Hi, i am melanogaster

Артем Тетюхин
Денис Сермухамедов
Илья Корвиго
Никита Шиляев

Drosophila melanogaster

Animalia > Arthropoda > Insecta > Diptera > Drosophilidae > Drosophila > Drosophila melanogaster

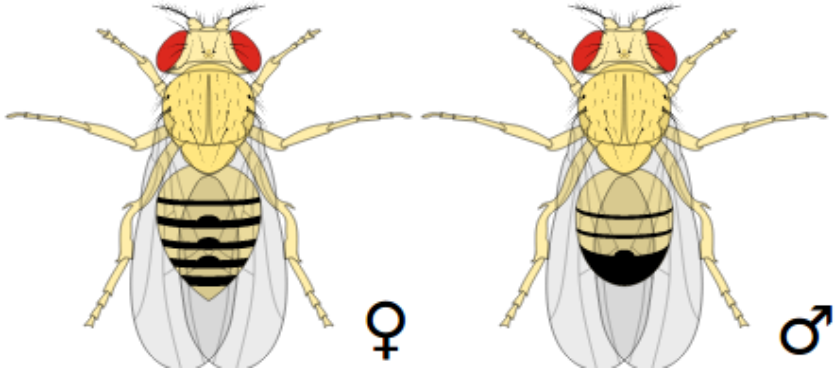
Количество хромосом: 4 пары

Размер генома : 139.5 Mb

Аннотированных генов: ~ 15.700

Митохондриальный геном: 8.5 kb

Половина генов сходны с генами человека.



Первоначальный план действий

Изначальные планы:

1. Используя RepeatScout, создать базу повторов. Замаскировать их с помощью RepeatMasker. Сравнить нашу базу повторов и официальную.
2. Отобрать все белки и мРНК отряда Diptera. Провести кластеризацию. Выделить центроиды.
3. Используя exonerate и tblastn, сопоставить библиотеку репрезентативов с неаннотированным геномом. Исключить пересекающиеся группы. Сопоставить нашу аннотацию с официальной
4. Используя найденные гены, создать базу данных для тренировки Augustus. Предсказать гены de novo.
5. Провести анализ разложения ортологических семейств генов.

Поиск генов

1)Выкачали ~ 400к белков - через api ncbi, иначе обрывается

2)Откластили - blustclust -> не взлетел

kClust -> 190к

3) Tblastn - 20к, хорошо вошли
4) Exonerate ~ 16к
5) Bedtools intersect ~ 97%

мРНК

...

