCG TCTTCGAC         736
## 94  TCGAGAGT TCTTCGAC         724
## 95  TCTTCGAC ACGATCAG         722
## 96  TATGGCAC CTCTGGAT         720
## 97  GCTACTCT TACCGGAT         712
## 98  TCTTCGAC TAGCCATG         701
## 99  TACCGGAT TCGGATTC         677
## 100 ATCGTGGT TCTTCGAC         662
## 101 AGAGTCCA GTAGCGTA         660
## 102 TCTTCGAC TCGGATTC         651
## 103 AGGATAGC TACCGGAT         645
```

```
## 104 TCTTCGAC AGAGTCCA       623
## 105 TCTTCGAC ATCATGCG       617
## 106 AGGATAGC TCTTCGAC       615
## 107 CTAGCTCA GTAGCGTA       608
## 108 TACCGGAT CACTTCAC       590
## 109 TATGGCAC TCGGATTC       585
## 110 TCGGATTC AACAGCGA       584
## 111 ATCATGCG CTCTGGAT       582
## 112 TCTTCGAC CGATCGAT       573
## 113 TCTTCGAC GTCCTAAG       573
## 114 CGATCGAT CTCTGGAT       567
## 115 GATCAAGG TACCGGAT       567
## 116 CTCTGGAT ACGATCAG       566
## 117 CTAGCTCA TCGAGAGT       564
## 118 CTCTGGAT TAGCCATG       563
## 119 CTCTGGAT GTCCTAAG       553
## 120 TCTTCGAC GTAGCGTA       551
## 121 AGAGTCCA TCTTCGAC       548
## 122 CTCTGGAT ATCATGCG       547
## 123 CGATCGAT TCTTCGAC       537
## 124 ATCGTGGT CTCTGGAT       534
## 125 TGTTCCGT TCGGATTC       533
## 126 CTCTGGAT GTAGCGTA       529
## 127 AGAGTCCA TCGAGAGT       527
## 128 TCTTCGAC CGGTAATC       525
## 129 TCTTCGAC GCTACTCT       525
## 130 TACCGGAT GATCTTGC       519
## 131 GTCCTAAG TCTTCGAC       516
## 132 GCTACTCT CTCTGGAT       514
## 133 AACAGCGA CTCTGGAT       511
## 134 TAGCCATG TCTTCGAC       507
## 135 TCTTCGAC GATCTTGC       499
## 136 AGGATAGC GTAGCGTA       495
## 137 GTCCTAAG TGTTCCGT       493
## 138 TACCGGAT TCGACAAG       493
## 139 CTCTGGAT AGGATAGC       490
## 140 ACGATCAG TCTTCGAC       483
## 141 CTAGCTCA AGAGTCCA       481
## 142 GTAGCGTA TCTTCGAC       481
## 143 TATGGCAC TCGAGAGT       471
## 144 AGAGTCCA CTAGCTCA       456
## 145 GTAGCGTA ACGATCAG       456
## 146 CTCTGGAT AACAGCGA       453
## 147 GTCCTAAG CTCTGGAT       451
## 148 CTCTGGAT ATCGTGGT       445
## 149 TGTTCCGT AACAGCGA       439
## 150 CACTTCAC TACCGGAT       435
## 151 CGGTAATC TCTTCGAC       435
## 152 ATCATGCG TATGGCAC       428
## 153 TAGCCATG CTCTGGAT       428
## 154 GCTACTCT TCTTCGAC       418
## 155 TGTTCCGT TCGAGAGT       418
## 156 ACGATCAG CTCTGGAT       415
## 157 CACTTCAC TCTTCGAC       410
```

```
