

7iM-^M-^T 03, 25 13:43

run-polap-r-pairs.R

Page 1/9

```

1  #!/usr/bin/env Rscript
2
3  #####
4  # This file is part of polap.
5  #
6  # polap is free software: you can redistribute it and/or modify it under the
7  # terms of the GNU General Public License as published by the Free Software
8  # Foundation, either version 3 of the License, or (at your option) any later
9  # version.
10 #
11 # polap is distributed in the hope that it will be useful, but WITHOUT ANY
12 # WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR
13 # A PARTICULAR PURPOSE. See the GNU General Public License for more details.
14 #
15 # You should have received a copy of the GNU General Public License along with
16 # polap. If not, see <https://www.gnu.org/licenses/>.
17 #####
18
19 # polaplib/run-polap-r-pairs.R
20 # Check: 2025-06-17
21
22 #####
23 # NOTE: this is the first of the two read-selection scripts.
24 # This script selects reads using some of minimap2 output files.
25 # We have multiple scripts that selects reads.
26 # This is the 2nd one.
27 #
28 # 1. run-polap-pairs.R -> the first version used in v0.2.6
29 # 2. run-polap-r-pairs.R -> used in test-reads in oga
30 # 3. run-polap-r-bridge.R -> used in test-reads in oga
31 # 4. run-polap-r-select-reads-polap.R -> the 2nd version not used
32 # 5. run-polap-r-select-reads-ptgaul.R -> a slim version of the 2nd used in dga
33 # 6. run-polap-r-directional.R -> used by an older version of dga
34 #    or _run_polap_original-directional-reads
35 #
36 # Subcommand test-reads uses this script.
37 #
38 # Used by:
39 # function _run_polap_test-reads { # selects reads mapped on a genome assembly
40 #
41 # See Also:
42 # run-polap-pairs.R
43 # run-polap-r-pairs.R
44 # run-polap-r-bridge.R
45 # run-polap-r-select-reads-polap.R
46 # run-polap-r-select-reads-ptgaul.R
47 # run-polap-r-directional.R
48 # run-polap-function-oga.sh
49 # run-polap-function-dga.sh
50 #
51 # Check: 2025-06-17

```

7iM-^M-^T 03, 25 13:43

run-polap-r-pairs.R

Page 2/9

```

52 #####
53
54 #####
55 # This script selects reads using minimap2 alignments of the reads on the seed
56 # contigs. It tries to use two main different approaches; one is due to ptGAUL,
57 # another is devised by Polap. Polap's read selection is more stringent in a
58 # way that it selects reads that are mapped on two seed contigs and those that
59 # are mapped completely within a contig.
60 # This has some benefit or better assembly for a smaller dataset; i.e., 10x.
61 # If reads are enough, ptGAUL method seems to work great.
62 #
63 # Check: 2025-06-16
64 #####
65
66 # name: select long reads
67 #
68 # synopsis:
69 #   run-polap-pairs.R mt.contig.name-1 contig.tab ${MTSEEDSDIR} $SINGLE_MIN $SINGLE_MIN
70 #
71 # requirement: executes Flye
72 # flye --nano-raw "$LR3K" --out-dir "${_arg_outdir}" \
73 #   --threads "${_arg_threads}" \
74 #   --stop-after contigger \
75 #   --asm-coverage 30 \
76 #   --genome-size "$EXPECTED_GENOME_SIZE"
77 #
78 # input:
79 #   1. mt.contig.name-1 - contig or edge numbers
80 #   2. contig.tab - minimap2 output modified
81 #   3. seeds directory for output
82 #   4. pair minimum length V11: 3000 for MT, 1000 for PT
83 #   5. bridge minimum length V7: depends: 3000 or 5000
84 #   6. single minimum length V11: 3000 for MT, 0 or 1000 for PT
85 #
86 # output:
87 #   single.names and <contig1-contig2.name> files in the output directory
88 #
89 # MTSEEDSDIR="${_arg_outdir}"/60-mt-${STEP4}/o${MR}/seeds
90 #   run-polap-pairs.R mt.contig.name-1 contig.tab ${MTSEEDSDIR} $PAIR_MIN $BRIDGE_MIN $SINGLE_MIN
91 #   run-polap-pairs.R mt.contig.name-1 contig.tab ${MTSEEDSDIR} $PAIR_MIN $BRIDGE_MIN
92
93 suppressPackageStartupMessages(library("optparse"))
94 suppressPackageStartupMessages(library("dplyr"))
95 suppressPackageStartupMessages(library("readr"))
96 suppressPackageStartupMessages(library("purrr"))
97 suppressPackageStartupMessages(library("tidyr"))
98 suppressPackageStartupMessages(library("ggplot2"))
99
100 debug <- Sys.getenv("_POLAP_DEBUG", unset = "0")
101
102 parser <- OptionParser()

