

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

Coffea canephora

Coffea canephora group

5 декабря 2014 г.

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

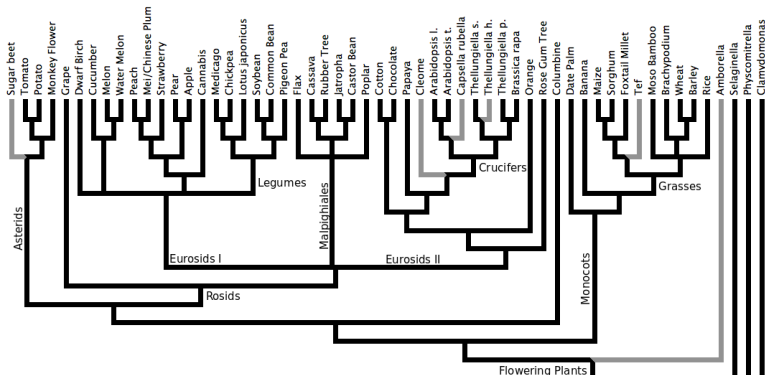
Clusters

Кофе (сем. Маревые).

В промышленности используют два вида: *C. arabica* и *C. canephora* [<http://en.wikipedia.org/wiki/Coffea>]

Геном *C. canephora* недавно секвенирован: 11 пар хромосом, 710 Mbp. [[http:](http://www.sciencemag.org/content/345/6201/1181.full)

[//www.sciencemag.org/content/345/6201/1181.full](http://www.sciencemag.org/content/345/6201/1181.full)].



*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

Ближайший к кофе секвенированный геном — это либо помидор (*Solanum lycopersicum*), либо картофель (*S. tuberosum*). Проекты по секвенированию генома картофеля и томата живут здесь:

<http://www.sgn.cornell.edu/>.

+ классический модельный объект генетики растений *Arabidopsis thaliana*. Основной кладезь информации про этот геном здесь:

<http://www.arabidopsis.org/>.

Самая главная страница:

[<http://coffee-genome.org/download>]



Coffee Genome Hub

a genomics and genetics resource for coffee

5 предложенных белков

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

p01 >pdb|4G2M|A Hydroxycinnamoyltransferase

p02 >pdb|5MDH|A Malate Dehydrogenase

p03 >pdb|1WGP|A Ion Channel

p04 >pdb|2WVJ|A Thymidine Kinase

p05 >pdb|1U1H|A Methionine Synthase

...

p01: 33 потенциальных локуса

p02: 6 потенциальных локусов

p03: 23 потенциальных локуса

p04: 7 потенциальных локусов

p05: 1 потенциальный локус

Аннотация

*Coffea
canephora*

*Coffea
canephora
group*

Intro

Annotation

Repeats

Clusters

Аннотация генов с помощью Augustus
Augustus по модели *Arabidopsis thaliana*:

chr1 – 3883 генов

Поиск по модели *Solanum lycopersicon*:

chr1 – 3575 генов.

====

Это одни и те же гены или нет?

intersectBed из bedtools

```
intersectBed -a \  
./chr1_by_Arabidopsis_augustus.gff -b \  
./chr1_by_tomato_Augustus.txt -r -f .9 \  
>./chr1_intersectBed_r_f_90.gff
```

chr 1 – 2058 генов

CEGMA: подготовка генов для Augustus

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

```
cegma -g pseudomolecules.fa -o coffeeCEGMAdef
```

```
#      Statistics of the completeness of the genome based on 248 CEGs      #

          #Prots  %Completeness  -  #Total  Average  %Ortho

Complete    225      90.73      -   371      1.65      39.56

Group 1      58      87.88      -    83      1.43      27.59
Group 2      49      87.50      -    80      1.63      38.78
Group 3      54      88.52      -    90      1.67      38.89
Group 4      64      98.46      -   118      1.84      51.56

Partial     240      96.77      -   452      1.88      49.17

Group 1      63      95.45      -   102      1.62      38.10
Group 2      54      96.43      -    94      1.74      42.59
Group 3      59      96.72      -   115      1.95      54.24
Group 4      64      98.46      -   141      2.20      60.94

#      These results are based on the set of genes selected by Genis Parra  #

#      Key:                                                                    #
#      Prots = number of 248 ultra-conserved CEGs present in genome          #
#      %Completeness = percentage of 248 ultra-conserved CEGs present        #
#      Total = total number of CEGs present including putative orthologs     #
#      Average = average number of orthologs per CEG                        #
#      %Ortho = percentage of detected CEGs that have more than 1 ortholog   #
```

Аннотация: обучение Augustus

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

***** Evaluation of gene prediction *****

			\	
			sensitivity	specificity
			/	
nucleotide level	0.865	0.442		

	#pred	#anno		FP = false pos.			FN = false neg.				
	total/	total/	TP	-----			-----			sensitivity	sp
	unique	unique		part	ovlp	wrng	part	ovlp	wrng		

exon level				1246			557				
	2151	1462	905	-----			-----			0.619	
	2151	1462		239	24	983	239	21	297		

\							
transcript	#pred	#anno	TP	FP	FN	sensitivity	specificity
/							
gene level	431	200	34	397	166	0.17	0.0789

\				
UTR	total pred	CDS bnd. corr.	meanDiff	medianDiff
/				
TSS	42	0	-1	-1
TTS	16	0	-1	-1
/				
UTR	uniq. pred	unique anno	sens.	spec.

