International Coffee Genomics Network (ICGN)
Report 8th Coffee Genomics Workshop held at the XXIII Plant and Animal Genome (PAG) Meeting
San Diego, California
January 9-15, 2015

Coffee Genomics Workshop Speakers

1. **Alvaro Gaitán** and **Marco Cristancho**, CENICAFE, Colombia. Long-Read deep sequencing and assembly of the allotetraploid *Coffea arabica* cv. Caturra and its maternal ancestral diploid species *Coffea eugenioides*.
2. **M. Morgante**, IGA Technology Services/ DISA, University of Udine, Italy. Progress report on the sequencing and assembly of the allotetraploid *Coffea arabica* var. Bourbon genome.
3. **Alexandre de Kochko**, IRD, France. Dihaploid *Coffea arabica* genome sequencing and assembly.
4. **Romain Guyot,** IRD-France. Transposable element distribution, abundance and impact in genome evolution in the genus *Coffea.*
5. **Elaine Silva Dias,** UNESP – Univ. Estadual Paulista, Departamento de Biologia, Brazil. Impact of transposable elements on the evolution of *Coffea arabica* (Rubiaceae).

Coffee abstract presented at Sequencing of Complex Genomes Workshop

6. **Perla Hamon,** IRD UMR DIADE, France. The international *Coffea* genome13 project: A way to understand the evolutionary history of *Coffea* genomes and unlock the potential use of wild species in breeding?

Coffee Genomics Workshop at PAG

The Plant and Animal Genome (PAG) meeting is the largest international scientific conference reporting on animal and plant genomics developments in the world, this year with 3002 participants from 65 countries. For those interested in participating in future meetings see http://www.intlpag.org. The XXIV Plant & Animal Genome Conference will be held in San Diego, January 9-13, 2016.

More than 60 scientists participated in our 8th Coffee Genomics Workshop held as part of the PAG Meeting in San Diego on January 11, 2015. The co-organizers of the workshop, Marcela Yepes (Cornell University, my11@cornell.edu), Philippe Lashermes (IRD-CIRAD, France, philippe.lashermes@ird.fr), and Rod Wing (University of Arizona) thank the speakers for their participation and contributions. Abstracts presented at the workshop and poster presentations on coffee are included as an appendix at the end of this report. The abstracts and some pdfs of the presentations can also be accessed at the PAG Archives at https://pag.confex.com/pag/xxiii/webprogram/Session2616.html. Our next Coffee Genomics Workshop will be held January 10, 2016 as part of the XXIV PAG meeting in San Diego, January 9-13, 2016. Please contact one of the organizers if interested on presenting a talk or poster, or with suggestions for new topics for workshop presentations or for round table discussions organized by ICGN in conjunction with the PAG meeting. The coffee genomics workshop is an excellent opportunity to present advances in coffee genomics research to the International Plant and Animal Genomics Community and is helping our community explore new collaborations as well as funding opportunities.

ICGN survey and collaboration with the International Coffee Organization

ICGN is celebrating 10 years in 2015 (see pdf presentation of history of our network and accomplishments at http://www.coffeegenome.org/communications/publications.htm). ICGN is conducting a survey to help us update our mailing list, identify future priority projects for the community as well as new leadership to help secure funding for new proposals. Please help us contribute to this effort by completing and submitting the survey available at our www site (http://www.coffeegenome.org). Survey results will be discussed at the next ICGN meeting held in conjunction with the 2016 PAG meeting in San Diego.

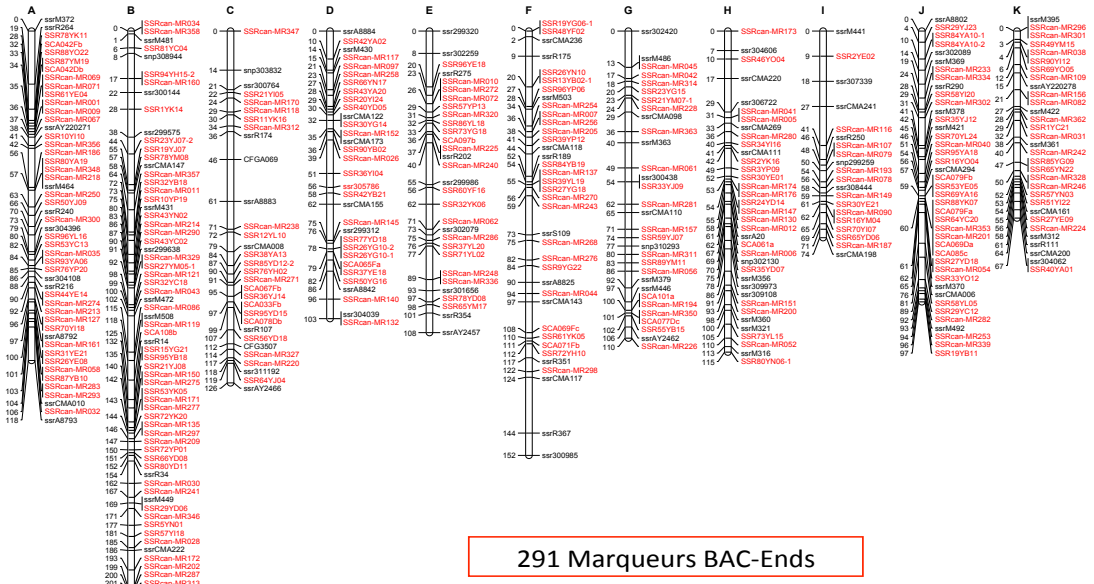
As the first *de novo* coffee genome references and assemblies become available (see report on the status of on going projects below), we would like to take advantage of the momentum to identify new priority projects of interest that ICGN can develop as a community to help mine the data generated and develop innovative tools and advanced resources in coffee genomics to address challenging issues for our community such as climate change adaptation and sustainability that could be accelerated with transforming genomic technologies and strategies. The African Coffee Research Network (ACRN) joined ICGN in 2011 as an institutional member, and its Director of Research and Development, Dr. Bayeta Bellachew helped us conduct the ICGN survey among ACNR members at several Coffee Research Institutions in Africa. We received through ACRN responses from scientists and scientific groups from the following countries: Ethiopia, Kenya, Rwanda, Uganda and Ghana with strong interest to work with ICGN on a global initiative to develop advanced genomic tools to speed up diversity characterization, enhanced utilization and conservation of *Coffea* germplasm in the context of climate change. In addition with support from the International Coffee Organization (ICO), ICO member countries have been contacted to discuss possible interest in developing a global initiative in collaboration with ICGN/ICO aiming at improving conservation and characterization of the world coffee gene pool for varietal development in a world of changing farming systems and climate. Other ICO member countries that have expressed strong interest in working on an ICGN/ICO collaborative proposal include, for Europe: France (IRD-CIRAD); for Latin America: Brazil, Colombia, Guatemala, Costa Rica, Mexico; for Africa: Cote D'Ivoire, Ethiopia, Kenya, Malawi, as well as the Inter-African Coffee Organization; and for Asia: India and Vietnam.

