International Coffee Genomics Network (ICGN)
Report Coffee Genomics Workshop

Plant and Animal Genome (PAG-30) Meeting,
San Diego, California, January 13-18, 2023
https://plan.core-apps.com/pag_2023/event/8cd42c3e9afb9e0055487a3968788052

Abstracts

How Haplotype-Resolved Assemblies for the Allotetraploid *Coffea arabica* Can Accelerate Genomics Assisted Breeding for Climate Change Adaptation

Marcela Yepes[1], **Aleksey Zimin**[2,3], Carlos Ernesto Maldonado[4], Carmenza E. Góngora[4], Claudia Flórez[4], Alvaro Gaitán[4] and Herb Aldwinckle[1]

[(1)]Cornell University/ School of Integrative Plant Sciences/ Plant Pathology and Plant Microbe Biology Section, Geneva, NY, [(2)] University of Maryland, College Park, MD, [(3)] Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [(4)]FNC/CENICAFE Colombian National Coffee Research Center, Chinchiná, Colombia

Abstract Text:

Coffea arabica is a commercially very important perennial polyploid species, representing 60-70% of the world's coffee market, with underdeveloped genomic resources yet highly appreciated by coffee consumers for its superior quality and health benefits. To streamline coffee genome analysis and breeding efforts for climate change adaptation, we sequenced, assembled, and chromosome scaffolded the genome of a highly homozygous di-haploid *C. arabica* genotype, and the genome of a highly heterozygous tetraploid *C. arabica* genotype to generate the first haplotype-resolved assemblies for this important species. For this effort, we collaborated with Dovetail Genomics (now Cantata Bio), and used their newest workflow which includes PacBio HiFi long reads assembled using HiFiasm1 and Omni-C® long range proximity ligation data. The resulting haplotype-resolved assemblies were scaffolded to chromosome-scale using Dovetail's proprietary HiRise software.

Previous *C. arabica* genome assemblies built by our group and others are mosaics of the parental genomes which greatly hindered our ability to properly identify important variants within the genome. Mis-assignment of contigs to the *C. eugenioides* and *C. canephora* parental sub-genomes was due to the high homology (>95%) of the two sub-genomes. Since the non-haplotype resolved reference was not representative of the actual variation present in the species and was mosaic in nature, it was also difficult to relate to the genetic mapping data previously obtained by our team and others. Omni-C®/HiRise scaffolding allowed us to scaffold both a diploid-aware genome for a target di-haploid, highly homozygous *C. arabica* genotype, and a tetraploid-aware genome for a target highly heterozygous *C. arabica* allotetraploid genotype. The ability to separate the phases

into separate pseudomolecules for the di-haploid and allotetraploid *C. arabica* genome assemblies is allowing us to unveil for the first time the extent of variation between both sub-genomes as well as between homeologs. So far, *Coffea* diversity has not been effectively characterized at the genome sequence level, nor widely utilized in cultivated coffee varieties, particularly for *C. arabica* which suffers from very limited diversity due to population bottle necks. However, haplotype-resolved assemblies like those generated here should accelerate diversity characterization, further enhancing its utilization for climate change adaptation of this important endangered species. We continue to search for innovative solutions to adapt the crop to climate change to help coffee farmers achieve sustainable production, and ensure the sustainability of high-quality *C. arabica* coffee production for the world market.

Note: This presentation will have an extended time and will be presented by co-authors M. Yepes (introduction), C. Maldonado (di-haploid genome assembly), and A. Zimin (new tools for genome assembly, phasing, and annotation of polyploid genomes).

Building Haplotype-Resolved Assemblies of Diploid and Polyploid Plant Genomes

Mark Daly

Dovetail Genomics, now CantataBio,
North Vancouver, BC, Canada

Abstract Text:

Traditional *de novo* genome assemblies are haploid; that is, a mosaic of the two parental haplotypes. Here we review our latest assembly workflow that delivers two distinct assemblies for a diploid organism - one for each haplotype. A unique combination of PacBio HiFi and Dovetail® Omni-C® proximity ligation enables highly accurate and contiguous assemblies scaffolded to chromosome scale. We will also discuss how this workflow has been applied to polyploid genome assemblies, and provide case studies of genome assemblies of varying size and heterozygosity levels.

