# Описание отчета (что получает пользователь)

|  |  |
| --- | --- |
| **Вход** | **Выход** |
| Xml по плашкам  \*Параметры качества | Частота по качеству  Статистика и частоты по капиллярам  Статистика и частоты по стрипу |
| \*Таблица с разметкой праймеров | \*Статистика и частоты по праймерам |
| \*Несколько плашек | \*Статистика и частоты по плашкам |

Отчет показывает частотное распределение качества загруженных ридов.

Если данные были проанализированы, в отчете отражается сводка качества, а та же информация о параметрах, на него влияющих.

Наша программа автоматически извлекает данные из файла отчета об анализе Sequencing Analysis Software 6, что сводит к минимуму необходимость ручного поиска интересующих параметров пользователем. Пороговые значения могут быть настроены с использованием диалогового окна.

Влияние выбранных параметров на качество исследования определяется методом сравнения 95%-х доверительных интервалов выборки. В случае, если они не пересекаются – имеет место статистически значимых различий между выборками, т.е. наличие влияния выбранного параметра на качество постановки. В случае, если доверительные интервали параметра лежат в пределах близких значений, влияния на качество не отмечается.

При желании пользователь может указать сопоставление имен с используемыми праймерами и получить дополнительный статистический анализ их влияния на качество постановки.

Кроме того, программа позволяет вносить данные различных постановок и объединять их в один отчет, получать дополнительную информацию о влиянии конкретное постановки на качество исследования, что может быть интерпретировано как участие конкретного исследователя. Имена входных файлов автоматически сокращаются или переименовываются, чтобы обеспечить компактность аннотаций по осям фигур на графиках. При желании пользователь может изменить сопоставление имен и назначить новые имена файлов глобально

Отчет используется для устранения неполадок и простой оценки качества данных. Его можно экспортировать в виде файла с разделителями табуляции и открыть в программном обеспечении Microsoft ® Excel.

Оценить состояние оборудования (капилляры и время забора – стрипы)

# Описание методологии

Получая данные отчета об анализе Sequencing Analysis Software 6, наша программа использует ряд параметров для оценки качества каждого рида. Так, при стандартных настройках параметров, рид считается качественным, если выполняются оба условия: sample score выше 25 и длина Clear region выше 350 п.н.

Длина Clear region рассчитывается с использованием аргументов CR\_start и CR\_stop, заложенных в схему отчета об анализе Sequencing Analysis Software 6.

Значение аргумента Качество является бинарным и может принимать два значения: 1 – рид высокого качества, 0 – рид низкого качества.

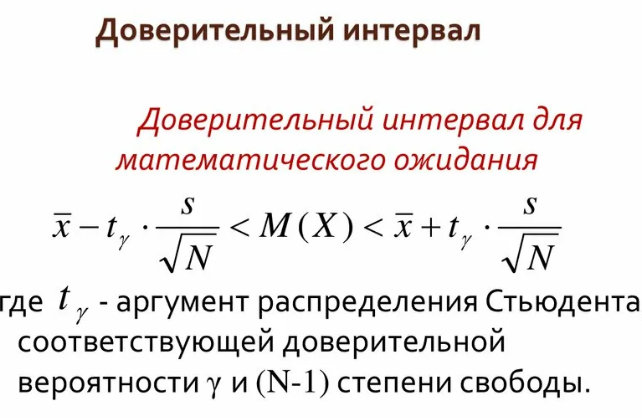
На экран выводится 2 группы ридов: с высоким и низким качеством, с указанием частот и встречаемости в общей выборке.

В условиях наличия всевозможных данных (соответствие праймера имени файла, наличие нескольких постановок) программа может оценить влияние на качество 4 параметров: используемый анализатором капилляр для данной пробы, положение ее на плашке, т.е. в определённом стрипе, соответствие пробы конкретной постановке, используемый праймер в предшествующих реакциях.

Влияние конкретных параметров на качество исследования определяется методом сравнения 95%-х доверительных интервалов выборки. В случае, если они не пересекаются – имеет место статистически значимых различий между выборками, т.е. наличие влияния выбранного параметра на качество постановки. В случае, если доверительные интервалы параметра лежат в пределах близких значений, влияния на качество не отмечается.

Чем меньше область доверительного интервала – тем больше точность измерений.