ICGN is grateful for the invitation by the ICO Executive Director Dr. Robeiro Oliveira Silva to participate as an observer in the ICO Council meetings in 2015, and we are looking forward to working closely with ICO officials on the preparation and submission of a first ICGN/ICO proposal, and to explore potential sources of finance for such joint initiative. Support from ICO and private sector will be key for ICGN to secure future funding for diversity conservation efforts in *Coffea* with a broader funding base, and to promote coffee genomics research for coffee improvement targeting priority traits for different regions as well as for the coffee industry. Capacity building in developing countries to participate in coffee genomic research can be supported through ICGN networking to help us secure international funding for those efforts.

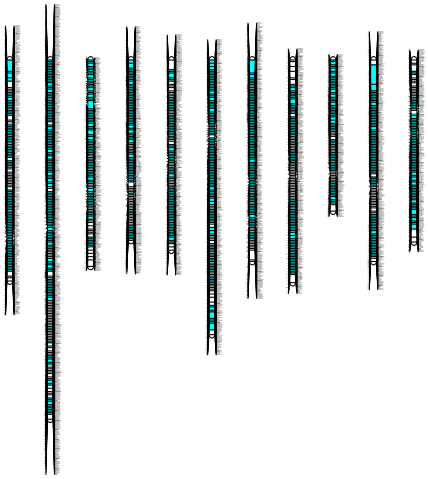
Update Status High-Density Mapping of the diploid species *Coffea canephora*

A high-density reference genetic map for *Coffea canephora* Pierre was constructed in the frame of International Coffee Genomics Network (ICGN) in collaboration with IRD/CIRAD, Nestlé R&D Centre and the Indonesian Coffee and Cocoa Research Institute. The population mapped was from a cross between two highly heterozygous genotypes, a Congolese group genotype (BP409) and a Congolese-Guinean hybrid parent (Q121). The segregating population is composed of 93 F1 individuals. DNA from the two parental clones and the segregating progeny were distributed to several ICGN members (on request). First, a high-density genetic map was constructed including 1481 loci covering 1400 cM markers, with a third of the SSR markers derived from BAC end sequences (see figure in page 3). The first set of markers mapped included: 360 RFLPs, >890 SSRs, and 213 SNPs, and were mined from genomic or EST libraries from different institutes (IRD, CIRAD, Trieste University, Cornell University, CENICAFE, and Nestlé).

As a second approach, Restriction Associated DNA sequencing (RADseq), which enables synchronous SNP marker discovery and genotyping using massively parallel sequencing, was used. The RAD libraries were made from digestion of DNA using two restriction enzymes, *Nsi*I (6 base cutter) and *Mse*I (4 base cutter). The fragments (150 - 500 bp) were selected to ligate to two adaptors, and one of them with tag for each progeny. Equal amount of amplicons from each individual were pooled to make Illumina libraries with individual tags for each library. Co-segregating markers within 50 kb region (< 1 cM) based on the aligned template scaffold were sorted as bin. One marker from each bin was selected for mapping. The linkage analysis and map construction were performed using JoinMap software version 4.1 using LOD threshold of 5 and Kosambi's function to calculate genetic distance between two loci. The Robusta consensus genetic map was built using the F2 segregating loci as anchor markers in order to merge the two homologous parental linkage groups. Using RAD sequence data from the segregating population previously selected 1747 RADseq markers were added.



The final high density Robusta map (see figure below) comprises 3230 loci, genetic size 1471 cM (1cM ~ 500 Kb), with an average density close to one marker every 220 Kb. The F1 high density genetic map will facilitate comparative genomic studies based on synteny and provided the opportunity for anchoring and ordering the numerous scaffolds arising from the *Coffea canephora* genome sequencing (see report below). Initially, the DNA sequences (scaffolds) anchored covered approximately 75% of the genetic map (1023 cM). Additional mapping efforts focused on the identification and mapping of SSRs from the *C. canephora* sequence scaffolds that were not or were insufficiently anchored. Both the high-density genetic map and the marker information will be freely available on a dedicated web-site once the construction of the map is completed. Please send information or comments to Philippe Lashermes (philippe.lashermes@ird.fr), Dominique Crouzillat, Nestlé (dominique.crouzillat@rdto.nestle.com), or Ray Ming (rming@life.illinois.edu). The international *C. canephora* high density map is a highly valuable resource for different applications including transposition to other mapping populations, as genetic framework that can be used for various QTL studies, as well as genome structure comparisons. RAD sequencing is a powerful strategy for genotyping in coffee to provide access to high-throughput SNP detection.



Update status of the *Coffea canephora* genome sequencing

In the frame of the International Coffee Genomics Network (ICGN) through a proposal funded by the Agence Nationale de la Recherche (ANR; Genoplante ANR-08-GENM-022-001), France, several Institutes (Genoscope-CEA, IRD and CIRAD) combined their scientific resources and expertise to sequence, assemble, and annotate the entire genome of *C. canephora*. Additional partners included several ICGN members (EMBRAPA/Brazil, ENEA/Italy, University of Trieste/Italy, University of Queensland/Australia, CCRI/India,

University of Illinois, Urbana/USA, Hawaii Agriculture Research Center HARC/USA, SUNY Buffalo/USA, University of Ottawa/Canada). A community effort for genome annotation is on going. The *C. canephora* genome consists of 11 chromosomes, is about 710 Mb in size, and was sequenced *de novo* with deep coverage using different sequencing platforms. Genoscope lead the sequencing and assembly of the *C. canephora* genome. Patrick Wincker, Head of Sequencing and Coordinator of Eukaryote Annotation and Analysis at Genoscope, presented the sequencing strategy and the status of the project during the 4th ICGN Coffee Genomics Workshop at PAG in San Diego in 2011. In 2013, France Denoeud from Genoscope presented at the 6th ICGN Coffee Genomics Workshop an update on the first genome assembly, and Alexis Dereeper presented at the 7th ICGN Coffee Genomics Workshop the **Coffee Genome Hub**, an integrative genome information system accessible through the South Green Bioinformatics Platform, developed to provide centralized access to all the coffee scientific community of the full *C. canephora* genome sequence, as well as genomics, genetics, mapping, and breeding data and analysis tools to facilitate basic, translational and applied research in coffee (Dereeper *et al*. 2014. Nucleic Acids Research. 43: D1028-D1035, http://nar.oxfordjournals.org/content/43/D1/D1028.full.pdf+html) . **The manuscript for the sequencing of the *C. canephora* genome was published in 2014** (Denoed *et al.* 2014. Science 345: 1181-1184; http://www.sciencemag.org/content/345/6201/1181.full, and the genome assembly can be freely accessed at: http://coffee-genome.org.

