Выполнила: Таран Юлия,
студентка БКМБ-191

Домашнее задание “ATAC-Seq”

Задание 0. Кратко опишите суть эксперимента.

В статье “Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions” проводится исследование применимости метода ATAC-seq - NucleoATAC для анализа изменений расположения и занятости нуклеосом во время динамического процесса. Модельный эксперимент включал выполнение ATAC-seq на *Saccharomyces cerevisiae*, подвергшихся осмотическому стрессу (увеличивали концентрацию NaCl в среде на 0,6 М в течение 60 минут), для исследования динамики хроматина. Для анализа проводилось сопоставление *Saccharomyces cerevisiae* подвергавшихся (пробы отбирали с интервалом в 15 минут) и не подвергавшихся осмотическому стрессу.

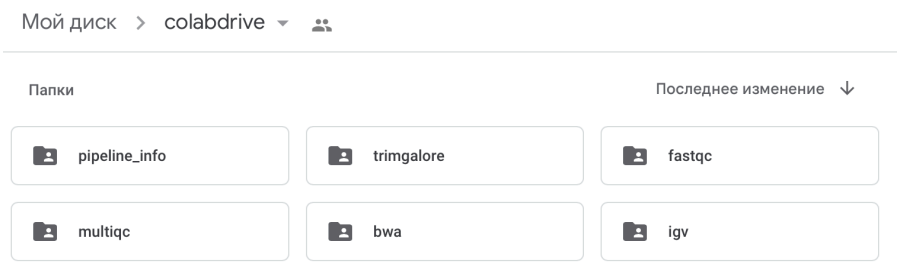
Задание 1. Сколько образцов и сколько реплик используется.

Использовано 32 образца всех 3-х организмов (NCBI Gene Expression Omnibus с accession number GSE66386). Для экспериментов по осмотическому стрессу использовали 6 образцов по 2 повтора.

GSM1621323 S. cerevisiae lin0a
GSM1621324 S. cerevisiae lin0c
GSM1621325 S. cerevisiae lin2a
GSM1621326 S. cerevisiae lin2b
GSM1621327 S. cerevisiae lin2c
GSM1621328 S. cerevisiae lin3a
GSM1621329 S. cerevisiae lin3b
GSM1621330 S. cerevisiae lin3c
GSM1621331 S. cerevisiae lin6a
GSM1621332 S. cerevisiae lin6b
GSM1621333 S. cerevisiae lin6c
GSM1621334 S. pombe_1M_5min
GSM1621335 S. pombe_5M_5min
GSM1621336 S. pombe_20M_5min
GSM1621337 S. pombe_5M_20min
GSM1621338 S. pombe_20M_20min
GSM1621339 Osmotic Stress Time 0 A rep1
GSM1621340 Osmotic Stress Time 0 A rep2
GSM1621341 Osmotic Stress Time 0 B rep1
GSM1621342 Osmotic Stress Time 0 B rep2
GSM1621343 Osmotic Stress Time 15 C rep1
GSM1621344 Osmotic Stress Time 15 C rep2
GSM1621345 Osmotic Stress Time 30 D rep1
GSM1621346 Osmotic Stress Time 30 D rep2
GSM1621347 Osmotic Stress Time 45 E rep1
GSM1621348 Osmotic Stress Time 45 E rep2
GSM1621349 Osmotic Stress Time 60 F rep1
GSM1621350 Osmotic Stress Time 60 F rep2

Задание 2. Приведите содержание вашей директории.

https://drive.google.com/drive/folders/1lUMt6hmz39Q0RzkXLqTsk4yOnCtehIsC?usp=sharing



Задание 3. Приведите имя вашего запуска.

evil_knuth

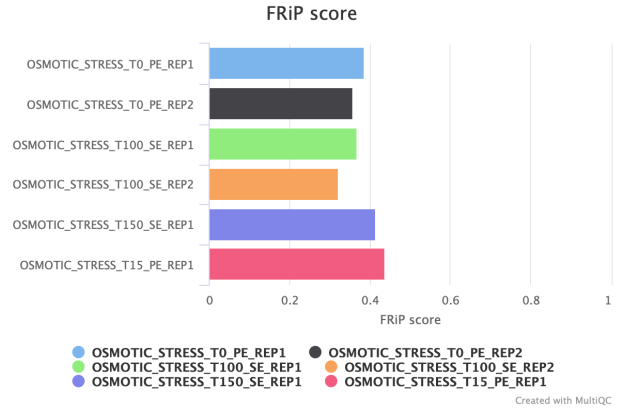
Задание 4. Приведите добавочный параметр для поиска транскрипционных факторов.

