Summary of the Contents

Toward complete reconstruction of repetitive sequences in metagenome and centromere

(メタゲノムおよびセントロメアにおける反復配列の完全な再構成に向けて)

Yoshihiko Suzuki (鈴木慶彦)

Sequence assembly, i.e., reconstruction of the original DNA sequences from their subsequences observed, has been one of the most principal challenges in bioinformatics since the dawn of its history. Repetitive sequences (*repeats*) are a major intrinsic source of the difficulty in assembling sequence fragments because they cause a serious ambiguity. Repetitiveness in a genome sequence takes several forms. The most typical repeats are interspersed repeats mainly caused by mobile genetic elements. Tandem repeats are abundant especially in eukaryotic genomes and known to be associated with many diseases and also as major components of centromeres in a variety of organisms. In a broad sense, ploidy of a genome is regarded as chromosome-scale repeats, and similar strains in a metagenome would be deemed as real-valued ploidy. The continuous evolution of sequencing technologies and algorithms for overcoming repeats has made it possible to achieve near-complete assembly of large genomes. Still now, however, there remain genomes and genomic regions difficult to perfectly reconstruct.

Various instances of the above-mentioned repeat types exist in real genomes. In this thesis, we especially focus on the following two relevant, unresolved research topics in genome science:

1) complete extrachromosomal mobile genetic elements (eMGEs), specifically plasmids and bacteriophages, in 12 human gut metagenomes; and
2) tandem repeats of ∼360 bp units of the 1.688 gm/cm^3 satellite family in centromeres of *Drosophila melanogaster* F1 (A4×ISO strains) females.

In both cases, we assume no prior (incomplete) sequence databases and thus assemble the reads *de novo*. The types of the reads assembled are continuous long reads of the PacBio RS II sequencer with an error rate ∼15% for metagenomes, and public circular consensus reads of the PacBio Sequel II sequencer with an error rate up to ∼1% for centromeres.

A common strategy in genome assembly is the *overlap-layout-consensus* paradigm that first detects true overlaps between reads from all possible read pairs. We remark why the traditional pairwise method of overlapping reads cannot resolve (approximate) repeats to rebuild the original sequences properly.

One can simply resolve approximate/exact repeats by spanning each repeat by reads. Finding a true overlap can be achieved with a high probability if the overlap is not contained in

maximal approximate repeats of the genome sequence. One approach for locating unsettled repeats is to utilize some additional guidance on true read pairs such as optical mapping and Hi-C contact information. On the other hand, read overlap with a model on the background genome sequence is promising in finding more true overlaps by integrating out the sequencing errors on the conserved regions among repeats in the genome model. This provides a rationale for variant-detection-based methods employed in some recent works. There is, however, an inevitable limit due to sequencing errors where some of the pairwise read overlaps could fail even if we employ an appropriate genome model. Therefore, simultaneous estimation of the genome model and read overlaps is required ultimately.

Chapter 2: Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut

For the reconstruction of complete eMGEs from metagenomes, we took the approach based on external information. We first generated conservative *unitigs* and then mapped the labels of unitig binning onto the string graph to untangle edges on chimeric nodes by concatenating edges consistently labeled with the same bin ID.

De novo assembly of continuous long reads from 12 human faecal samples (∼11 Gb total bases and ∼8 kb mean subread length per sample) generated 82 eMGE contigs (2.5-666.7 kbp), which were classified as 71 plasmids and 11 bacteriophages, including 58 novel plasmids and six bacteriophages, and complete genomes of five diverse crAssphages with terminal direct repeats. In a dataset of 413 gut metagenomes from five countries, many of the identified plasmids were highly abundant and prevalent. Host microbes of the assembled eMGEs were predicted using several signatures including DNA methylation motifs reported by a PacBio's official tool, and the result suggested that Bacteroidetes-associated plasmids predominated, regardless of microbial abundance.

Chapter 3: Assembling centromeric complex satellites in *Drosophila melanogaster* using a Bayesian repeat model

We developed a Dirichlet process mixture model of unit sequences as a background genome model for complex satellites to accurately detect overlaps between reads. Experiments with simulation data showed that our method described below perfectly reconstructed the entire sequence of a ∼100 kbp *in silico* synthetic centromere from ∼30× reads with 1% errors, while state-of-the-art assemblers did not.

