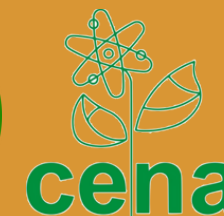
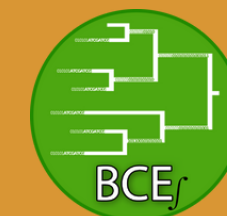


Gerando e explorando recursos genômicos computacionais para cana-de-açúcar.

Prof. Dr. Diego Mauricio Riaño-Pachón
diego.riano@cena.usp.br – <https://labbbces.netlify.app/>
Laboratório de Biologia Computacional, Evolutiva e de Sistemas
Centro de Energia Nuclear na Agricultura
Universidade de São Paulo



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa



Equipe



Diego Mauricio Riaño-Pachón

Assistant Professor (MS3.2) in
Computational, Evolutionary and
Systems Biology



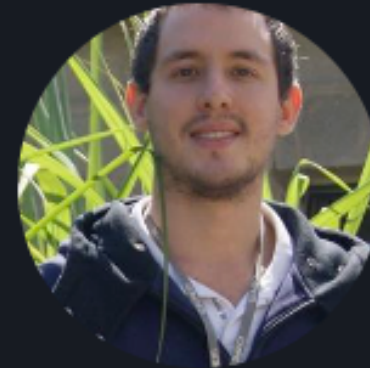
Felipe Vaz Peres

Master student - PPG - Ciências
CENA/USP - Long ncRNAs in
sugarcane



**Hellen Regina Silvestre
Silva**

Undergrad student - Biological
Sciences(UNESP - Rio Claro)



**Renato Augusto Correa
dos Santos**

Postdoct Gene regulatory
networks in grasses



Isabella Vitoria Alonso

Undergrad student - Biotechnology
(UFSCar)



Bianca Viana

Undergrad student - Agronomy
(ESALQ/USP)



Bianca Sagiorato

Undergrad student - Biotechnology
(UFSCar)



Guilherme Furlan

Technical assistant



**Isabella Gallego
Rendón**

Undergrad student - Biology (USP)
- LLMs



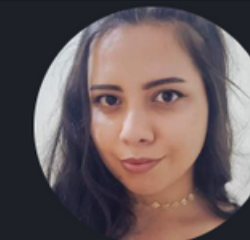
**Jorge Mario Muñoz
Perez**

PhD student - PPG - Ciências
CENA/USP - Sugarcane Co-
expression networks



**Beatriz Rodrigues
Estevam**

Undergrad student - Biological
Sciences (ESALQ/USP)



**Danielli Teixeira Brito
Machado**

Undergrad student - Biotechnology
(UFSCar)



Bianca Pastos

Undergrad student - Computer
Science (UNESP Rio Claro -IGCE)



**Arthur Shuzo Owtake
Cardoso**

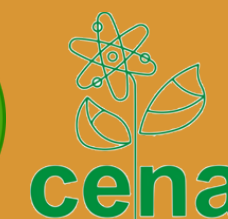
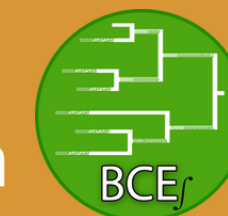
Undergrad student - Biotechnology
(USP)

Colaboradores

Marcos Buckeridge
Marcelo Menossi



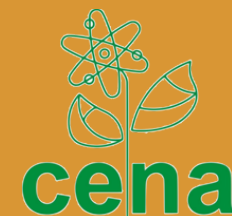
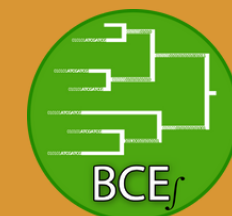
Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa



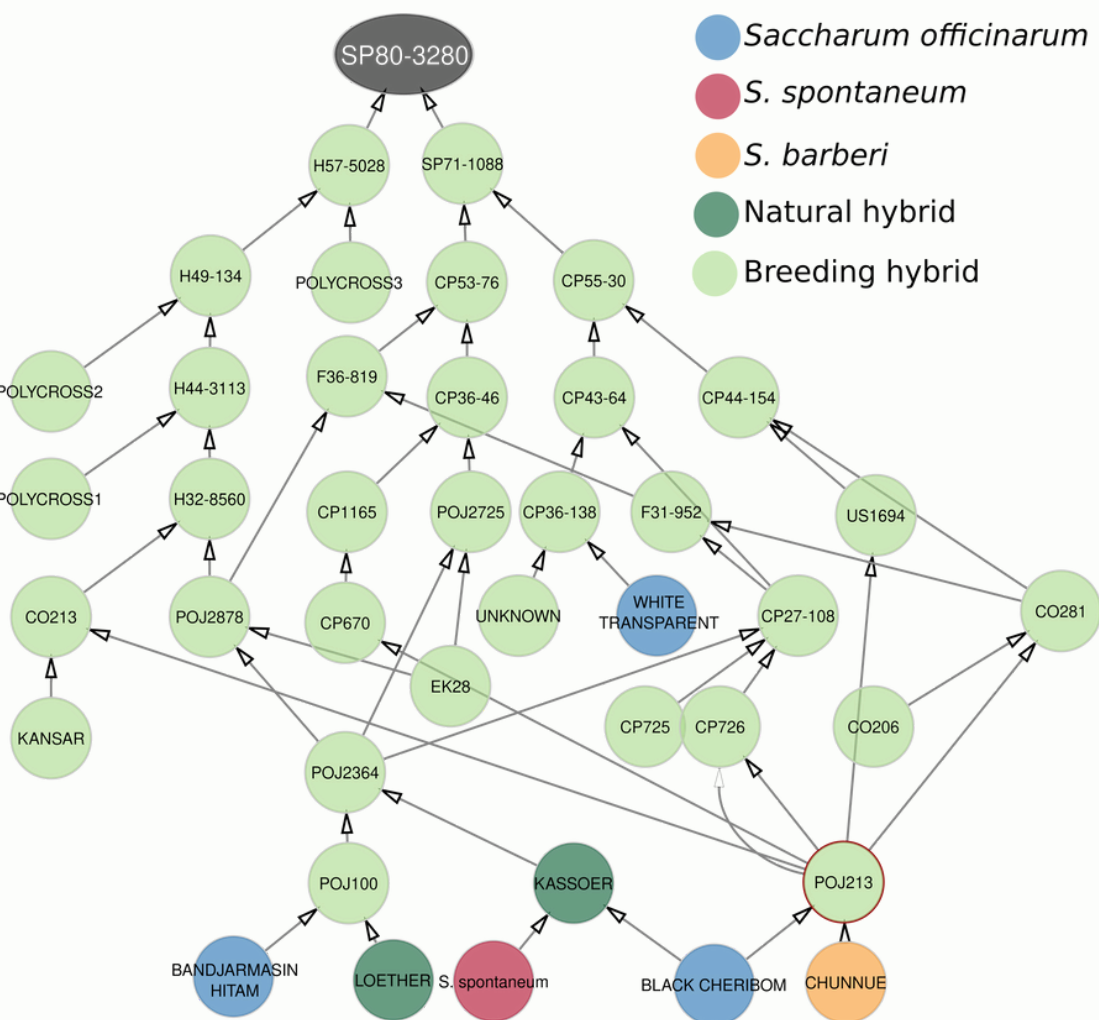
- 1. Sequenciamento, montagem e anotação do genoma da cana-de-açúcar variedade SP80-3280**
- 2. Captura da variabilidade genética de cana a partir de dados de transcriptoma**
- 3. Desenvolvimento de plataformas para mineração de dados genômicos, de expressão e de apoio a edição do genoma de cana**
- 4. Desenvolvimento de software para apoiar a edição do genoma de cana**
- 5. Redes de co-expressão em cana e sorgo**



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa

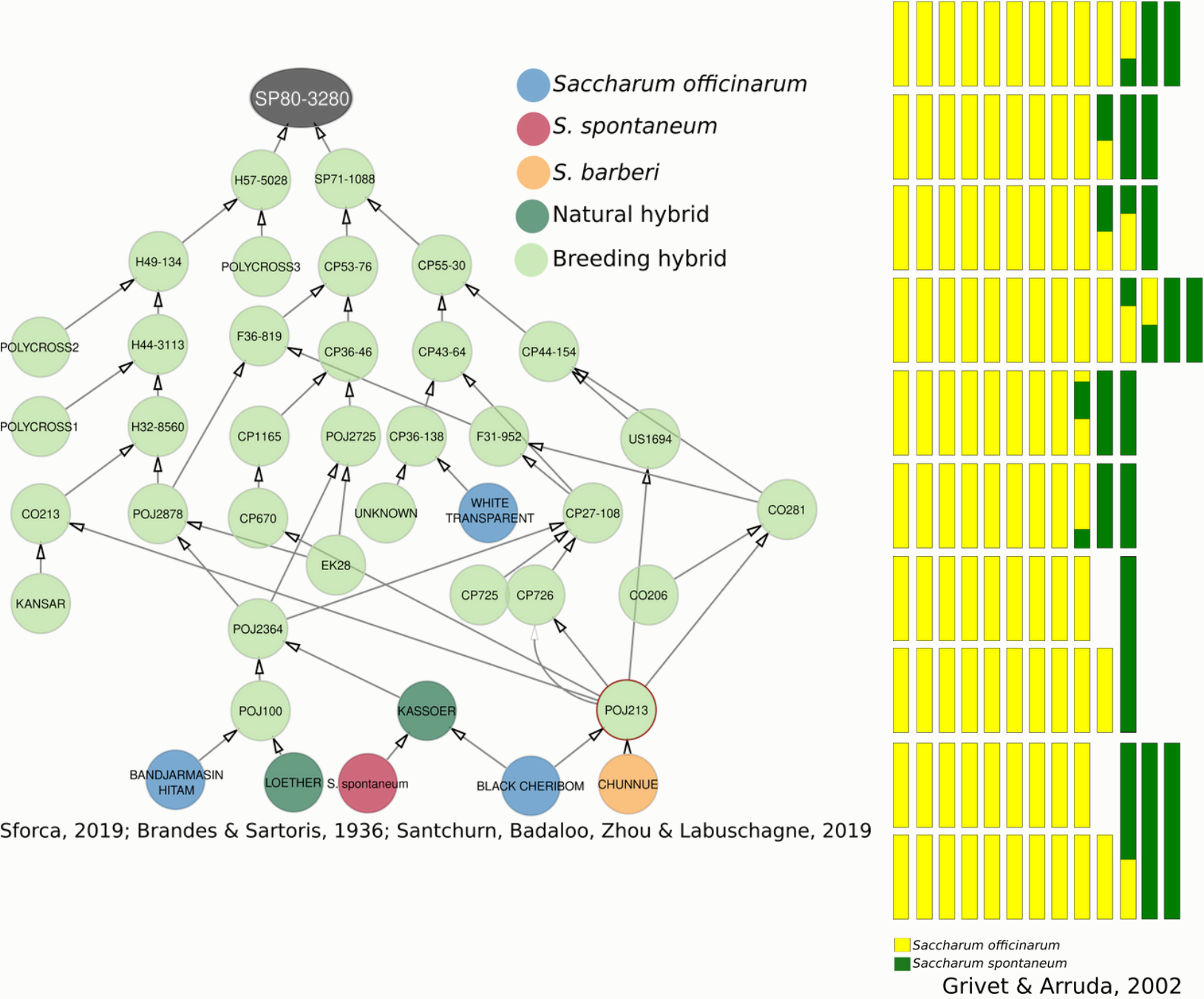


O melhoramento da cana-de-açúcar e seu genoma

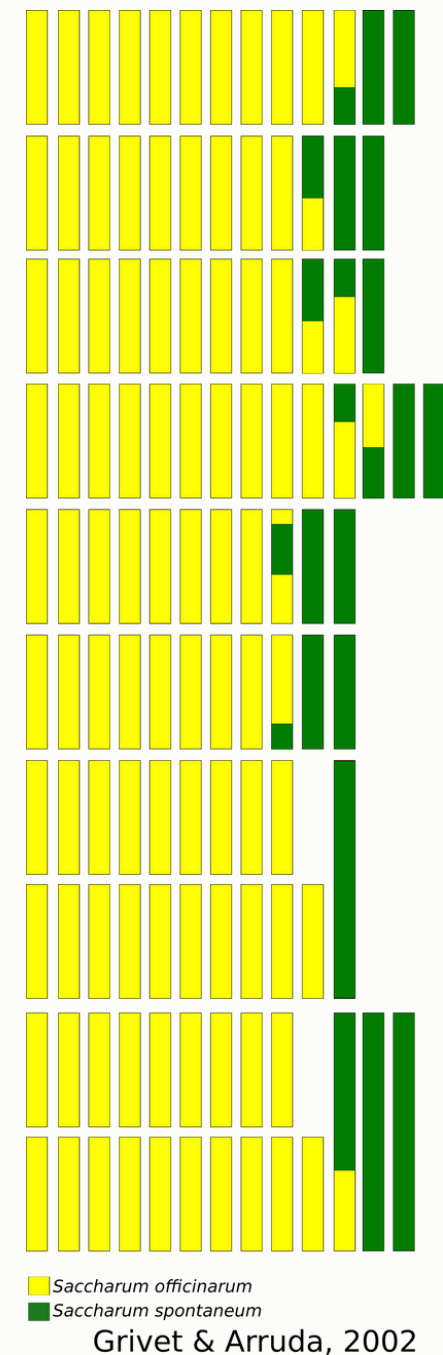
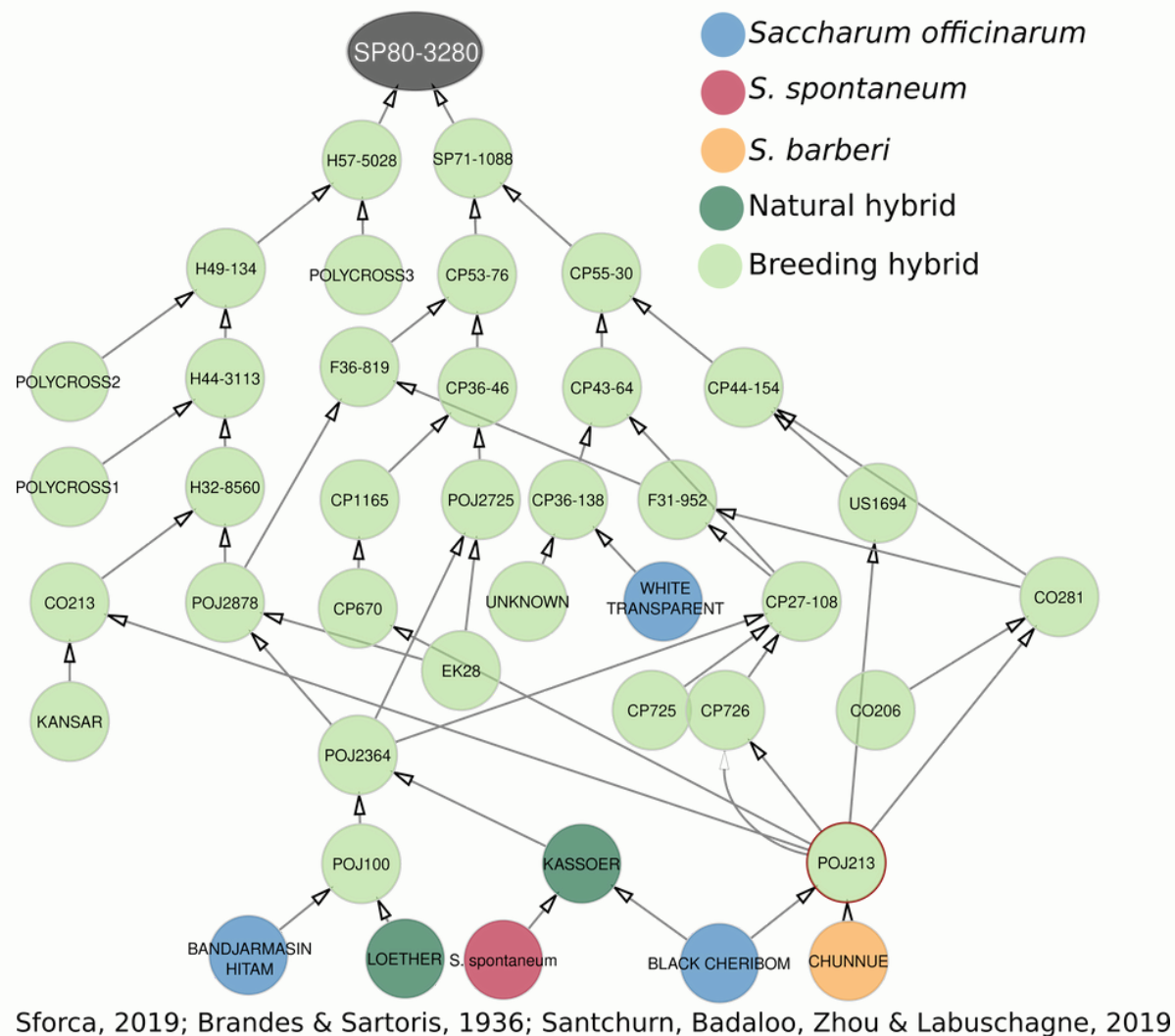


Sforca, 2019; Brandes & Sartoris, 1936; Santchurn, Badaloo, Zhou & Labuschagne, 2019

O melhoramento da cana-de-açúcar e seu genoma



O melhoramento da cana-de-açúcar e seu genoma



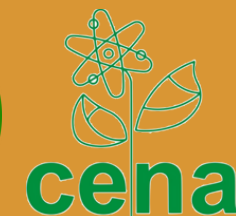
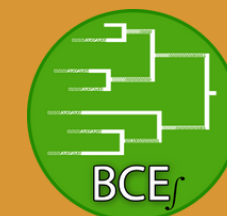
Available genome assemblies

Assembly	Contigs	N50 (Kbp)	Size (Mbp)	Genes	Technologies	Publication
SP80-3280-CTBE	199,028	8.4	1,167	153,078	TruSeq Synthetic Long Reads	Riaño-Pachón & Mattiello, 2017
SP80-3280-IQ/USP	450,608	13.2	4,259	374,774	TruSeq Synthetic Long Reads	Souza, et al., 2019
R570-MTP-CIRAD	5,708	116.7	530	41,223	BAC, PacBio RSII	Garsmeur, et al., 2018
R570-STP-CIRAD	211	45,576.6	427	25,316	BAC, PacBio RSII	Garsmeur, et al., 2018
CC-01-1940-CENICANA	35,089	34,980.0	904	68,260	PacBio RSII, HiC, genetic map	Trujillo-Montenegro, et al., 2021

