

Сборка района центромер

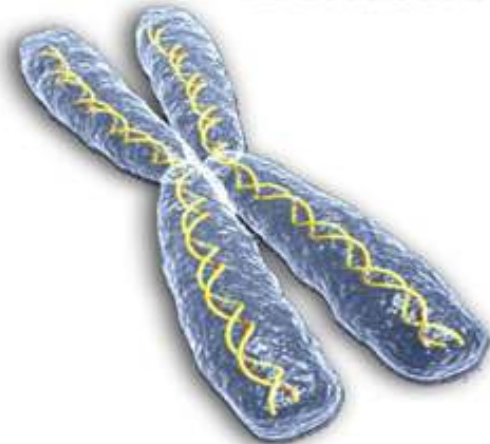
Студенты:

Николай Кочкин

Олег Магнес

Руководитель проекта:

Алексей Комиссаров



Центромеры

- Участки хромосом, не имеют характерной последовательности, в основном образованы повторами (но не только ими!). Например: α -сателлитная ДНК у человека и FA-SAT у домашней кошки являются преобладающими, но не единственными
- Важная роль в клеточном делении: сборка кинетохора, участие в соединении сестринских хроматид, конъюгации гомологичных хромосом
- В норме представляют собой конститутивный гетерохроматин, такое состояние необходимо для корректного выполнения функций
- ! Район центромер плохо собран или не собран вовсе

Задачи



Centromere reference models for human chromosomes X and Y satellite arrays

Karen H. Miga, Yulia Newton, Miten Jain, et al.

Genome Res. published online February 5, 2014

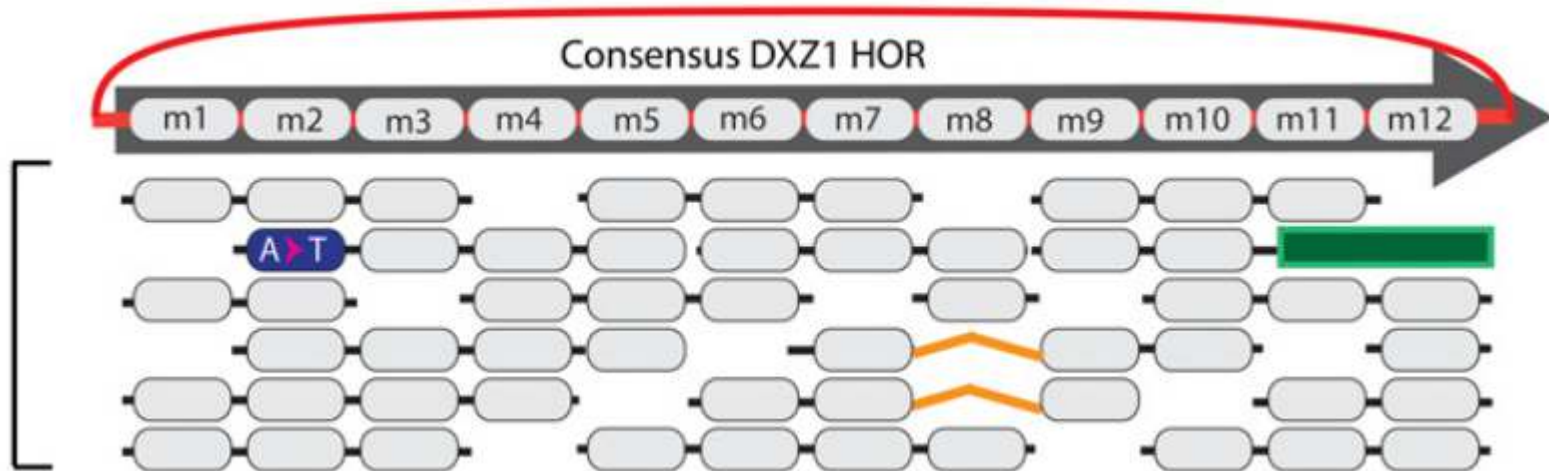
- Повторить алгоритм на человеке
- Сравнить получившиеся данные с данными из статьи
- Сделать то же самое для домашней кошки

План работы

- Из Repbase мы получили последовательности в несколько килобаз с интересующими нас повторами
- При помощи bowtie2 и samtools выравнивали риды на последовательности с повторами, отобрали только те, которые подошли
- Собрали статистику 23- и 31- меров в выбранных ридах, пользовались jellyfish
- Отбросили ошибки сиквенса (к-меры, встречающиеся одиножды)

План работы

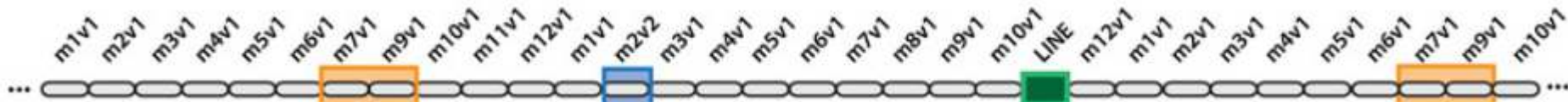
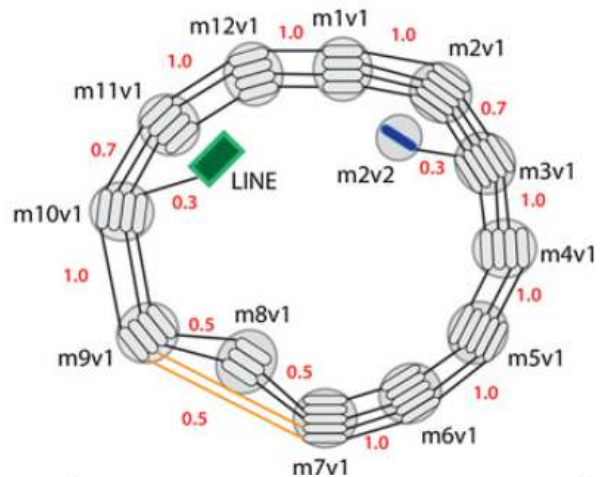
- Представление ридов в виде последовательностей k-меров с редкими вставками информации, не относящейся к повторам.



