Coffea canephora

Coffea canephora group
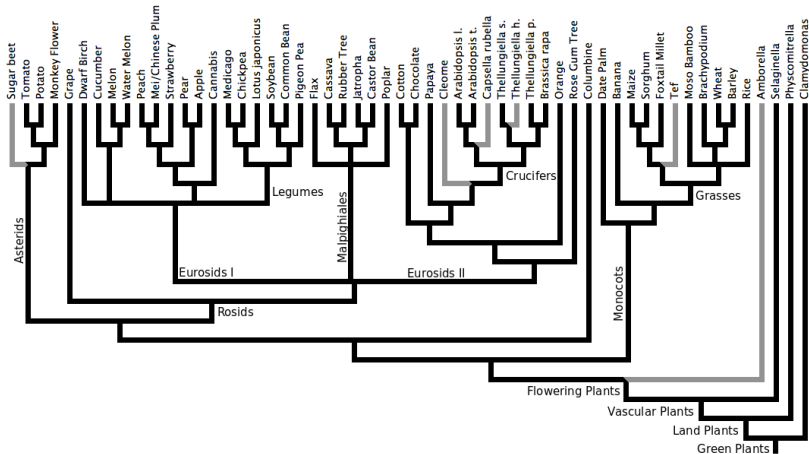
25 октября 2014 г.

Кофе (сем. Маревые).

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

Геном *C. canephora* недавно отсеквенирован: 11 пар хромосом, 710 Mbp.

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

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

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

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

Поиск повторов с помощью RepeatMasker на основе генома *Arabidopsis*

```
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)

```
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,           0            0 bp    0.00 %
    P-element, Transib)

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
```

Поиск повторов с помощью RepeatMasker на основе генома риса

```
==================================================
file name: chr1.fna * поиск по повторам из генома риса
sequences:                 1
total length:    38193400 bp  (30728063 bp excl N/X-runs)
GC level:          36.14 %
bases masked:      713556 bp ( 1.87 %)
==================================================
                number of        length    percentage
                elements*      occupied    of sequence
--------------------------------------------------
SINEs:                17          1008 bp    0.00 %
      ALUs             0             0 bp    0.00 %
      MIRs             3           183 bp    0.00 %

LINEs:               105          6967 bp    0.02 %
      LINE1           10           449 bp    0.00 %
      LINE2           11           727 bp    0.00 %
      L3/CR1          60          4183 bp    0.01 %
```

```
LTR elements:            19           2445 bp     0.01 %
     ERVL                 2            135 bp     0.00 %
     ERVL-MaLRs           0              0 bp     0.00 %
     ERV_classI           5            470 bp     0.00 %
     ERV_classII          1             54 bp     0.00 %

DNA elements:            29           1731 bp     0.00 %
     hAT-Charlie          4            165 bp     0.00 %
     TcMar-Tigger         6            380 bp     0.00 %

Unclassified:            24           2999 bp     0.01 %

Total interspersed repeats:         15150 bp     0.04 %


Small RNA:               77           6918 bp     0.02 %

Satellites:               0              0 bp     0.00 %
Simple repeats:       12270         535071 bp     1.40 %
Low complexity:        3016         157382 bp     0.41 %
==================================================

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


The query species was assumed to be homo sapiens
RepeatMasker version open-4.0.5 , default mode

run with rmblastn version 2.2.27+
```

Поиск повторов с помощью RepeatMasker на основе генома кукурузы

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

```
==================================================
file name: chr1.fna
sequences:            1
total length:   38193400 bp  (30728063 bp excl N/X-runs)
GC level:         36.14 %
bases masked:    1632178 bp ( 4.27 %)
==================================================
               number of      length   percentage
               elements*    occupied   of sequence
--------------------------------------------------
Retroelements       841      849585 bp    2.22 %
   SINEs:             0           0 bp    0.00 %
   Penelope           0           0 bp    0.00 %
   LINEs:           111       32297 bp    0.08 %
     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        111       32297 bp    0.08 %
   LTR elements:    730      817288 bp    2.14 %
     BEL/Pao          0           0 bp    0.00 %
     Ty1/Copia      296      328851 bp    0.86 %
     Gypsy/DIRS1    431      487305 bp    1.28 %
       Retroviral     0           0 bp    0.00 %
```

```
DNA transposons          246         49637 bp     0.13 %
   hobo-Activator         77         16928 bp     0.04 %
   Tc1-IS630-Pogo         18          1956 bp     0.01 %
   En-Spm                  0             0 bp     0.00 %
   MuDR-IS905              0             0 bp     0.00 %
   PiggyBac                0             0 bp     0.00 %
   Tourist/Harbinger       1           226 bp     0.00 %
   Other (Mirage,          0             0 bp     0.00 %
    P-element, Transib)

Rolling-circles            0             0 bp     0.00 %

Unclassified:            143         48891 bp     0.13 %

Total interspersed repeats:        948113 bp     2.48 %


Small RNA:                12          1876 bp     0.00 %

Satellites:                0             0 bp     0.00 %
Simple repeats:        12125        527308 bp     1.38 %
Low complexity:         2993        155668 bp     0.41 %
==================================================

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


The query species was assumed to be oryza sativa
RepeatMasker version open-4.0.5 , default mode
```

Поиск повторов с помощью RepeatMasker на основе генома *C. elegans*

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

```
==================================================
file name: chr1.fna
sequences:              1
total length:   38193400 bp  (30728063 bp excl N/X-runs)
GC level:        36.14 %
bases masked:    1653432 bp ( 4.33 %)
==================================================
                number of      length    percentage
                elements*    occupied   of sequence
--------------------------------------------------
Retroelements       1029       869163 bp     2.28 %
   SINEs:              0            0 bp     0.00 %
   Penelope            0            0 bp     0.00 %
   LINEs:            145        22491 bp     0.06 %
    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          145        22491 bp     0.06 %
   LTR elements:     884       846672 bp     2.22 %
    BEL/Pao            0            0 bp     0.00 %
    Ty1/Copia        349       361735 bp     0.95 %
    Gypsy/DIRS1       535       484937 bp     1.27 %
      Retroviral       0            0 bp     0.00 %
```

```
DNA transposons           220        46689 bp    0.12 %
   hobo-Activator         131        28258 bp    0.07 %
   Tc1-IS630-Pogo           0            0 bp    0.00 %
   En-Spm                   0            0 bp    0.00 %
   MuDR-IS905               0            0 bp    0.00 %
   PiggyBac                 0            0 bp    0.00 %
   Tourist/Harbinger       17          844 bp    0.00 %
   Other (Mirage,           0            0 bp    0.00 %
    P-element, Transib)

Rolling-circles             0            0 bp    0.00 %

Unclassified:              56        41852 bp    0.11 %

Total interspersed repeats:         957704 bp    2.51 %


Small RNA:                 12         1876 bp    0.00 %

Satellites:                 7          508 bp    0.00 %
Simple repeats:         12233       536270 bp    1.40 %
Low complexity:          3014       157560 bp    0.41 %
==================================================

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

The query species was assumed to be zea
RepeatMasker version open-4.0.5 , default mode

run with rmblastn version 2.2.27+
```

Поиск повторов с помощью RepeatScout

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

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


wc -l chr1.lt
1320654 chr1.lt


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