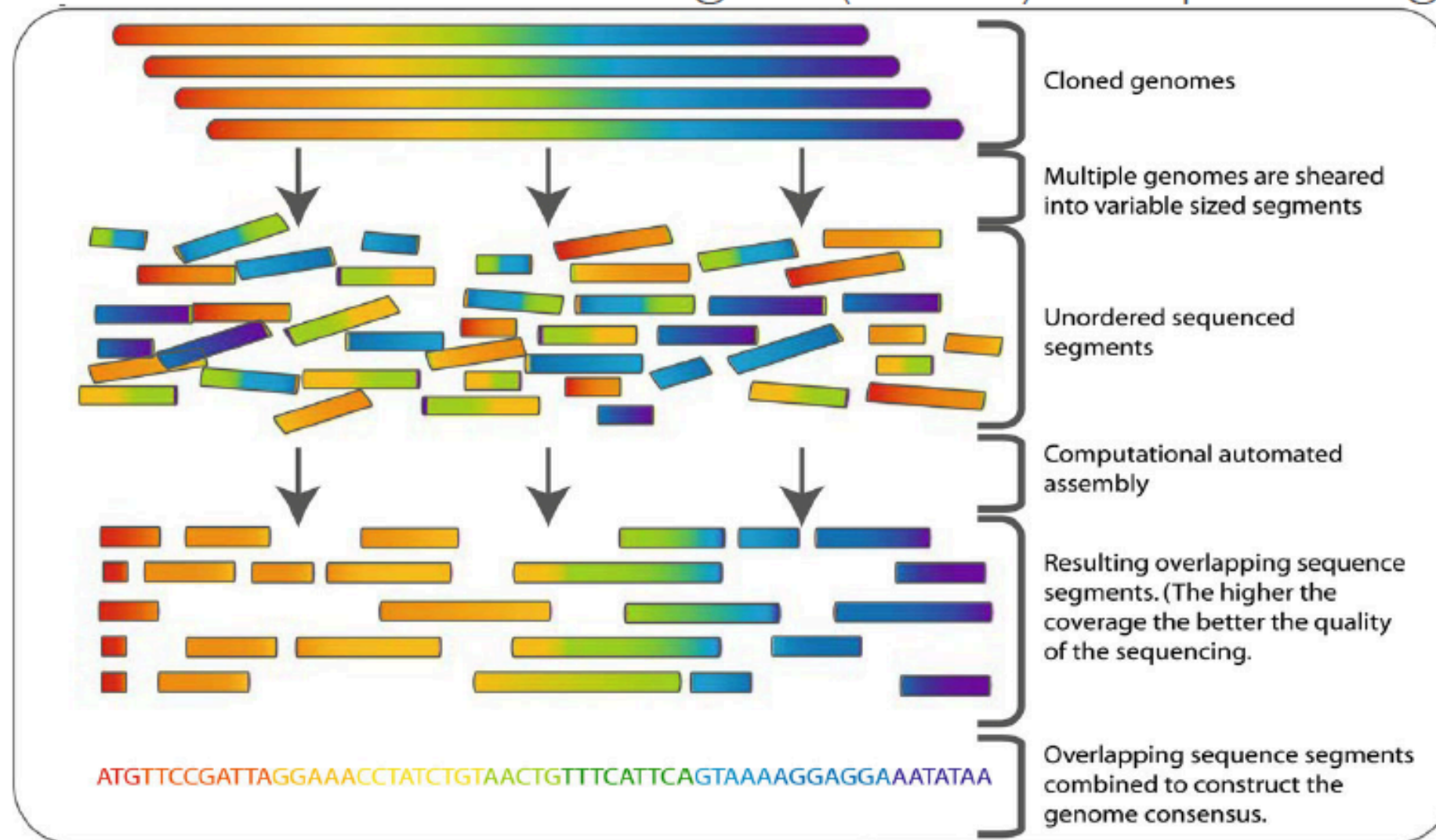
1. As nossas primeiras montagens (draft) do genoma da SP80-3280



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa



Whole Genome Shotgun (WGS) Sequencing



Genome

Reads

Contigs

Scaffolds



http://bch709.plantgenomicslab.org/Genome_assembly_annotation/index.html

PacBio HiFi. Error rate ~0.05%

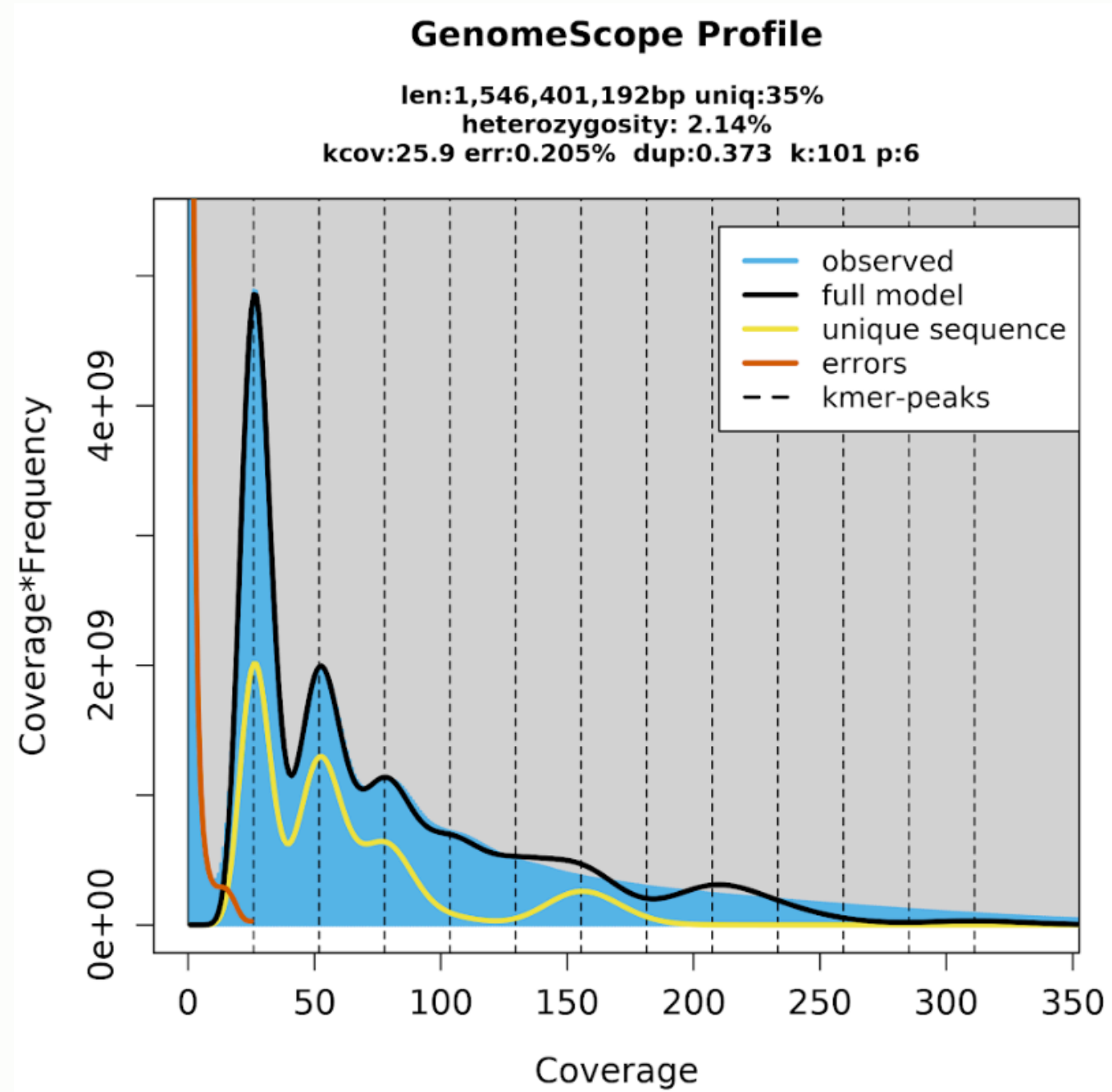
312 Gbp ou ~312x do genoma monoploide, ou 31x each haplofase

Illumina HiC. Error rate ~ 0.08%

484 Gbp ou ~484x do genoma monoploide, ou 48x each haplofase

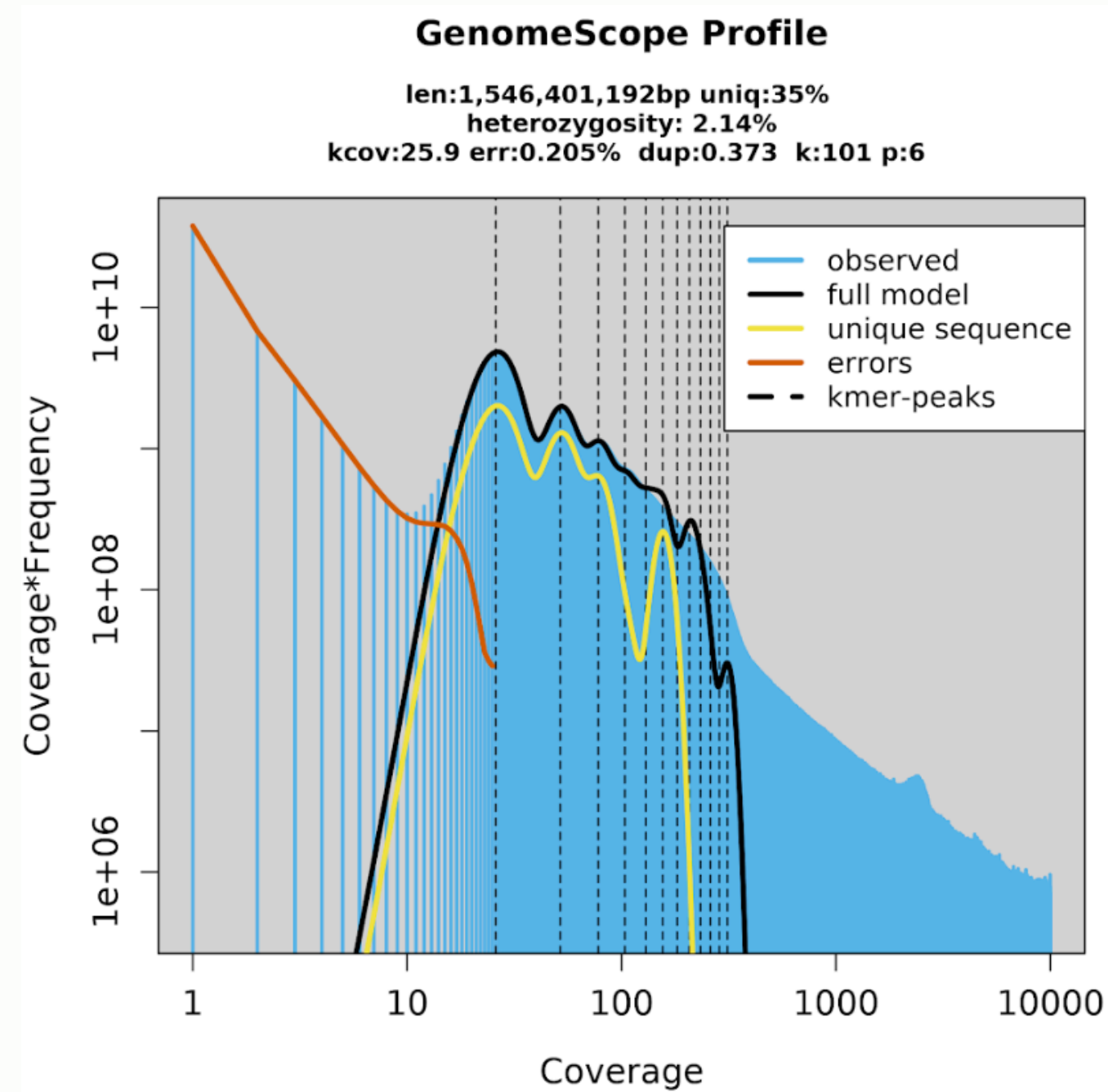
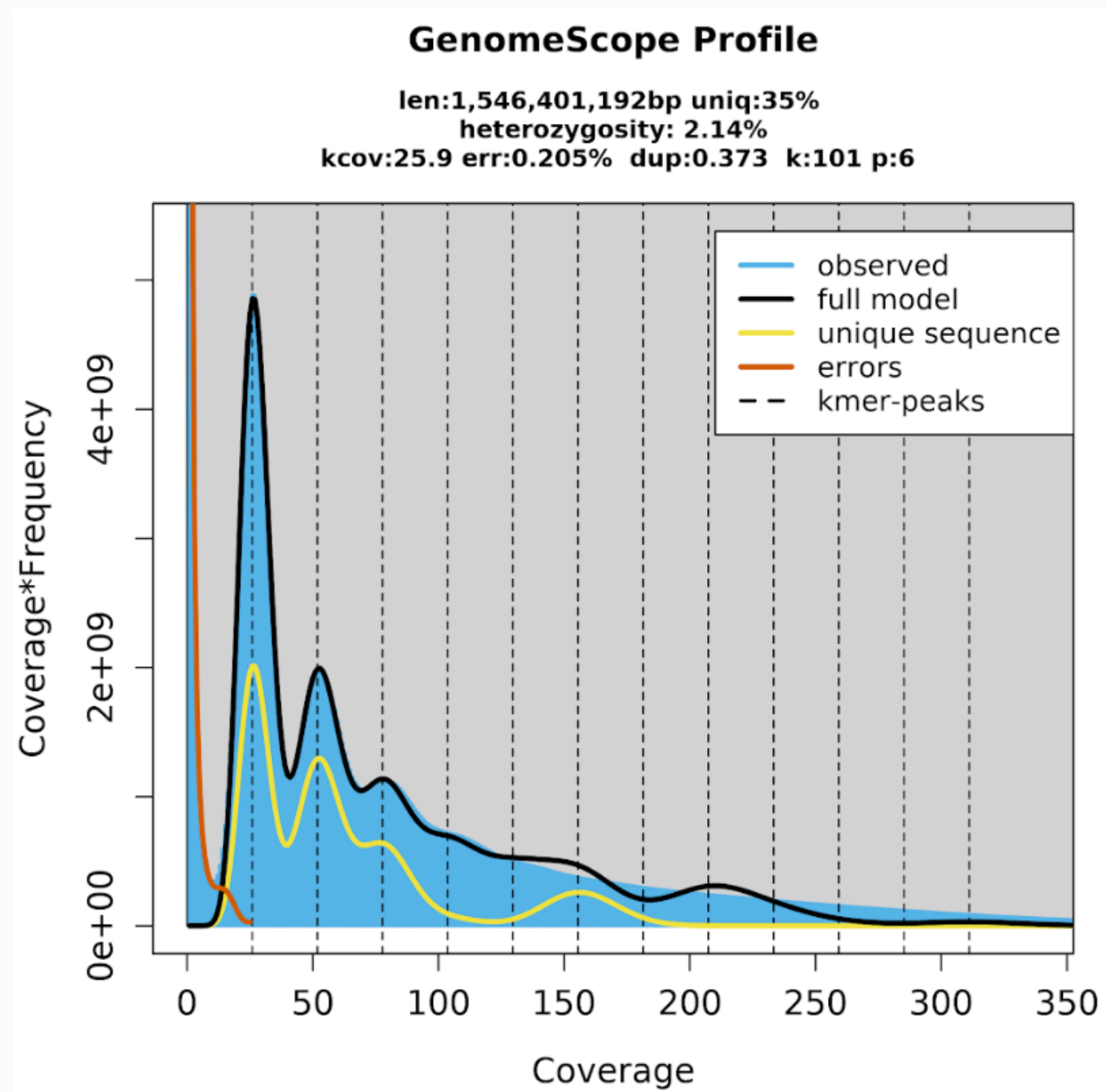
Genome statistics with K-mers

PacBio HiFi. Error rate ~0.05% 312 Gbp ou ~312x do genoma monoploide, ou 31x each haplofase

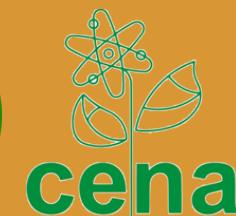
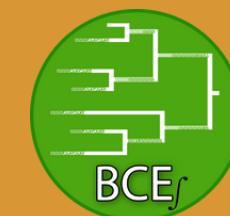


Genome statistics with K-mers

PacBio HiFi. Error rate ~0.05% 312 Gbp ou ~312x do genoma monoploide, ou 31x each haplofase

