International Coffee Genomics Network (ICGN)
Report Coffee Genomics Workshop

Plant and Animal Genome (PAG-30) Meeting,
San Diego, California, January 12-16, 2019
https://plan.core-apps.com/pag_2019/event/9441a7255c56cf0ce04b90bfc40d725f

Abstracts

Release of the *Coffea arabica* Variety Caturra Genome and That of Its Maternal Diploid Ancestor *C. eugenioides* to Provide a Strong Foundation for Breeding and Functional Genomics Studies in Coffee

Alvaro Gaitán[1], Marcela Yepes[2], Aleksey Zimin[3,4], Carlos Ernesto Maldonado[1], Lucio Navarro[1], Claudia Flórez[1], Carmenza E. Góngora[1], Pilar Moncada[1], James Yorke[4] and Herb Aldwinckle[2]

[(1)]Centro Nacional de Investigaciones de Café, CENICAFE, Chinchiná, Colombia, [(2)]Cornell University/ School of Integrative Plant Sciences/ Plant Pathology and Plant Microbe Biology Section, Geneva, NY, [(3)]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, [(4)]University of Maryland, College Park, MD

Abstract Text:

The world's most widely cultivated coffee species, representing 70% of the global market, is the allotetraploid, *Coffea arabica* (2n=4x=44; genome size ~1.1 Gb). *C. arabica* evolved through the interspecific hybridization of the ancestors of two diploid *Coffea* species: *C. eugenioides* (2n=2x=22, maternal donor, genome size ~0.67 Gb) and *C. canephora* (2n=2x=22, paternal donor, genome size ~0.71 Gb previously sequenced, Denoeud *et al.* 2014).

Due to extreme bottlenecks created by human migration in the 6th and 17th centuries, throughout the world cultivated *C. arabica* varieties benefit very little from genetic diversity. The very narrow gene pool keeps cultivated *C. arabica* constantly vulnerable to diseases and insect pests [coffee berry disease (*Colletotrichum kahawae),* coffee leaf rust (*Hemileia vastatrix),* wilt *(Gibberella xylarioides,* anamorph *Fusarium xylarioides),* coffee berry borer (*Hypothenemus hampei*), leaf miner (*Leucoptera coffeella*), stem borer (*Xylotrechus quadripex*), and nematodes (*Meloidogyne, Pratylenchus* sps.)]. These vulnerabilities are exacerbated in the context of environmental stress due to climate change: excess rain, drought, and increased temperature.

To stream-line coffee genome analysis and breeding efforts for climate change adaptation, we sequenced, assembled, and chromosome-scaffolded the genome of allotetraploid *C. arabica* variety Caturra, as well as the genome of its diploid maternal ancestor *C. eugenioides*, to generate high- quality reference genome assemblies. To celebrate the public release of our coffee reference genomes, we have several presentations describing the importance of the target genotypes to the coffee community, the status of their release, as well as our progress on gene annotation.

This work was co-funded by the US National Science Foundation (NSF Award#1444893), the InterAmerican Development Bank through FONTAGRO, the Colombian National Coffee Growers Federation (FNC) and its National Coffee Research Center CENICAFE.

Coffea arabica variety Caturra and *C. eugenioides* Genome Assembly and Annotation

Aleksey Zimin[1,2], Marcela Yepes[3], Carlos Ernesto Maldonado[4], Lucio Navarro[4], Sam Kovaka[2], Mihaela Pertea[5], Claudia Flórez[4], Carmenza E. Góngora[4], Pilar Moncada[4], James Yorke[1], Alvaro Gaitán[4] and Herb Aldwinckle[3]

[(1)]University of Maryland, College Park, MD, [(2)]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, [(3)]Cornell University/ School of Integrative Plant Sciences/ Plant Pathology and Plant Microbe Biology Section, Geneva, NY, [(4)]Centro Nacional de Investigaciones de Café, CENICAFE, Chinchiná, Colombia, [(5)]McKusick-Nathans Institute of Genetic Medicine, JHU, Baltimore, MD