## 158 CTCTGGAT CGGTAATC         410
## 159 AGAGTCCA CTCTGGAT         409
## 160 TCGGATTC TACCGGAT         404
## 161 CTAGCTCA TGTTCCGT         401
## 162 TCTTCGAC AACAGCGA         400
## 163 TCTTCGAC TCGACAAG         398
## 164 TATGGCAC GTCCTAAG         395
## 165 TCGAGAGT TGTTCCGT         394
## 166 CTCTGGAT TCGGATTC         391
## 167 GTAGCGTA TCGAGAGT         391
## 168 TCTTCGAC CACTTCAC         389
## 169 TGTTCCGT CTAGCTCA         388
## 170 TCGGATTC TCTTCGAC         386
## 171 ATCATGCG TAGCCATG         384
## 172 TGTTCCGT AGAGTCCA         383
## 173 ATCGTGGT GATCAAGG         380
## 174 GCTACTCT GTAGCGTA         377
## 175 AACAGCGA GATCTTGC         373
## 176 GTAGCGTA CTCTGGAT         373
## 177 ATCATGCG ACGATCAG         370
## 178 GTAGCGTA CTAGCTCA         363
## 179 CTAGCTCA TAGCCATG         356
## 180 CTCTGGAT GATCAAGG         350
## 181 TATGGCAC CTAGCTCA         346
## 182 CTAGCTCA TATGGCAC         345
## 183 CTAGCTCA ACGATCAG         338
## 184 TATGGCAC AACAGCGA         335
## 185 TCGAGAGT CTAGCTCA         330
## 186 GATCTTGC TACCGGAT         329
## 187 GATCTTGC TCTTCGAC         320
## 188 TCGACAAG TACCGGAT         320
## 189 TGTTCCGT ATCATGCG         317
## 190 AGGATAGC CTCTGGAT         310
## 191 CACTTCAC CTCTGGAT         300
## 192 ATCATGCG TGTTCCGT         298
## 193 CTCTGGAT CACTTCAC         297
## 194 ATCATGCG CTAGCTCA         294
## 195 TCGAGAGT ACGATCAG         294
## 196 AGAGTCCA TGTTCCGT         293
## 197 ATCATGCG GTCCTAAG         292
## 198 CTCTGGAT GATCTTGC         285
## 199 GCTACTCT CTAGCTCA         283
## 200 CTAGCTCA GTCCTAAG         281
## 201 TCGAGAGT GCTACTCT         276
## 202 TCGGATTC CTCTGGAT         275
## 203 TATGGCAC AGGATAGC         273
## 204 GATCAAGG CTCTGGAT         270
## 205 AGAGTCCA ACGATCAG         268
## 206 AACAGCGA CTAGCTCA         260
## 207 AGGATAGC GATCTTGC         259
## 208 ATCATGCG AGGATAGC         259
## 209 TGTTCCGT CGATCGAT         259
## 210 TGTTCCGT TAGCCATG         257
## 211 AACAGCGA ATCATGCG         256
```

```
## 212 CTCTGGAT TCGACAAG          256
## 213 GCTACTCT TGTTCCGT          256
## 214 CTAGCTCA AACAGCGA          254
## 215 CGGTAATC CTCTGGAT          246
## 216 TGTTCCGT GTAGCGTA          246
## 217 CGATCGAT TGTTCCGT          245
## 218 CTAGCTCA AGGATAGC          244
## 219 ACGATCAG TCGAGAGT          239
## 220 GTAGCGTA AGAGTCCA          238
## 221 TAGCCATG CTAGCTCA          234
## 222 TATGGCAC TAGCCATG          234
## 223 TCGACAAG TCTTCGAC          234
## 224 ACGATCAG TGTTCCGT          233
## 225 