hetDNA Mapping: Computer Pipeline Instructions for 2 or More SMRT cells

1. Copy each of the files listed below into a desktop folder that contains the SMRT data (CCS reads). The italicized files must be provided by the user. (See Note 1)

   "Barcodes.txt" contains all barcodes used for PCR amplification. Each barcode must be listed in forward and reverse orientations; see format in sample_files/Barcodes.txt.

   "Blastn" obtained from https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=ProTeins

   "BLASTscript.pl"

   "ccs.fasta" fasta files containing the CCS reads. If fastq files are provided, convert to fasta format (http://hannonlab.cshl.edu/fastx_toolkit/) and make sure that the file is renamed "ccs.fasta".

   "donor.txt" is the reference sequence of the donor allele in fasta format; see required format in sample_files/donor.txt.

   "PositionsWithSnps.pl"

   "ReadPlotterNew.R"

   "readSummary.pl"

   "recipient.txt" is the reference sequence of the recipient allele in fasta format; see required format in sample_files/recipient.txt.

   "SimpleFlagParser.pl"

2. Open the "Terminal" file found in Applications>Utilities (Figure 4a in publication)

   Type: **cd desktop**<return> (change directory; this directs the program to where the relevant folder is located)

   Type: **cd folder name**<return> (type the folder name to direct the program to the folder with the CCS data)

3. Locate the SNP positions in the donor and recipient (Figure 4b in publication):

   Command: **perl PositionsWithSnps.pl**<return>

   The program automatically inputs the "donor.txt" and "recipient.txt" files that are in the folder (labeled "Input" in subsequent steps).

New files named “PositionsWithSNPs.txt” and “SNPPositionsAlongRef.txt” are automatically created (labeled “Output” in subsequent steps). “PositionsWithSNPs.txt” is a file in which the end has the numbered positions of SNPs between the donor and recipient.

Open the “SNPPositionsAlongRef.txt” file and copy the line of SNP positions into the first line of the “ReadPlotterNew.R” file. Retain the “SNPS<-c()” part of “ReadPlotterNew.R” file and insert the numbered positions from “SNPPositionsAlongRef.txt” between the parentheses. Note that the next line in “ReadPlotterNew.R” designates the position of the initiating DSB (45.5, which is between the 45th and 46th SNPs). Depending on the substrates used, this may need to be changed.

4. Separate reads from different SMRT cells:

   Each circular read is labeled with an ID number assigned by Pacbio. ID numbers are assigned 1 to the total number of reads (per SMRT cell) and reset for every new SMRT cell used for a library. The SMRTCellParser.pl script will separate the reads, in a single ccs file, from multiple SMRT cells.

   Command: **perl SMRTCellParser.pl ccs.fasta**<return>

   Input: Combined fasta file (ccs.fasta)

   Output: Separate .txt files containing sequences from the different SMRT cells. These

   files will be labeled SMRTCell1, SMRTCell2, SMRTCell 3, etc.

   * For each SMRT cell, create a new folder and label as SMRT#, for instance (SMRT1). Copy and paste all the required files (see step 1) including the newly separated fasta file into each SMRT folder.

<u>For each SMRT cell, repeat the following steps 5-7.</u>

5. Align sequencing results by blasting against the donor sequence:

   Command: **./blastn -subject donor.txt -query ccs.fasta -out CONSENSUSDONOR**<return>

   Input: “donor.txt” and “ccs.fasta” files

   Output: “CONSENSUSDONOR” file

6. Align sequencing results by blasting against recipient reference sequence:

Command: **./blastn -subject recipient.txt -query ccs.fasta -out CONSENSUSRECIPIENT**<return>

Input: "recipient.txt" and "ccs.fasta" files

Output: "CONSENSUSRECIPIENT" file

7. SNP identity calling and barcode assignment:

   Command: **perl BLASTscript.pl CONSENSUSDONOR CONSENSUSRECIPIENT ccs.fasta**<return>

   Input: "CONSENSUSDONOR" and "CONSENSUSRECIPIENT" files

   Output: "OUTPUT" and "NoEmptyFlags.txt" files

8. Parsing of sequences containing barcodes (Figure 5 in publication; see Note 2):

   Command: **perl SimpleFlagParser.pl**<return>

   Input: "Barcodes.txt" and "OUTPUT" (must be in the same folder as the script)

   Output: "Matches.txt", "fileNames", and individual text files (e.g. "NCF1R1.txt", "NCF1R1count.txt", etc.) containing all the reads for each unique barcode pair

9. Combining all sequencing reads with the same forward and reverse barcodes from different SMRT cells

   Under each SMRT cell folder, move all individual barcoded files (F1R1.txt etc) into a new folder and name the folder as follows: "SM1" from the SMRT1 folder, "SM2" from the SMRT2 folder, etc.

   a. Create a new folder under the main folder on the desktop called "results".

   b. Move "SM1" and "SM2" out of their respective folders ("SMRT1" and "SMRT2") to the "results" folder. Copy and paste "find_merge_by_title.sh" script into the "results" folder.

   c. Change the directory of the terminal to the "results" folder:

      Command: **chmod +x find_merge_by_title.sh**<return>

      Command: **./find_merge_by_title.sh**<return>

Output: All combined files with the same barcodes should be listed in a newly created folder named “finalresult”

d. Change the directory of the terminal to the “finalresult” folder

Command: **ls > FN**

This command will create a file named “FN”, which is an input file for the next command to quantify the identities of each SNP position (see Note 3)

e. Copy “FlagCount.pl” and “SNPPositionsAlongRef.txt” into the “finalresult” folder

Command: **perl FlagCount.pl**

This step will finalize the total number of events associated with a specific barcode combination

9. Sequence analysis:

Command: **perl readSummary.pl**<return>

Input: All files listed in “fileNames” (“NCF1R1.txt”, “NCF1R1count.txt”, etc.)

Output: Ten informative data files each sample (see Figure 5 in publication).

10. SNP plots for the reads of each unique barcode pair:

Command: R<Enter>ReadPlotterNew.R<enter> (see Note 4)

Input: “fileNames”

Output: Individual PDF files (Figure 5 in publication) that display the species of reads for each barcode pair after filtering and condensation that is explained in “readSummary.pl”.

There are three separate displays for each sample in the corresponding pdf file (Figure 6 in publication). The first shows the total number of CCS reads and the counts of the recipient and donor SNPs at each position (yellow and blue symbols, respectively); the second presents the numbers of each read species after condensation and filtering; and the third shows the counts of each SNP after the final filtering (see Note 5).

Notes

1. The time for each step varies depending on the size of the library. In general, steps 6-8 take longer than other steps. Step 6 will take 1-4 h; steps 7 and 8 each take 1-2 h.
2. Before the barcode parsing step, check the main folder and delete any "F1R1.txt" or "matches.txt" files that are present; new ones need to be freshly created.
3. Click into the "FN" file to manually delete the FN.txt listed at the top of the file
4. If the program stops and gives an error on a particular barcode, open "fileNames", delete that barcode and re-run the command.
5. In analyses of the hetDNA patterns from this pipeline, only data from barcode pairs that have a minimum of 20 reads after the initial condensation and only two major species are considered informative. If a DSB occurs after DNA replication, the repair of each broken sister chromatid will generate four major species and these are omitted in the final analysis.