Оценка качества секвенирования

Старостина Екатерина
Суворов Владимир

Руководители проекта:
Добрынин Павел
Комиссаров Алексей

2013г.

Постановка задачи

Цель проекта - создание инструмента для контроля качества и фильтрации ридов.
Типы загрязнений:

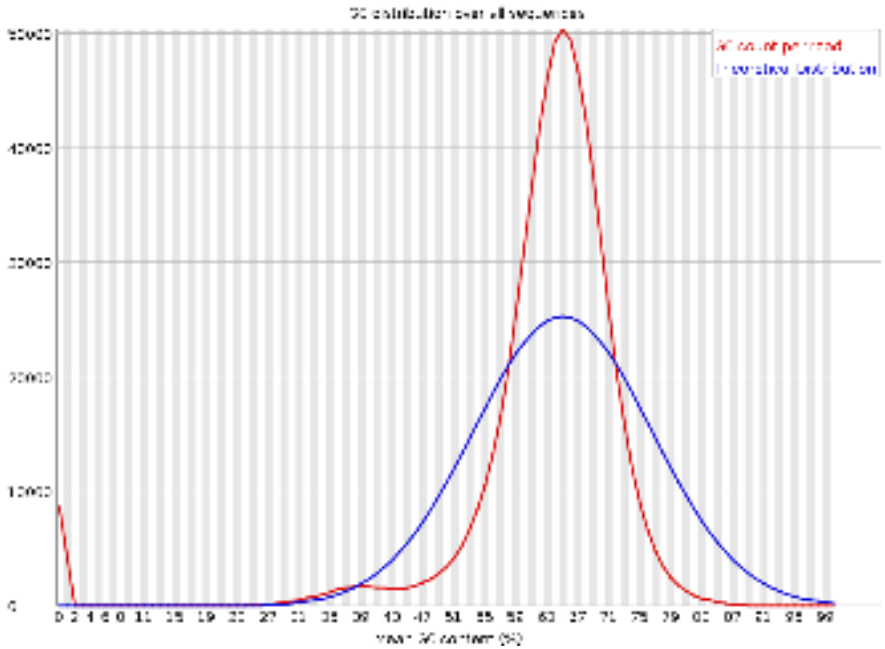
- **Загрязнения чужеродным организмом**
- **Праймеры и адаптеры**
- Артефактные риды
- Ошибки, специфичные для технологии секвенирования

Загрязнения чужеродным организмом

Загрязнение туберкулеза человеком
GC- human ~ 40 , GC HPV ~ 66

2 вопроса:
1. Что за загрязнение?
2. Как убрать?

3. и побыстрее!!!...



Что за загрязнение?

- Экспертное решение
- GC- распределение (FastQC)
  - Не точно
  - Не всё можно определить

  Human – 41.6%, Cat – 41.7% (NCBI)
- BLAST
  - Надо правильно выбрать риды
  - Медленно

Как убрать

- Кластеризация по GC-контенту
- Выравнивание на референс загрязнителя
- Собрать в контиги, а потом использовать всякие алгоритмы на графах и клевую визуализацию

  (http://img.jgi.doe.gov/w/doc/SingleCellDataDecontamination.pdf)
- Кластеризация ридов по длинным уникальным k-мерам

  (Algorithms Mol Biol. 2012 Separating metagenomic short reads into genomes via clustering)

Кластеризация по коротким k-мерам

Пусть известны загрязнители

- Выбираем n/2 k-меров (из $n = 2^k$), наиболее различающих загрязнителей (сортировка по diff)
- Считаем распределение ридов по выбранному множеству k-меров
- Считаем параметры распределений используя Vector Generalized Linear and Additive Model (например, используя библиотеку R)
- Осуществляем префильтр по наивной Баесовской модели с отсечкой
- То что под сомнением - BLAST

Достоинства и недостатки

Достоинства

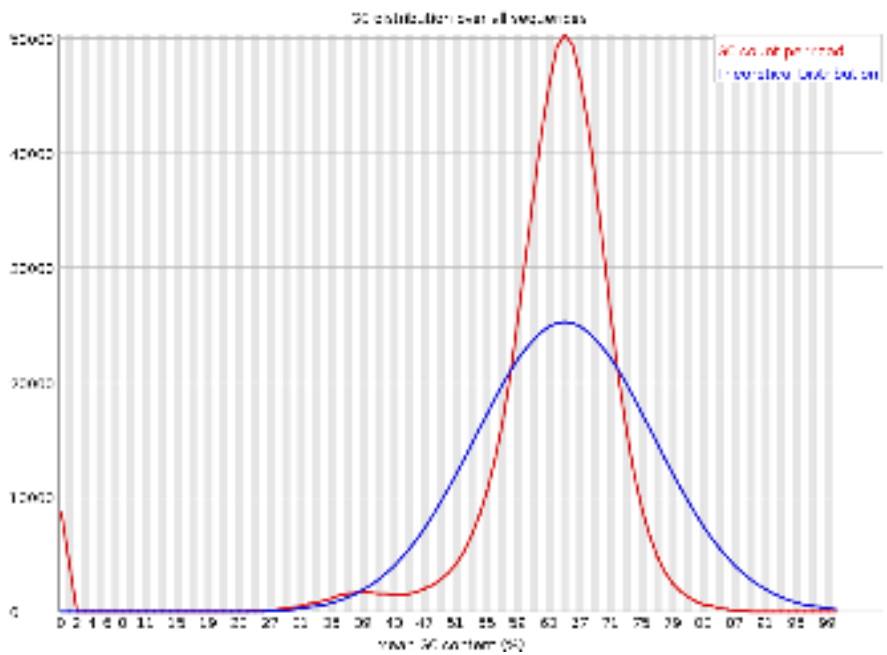
- Скорость
- Решает проблему неоднозначности GC-контента
- Решает проблему коротких ридов – маленькой избирательности распределений больших k-меров
- Сильно помогает разнести распределения
- Выступает для предварительной кластеризации ридов