ATCATGCG AGAGTCCA          233
## 226 CTAGCTCA ATCGTGGT          232
## 227 AACAGCGA ACGATCAG          231
## 228 CTAGCTCA CGATCGAT          230
## 229 TATGGCAC ACGATCAG          230
## 230 TGTTCCGT ACGATCAG          230
## 231 AACAGCGA AGGATAGC          229
## 232 AACAGCGA TGTTCCGT          229
## 233 AGGATAGC ACGATCAG          228
## 234 GTAGCGTA TGTTCCGT          227
## 235 ACGATCAG CTAGCTCA          225
## 236 TAGCCATG TGTTCCGT          225
## 237 TCGAGAGT AGGATAGC          225
## 238 AACAGCGA AGAGTCCA          223
## 239 GTCCTAAG AACAGCGA          223
## 240 ACGATCAG ATCATGCG          220
## 241 GATCAAGG ACGATCAG          220
## 242 ACGATCAG TAGCCATG          219
## 243 ATCATGCG AACAGCGA          219
## 244 ATCATGCG GATCAAGG          219
## 245 TGTTCCGT AGGATAGC          217
## 246 AGAGTCCA AACAGCGA          216
## 247 AGGATAGC ATCATGCG          213
## 248 GATCAAGG ATCGTGGT          213
## 249 GCTACTCT AGAGTCCA          212
## 250 CTAGCTCA GATCAAGG          210
## 251 ACGATCAG GTCCTAAG          209
## 252 TATGGCAC ATCATGCG          209
## 253 AACAGCGA TATGGCAC          207
## 254 ACGATCAG AGAGTCCA          207
## 255 AACAGCGA GTAGCGTA          204
## 256 TATGGCAC CACTTCAC          203
## 257 GTCCTAAG CTAGCTCA          202
## 258 TGTTCCGT ATCGTGGT          202
## 259 AACAGCGA TCGGATTC          201
## 260 ATCGTGGT TCGAGAGT          201
## 261 TAGCCATG GTCCTAAG          200
## 262 AGGATAGC TATGGCAC          198
## 263 TCGAGAGT ATCGTGGT          198
## 264 GATCTTGC AGGATAGC          197
## 265 ACGATCAG CGATCGAT          196
```

9

```
## 266 AGGATAGC AGAGTCCA         196
## 267 AGGATAGC TGTTCCGT         195
## 268 ATCATGCG ATCGTGGT         195
## 269 GCTACTCT TCGAGAGT         194
## 270 TAGCCATG TATGGCAC         193
## 271 AGAGTCCA TATGGCAC         192
## 272 CTAGCTCA TCGGATTC         192
## 273 TCGACAAG CTCTGGAT         191
## 274 AGGATAGC CTAGCTCA         190
## 275 CGATCGAT CTAGCTCA         189
## 276 ATCATGCG TCGAGAGT         188
## 277 TGTTCCGT GCTACTCT         188
## 278 AGAGTCCA AGGATAGC         187
## 279 TATGGCAC AGAGTCCA         186
## 280 TCGAGAGT TATGGCAC         184
## 281 ACGATCAG AGGATAGC         183
## 282 GATCTTGC CTCTGGAT         183
## 283 AACAGCGA ATCGTGGT         182
## 284 CACTTCAC ACGATCAG         182
## 285 TGTTCCGT GATCTTGC         182
## 286 ATCGTGGT CTAGCTCA         179
## 287 GTCCTAAG TAGCCATG         179
## 288 TAGCCATG ACGATCAG         