Scientific Report

ESF GENOMIC-RESOURCES Exchange Grant 4115

Kisun Pokharel

July 10, 2013

Research Title: Transcriptome analysis to characterize prolific sheep *(Ovis aries)* breeds

Host Researcher: Prof. Goran Andersson, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden.

Duration: 1 month (June 2013)

1 Purpose of the visit

The objective of my visit was to learn transcriptome analysis methods. In my project, I am analyzing miRNA and mRNA data from different sheep breeds with varying litter sizes. Genes affecting fertility traits in sheep are complex and in addition, the sheep genome is poorly annotated. Therefore, I wanted to get overview of the methods that are employed in analysing sheep transcriptome. Also, I wanted to get exposure of good practices in working with big genomic data - how to manage big data and get the most out of it.

2 Description of the work carried out during the visit

During my exchange period, I discussed results from mRNA data analysis and annotated transcripts resulting from assembly of reads that were mapped against sheep reference genome. I also made de novo assembly of RNA-Seq data. In addition, I designed a pipeline to analysis miRNA data.

2.1 De novo assembly

Trinity [1] program was employed for de novo assembly of sheep transcriptome. Trinity combines three different software modules namely Inchworm, Chrysalis, and Butterfly. Inchworm assembles the RNA-seq reads into unique sequences of transcripts also known as contigs. Chrysalis clusters contigs and constructs de Brujin graphs. Finally, Butterfly processes each graphs and generates full-length transcripts.

Mate-ends and paired-ends from two data set were combined separately. The trinity job was run in Taito supercluster at CSC-IT Center for Science, Finland.

2.2 Transcriptome annotation

Blast2GO (B2G) [2] was used to annotate assembled reads resulting from genome/transcriptome based assembly as well as de novo assembly. The program makes blast search for query sequences and extracts gene ontology (GO) terms for each sequences. However, when we have large transcriptome data, it is useful to make local blast and import blast results into B2G. In addition, results from InterProScan can also be imported for motif analysis as it takes significant amount of time to query sequences against the database.

Contigs assembled by Trinity as well as transcripts from sheep-genome based assembly were blasted against local non-redundant (nr) database in Taito supercluster at CSC. Blast results were imported into Blast2GO program and were mapped against Gene Ontology (GO) terms and annotated against reliable functions.

3 Description of the main results obtained

A total of 277425 transcripts were assembled by Trinity with N50 value of 4568. The length of assembled transcripts range from 200nts to 41905nts. Due to the large number of assembled contigs, the transcripts were divided into smaller chunks based on the sequence length. All transcripts were blasted against local non-redundant (nr) protein database using blastx. The blast results were imported into Blast2Go for further annotation and analysis.

Mapping reads against sheep reference assembly resulted 32,503 transcripts, out of which 23,249 were annotated. A total of 1,047 transcripts did not find any hits in sequence database. Although majority of sequences were found in UniProtKB database, two other databases which consisted queried sequences were mouse genome database (MGI) and rat genome database (RGD). 21,468 transcripts also have functional classification in InterPro database.

4 Future collaboration with host institution

Researchers from both host and guest institution participated in a one-day meeting at host institution on the last day of my stay. Research projects from both parties were presented and discussed in detail. As there were many common research questions, both parties agreed to strengthen collaboration. I am planning to make another visit during Autumn 2013 for working with larger data set.

5 Projected publications/articles resulting from or to result from the grant

In addition to the results herein described, there are some tasks that need to be accomplished in next couple of months. A draft-paper has already been composed and we are planning to submit the work within next few months.

6 Other Comments

Prof. Goran Andersson, my host researcher, not only provided valuable suggestions regarding my project work, he also introduced other fellow bioinformaticians and group leaders from various genomics projects which helped me in many ways. Although one month was not enough for analysing next generation sequencing data, I had a very fruitful visit in terms of expanding the knowledge I had before. The results from this visit will be added to the manuscript that was in preparation before leaving to host institute.

My sincere appreciation goes to Prof. Goran Andersson for providing such a good learning environment, fruitful suggestions and collaboration. I would also like to thank Prof. Eric-Boncgam Rudolf, Dr. Fernando Lopes Pinto, Shahid Manzoor, Prof. Anna Nasholm, Dr. Anna Maria Johansson and Sangeet Lamichhaney for all their help and discussions. Last but not least, I would like to acknowledge the Advances in Farm Animals Genomic resources Research Networking Programme and the European Science Foundation for providing an exciting opportunity.

References

[1] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, and F. Palma, "Full-length transcriptome assembly from RNA-Seq data without a reference genome.," *Nature biotechnology*, vol. 29, pp. 644–52, July 2011.

[2] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.," *Bioinformatics (Oxford, England)*, vol. 21, pp. 3674–6, Sept. 2005.