De novo genome assemblies have traditionally been pseudo-haploid in nature. Newer, more accurate long read sequencing coupled with unbiased, restriction-enzyme-free proximity ligation technology is enabling high-quality haplotype-phased genome assemblies from a single individual. Phased haplotype blocks are now of chromosome size and are beginning to uncover the true structure of plant and animal diploid genomes. At Dovetail, we are continually looking for ways to enrich our datasets to provide our customers with genome assemblies that will stand the test of time. Our newest workflow includes PacBio HiFi long reads assembled using HiFiasm1 and Omni-C® long range proximity ligation data. The resulting two haplotype-resolved assemblies are then scaffolded to chromosome-scale using Dovetail's proprietary HiRise software. Haplotype-resolved assemblies offer many advantages for genomic-based studies in evolution, conservation,

agricultural biology, and human disease. We launched our haplotype-resolved genome assembly service for diploid organisms in late 2021. Building on the great reception to our diploid service, we are now pleased to offer a similar service for polyploids. Polyploids are the most challenging genomes to assemble. At Dovetail Genomics®, we use a unique combination of PacBio HiFi long reads and Dovetail® Omni-C® proximity ligation, plus a proprietary and bespoke bioinformatics workflow, to build high quality polyploid assemblies from scratch.

Highly Contiguous, Chromosome Scale Scaffolded Genome Assembly of the Coffee Leaf Rust (*Hemileia vastatrix*), One of the Largest Fungal Pathogen Genomes Known so Far

Carlos Ernesto Maldonado, FNC/CENICAFE Colombian National Coffee Research Center, Chinchiná, Caldas, Colombia, Gustavo Adolfo Marin, FNC/CENICAFE Colombian National Coffee Research Center, Chinchiná, Caldas, Colombia and Carlos Ariel Angel, PhD, Manizales, Caldas, Colombia

Abstract Text:

Coffee leaf rust (CLR), caused by *Hemileia vastatrix*, is the most important coffee disease worldwide, generating recurrent and devastating epidemics with yield losses up to 80%. This pathogen belongs to the order Pucciniales, class Basidiomycetes, that contains the causal agents of rust, a group of biothrophic fungal pathogens that cause destructive plant diseases in many crops. The life cycle of CLR occurs only on plants of the genus *Coffea*, and as an obligate parasitic fungus, it is not possible to culture a pure isolate on artificial growth media. A CLR isolate from *Coffea arabica* var. Caturra was increased by several cycles of inoculation on the same host to obtain enough urediniospores for high molecular weight DNA isolation and PacBio HiFi genome sequencing combined with Dovetail® Omni-C® proximity ligation and chromosome-scale scaffolding using Dovetail's proprietary HiRise pipeline. A highly contiguous, chromosome scaffolded genome assembly was obtained with L75 in 15 scaffolds, a genome size of 773 Mb, and transposable element (TE) content higher than 80%. This is the fourth largest genome among plant pathogenic fungi sequenced so far, as Pucciniales represent a group where genome size expansion appears to be a common characteristic. The isolate's race was characterized phenotypically according to virulence reactions on differential coffee clones, a series of plant genotypes carrying different rust resistance genes (*SH*) developed by the Coffee Leaf Rust Research Center at Oeiras, Portugal (Centro de Investigacao das Ferrugens do Cafeeiro, CIFC). Overall, our high quality, chromosome scaffolded genome assembly for coffee leaf rust constitutes a very important robust resource to study fungal diversity, detect the emergence of more aggressive races, study molecular mechanisms of host-pathogen interaction, and support coffee breeding efforts towards the development of more durable genetic resistant varieties in our efforts to adapt coffee to climate change.