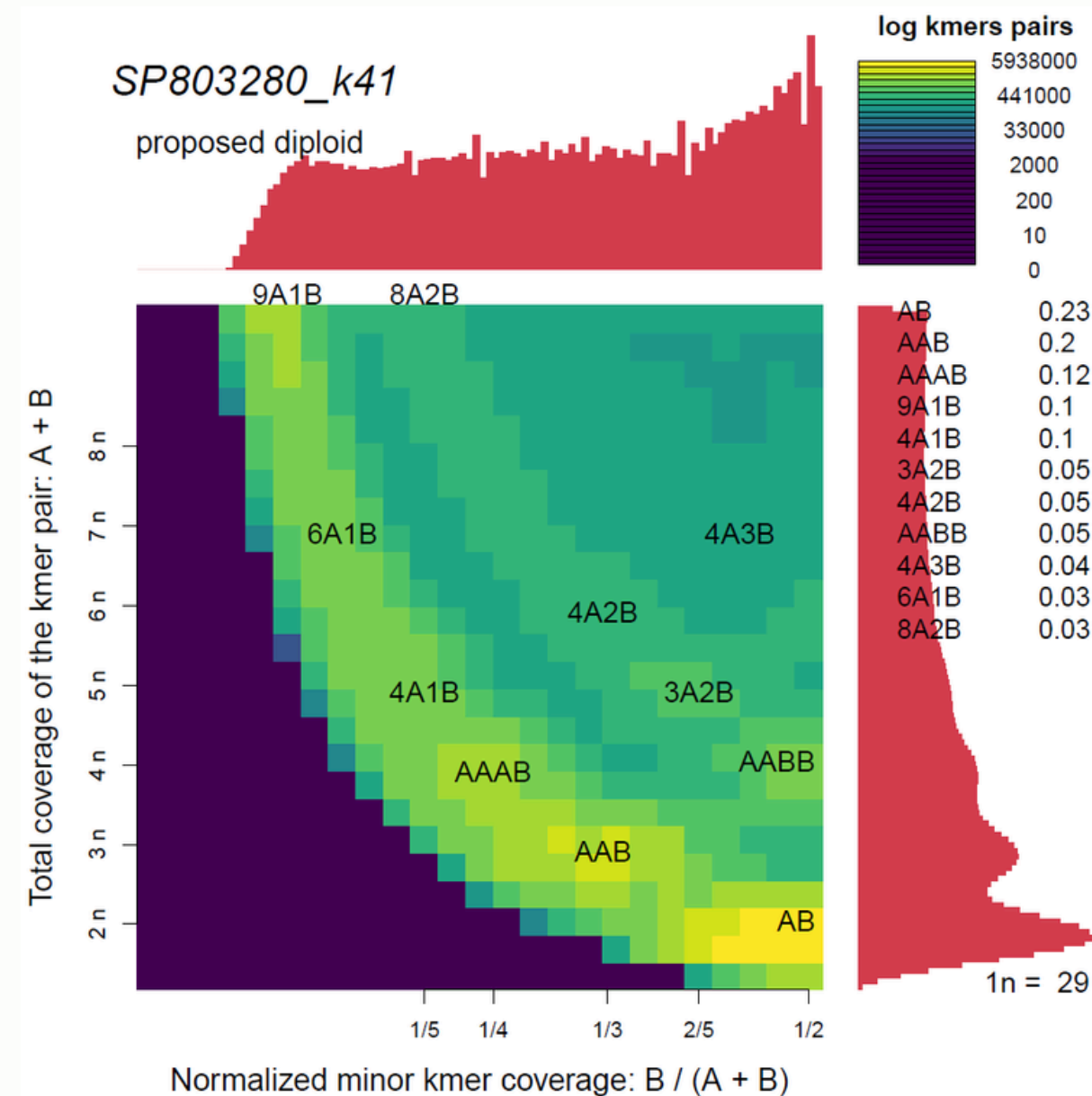
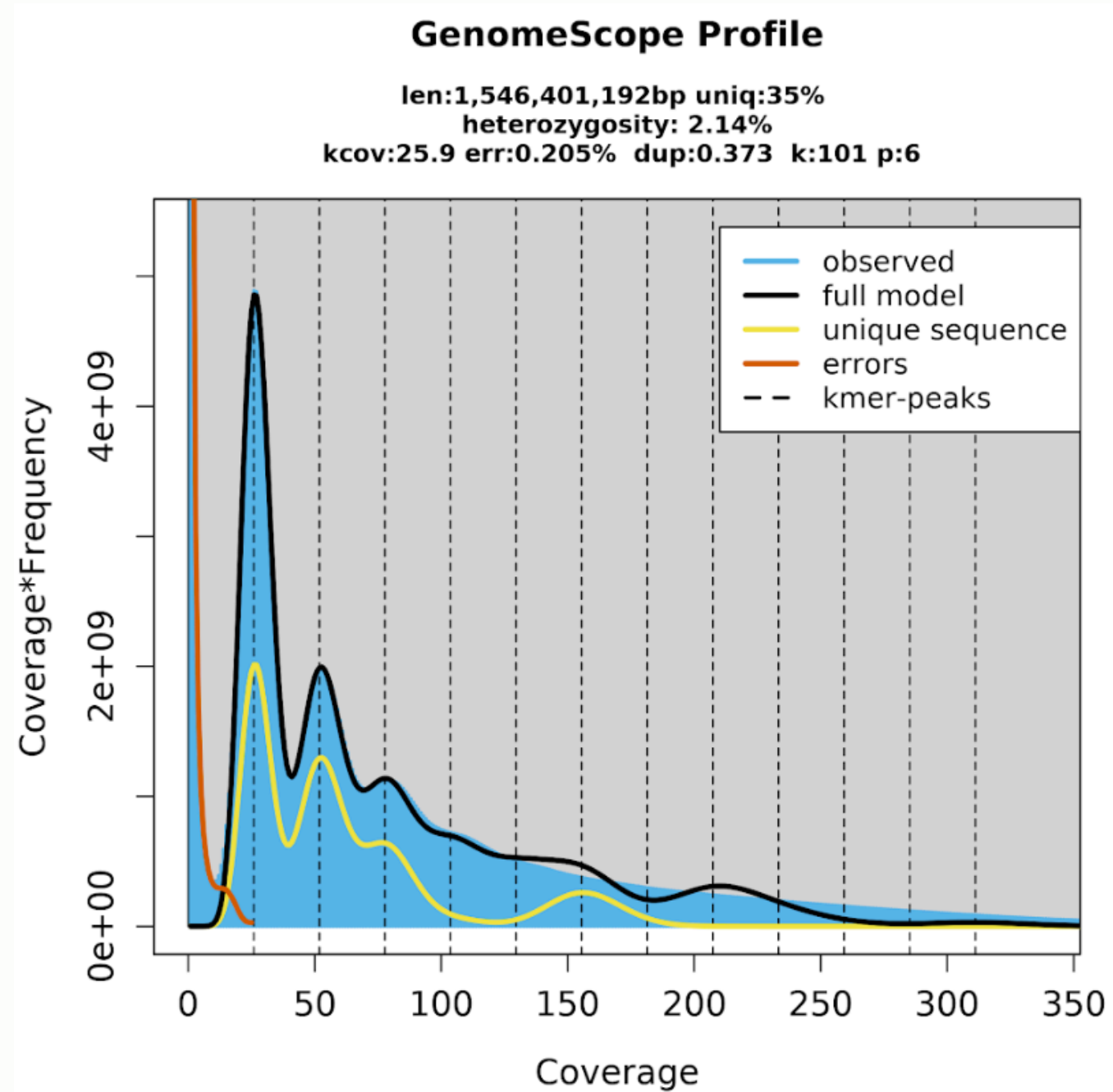


Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa

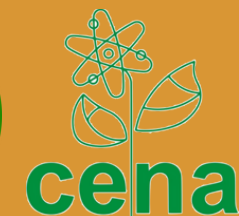
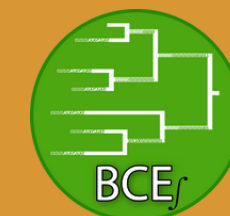


Genome statistics with K-mers

PacBio HiFi. Error rate ~0.05% 312 Gbp ou ~312x do genoma monoploide, ou 31x each haplofase



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa



Assembly	Contigs	N50 (Kbp)	Size (Mbp)	Genes
SP80-3280-CTBE	199,028	8.4	1,167	153,078
SP80-3280-IQ/USP	450,608	13.2	4,259	374,774
R570-MTP-CIRAD	5,708	116.7	530	41,223
R570-STP-CIRAD	211	45,576.6	427	25,316
CC-01-1940-CENICANA	35,089	34,980.0	904	68,260

As métricas de contiguidade são muito melhores, do que quaisquer das montagens anteriores poliploides

ASM	Contigs > 3 Kbp	Contigs > 50 Kbp	N50 (Kbp) > 3 Kbp	N50 (Kbp) > 50 Kbp	Largest contig (Kbp)	Size (Mbp) > 3 Kbp	Size (Mbp) > 50 Kbp
B0.1	152.5	35.3	198.6	505.6	4,791.7	11,322.6	7,877.5
B1.2	48.5	36.6	211.8	224.7	3,451.1	6,822.6	6,473.2
B2.p_utg	128.2	29.5	645.0	4,038.1	63,914.8	11,799.0	8,656.2

Assembly	Contigs	N50 (Kbp)	Size (Mbp)	Genes
SP80-3280-CTBE	199,028	8.4	1,167	153,078
SP80-3280-IQ/USP	450,608	13.2	4,259	374,774
R570-MTP-CIRAD	5,708	116.7	530	41,223
R570-STP-CIRAD	211	45,576.6	427	25,316
CC-01-1940-CENICANA	35,089	34,980.0	904	68,260

As métricas de contiguidade são muito melhores, do que quaisquer das montagens anteriores poliploides

ASM	Contigs > 3 Kbp	Contigs > 50 Kbp	N50 (Kbp) > 3 Kbp	N50 (Kbp) > 50 Kbp	Largest contig (Kbp)	Size (Mbp) > 3 Kbp	Size (Mbp) > 50 Kbp
B0.1	152.5	35.3	198.6	505.6	4,791.7	11,322.6	7,877.5
B1.2	48.5	36.6	211.8	224.7	3,451.1	6,822.6	6,473.2
B2.p_utg	128.2	29.5	645.0	4,038.1	63,914.8	11,799.0	8,656.2

Assembly	Contigs	N50 (Kbp)	Size (Mbp)	Genes
SP80-3280-CTBE	199,028	8.4	1,167	153,078
SP80-3280-IQ/USP	450,608	13.2	4,259	374,774
R570-MTP-CIRAD	5,708	116.7	530	41,223
R570-STP-CIRAD	211	45,576.6	427	25,316
CC-01-1940-CENICANA	35,089	34,980.0	904	68,260

As métricas de contiguidade são muito melhores, do que quaisquer das montagens anteriores poliploides

AUMENTO DE 1.7X - 480X

ASM	Contigs > 3 Kbp	Contigs > 50 Kbp	N50 (Kbp) > 3 Kbp	N50 (Kbp) > 50 Kbp	Largest contig (Kbp)	Size (Mbp) > 3 Kbp	Size (Mbp) > 50 Kbp
B0.1	152.5	35.3	198.6	505.6	4,791.7	11,322.6	7,877.5
B1.2	48.5	36.6	211.8	224.7	3,451.1	6,822.6	6,473.2
B2.p_utg	128.2	29.5	645.0	4,038.1	63,914.8	11,799.0	8,656.2

Assembly	Contigs	N50 (Kbp)	Size (Mbp)	Genes
SP80-3280-CTBE	199,028	8.4	1,167	153,078
SP80-3280-IQ/USP	450,608	13.2	4,259	374,774
R570-MTP-CIRAD	5,708	116.7	530	41,223
R570-STP-CIRAD	211	45,576.6	427	25,316
CC-01-1940-CENICANA	35,089	34,980.0	904	68,260

As métricas de contiguidade são muito melhores, do que quaisquer das montagens anteriores poliploides

ASM	Contigs > 3 Kbp	Contigs > 50 Kbp	N50 (Kbp) > 3 Kbp	N50 (Kbp) > 50 Kbp	Largest contig (Kbp)	Size (Mbp) > 3 Kbp	Size (Mbp) > 50 Kbp
B0.1	152.5	35.3	198.6	505.6	4,791.7	11,322.6	7,877.5
B1.2	48.5	36.6	211.8	224.7	3,451.1	6,822.6	6,473.2
B2.p_utg	128.2	29.5	645.0	4,038.1	63,914.8	11,799.0	8,656.2

Assembly	Contigs	N50 (Kbp)	Size (Mbp)	Genes
SP80-3280-CTBE	199,028	8.4	1,167	153,078
SP80-3280-IQ/USP	450,608	13.2	4,259	374,774
R570-MTP-CIRAD	5,708	116.7	530	41,223
R570-STP-CIRAD	211	45,576.6	427	25,316
CC-01-1940-CENICANA	35,089	34,980.0	904	68,260

As métricas de contiguidade são muito melhores, do que quaisquer das montagens anteriores poliploides

AUMENTO DE 1.5X - 22X

ASM	Contigs > 3 Kbp	Contigs > 50 Kbp	N50 (Kbp) > 3 Kbp	N50 (Kbp) > 50 Kbp	Largest contig (Kbp)	Size (Mbp) > 3 Kbp	Size (Mbp) > 50 Kbp
B0.1	152.5	35.3	198.6	505.6	4,791.7	11,322.6	7,877.5
B1.2	48.5	36.6	211.8	224.7	3,451.1	6,822.6	6,473.2
B2.p_utg	128.2	29.5	645.0	4,038.1	63,914.8	11,799.0	8,656.2

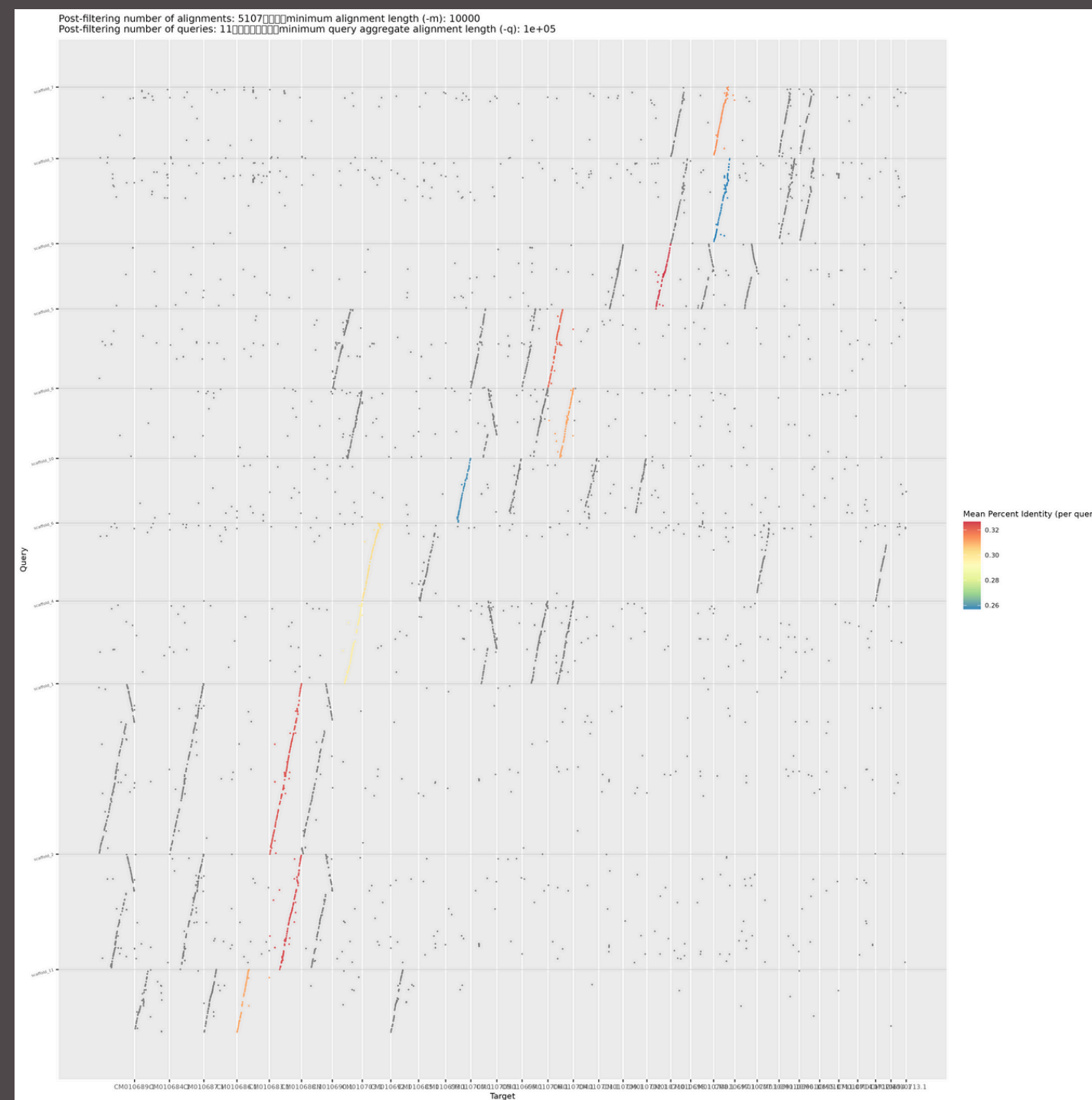
Assembly	Contigs	N50 (Kbp)	Size (Mbp)	Genes
SP80-3280-CTBE	199,028	8.4	1,167	153,078
SP80-3280-IQ/USP	450,608	13.2	4,259	374,774
R570-MTP-CIRAD	5,708	116.7	530	41,223
R570-STP-CIRAD	211	45,576.6	427	25,316
CC-01-1940-CENICANA	35,089	34,980.0	904	68,260

As métricas de contiguidade são muito melhores, do que quaisquer das montagens anteriores poliploides

Será um cromossomo?

ASM	Contigs > 3 Kbp	Contigs > 50 Kbp	N50 (Kbp) > 3 Kbp	N50 (Kbp) > 50 Kbp	Largest contig (Kbp)	Size (Mbp) > 3 Kbp	Size (Mbp) > 50 Kbp
B0.1	152.5	35.3	198.6	505.6	4,791.7	11,322.6	7,877.5
B1.2	48.5	36.6	211.8	224.7	3,451.1	6,822.6	6,473.2
B2.p_utg	128.2	29.5	645.0	4,038.1	63,914.8	11,799.0	8,656.2

**Temos cromossomos
inteiros, mas ainda
não todas as cópias
de cada
cromossomo. A
figura mostra pelo
menos duas cópias
inteiras de cada
cromossomo. Só
mostrei os 11
scaffolds maiores!**



**Comparison against *S.
spontaneum* full genome (32
chromosomes)**

**9 desses 11
cromossomos tem
milhares de
repetições
teloméricas em pelo
menos uma das
pontas!**

Work in progress!

Anotação estrutural (previsão de genes) na montagem B2.p_utg

Avaliação

Na avaliação das anotações, estabelecemos dois critérios importantes relacionados à presença e à semelhança dos genes BUSCO (4896 em total), que são genes conservados e geralmente encontrados em uma única cópia nos genomas de Poales. Os critérios são os seguintes:

- Presença de Genes BUSCO:** Verificar se os genes BUSCO estão presentes nas anotações. Isso implica que a anotação deve incluir esses genes conservados, o que é crucial para garantir a integridade da anotação genômica.
- Porcentagem de Genes BUSCO Semelhante ou Maior que nos Genomas de Comparação:** Comparar a porcentagem de genes BUSCO presentes nas anotações com a porcentagem de genes BUSCO encontrados no genoma de referência. O ideal é que a porcentagem seja igual ou maior nas anotações, indicando que a anotação está em conformidade com a expectativa de conservação desses genes em Poales.

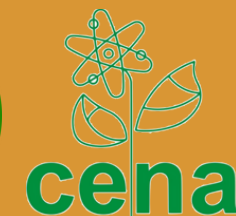
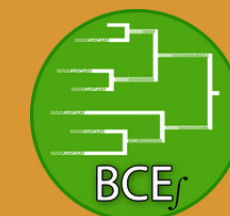
Montagem/Métodos	Total de genes preditos	Número de genes completos (%)	Número de genes duplicados (%)
IQ/USP GENOME (2019)	373869	4416 (90.19)	3511(71.71)
CTBE GENOME (2017)	153078	3704 (75.66)	1545 (31.56)
B2.p_utg GENOME	NA	4887 (99.82)	4863 (99.33)
B2.p_utg GALBA	722191	4862 (99.3)	2501 (51.1)
B2.p_utg BRAKER3	257198	4663 (95.2)	4525 (92.4)

O que é importante?