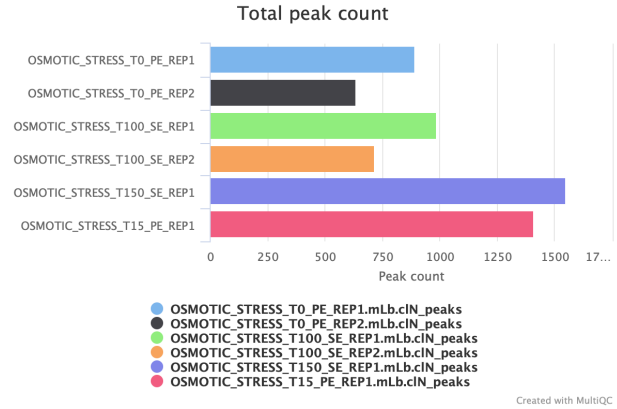
Добавочный параметр – narrow_peak.

Измененная строка запуска:

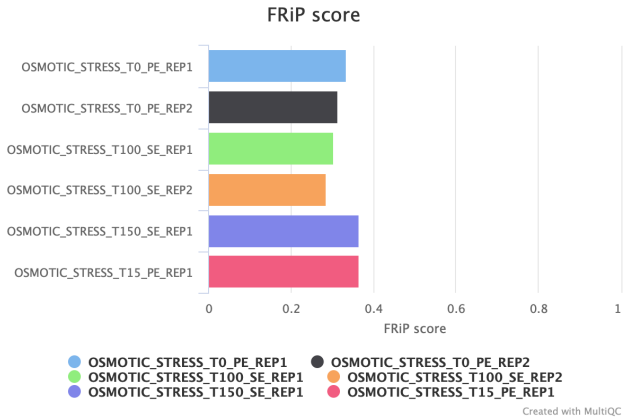
! nextflow run nf-core/atacseq -profile test,conda --outdir /googledrive/MyDrive colabdrive_narrow_peak --narrow_peak --skip_consensus_peaks -resume

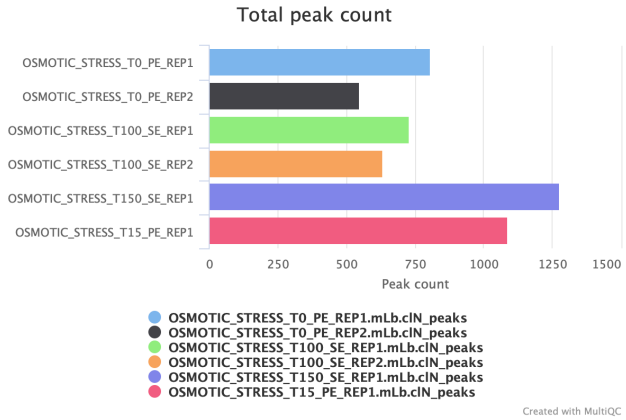
Задание 5. Какова Fraction of Reads in Peaks для двух запусков? Почему? Удовлетворяют ли критерию? Связано ли с количеством пиков?





1) Fraction of Reads in Peaks, 2) Total Counts для запуска с broad peaks.





1) Fraction of Reads in Peaks, 2) Total Counts для запуска с narrow peaks.

Мой образец – OSMOTIC_STRESS_T100_SE_REP1:

FRiP (broad peak) = 0.37,

Peak Counts (broad peak) = 984,

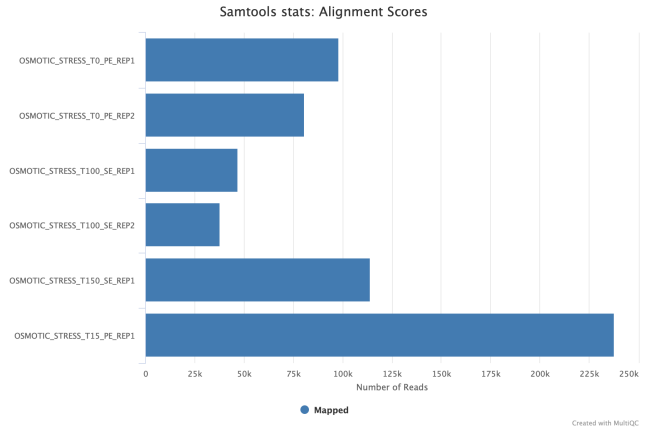
FRiP (narrow peak) = 0.3,

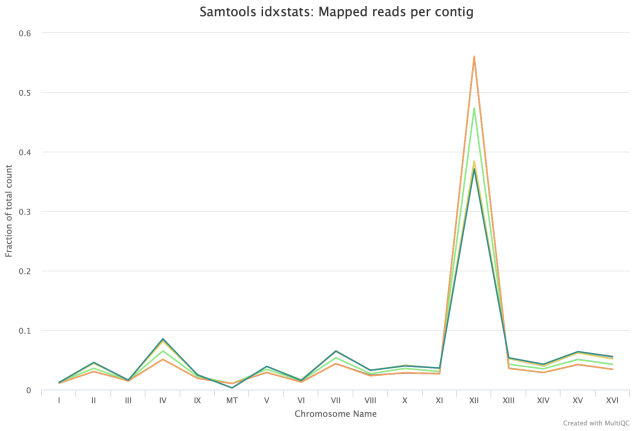
Peak Counts (narrow peak) = 728.