Summary of the sequencing strategy used for C. canephora:

C. canephora is one of the ancestral progenitors of the widely cultivated, *C. arabica,* a recent allotetraploid species formed from the merger of the diploid species *C. canephora* and *C. eugenioides*. The accession DH200-94, a doubled haploid genotype was selected for sequencing because of its homozygous nature to facilitate genome assembly. *De novo* genome sequencing with deep coverage was performed using both 454 Roche and Illumina next generation sequencing technologies. Direct whole genome shotgun (WGS) sequencing and paired-end sequencing of large insert libraries 8kb and 20 kb insert libraries was conducted. Furthermore, 73,728 BAC clones from two *C. canephora* BAC libraries *Hin*d III and *Bst*YI constructed in collaboration with Rod Wing at University of Arizona were BAC-end sequenced using Sanger technology. Average inserts for each BAC library were 166 and 121 kb in size, with 36,864 BAC clones per library for an estimated coverage of ~8.6X and 6.3 X per library, respectively. Both *C. canephora* BAC libraries are publicly available at Arizona Genomics Institute Resource Center (http://www.genome.arizona.edu.orders). BAC end sequences (BES) are also publicly available and were deposited in EMBL-EBI Bank (accession numbers FO535330, FO538768 to FO624989, and FO624992 to FO680656) (A. Dereeper *et al*. 2013. BAC-end sequences analysis provides first insights into coffee (*Coffea canephora* P.) genome composition and evolution. Plant Molecular Biology 83: 177-189. The full manuscript can be accessed at: http://www.researchgate.net/publication/257120929_Dereeper_et_al_canephora_BAC_ends.

The genome sequencing data generated for the *C. canephora* genome assembly included an estimated coverage of 28.87 X (454 Roche/Sanger) and Illumina 69.7 X, as follows:

Roche/454 Titanium:

Whole Genome Shotgun (WGS) 454 sequencing: 28.9X coverage (assuming a genome size of 710 Mb) including:

23X single end 454 Titanium reads

– Reads single end: 14.83X , Mean size: 359 bp

– Long reads single end: 8.24X , Mean size : 462 bp

5.8X Paired-end sequencing of long insert libraries (8 and 20 Kb): 5.8X (2.2X for 8kb, 3.6X for 20kb), Mean Size: 252 bp.

Sanger end sequencing of Bacterial artificial chromosome (BACs):

Two BAC libraries (*Hind* III and *Bst*YI) were constructed in collaboration with Rod Wing at Arizona Genomics Institute. The BAC libraries have 73,728 clones (>11X coverage).

Sanger BAC-end sequencing: 131,412 BES were generated (73,728 BAC clones x 2 ends 5' and 3'): 0,27X

Mean BAC insert size : 135 Kb, range: 63,2Kb < insert < 253,6 Kb

Illumina sequencing was done at deep coverage (~70X) to correct bias of different sequencing platforms.

Single reads coverage 7.3X: read size 76 bp (4.8X) and read size 150 bp (2.5X)

Paired end reads coverage 62.4X: read size 76 bp (42.4X) and read size 108 bp (20X)

Assembler used Newbler and assembly statistics for the first raw assembly and for the final assembly were summarized by Genoscope see Denoeud *et al.* 2014, Supplemental Table 2, as follows:

Table S2. Assembly statistics.

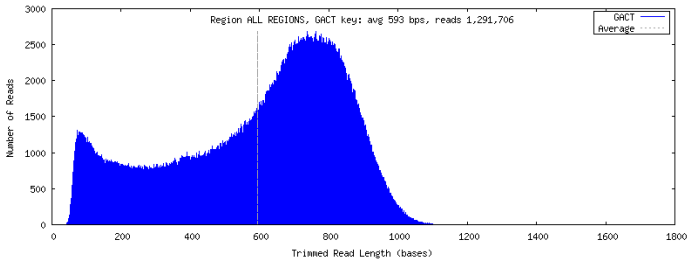
	Raw assembly		Final Assembly	
	Contigs	Scaffolds	Contigs	Scaffolds
Number	91,439	13,345	25,216	13,345
Cumulative size (Mb)	475.6	569.4	471.3	568.6
Average size (kb)	5.2	42.6	18.7	42.6
N50 size (kb)	14.8	1,261	51.1	1,261
N50 number	8,509	108	2,290	108
N80 size (kb)	4.3	65.2	15.5	65.3
N80 number	26,145	637	7,259	635
Largest size (kb)	193.8	9,035	817.6	9,028

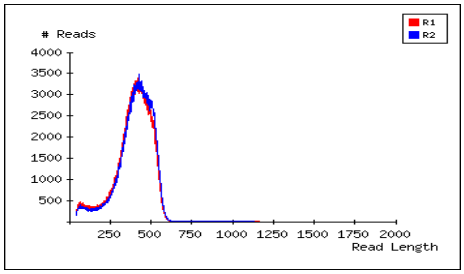
Update status of the *Coffea eugenioides* genome sequencing

This project is also being developed in frame with the ICGN, co-funded by the Inter American Development Bank (FONTAGRO/SECCI), and the Colombian National Coffee Growers Federation and its National Coffee Research Center, CENICAFE. Genome sequencing was started towards the end of 2012. The project is being developed collaboratively by CENICAFE and Cornell University. Funding for this project was secured jointly through a proposal prepared and submitted by Cornell University and CENICAFE.