- A montagem B2.p_utg e suas previsões de genes destacam-se como a representação mais precisa do conteúdo genético do SP80-3280, com uma notável margem de superioridade com as montagens anteriores.
- Tanto o GALBA quanto o BRAKER3 são capazes de recuperar a maioria dos genes esperados (um conjunto avaliado de 4.896 genes). No entanto, o BRAKER3 se destaca por seu maior êxito na recuperação de versões alélicas, o que explica a significativa porcentagem de genes duplicados identificados.
- Essa montagem está pronta para ser explorada por nossos parceiros.

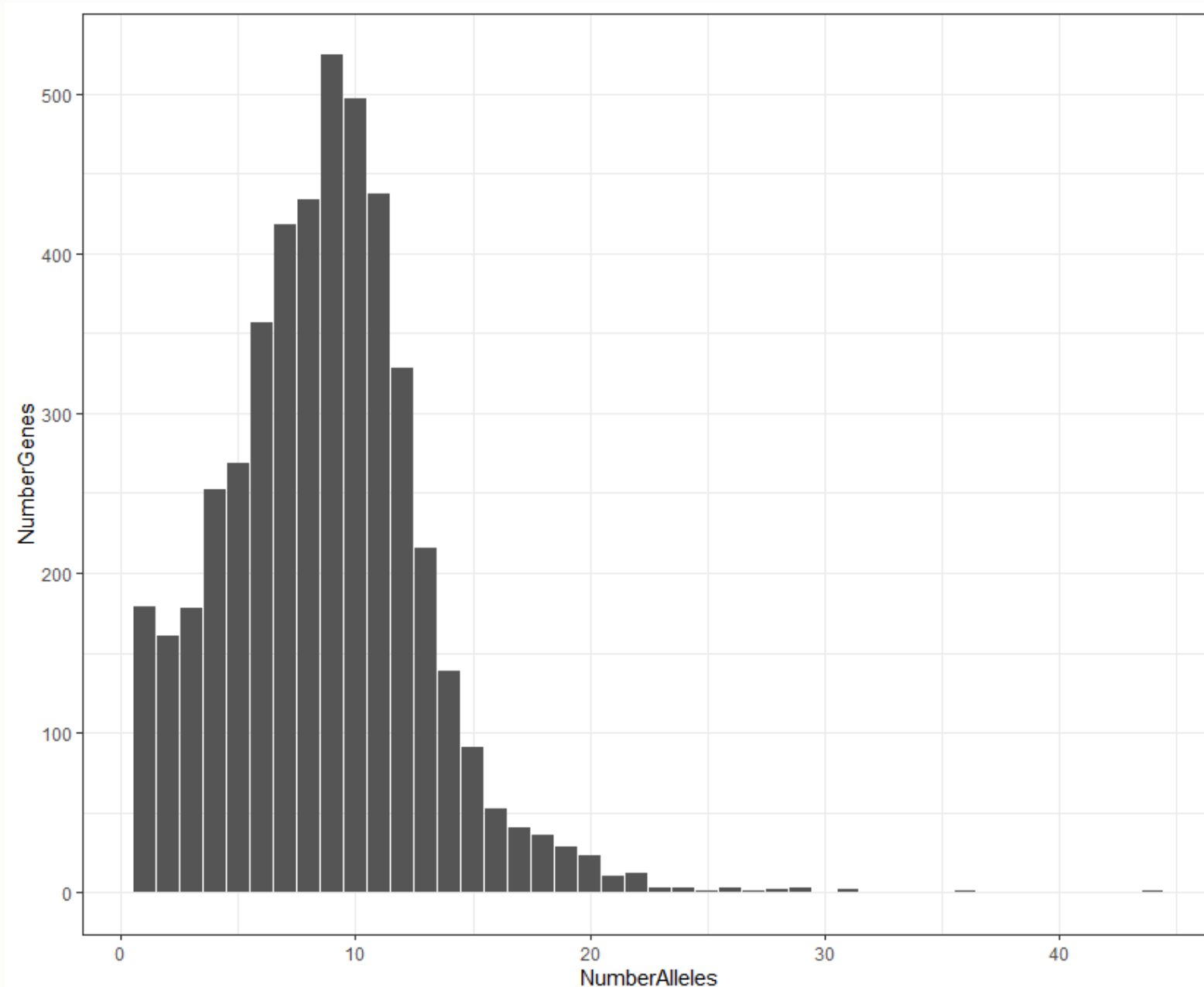


Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa



Anotação estrutural (previsão de genes) na montagem B2.p_utg

Avaliação - Número de alelos BRAKER3

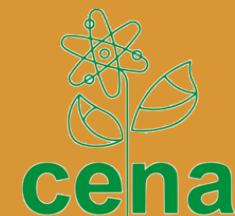
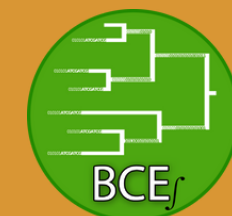


1. Dado que a previsão de genes com BRAKER3 tem sucesso na identificação de uma proporção significativa de duplicatas, que correspondem a alelos, a figura à esquerda ilustra a relação entre o número de alelos e o número de genes. Como pode ser visto, a maioria dos genes é encontrada com 10 ou 11 alelos, o que se alinha com a possível ploidia desse cultivar.

**Estamos próximos a ter a primeira
sequencia de um genoma de uma
variedade comercial brasileira
(hibrido) de cana em escala
cromossômica e poliploide!**



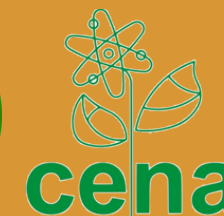
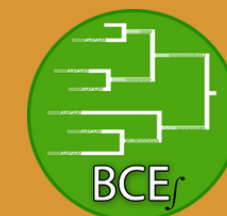
**Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa**



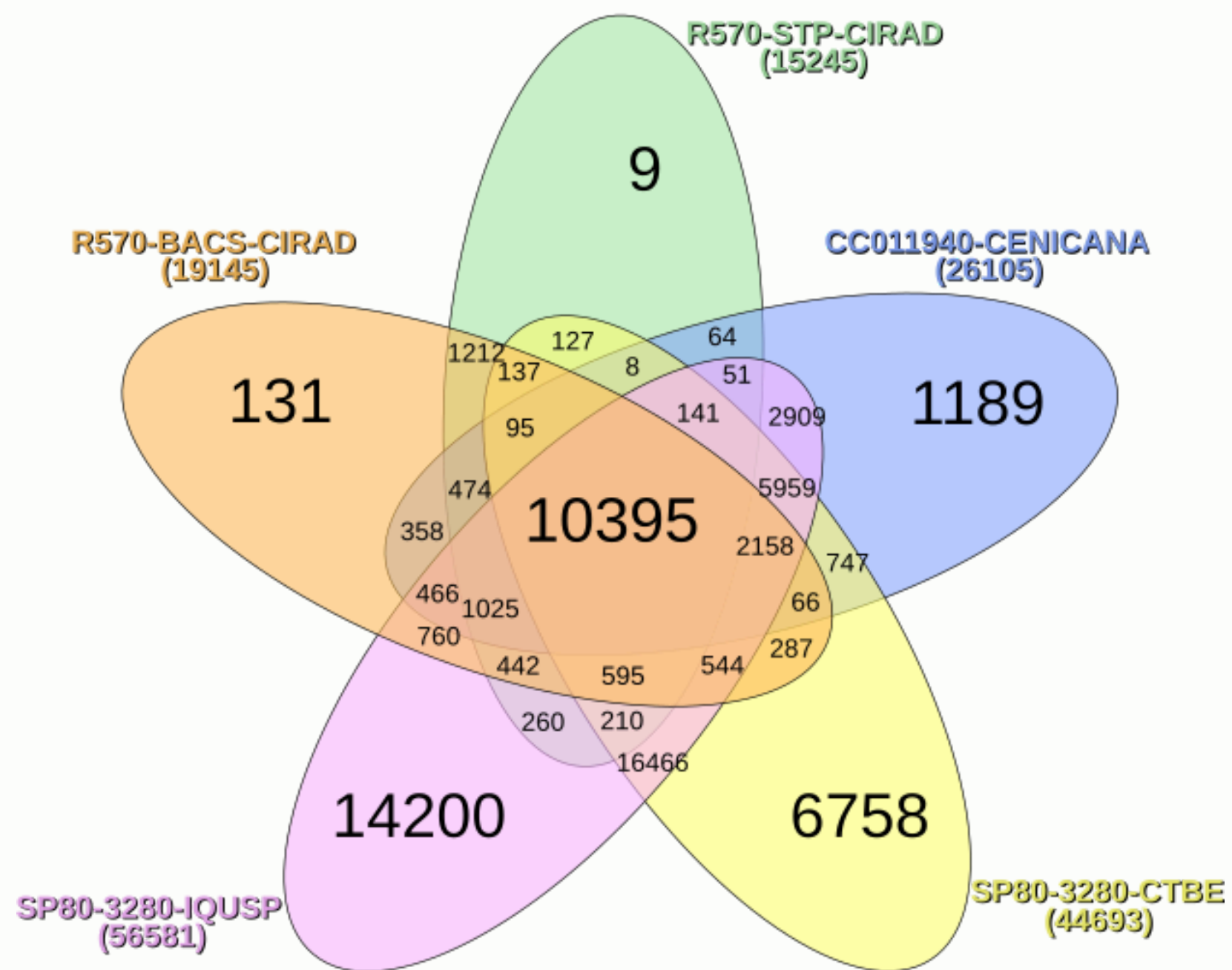
2. Capturando a variabilidade genética da cultura a partir de dados públicos de transcriptoma



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa

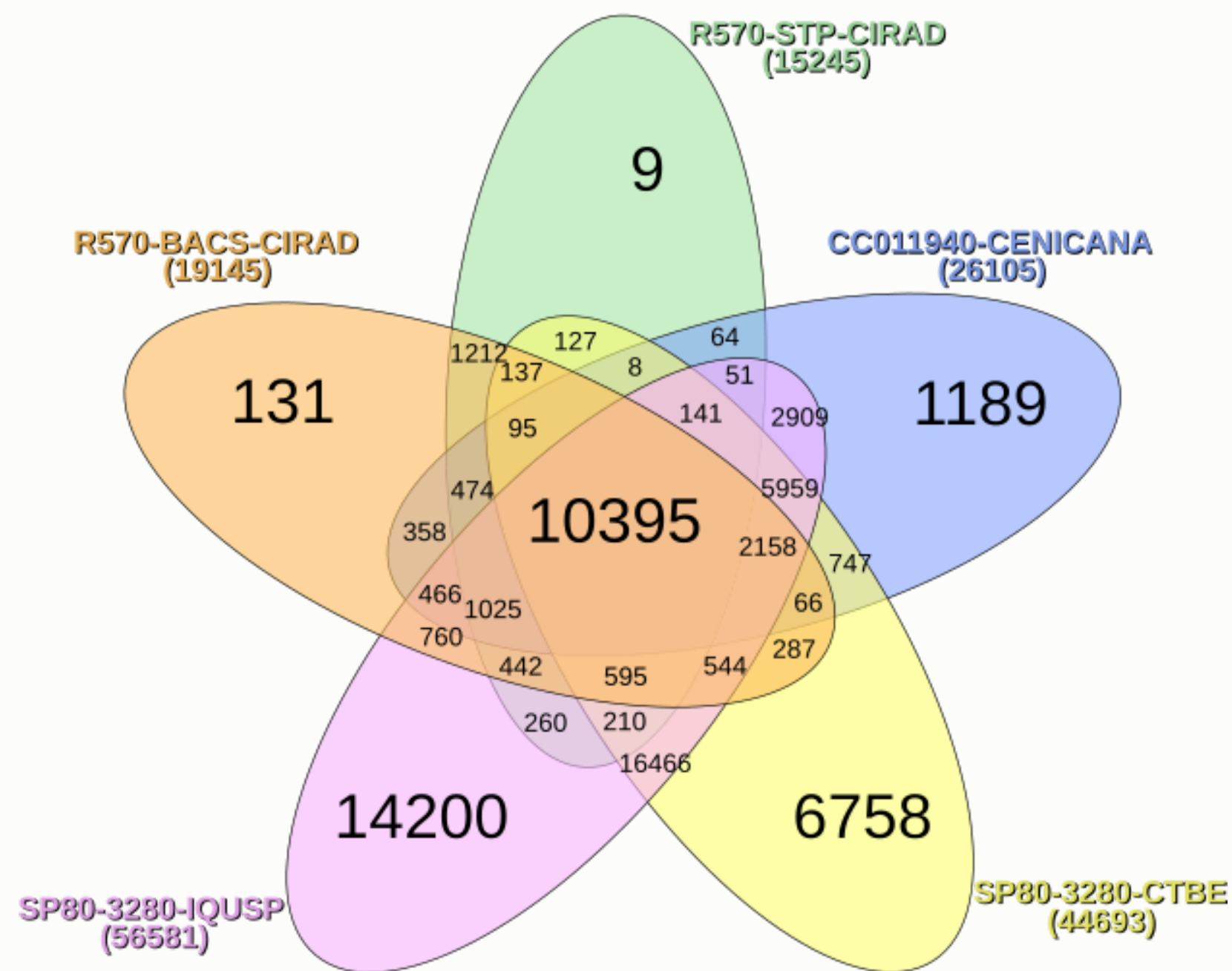


Do we have a complete sugarcane genome yet?

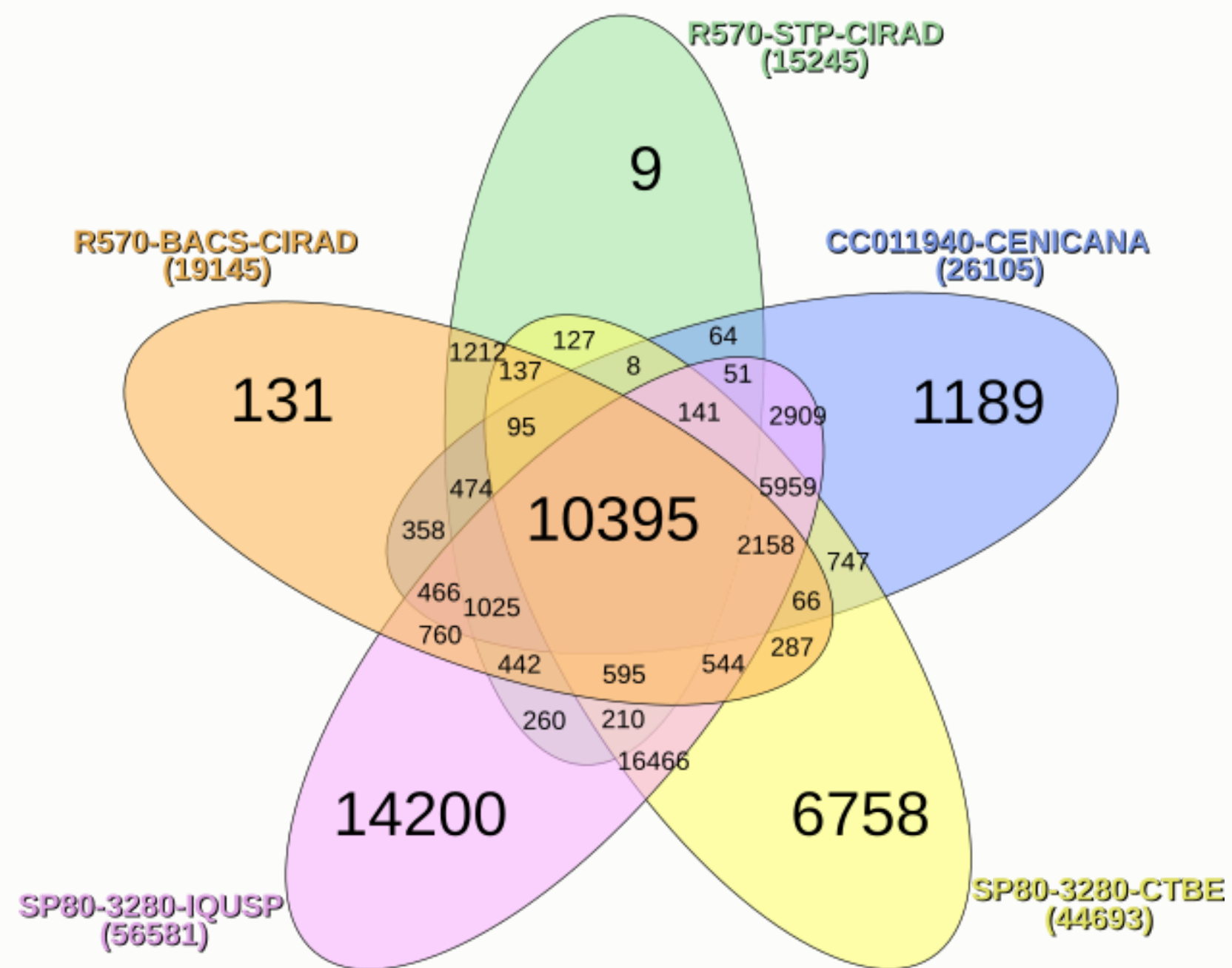


Do we have a complete sugarcane genome yet?

A resposta é não, e talvez a gente não vai ter um genoma completo, na forma de um genoma de uma única variedade!



