論文の内容の要旨

論文題目 Assembly of nucleotide sequences of the multipartite structures in plant mitochondria genomes and the diploid human genomes
(植物ミトコンドリアゲノムおよび2倍体ヒトゲノムの塩基配列のアセンブリ)

氏　名 舛谷　万象

Genome assembly is the process of reconstructing the DNA sequence of a genome from observed subsequences (*reads*). It is one of the fundamental processes in biology because genome sequences are powerful tools to analyze how we have evolved from ancestors, how genomes relate to phenotype, and how much genetic diversity we have.

As a result of advances in technology and algorithms, the gapless assembly of a human haploid genome and nearly complete assemblies of other organisms are now available. These assemblies are called *reference genomes*. We typically analyze genomes of interest (e.g., individual genomes in a population) by representing them as single nucleotide variants (SNVs) and structural variations (SVs) on the references.

However, the difference between a genome sequence of a sample and a reference sequence can be so high that there are no obvious ways to represent the genome as SNVs and SVs on the reference. For example, the mitochondrial genomes in plants recombine considerably from strain to strain and organelle to organelle, shaping multipartite structures. These recombinations are complex, and SVs are insufficient to represent them (Fig. 1a). They are worth investigating because some genes in these genomes inhibit a plant from making mature pollen, called cytoplasmic male sterility. To fully represent the plant mitochondrial genomes, we need to assemble the genome and catalog the recombinations in each strain.

Another example is the major histocompatibility complex (MHC) region and the leukocyte receptor complex (LRC) region in human genomes. These immune-related regions have diverged, and the two haplotypes in an individual often have significant differences, which are hard to represent as SNVs and SVs on a single reference (Fig. 1b). Although many variants in these regions show correlations with diseases, the responsible genes are still ambiguous due to these differences and high linkage disequilibrium. To understand these regions, we need the complete assembly for each haplotype of these regions. In short, the first topic is: a genome is more than a set of SNVs and SVs on a reference.

I focus on (1): multipartite structures of the mitochondrial genomes in the eight strains of *Arabidopsis thaliana*, and (2): the MHC and the LRC region in two human diploid genomes.

Nevertheless, these cases contain nearly identical regions in addition to highly divergent regions. Specifically, the plant mitochondrial genome recombines frequently, but the mutation rate is low. Similarly, the MHC and LRC regions contain nearly identical regions with as low as 0.1% difference between haplotypes.

As sequencing technologies improve, DNA sequencers produce reads of tens to hundreds of thousands of base pairs, which can span these nearly identical regions. However, high sequencing error rates in these reads (5-15%) blur true SNVs in these regions and make it

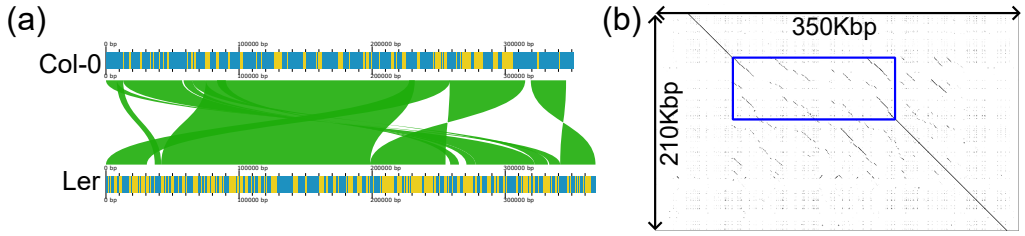

Figure 1: Examples of highly diverged regions in plant mitochondrial genomes (Col-0 and L*er* strain, a) and the human MHC region (two haplotypes, a).

difficult to distinguish SNVs from the sequencing errors. Can we correctly assemble the highly diverged regions and the nearly identical regions simultaneously by the reads with high error rates? This is the second topic I answer positively by implementing software named JTK. In short, genome assembly of difficult regions is possible with error-prone long reads.

JTK – regional genome assembler

JTK assembles a target region, such as the MHC region in a diploid genome, from erroneous Oxford Nanopore Technology (ONT) or PacBio reads. It takes three steps to assemble highly diverged regions and nearly identical regions simultaneously.

