Computational methods for resolving genomic complexity using genetic variation

Haynes Heaton

June 14, 2021

New sequencing technologies have driven biological research in the past several decades allowing for highly accurate measurements of genomes and transcriptomes. But much of the complexity of these systems is still out of our reach. Here I use properties of genetic variation in samples to uncover some of these features.

Single cell RNA-seq (scRNAseq) has revolutionized transcriptomics by enabling researchers to measure the transcriptional landscape at the cellular level instead of an average transcript level across all of the varied cells and cell types in the sample. But several artifacts make the interpretation of these experiments challenging. They contain technical artifacts that make comparing scRNAseq results across multiple experiments difficult. Also, due to the poisson loading process of the cells into the nanodroplets, as well as the random sampling of a finite barcode set, some barcodes represent multiple cells. And finally, cells may lyse in solution before the droplet partitioning is done which means that some measured RNA in cell-barcodes will come from this ambient RNA source and not the cell of interest. One solution that promises to solve multiple problems is multiplexing multiple individuals' cells into the same experiment. This then introduces a further problem—the need to demultiplex the samples from this mixture. I describe a method for using the genetic variation between multiple individuals to deconvolve the cells to their individual of origin, detect doublet cell barcodes, and measure the amount of ambient RNA in the system.

While the technology improvements and cost reductions of third generation sequencing techniques have revolutionized genome assembly, problems remain especially for certain challenging organisms such as very small organisms and highly heterozygous organisms. When assembling a genome, reads with inexact homology must be determined to have arisen from sequencing errors, paralogous repeat sequences from different regions of the genome, or from the differences between haplotypes. Until recently, relatively high DNA input requirements (5μg) have made 3rd generation long read sequencing out of reach for single individuals of small organisms. One solution to this is to sequence a mixture of individual's DNA. However, this further exacerbates the problem of heterozygosity in the assembly process as there are many more than two haplotypes. I demonstrate how a new low-input library prep can be used to create a high quality assembly of a single *Anopheles coluzzii* genome and compare it to the current PEST reference genome for the closely related *Anopheles gambiae* species.

Continuing to address the remaining problems with genome assembly in the modern era of high quality long reads, linked reads, and 3D genome distance measuring Hi-C technologies, I use phasing consistency of multiple heterozygous loci as a signal for physical linkage. I describe a suite of methods for phased genome assembly, haplotype phasing of existing assemblies, and phasing aware assembly scaffolding.