Final Report

Undergraduate Research

2015 Spring Semester

Name: Yuh Jen, Hwong

Student ID: R04945024

Advisor: Eric Y., Chuang

INTRODUCTION

It is believed that diseases are caused by pathogens and pathogens can be found in patients' RNA reads. Samples are taken from patient's tissue. It is assumed that the samples will contain human reads that belong to the patient, as well as non-human reads, that might come from virus or bacteria. The pathogens should lie among the non-human reads, therefore we hope to find a systematic way to analyze these pathogens from the samples.

METHODOLOGY

Tool

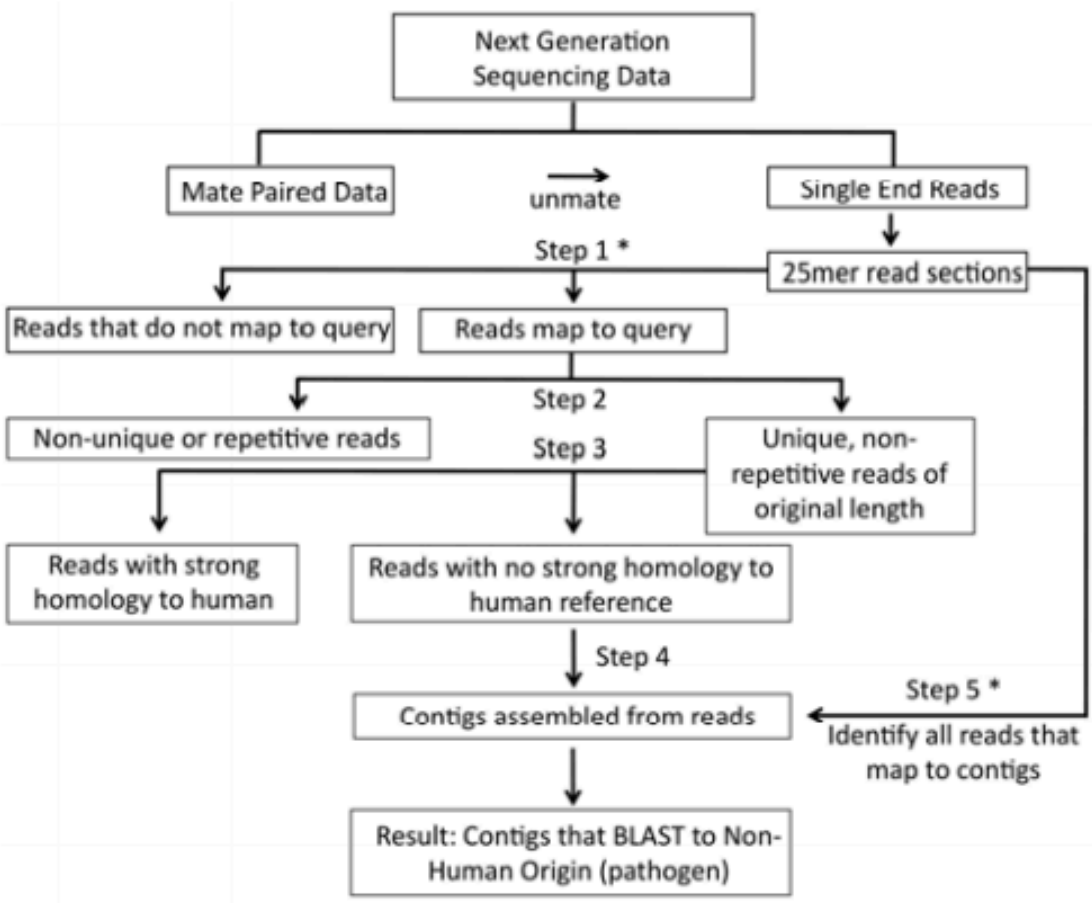
RINS, Rapid Identification of Non-human Sequences, is an intersection-based pathogen detection workflow that utilizes a user-provided custom reference genome set for identification of non-human sequences in deep sequencing datasets. It is invented in Stanford. Like its name, it has features of speed and local computing based nature. RINS' efficiency outperforms PathSEQ and thus serves as an alternative way to identify pathogens.

Sample

The sample I used in this test is generated by flux-simulator. The generated sample has a read length of 150bp and it is mixed with 30 million human reads and 0.1 millions pathogen reads.

Procedure

First, 25 mers of each read will be intersected with the non-human genomes dataset. Read sections that do not map to query will be discarded (1st removal). The remaining raw reads will be compared to the human genomes dataset. Reads with 97% homology are removed from the read set (2nd removal). The remaining non-human reads are first compressed and then matched with non-human genome dataset to be assembled into possible pathogen sequence contigs. This procedure is called iteration. After iteration, there will be some outlier reads that did not match with references to form contigs. These will be treated as noises and discarded (3rd removal). The default iteration frequency is set to 2, but users are free to modify the number. Lastly, RINS will spit out a table that contains contigs that are possible pathogens.