Abstract Text:

We sequenced, assembled, and chromosome scaffolded the genome of the allotetraploid *Coffea arabica* variety Caturra, as well as the genome of its diploid maternal ancestor *C. eugenioides*, to generate high-quality reference genome assemblies for these two species that have now been deposited to NCBI Genbank under:

Coffea arabica: biosample/ genome assembly and annotation

https://www.ncbi.nlm.nih.gov/biosample/SAMN10272287/

https://www.ncbi.nlm.nih.gov/assembly/GCF_003713225.1/

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Coffea_arabica/100/

***C. eugenioides:* Biosample, genome assembly and annotation**

https://www.ncbi.nlm.nih.gov/biosample/SAMN10269643/

https://www.ncbi.nlm.nih.gov/assembly/GCF_003713205.1/

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Coffea_eugenioides/100/

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 genomes were fully annotated using NCBI's automated

Eukaryotic Annotation pipeline and publicly available transcriptomics data 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.

For the genome assemblies, we assembled the data with Falcon and MaSuRCA assemblers and combined the two assemblies to produce a highly contiguous *C. arabica* (contig N50 ~4 Mb) genome, which is a substantial improvement compared to other on-going efforts co-funded by private companies and their consortia. Assembled contigs were mid-range scaffolded using 10X Genomics data with fragscaff software, and long-range scaffolded into super-scaffolds using Hi-C data and HiRise scaffolder by Dovetail Genomics. Finally, the super-scaffolds were placed onto the chromosomes with the help of genetic and physical maps for the target genotypes (Moncada *et al*. 2009, 2016; López *et al*. 2013) and existing published pseudo-chromosomes of *C. canephora* (Denoeud *et al.* 2014).

To improve the quality of the genome annotation, we have also generated high coverage PacBio long-reads, error corrected with high coverage Illumina paired end reads.. We are currently using several RNASeq datasets generated by our group with Illumina and 454 to re-annotate our *C. arabica* and *C. eugenioides* genomes, as well as, PACBio IsoSeq data to generate more robust gene models. Genome guided transcript assembly (using StringTie v. 2) to integrate short read Illumina RNAseq and long read PacBio Iso-Seq data is being used to yield a near complete gene set (protein-coding genes), and isoforms covering 94% complete single-copy plantae BUSCOs. Assembled transcripts along with trained gene prediction methods within MAKER-P are being used to create our genome annotation. We will discuss tools that we used to create our assemblies and annotation.

This work is co-funded by the US National Science Foundation (NSF Award Award#1444893), the InterAmerican Development Bank through FONTAGRO, the Colombian National Coffee Growers Federation (FNC) and its National Coffee Research Center CENICAFE.

Adaptive Horizontal Transfer of a Bacterial Gene to an Invasive Insect Pest of Coffee

Ricardo Acuña[1], Beatriz Padilla[1], Claudia Flórez[1], Jose Rubio[1], Juan Carlos Herrera[1], Pablo Benavides[1], Jeffrey J. Doyle[2] and Jocelyn KC Rose[2]

[(1)]Centro Nacional de Investigaciones de Café, CENICAFE, Chinchiná, Colombia, [(2)]Plant Biology Department, Cornell University, Ithaca, NY

Abstract Text:

Horizontal gene transfer (HGT) involves the nonsexual transmission of genetic material across species boundaries. Although often detected in prokaryotes, examples of HGT involving animals

are relatively rare, and any evolutionary advantage conferred to the recipient is typically obscure. We identified a gene (HhMAN1) from the coffee berry borer beetle, *Hypothenemus hampei*, a devastating insect pest of coffee, which shows clear evidence of HGT from bacteria. HhMAN1 encodes a mannanase, representing a class of glycosyl hydrolases that has not previously been reported in insects. Recombinant HhMAN1 protein hydrolyzes coffee berry galactomannan, the major storage polysaccharide in this species and the presumed food of *H. hampei*. HhMAN1 was found to be widespread in a broad biogeographic survey of *H. hampei* accessions, indicating that the HGT event occurred before radiation of the insect from West Africa to Asia and South America. However, the gene was not detected in the closely related species *H. obscurus* (the tropical nut borer or "false berry borer"), which does not colonize coffee beans. Thus, HGT of HhMAN1 from bacteria represents a likely adaptation to a specific ecological niche.