Repeats, repeats, repeats

RepeatMasker lib - 27%

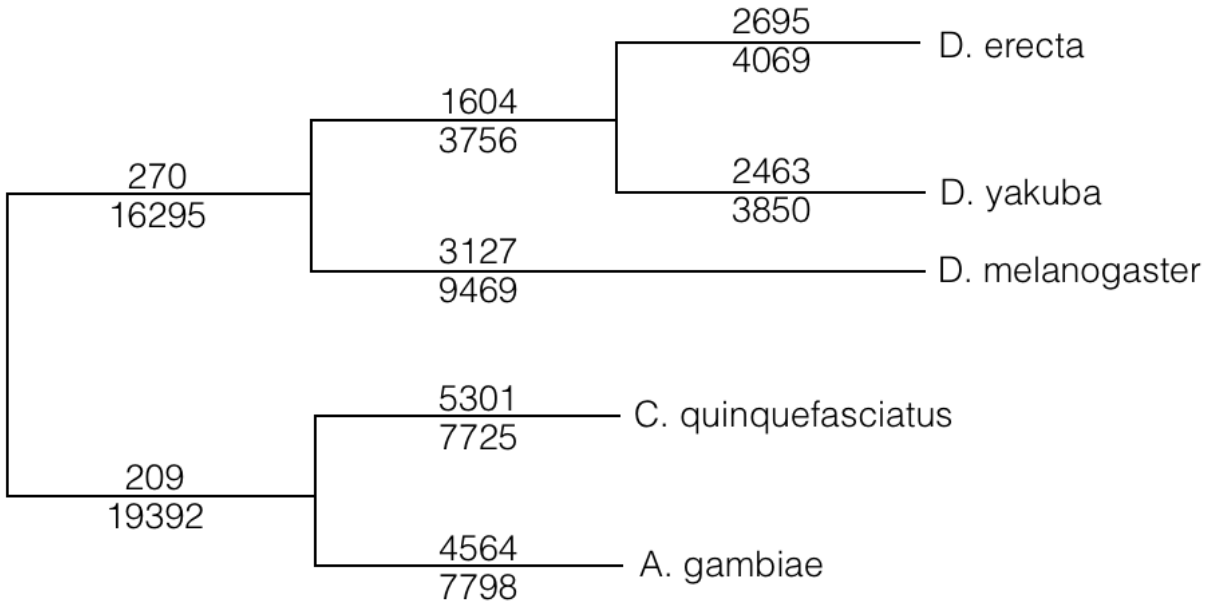
RepeatScout our lib - 31%

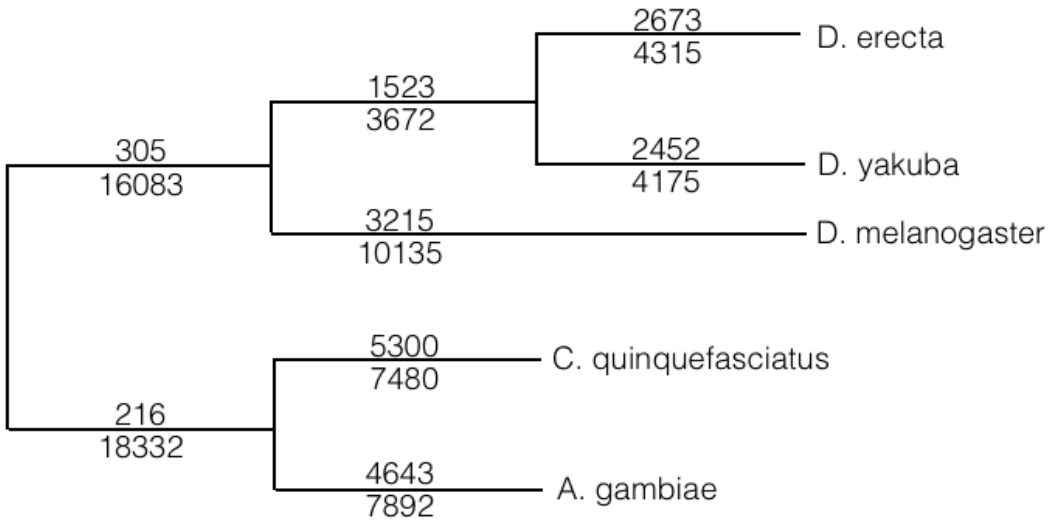
Поиск генов de novo

1. Использовали две программы GeneMark ES и Augustus. Первую пытались починить (безуспешно). Для второй смогли починить конвертацию gff в gb (написали скрипт, который чинит формат gff с exonerate).
2. Отобрали самые длинные и качественные аннотации с exonerate (300 штук). Оптимизировали по 3 цикла. Дотренировали до 18% чувствительности на генетическом уровне. Добавили 100 аннотаций близких родственников с ncbi. Запустили оптимизацию на 20 циклов. Получили чувствительность 53%.
3. Определили > 12к utyjd

Семейства генов

1. Выбрали 4 двукрылых на раздном уровне родства с нашим объектом: D. erecta, D. yakuba, Culex quinquefasciatus, Anopheles gambiae. Выкачали все их белки. Провели две кластеризации: 1. разредили библиотеки каждого вида; 2. найшли родственные группы.
2. Использовали два типа нормализации библиотек: до и после кластеризации.
3. Разобрали клатеры и преобразовали их в формат таблицы CAFE.
4. Построили филогенетическое


2695
4069
D. erecta
1604
3756
2463
3850
D. yakuba
270
16295
3127
9469
D. melanogaster
5301
7725
C. quinquefasciatus
209
19392
4564
7798
A. gambiae


2673
4315
D. erecta
1523
3672
305
16083
2452
4175
D. yakuba
3215
10135
D. melanogaster
5300
7480
C. quinquefasciatus
216
18332
4643
7892
A. gambiae

Good bye,