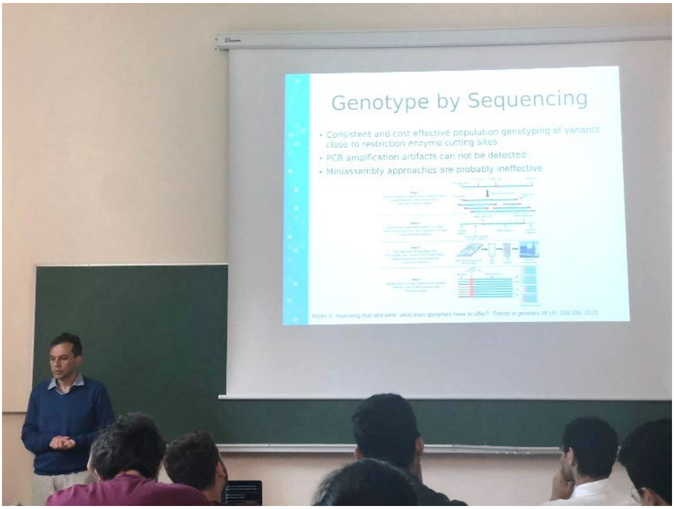
July 18th, 2019

Software development for comparative and population genomics applied to Lima bean

Jorge Duitama
UNIANDES, Colombia

Abstract

Recent developments on short and long read high throughput sequencing technologies have enabled high quality genome assemblies and development of large databases of inter and intraspecies genomic variability for an unprecedented number of species. In this talk I present the latest algorithms and functionalities implemented in our software solution NGSEP for analysis of different kinds of genomic datasets. This includes the latest haplotype clustering algorithms to perform variants detection and genotyping from aligned reads, achieving comparable accuracy and better efficiency compared to widely used software tools such as the GATK haplotype caller and Strelka2. It also includes our new solution for whole genome alignment through efficient identification of synteny blocks built from large chains of orthologous genes. Our algorithm performs k-mer searches on FM- indexes built from the proteomes of annotated genomes to efficiently identify paralogs and orthologs. Benchmark experiments against commonly used tools for ortholog identification and synteny analysis show that construction of ortholog chains enables alignments between chromosomes of large genomes within minutes of computation. Using state-of-the-art data visualization technologies, we provide novel interactive views of the alignments provided by our software.



Genotype by Sequencing

- Consistent and cost effective population genotyping of variants close to restriction enzyme cutting sites
- PCR amplification artifacts can not be detected
- Miniassembly approaches are probably ineffective

Myles S. Improving fruit and wine: what does genomics have to offer?. Trends in genetics 29 (4): 190-196. 2013

The application of these developments is illustrated through the results of a large sequencing effort to provide a chromosome level assembly of the Lima bean genome. Lima bean (*Phaseolus lunatus L.*) is the second most important Phaseolus crop for human consumption and, compared to common bean, it shows a wider range of ecological adaptations along its range of distribution from Mexico to Argentina. These adaptations plus its phenotypic plasticity make Lima bean a promising crop for food security under predicted scenarios of climate change in Latin America. Combining PacBio long reads with Illumina short reads following the paired-end, 10x, Genotype-By-Sequencing (GBS) and RNA-seq protocols, we achieved a chromosome level assembly of 516 Mbp with 43,997 annotated gene models. Population genomic analysis of GBS data for over 500 Mesoamerican and Andean accessions confirmed the presence of four wild gene pools and two domestication events that gave rise to Mesoamerican and Andean landraces and provided evidence for a new center of diversity. Structural comparison of the lima bean genome with that of common bean revealed extensive synteny between both species and two large structural rearrangements in chromosomes Pl02 and Pl10. Analysis of

RNA-seq data obtained from wild and cultivated accessions at two different developmental stages revealed 1,887 transcripts differentially expressed either between stages or between wild and cultivated accessions that could be related to the pod dehiscence domestication trait.

Short bio

Jorge Duitama is a software engineer from Universidad de los Andes with over 12 years of experience in bioinformatics. Jorge finished his Ph.D. at the computer science department of University of Connecticut working on prediction of neo-epitopes for cancer immunotherapy. He has worked as a postdoc at the lab of Kevin Verstrepen in KU Leuven and as a researcher in bioinformatics at the International Center for Tropical Agriculture (CIAT). Currently Jorge is assistant professor at the systems and computing engineering department of Universidad de los Andes. Through these works, Jorge has contributed different open source bioinformatic tools for different genomic analyses, including variants detection and genotyping, molecular haplotyping, prediction of epitopes for cancer immunotherapy, virus subtype identification and QTL mapping in pools of segregants. Jorge also collaborated in different projects involving analysis of large genomic datasets for yeast, rice, beans, sugar cane, humans and several other species.