Identification of Dof1 Transcription Factor in Coffee (*Coffea arabica* L.) and Its Expression in Response to 2-Oxoglutarate

Ricardo Acuña[1], Jefferson Medina[1], Monica Quintero[1], Carlos Ernesto Maldonado[1], Marcela Yepes[2], Herb Aldwinckle[2] and Alvaro Gaitán[1]

[(1)]Centro Nacional de Investigaciones de Café, CENICAFE, Chinchiná, Colombia, [(2)]Cornell University/ School of Integrative Plant Sciences/ Plant Pathology and Plant Microbe Biology Section, Geneva, NY

Abstract Text:

Coffea arabica L. is a worldwide economic crop and nitrogen is one of the most important mineral elements for its growth and production. 2-oxoglutarate (2-OG) is an important regulator of carbon and nitrogen metabolism in higher plants. Feeding experiments were designed to investigate the role of 2-OG in regulation of transcription and DNA binding with the one zinc finger (Dof1) transcription factors (TFs) involved in nitrogen metabolism. These TFs participate widely in plant responses to nitrogen assimilation, but there are no reports of their activity in coffee. Using bioinformatics tools a complete sequence for *Coffea arabica* transcription factor *Cardof1* was obtained. Quantitative real- time polymerase chain reaction (qRT-PCR) analysis showed that expression levels of *CarDof1* were higher in coffee roots than in leaves. This work lays the foundation for further analysis of the function of *CarDof1* in *Coffea arabica,* which will be helpful for improving the metabolism of nitrogen assimilation in coffee.

Developing a US STEM Workforce That Is Globally Competitive

Stephanie Fuchs[2], **Matthieu Fuchs**[2], **Marcela Yepes**[1], Aleksey Zimin[3,4], James Yorke[3] and Herb Aldwinckle[1]

[(1)]Cornell University/ School of Integrative Plant Sciences/ Plant Pathology and Plant Microbe Biology Section, Geneva, NY, [(2)]Cornell University/ Biological Engineering, Ithaca, NY, [(3)]University of Maryland, College Park, MD, [(4)]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD

Abstract Text:

As part of our coffee genomics NSF funded project (NSF Award Award#1444893), we organize every year a STEM career exploration workshop. The participants are rising undergraduates and High School minority students as well as Biology/Science instructors and teachers. The program is designed to explore the intersection of science, technology, engineering and math (STEM) as it pertains to applications in biological research including big data analysis, gene editing, genomics, high through put sequencing, nanotechnology, among others. Our NSF funded STEM exploration workshops are designed for undergraduate students, high school students, and rising undergraduates, who may be interested in exploring applications of STEM paths to medicine and biological research with emphasis in genomics and new cutting-edge science. Discussions and workshop exercises allow students to gain insight into applications of STEM while they interact with faculty and guest speakers. Students have the opportunity to gain a broad overview of STEM applications while they learn about current biological and medical cutting-edge research topics that are at the intersection of medicine, biology, and engineering career paths that will equip students to lead in increasingly complex, interconnected, and diverse fields. The primary workshop objective is to give participants a greater understanding of interdisciplinary fields and motivate students to follow STEM paths to produce future leaders prepared to influence their communities and the world in positive ways. Exposure to STEM is critical for high school students and for inspiring future scientists in the US. The spark is the discovery of what science and technology have to offer them in the future. **In this presentation, we will highlight one of our outreach STEM modules on the origin of the Pacific BioSciences (PACBio) long-read sequencing technology at Cornell University, with one of the co-inventors Dr. Harold Craighead, co-inventor of the technology and former Director of the National Nano-fabrication Facility at Cornell University.**

Our STEM outreach workshops are funded by the US National Science Foundation (NSF Award Award#1444893).