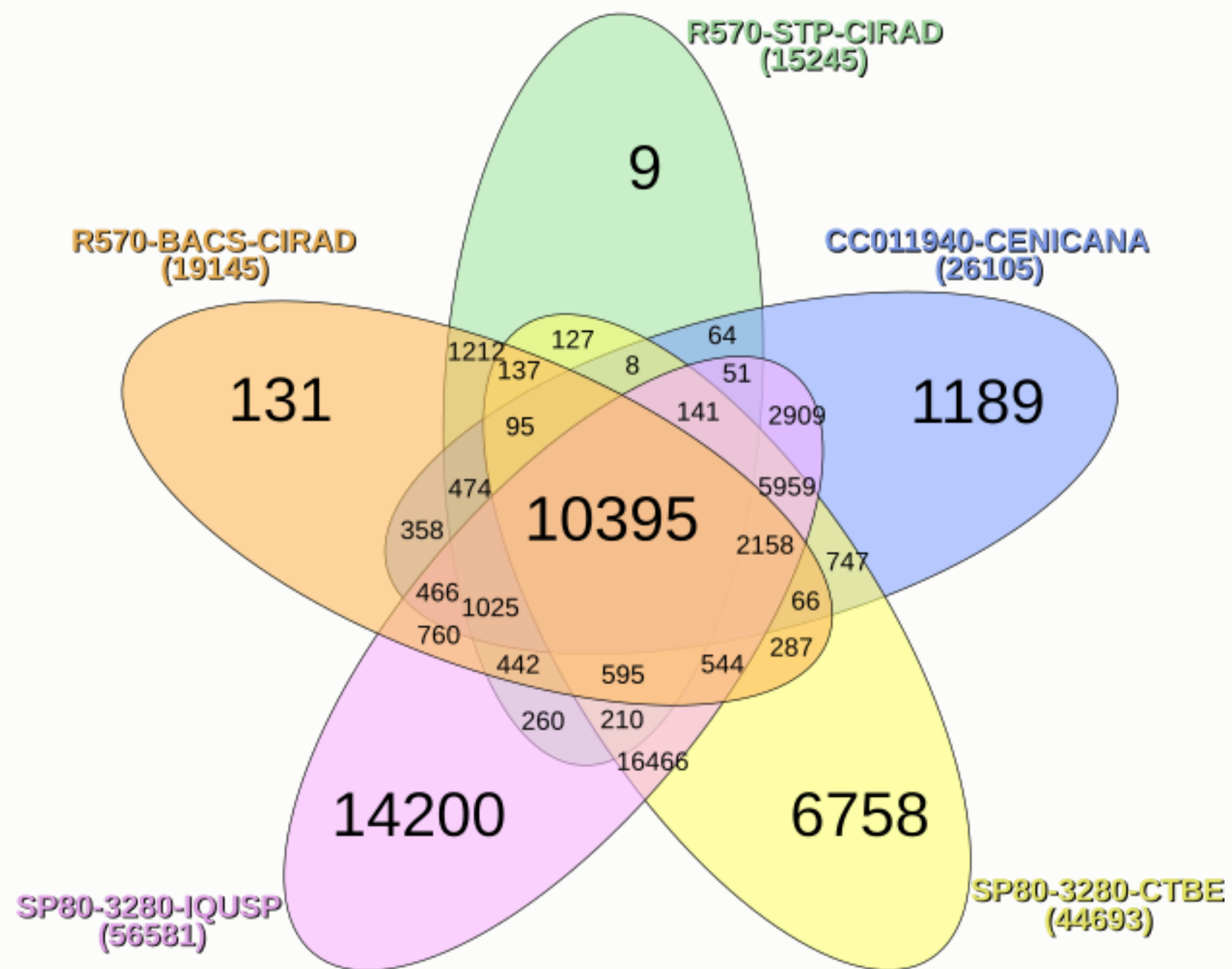
Do we have a complete sugarcane genome yet?



A resposta é não, e talvez a gente não vai ter um genoma completo, na forma de um genoma de uma única variedade!

Diferentes indivíduos/variedades/populações de uma cultura/especies tem regiões do genoma exclusivas, ou compartilhadas com alguns outros poucos.

Do we have a complete sugarcane genome yet?

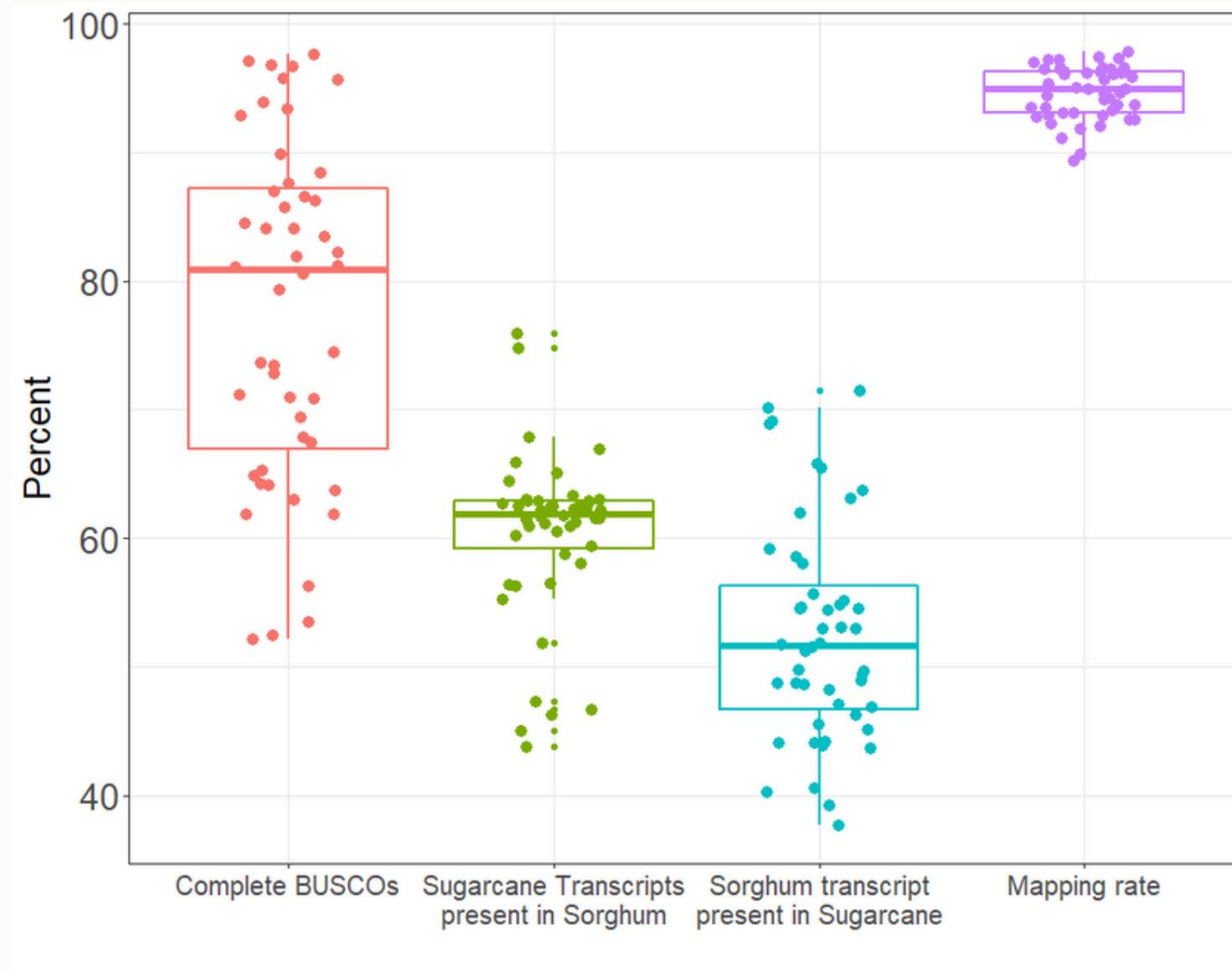


A resposta é não, e talvez a gente não vai ter um genoma completo, na forma de um genoma de uma única variedade!

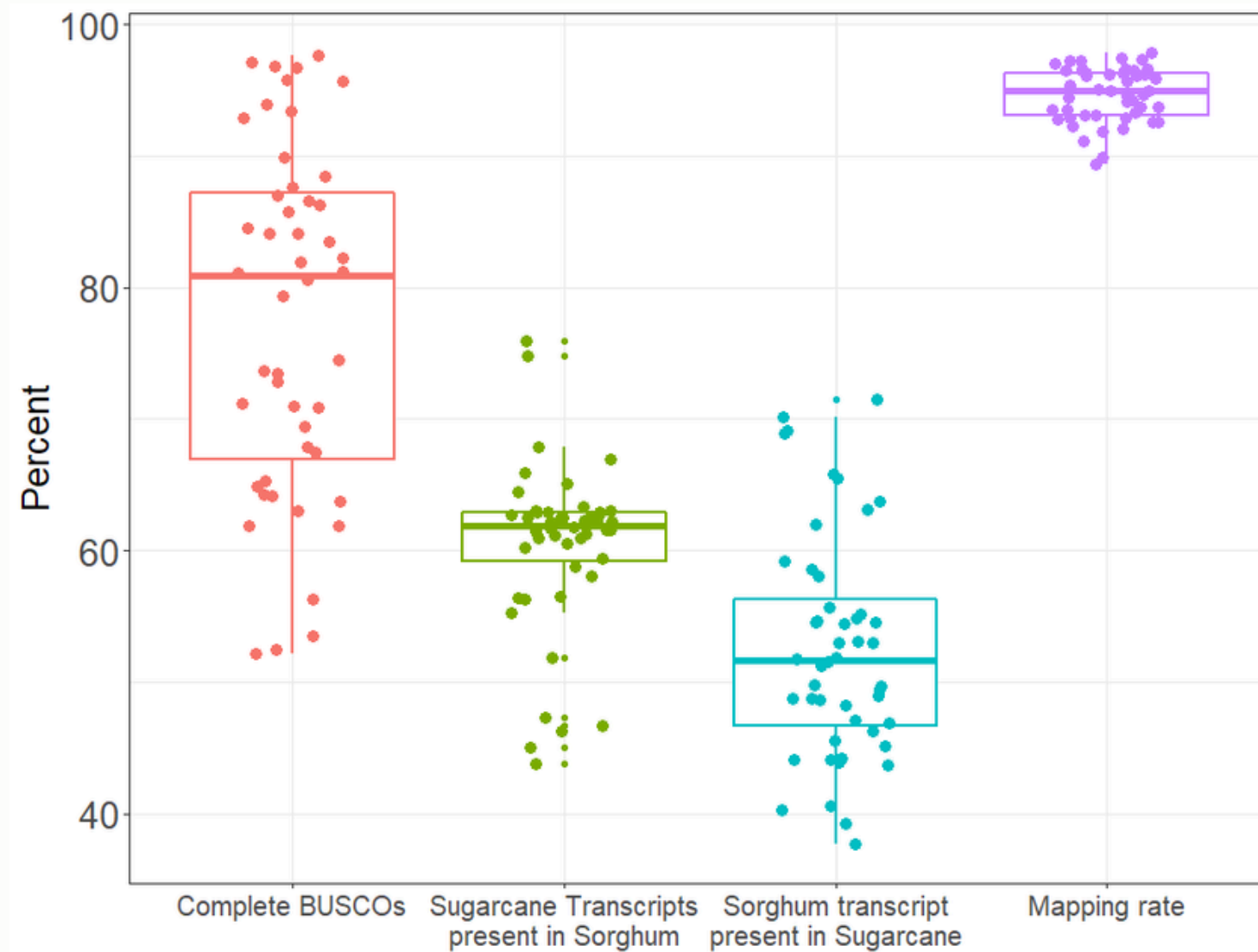
Diferentes indivíduos/variedades/populações de uma cultura/especies tem regiões do genoma exclusivas, ou compartilhadas com alguns outros poucos.

Avaliamos isso em cana explorando dados públicos de transcriptômica

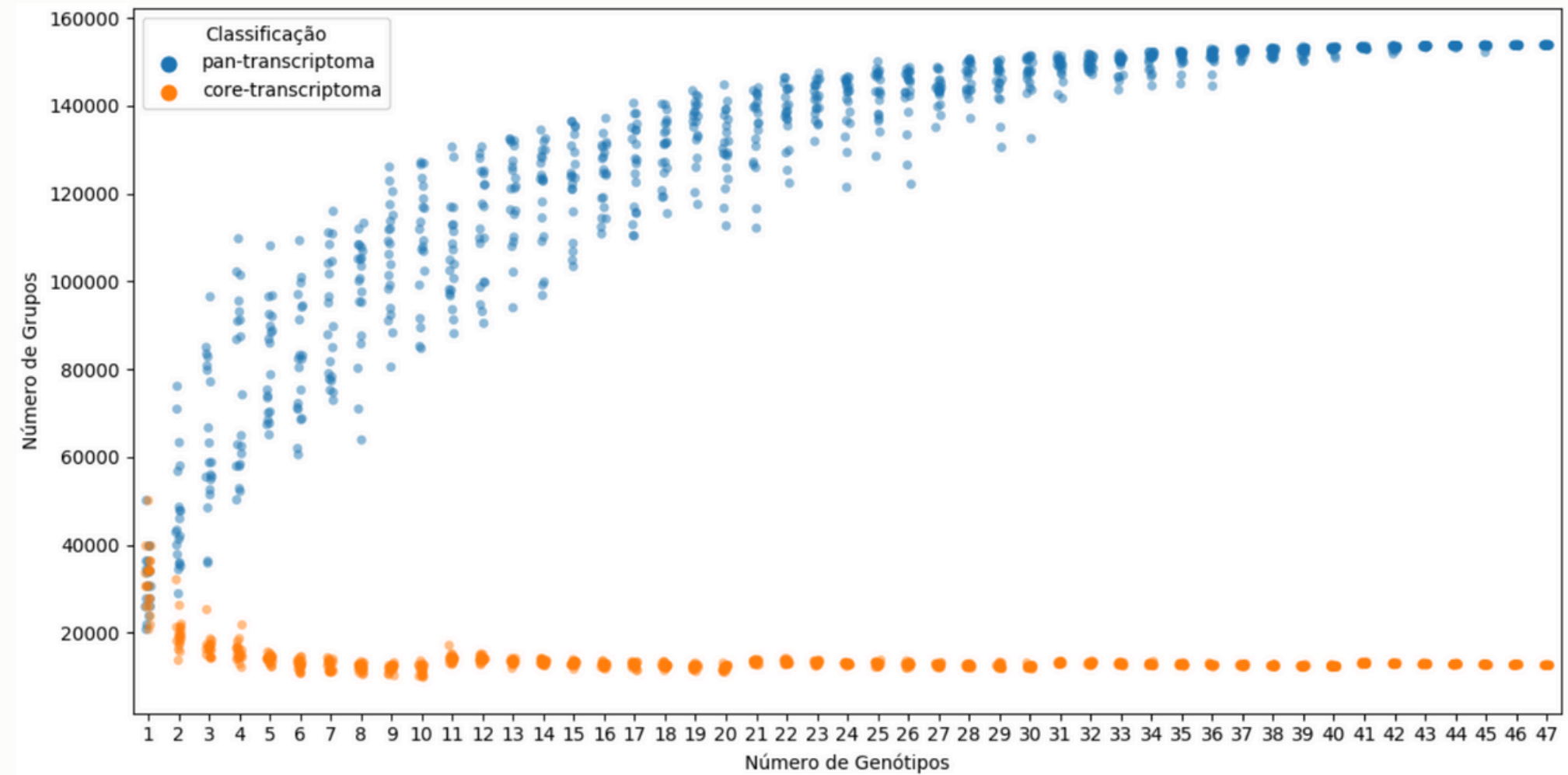
Temos montagens de transcriptomas de 48 variedades de cana-de-açúcar de diferentes lugares ao redor do mundo



Temos montagens de transcriptomas de 48 variedades de cana-de-açúcar de diferentes lugares ao redor do mundo

