

# Про ДНК, исправление ошибок и Echo

Дмитрий Грошев



СПбГУ

26.10.2012

# Общие понятия

- ▶ ...ACGGGCACACTTACG...
- ▶ 4 нуклеотида: аденин, тимин, цитозин, гуанин (но это неважно)
- ▶ ДНК → РНК → белки → жизнь
- ▶ 4 миллиарда нуклеотидов у человека, 4Gb данных

- ▶ чтение последовательности ДНК
- ▶ текущая технология — прочтение случайных участков по 100
- ▶ хорошее покрытие — 30+ (4Gb→120Gb)
- ▶ очень сложно собирать
- ▶ неидеальное прочтение

# Ошибка прочтения

\_\_\_GCTG\_\_\_ <- рид  
GCGC\_\_\_  
\_\_\_GCAC  
\_\_\_CTAC\_\_\_  
\_\_\_TGCA\_

# Ошибка прочтения

GCGC\_\_\_\_  
\_\_GCTG\_\_  
\_\_CTAC\_\_ < - - -  
\_\_\_\_TGCA\_  
\_\_\_\_GCAC

# Ошибка прочтения

GCGCTGCAC

GCGC\_\_\_\_\_

\_\_\_GCTG\_\_\_

\_\_\_CTAC\_\_\_ < - - -

\_\_\_TGCA\_

\_\_\_GCAC

- ▶ k-мер
- ▶ замена
- ▶ индел (indel, insertion/deletion)



- ▶ Геном достаточно неравномерен, чтобы риды в большинстве случаев сильно отличались друг от друга
- ▶ Покрывание каждого нуклеотида значительно больше 1
- ▶ Секвенаторы Illumina редко делают инделов, чаще замены
- ▶ Секвенаторы Illumina делают больше ошибок ближе к концу рида
- ▶ Ошибки обычно независимы друг от друга и зависят от исходного нуклеотида ( $A \rightarrow T \neq A \rightarrow C$ )

Один из достаточно успешных тулов — ECHO (Wei-Chun Kao, Andrew H. Chan, and Yun S. Song, University of California, Berkeley)

# Как работает ECHO

## 2 стадии

- ▶ поиск пересекающихся ридов (соседей)
- ▶ исправление ошибок



# Точные соседи

\_\_\_\_GCGCTGCACAGTTCGAG  
ATATGCGCTGCACAG\_\_\_\_  
^^^^^^^^^^

# Неточные соседи

\_\_\_\_ GCGCTTCACAGTTCGAG  
ATATGAGCTGCACAG \_\_\_\_\_  
  ^  ^^^  ^^^^^

# Неточные соседи

\_\_\_\_\_GCGCTTCACAGTTCGAG  
ATATGAGCTGCACAG\_\_\_\_\_

^ ^^^ ^^^^^

12345678901

- ▶  $\omega = 11$
- ▶  $\epsilon = 2$

Для каждого ряда:

- ▶ поиск точно совпадающего k-мера hashmap'ом (k подобран полуэмпирически)
- ▶ подбор  $\omega$  и  $\epsilon$  (покрытие конкретного нуклеотида должно иметь распределение Пуассона)
- ▶ поиск неточных соседей по  $\omega$  и  $\epsilon$



- ▶ Для каждой позиции есть набор значений, нужно наиболее вероятное
- ▶ Вероятное в данном случае = Maximum A Posteriori (MAP)
- ▶ Вводится понятие confusion matrix
- ▶ Ищутся наиболее вероятные нуклеотиды в перекрывающихся областях

## Maximum A Posteriori

$$P(A_i) = \prod_{j=1}^N \Phi_{READS_{j,i}, A}$$

$$\arg \max_{X=A,T,G,C} P(X)$$

$$\Phi_{b,b'}^{(m)} = P(r_m = b' | H_m = b)$$

$r_m$  — нуклеотид в ряде в позиции  $m$

$H_m$  — нуклеотид в истинном сиквенсе в позиции  $m$

- ▶ Берётся prior confusion matrix (из общих рассуждений или предыдущих экспериментов)
- ▶ В группе соседей делается MAP по каждому нуклеотиду с помощью существующей confusion matrix и строятся исправленные риды
- ▶ На основе исправленных ридов пересчитывается confusion matrix
- ▶ Повторять до схождения (исправленный рид не меняется)

Это EM-алгоритм



Вопросы?  
Слайды:

Использовались картинки:

- ▶ [http://en.wikipedia.org/wiki/File:Thomas<sub>B</sub>ayes.gif](http://en.wikipedia.org/wiki/File:Thomas_Bayes.gif)[http : // mlp.wikia.com/wiki/ File :](http://mlp.wikia.com/wiki/File:Happy_Rainbow_Dash_1E1.png)  
*Happy<sub>R</sub>ainbow<sub>D</sub>ash<sub>S</sub>1E1.png*