

7iM-^M-^T 01, 25 10:17

run-polap-r-bridge.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-bridge.R
20 # Check: 2025-06-17
21
22 #####
23 # NOTE: this is the second 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 3rd one.
27 #
28 # 1. run-polap-pairs.R or polap-r-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 # See Also:
39 # run-polap-pairs.R
40 # run-polap-r-pairs.R
41 # run-polap-r-bridge.R
42 # run-polap-r-select-reads-polap.R
43 # run-polap-r-select-reads-ptgaul.R
44 # run-polap-r-directional.R
45 # run-polap-function-oga.sh
46 # run-polap-function-dga.sh
47 #
48 # Check: 2025-06-17
49 #####
50
51 # name: select long reads

```

7iM-^M-^T 01, 25 10:17

run-polap-r-bridge.R

Page 2/9

```

52 #
53 # synopsis:
54 #   run-polap-pairs.R mt.contig.name-1 contig.tab ${MTSEEDSDIR} $SINGLE_MIN $SINGLE_MIN
55 #
56 # requirement: executes Flye
57 # flye --nano-raw "$LR3K" --out-dir "${_arg_outdir}" \
58 #   --threads "${_arg_threads}" \
59 #   --stop-after contigger \
60 #   --asm-coverage 30 \
61 #   --genome-size "$EXPECTED_GENOME_SIZE"
62 #
63 # input:
64 #   1. mt.contig.name-1 - contig or edge numbers
65 #   2. contig.tab - minimap2 output modified
66 #   3. seeds directory for output
67 #   4. pair minimum length V11: 3000 for MT, 1000 for PT
68 #   5. bridge minimum length V7: depends: 3000 or 5000
69 #   6. single minimum length V11: 3000 for MT, 0 or 1000 for PT
70 #
71 # output:
72 #   single.names and <contig1-contig2.name> files in the output directory
73 #
74 # MTSEEDSDIR="${_arg_outdir}"/60-mt-${STEP4}/o${MR}/seeds
75 #   run-polap-pairs.R mt.contig.name-1 contig.tab ${MTSEEDSDIR} $PAIR_MIN $BRIDGE_MIN $SINGLE_MIN
76 #   run-polap-pairs.R mt.contig.name-1 contig.tab ${MTSEEDSDIR} $PAIR_MIN $BRIDGE_MIN
77
78 suppressPackageStartupMessages(library("optparse"))
79 suppressPackageStartupMessages(library("dplyr"))
80 suppressPackageStartupMessages(library("readr"))
81 suppressPackageStartupMessages(library("purrr"))
82 suppressPackageStartupMessages(library("tidyr"))
83 suppressPackageStartupMessages(library("ggplot2"))
84
85 debug ← Sys.getenv("_POLAP_DEBUG", unset = "0")
86
87 parser ← OptionParser()
88 parser ← add_option(parser, c("-t", "--table"),
89   action = "store",
90   help = "minimap2 PAF",
91   metavar = "<FILE>"
92 )
93 parser ← add_option(parser, c("-m", "--mtcontigname"),
94   action = "store",
95   help = "mt.contig.name-1",
96   metavar = "<FILE>"
97 )
98 parser ← add_option(parser, c("-o", "--out"),
99   action = "store",
100   help = "output directory"
101 )
102 parser ← add_option(parser, c("-r", "--pair-min"),

```

7iM-^M-^T 01, 25 10:17

run-polap-r-bridge.R

Page 3/9

```

103   type = "integer",
104   default = 3000,
105   help = "Minimum length of pair-mapping alignment",
106   metavar = "number"
107 )
108 parser ← add_option(parser, c("-w", "--single-min"),
109   type = "integer",
110   default = 3000,
111   help = "Minimum length of single-mapping alignment",
112   metavar = "number"
113 )
114 parser ← add_option(parser, c("-x", "--bridge-min"),
115   type = "integer",
116   default = 3000,
117   help = "Minimum length of bridging reads",
118   metavar = "number"
119 )
120 parser ← add_option(parser, c("--intra-base-ratio"),
121   type = "double",
122   default = 0.7,
123   help = "Minimum of V10/V11",
124   metavar = "number"
125 )
126 parser ← add_option(parser, c("--intra-read-ratio"),
127   type = "double",
128   default = 0.7,
129   help = "Minimum of (V4-V3)/V2",
130   metavar = "number"
131 )
132 parser ← add_option(parser, c("--inter-base-ratio"),
133   type = "double",
134   default = 0.7,
135   help = "Minimum of V10/V11",
136   metavar = "number"
137 )
138 parser ← add_option(parser, c("--use-strand"),
139   action = "store_true",
140   default = FALSE, help = "will create ptgaul.names"
141 )
142 parser ← add_option(parser, c("--create-ptgaul"),
143   action = "store_true",
144   default = FALSE, help = "will create ptgaul.names"
145 )
146 parser ← add_option(parser, c("--create-single"),
147   action = "store_true",
148   default = FALSE, help = "will create single.names"
149 )
150 parser ← add_option(parser, c("--create-pair"),
151   action = "store_true",
152   default = FALSE, help = "will create pair.names"
153 )

```

7iM-^M-^T 01, 25 10:17

run-polap-r-bridge.R

Page 4/9

```

154 parser <- add_option(parser, c("--create-combined"),
155   action = "store_true",
156   default = FALSE, help = "will create combined.names"
157 )
158 parser <- add_option(parser, c("--all"),
159   action = "store_true",
160   default = FALSE, help = "will create all the 4 names files."
161 )
162 parser <- add_option(parser, c("--outptgaul"),
163   action = "store",
164   default = "ptgaul.names",
165   help = "ptgaul output file"
166 )
167 parser <- add_option(parser, c("--outsingle"),
168   action = "store",
169   default = "single.names",
170   help = "single output file"
171 )
172 parser <- add_option(parser, c("--outpair"),
173   action = "store",
174   default = "pair.names",
175   help = "pair output file"
176 )
177 parser <- add_option(parser, c("--outcombined"),
178   action = "store",
179   default = "combined.names",
180   help = "combined output file"
181 )
182
183 args1 <- parse_args(parser)
184
185 if (is_null(args1$table)) {
186   s <- "bioprojects"
187   o <- "PRJNA817235-Canavalia_ensiformis"
188
189   input_dir0 <- file.path(".")
190   input1 <- file.path(input_dir0, "contig.tab")
191   input2 <- file.path(input_dir0, "mt.contig.name-1")
192   output1 <- file.path(input_dir0, "02-reads")
193
194   args1 <- parse_args(parser, args = c(
195     "--table", input1,
196     "--mtcontigname", input2,
197     "-o", output1,
198     "-r", 10000,
199     "-w", 10000,
200     "-x", 10000,
201     "--all",
202     "--use-strand",
203     "--intra-base-ratio", 0.7,
204     "--intra-read-ratio", 0.7,

```

7iM-^M-^T 01, 25 10:17

run-polap-r-bridge.R

Page 5/9

```

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

```

7iM-^M-^T 01, 25 10:17

run-polap-r-bridge.R

Page 6/9

```

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

```

7iM-^M-^T 01, 25 10:17

run-polap-r-bridge.R

Page 7/9

```

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

```

7iM-^M-^T 01, 25 10:17

run-polap-r-bridge.R

Page 8/9

```

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

```



7iM-^M-^T 01, 25 10:17

run-polap-r-bridge.R

Page 9/9

```

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

```