Circularized Visualisation of Genetic Interactions

Irina Kuznetsova
PhD candidate
Graz University of Technology,
Graz, AUSTRIA
Institute of Interactive Systems and Data Science
i.kuznetsova@hci-kdd.org

Aleksandra Filipovska
Harry Perkins Institute of Medical Research and Centre for Medical Research and School of Molecular Sciences
The University of Western Australia,
Nedlands, Western Australia 6009
AUSTRALIA
aleksandra.filipovska@uwa.edu.au

Oliver Rackham
Harry Perkins Institute of Medical Research, Centre for Medical Research and School of Molecular Sciences
The University of Western Australia,
Nedlands, Western Australia 6009
AUSTRALIA
oliver.rackham@uwa.edu.au

Artur Lugmayr
Curtin University,
Perth, AUSTRALIA
Visualisation and Interactive Media Lab. (VisLab)
lartur@acm.org
http://www.artur-lugmayr.com

Andreas Holzinger
Graz University of Technology,
Graz, AUSTRIA
Institute of Interactive Systems and Data Science
a.holzinger@hci-kdd.org
http://hci-kdd.

ABSTRACT

Next Generation Sequencing (NGS) has been a powerful tool to investigate gene networks in biological sciences [1]. Visualisation of data produced by NGS is essential for the interpretation of the findings by biological scientists. Here we describe a workflow to image findings from a NGS sequencing methodology to investigate gene expression that can be visualised with Circus software [2]. Visualisation of these processes has provided biological scientists with valuable interpretation of high throughput data and identification of new transcripts.

Keywords

Circular RNA; Next Generation Sequencing; Big Data; Data visualization; Data Mining

1. INTRODUCTION

Modern technology has become a necessity in a range of different fields including chemistry, physics, computer science, statistics and biological sciences. Currently the cross-discipline driven research requires visualisation of findings and results in a succinct and clear manner to deliver the message to wide audiences. This is particularly valued in biological sciences where complex systems and genetic networks can be represented visually to identify their interactions and connections.

NGS generates massive datasets that contain fragments of genetic information [1]. Analysis of NGS is carried out by specialists with computer science backgrounds who generate large datasets that contain numerical information that is often not understood by biological scientists. Therefore visualisation of NGS results is essential to clearly summarize valuable findings about gene structure, regulation and genetic networks for the biological community.

2. RELATED WORKS

Deoxyribonucleic acid (DNA) is converted into *Ribonucleic acid (RNA)* in a process known as transcription. RNA is used as a blueprint to produce proteins [3], or to regulate gene expression by splicing RNA, silencing specific genes and act as a structural scaffold [3-7]. Changes in the levels of RNA can be a useful readout of gene expression under different conditions and in different cell types, tissues and organisms.

Our group is focused on investigation of RNA processes driven in parts of the cell known as mitochondria that contain their own genetic information. Recently, we analysed the regulation of RNA inside mitochondria using a new method that we developed that required visualisation of the genetic interactions that we identified. We developed a new pipeline that enables characterization of new transcripts and a new method to visualise these transcripts and their origin from the genome that is easy for biologists to interpret [8].

3. WORKFLOW DESCRIPTION

We used NGS for RNA data, and developed the pipeline described in our paper [8]. First of all, adaptor sequences were removed by applying cutadapt software [9]. Then the pair-end reads were merged into a single read with FLASH [10]. The output of FLASH is three files, where one contains merged reads, and two remaining files contain reads that were not merged. The next step is to find repetitive read parts. *Tandem Repeats Finder (TRF)* software was used to find the repetitive parts of a read [11]. This data was used for alignment against the mouse mitochondrial genome with bowtie2 in soft-clipping mode [12]. Unaligned read parts were extracted according to CIGAR information [13], and realigned with bowtie2, but with end-to-end mode [12]. The generated data that was converted to circus format was illustrated with Circos software [2] Figure 1.

© 2017 International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License.
WWW 2017, April 3-7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
DOI: http://dx.doi.org/10.1145/3041021.3058587

Figure 1 shows connection between genes in a normal and disease state. Figure1 (A) shows a normal state of mitochondrial RNA, and (B) illustrates changes in heart disease. Genes are shown in lilac, yellow, green and grey. For illustrative purposes we selected the Co1 gene, and shown its connection to other genes within the entire mitochondrial genome. Visualization is made with Circos [2].

