Pattern matching with different mappers to improve LINE-1 recall

Juan O. Lopez, Mariliana T. Barrientos, Universidad de Puerto Rico Recinto de Arecibo

mariliana.Barrientos@upr.edu juano.lopez@upr.edu

This work uses the pattern matching for the identification of L1s made by the mentor. In its results we have the precision, recall and f1 score. the recall was relatively low, so that's my part in this investigation. my goal is to try different mappers to see an improvement in the result. Although some did work, a greater amount of data could not be obtained. I learned to use the aligners structure and to work with Linux commands. My future direction would be to test new mappers, fine-tune the parameters and on only the best of them, test how they work with the 75mers.

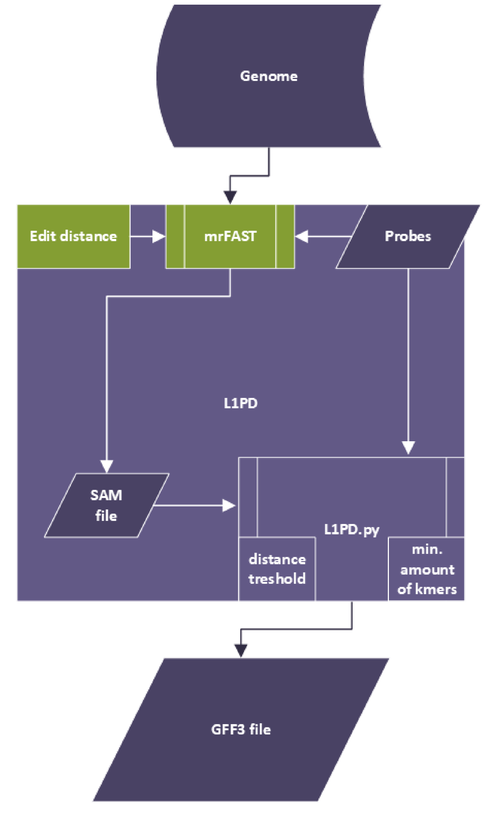
INTRODUCTION

In this research will attempt to improve the recall by studying how different mappers perform when using the same probes with the same pattern matching strategy. Changing the mapper could have different results since each one has a different algorithm and a specific way of working. Besides that, they have different parameters to be more specific, but others are already rigorous, and their parameters cannot be changed.

METHODOLOGY

Four mappers were used in this study: bwa, bowtie2, mummer4 and hisat2. These are compared to the one used previously which is mrFAST. Some have an index where the reference file and the query file are placed (GRCh38 and 50mers), and this index in some mappers are used in the search. Then the search engine is used to make a SAM format file, which will later be used with the algorithm that the mentor created. The same script is used, only the mapper is changed to then compare results.

LINE 1 DETECTION



RESULTS

In the results what we want is to improve the recall but also we consider the time. Unfortunately, only one aligner had a decent result while the others failed for different reasons. It is not known exactly what is wrong with these mappers since they have a predetermined algorithm. I tried to configure their parameters to make them like the aligner that gave result, even so it did not work. Despite the few results that are available, it can be seen that mrFAST had a better overall result.

Mappers	Presicion	Recall	F1 Score	Time
Bowtie2	.9036	.3651	.5201	1h 38min
Hisat2	[Failed] Good amount of data but syntax error.			75min
BWA	[Failed] Little amount of data.			116 in
mrFAST	.7885	.5949	.6782	33min

CONCLUSION

For this study we tested 4 mappers: BWA, mummer, Bowtie2 and hisat2. Just Bowtie2 gave good result, but less than expected. The algorithm created by the mentor needs more data, so that's why I don't consider that they were good results. it is notable that mrFAST was better on each outcome.

FUTURE WORK

In this research will attempt to improve the recall by studying how different mappers perform when using the same probes with the same pattern matching strategy. Changing the mapper could have different results since each one has a different algorithm and a specific way of working. Besides that, they have different parameters to be more specific, but others are already rigorous, and their parameters cannot be changed.

REFERENCES

Juan O. Lopez study: https://rdcu.be/c5b1c
Bowtie2 manual: https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
Hisat2 manual: http://daehwankimlab.github.io/hisat2/manual/
BWA manual: https://bio-bwa.sourceforge.net/bwa.shtml

ACKNOWLEDGEMENTS