Coffea canephora

Coffea canephora group

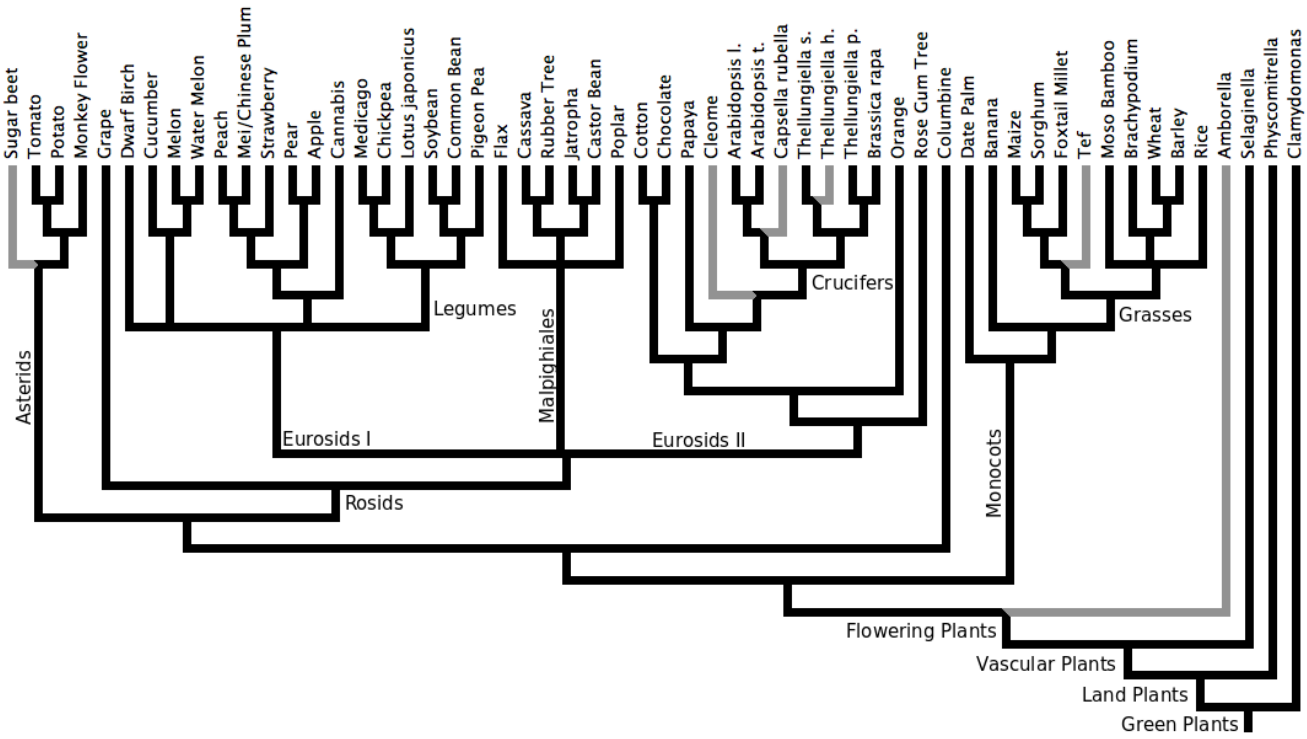
11 октября 2014 г.

Кофе (род *Coffea*) — это растение семейства Маревые (Rubiaceae), порядка Горечавкоцветные (Gentianales), клады Asterids. Для производства любимых нами всеми зёрен используют в основном два вида: *C. arabica* и *C. canephora* [`http://en.wikipedia.org/wiki/Coffea`]

Недавно (статью в Science вышла 5 сентября 2014 года) был отсеквенирован геном *C. canephora*. 11 пар хромосом, 710 Mbp. [`http://www.sciencemag.org/content/345/6201/1181.full`]. Авторы не пишут прямо, но есть подозрение, что этот вид предпочли потому, что он диплоид, а *C. arabica* — тетраплоид.