We mimicked the strategy used for the *C. canephora* genome sequencing to generate a high quality reference assembly for *C. eugenioides* using mixed next generation sequencing platforms: Roche 454 FLX+ and Illumina HiSeq 2500. We collaborated with Roche to construct and sequence a whole genome shotgun (WGS) library (fragment size >1,100 bases and <2,000 bases) using 454 FLX+ single end reads with mode length of 763 bases to generate a total of 6,082,341,937 bases for an estimated 9.2X coverage of the *C. eugenioides* genome (~estimated genome size of 660 Mb). See figure on quality control results of a typical 454 FLX+run. In addition, we collaborated with Roche to construct and sequence twelve 20 Kb long insert libraries (3.1X coverage) using paired end sequencing and Roche 454 FLX Titanium. See graph of the paired end statistics profile for a run: Read length distribution of high quality paired-end reads. R1 = Coffee 20kb 1-1, R2 = Coffee 20kb 1-1. Linker positive displays statistics of reads with paired end linker sequence 71.47% and 73.62%. Linker Negative refers to reads with no paired end linker

Region	Passed Filter	Total Bases	Reads ≥ 1000 bases (% Passed Filter)	Read Length (Bases)				
				Mean	Median	Mode	Min	Max
1	634,065	373,033,536	4,712 (0.74%)	588.3	653	763	40	1,735
2	657,641	392,830,440	3,723 (0.57%)	597.3	660	782	40	1,692
ALL REGIONS	1,291,706	765,863,976	8,435 (0.65%)	592.9	657	763	40	1,735





sequence 28.53% and 26.38%. The overall data generated for the twelve 20 Kb-insert libraries paired end-sequenced using Roche 454 FLX Titanium included:

Paired end reads 57% 5,876,463 reads x 170 bases x 2= 1,997,997,420 bases/660,000,000 = **3.03 X Coverage**
Non paired end reads 43% 4,225,283 reads x 298 bases = 1,259,134,334 bases/660,000,000 = **1.91 X Coverage**

The first genome assembly for *C. eugenioides* using the 454 data described above and Newbler v.3.0 was completed in collaboration with Roche in 2014, and was presented by Marco Cristancho from CENICAFE at the 8th ICGN Coffee Genomics Workshop, and also at the Roche Workshop during the PAG meeting in San Diego in January, 2015. Below is a summary table comparing the initial raw assembly for *C. eugenioides* with the initial raw assembly for *C. canephora* (statistics for *C. canephora* are from Genoscope data presented previosuly in San Diego and from Table S2 of manuscript Denoeud *et al.* 2014):

Initial Raw Assembly Comparison:

	C. eugenioides Heterozygous genotype Raw assembly	***C. canephora*** Homozygous Double haploid Raw assembly
No. of contigs (>100 bp)	364,530	211,157
No. of contigs (>500 bp) raw assembly	146,520	96,182
No. of scaffolds	30,263	13,345
Average scaffold size	**15,897 bp**	**42,606 bp**
Genome size assembled	508 Mb	569 Mb
Estimated genome size	630-660 Mb	710 Mb
% genome assembled	77-80.6%	80.14%
N50	**209,891 bp (502 scaffolds)**	**1,260,636 bp (108 scaffolds)**
Largest scafold	4.0 Mb	9.0 Mb
Assembler Newbler	v.3	v.2.3
Q40	97.85%	
Inferred read error	1.41%	

- For this first raw assembly of *C. eugenioides*, the average depth of the alignment for the contigs was very good with coverage of X15-25 for individual contigs, and an estimated 77-80.6% of the genome assembled.
- The Q40 indicated very good data quality for the 454 *C. eugenioides* raw assembly using large contig consensus sequences, with **very high Q40= 97.85%** of consensus sequences of Q40 or higher, indicating a very **low error rate: probability of an incorrect base 1/10,000 bp**. The inferred error rate was also very low 1.41%.
- The 20 Kb insert size library appeared to cover most of the repeat regions for *C. eugenioides* (99.8%), with only 0.22% of the paired reads being unmapped (12,644 out of 5,727,383).

In addition, we have constructed and high throughput sequenced a *C. eugenioides* library using PACBio P6/C4 chemistry (read N50 10 Kb) and generated also Illumina Moleculo (synthetic 10 Kb fragments) to help us connect and reduce the overall number of contigs and scaffolds in the *C. eugenioides* raw assembly, as well as to increase the overall percent of genome assembled (currently 77-80.6%). The reference genomes of the diploid species *C. canephora* and *C. eugenioides* (parental diploid ancestors of the allotetraploid species *Coffea arabica)* will serve as frames for assembly of *C. arabica*, the major cultivated coffee species worldwide.

Update *Coffea arabica* sequencing

With funding from the Inter American Development Bank (IDB/FONTAGRO), Cornell University and CENICAFE have also sequenced the allotetraploid *Coffea arabica* genome to generate a deep coverage PACBio only assembly of *C. arabica*. The allotetraploid assembly will be validated using the high quality genome assemblies of its two diploid ancestral species. We collaborated with Pacific BioSciences to

generate three *C. arabica* genomic libraries that were high-throughput sequenced using PACBio to generate a 56-60X coverage of the allotetraploid *C. arabica* genome (73.45 Gb). We have also generated Illumina Moleculo (synthetic 10 Kb fragments) data for *C. arabica* that will be used to validate the assembly. Transcriptome assembly and genotype-by-sequencing (GBS) studies to validate genome assemblies and anchor scaffolds to chromosomes for *C. arabica* and *C. eugenioides* are on going, and should dramatically improve our current understanding of coffee genetics and genomics providing direct applications to breeders for climate change adaptation. Integration of genomic studies of equivalent quality among the allotetraploid *C. arabica* and it diploid progenitors will maximize scientific insights into the complex biology of polyploids.

Updates on the status of the projects funded by Nestlé, and IllyCaffé/Lavazza were presented at the 8th ICGN Coffee Genomics Workshop. See abstracts included in the appendix.

Perspectives

High quality coffee genome and transcriptome assemblies for *C. arabica* and its diploid ancestral species will provide an invaluable resource for future coffee improvement strategies. It will speed up the identification of genes involved in important agricultural traits, and help build crucial information on the structural variation between *Coffea* wild species and cultivated accessions. They will facilitate positional cloning of agriculturally important genes, re-sequencing and deep diversity analysis in the *Coffea* gene pool, and will support the development of genomic tools for whole-genome expression analysis. They will also provide a foundation to study the evolution of euasterids and accompanying genomic changes.

To ensure full benefit from the generated coffee genomic sequences and resources by the coffee sector, ICGN continues to explore additional funding from International Funding Agencies to support our community efforts to accelerate the use of rapidly eroding diversity for coffee improvement and to establish *Coffea* as a model for genome-guided studies of trait-evolution, speciation and domestication toward adaptation of the crop to climate change.

Acknowledgements:

ICGN is particularly grateful to all our workshops speakers who kindly accepted our invitation to participate in our 8th ICGN coffee genomics workshop at PAG. Abstracts of their presentations are enclosed below in the appendix, as well as other abstracts and posters presented on coffee genomics during the PAG meeting.

See below also pictures of the ICGN workshop speakers and some of the workshop participants as well as information on future meetings of interest to our ICGN Community.

