SEMINAR

ALL ARE WELCOME

TEMASEK LIFESCIENCES LABORATORY

12 April 2016 (Tuesday), 3pm
Seminar Rm 05-41 (Level 5)

Hosted by: Prof Laszlo Orban

Unzipping Diploid Genomes with PacBio Continuous Long Reads: From Contigs to "Haplotigs"

Dr Jason Chin
Pacific Biosciences, USA



Jason Chin joined Pacific Biosciences in 2008 starting with developing bioinformatics methods and automated analytic platform for scientists and engineers to use for improving systems' performance. Since 2013, his major work has been on developing genome assembly algorithms and softwares for sequences generated on the Pacific Biosciences' (PacBio) platform. It led to a new ground breaking method, hierarchical genome assembly process (HGAP), to effectively use PacBio to assemble genomes. In collaboration with a Google engineer, he had assembled the first human genome from 54X coverage of PacBio data in 2014, which produced a highly contiguous assembly exceeding all the previous human WGS assembly attempts in the post sequencing era. Since then, Jason has focused on (1) improving assembly computational performance, (2) developing algorithms for assembling diploid genome and (3) understanding the fundamental mathematical properties in the genome assembly problem.

It is often conveniently ignored that with only few exceptions on simplest cases like haploid, bacteria, or inbreds, genomic information is never carried in a single reference per individual manner, but rather has higher ploidy ($n \geq 2$) for almost all organisms. So a more complete view of both haplotypes can help to generate better annotations, provide useful biological insights, including information about all kinds of heterozygous variants, and be used for studying differential allele expression in organisms. Unfortunately, the current dominant sequencing technologies as well as assembler algorithms lack the capability to address this bottleneck. In alignment with the raising awareness of the advantage of long read sequencing, we have developed a new software - FALCON for de novo haplotype reconstructions which allows polyploidy aware assembly (https://github.com/PacificBiosciences/FALCON). We applied the algorithms and the prototype software for (1) a highly repetitive diploid fungal genome (Clavicorona pyxidata) and (2) a F1 hybrid from two inbred Arabidopsis strains: Cvi-0 and Col-0. For the fungal genome, we achieve an N50 of 1.53 Mb (of the 1n assembly contigs) of the ~42 Mb 1n genome and an N50 of the haplotigs of 872 kb from a 95x read length N50 ~16 kb dataset. We find that ~45% of the genome is highly heterozygous and ~55% of the genome is highly homozygous. For the Arabidopsis F1, we achieved contig N50=7.96Mb and haplotig N50=6.92Mb (equivalent to haplotype block N50=6.92Mb). Effectively, we have reconstructed the two parental genomes using the F1 sequencing data. Finally, we apply FALCON/FALCON-Unzip on three other organisms (grape, zebra finch & human) to demonstrate this is a general approach to solve diploid assembly problem with PacBio long read data.

Recent Publications:

1. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A, Dai H, Fritz MH, Cao H, Cohain A, Deikus G, Durrett RE, Blanchard SC, Altman R, **Chin CS**, Guo Y, Paxinos EE, Korbel JO, Darnell RB, McCombie WR, Kwok PY, Mason CE, Schadt EE, Bashir A. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 12:780-6.
2. Berlin K, Koren S, **Chin CS**, Drake JP, Landolin JM, Phillippy AM. 2015 Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 33:623-30.
3. **Chin CS**, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013 Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 10(6):563-9.