# International Coffee Genomics Network (ICGN)
# Report Coffee Genomics Workshop

XXIX Plant and Animal Genome Meeting,
San Diego, California, January 8-12, 2022
https://pag.confex.com/pag/xxix/meetingapp.cgi/Session/7565

**Note due to the pandemic the 2021 PAG Meeting was cancelled, and the 2022 PAG Meeting and Coffee Genomics Workshop were held virtually**

## Abstracts

### Bringing Coffee (*Coffea arabica*) to the Forefront of Plant Genomics Research for Polyploid Species

**Marcela Yepes[1], Aleksey Zimin[2,3], Carlos Ernesto Maldonado[4],**
Carmenza E. Góngora[4], Claudia Flórez[4], Alvaro Gaitán[4] and Herb Aldwinckle[5]

[(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)]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, [(4)]FNC/CENICAFE Colombian National Coffee Research Center, Chinchiná, Colombia, [(5)]Cornell University/ School of Integrative Plant Sciences/ Plant Pathology and Plant Microbe Biology Section, Geneva, NY

**Abstract Text:**

*Coffea arabica* is a genetically complex, economically important, perennial, polyploid species, comprising 60-70% of the world's coffee market, and highly appreciated by coffee consumers for its superior quality and health benefits. However, its genomic resources are as yet grossly underdeveloped. To streamline genome analysis and breeding for climate change adaptation of *C. arabica*, we sequenced, assembled, and chromosome-scaffolded the genome of the allotetraploid variety Caturra, as well as the genome of its diploid maternal ancestor *C. eugenioides*, to generate high-quality reference genome assemblies. Our first genome assembly of the allotetraploid *C. arabica* Caturra represents the most contiguous publicly available genome assembled for this economically most important species, with contig N50 ~4 Mb, and scaffold N50 ~42.36 Mb with 90% of the genome assigned to chromosomes (22 pseudo-chromosomes split by sub- species) (GenBank assembly accession: GCA_003713225.1). Our high-quality *C. eugenioides* chromosome-scaffolded reference genome assembly is the first publicly available for the maternal diploid ancestor of the allotetraploid *C. arabica* (accession number GCA_003713205.1). Both genome assemblies were fully annotated using NCBI's automated Eukaryotic Annotation pipeline revealing 44,482 protein coding genes for *C. arabica*, and 29,100 protein-coding genes for *C. eugenioides.* Protein alignments with other Rubiaceae illustrate the very limited publicly available information for this large Angiosperm family.

Validation of our first *C. arabica* reference genome assembly using high density genetic mapping and skim sequencing data on an $F_2$ mapping population, revealed areas of mosaic assembly due to the high similarity of the two parental sub-genomes. We are in the process of generating and validating new genome assemblies with improved contiguity and accurate haplotype resolution and phasing. In addition, we have generated large RNASeq data sets using long read IsoSeq and/or short read Illumina for a wider range of tissues/organs /developmental stages/biotic and abiotic stress treatments for climate change adaptation for both species in order to improve genome annotation. Contiguity of our new genome assemblies for both *C. arabica* and *C. eugenioides* has been optimized using SAMBA, a novel tool developed by our group for scaffolding assemblies with multiple big alignments (Zimin and Salzberg, 2021). In addition, we used StringTie2 (Kovaka *et al*. 2019), a reference guided transcript assembly developed by our group capable of assembling both short and long RNASeq reads, as well as full- length super-reads, and an optimized annotation pipeline that significantly increases isoform and gene discovery for both target species. Our long-term goal is to support a more sustainable and resilient future for producers and the coffee sector overall. In order to shape the industry and coffee production for the 21$^{st}$ century and beyond, we continue to be committed to using the best possible technologies available to move this orphan crop to the forefront of plant genomics research for polyploid species.

**Note: This presentation had an extended time and was presented by coauthors M. Yepes (introduction), A. Zimin (New tools for genome assembly and annotation), C. Maldonado (reference genome assembly validation).**

---

## Exploring the Genetic Diversity of the Genus *Coffea*:

## A Quest for Genes of Agronomic Importance

**Carlos Ernesto Maldonado[1]**, Carlos Ariel Angel[1], Alvaro Gaitán[1], Aleksey Zimin[2,3], Marcela Yepes[4] and Herb Aldwinckle[4]