177
## 289 TCGAGAGT TCGGATTC         177
## 290 ACGATCAG TATGGCAC         176
## 291 AGAGTCCA ATCATGCG         176
## 292 GTCCTAAG GATCAAGG         174
## 293 TCGAGAGT AGAGTCCA         173
## 294 TGTTCCGT GTCCTAAG         172
## 295 AGGATAGC GCTACTCT         171
## 296 TATGGCAC GATCTTGC         170
## 297 GTAGCGTA GTCCTAAG         169
## 298 GTCCTAAG ATCATGCG         169
## 299 AACAGCGA TCGAGAGT         168
## 300 AGGATAGC TCGAGAGT         167
## 301 ACGATCAG TCGACAAG         166
## 302 CGATCGAT AGGATAGC         166
## 303 CGGTAATC ATCGTGGT         166
## 304 TAGCCATG ATCATGCG         165
## 305 CTAGCTCA CGGTAATC         163
## 306 TCGAGAGT GTAGCGTA         163
## 307 GTCCTAAG TCGAGAGT         162
## 308 TAGCCATG AGAGTCCA         161
## 309 TAGCCATG TCGAGAGT         161
## 310 ATCGTGGT TGTTCCGT         160
## 311 CACTTCAC TATGGCAC         160
## 312 TCGGATTC TATGGCAC         159
## 313 AACAGCGA CGATCGAT         158
## 314 GATCAAGG GTCCTAAG         158
## 315 GTCCTAAG TCGGATTC         158
## 316 GCTACTCT ACGATCAG         157
## 317 GTCCTAAG ACGATCAG         156
## 318 TCGAGAGT TAGCCATG         156
## 319 TCGGATTC GATCTTGC         156
```

```
## 320 TCGAGAGT ATCATGCG          154
## 321 CTAGCTCA GATCTTGC          152
## 322 TAGCCATG GATCAAGG          152
## 323 TATGGCAC GATCAAGG          152
## 324 TATGGCAC GTAGCGTA          152
## 325 ATCATGCG TCGACAAG          151
## 326 TCGAGAGT CGATCGAT          151
## 327 AGAGTCCA TAGCCATG          149
## 328 GTAGCGTA TATGGCAC          148
## 329 TCGAGAGT GTCCTAAG          148
## 330 AACAGCGA GATCAAGG          147
## 331 GTCCTAAG GTAGCGTA          147
## 332 TGTTCCGT GATCAAGG          147
## 333 AACAGCGA TAGCCATG          146
## 334 CGGTAATC TAGCCATG          146
## 335 ATCGTGGT ACGATCAG          144
## 336 TATGGCAC CGGTAATC          143
## 337 GTAGCGTA AACAGCGA          142
## 338 TATGGCAC ATCGTGGT          142
## 339 TCGACAAG TCGAGAGT          142
## 340 AACAGCGA GTCCTAAG          140
## 341 AGGATAGC AACAGCGA          140
## 342 ATCGTGGT ATCATGCG          140
## 343 GTAGCGTA ATCATGCG          140
## 344 ATCGTGGT TATGGCAC          139
## 345 TCGGATTC CTAGCTCA          139
## 346 ATCATGCG GCTACTCT          138
## 347 GATCAAGG TGTTCCGT          138
## 348 AGAGTCCA ATCGTGGT          137
## 349 CGATCGAT TCGAGAGT          137
## 350 TCGAGAGT AACAGCGA          137
## 351 ACGATCAG GCTACTCT          136
## 352 ATCGTGGT AGAGTCCA          