We also assembled a real dataset. We downloaded public circular consensus reads (11 kb insert length, 5.5 Gb total bases) of *Drosophila melanogaster* F1 females. We first detected tandem repeats and their unit sequences from reads via self-vs-self read alignments, and collected centromeric reads *de novo* based on the assumption that centromeric tandem repeats are most abundant in a genome.

Since the inference of the generative model is intractable with all the centromeric units, we performed an initial overlap filtering based on sequence similarity calculated with pairwise sequence alignment. Initial all-vs-all read overlap with 2% maximum sequence dissimilarity and 5 kb minimum overlap length generated ∼7,000 overlaps in total (Fig. 1a). After that, for each read, we performed the inference of the mixture model, i.e., clustering of the units occurring in the read and other reads overlapped to it. The inference was performed with the split-merge sampling plus several heuristics for feasibility and speed. Each unit in the reads was encoded to the consensus unit estimated by the unit clustering, which corresponds to the maximum a posterior estimation of true units. Then, overlaps were re-calculated based on the encoded units with more strict maximum sequence dissimilarity: 0.3% (Fig. 1b). The total length of the remaining 36 contigs >30 kbp was ∼1.6 Mbp. There still remain several false overlaps probably due to higher-order (interspersed) repeats and structural variations. Compared with >30-kbp contigs produced by Canu with the same reads, our contigs connected the separate Canu contigs while the sequence contents mostly agreed, except for one case (Fig. 2).

Conclusion

Overall, we explored a framework and methods for sequence assembly of repeats and presented that long-read sequencing was effective for the identification of eMGEs as complete contigs and for the characterization of centromeric tandem repeats for better read overlap.

[1] Suzuki, Y. et al. Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. *Microbiome* 7, 119 (2019)

[2] Suzuki, Y. et al. Assembling centromeric complex satellites in *Drosophila melanogaster* using a Bayesian repeat model. *Manuscript in submission to Genome Research.*

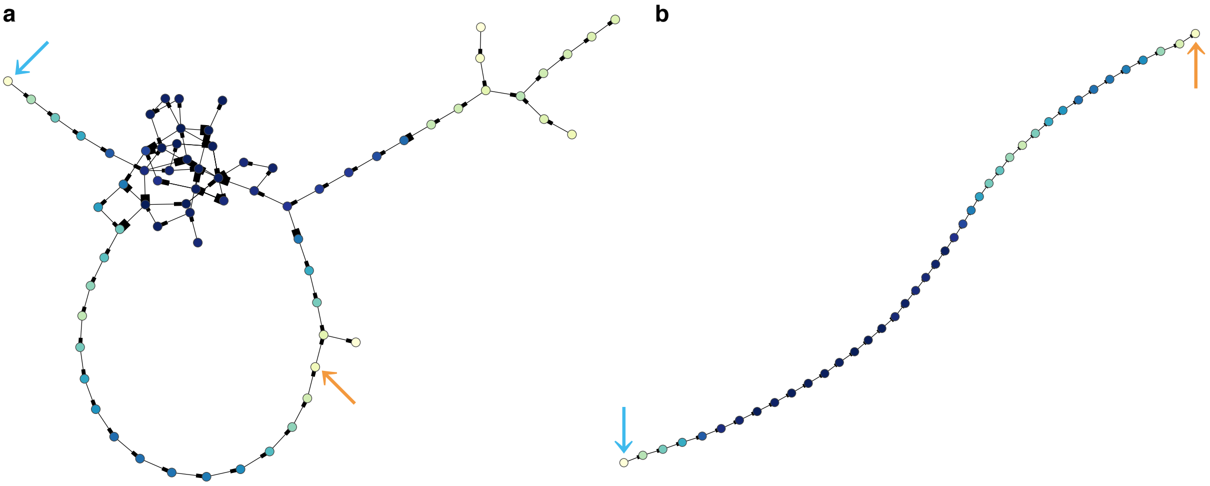

Figure 1: A component of the string graph constructed from overlaps between raw reads (**a**) and corrected reads (**b**). Arrows of the same color indicate identical nodes, meaning our method resolved this component.

