Шпаргалка.
Intoshia: долгий путь от сырых ридов к эволюционной истории…

Шпаргалка. Основные термины

- Чтения (риды, reads) – фрагменты ДНК, которые выдаёт секвенатор. Для HiSeq типичная длина 100 нуклеотидов, для MiSeq 250 нуклеотидов.
- Секвенирование pair-end – метод секвенирования, при котором кроме рида читается ещё один рид, на расстоянии обычно несколько сотен нуклеотидов от первого.
- Секвенирование mate-pair – вариант секвенирования pair-end, при котором расстояние между ридами не несколько сотен, а несколько тысяч нуклеотидов. Сборки с использованием ридов mate-pair получаются особенно хорошими.
- FASTA – популярный формат записи последовательностей белков и ДНК.
- FASTQ – специальная модификация FASTA для записи последовательностей ридов.
- Тримминг – удаление из ридов адаптеров и мест низкого качества (с высокой вероятностью ошибки в нуклеотидах).
- Сборка генома – объединение подряд идущих ридов с целью реконструировать как можно более длинные фрагменты генома.
- Контиг – протяжённая последовательность ДНК, которая получается в результате сборки ридов.
- Скаффолд – несколько контигов, которые не удалось соединить в один, но для которых известно, что они в геноме идут подряд.
- N50 – если суммарная длина всех контигов L, то N50 называется длина такого контига, что все контиги больше него дают в сумме L/2. Хотя и звучит сложно, но N50 это удобная мера качества сборки. Смысл примерно такой же, как у «средней длины контига», но более грамотно для оценки качества сборки. Чем больше N50, тем лучше сборка. Обычные N50 порядка 10 000 – 100 000 нуклеотидов.
- Структурная аннотация – поиск генов в последовательностях ДНК, определение их экзонов и интронов.
- Функциональная аннотация – определение, какую функцию выполняет данный ген.
- Ортологи – гены-гомологи между разными видами. Например, ген определённой субъединицы ДНК-полимеразы у человека и ген той же субъединицы ДНК-полимеразы у мыши.
- Паралоги – гены гомологи в пределах одного генома. Например, ген гемоглобина A и ген гемоглобина B - паралоги, поскольку произошли через дупликацию одного гена.