Порог FRiP равен 0.3, в некоторых случая допускается 0.2 (ATAC-seq Data Standards and Processing Pipeline). Для всех образцов FRiP > 0.3, что удовлетворяет критерию. При использовании параметра narrow_peak мы уменьшаем допустимую ширину пиков, из-за этого общее число пиков уменьшается, соответственно, и FRiP доля ридов, покрывающих пики уменьшается.

Задание 6. Достаточна ли глубина секвенирования? У какой хромосомы наибольшее покрытие?

Каждый повтор должен иметь 25 миллионов неповторяющихся, не митохондриальных ридов для single-end секвенирования и 50 миллионов для paired-ended секвенирования (т. е. 25 миллионов ридов, независимо от типа запуска секвенирования) (ATAC-seq Data Standards and Processing Pipeline). Этот критерий применим для млекопитающих, чей размер генома ~ 3.3 Gb пар оснований. Размер генома *Saccharomyces cerevisiae* 12 млн пар оснований, поэтому число ридов должно быть кратно меньше (примерно в 300 раз). По такой очень примерной оценке, мы должны получить иметь по 84 тыс ридов для single-end секвенирования.



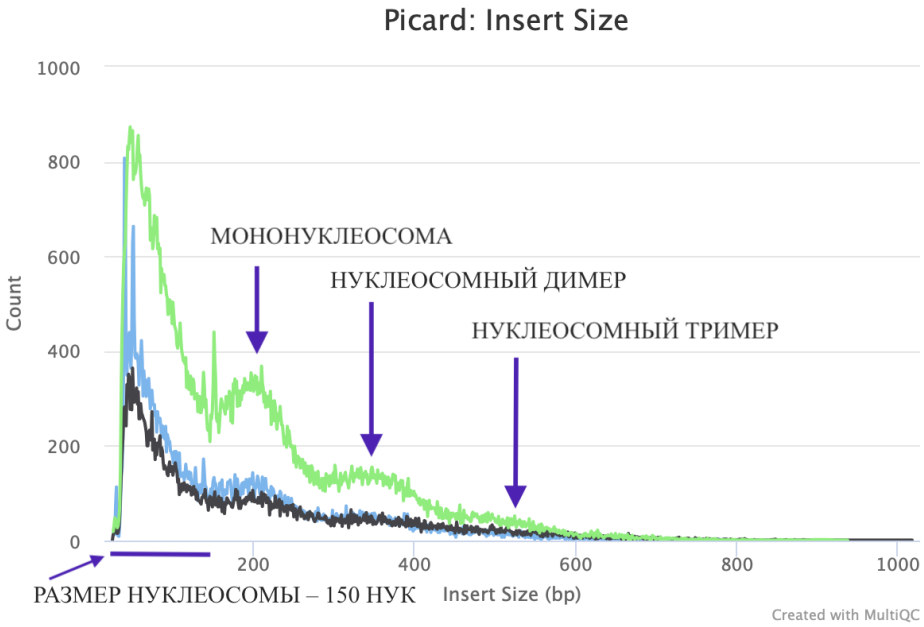


Для моего образца (OSMOTIC_STRESS_T100_SE_REP1) 46 тыс откартированных ридов (после фильтрации) (График alignment scores), что может не соответствовать достаточной глубине секвенирования по моей оценке.

У 12 хромосомы самое хорошее покрытие, для моего образца оно составляет 47% от общего покрытия ридами.

Задание 7. Какой размер нуклеосомы исходя из вашего графика. Отметьте это на графике, добавьте в отчет и поясните.

Размер нуклеосомы можно оценить по пикам длин ридов на Insertion Size графике. Второй пик графика соответствует мононуклеосоме. Длина ридов, соответствующая началу второго пика является приближенной оценкой размера нуклеосомы (т.е. размер нуклеосомы примерно 150 п.н.).



Задание 8. Найдите нуклеосомные димеры, тримеры, тетрамеры. Опишите результаты в отчете и поясните почему.