135
## 353 CACTTCAC CTAGCTCA          135
## 354 GATCAAGG TAGCCATG          135
## 355 GTCCTAAG AGAGTCCA          135
## 356 AGGATAGC CACTTCAC          134
## 357 ATCATGCG GTAGCGTA          134
## 358 TAGCCATG AGGATAGC          134
## 359 GTAGCGTA AGGATAGC          133
## 360 AGAGTCCA GCTACTCT          131
## 361 CGATCGAT ACGATCAG          131
## 362 TCGAGAGT TCGACAAG          131
## 363 TCGGATTC CGATCGAT          130
## 364 ACGATCAG GTAGCGTA          129
## 365 ATCGTGGT AGGATAGC          128
## 366 GTAGCGTA ATCGTGGT          127
## 367 ACGATCAG GATCAAGG          126
## 368 GCTACTCT TATGGCAC          125
## 369 GTCCTAAG CGATCGAT          125
## 370 AGGATAGC CGATCGAT          124
## 371 AGAGTCCA GTCCTAAG          123
## 372 GATCTTGC TATGGCAC          121
## 373 TCGACAAG TAGCCATG          121
```

```
## 374 AGAGTCCA CGATCGAT        120
## 375 CACTTCAC TGTTCCGT        120
## 376 GTCCTAAG AGGATAGC        120
## 377 TAGCCATG GTAGCGTA        120
## 378 TATGGCAC GCTACTCT        119
## 379 TGTTCCGT CACTTCAC        118
## 380 ACGATCAG ATCGTGGT        117
## 381 CGATCGAT AGAGTCCA        117
## 382 CGATCGAT TAGCCATG        117
## 383 CTAGCTCA CACTTCAC        117
## 384 GATCAAGG CTAGCTCA        117
## 385 GCTACTCT GTCCTAAG        117
## 386 TCGACAAG ACGATCAG        117
## 387 ACGATCAG AACAGCGA        116
## 388 AGGATAGC TAGCCATG        116
## 389 TAGCCATG CACTTCAC        115
## 390 CGGTAATC CTAGCTCA        114
## 391 GTCCTAAG GCTACTCT        114
## 392 AGGATAGC TCGGATTC        113
## 393 GCTACTCT ATCATGCG        113
## 394 TATGGCAC CGATCGAT        113
## 395 ATCGTGGT GTAGCGTA        112
## 396 GATCTTGC CGATCGAT        112
## 397 TCGACAAG GTCCTAAG        112
## 398 TCGAGAGT CGGTAATC        112
## 399 GTAGCGTA TAGCCATG        111
## 400 TGTTCCGT TCGACAAG        111
## 401 ATCATGCG GATCTTGC        110
## 402 CACTTCAC CGGTAATC        109
## 403 GATCAAGG GTAGCGTA        109
## 404 TCGAGAGT GATCAAGG        109
## 405 GATCAAGG TCGAGAGT        108
## 406 ATCATGCG CGATCGAT        107
## 407 GCTACTCT TAGCCATG        107
## 408 TATGGCAC TCGACAAG        107
## 409 TCGGATTC AGGATAGC        107
## 410 GATCAAGG ATCATGCG        106
## 411 TGTTCCGT CGGTAATC        104
## 412 CGGTAATC AGGATAGC        103
## 413 GCTACTCT AACAGCGA        103
## 414 TCGGATTC TGTTCCGT        103
## 415 AGGATAGC CGGTAATC        101
## 416 ATCGTGGT CGATCGAT        101
## 417 TCGACAAG CTAGCTCA        101
## 418 GCTACTCT AGGATAGC        100
## 419 CACTTCAC AGGATAGC         99
## 420 GATCTTGC CTAGCTCA         99
## 421 GTAGCGTA GATCAAGG         99
## 422 AACAGCGA GCTACTCT         98
## 423 CGATCGAT CGGTAATC         98
## 424 TAGCCATG TCGACAAG         98
## 425 TAGCCATG AACAGCGA         97
## 426 ACGATCAG CGGTAATC         96
## 427 ATCGTGGT TAGCCATG         96
```