```

7iM-^M-^T 03, 25 13:43

run-polap-r-pairs.R

Page 3/9

```

103 parser ← add_option(parser, c("-t", "--table"),
104   action = "store",
105   help = "minimap2 PAF",
106   metavar = "<FILE>"
107 )
108 parser ← add_option(parser, c("-m", "--mtcontigname"),
109   action = "store",
110   help = "mt.contig.name-1",
111   metavar = "<FILE>"
112 )
113 parser ← add_option(parser, c("-o", "--out"),
114   action = "store",
115   help = "output directory"
116 )
117 parser ← add_option(parser, c("-r", "--pair-min"),
118   type = "integer",
119   default = 3000,
120   help = "Minimum length of pair-mapping alignment",
121   metavar = "number"
122 )
123 parser ← add_option(parser, c("-w", "--single-min"),
124   type = "integer",
125   default = 3000,
126   help = "Minimum length of single-mapping alignment",
127   metavar = "number"
128 )
129 parser ← add_option(parser, c("-x", "--bridge-min"),
130   type = "integer",
131   default = 3000,
132   help = "Minimum length of bridging reads",
133   metavar = "number"
134 )
135 parser ← add_option(parser, c("--intra-base-ratio"),
136   type = "double",
137   default = 0.7,
138   help = "Minimum of V10/V11",
139   metavar = "number"
140 )
141 parser ← add_option(parser, c("--intra-read-ratio"),
142   type = "double",
143   default = 0.7,
144   help = "Minimum of (V4-V3)/V2",
145   metavar = "number"
146 )
147 parser ← add_option(parser, c("--inter-base-ratio"),
148   type = "double",
149   default = 0.7,
150   help = "Minimum of V10/V11",
151   metavar = "number"
152 )
153 parser ← add_option(parser, c("--create-ptgaul"),

```

7iM-^M-^T 03, 25 13:43

run-polap-r-pairs.R

Page 4/9

```

154   action = "store_true",
155   default = FALSE, help = "will create ptgaul.names"
156 )
157 parser ← add_option(parser, c("--create-single"),
158   action = "store_true",
159   default = FALSE, help = "will create single.names"
160 )
161 parser ← add_option(parser, c("--create-pair"),
162   action = "store_true",
163   default = FALSE, help = "will create pair.names"
164 )
165 parser ← add_option(parser, c("--create-combined"),
166   action = "store_true",
167   default = FALSE, help = "will create combined.names"
168 )
169 parser ← add_option(parser, c("--all"),
170   action = "store_true",
171   default = FALSE, help = "will create all the 4 names files."
172 )
173 parser ← add_option(parser, c("--outptgaul"),
174   action = "store",
175   default = "ptgaul.names",
176   help = "ptgaul output file"
177 )
178 parser ← add_option(parser, c("--outsingle"),
179   action = "store",
180   default = "single.names",
181   help = "single output file"
182 )
183 parser ← add_option(parser, c("--outpair"),
184   action = "store",
185   default = "pair.names",
186   help = "pair output file"
187 )
188 parser ← add_option(parser, c("--outcombined"),
189   action = "store",
190   default = "combined.names",
191   help = "combined output file"
192 )
193
194 args1 ← parse_args(parser)
195
196 if (is_null(args1$table)) {
197   s ← "bioprojects"
198   o ← "PRJNA817235-Canavalia_ensiformis"
199
200   input_dir0 ← file.path(".")
201   input1 ← file.path(input_dir0, "contig.tab")
202   input2 ← file.path(input_dir0, "mt.contig.name-1")
203   output1 ← file.path(input_dir0, "02-reads")
204