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

**Abstract Text:**

The generation of a high-quality chromosome scaffolded reference genome assembly for *Coffea arabica* variety Caturra by our team has provided a solid foundation to explore the genetic diversity of the *Coffea* genus, allowing us to integrate phenotypic and genetic data to identify genomic regions and genes associated with important agronomic traits such as disease tolerance. Our highly curated chromosome- scale genome assembly for *C. arabica*, the only allotetraploid and most widely cultivated species of the genus, was separated into subgenomes using the genome

assemblies of its ancestral diploid species*: C. eugenioides*, maternal ancestor, genome assembly generated *de novo* by our group, and *C. canephora*, paternal ancestor, genome assembly previously generated by Denoeud *et al*. 2014. Coffee leaf rust (CLR) caused by *Hemileia vastatrix* and coffee berry disease (CBD) caused by *Colletotrichum kahawae* are the most devastating diseases for the crop, multiple disease resistance genes for those diseases have been found within the genus *Coffea* and a portion of those have been mapped. Using a combined strategy of *de novo* assembly and reference guided scaffolding, the draft genomes of *C. liberica* and Timor Hybrid were obtained. By integration of genetic maps and genomic sequences the regions containing the genes *SH*3 and *Ck*-1, conferring resistance to CLR and CBD, respectively were found. Resequencing of 13 *C. arabica* accessions carrying CLR resistance genes *SH*1, *SH*2, and *SH*4 allowed the identification of allelic forms of candidates for these disease resistance genes. The exploration of the genetic diversity of *C. arabica* in the Colombian *Coffea* germplasm collection at FNC/CENICAFE and its integration with phenotypic data is in progress and will provide tools for breeders with the goal of developing varieties with new traits of tolerance to pests and diseases, beverage quality and high yield, also expressing climate change resilience.

---

## Updates on Pathogenic, Genetic and Genomic Variation of Coffee Leaf Rust (*Hemileia vastatrix* Berk. & Br.) in Colombia

**Carlos Ariel Angel**, FNC/CENICAFE Colombian National Coffee Research Center, Chinchiná, Caldas, Colombia, Lina Maria Del Mar Escobar, Colombian Ministry of Science, Colombia, Carlos Ernesto Maldonado, FNC/CENICAFE Colombian National Coffee Research Center, Chinchiná, Colombia and Gustavo Adolfo Marin, FNC/CENICAFE Colombian National Research Center, Chinchina, Caldas, Colombia

**Abstract Text:**

Coffee leaf rust (CLR) is the most important coffee disease worldwide, generating recurrent and devastating epidemics with yield losses up to 80%. FNC– Cenicafé's (Colombia) main management strategy is breeding composite resistant varieties based on *C. arabica* cv. Caturra (susceptible) x Timor Hybrid (HdT, resistant), reaching 85% out of 850,000 ha coffee growing area. This achievement is a world particular scenario that has allowed Colombia to use composite resistance varieties and maintain field resistance to coffee leaf rust for over 40 yrs, but also encourages co-evolution and greater pressure on the pathogen to develop new races and complex variants. At Cenicafé, we are studying changes in pathogenicity and their relationship with CLR epidemics, from populations to single pustule isolates, surveying genomic diversity. Aggressive populations on individual coffee lines and varieties are inoculated on 106 coffee genotypes including CIFC differentials (*SH*1 to 9?) to characterize physiological races, pathotypes, and virulence genotypes. To assess diversity after screening 752,000 SSR markers, approximately 2,500 high quality were selected and eight used on 89 CLR single pustule isolates from contrasting

coffee genotypes, and 33 inoculated on CIFC differentials. To study the structure of an individual population, 37 single pustule isolates collected on a Caturra x HdT advanced line from Castillo® varieties were characterized for incomplete resistance variables on parental genotypes, with low coverage genome sequencing, and three by CIFC differentials. Differences were detected between isolates from the same leaf, same plant or several plants of the same line, and from two same location fields. A high-quality reference genome assembly was obtained sequencing a characterized CLR population Race I (*v*2,5), assembling 707 - 718 Mbp of the estimated genome size, at 70X coverage, with 96% BUSCO completeness, and genome annotation.

---

## Building a Chromosome-Scale Haplotype-Resolved Assembly with Omni-C® Data

**Mark Daly**

Dovetail Genomics, Scotts Valley, CA

**Abstract Text:**

The tools available for capturing genomic information have evolved dramatically over the past decade, yet reference genomes remain haploid. Through advances in Dovetail® proximity ligation (Hi-C) technology, the capture of genetic variation along with ultra-long-range genomic sequence information in a single assay is now possible. A notable benefit to this enhanced data type is the ability to phase SNPs over extremely large distances. The newest Dovetail® *de novo* assembly workflow utilizes a unique combination of: PacBio HiFi sequencing for best-in- class base calling accuracy, and Dovetail® Omni-C® scaffolding for unprecedented SNP coverage and long-range information. Unlike traditional Hi-C approaches that digest chromatin with sequence-specific and biased restriction enzymes, Dovetail® Omni-C® technology utilizes *DNase* I to randomly fragment chromatin. This generates unbiased coverage of the genome, with no blind spots, enabling SNP detection rates approaching those achieved with standard shot-gun libraries. As a result, assemblies produced from this data display extremely long haplotype blocks when phased, up to full chromosome-length. The benefit of a chromosome-scale reference genome with phased SNPs is to gain a better understanding of: hybridization, distribution of heterozygosity, inbreeding, allele- specific epigenetic events, *cis* versus *trans* mutations, etc. We will discuss advantages of this application for the coffee community, in particular for *Coffea arabica*.