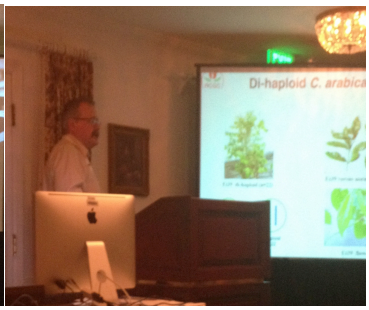
Pictures coffee genomics workshop speakers and participants:

8th ICGN Coffee Genomics Workshop speakers at PAG

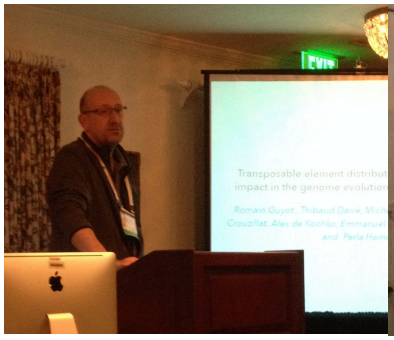
Alvaro Gaitán

Marco Cristancho

Michelle Morgante



Alexandre de Kochko



Romain Guyot

Elaine Silva Dias

Other ICGN Coffee Genomics Workshop participants pictured below:

From left to right: Marcela Yepes (Cornell University, USA), Elaine Silva Dias (UNESP, Brazil), Carmenza Góngora, Alvaro Gaitán, and Marco Cristancho (CENICAFE, Colombia)

From left to right: Alvaro Gaitán (CENICAFE), Marcela Yepes and Herb Aldwinckle (Cornell University), Marco Cristancho (CENICAFE)

From left to right: Marco Cristancho and Alvaro Gaitán (CENICAFE) invited Speakers to the Roche workshop at PAG co-hosted by William J. LaRochelle (center) and Susan Ulanowicz (far right) from Roche.

From left to right: Elaine Silva Dias (UNESP, Brazil), Howard Laten, Loyola University, Chicago, USA, and Romain Guyot (IRD, France)

From left to right: Alexandre de Kochko (IRD, France), Dominique Cruzillat (Nestlé), Marcela Yepes (Cornell University), Perla Hamon (front center- IRD, France), and Romain Guyot (far right- IRD, France)

Appendix

Abstracts 8th ICGN Coffee Genomics Workshop 2015

Workshop Co-Organizers:

Marcela Yepes, Cornell University (my11@cornell.edu)
Philippe Lashermes, L'Institut de Recherche pour le Développement (IRD), France (philippe.lashermes@ ird.fr)
Rod Wing, University of Arizona (rwing@Ag.arizona.edu)

(Program and abstracts with pdfs also posted at:
https://pag.confex.com/pag/xxiii/webprogram/Session2616.html

Long-Read Deep Sequencing and Assembly of the Allotetraploid *Coffea arabica* cv. Caturra and its Maternal Ancestral Diploid species *Coffea eugenioides*

Alvaro Gaitán[1], Marco Cristancho[1], Carmenza Góngora[1]*, Pilar Moncada[1], Huver Posada[1,2], Fernando Gast[1], Marcela Yepes[3], and Herb Aldwinckle[3]*

[1]Colombian National Coffee Research Center, CENICAFE, Chinchiná, Caldas, Colombia, [2]Federación Nacional de Cafeteros de Colombia, Bogotá, Colombia, and [3]Cornell University, Geneva, New York, USA

Over the last decade, climate change has caused major reductions in coffee production due to increased incidence of insect pests and diseases, as well as abiotic stresses that are threatening sustainable coffee production at a global scale. In Latin America, the coffee leaf rust epidemic has had a devastating effect with losses of more than 1 billion dollars that are threatening coffee production and food security for many small coffee farmers. During 2008-2011, coffee leaf rust caused a reduction of nearly one third of the coffee harvest in Colombia, and between 2012-2014, it has caused >50% reduction in production in Central America, affecting more than 5 million people. Peru has been hit particularly hard in 2014, and climatic conditions continue to be favorable for spread of the fungus. Our research is focus on *de novo* sequencing and assembly of the coffee genome to speed up adaptation of the crop to climate change.

We collaborated with Pacific BioSciences for the construction of three long insert coffee genomic libraries, Blue Pippin size selection, QC, and high-throughput PACBio sequencing as well as *de novo* assembly of the allotetraploid *Coffea arabica* cv. Caturra genome. This genotype was selected since we have already developed a high density molecular genetic map (>1,000 sequenced based markers), and a BAC library (11X coverage, 114,816 clones) completely BAC end-sequenced using Sanger and fingerprinted, as well as comprehensive transcriptomic data (>125,000 Sanger-sequenced ESTs), and several populations that are currently being phenotyped for climate change adaptation. High-throughput PACBio sequencing of the libraries generated 73.54 Gb of post-filtered data with a N50 read of 12-15Kb for ~57-60X coverage of the *C. arabica* genome (estimated genome size 1,300 Mb). This data has been used for a first assembly of the allotetraploid *C. arabica* genome.

To guide and validate the allotetraploid assembly, we will use high quality assemblies of the genomes of its two diploid ancestral species. We collaborated with Roche to generate the first high quality *de novo* assembly of the maternal diploid ancestor of *Coffea arabica*, the diploid species *C. eugenioides* (estimated genome size 660 Mb). We used Roche 454 FLX plus long reads (average read length 750 bp) to sequence a WGS library at 9X coverage, and Roche 454 Titanium (paired end reads) to sequence twelve 20 Kb insert libraries at 3.03X coverage, as well as Illumina Moleculo (synthetic 10 Kb fragments) paired end sequenced at 3.43X coverage. Our strategy mimicked the sequencing strategy for the high quality assembly of the diploid paternal ancestor of *C. arabica*, the cultivated species *C. canephora* that was recently published (Denoed *et al.* 2014. Science 345: 1181-1184; genome assembly can be accessed at: http://coffee-genome.org). Transcriptome assembly and genotype-by-sequencing (GBS) studies to validate genome assemblies and anchor scaffolds to chromosomes for *C. arabica* and *C. eugenioides* are on going, and should dramatically improve our current understanding of coffee genetics and genomics providing direct applications to breeders for climate change adaptation. Integration of genomic studies of equivalent

quality among the allotetraploid *C. arabica* and it diploid progenitors will maximize scientific insights into the complex biology of polyploids.

This work is funded by the InterAmerican Development Bank (IDB) and the Fondo Regional de Tecnologia Agropecuaria (FONTAGRO), as well as by the Colombian National Coffee Growers Federation and its National Coffee Research Center, Centro Nacional de Investigaciones de Café, (CENICAFE), and by the Colombian Ministry of Agriculture.