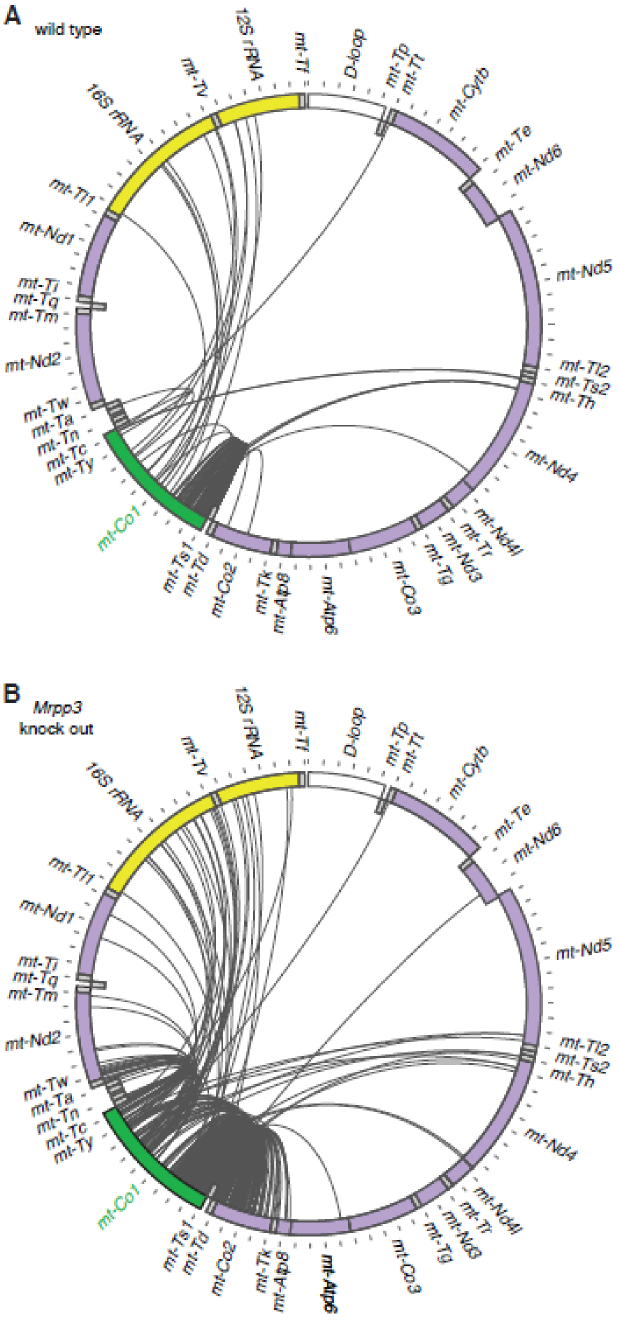

Figure 1. Connection of mtCo1 gene to other genes on mitochondrial genome. Visualization is made with Circos software [2].

4. CONCLUSIONS

The major biological conclusions are discussed at [8]. The presented workflow provides precise step by step explanation of data analysis for the dataset that derived from sequencing to the final stage of generating data format that is required for Circos software [2]. Such visualization provides a biologist with a map of genetic interactions and is essential in making conclusions about changes in gene interactions in disease.

5. ACKNOWLEDGEMENTS

We thank Leonid Brodsky, Yongjun Chu, David Corey for advice on the bioinformatics analyses

6. REFERENCES

[1] (2016). Next-Generation Sequencing (NGS). Available: http://www.illumina.com/technology/next-generation-sequencing.html

[2] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, et al., "Circos: an information aesthetic for comparative genomics," Genome Res, vol. 19, pp. 1639-45, Sep 2009.

[3] H. F. Lodish, Molecular cell biology, 4th ed. New York: W.H. Freeman, 2000.

[4] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," Cell, vol. 75, pp. 843-54, Dec 03 1993.

[5] B. Wightman, I. Ha, and G. Ruvkun, "Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans," Cell, vol. 75, pp. 855-62, Dec 03 1993.

[6] C. Suzanne, "RNA splicing: introns, exons and spliceosome," Nature Education 2008.

[7] J. L. Rinn and H. Y. Chang, "Genome regulation by long noncoding RNAs," Annu Rev Biochem, vol. 81, pp. 145-66, 2012.

[8] I. Kuznetsova, S. J. Siira, A. J. Shearwood, J. A. Ermer, A. Filipovska, and O. Rackham, "Simultaneous processing and degradation of mitochondrial RNAs revealed by circularized RNA sequencing," Nucleic Acids Res, Feb 15 2017.

[9] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," EMBnet.journal vol. volume 17, pp. 10-12, 2011.

[10] T. Magoc and S. L. Salzberg, "FLASH: fast length adjustment of short reads to improve genome assemblies," Bioinformatics, vol. 27, pp. 2957-63, Nov 1 2011.

[11] G. Benson, "Tandem repeats finder: a program to analyze DNA sequences," Nucleic Acids Res, vol. 27, pp. 573-80, Jan 15 1999.

[12] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," Nat Methods, vol. 9, pp. 357-9, Apr 2012.

[13] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., "The Sequence Alignment/Map format and SAMtools," Bioinformatics, vol. 25, pp. 2078-9, Aug 15 2009.