Flavor to Cup (F2C): A Genetical Roadmap to Improved Coffee Flavor

Luis Felipe V. Ferrão,
Horticultural Sciences Department, University of Florida, Gainesville, FL

Abstract Text:

Coffee is a widely consumed beverage that drives a vibrant industry. Better-tasting is arguably one of the most important coffee attributes, with many consumers willing to pay for premium products. Therefore, the flavor phenotype is a key target for coffee breeding. Despite its significance, making headway in this field is not simple, and the slow progress has been largely due to the complexity of measuring the "flavor phenotype". In this scenario, modern breeding programs have paired metabolomics, genomics, and sensory information to predict consumer preference and guide marker-assisted selection for flavor. While examples of success have been recently reported in tomato and blueberry, coffee breeding studies integrating multi-omics data are still in their infancy. Bringing together the expertise of key international coffee institutions, the F2C is a project lead by the University of Florida with the specific goal of providing a genetic and metabolic roadmap for coffee flavor improvements. The research project was conceptualized in three main stages. First, we will genotype by sequencing a diversity panel containing 384 materials from *Coffea canephora* and *C. arabica*. Next, flavor scores will be collected through sensory panel analyses assessed by certified Q graders. Then, at the metabolomics level, we will use gas chromatography–mass spectrometry (GC–MS) to identify key volatile organic compounds associated to coffee aroma. Finally, we propose to map metabolites modulating consumer preference, and the genes underlying the variability of flavor-related metabolites, thereby providing tools for molecular breeding in coffee.

Chromosome-Level *De Novo* Genome Assembly and Comparative Genome Analysis of the Coffee Bee Hawk Moth, *Cephonodes Hylas*

Takahiro Yamabe,
School of Life Science and Technology, Tokyo Institute of Technology, Tokyo, Japan

Abstract Text:

Cephonodes hylas, known as the coffee bee hawkmoth, pellucid hawk moth, or coffee clearwing is a hawkmoth of the subfamily Macroglossinae that has a large distribution in the warmer regions of the Middle East, Asia, and Africa. The adults have transparent wings, and the caterpillars feed mainly on a wide range of Rubiaceae plants, including but not limited to *Adina* sp., *Guettarda*, *Pavetta*, *Haldina*, *Mitragyna*, *Mussaenda*, *Morelia*, *Canthium*, *Randia*, *Catunaregam*, *Wendlandia*, *Ixora*, *Kraussia*, and also a few host plants of agricultural value such as *Gardenia*, and *Coffea* (coffee).

C. hylas has evolved and adapted with unique characteristics, such as larvae feeding on gardenia, overcoming the toxicity of its iridoid glycosides, diurnal adults, and transparent wings. Although

C. hylas is a fascinating model for molecular biological research, genome sequence analysis-based genetic approaches to elucidate these peculiarities have not yet been undertaken. We successfully achieved *de novo* genome assembly at the chromosome level of *C. hylas* comparable to the Lepidoptera model organism, silkworm. Additionally, 16,854 protein-coding genes were annotated, and the constructed genome sequence and annotated genes were of the highest quality BUSCO completion compared to closely related species. Comparative genome analysis revealed the process of chromosomal evolution from the Bombycoidea ancestral (n = 31) genome and changes in turnover at the chromosome level associated with chromosomal fusion events, such as the rate of repetitive sequence insertion. These analyses were only possible because the genome was constructed at the chromosome level. Additionally, increased dN/dS ratios were observed in multiple photoreceptor-related genes that were strongly associated with the acquisition of diurnal activity. Furthermore, tandemly duplicated expanded genes containing many digestive and other enzymes and larval midgut-specific expression were also confirmed. These genes may be involved in the metabolism of genipin, a toxin found in gardenias. The genome sequence of *C. hylas* determined at the chromosomal level is expected to be an important resource for the molecular elucidation of the unique phenotype and chromosomal evolution of not only *Sphingidae* species but also of the entire Lepidoptera order.