Ещё некоторые полезные ссылки:

`https://genomevolution.org/wiki/index.php/File:Totalsequencedtree.png`



`http://www.coffeegenome.org/`

`http://coffee-genome.org/`

Coffee Genome Hub

a genomics and genetics resource for coffee

Я утверждаю, что ближайший к кофе секвенированный геном — это либо помидор (*Solanum lycopersicum*), либо картофель (*S. tuberosum*), двух представителей сем. Паслёновые (Solanaceae), входящего в ту же кладу Asterids. Всего десяток лет назад кофе относили к тому же семейству Паслёновые. Проекты по секвенированию генома картофеля и томата живут здесь:

`http://www.sgn.cornell.edu/`.

Ещё можно попробовать поработать с генами классического модельного объекта генетики растений *Arabidopsis thaliana*. Насколько я знаю, основной кладезь информации про этот геном здесь:

`http://www.arabidopsis.org/`.

Арабидопсис намного дальше от кофе, чем томат и картофель, зато изучен несравнимо лучше.

Я вот нашла статью о том, что у кофе и помидора схожие наборы генов, а на арабидопсисные они как раз не очень похожи: [`http://link.springer.com/article/10.1007%2Fs00122-005-0112-2`].

А это самая главная страничка, с которой и стоит скачивать данные:

[`http://coffee-genome.org/download`] Мы можем работать с файлами .fasta с каждой хромосомой. Разбираться со скэффолдами отдельно веселее, но сложнее.

Повторить файл .gff3 в самом начале — это и есть цель нашей работы.

Пока работаем с генами *Arabidopsis thaliana*, *Solanum lycopersicum* и *Solanum tuberosum* и посмотреть, что лучше получится.

Результаты (code + output)

Ищем 5 предложенных белков

```
tblastn -query ./raw_data/sample-proteins.fa -db nr -remote -out ./blastx.out
```

p01 >pdb|4G2M|A Chain A, Structure Of A Lys-Hct Mutant From Coffea Canephora (Hc t for Hydroxycinnamoyltransferase) (Crystal Form 2)

p02 >pdb|5MDH|A Chain A, Crystal Structure Of Ternary Complex Of Porcine Cytoplasmic Malate Dehydrogenase Alpha-Ketomalonate And Tnad At 2.4 Angstroms Resolution

p03 >pdb|1WGP|A Chain A, Solution Structure Of The Cnmp-Binding Domain From Arabidopsis Thaliana Cyclic Nucleotide-Regulated Ion Channel

p04 >pdb|2WVJ|A Chain A, Mutation Of Thr163 To Ser In Human Thymidine Kinase Shifts The Specificity From Thymidine Towards The Nucleoside Analogue Azidothymidine

p05 >pdb|1U1H|A Chain A, A. Thaliana Cobalamine Independent Methionine Synthase

```
exonerate --model protein2genome sample-proteins.fa coffeechrom.fna >../exonerate_sam
```

p01: 33 потенциальных локуса

p02: 6 потенциальных локусов

p03: 23 потенциальных локуса

p04: 7 потенциальных локусов

p05: 1 потенциальный локус

Аннотация генов с помощью Augustus

Поиск без тренировки по модели *Arabidopsis thaliana*:
chr1 - 3883 генов, chr2 - 8246 генов, chr10 - 4020 генов
Поиск по модели *Solanum lycopersicon*:
chr1 - 3575 генов.
Что нашел бласт и что augustus:
Blast_out vs. augustus_chr1
pdb |4G2M|A Chain A—-> gene g4|
pdb |4G22|A Chain A—-> gene g4|
pdb |4G0B|A Chain A—-> gene g4|
Код для Augustus (draft):
Сначала по *Arabidopsis thaliana*:

```
augustus --species=arabidopsis --UTR=off --strand=both --sample=100 --keep_viterbi=tr
/home/drozdovapb/apt/augustus-3.0.3/bin/augustus --species=arabidopsis ./raw_data/chr
```

Потом, как и обещали, по геному томата:

```
/home/drozdovapb/apt/augustus-3.0.3/bin/augustus --species=tomato ../coffee_raw_data
```