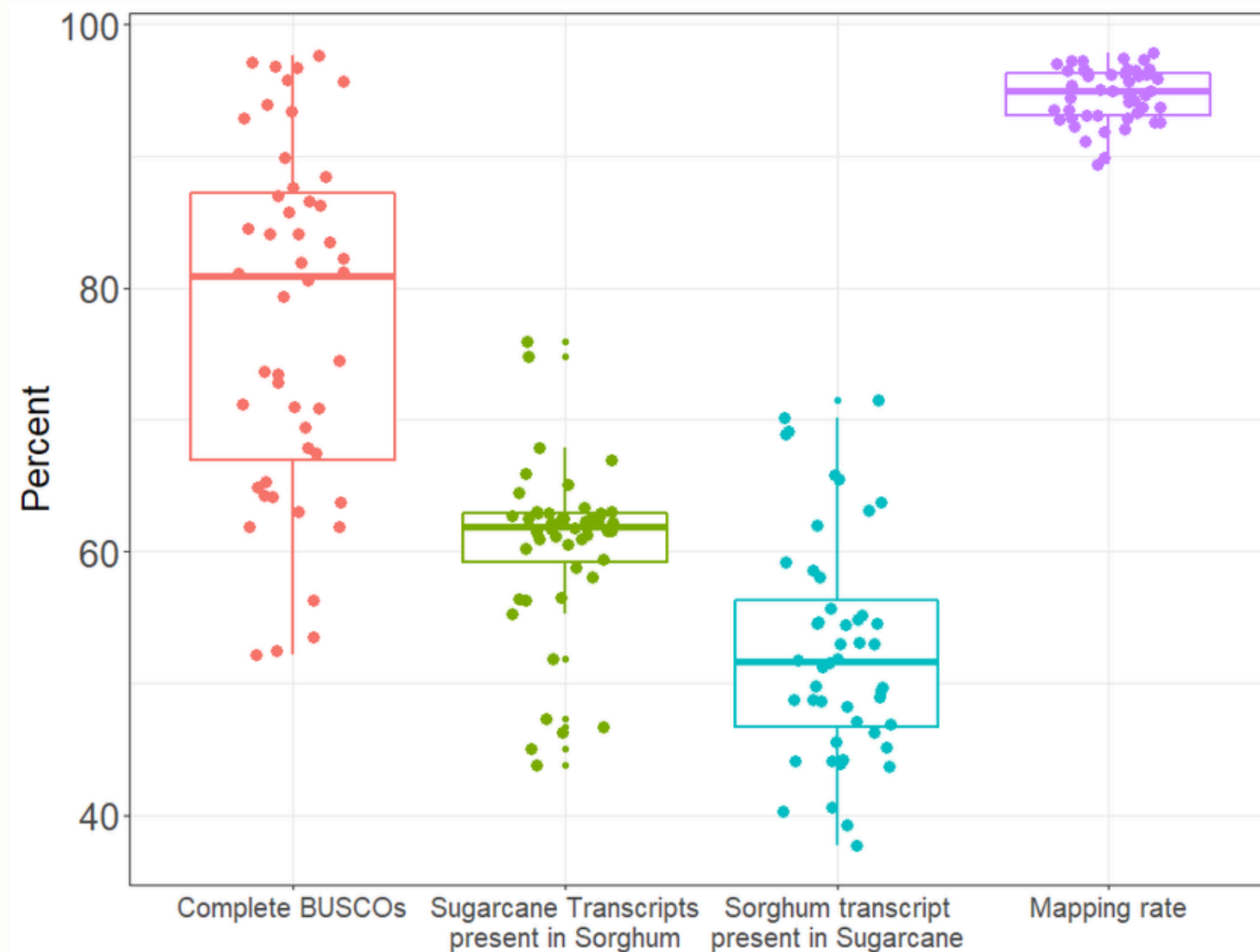


Pan-transcriptoma da cana-de-açúcar

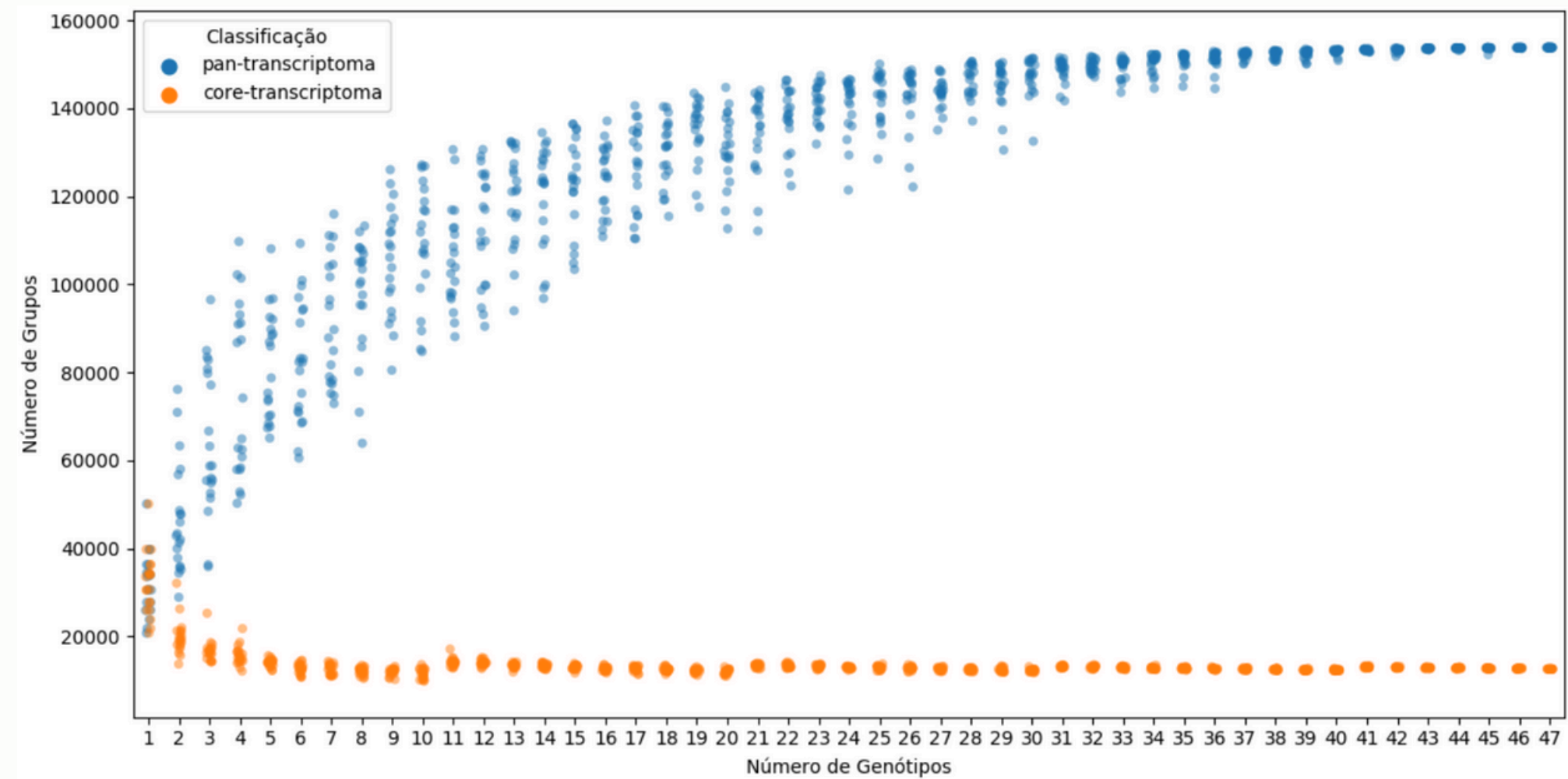


~5 milhões de transcritos (protein-coding) em
153.000 grupos
E os ncRNAs?

Temos montagens de transcriptomas de 48 variedades de cana-de-açúcar de diferentes lugares ao redor do mundo



Pan-transcriptoma da cana-de-açúcar



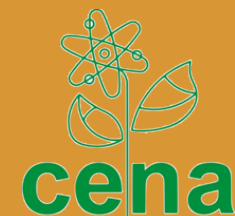
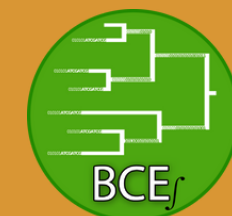
~5 milhões de transcritos (protein-coding) em
~400K grupos
E os ncRNAs?

Montagens e grupos disponíveis https://figshare.com/projects/Sugarcane_Pan-transcriptome/130586

3. Desenvolvimento de plataformas para mineração de dados genômicos e de expressão



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa



CoNekT Bioenergy ou CaneMine

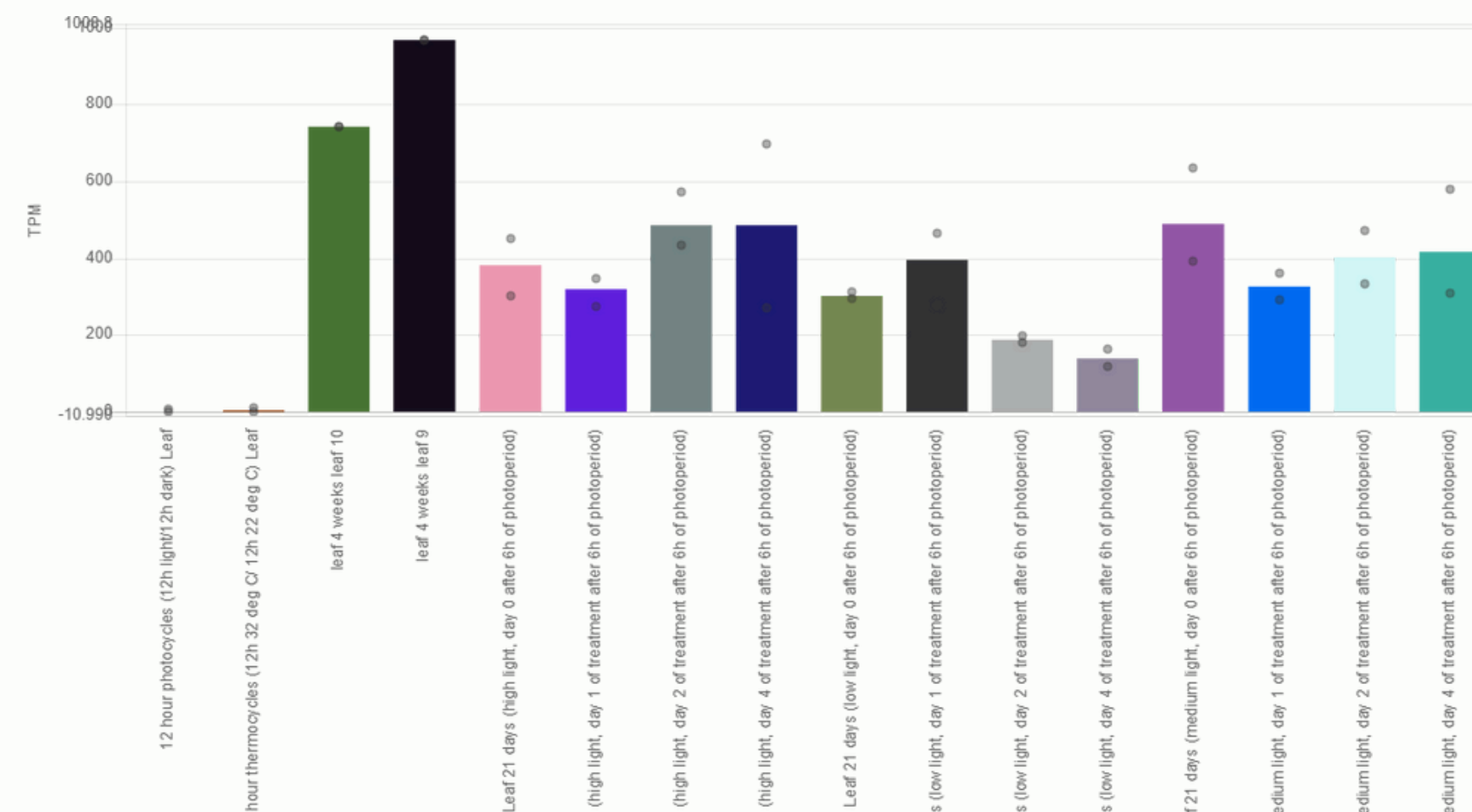
Incluindo dados públicos de espécies de importância em bioenergia, como *Setaria* spp, *Sorghum*, *Brachypodium*, arroz, milho, e cana-de-açúcar (genomas e transcriptomas)

Identificação facilitada de grupos de genes com padrões de expressão semelhantes.

Coming soon:

- 48 transcriptomas de cana-de-açúcar
- O genoma da SP80-3280
- Análise de promotores de genes
- Categorização dos genes envolvidos no metabolismo da parede celular

<http://conekt.cena.usp.br:82/conekt/>

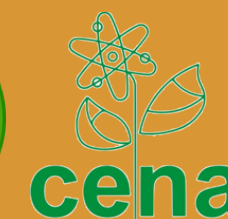
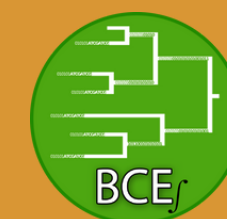


Visualização de perfis de expressão de conjuntos de genes. Identificação de padrões de expressão específicos em condições de interesse.

A partir de um gene numa espécie, pode encontrar seu ortólogo numa outra espécie



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa

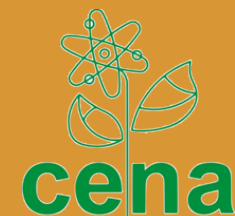
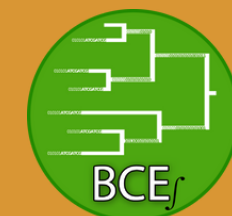


4. Desenvolvimento de software para apoiar a edição do genoma de cana

Colaboração com Prof. Menossi, UNICAMP



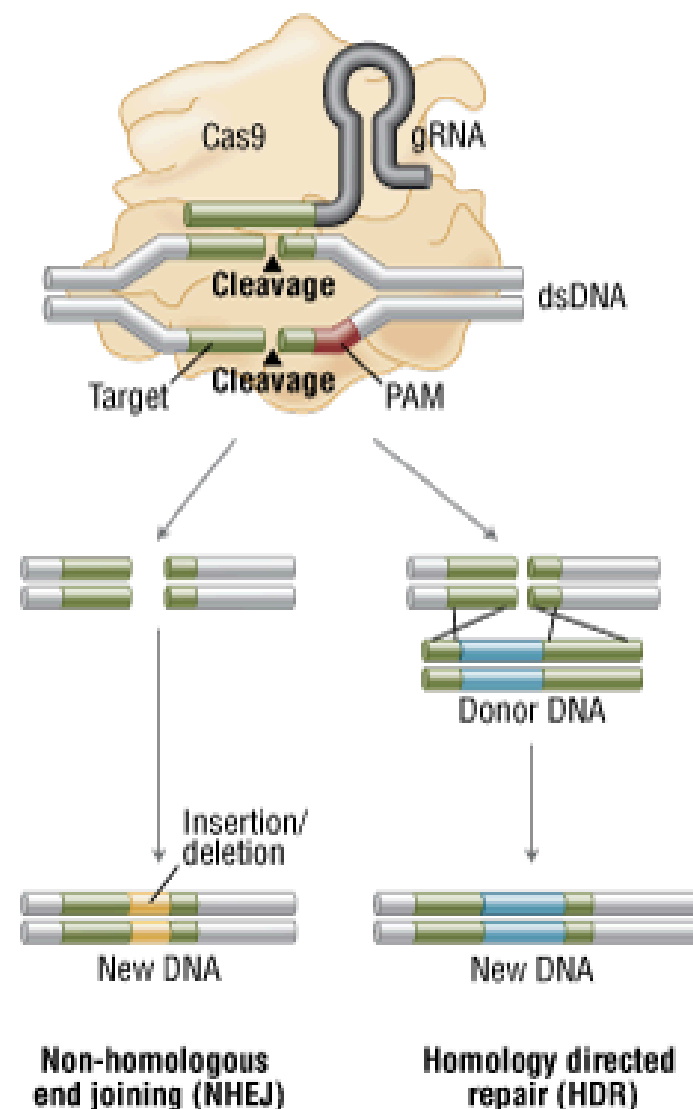
Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa



Introns no gene da Cas9

Foi comprovado em *Arabidopsis* que a inclusão de introns da espécie alvo no gene Cas9 aumenta a taxa de sucesos de edição de genoma

A. Genome Engineering With Cas9 Nuclease



Construto CRISPR/Cas



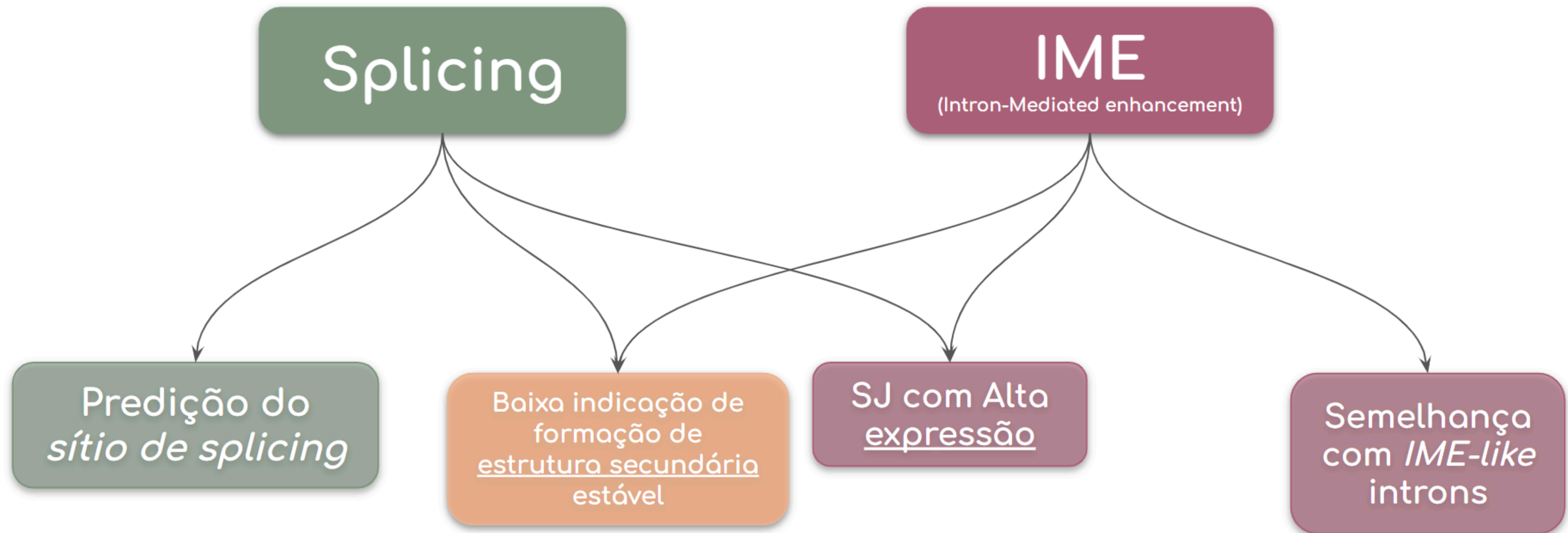
Quais introns usar?

Desenvolvendo um software para selecionar

Genoma	CTBE	IQ-USP	TOTAL
Quantidade de genes que codificam proteínas	153.078	374.774	527.852
Quantidade total de introns	229.857	666.962	896.819

<https://international.neb.com/tools-and-resources/feature-articles/crispr-cas9-and-targeted-genome-editing-a-new-era-in-molecular-biology>

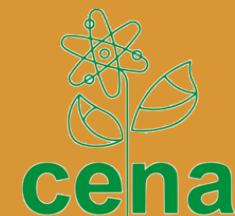
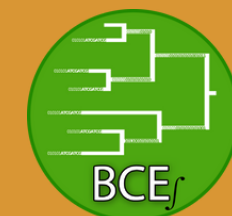
Explorando informação genômica para identificar introns de cana que ajudem a aumentar a taxa de edição de genoma



4. Redes de co-expressão em cana e sorgo



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa



Pan-transcritomas, redes de co-expressão comparativas

Pan transcriptoma
Cana
Sorgo

Redes de co-
expressão
explorando dados
publicos de
acumulo de fibra e
açúcar

Comparação das
redes das duas
especies

Conclusões

Estamos próximos a ter sequencias de cromossomos de telomero até telomero de uma variedade comercial de cana-de-acúcar, i.e., podemos enfrentar genomas poliploides e gerar montagens de altíssima qualidade.

Temos plataformas que podem ajudar na mineração de dados genômicos e de expressão para identificar alvos para estudos de fisiologia e de edição de genoma.

Funding

Obrigado!



