Summary of Loh et al., 2012

Will Shoemaker

I519

The paper could be summarized as a "big data" paper, that is, with the seemingly immense amount of genomic data out there, how can researchers effectively use it while decreasing the amount of memory required? To start the argument on how this is possible, the authors invoke an argument similar to the law of diminishing returns in regards to their prediction of the future of bioinformatics. That is, the more we sequence, the less data we collect that is unique. This applies for the size of individual datasets (i.e. the more you sequence a genome, the more redundant your dataset becomes). This is a trend that can be exploited using more computationally tractable algorithms that work on compressed data. However, there is a limit to this. For example, while the genomes of each individual in a population can be compressed into a much more manageable size due to redundancy across individuals, a single genome can not be compressed that much. This is analogous to aligning reads to a reference from a population and retaining only the reference and a Variant Call Format file.

This sequence similarity allows for a reduction in the effective size of the data, which in turn allows for the data to be more easily managed. However, there's still the issue of analyzing the compressed data. While it's easier to transfer the data, it can't be effectively analyzed using current tools. Therefore, there is a great need for tools that analyze data in a compressed form. The need for this is discussed in the paper, however, some of the statements in the paper make it unclear whether the data compression schemes are lossy or lossless. Whether they are or not, the authors makes it clear that many of the current algorithms do not sufficiently recover the structure of the compressed data, complicating downstream analyses. The authors interprets this problem as a trade-off between the speed of the algorithm and the quality of the output.

One aspect of the paper that I particularly liked was how the authors explicitly brought up how algorithms that work on compressed data could help small, independent researchers effectively use genetic data as a resource. This was an important point for the authors to bring up, as often it is difficult to convince researchers who are more biological than computational that developing new algorithms and improving existing ones is something that is in their best interest. Overall it will be interesting to see how widespread the use of algorithms capable of handling compressed data becomes. This paper is approximately three years old, and while I am only starting to effectively use bioinformatics in my research, I had not heard of compressive genomics until reading this paper. Nevertheless, I predict that this is a topic that will only become more prominent in analyses that use bioinformatic tools as more and more smaller labs generate their own genomic datasets.