```
## 428 CGATCGAT TATGGCAC        96
## 429 CGGTAATC TGTTCCGT        96
## 430 GATCAAGG AGGATAGC        95
## 431 GATCAAGG TATGGCAC        95
## 432 GATCTTGC ATCATGCG        95
## 433 GTCCTAAG ATCGTGGT        95
## 434 GTCCTAAG GATCTTGC        95
## 435 CGATCGAT ATCATGCG        93
## 436 CGATCGAT GCTACTCT        92
## 437 GTCCTAAG TCGACAAG        92
## 438 AGGATAGC ATCGTGGT        91
## 439 CGATCGAT GTAGCGTA        91
## 440 CGATCGAT GTCCTAAG        91
## 441 AGAGTCCA TCGGATTC        89
## 442 ATCGTGGT GCTACTCT        89
## 443 GCTACTCT CGATCGAT        89
## 444 GTAGCGTA CGATCGAT        89
## 445 TCGGATTC TCGAGAGT        89
## 446 GATCAAGG GATCTTGC        87
## 447 AGAGTCCA GATCAAGG        86
## 448 AGGATAGC GTCCTAAG        85
## 449 GATCAAGG AGAGTCCA        85
## 450 GCTACTCT ATCGTGGT        85
## 451 CACTTCAC CGATCGAT        84
## 452 GCTACTCT GATCAAGG        84
## 453 ATCGTGGT AACAGCGA        83
## 454 CGGTAATC TATGGCAC        83
## 455 TAGCCATG ATCGTGGT        83
## 456 GATCTTGC AACAGCGA        82
## 457 GATCTTGC AGAGTCCA        82
## 458 CGGTAATC CGATCGAT        81
## 459 GCTACTCT GATCTTGC        81
## 460 TCGACAAG TGTTCCGT        81
## 461 GATCAAGG CGATCGAT        80
## 462 ATCATGCG CACTTCAC        79
## 463 ATCATGCG CGGTAATC        79
## 464 GATCTTGC GTCCTAAG        79
## 465 GATCTTGC TGTTCCGT        78
## 466 TCGAGAGT GATCTTGC        78
## 467 TCGGATTC ACGATCAG        78
## 468 ACGATCAG CACTTCAC        77
## 469 ATCATGCG TCGGATTC        77
## 470 ATCGTGGT CGGTAATC        77
## 471 AGAGTCCA CGGTAATC        76
## 472 GATCAAGG GCTACTCT        76
## 473 GATCTTGC GTAGCGTA        76
## 474 TAGCCATG TCGGATTC        76
## 475 TCGGATTC AGAGTCCA        76
## 476 ATCGTGGT GTCCTAAG        75
## 477 CACTTCAC TCGGATTC        75
## 478 CGATCGAT ATCGTGGT        75
## 479 TCGACAAG AGGATAGC        75
## 480 GATCAAGG AACAGCGA        74
## 481 CGGTAATC TCGAGAGT        73
```

```
482 GATCTTGC GATCAAGG   72
483 AGAGTCCA GATCTTGC   71
484 CACTTCAC AGAGTCCA   71
485 CGGTAATC CACTTCAC   70
486 GTCCTAAG CGGTAATC   70