Research Centre for
Greenhouse Gas Innovation

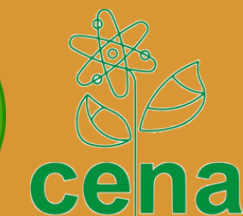
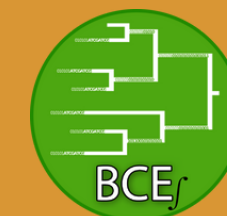
BIOENERGY WITH CARBON CAPTURE
AND STORAGE (BECCS)



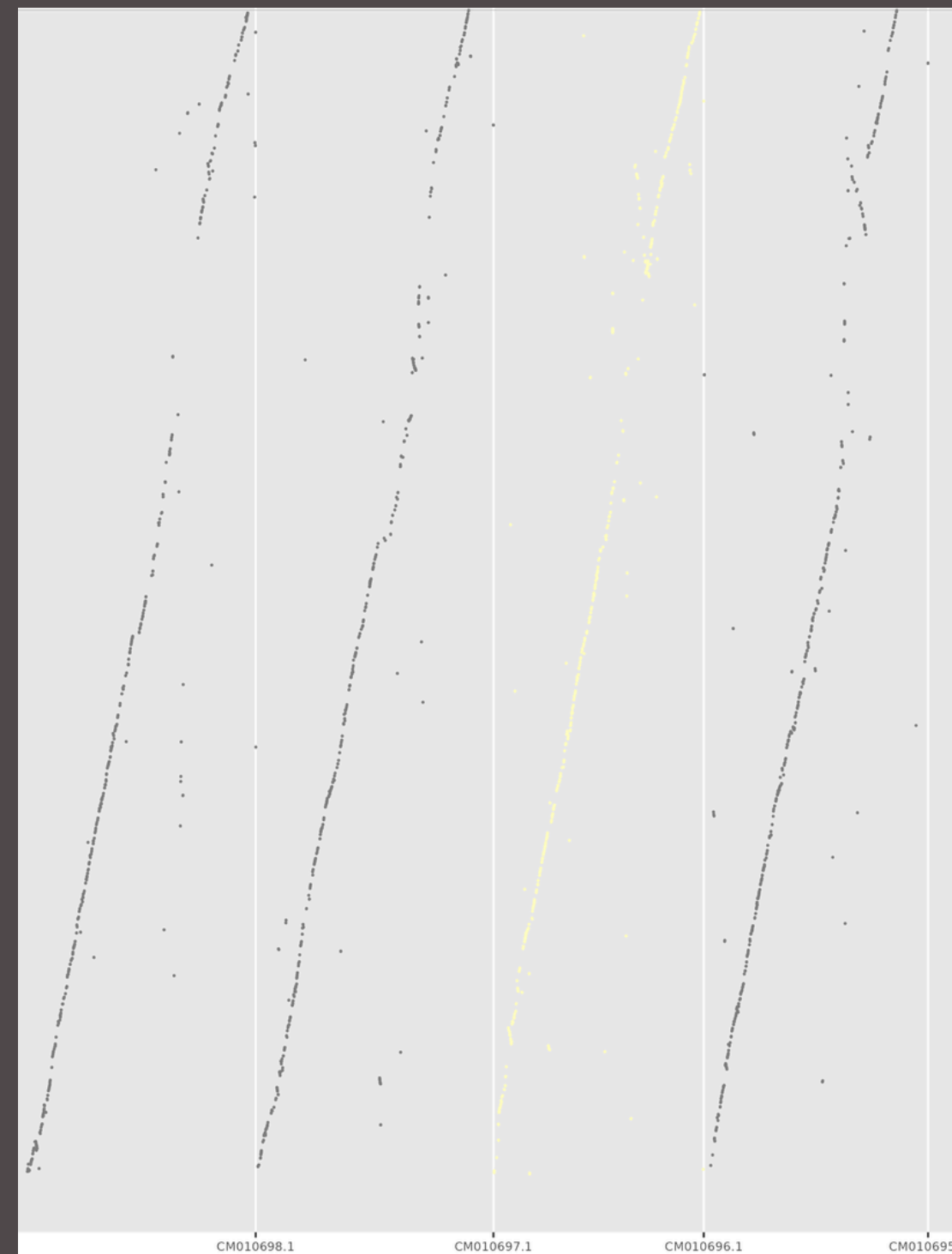
CNPq



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa



**O maior contig
montado até o
momento corresponde
ao cromossomo 4 de
S. officinarum e *S.
spontaneum* quase
interiro. E reparem,
ainda não realizamos o
passo de scaffolding!**

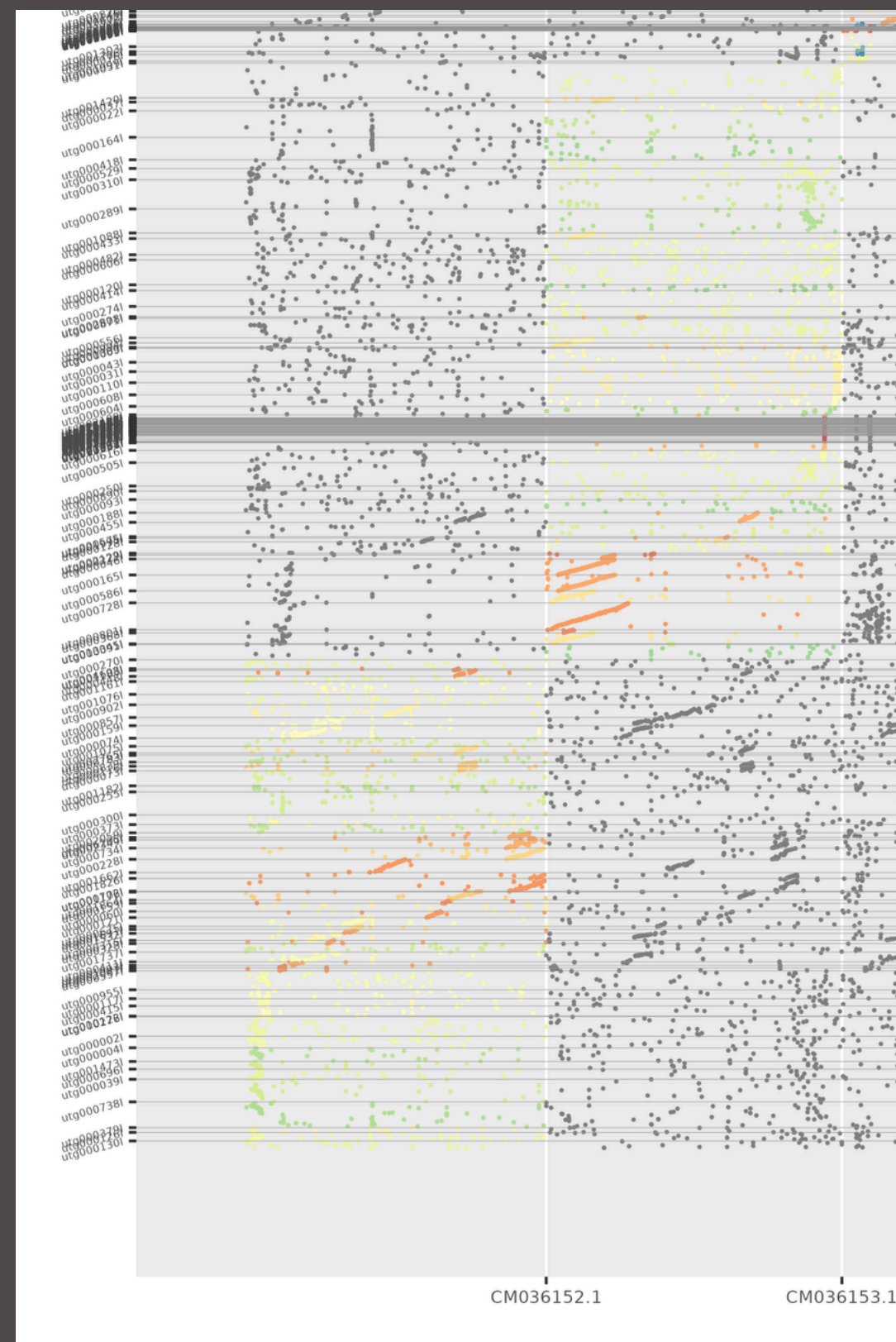


**The largest contig (63.9Mbp) vs
S. spontaneum chromosome 4
(76Mbp - 82Mbp)**

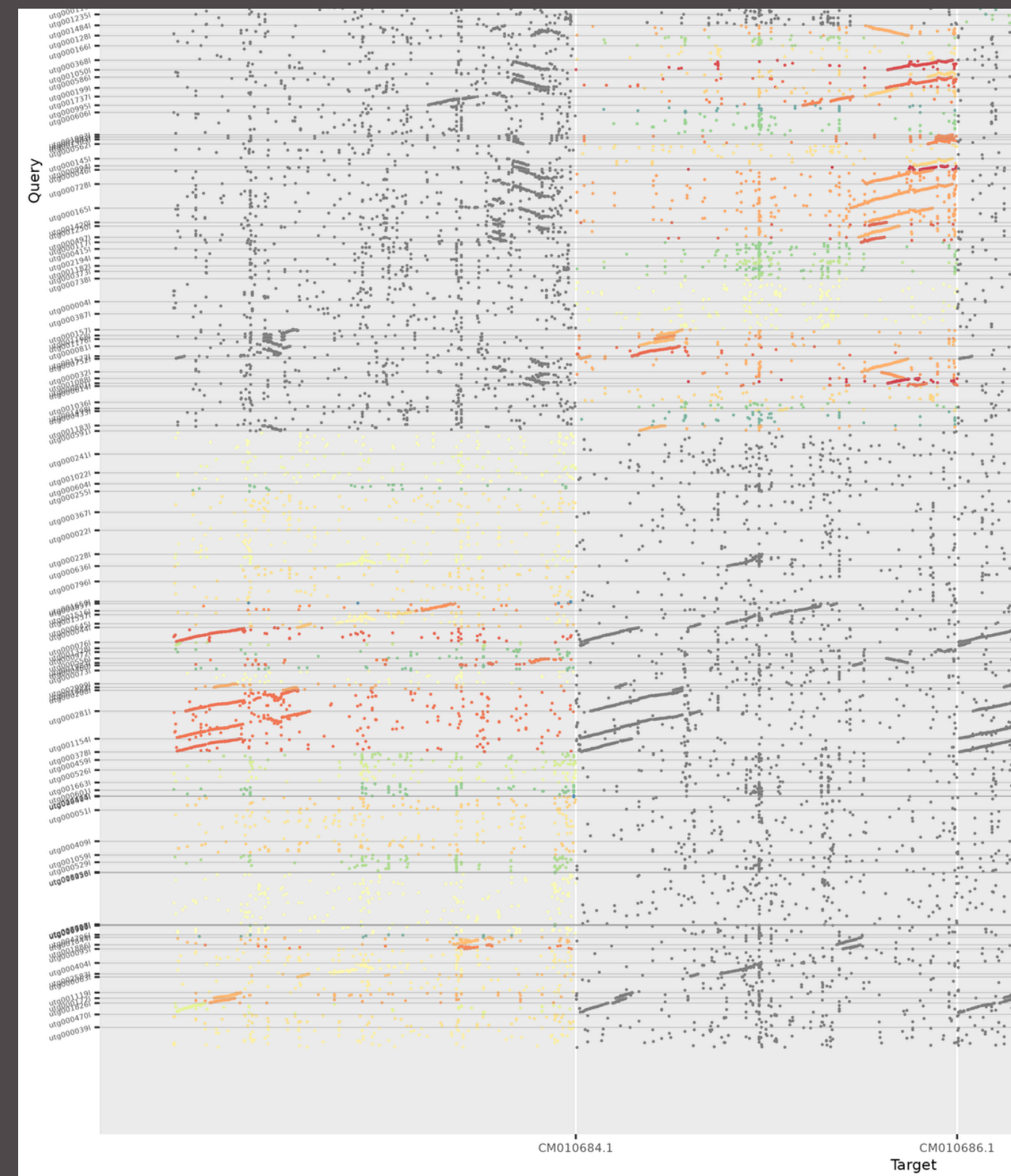


**The largest contigs (63.9Mbp) vs
S. officinarum chromosome 4
(65Mbp - 114Mbp)**

**Nos contigs "menores"
também conseguimos
recuperar regiones
sinténicas aos
chromosomos das
especies parentais.
A montagem da SP80-3280
é uma montagem
poliploide!**



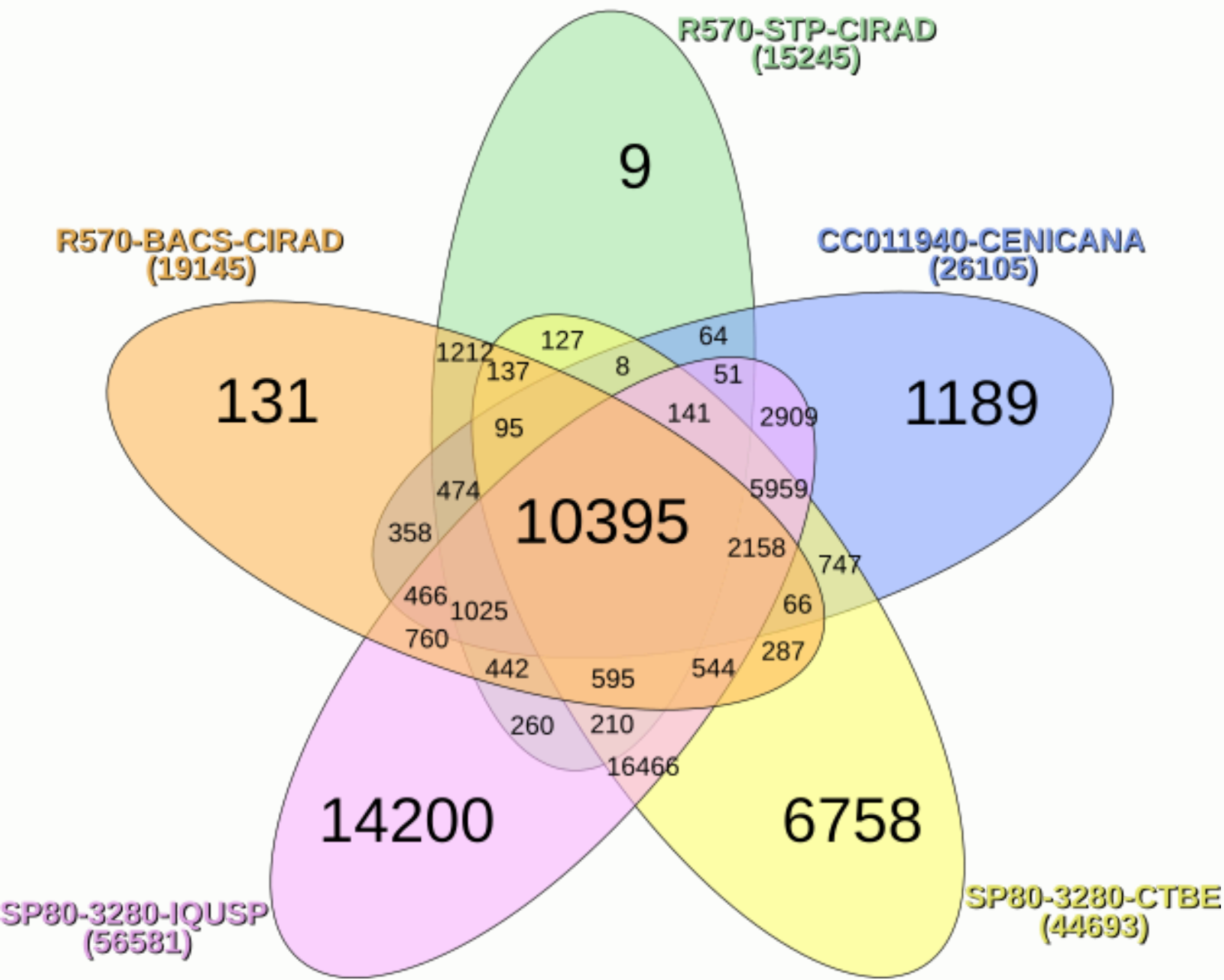
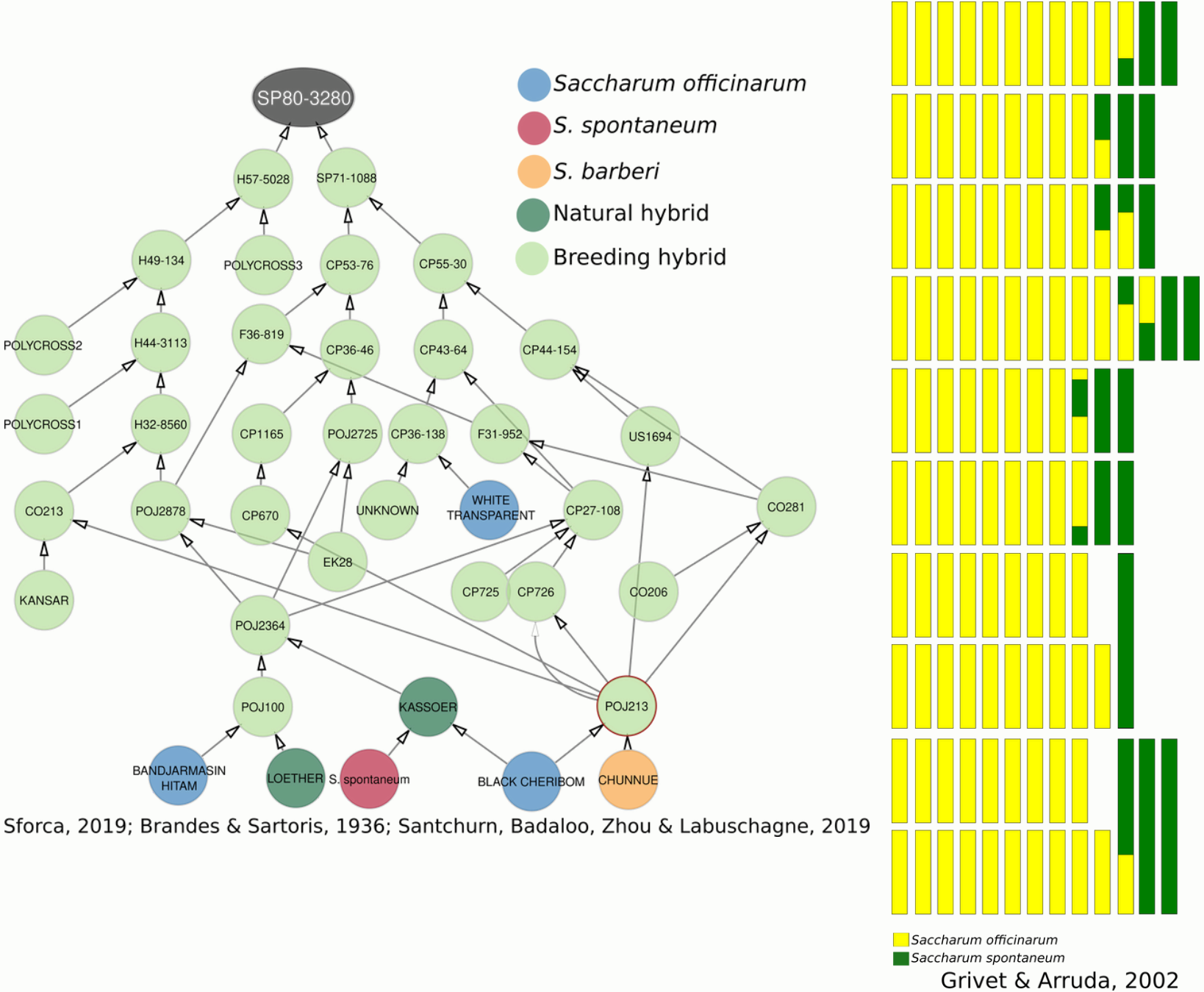
***S. officinarum* chr1**



***S. spontaneum* chr1**

O melhoramento da cana-de-açúcar e seu genoma

Do we have a complete sugarcane genome yet?

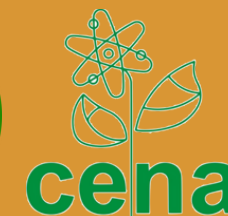
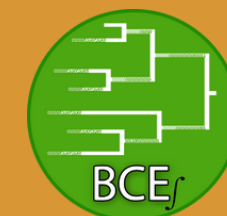


Mas . . .