This abstract had an extended time slot as it was the merge of two abstracts and was presented by co-authors A. Gaitán, and M. Cristancho. **Note:** ***This abstract was also presented at the Roche Workshop at PAG by co-author M. Cristancho.***

Progress report on the sequencing and assembly of the allotetraploid *Coffea arabica* var. Bourbon genome

M. Morgante[2,7]*, S. Scalabrin*[1]*, F. Cattonaro*[1]*, D. Scaglione*[1]*, F. Magni*[1]*, I. Jurman*[2]*, M. Cerutti*[3]*, F. Suggi Liverani*[4]*, L. Navarini*[4]*, L. Del Terra*[4]*, G. Pellegrino*[3]*, N. Vitulo*[5]*, G. Valle*[5]*, G. Graziosi*[6]

[1]IGA Technology Services, Via J. Linussio, 51 - Udine 33100, Italy; [2]IGA, Via J. Linussio, 51 - Udine 33100, Italy; [3]Lavazza Spa, Strada Settimo 410 - Torino, Italy; [4]illycaffè S.p.A., via Flavia, 110 - Trieste 34147 – Italy; [5]Dip. di Biologia; Università di Padova, Via U. Bassi 58/B - Padova 35121, Italy; [6]DNA Analytica Srl - Area Science Park, Padriciano, 99 - 34149 Trieste, Italy; [7]DISA, University of Udine, Via delle Scienze 208 - Udine 33100, Italy

It is well known that *Coffea arabica* is the result of a cross pollination between two *Coffea* species, very likely *Coffea canephora* and *Coffea eugenioides*. Moreover, *arabica* can set flowers and fruits by self-fertilization, and indeed beans can be obtained by a single and isolated plant. Such reproductive behaviour should find some justification in its genome. A genome sequencing project has been initiated to investigate the structure of the allotetraploid genome of Arabica.
High molecular weight genomic DNA was obtained from entire plantlets of *Coffea arabica* var. Bourbon and a BAC library was constructed. 175,872 BAC clones were pooled into 96 pools of 384 clones each and the pools underwent DNA sequencing on next generation sequencing Illumina platform. Whole genome shotgun sequencing was also performed on two Illumina libraries with 500 and 800 bp insert size and on one mate-pair library with inserts of two kbp. These libraries were supplemented by the sequencing of cDNA libraries (RNA-seq on Illumina platform) obtained from leaves, root and cherries to use for gene prediction. A preliminary assembly of the genome has been carried out. The assembly is now mapped on the available *Coffea canephora* genome to obtain pseudomolecules. The preliminary bioinformatics analysis of the *arabica* genome suggests a high degree of polymorphism between its sub-genomes, in line with the allotetraploid constitution of the *Coffea arabica* genome.

Dihaploid *Coffea arabica* Genome Sequencing and Assembly

Alexandre de Kochko[2]*, Dominique Crouzillat*[1]*, Michel Rigoreau*[1]*, Maud Lepelley*[1]*, Laurence Bellanger*[1]*, Virginie Mérot-l'Anthoëne*[1]*, Céline Vandecasteele*[1]*, Romain Guyot*[2]*, Valérie Poncet*[2]*, Christine Tranchant*[2]*, Perla Hamon*[2]*, Serge Hamon*[2]*, Emmanuel Couturon*[2]*, Patrick Descombes*[3]*, Déborah Moine*[3]*, Lukas Müeller*[4]*, Suzy Strickler*[4]*, Alan Andrade*[5]*, Luiz Filipe Pereira*[5]*, Pierre Marraccini*[6]*, Giovani Giuliano*[7]*, Alessia Fiore*[7]*, Marco Pietrella*[7]*, Giuseppe Aprea*[7]*, Ray Ming*[8]*, Jennifer Wai*[8]*, Douglas Silva Domingues*[9]*, Alexandre Paschoal*[10]*, Gerrit Kuhn*[11]*, Jonas Korlach*[11]*, Jason Chin*[11]*, David Sankoff*[12]*, Chunfang Zheng*[12]*, Victor Albert*[13]

[1]Nestlé R&D Tours France, [2]IRD, Montpellier France, [3]NIHS, Lausanne Switzerland, [4]Boyce Thompson Institute, USA, [5]EMBRAPA Brazil, [6]CIRAD France, [7]ENEA Rome Italy, [8]University of Illinois Urbana-Champaign USA, [9]IAPAR Londrina Brazil, [10]University of Londrina Brazil, [11]Pacific Biosciences USA, [12]University of Ottawa Canada, [13]University of Buffalo USA

Coffea arabica which accounts for 70% of world coffee production is an allotetraploid with a genome size of approximately 1.3 Gb and is derived from the hybridization of *C. canephora* (710 Mb) and *C. eugenioides* (670 Mb). To elucidate the evolutionary history of *C. arabica*, and generate critical information for breeding programs, a sequencing project is underway to finalize a reference genome using a dihaploid line and a set of 30 *C. arabica* accessions. For the reference genome, we have generated two assemblies, one from Illumina data (>150x coverage) and a second from PacBio sequences (>50x coverage). The present assemblies cover 1,031 and 1,042 Mb, respectively. After further refinement, using Illumina mate pairs and optical mapping, the genome assemblies will be annotated using RNA-Seq. Re-sequencing of *C. eugenioides* and *C. canephora* has been completed and is being used to better assess homeologs within the sub-genomes. Furthermore, 30 *C. arabica* accessions, representing wild and cultivated genotypes, are being resequenced (20x coverage) using Illumina. A *C. arabica* genetic map, currently including over 600 SSR markers, that differentiate between the two sub-genomes, is used to anchor the assemblies. Newly identified SNP markers are being added to the map. The final goals of the project are to produce a high quality reference genome, assess an eventual neo-diversification occurring in the cultivated varieties, have a better understanding of the species formation and evolution, and develop tools that will make the finished genome accessible and useful to breeders and researchers. **Note: This abstract was also presented at the PACBio workshop at PAG by coauthor Suzy Strickler.**

Transposable Element Distribution, Abundance and Impact in Genome Evolution in the genus *Coffea*

Romain Guyot [3*], *Thibaud Darré* [1], *Michel Rigoreau* [2], *Dominique Crouzillat* [2], *Emmanuel Couturon* [1], *Serge Hamon* [1] *and Perla Hamon* [1]

[1] Institut de Recherche pour le Développement (IRD), UMR DIADE (CIRAD, IRD, UM2), BP 64501, 34394 Montpellier Cedex 5, France; [2] Nestlé R&D Tours, 101 AV. G. Eiffel, Notre Dame d'Oé, BP 49716 37097, Tours, Cedex 2, France; [3] Institut de Recherche pour le Développement (IRD), UMR IPME (CIRAD, IRD, UM2), BP 64501, 34394 Montpellier Cedex 5, France

- Corresponding author: Romain Guyot romain.guyot@ird.fr

The genus *Coffea* is composed of 124 distinct species: 123 diploid and one polyploid species (*Coffea arabica*). They are classified into three botanical sections according to their geographic distribution: Eucoffea (West and Middle Central Africa), Mozambicoffea (East Africa) and Mascarocoffea (Indian Ocean Islands), and genome size variations are about 2-fold (*C. humblotiana*: 469 Mb and *C. heterocalyx*: 863 Mb for diploid species and 1,240 Mb for *C. arabica*). Despite the fact that the *Coffea* genus includes the world's most traded agricultural product: coffee beans, the phylogeny, origin and evolution of species as well as their genome size variations remain particularly poorly understood.

