Pan-genome analysis of *Theobroma cacao* reveals new genes and provides new insights into the diversity of the species.

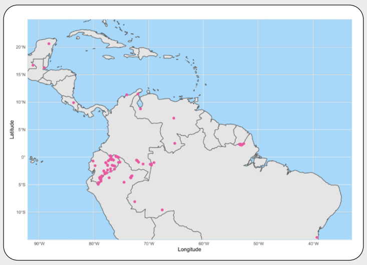
X. Argout[1,2], G. Droc[1,2], O. Fouet[1,2], A. Lemainque[3], B. Rhoné[1,2], G. Loor[4], C. Lanaud[1,2]

[1] *CIRAD, UMR AGAP Institut, Montpellier, France*
[2] *AGAP, Univ. Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France*
[3] *Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France*
[4] *Instituto Nacional de Investigaciones Agropecurias, INIAP, Quito, Ecuador*

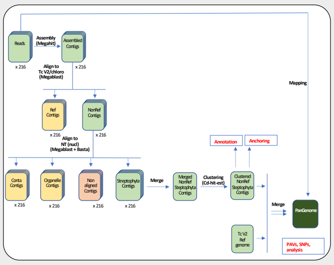
While two reference genomes of *T. cacao* are available (B97-61/B2 and Matina1-6), a reference sequence cannot capture the entire gene content of a species owing to structural variants. In the recent years, high-throughput resequencing data of *T. cacao* genotypes have provided tools to discover allelic variants in the species or characterize the level of expression of genes, but much of the genotype-specific information is often lost by direct mapping of short sequence reads onto a single reference genome. For *T. cacao*, very little is known about the complete genomic content across the species.

To address this issue, we fully re-sequenced (58X) a collection of cocoa tree representative of the current known genetic diversity and built a pan-genome of *Theobroma cacao.* We comprehensively investigated gene PAVs inside the diversity of the species and captured new protein coding genes not included in the reference genomes. We also examined how the processes of diversification of the *Theobroma cacao* genetic groups have shaped their gene content and investigated putative protein-coding genes under selection during the domestication process of the Criollo group.

❖ Plant material and pan-genome construction



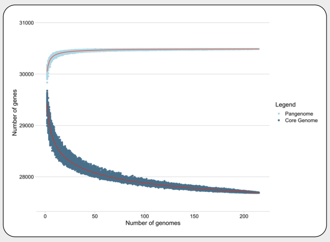
Geographic distribution of the 216 accessions



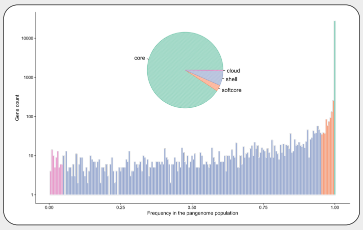
Bioinformatic pipeline

❖ Assembly, Gene prediction, Core and dispensable genome

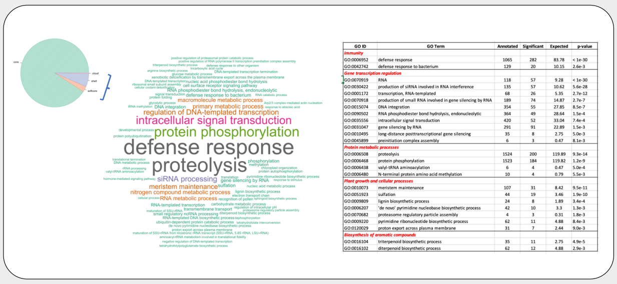
- The <u>non reference genome</u> dataset resulting from the assembly of 216 *T. cacao* accessions comprised **48.2 Mbp** and **1,407** candidate protein-coding gene models.
- The <u>*Theobroma cacao* pan-genome</u>, including the B97-61/B2 reference and non-reference genome, has a total size of **372.9 Mbp** and contained **30,489** protein-coding gene models.
- A vast majority of the genes, **27,687** (90,8%), are core genes, shared by all accessions while **2802** are variable genes.