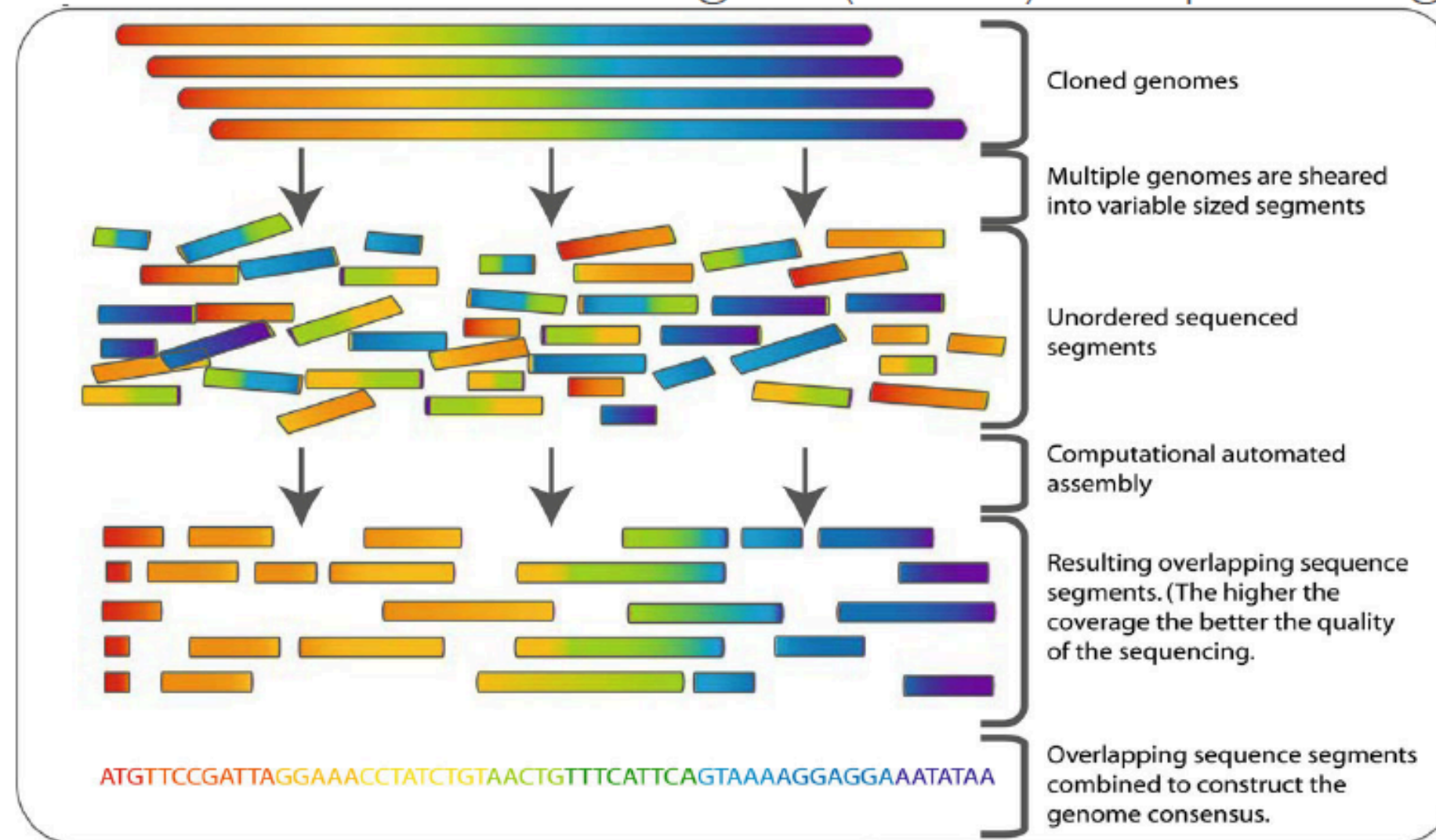
O que é a montagem de genoma?



Instituto Nacional de Ciência e Tecnologia do Bioetanol
Centro de Pesquisa e Inovação de Gases de Efeito Estufa

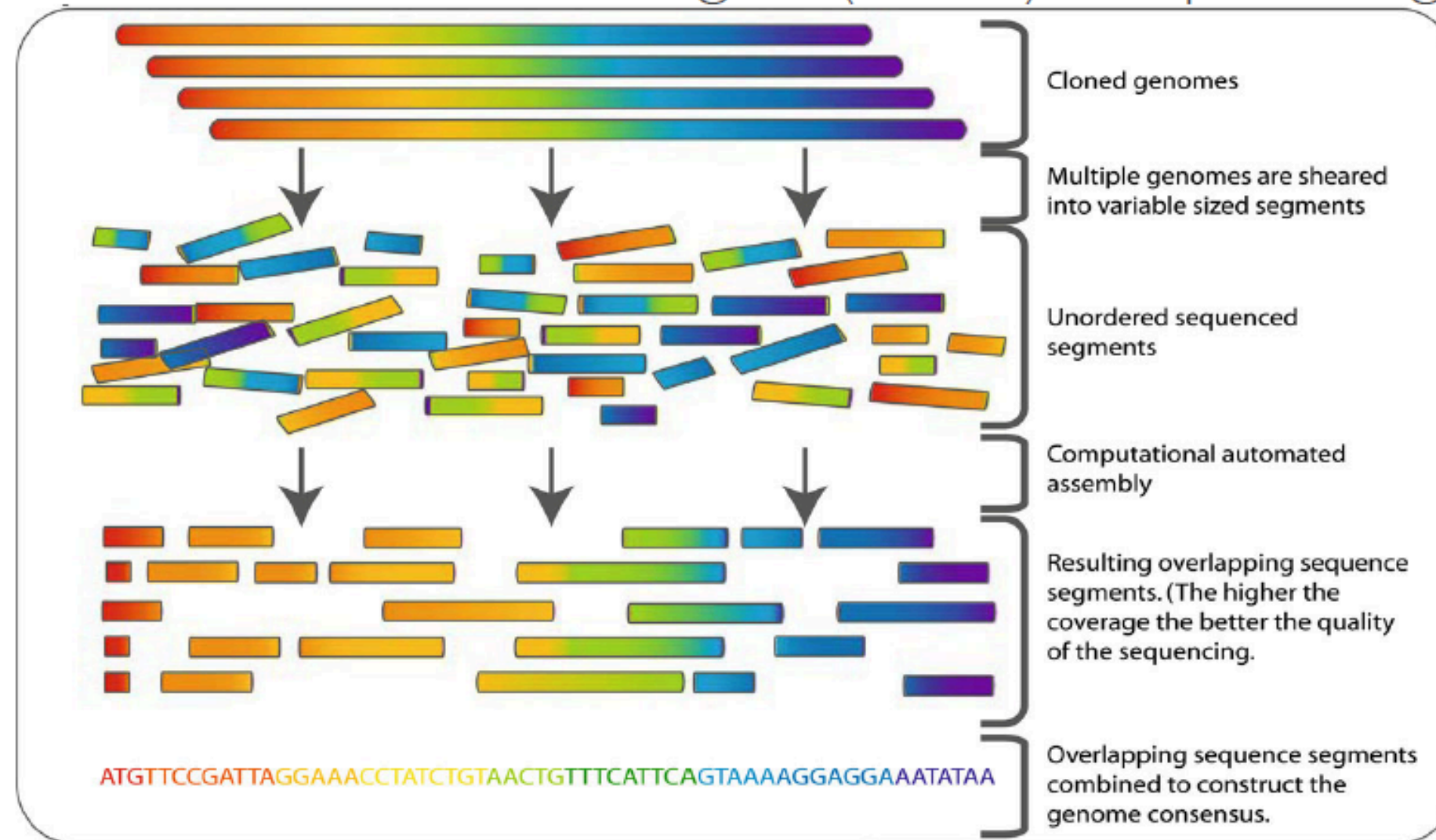


Whole Genome Shotgun (WGS) Sequencing

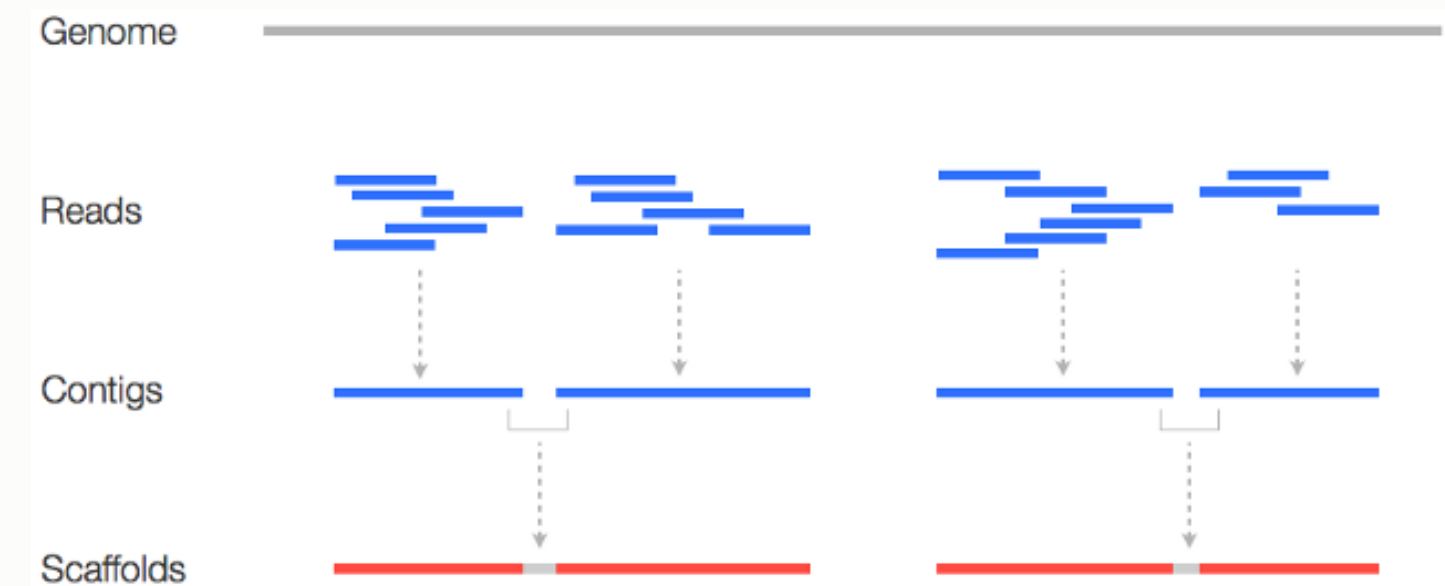


http://bch709.plantgenomicslab.org/Genome_assembly_annotation/index.html

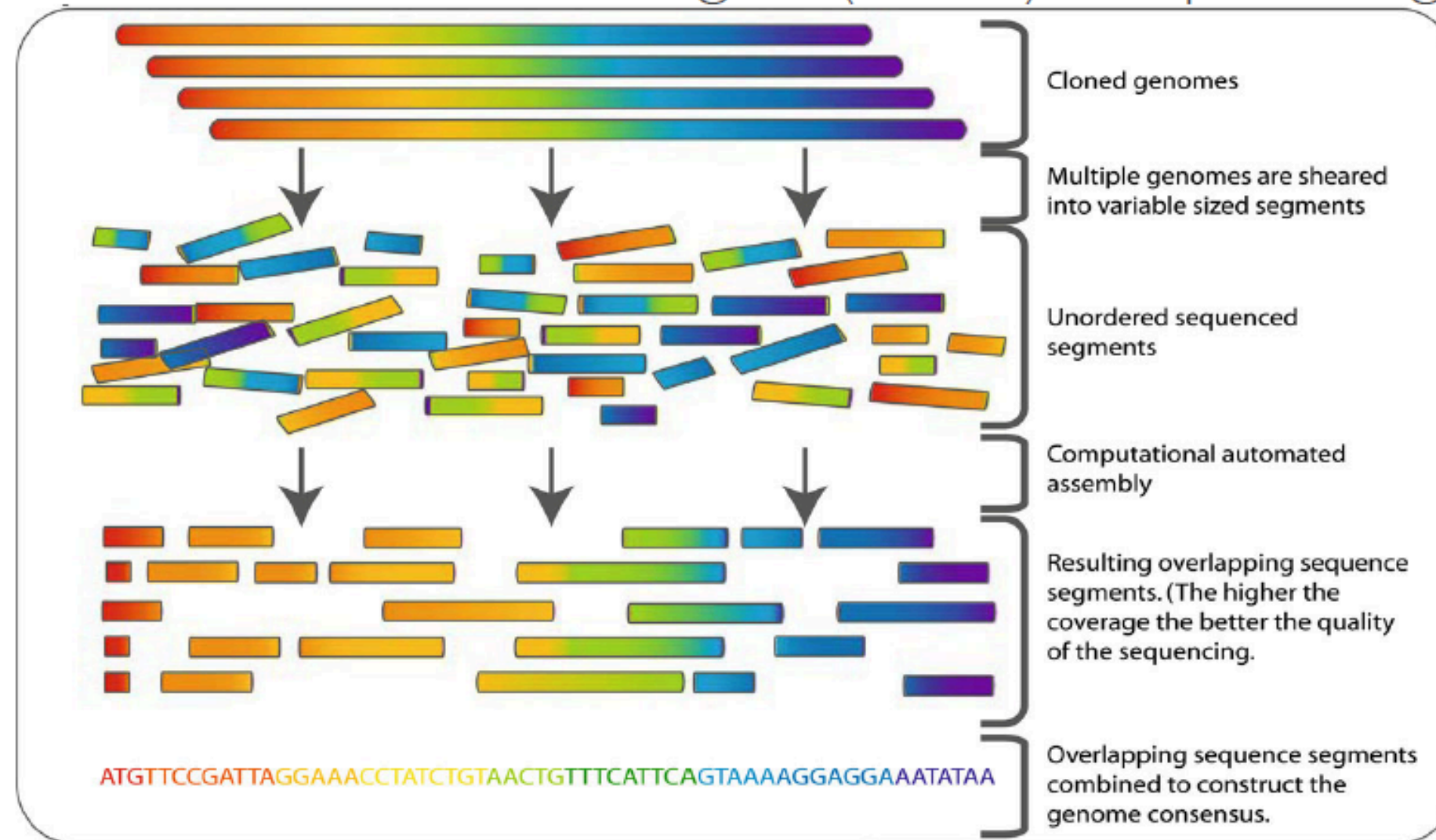
Whole Genome Shotgun (WGS) Sequencing



http://bch709.plantgenomicslab.org/Genome_assembly_annotation/index.html



Whole Genome Shotgun (WGS) Sequencing



http://bch709.plantgenomicslab.org/Genome_assembly_annotation/index.html

Genome

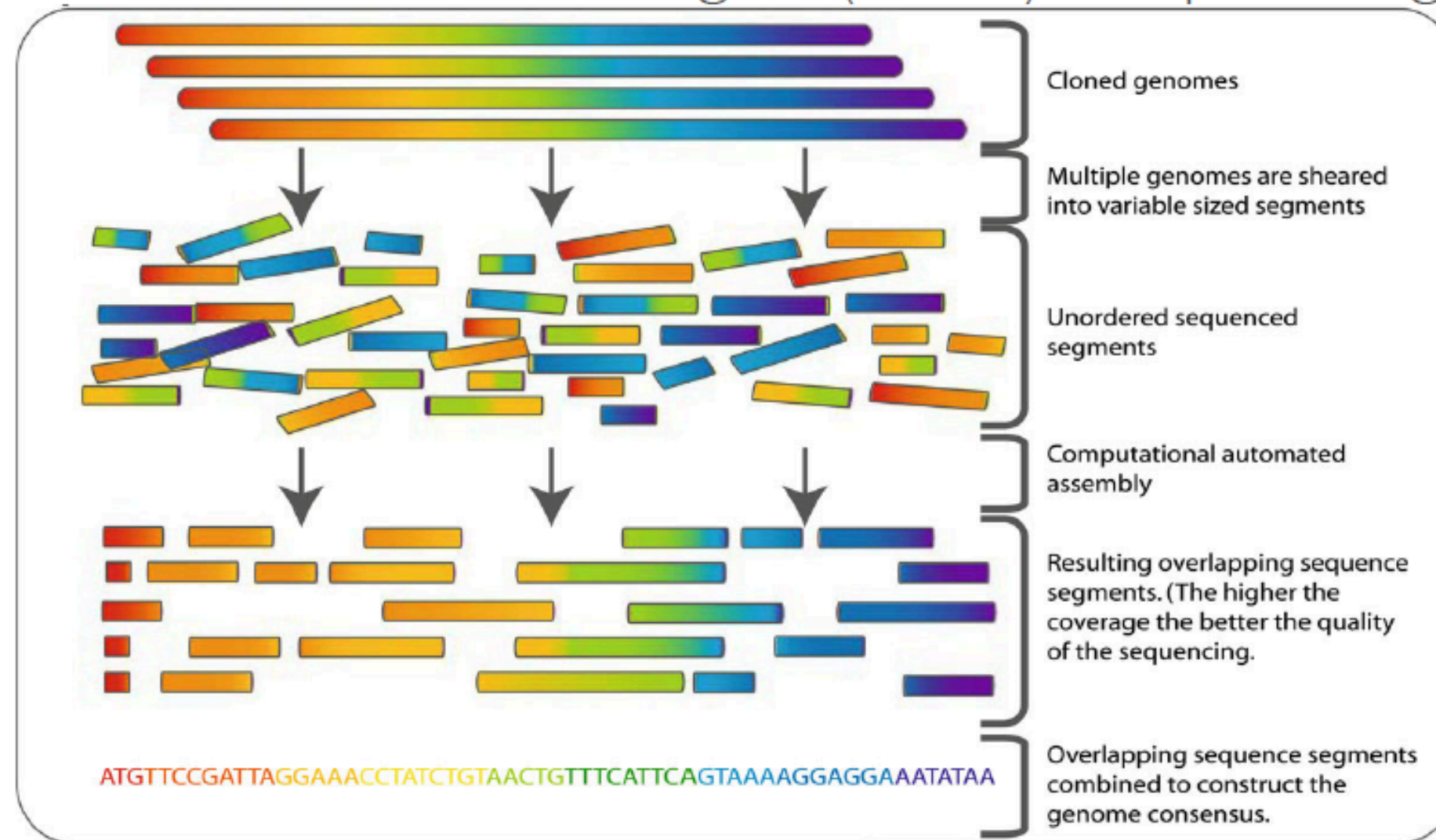
Reads

Contigs

Scaffolds



Whole Genome Shotgun (WGS) Sequencing



Genome

Reads

Contigs

Scaffolds



http://bch709.plantgenomicslab.org/Genome_assembly_annotation/index.html

PacBio HiFi. Error rate ~0.05%

312 Gbp ou ~312x do genoma monoploide, ou 31x each haplofase