Сравнение наших результатов с официальной аннотацией

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

```
intersectBed -a coffea_canephora.chr1.gff3 \  
-b chr1_by_coffee.gff -wa -u  
>chr1_official_to_our_model_report_off_once.gff
```

```
grep -c gene chr1_official_to_our_model_report_off_once.gf  
3780
```

```
grep -c gene coffea_canephora.chr1.gff3  
11208
```

```
grep -c gene chr1_by_coffee.gff  
39816
```


Поиск повторов с помощью RepeatMasker на основе генома *Arabidopsis*

Coffea
canephora

Coffea
canephora
group

Intro

Annotation

Repeats

Clusters

```
RepeatMasker -species arabidopsis -xsmall -gff ./chr1.fna
```

```
=====
file name: chr1.fna
sequences:          1
total length:    38193400 bp (30728063 bp excl N/X-runs)
GC level:        36.14 %
bases masked:    1673760 bp ( 4.38 %)
```

```
=====
              number of      length  percentage
              elements*  occupied  of sequence
-----
Retroelements      1027      921072 bp    2.41 %
  SINEs:             4         213 bp     0.00 %
  Penelope           0           0 bp     0.00 %
  LINEs:            164      66524 bp     0.17 %
    CRE/SLACS        0           0 bp     0.00 %
    L2/CR1/Rex       0           0 bp     0.00 %
    R1/LOA/Jockey    0           0 bp     0.00 %
    R2/R4/NeSL       0           0 bp     0.00 %
    RTE/Bov-B        0           0 bp     0.00 %
    L1/CIN4          164      66524 bp     0.17 %
LTR elements:      859      854335 bp    2.24 %
  BEL/Pao            0           0 bp     0.00 %
  Ty1/Copia          380      386105 bp    1.01 %
  Gypsy/DIRS1        478      468183 bp    1.23 %
    Retroviral        0           0 bp     0.00 %
```

Поиск повторов с помощью RepeatMasker на основе генома *Arabidopsis* (continued)

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

```
RepeatMasker -species rice -no_is chr1.fna
```

DNA transposons	203	39568 bp	0.10 %
hobo-Activator	94	18339 bp	0.05 %
Tc1-IS630-Pogo	33	5832 bp	0.02 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	3	477 bp	0.00 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %
Rolling-circles	0	0 bp	0.00 %
Unclassified:	71	18327 bp	0.05 %
Total interspersed repeats:		978967 bp	2.56 %
Small RNA:	12	1876 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	12222	536197 bp	1.40 %
Low complexity:	3011	157206 bp	0.41 %
=====			

* most repeats fragmented by insertions or deletions
 have been counted as one element

The query species was assumed to be arabidopsis

Поиск повторов с помощью RepeatScout

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

```
build_lmer_table -l 15 -sequence chr1.fna \  
-freq chr1.lt -v
```

```
RepeatScout-1/RepeatScout -sequence chr1.fna \  
-output chr1.rs -freq chr1.lt -l 15 -v
```

```
wc -l chr1.lt  
1320654
```

```
grep -c \> chr1.rs  
2223
```

Поиск повторов с помощью RepeatMasker на основе данных RepeatScout

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

```
=====
file name: chr1.fna
sequences:          1
total length:    38193400 bp (30728063 bp excl N/X-runs)
GC level:        36.14 %
bases masked:    11761995 bp ( 30.80 %)
=====
              number of      length  percentage
              elements*  occupied  of sequence
-----
....
Unclassified:    41342      12121768 bp   31.74 %

Total interspersed repeats: 12121768 bp   31.74 %


Small RNA:        0          0 bp   0.00 %

Satellites:        0          0 bp   0.00 %
Simple repeats:   11017      449018 bp   1.18 %
Low complexity:    0          0 bp   0.00 %
=====

* most repeats fragmented by insertions or deletions
  have been counted as one element
```

Анализ семейств генов

*Coffea
canephora*

*Coffea
canephora*
group

Intro

Annotation

Repeats

Clusters

```
cat <3 genomes> => kClust =>  
kclust_out_for_cafe.py => CAFE3.0
```

```
load -i cafe_input.out  
tree (>AT:5,(>Cc:1,>So:2):3)  
lambda -s  
lambda -l 0.17  
report resultsfile
```