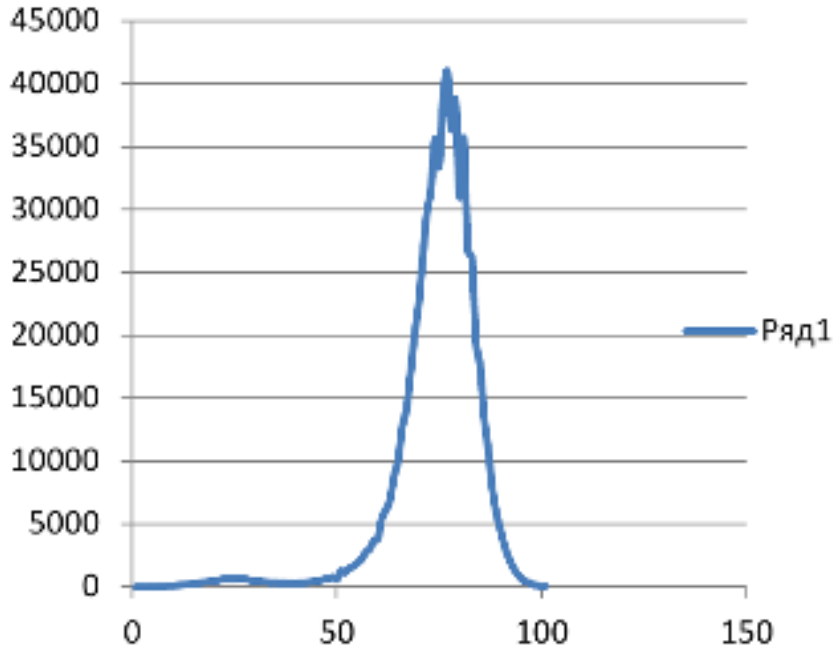
Недостатки

- Сильно растет пространство состояний C(n/2,n)
- Нужна база для существующих геномов

Немного результатов

- Туберкулез vs Человек (mean): GC – 66 vs 41 ; 2-mer 66 vs 37 ; 3-mer – 36 vs 73 4-mer – 76 vs 32





- Кошка vs Человек (mean) : GC – 41.7 vs 41.6 ; 3-mer – 43.6 vs 41.3 ; 4-mer: 43.7 vs 40.0

Борьба с недостатками

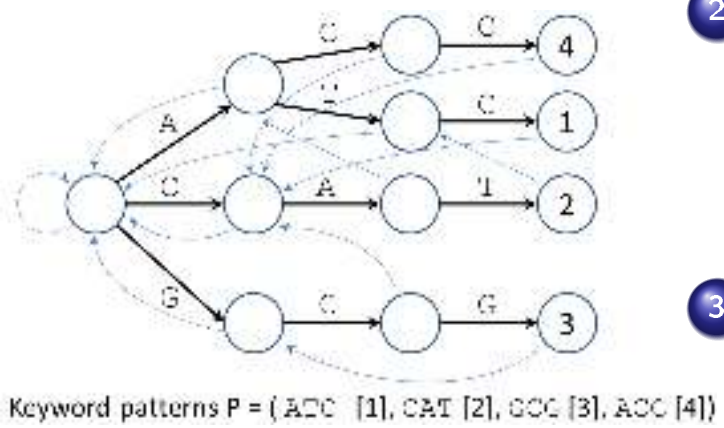
- GC – контент это частный случай этого метода
- На самом деле нам не нужно ВСЁ пространство состояний. Достаточно ограниченного набора комбинаций
- Скорость не сильно зависит от набора комбинаций и размерности k-меров, пока они короткие =)

Праймеры и адаптеры

Разработана программа, которая умеет:

- Искать и удалять риды с адаптерами - точно и с 1-2 ошибками замены
- Искать и удалять polyG, polyC последовательности
- Фильтровать low-complexity риды
- Фильтровать риды с N
- Фильтровать короткие риды

Алгоритм



Keyword patterns P = { ACC [1], CAT [2], GGG [3], ACC [4]}

1. Строим бор из адаптерных, polyG и polyC - последовательностей
2. В каждом риде ищем адаптеры по бору, отбрасываем риды с адаптерами
3. Для фильтрации low-complexity - ридов реализован dust-фильтр
4. Для поиска с ошибками рид, адаптеры делятся на сиды, которые ищутся точно, а затем расширяются

Спасибо за внимание!