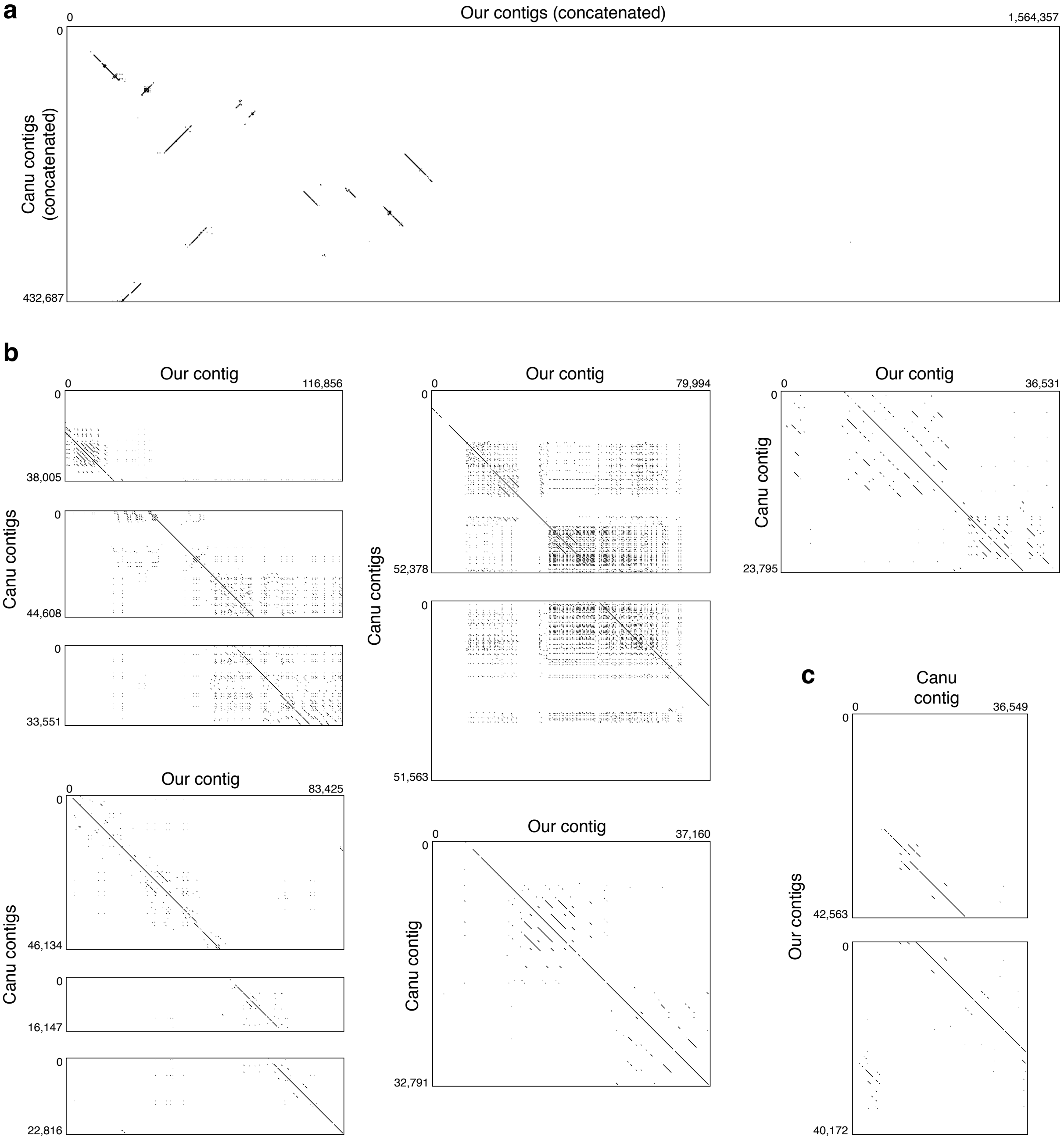

Figure 2: a. Dot plot between the concatenated sequence of the Canu contigs and that of our contigs (word size is 1 kbp). The sequences of the Canu contigs were mostly contained in our contigs. **b**. Dot plots of contigs that are similar between Canu and our method (word size is 300 bp). Our contigs resolved the breakpoints of Canu and produced more contiguous contigs. **c**. A case where contig sequences disagreed between Canu and our method.