```

7iM-^M-^T 03, 25 13:43

run-polap-r-pairs.R

Page 5/9

```

205   args1 <- parse_args(parser, args = c(
206     "--table", input1,
207     "--mtcontigname", input2,
208     "-o", output1,
209     "-r", 10000,
210     "-w", 10000,
211     "-x", 10000,
212     "--all",
213     "--intra-base-ratio", 0.7,
214     "--intra-read-ratio", 0.7,
215     "--inter-base-ratio", 0.7
216   ))
217 }
218
219 # mt.contig.name-x
220 x1 <- as_tibble(read.table(args1$mtcontigname))
221
222 # https://lh3.github.io/minimap2/minimap2.html
223 # Li 2008 - "OUTPUT FORMAT: Minimap2 outputs mapping positions in the
224 # Pairwise mApping Format (PAF) by default. PAF is a TAB-delimited text format
225 # with each line consisting of at least 12 fields as are described in the
226 # following table:
227 #
228 # Col Type Description
229 # 1 string Query sequence name
230 # 2 int Query sequence length
231 # 3 int Query start coordinate (0-based)
232 # 4 int Query end coordinate (0-based)
233 # 5 char  âM-^@M-^X+âM-^@M-^Y if query/target on the same strand; âM-^@M-^X-âM-^@M-^Y if opposite
234 # 6 string Target sequence name
235 # 7 int Target sequence length
236 # 8 int Target start coordinate on the original strand
237 # 9 int Target end coordinate on the original strand
238 # 10 int Number of matching bases in the mapping
239 # 11 int Number bases, including gaps, in the mapping
240 # 12 int Mapping quality (0-255 with 255 for missing)
241 #
242 # Read in the data with assigned column names
243 data <- read_tsv(args1$table,
244   show_col_types = FALSE,
245   col_names = c(
246     "rname",
247     "rlen",
248     "rstart",
249     "rend",
250     "strand",
251     "cname",
252     "clen",
253     "cstart",
254     "cend",
255     "match",

```

7iM-^M-^T 03, 25 13:43

run-polap-r-pairs.R

Page 6/9

```

256     "base"
257   )
258 )
259 mtdir <- args1$out
260 pair_min <- as.numeric(args1$`pair-min`)
261 brigde_min <- as.numeric(args1$`bridge-min`)
262 single_min <- as.numeric(args1$`single-min`)
263 a <- as.numeric(args1$`intra-read-ratio`)
264 b <- as.numeric(args1$`intra-base-ratio`)
265 c <- as.numeric(args1$`inter-base-ratio`)
266 ptgaul_option_base <- paste("ptgaul",
267   a,
268   b,
269   single_min,
270   c,
271   pair_min,
272   brigde_min,
273   sep = "-"
274 )
275 ptgaul_option_base <- paste0(ptgaul_option_base, ".names")
276
277 single_option_base <- paste("single",
278   a,
279   b,
280   single_min,
281   c,
282   pair_min,
283   brigde_min,
284   sep = "-"
285 )
286 single_option_base <- paste0(single_option_base, ".names")
287
288 pair_option_base <- paste("pair",
289   a,
290   b,
291   single_min,
292   c,
293   pair_min,
294   brigde_min,
295   sep = "-"
296 )
297 pair_option_base <- paste0(pair_option_base, ".names")
298
299 combined_option_base <- paste("combined",
300   a,
301   b,
302   single_min,
303   c,
304   pair_min,
305   brigde_min,
306   sep = "-"

```

7iM-^M-^T 03, 25 13:43

run-polap-r-pairs.R

Page 7/9

```

307 )
308 combined_option_base ← paste0(combined_option_base, ".names")
309
310 # 0. Check preconditions
311 # rstart is not greater than rend.
312 # match is not greater than base.
313 result ← data %>%
314   summarize(all_rows_meet_condition = all(rstart ≤ rend & match ≤ base))
315
316 # Check the result
317 stopifnot(result$all_rows_meet_condition)
318 stopifnot(pair_min ≥ 0, bridge_min ≥ 0, single_min ≥ 0)
319
320 # 1. ptGAUL: we use ptGAUL for the case of single edge reference
321 if (args1$`create-ptgaul` || args1$all) {
322   ptgaul_file ← file.path(mtdir, args1$outptgaul)
323   ptgaul_mapped_reads ← data |>
324     filter(match / base > args1$`intra-base-ratio`, base > single_min) |>
325     select(rname) |>
326     distinct(rname)
327
328   ptgaul_mapped_reads |>
329     write.table(
330       ptgaul_file,
331       row.names = FALSE,
332       col.names = FALSE,
333       quote = FALSE
334     )
335   ptgaul_option_file ← file.path(mtdir, ptgaul_option_base)
336   print(ptgaul_file)
337   file.copy(ptgaul_file, ptgaul_option_file)
338   # file.copy(ptgaul_file, ptgaul_option_file, showWarnings = FALSE)
339 }
340
341 # 2. intra-contig mapping
342 if (args1$`create-single` || args1$all) {
343   single_file ← file.path(mtdir, args1$outsingle)
344   intra_contig_mapped_reads ← data |>
345     filter(
346       (rend - rstart) / rlen > args1$`intra-read-ratio`,
347       match / base > args1$`intra-base-ratio`,
348       base > single_min
349     ) |>
350     select(rname) |>
351     distinct(rname)
352
353   intra_contig_mapped_reads |>
354     write.table(single_file, row.names = FALSE, col.names = FALSE, quote = FALSE)
355
356   single_option_file ← file.path(mtdir, single_option_base)
357   print(single_file)