To study the evolution of *Coffea* species and analyze their transposable elements composition, we used the Roche 454 technology to perform a low-coverage sequencing (ranged between 5-10X) of 11 species representative of the geographic distribution of the genus. The composition and distribution of transposable elements were studied. Despite the previous assumption that transposable elements don't seem to play an important role in genome size variation in *Coffea*, we found a clear variation of one LTR Retrotransposon family called SIRE: SIREs are abundant in Eucoffea but almost absent from Mozambicoffea while Mascarocoffea species are completely devoid of these elements. These results suggest that SIRE are involved in *Coffea* genome differentiation and the restructuring of the Eucoffea genomes.

Impact of Transposable Elements on the Evolution of *Coffea arabica* (Rubiaceae)

Elaine Silva Dias[1,2], *Serge Hamon*[2], *Perla Hamon*[2], *Romain Guyot*[2], *Alexandre de Kochko*[2], *Claudia Carareto*[1]

[1] UNESP – Univ. Estadual Paulista, Departamento de Biologia, São José do Rio Preto, SP, Brazil; [2] IRD UMR DIADE, EVODYN, BP 64501, 34394 Montpellier Cedex 5, France

Transposable elements (TEs) are able to promote not only point mutations at the locus level but also large genomic rearrangements. They are prone to contribute to duplications, deletions, inversions and/or translocations, resulting in genomic structural changes, and sometimes, functional modifications. LTR-retrotransposons (RTs) are particularly involved in these kinds of modifications, and because of their mode of transposition following the "copy and paste" model, they may also be responsible for changes in genome sizes. We investigated the transcriptional activity and diversity of insertion sites for nine RTs, two belong to the *Copia* superfamily and seven to the *Gypsy*, using RT-PCR, IRAP and REMAP on 22 genotypes of the tetraploid *C. arabica*, and 10 of *C. canephora* and 2 of *C. eugenioides*, the *C. arabica* parental species. A differential transcriptional activity of the RTs was observed, being some of them active in the three species and others only in the hybrid. Exclusive insertion sites were observed in *C. arabica* suggesting that these RTs could have mobilized recently, less than 0.6 Mya (time of species origin). Additionally, the different patterns of band distribution comparing the parental species to the hybrid suggests that *C. arabica* could have undergone genomic structural changes associated to rearrangements mediated by some of these TEs, involving gains and losses of copies. In summary, the results show that TEs could be related to a structural genome evolution of *C. arabica* in a short time. Further analysis may reveal their potential to impact on some genome functions of this species.

Financial support: FAPESP, CAPES-Agropolis Foundation, IRD.

Sequencing Complex genomes Workshop

The International *Coffea* Genome13 Project: A Way to Understand the Evolutionary History of *Coffea* Genomes and Unlock the Potential Use of Wild Species in Breeding?

Perla Hamon , IRD UMR DIADE, Montpellier cedex 5, France

Comparative genomics between cultivated and wild species at the genus scale is a powerful tool to understand the evolutionary history of genomes and to unlock the genetic use of wild species in breeding programs. The recent availability of the *Coffea canephora* (Robusta coffee) genome allows for the first time an investigation of genome composition and evolution at the genus scale for tropical trees. *Coffea* genus (Rubiaceae) contains 124 species natives to inter-tropical forests of Africa, Indian Ocean islands, Indian sub-continent and Australasia. Only two African species are cultivated. Wild species grow in contrasting environments suggesting that they represent a potential source of diversity and a rich reservoir of genes for improving the world's most widely traded agricultural commodity. Based on a robust phylogenetic framework, we selected 13 diploid *Coffea* species with diverse geographical origins, environmental adaptations and genome sizes. All species were sequenced with the 454 (0.1X) and the Illumina technology (20X). The reads obtained were assembled into contigs and compared to the *C. canephora* reference genome sequence. These resources have allowed study of the maternal molecular phylogeny using chloroplast genome reconstruction and plastid insertion into the nuclear genomes. They will serve also as starting points to tackle among other issues, the origin of genome size variations, the evolutionary dynamics of transposable elements and non coding elements, the overall allelic diversity and the diversity of metabolic pathways that could be of interest for breeding targets especially to increase the sensory variability of the final product. The first results will be presented.

Analysis of Complex genomes Workshop

The Molecular Mechanisms of Plant Polyploidization Revealed By Systems Analysis of the Genomes and Transcriptomes of Wheat, Cotton and Related Species

Yang Zhang , Texas A&M University, College Station, TX
Yun-Hua Liu , Texas A&M University, College Station, TX
Meiping Zhang , Texas A&M University, College Station, TX
Qijun Zhang , North Dakota State University, Fargo, ND

Steven S. Xu , USDA-ARS, Fargo, ND
Wayne Smith , Texas A&M University, College Station, TX
Steve Hague , Texas A&M University, College Station, TX
James Frelichowski , USDA-ARS, College Station, TX
Hong-Bin Zhang , Texas A&M University, College Station, TX

Polyploidization is a predominant process in flowering plant speciation and evolution. The genomes of most angiosperms are thought to have incurred one or more polyploidization events during their evolution. It is estimated that approximately 70% of flowering plants, including numerous agriculturally important crops such as wheat, cotton, potato, canola, oats, peanut, tobacco, **coffee**, banana and sugarcane, are late-evolution polyploids. However, little is known about the underlying molecular mechanisms of how a polyploid species forms and evolves. These questions are addressed by systems analysis of the genomes and transcriptomes of a total of 634 lines of polyploid wheat, polyploid cotton and their related diploid species, including 112 newly synthetic polyploid wheat lines. We first scrutinized the variations of contents, family sizes and interactions of 29 fundamental function elements (FFE) constituting the genomes, including genes (GEN), DNA transposable elements (DTE), retro-transposable elements (RTE), simple sequence repeats (SSR) and low complexity repeats (LCR), in the processes of polyploidization from parental species to nascent polyploids to naturally-established polyploids. Then, we examined the variations of the transcriptomes of the species during the process. Finally, we investigated the variations of the contents and networks of the genes controlling cotton fiber length and wheat grain yield-related traits. The findings of this study have elucidated novel molecular mechanisms underlying plant polyploidization and significantly advanced our understanding of the process of plant speciation, variation and evolution.