1. JTK samples 2000-bp subsequences (*chunks*) from the reads. Then, it aligns these chunks to the reads with a relaxed sequence similarity threshold so that multiple copies of nearly identical regions in the target region are represented as the same chunk. JTK captures the SVs and highly diverged regions through these alignments by the presence or absence of the chunks in the reads. Also, JTK estimates the copy number of each chunk in the target region by using how many times it is aligned to the reads.

2. JTK finds SNVs on these chunks and separates the chunks into homologous copies. To exhaustively enumerate SNVs on a chunk, JTK introduces each possible SNV to the chunk and accepts it as an actual SNV if the alignment scores of many reads increase. I developed a sampling-based algorithm to separate these SNVs into each copy. JTK further improves the separation on a chunk by integrating the results of nearby chunks.

3. JTK constructs a graph representing the highly diverged regions found in the first step and the SNVs inferred by the second step and produces the assembly by traversing the graph.

JTK is available at GitHub (`https://github.com/ban-m/jtk`).

Complex genome arrangements in mitochondrial genomes in the six strains of *Arabidopsis thaliana*.

I assembled the major conformations of the mitochondrial genomes in the nine datasets, consisting of eight strains and one biological replicate, from 350-fold reads with ∼15 % errors sequenced by PacBio Sequel sequencer. These assemblies were highly concordant with the

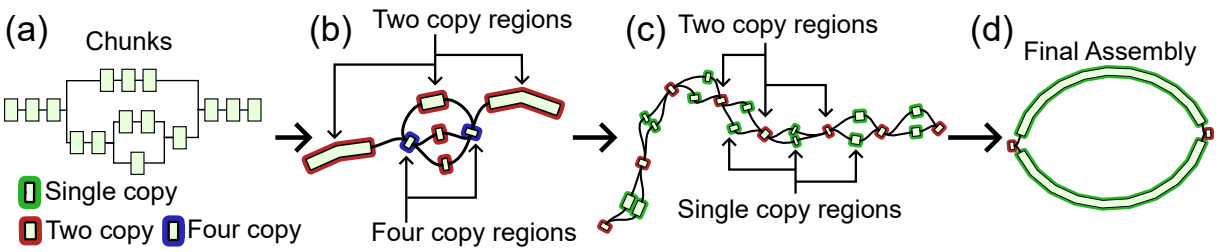

Figure 2: Overview of the **JTK** assembler

reference in the base pair level.

Nonetheless, there were many recombinations (Fig. 3). Their positions were close to the repetitive sequences in the genomes, confirming that the repeat-mediated recombinations have promoted these complex arrangements in this species. Also, I investigated the alignments between the reads and the assemblies and found putative recombinations and linear structure in the multipartite structures. These assemblies and the datasets are available at GitHub (`https://github.com/ban-m/reconstructmitogenome`).

Haplotype differences in the MHC and the LRC regions in human genomes.

JTK produced fully resolved assemblies for the MHC and the LRC region in HG002 and B080 samples (a Japanese B-cell) from 60-fold ONT reads, while other software based on the same datasets produced sub-optimal assemblies. The contiguities and the base-level accuracy of the JTK's assemblies were on par with those produced from high-coverage PacBio HiFi reads, Hi-C reads, and ultra-long ONT reads. The assemblies on the HG002 are available at Zenodo (`https://doi.org/10.5281/zenodo.7192214`).

In the LRC region in B080, the two haplotypes contained genes of the major haplotype in the Japanese population. However, there were around 0.2% substitutions between the haplotypes on average, and there was a 3Kbp expansion of a CT-rich region. Also, in the MHC region in the B080 (Fig. 4), a segmental duplication occurred in the class II region. These results revealed considerable variations among the haplotypes in the LRC region, which have been considered the same haplotype in course-grained resolution.

Conclusion

I could assemble the highly diverged regions with error-prone reads. In plant mitochondrial and human diploid genomes, newly assembling the genome provided insights into these genomes. Soon, complete genome sequences will be available on a population scale, and interpretation of them will be a fascinating field of research.