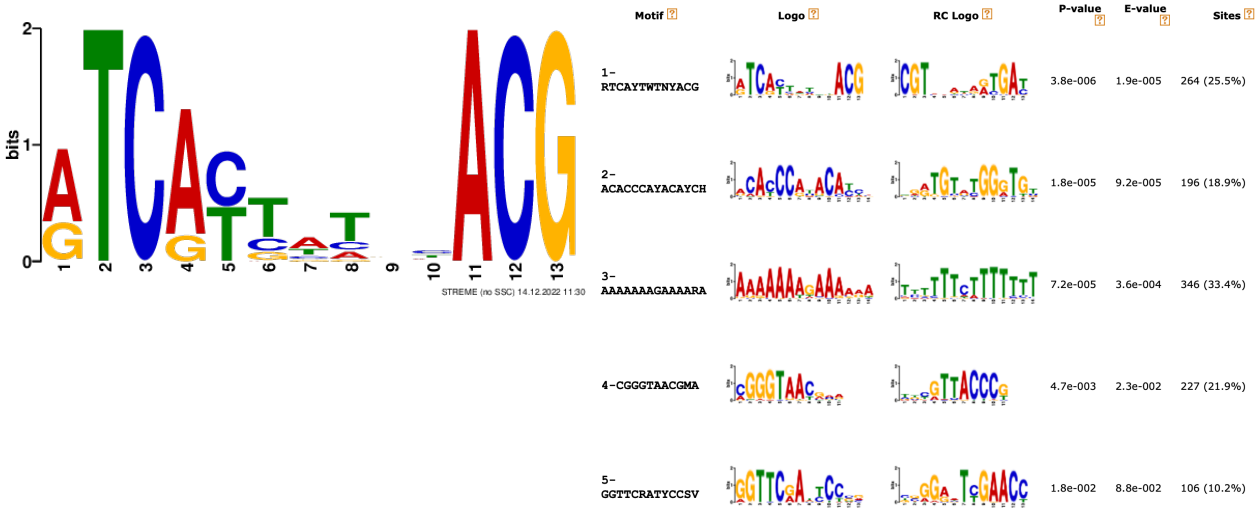
На insertion size графике можно увидеть пики соответствующие мононуклеосоме, нуклеосомным димерам и тримерам. При этом димеры и особенно тримеры уже видны менее отчетливо, пик, соответствующий тетрамерам, уже визуально не различим. Это может быть связано с затруднением посадки транспозазы на участки закрытого хроматина, со скоплением нуклеосом.

Задание 9. Какие нашлись мотивы? Приложите лого лучшего из них. Является ли эта находка статистически значимой?

Для анализа образца OSMOTIC_STRESS_T100_SE_REP1 был взят OSMOTIC_STRESS_T0_PE.mRp.clN_peaks.narrowPeak файл. Геном *Saccharomyces cerevisiae* был взят из Genbank, хромосомы в нем вручную привели в соответсвие с файлом пиков. С помощью пакета bioconda.bedtools перевели narrowPeak файл в FASTA-файл.

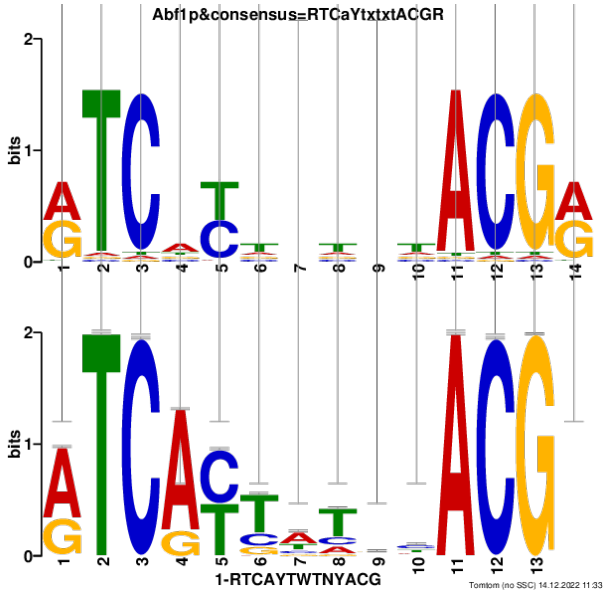
Инструмента MEME выделил следующие мотивы:

Наиболее значимым из них является первый мотив – RTCAYTWTNYACG, он также является статистически значимой.



Задание 10. Запустите поиск похожих мотивов с помощью Tomtom. Приложите лого наиболее похожего. Является ли эта находка статистически значимой? Опишите функцию данного транскрипционного фактора.

Был найден Abf1p&consensus=RTCaYtxtxtACGR (Abf1p) мотив с *p*-value = 7.19e-07. Данная находка является статистически значимой.



ABF1p фактор связывается с репликатором дрожжей, точнее со специфической последовательностью, которая найдена в домене B различных ARS и необходима для репликации.

Задание 11. Что отмечено стрелками? Обоснуйте.

На графике стрелками отмечены сайты посадки транспозазы на ДНК. Эти периодичные пики возникают из-за того, что домен транспозазы обладает сродством с большими бороздками ДНК. Виток спирали ДНК составляет 10.5 нуклеотидов, что соответствует периодичности пиков.