487 TAGCCATG GATCTTGC   70
488 AGGATAGC GATCAAGG   69
489 AACAGCGA CGGTAATC   68
490 CGATCGAT GATCAAGG   68
491 GATCTTGC TCGAGAGT   68
492 CGATCGAT AACAGCGA   67
493 ACGATCAG GATCTTGC   66
494 GATCTTGC ACGATCAG   66
495 GTAGCGTA GATCTTGC   66
496 TCGGATTC GTCCTAAG   66
497 CACTTCAC ATCATGCG   65
498 GTAGCGTA CGGTAATC   65
499 GTAGCGTA TCGGATTC   65
500 TCGGATTC ATCATGCG   65
501 ACGATCAG TCGGATTC   64
502 ATCGTGGT TCGGATTC   64
503 CGGTAATC TCGGATTC   64
504 TAGCCATG CGATCGAT   64
505 AGAGTCCA CACTTCAC   63
506 CGATCGAT TCGGATTC   63
507 AACAGCGA CACTTCAC   62
508 CACTTCAC TCGAGAGT   62
509 CGATCGAT GATCTTGC   62
510 GCTACTCT TCGGATTC   62
511 TAGCCATG GCTACTCT   62
512 TCGGATTC CGGTAATC   62
513 AACAGCGA TCGACAAG   61
514 CGGTAATC GATCAAGG   61
515 GATCTTGC GCTACTCT   61
516 TCGACAAG TATGGCAC   61
517 CGGTAATC GTAGCGTA   59
518 TAGCCATG CGGTAATC   59
519 ATCGTGGT GATCTTGC   57
520 CACTTCAC GTAGCGTA   57
521 GATCAAGG TCGACAAG   57
522 CGATCGAT CACTTCAC   56
523 CGGTAATC ATCATGCG   56
524 GATCTTGC TAGCCATG   56
525 GCTACTCT CGGTAATC   56
526 AGAGTCCA TCGACAAG   55
527 CGATCGAT TCGACAAG   55
528 TCGACAAG AGAGTCCA   55
529 TCGGATTC TAGCCATG   54
530 ATCGTGGT TCGACAAG   53
531 CGGTAATC GATCTTGC   53
532 GATCTTGC CGGTAATC   53
533 AGGATAGC TCGACAAG   52
534 TCGACAAG GATCAAGG   52
535 CGGTAATC ACGATCAG   50
```

```
## 536 TCGAGAGT CACTTCAC       49
## 537 CACTTCAC ATCGTGGT       48
## 538 CACTTCAC GATCAAGG       48
## 539 CGGTAATC AGAGTCCA       48
## 540 TCGGATTC GATCAAGG       48
## 541 GTCCTAAG CACTTCAC       47
## 542 TCGACAAG CGATCGAT       47
## 543 TCGACAAG GTAGCGTA       47
## 544 TCGGATTC ATCGTGGT       47
## 545 CACTTCAC GTCCTAAG       46
## 546 CGGTAATC GTCCTAAG       46
## 547 GATCAAGG CGGTAATC       46
## 548 GTAGCGTA CACTTCAC       46
## 549 CACTTCAC AACAGCGA       45
## 550 GATCTTGC CACTTCAC       45
## 551 TCGACAAG TCGGATTC       45
## 552 TCGGATTC CACTTCAC       45
## 553 GTAGCGTA TCGACAAG       44
## 554 TCGACAAG ATCGTGGT       44
## 555 TCGACAAG GCTACTCT       44
## 556 TCGGATTC GTAGCGTA       43
## 557 CACTTCAC GATCTTGC       41
## 558 CGGTAATC AACAGCGA       41
## 559 TCGACAAG AACAGCGA       41
## 560 TCGGATTC GCTACTCT       40
## 561 CACTTCAC GCTACTCT       38
## 562 CGGTAATC GCTACTCT       37
## 563 GATCAAGG TCGGATTC       37
## 564 GCTACTCT CACTTCAC       37
## 565 GATCTTGC TCGGATTC       36
## 566 GATCTTGC ATCGTGGT       33
## 567 GCTACTCT TCGACAAG       33
## 568 ATCGTGGT CACTTCAC       29
## 569 CACTTCAC TCGACAAG       27
## 570 CGGTAATC TCGACAAG       26
## 571 TCGGATTC TCGACAAG       25
## 572 GATCAAGG CACTTCAC       24
## 573 TCGACAAG GATCTTGC       23
## 574 TCGACAAG CGGTAATC       22
## 575 GATCTTGC TCGACAAG       19
## 576 TCGACAAG CACTTCAC       14
## 577  Unknown  Unknown   5495481
```