```

7iM-^M-^T 03, 25 13:43

run-polap-r-pairs.R

Page 8/9

```

358 file.copy(single_file, single_option_file)
359 # file.copy(single_file, single_option_file, showWarnings = FALSE)
360 }
361
362 if (nrow(x1) > 1) {
363   y <- t(combn(x1$V1, 2))
364
365   orders <- function(y) {
366     stopifnot(length(y) == 2)
367     z <- as.numeric(gsub("\\D", "", y))
368     bname <- paste0(z[1], "-", z[2])
369     oname <- paste0(mtdir, "/", bname, ".name")
370     y1 <- data |>
371       filter(
372         cname == y[1],
373         match / base > args1$`inter-base-ratio`,
374         base > pair_min,
375         rlen > brigde_min
376       ) |>
377       select(rname) |>
378       distinct(rname)
379     y2 <- data |>
380       filter(
381         cname == y[2],
382         match / base > args1$`inter-base-ratio`,
383         base > pair_min,
384         rlen > brigde_min
385       ) |>
386       select(rname) |>
387       distinct(rname)
388     intersect(y1, y2) |>
389     write.table(oname, row.names = FALSE, col.names = FALSE, quote = FALSE)
390   }
391
392
393   # 3. inter-contig mapping: individual pair
394   if (args1$all) {
395     apply(y, 1, orders)
396   }
397
398   # 4. inter-contig mapping
399   #
400   # Filter data based on the given criteria
401
402   if (args1$`create-pair` || args1$all) {
403     inter_contig_mapped_reads <- data |>
404       # Calculate the match/base ratio
405       mutate(match_base_ratio = match / base) |>
406       # Filter for match/base > 0.7 and base > 3000
407       filter(
408         match_base_ratio > args1$`inter-base-ratio`,

```


7iM-^M-^T 03, 25 13:43

run-polap-r-pairs.R

Page 9/9

```

409     base > pair_min,
410     rlen > brigde_min
411   ) |>
412   # Count the number of contigs for each read
413   group_by(rname) |>
414   # Keep only those reads that are mapped to at least two contigs
415   filter(n_distinct(cname) ≥ 2) |>
416   # Select unique reads
417   distinct(rname) |>
418   ungroup()
419
420   pair_file ← file.path(mtdir, args1$outpair)
421   inter_contig_mapped_reads |>
422     write.table(
423       pair_file,
424       row.names = FALSE,
425       col.names = FALSE,
426       quote = FALSE
427     )
428   pair_option_file ← file.path(mtdir, pair_option_base)
429   print(pair_file)
430   file.copy(pair_file, pair_option_file)
431   # file.copy(pair_file, pair_option_file, showWarnings = FALSE)
432 }
433
434 # 5. combined: single + pair
435 if (args1$`create-combined` || args1$all) {
436   combined_mapped_reads ← bind_rows(
437     inter_contig_mapped_reads,
438     intra_contig_mapped_reads
439   ) |>
440   distinct(rname, .keep_all = TRUE)
441
442   combined_file ← file.path(mtdir, args1$outcombined)
443   combined_mapped_reads |>
444     write.table(
445       combined_file,
446       row.names = FALSE,
447       col.names = FALSE,
448       quote = FALSE
449     )
450   combined_option_file ← file.path(mtdir, combined_option_base)
451   print(combined_file)
452   file.copy(combined_file, combined_option_file)
453   # file.copy(combined_file, combined_option_file, showWarnings = FALSE)
454 }
455 }

```