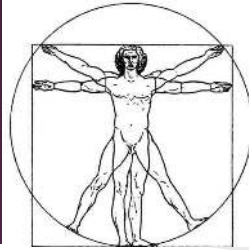
План работы

- Использование тула, созданного авторами статьи.
LinearSat использует марковские цепи второго порядка с модификацией для того, чтобы все элементы встречались хотя бы раз

- ❖ Сравнение наших результатов с результатами из статьи для человека.
- ❖ Анализ наших результатов, полученных на кошке



Local



Global

13906132 reads; of these:

13906132 (100.00%) were paired; of these:

13728712 (98.72%) aligned concordantly 0 times

23956 (0.17%) aligned concordantly exactly 1 time

153464 (1.10%) aligned concordantly >1 times

13728712 pairs aligned concordantly 0 times; of these:

4598 (0.03%) aligned discordantly 1 time

13724114 pairs aligned 0 times concordantly or discordantly;

27448228 mates make up the pairs; of these:

27422211 (99.91%) aligned 0 times

11449 (0.04%) aligned exactly 1 time

14568 (0.05%) aligned >1 times

1.40% overall alignment rate

13906132 reads; of these:

13906132 (100.00%) were paired; of these:

13895722 (99.93%) aligned concordantly 0 times

9211 (0.07%) aligned concordantly exactly 1 time

1199 (0.01%) aligned concordantly >1 times

13895722 pairs aligned concordantly 0 times; of these:

5339 (0.04%) aligned discordantly 1 time

13890383 pairs aligned 0 times concordantly or discordantly; of these:

27780766 mates make up the pairs; of these:

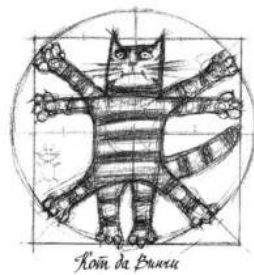
27727921 (99.81%) aligned 0 times

28878 (0.10%) aligned exactly 1 time

23967 (0.09%) aligned >1 times

0.30% overall alignment rate

Local



Global

315561392 reads; of these:

315561392 (100.00%) were paired; of these:

313227651 (99.26%) aligned concordantly 0 times

147330 (0.05%) aligned concordantly exactly 1 time

2186411 (0.69%) aligned concordantly >1 times

313227651 pairs aligned concordantly 0 times; of these:

11034 (0.00%) aligned discordantly 1 time

313216617 pairs aligned 0 times concordantly or discordantly; of these:

626433234 mates make up the pairs; of these:

626219407 (99.97%) aligned 0 times

55811 (0.01%) aligned exactly 1 time

158016 (0.03%) aligned >1 times

0.78% overall alignment rate

315561392 reads; of these:

315561392 (100.00%) were paired; of these:

315550442 (100.00%) aligned concordantly 0 times

10782 (0.00%) aligned concordantly exactly 1 time

168 (0.00%) aligned concordantly >1 times

315550442 pairs aligned concordantly 0 times; of these:

38680 (0.01%) aligned discordantly 1 time

315511762 pairs aligned 0 times concordantly or discordantly; of these:

631023524 mates make up the pairs; of these:

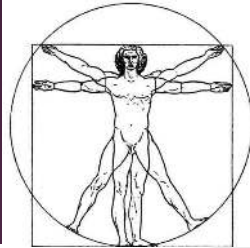
630091191 (99.85%) aligned 0 times

455877 (0.07%) aligned exactly 1 time

476456 (0.08%) aligned >1 times

0.16% overall alignment rate

Local



Global

23-меры:

Unique: 1131139
Distinct: 1875024
Total: 30763681
Max_count: 28825

31-меры:

Unique: 1695553
Distinct: 2624726
Total: 27634953
Max_count: 18844

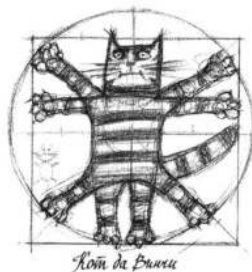
23-меры:

Unique: 132161
Distinct: 221742
Total: 6636474
Max_count: 17514

31-меры:

Unique: 188988
Distinct: 304564
Total: 5959698
Max_count: 13689

Local



Global

23-меры:

Unique: 9249262

Distinct: 12591094

Total: 382463406

Max_count: 5036261

31-меры:

Unique: 14046037

Distinct: 18516159

Total: 343236390

Max_count: 4257658

23-меры:

Unique: 609570

Distinct: 1049134

Total: 80464254

Max_count: 281206

31-меры:

Unique: 979755

Distinct: 1589909

Total: 72211510

Max_count: 235869

Использование linearSat

linearSat input:

- ряды, представленные как последовательности k -меров
- структура HOR - вариации k -меров
(получены, используя марковские цепи)

Проблема

Марковская модель слишком разрозненная.

Определить структуру NOR не получается.

Способы решения:

- попробовать граф де Брюйна, выкидывая редкие k -меры (что-то около 3%), отрезая небольшие куски ридов от начала и конца (чтобы начинался и заканчивался рид на часто встречающихся k -мерах). Далее, можно было бы посмотреть на длинные куски графа, игнорируя короткие ответвления и редкие k -меры, попробовать из этого понять, как устроены NOR.
- попробовать упростить марковскую модель

Спасибо за внимание!

