# Law of genetic linkage interdicts unlawful assembly

Limsoon Wong

National University of Singapore

**Abstract**

Modern genome assembly methods often produce “unlawful” assemblies that are fragmented, incomplete, and even contain misassembled segments. The quality of an assembly is often assessed by multiple metrics in terms of contiguity, completeness, and correctness. However, these metrics can be misleading and hard to combine into a single easy-to-interpret quality index. In the first part of this talk, we introduce and demonstrate a quality index that integrates contiguity, completeness, and correctness into a single number. This quality index is based on the law of genetic linkage; viz. the nearer two DNA sequences are to each other, the more likely they are inherited together. Specifically, this quality index measures what the average error is when one uses the assembly to estimate the distance between loci on the actual genome. A direct calculation of this index is obviously time-consuming, but we are able to derive an efficiently computable accurate approximation.

Unlawful assemblies can be corrected using additional data such as genetic maps and chromatin-interaction data. These additional data, when unavailable, may be difficult to construct and/or relatively costly to generate. However, in many application scenarios (e.g. progeny studies), the genome sequence and/or sequencing reads of multiple closely related strains are available. The law of genetic linkage governs the genome sequence of such progeny strains. In the second part of this talk, we show that this law is sufficient to produce better genome assembly in such scenarios, in the absence of a genetic map. In fact, it can even help correct misassembled long repeat regions. The improvement based on this approach is much better than the improvement made by several assembly-correction methods which use chromatin-interaction data.

This talk is based on the work of my student Luyu Xie.

**Brief bio**

Limsoon Wong is Kwan-Im-Thong-Hood-Cho-Temple Chair Professor in the School of Computing at the National University of Singapore (NUS). He currently works mostly on knowledge discovery technologies and their application to biomedicine. He is a Fellow of the ACM, named in 2013 for his contributions to database theory and computational biology.