The first 24 lines give the number of retained reads for each of our expected (matching) index pairs. In total, these account for 239750646 of the total 245598505; 97.6% of the retained reads. (Note: I've indcluded the Sorted_Index_counts.txt file with the breakdown of reads by index pair with this document, since the output is slightly unwieldy here.)

  b. There are a total of 352378 reads that have swapped indexes. Note that this does not include the counts of reads that contained indexes that did not match to any expected library (our "Unknown" counts)

  c. We can visualize the read counts per swapped index-pairs by plotting the first index against the second index in a dotplot and a heat map (Note: to keep the heatmap from oversaturation for the outlier read counts, all read counts have been log corrected such that the heatmap represents the distribution of log(Read_Counts)):

15

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```


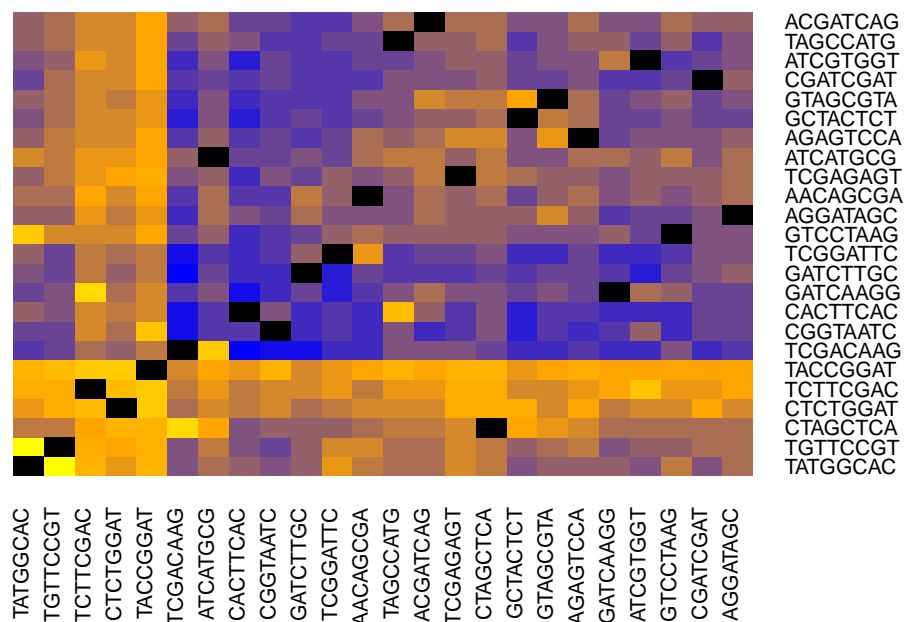Read Counts for unmatched Index Pairs
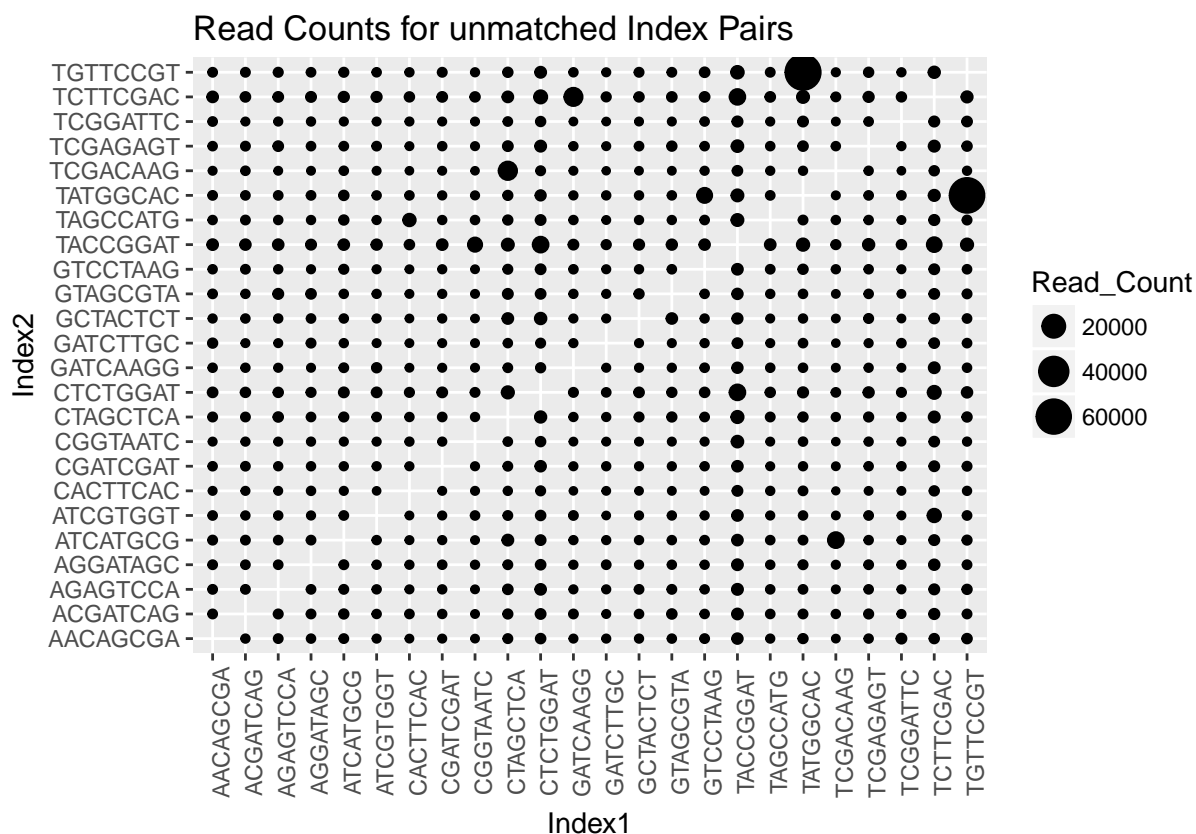
```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

Color Key and Histogram
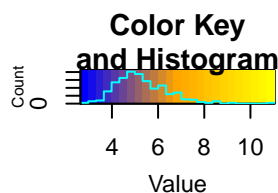
Map of log−corrected Read Counts for Index Swapped pairs

These two visualizations illustrate that certain index swaps is much more likely to occur. The TATGGCAC (C4) index had the highest counts among the swapped index pairs, with TCTTCGAC (C10) and CTCTGGAT (B3) having very high representation as well. This makes sense, as these were also the libraries that had shared the largest proportion of the total read counts; however, the converse does not appear to be true, with the lowest represented libraries (CACTTCAC, TCGACAAG, and GATCTTGC) not appearing to be represented that much lower than some of the libraries with 2x or more reads. This could indicate a potential "read representation limit" below which index swapping does not become significantly more common.

3. No data to report – read files were not created due to file limit constraints.