Washington University in St. Louis

Washington University Open Scholarship

Volume 13

Washington University
Undergraduate Research Digest

Spring 2018

Tenxsim: Simulator for Pure and Heterogeneous Genomic Sequence with 10X Genomics

Guanlan Dong
Washington University in St. Louis

Tenxsim: Simulator for Pure and Heterogeneous Genomic Sequence with 10X Genomics

Guanlan Dong

Mentor: Li Ding

Next-generation sequencing has become the major sequencing technology nowadays, however, limitations still exist, such as the short read length in Illumina and the high error rate in PacBio. 10X Genomics addresses both problems with its microfluidic droplet technology where an additional barcoding system is attached to Illumina reads, so that one can obtain long and accurate DNA fragments. There have been numerous simulators developed for sequencing technologies. However, these tools mainly assume a haploid reference genome and a simulator for 10X sequencing has not yet been developed. Therefore, we present tenxsim, a software written in python that can perform *in silico* 10X sequencing simulation for both pure and heterogeneous genomes. It can serve to benchmark 10X experiments and optimize relevant software. We used the experimental data from Zheng et al. (2016) as a bench mark to design the simulation process. We synthesized a random DNA genome as the reference, out of which we created two FASTA files with single nucleotide variants (SNVs). These two FASTA files represented two alleles in a diploid sample genome. From the sample genome, high molecular weight (HMW) DNA regions were generated with sizes selected from a normal distribution. Then we sonicated HMW DNA into fragments and attached a unique 16bp barcode to fragments from the same HMW DNA. Finally, fragments with attached barcodes would go through in silico Illumina sequencing and be aligned to the reference genome to create a BAM file with barcode information. Our next goal is to use the barcode information to link short reads into long DNA molecules, so that we can reconstruct sample haplotype tree and phase discovered SNVs. Furthermore, we will introduce heterogeneous cancer genome simulation to see if tenxsim can identify cancer clonality.