RESULTS

Sample type	pair-end
Sample source	manually generated
Sample (leftlane)	flux_homo_add20_R1_10000.fasta
Sample (rightlane)	flux_homo_add20_R2_10000.fasta
Sample content	30 million human reads 20*5000 non-human reads
Modifiable parameters	**Value**
iteration	2
raw_read_length	100
chop_read_length	50
minIdentity	80
compress_ratio_thrd	0.5
Total running time (RINS)	7 hours 19 minutes (26,340 s)
Total running time (PathSEQ)	3 days 16 hours 33 minutes (318,562 s)
Expected pathogens	20
Actual identified pathogens	20 (refer to Table 1)
Accuracy	100%

1. From the results, we can see that RINS successfully identified all 20 of the non-human contigs from among 30 million reads in a comparatively short period of time. For the same sample, PathSEQ took 318,562 seconds while RINS took only 26340 seconds, which is only one tenth of the running time of PathSEQ.
2. From the table we can see that every contigs matches a certain non-human sequence. This is because RINS only keeps those sequences that successfully match with non-human genome dataset. Human reads, unknown reads and noises are all filtered out in the previous steps.
3. The e-value gives a similarity of sequences. E-values of zero mean that there is an exact match for the sequence. The result is so because the exact pathogen contigs are manually added into the original non-human genome dataset.

contig_name	num	non_human_species	E-value	bit_score
comp6_c0_seq1	5000	Pseudo_7_length_6804	0.0	2468
comp4_c0_seq1	5000	Pseudo_5_length_6212	0.0	2372
comp7_c0_seq1	5000	Pseudo_10_length_6671	0.0	2669
comp1_c0_seq1	5000	Pseudo_20_length_6489	0.0	2015
comp17_c0_seq1	5000	Pseudo_3_length_6816	0.0	3506
comp8_c0_seq1	5000	Pseudo_6_length_7688	0.0	2708
comp3_c0_seq1	5000	Pseudo_8_length_5213	0.0	2239
comp5_c0_seq1	5000	Pseudo_1_length_7324	0.0	2385
comp0_c0_seq1	5000	Pseudo_14_length_6693	0.0	2047
comp9_c0_seq1	5000	Pseudo_12_length_7213	0.0	2922
comp11_c0_seq1	5000	Pseudo_9_length_7051	0.0	3227
comp19_c0_seq1	5000	Pseudo_2_length_8569	0.0	3596
comp15_c0_seq1	5000	Pseudo_17_length_5373	0.0	3419
comp2_c0_seq1	5000	Pseudo_19_length_3658	0.0	2442
comp12_c0_seq1	5000	Pseudo_4_length_6114	0.0	3157
comp18_c0_seq1	5000	Pseudo_16_length_7296	0.0	3535
comp13_c0_seq1	5000	Pseudo_13_length_6353	0.0	3332

comp16_c0_seq1	5000	Pseudo_11_length_5779	0.0	3626
comp14_c0_seq1	5000	Pseudo_15_length_4883	0.0	3262
comp10_c0_seq1	5000	Pseudo_18_length_5490	0.0	3042

Table 1. Final RINS output. Explanation for each column as below.

- **Column 1 – "contig_name"**
  shows the contigs that are matched to the non-human genome dataset. The names are randomly assigned by RINS.
- **Column 2 – "num"**
  represents the number of raw reads that fall on the contig (column 1). The values under this column are all 5000 because the sample is manually generated. We specially created 20*5000 non-human reads.
- **Column 3 – "non_human_species"**
  shows the matches that correspond with the contig in the non-human genome dataset.
- **Column 4 – "E-value"**
  shows how many times you would expect a result at least as extreme as the one observed occurring by chance. In this case, the lower the better, so we are glad to see zeros.
- **Column 5 – "bit_score"**
  represents the length of the contig.

DISCUSSION

1. **Cannot identify unknown pathogen (unless it is known and can be added to the non-human genome dataset).**

    Pathseq works by figuring out the identity of each read, whether it matches a human genome, a virus genome or neither. If it matches neither, the read will be labeled as unknown. On the other hand, RINS seeks for speed. In the first step, it resolutely filters out all the reads that do not match any of the references in non-human genome dataset. Suppose there is a new pathogen in which its sequence is unknown. The new pathogen will not be identified because it will be falsely discarded in the first step of RINS since it does not exist in the non-human genome dataset.

2. **Cannot evaluate new disease.**

    In RINS, all parameters can be set manually in "config.txt", including minimum contig length, iteration frequency... and different parameters generate different outcomes. It is hard to determine which results are the closest to the fact when there isn't an answer sheet to reference to. This is a problem when it comes to identifying a new disease.

    Suppose there is a new disease. The number of pathogens contained in the sample is unknown. RINS outputs 3 non-human contigs instead of 20 if you, perhaps, increase the number of 'minimum contig length' parameter from 5000 to 7000. With no expected result, nobody can determine whether there are 20 pathogens that causes this disease, or just 3.

3. **Unfriendly for amateurs.**

    In RINS, you are able to modify these parameters: iteration, raw read length, chop read length, minIdentity and compress ratio thread. However, it does not provide you with a range for each parameter or set of values, which you can reference to when inputting different type of samples. Compared to Pathseq, which carefully defines each read;

RINS provides a frame for the user. It is a problem when the user doesn't know what to expect. They have no idea how to set the values and they do not know how to interpret the outcomes. This is a drawback that comes with the flexibility of the tool.

REFERENCE

- Rapid identification of non-human sequences in high-throughput sequencing datasets.
  - http://www.ncbi.nlm.nih.gov/pubmed/22377895
- What is e-value
  - http://www.protocol-online.org/biology-forums/posts/5426.html
- PathSeq: software to identify or discover microbes by deep sequencing of human tissue
  - http://www.nature.com/nbt/journal/v29/n5/full/nbt.1868.html