Данные секвенирования IonTorrent

Руководитель: Брагин А.Г.

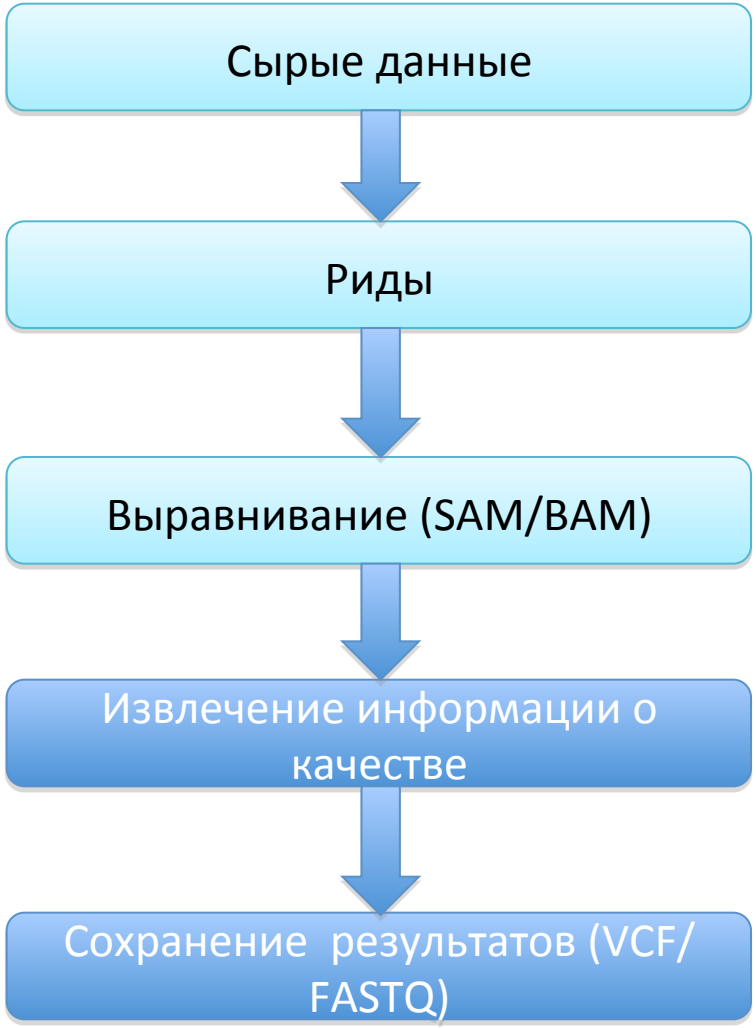
Цель

Разработка модулей для автоматизации работы с данными, потребляемыми и получаемыми в процессе секвенирования с использованием Ion PGM (Life Technologies), и реализация полученного функционала в формате плагина для Torrent Suite/Torrent Server.