Pan-genome saturation curve



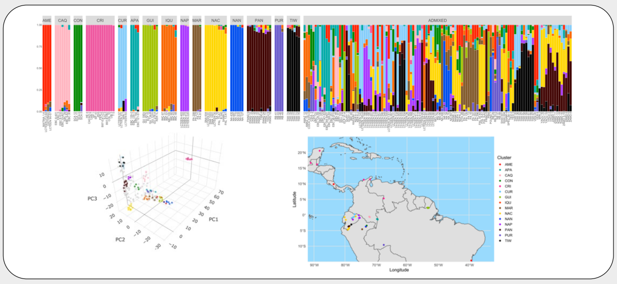
Core and dispensable genome distribution



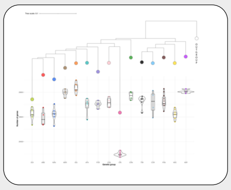
GO ID	GO Term	Annotated	Significant	Expected	p-value
Immunity					
GO:0006952	defense response	1065	282	83.78	< 1e-30
GO:0042742	defense response to bacterium	129	20	10.15	2.6e-3
Gene transcription regulation					
GO:0070919	RNA	118	57	9.28	< 1e-30
GO:0030422	production of siRNA involved in RNA interference	135	57	10.62	5.6e-28
GO:0001172	transcription, RNA-templated	68	26	5.35	2.7e-12
GO:0070918	production of small RNA involved in gene silencing by RNA	189	74	14.87	2.7e-7
GO:0015074	DNA integration	354	55	27.85	8.5e-7
GO:0090502	RNA phosphodiester bond hydrolysis, endonucleolytic	364	49	28.64	1.5e-4
GO:0035556	intracellular signal transduction	420	52	33.04	7.4e-4
GO:0031047	gene silencing by RNA	291	91	22.89	1.5e-3
GO:0010495	long-distance posttranscriptional gene silencing	35	8	2.75	5.0e-3
GO:0045899	preinitiation complex assembly	6	3	0.47	8.1e-3
Protein metabolic processes					
GO:0006508	proteolysis	1524	200	119.89	9.3e-14
GO:0006468	protein phosphorylation	1523	184	119.82	1.2e-9
GO:0006438	valyl-tRNA aminoacylation	6	4	0.47	5.0e-4
GO:0006480	N-terminal protein amino acid methylation	10	4	0.79	5.5e-3
Plant growth and cellular processes					
GO:0010073	meristem maintenance	107	31	8.42	9.5e-11
GO:0051923	sulfation	44	19	3.46	1.9e-10
GO:0009809	lignin biosynthetic process	24	8	1.89	3.4e-4
GO:0006207	'de novo' pyrimidine nucleobase biosynthetic process	42	10	3.3	1.3e-3
GO:0070682	proteasome regulatory particle assembly	4	3	0.31	1.8e-3
GO:0009220	pyrimidine ribonucleotide biosynthetic process	62	11	4.88	8.4e-3
GO:0120029	proton export across plasma membrane	31	7	2.44	9.0e-3
Biosynthesis of aromatic compounds					
GO:0016104	triterpenoid biosynthetic process	35	11	2.75	4.9e-5
GO:0016102	diterpenoid biosynthetic process	62	12	4.88	2.9e-3

Gene Ontology (GO) enrichment of protein-coding genes of the variable genome

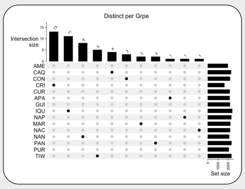
❖ Selection of gene PAVs during T. cacao diversification



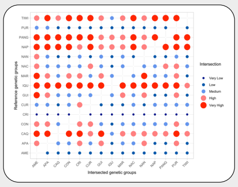
Genetic diversity of the population based on gene PAVs



Distribution of gene PAVs among the 15 groups



Gene space specificity for the 15 groups

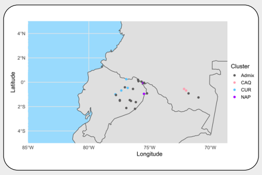


Intersection between the gene space of the 15 groups

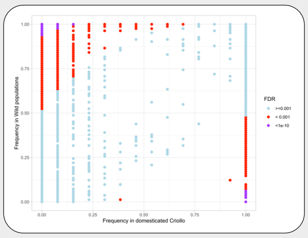
Key findings :

- 15 genetic groups
- Accessions of Iquitos and Napo ancestry have the highest gene number
- Gene content of the Criollo group shows gene loss during domestication process
- Very few genes are specific to a genetic group
- Iquitos gene space overlap almost all genetic groups
- Criollo is most closely related to Caquetá group and Caquetá group contains more than 95% of the genes found in Criollo group

❖ Protein-coding genes under selection during Criollo domestication



Geographic distribution of putative Criollo wild relative population



Gene frequencies comparison between Criollo accessions and accessions native from the putative center of origin of the domesticated Criollo.

Key findings :

- 71 genes selected during Criollo domestication
- “Defense response” enrichment in non selected genes

❖ Conclusion / recommendation

- The *Theobroma cacao* pan-genome adds depth and completeness to the reference genomes, and is useful for future biological discovery
- The understanding of gene PAV, which contribute to trait variation, can support applications for the development of new varieties

Contact : xavier.argout@cirad.fr