Opinion

NGS technology will revolutionize medical and biological research

Zang Li*

Biochemistry and Bioinformatics, Guangxi University, Guangxi, China

*Corresponding Author's E-mail: zangli@rediffmail.com

INTRODUCTION

Next-generation sequencing (NGS) technology offers a great opportunity to revolutionise a wide range of medical and biological research as well as their induced application fields, such as medical diagnosis, biotechnologies, virology, etc. (Alic et al., 2016). This is because NGS technology has an ever-increasing high throughput and dramatically decreasing cost. The sequences are not flawless, and there are a variety of errors present, including substitutions, insertions, deletions, and uncalled bases. For example, substitution error rates range from 1% to 2.5%, and insertion and deletion error rates can reach 40% (Kelley et al., 2010). The sequencing flaws in the data have made data analysis very difficult. So, the first and most important task is to fix these mistakes. a lot of downstream applications Corrected sequencing reads can be useful for many downstream applications, including sequence assembly, variant calling, read mapping, etc. (Salmela et al., 2011). Numerous strategies have been put out to fix faults, just a few of them are Coral, BLESS (Heo et al., 2014), and MEC (Zhao et al., 2017). These strategies strongly rely on k-mer.

K consecutive nucleotides make up a k-mer, which is a substring of a sequencing read. Mining of solid kmers is typically the first and most important phase in a k-mer-based technique. When a k-frequency mer's exceeds a certain minimal threshold, it is said to be solid, whereas the others are weak. This straightforward description successfully helps to distinguish between strong and weak k-mers, although it still has clear limits.

The main flaw is that a k-mer with a low frequency could not actually be weak. This is because system biases, such as the difficulty of sequencing sections with high GC concentration, lead the distribution of sequencing depth to be uneven. In order to do this, we concentrate on the significant yet understudied topic of refining solid k-mers utilizing.

Our model divides all k-mers into solid and liquid k-mers after counting k-mers using KMC2 depending on their frequency, provisionally set to be weak. At a later time, each k-and merits, the z-score and other factors are used to determine solidity jointly its frequency.

CONCLUSION

An essential component of many sequencing analyses, a k-mer is especially helpful for error correction, sequence assembly, variant calling, etc. K-mers are the cornerstone of several NGS applications. However, k-mers are prone to errors, which presents significant difficulties for further data processing. We provide a statistical method for clearly separating strong kmers from weak k-mers. For each k-mer, we precisely calculate a z-score, and using the z-score and frequency, we jointly decide if the k-mer is indeed solid. Studies reveal that our method successfully identifies solid kmers with low frequency.

A k-mer with a low frequency, however, may not be inaccurate due to bias and sequencing mistakes. Therefore, it is not ideal to discriminate solid k-mers just by frequency. We suggest a unique strategy of using z-score to identify incorrectly categorised weak and solid kmers in order to address this issue rather than ignoring it by existing ways. Research demonstrates that the z-score may be used to discriminate between genuine solid k-mers.

REFERENCES

1. Alic AS, Blanquer I, Dopazo J and Ruzafa D (2016). Objective review of de novo stand-alone error correction methods for NGS data. WIREs Comput Mol Sci. 6: 111-146.

2. Kelley DR, Schatz MC, Salzberg SL (2010). Quake: Quality-aware detection and correction of sequencing errors. Genome Biol. 11: R116.

3. Salmela L, Schroder J (2011). Correcting errors in short reads by multiple alignments. Bioinformatics. 27: 1455-1461.

4. Chen D, Hwu WM, Heo Y, Ma J, Wu XL (2014). BLESS: Bloom filter-based error correction solution for high-throughput sequencing reads. Bioinformatics. 30: 1354-1362.

5. Chen Q, Jiang P, Li W, Li J, Wong L, et al (2017). MapReduce for accurate error correction of next generation sequencing data. Bioinformatics. 33: 3844-3851.