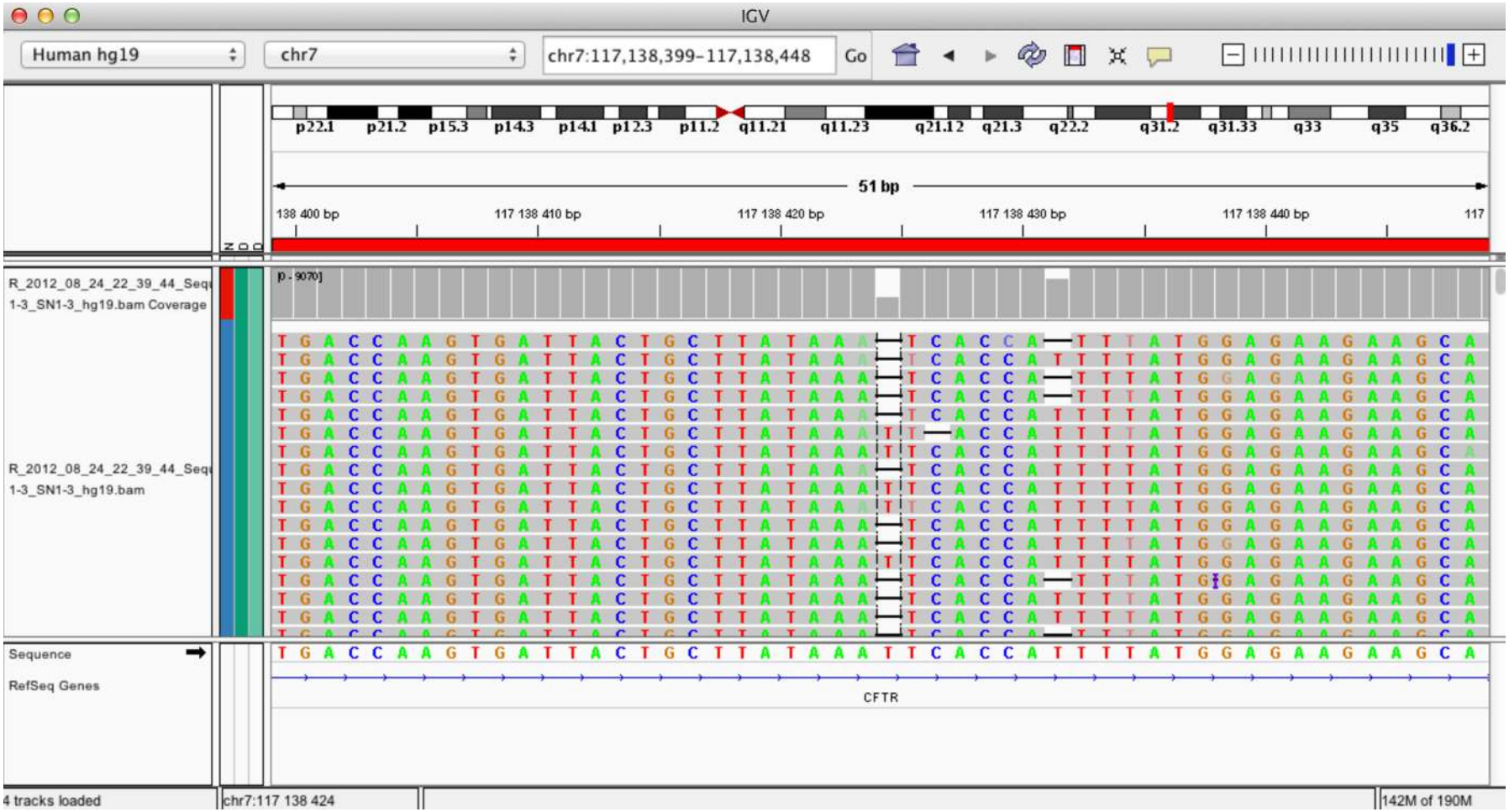
Задачи

1.Создать приложение, позволяющее автоматизировать процесс извлечения информации о качестве определения последовательности маркерных участков, описанных в файле аннотации (BED), из файлов, полученных в результате секвенирования и выравнивания (BAM), и экспортировать эту информацию в виде fastq файла.
2.Интегрировать приложение в Torrent Suite в виде плагина.
3.Подготовить пользовательскую документацию для релиза данного плагина.

Основные этапы:



Выравнивание



Integrative Genomic Viewer с ридами, выравненными относительно референсного генома

Извлечение информации о качестве

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Структура SAM файла

Сохранение результатов

@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;9;7;;.7;393333м

Пример FASTQ файла

Основные этапы обработки:

1. **Определение интересующих диапазонов в геноме на основании данных BED файла.** Многие риды могут приложиться несколько раз, и расстояние между подобными участками может достигать несколько сотен kbp. Для решения данной проблемы используется BED-файл, задающий интересующий диапазон, например:

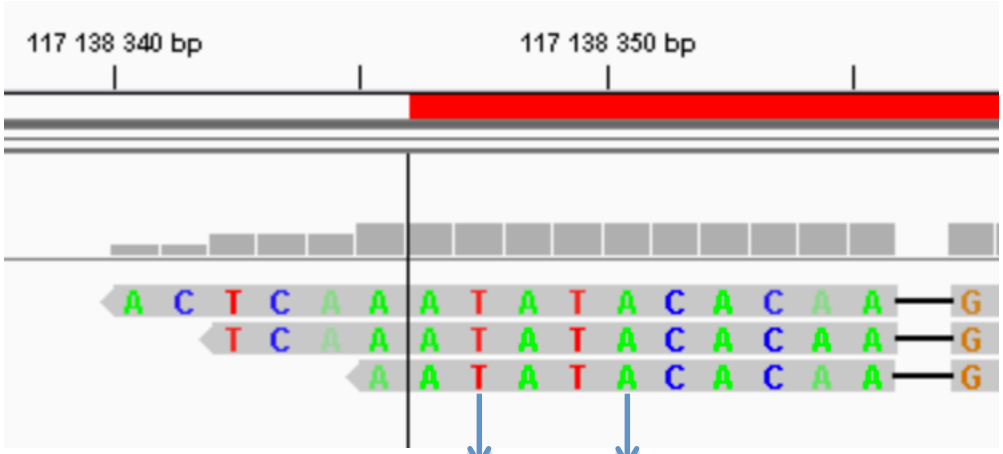
   chr7 117138345 117138546
   chr7 117149055 117149200
   chr7 117159429 117159636

2. **Извлечение данных о качестве нуклеотидов, попадающих в диапазон, из SAM-файла.** Использовалась библиотека samtools, которая позволяла извлекать все риды, пересекающие заданный диапазон

3. **Восстановление рида и его качества по CIGAR-строке (добавление делеций)**
   4M1D3M TCAAATA -> TCAA-ATA

Основные этапы обработки:

4. Вычисление среднего качества для данной позиции



$$\frac{1}{n}\sum_{i \in reads}^{n} QUAL_i$$

5. Вывод значения среднего качества в fastq:

@BED:117138345 - 117138546(A)
AATATACACAAGGCTTGTCTTTAGCGAGCATATACTCCCTAAAGTTGATTAAGCTGACCAAGTGATTACTGCTTATAA
ADTCACCATTTTATGGAGAAGAAGCAAACACTGCTAAATACCTTGTGGAATCAGAGGAGGGGAAATTAGTAACTTG
ACCCCAATACTGCGATTTTAAATTGAATTCTTGAAGCCTACAAGTTT
+
@9.?=96:*?<2@CAD>A*=*6?9C?A?6AFAE?A47A>7@B=:0??@*D@?:**B@*@?=:7>A-AA5AA8D*?4?
@5-37>7D929649*??=B?>?@>74?A<4@*B==*>@;3@3?=D?>>;=;4+@@?:7=58?7<@:
12>94><166;=24;3=<3>*<>A;;36>4:>6*-6=8+**6*7><76*4>4=:06=@

Полученные навыки:

1. Изучение форматов хранения данных: SAM/BAM/BED/FASTQ/VCF
2. Работа с библиотеками для обработки геномных данных: samtools/ picardtools
3. Написание модулей для IonTorrent Server (Ubuntu)

Обнаруженные трудности:

1. Большой объем данных
2. Риды могут выравниваться на разные участки ДНК. Необходимо задавать интересующие участки с помощью BED-файла
3. Структуру каждого рида (наличие вставок/удалений/замен) из BAM необходимо восстанавливать по CIGAR-string

Спасибо за внимание