Нижняя и верхняя граница доверительного интервала считается по формуле:



Степень различия крайних значений рассчитывается с помощью критерия хи-квадрат Стьюдента и на его основе делается вывод о достоверности этих различий. Если значение p-критерия меньше 0,05, то с 95% вероятностью мы утверждаем, что различие есть. В случае множества попарных сравнений p-уровень значимости (0,05) делится на количество сравнений и сравнение происходит уже с ним.

Подход к определению качества   
Отчет об анализе содержит большое количество признаков, но не все из них могут наглядно отразить качество образца.

Так, номер Peak\_1

N\_Low\_QV, N\_Med\_QV, N\_High\_QV, LOR коррелируют с длиной Clear region, рассчитываемому по CR\_Start и CR\_Stop.

Sample\_Score же в свою очередь отражает вероятность ошибки в определении нуклеотида по данному сайту что, гипотетически связано с уровнями A\_\_Signal\_to\_Noise, C\_\_Signal\_to\_Noise, G\_\_Signal\_to\_Noise, T\_\_Signal\_to\_Noise, Avg\_Signal\_to\_Noise , Base\_Spacing.

Согласно руководству к программному обеспечению DNA Sequencing Analysis Software 6 [1]

## Sample Score

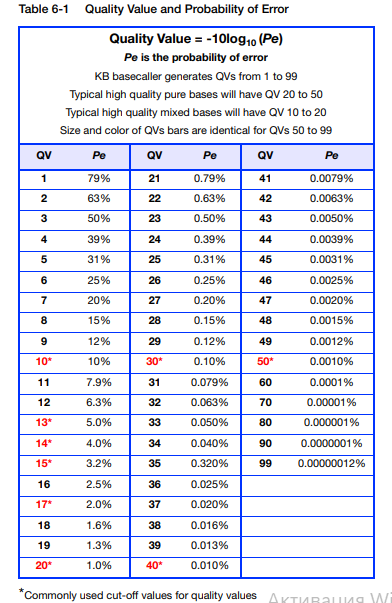
Среднее значение качества оснований в последовательности четких диапазонов.

Выборочная оценка генерируется на основе QVs. Это среднее значение качества оснований в последовательности четких диапазонов для этого образца

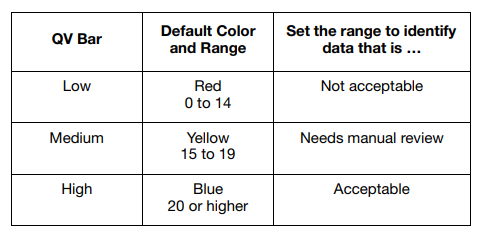
QV 10 –= log10(Pe),

где Pe это вероятность ошибки

Смешанные основания имеют меньший QVs чем чистые.



Рекомендуется считать качество рида высоким, если значение sample score для него превышает значение 20. В нашем анализе уровень отсечения по данному признаку установлен на значении 25, которое было выбрано посредством эмпирического анализа.

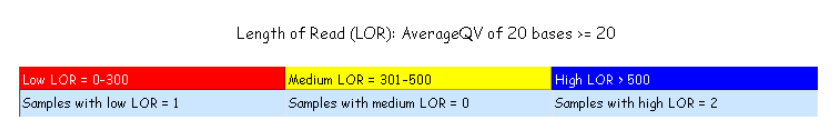


## Clear region

Чистая зона вычисляется исходя из показателей качества для каждого нуклеотида по правилу: удаляйте основания с концов до тех пор, пока не менее чем у 4 оснований из 20 не будет QVs меньше 20.

LOR вычисляется по правилу: Среднее QV для 20 оснований больше или равно 20. Это диапазон высококачественных или высокоточных баз, определяемый значениями качества.

Руководство по эксплуатации рекомендует считать рид качественным, если его длина выше 500 нуклеотидов и удовлетворяющим, если выше 300. На практике во время сбора генома ВИЧ длина рида была достаточной и исчерпывающей, если она составляла не менее 300 нуклеотидов, поэтому стандартное значение отсечения по длине чистой зоны составляет 350 п.н.



При желании пользователь может самостоятельно выставить значения sample score и clear region, зная типичную длину ридов, свойственных для исследования. Вычислить ее можно, используя метод электрофореза.