Abstracts of Poster Presentations

Functional Analysis of the Coffee (*Coffea arabica*) *14-3-3* Gene Promoter

Fabiola Ocampo, IB - UNESP - Department of Genetics, Botucatu, Brazil
Ivan G. Maia, IB - UNESP - Department of Genetics, Botucatu, SP, Brazil

Coffee is one of the most important agricultural products in the world, being widely cultivated in tropical countries, for both internal consumption and export. In the last thirty years, several research efforts in genetics and biotechnology have been performed aiming to obtain improved coffee plants with greater disease resistance, high productivity and better quality drink. However, despite the molecular tools currently available, advances in coffee biotechnology are limited. In this context, the availability of highly active coffee promoters with known expression patterns is scarce and further research is needed. With the aim of providing tools for constitutive transgene expression in coffee plants, the present study undertook the functional characterization of the *14-3-3* gene promoter from *C. arabica*. To this end, the 5`-flaking region of the mentioned gene was transcriptionally fused to the GUS reporter gene and stably transformed in *Arabidopsis thaliana* Col-0. Histochemical and qPCR analysis of GUS expression in the generated transgenic plants showed variable reporter expression through the three tested developmental stages. In comparison to control plants transformed whit a 35S:GUS cassette, the 14-3-3 promoter drove a less intense and ubiquitous expression of the reporter gene than the one promoted by 35S.

Transcriptome Analysis in Leaves, Flowers, and Initial Fruit Development of *Coffea arabica*

Suzana Tiemi Ivamoto , Instituto Agronômico do Paraná, Londrina, Brazil
Osvaldo Reis Júnior , Universidade Estadual de Campinas, Campinas, Brazil
Leonardo Murai Sakuray , UEL, Londrina, Brazil
Priscila Mary Yuyama , UFRGS, Porto Alegre, Brazil

Marcelo Falsarella Carazzolle , Genomics and Expression Laboratory - State University of Campinas - UNICAMP, Campinas, Brazil
Gonçalo Amarante Guimarães Pereira , Genomics and Expression Laboratory - State University of Campinas - UNICAMP, Campinas, Brazil
Douglas S. Domingues , Instituto Agronômico do Paraná, Londrina, Brazil
Luiz F. P. Pereira , Embrapa Café, Londrina, Brazil

Coffee oil is rich in kaurane family diterpenes, mainly cafestol (CAF) and kahweol (KAH), which are related with plant defense mechanisms, nutraceutical and sensorial beverage characteristics. In plants, the cytochrome P450s gene family (CYPs) is usually involved in most of plant secondary metabolites, which probably includes the diterpenes. We measured CAF and KAH by HPLC in flowers as well as fruit perisperm in several stages (30 to 210 days after flowering – DAF). CAF levels were detected mainly in flowers as well as in the perisperm decreasing after 120 DAF. On the other hand, KAH concentration increased with perisperm development reaching a peak at 120 DAF. Based on this HPLC analysis of diterpenes, 12 RNA-Seq libraries were obtained for *Coffea arabica*: leaves, flowers and perisperm tissue from fruits. 41.881.572 high quality sequences were generated using Illumina HiSeq2000 technology. De novo assembly generated 65,480 unigenes, which includes 242 CYPs candidate genes. For five CYPs genes we observed a similar pattern between gene transcription and diterpenes concentration levels. Three CYPs (*CaCYP76F2_1*, *CaCYP82C4*, *CaCYP74A1*) had transcriptional patterns similar to CAF accumulation and two CYPs (*CaCYP71A4_1*and *CaCYP701A3*) were related with KAH accumulation. These five CYPs warrant further investigation as potential candidate genes involved in the final stages of CAF and KAH biosynthetic pathway providing us important clues and valuable information for future analysis of coffee diterpenes synthesis. This is the first work with Illumina sequencing of coffee fruit during their initial development stages.

Exploring Metabolic Networks: Metacyc and SolCyc as Examples for High-Level Data Curation, Depository and Management Across and within Species

Hartmut Foerster , Boyce Thompson Institute for Plant Research, Ithaca, NY
Lukas Mueller , Boyce Thompson Institute for Plant Research, Ithaca, NY
Noe Fernandez-Pozo , Boyce Thompson Institute for Plant Research, Ithaca, NY
Aureliano Bombarely, Virginia Polytechnic Institute and State University, Department of Horticulture, Blacksburg, VA
Ron Caspi , SRI International, Menlo Park, CA
Peter D Karp , SRI International, Menlo Park, CA

The multi-species database MetaCyc (http://metacyc.org) and species-specific databases such as SolCyc (http://solcyc.solgenomics.net/) represent data repositories which provide access to non-redundant metabolic data in the context of relevant pathways ideally furnished with detailed information about affiliated compounds, genes and enzymes. The curation process is tapping widescale resources which allow integrating genomics, proteomics and metabolomics data into one information hub, e.g. MetaCyc. The most recent version of MetaCyc (release 18.1) contains 2203 pathways and close to 12.000 enzymatic reactions associated with about 10.000 enzymes and genes. The pathways are experimentally elucidated and supported by 41.532 citations. The ongoing effort to improve, complement and expand the existing scope of curated pathways is crucial for the importance of metabolic databases to serve as significant reference and research tools for plant metabolism researchers. The primary goal of metabolic databases to capture the universe of primary and specialized (secondary) metabolism has been extended towards the presentation of plasticity and interconnectivity of biochemical networks. MetaCyc has been the principal matrix for many Pathway/Genome Database's (PGDB's) which have been predicted based on their sequenced genomes. Using MetaCyc as reference database, SolCyc, a collection of single-species databases of primarily Solanacea (tomato, pepper, eggplant, petunia) and close relatives such as **coffee** has been created which comprises and displays the fraction of metabolism applicable for the respective individual organism. The cellular overview permits overlaying large-scale functional genome data such as gene expression, proteomic, and metabolite profiling data onto the metabolic map to illustrate the results in a broad metabolic context.

Upcoming Meetings of interest to the ICGN community

- 12th Solanaceae Genomics Network SOL Meeting, Bordeaux, France, 2015 http://www.solgenomics.net
- 9th ICGN Coffee Genomics Workshop at XXIV Plant and Animal Genome (PAG) Meeting, San Diego, California, January 9-13, 2016 http://www.intlpag.org/
- 26th ASIC International Conference on Coffee Science, Yunnan, China, 2016 http://www.asic-cafe.org/
- 4th World Coffee Conference, Addis Ababa, Ethiopia, February 22-23, 2016. http://www.ico.org/wconference.asp?section=Meetings