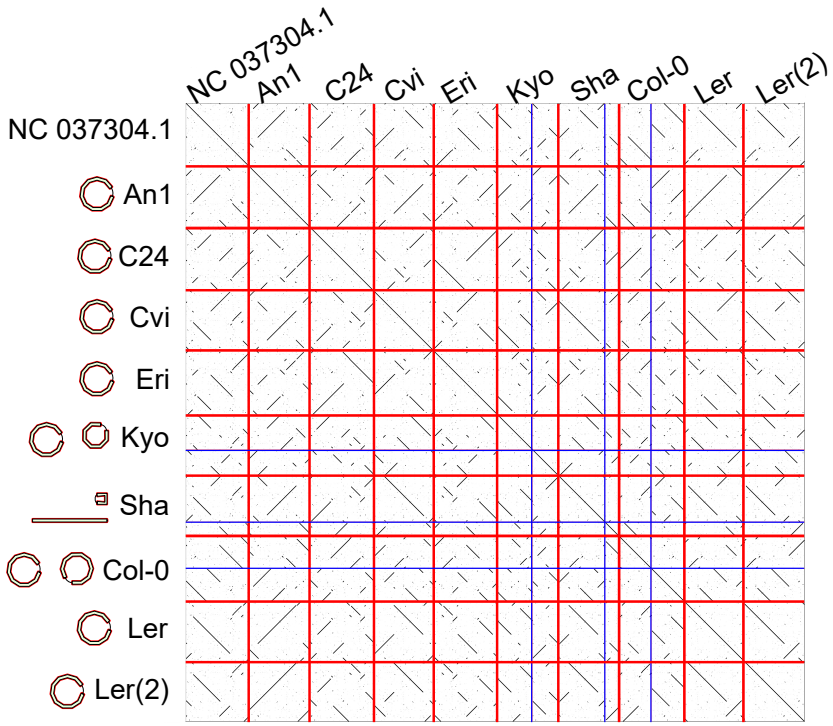

Figure 3: The dotplots between the major structures of mitochondrial genomes in eight strains. The first sequence, NC_037304.1 is the reference sequence, and the rest are assemblies produced by JTK. We assembled two L*er* strain as a biological replicates (L*er* and L*er*(2)). The thick red lines indicate the boundaries between strains, and the thin blue lines indicate the boundaries of assemblies in a strain. On the vertical axis, we put the assembly graph for each strain.

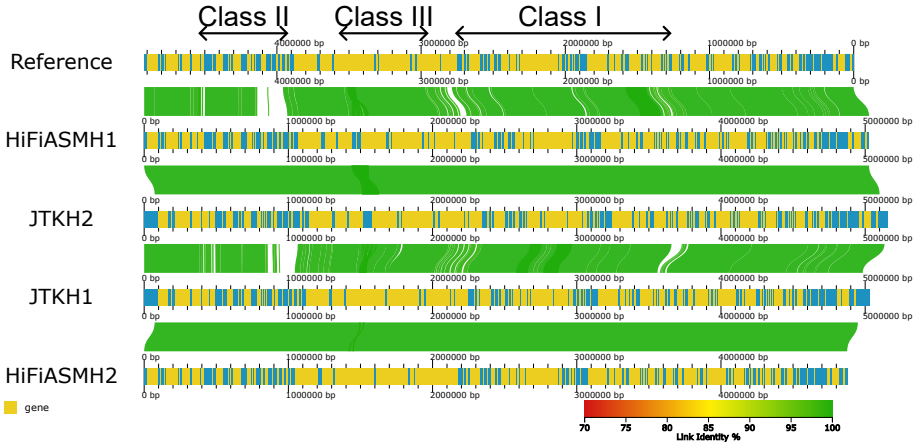

Figure 4: Assemblies on the MHC region in the Japanese sample. I compared the reference sequence, the haplotypes assembled `HiFiASM`, and the haplotypes assembled by JTK. The bars consisting of yellow bands (genes) and blue bands (intergenic regions) indicate the assemblies, and the green ribbons between assemblies are alignments.