Дудковская Настя, 3 группа

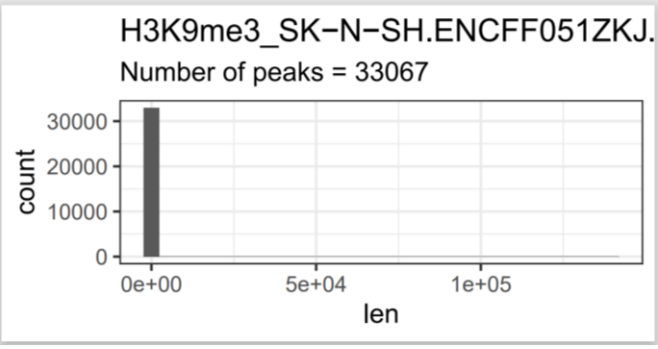
Проект

Для своей работы я выбрала человеческую ДНК, клетку типа SK-N-SH, гистоновую метку H3K9me3, и посмотрела для двух экспериментов места пересечения пиков с вторичной структурой ДНК Z-DNA.

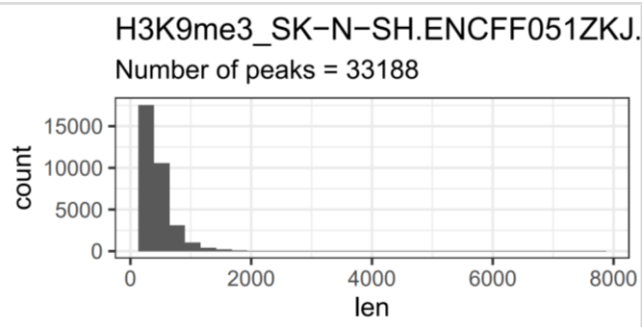
Для проекта я выбрала эксперименты ENCFF051ZKJ (https://www.encodeproject.org/files/ /ENCFF051ZKJ/) и ENCFF231PXT (https://www.encodeproject.org/files/ENCFF231PXT/). В этих файлах находятся пики для генома hg38, однако нам необходим hg19. Для того, чтобы привести в hg19, воспользуемся программой liftOver. Напишем такую строку для файла H3K9me3_SK-N-SH.ENCFF051ZKJ.hg38.bed (аналогично для H3K9me3_SK-N-SH.ENCFF231PXT.hg38.bed):

liftOver H3K9me3_SK-N-SH.ENCFF051ZKJ.hg38.bed hg38ToHg19.over.chain.gz H3K9me3_SK-N-SH.ENCFF051ZKJ.hg19.bed H3K9me3_SK-N-SH.ENCFF051ZKJ.unmapped.bed

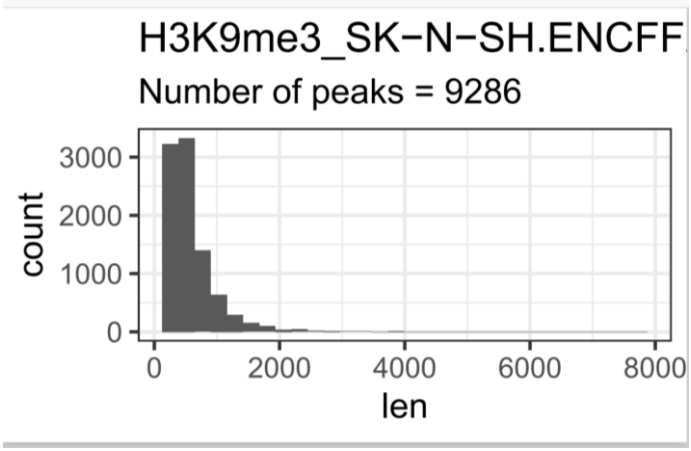
Теперь построим гистограммы длин пиков для каждого файла с помощью R. Код лежит в ‘scr/len_hist’. Получаем такие гистограммы:



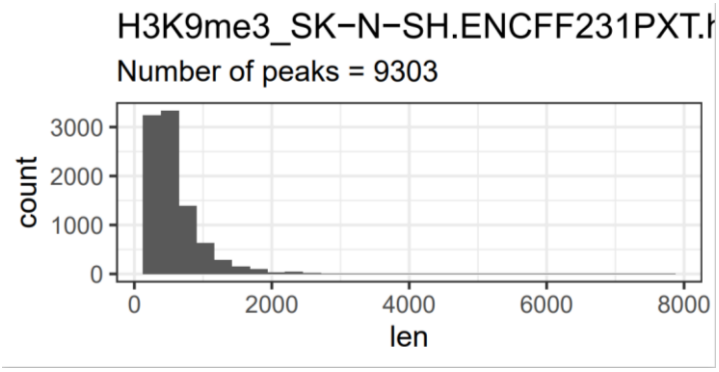
len_hist.H3K9me3_SK-N-SH.ENCFF051ZKJ.hg19



len_hist.H3K9me3_SK-N-SH.ENCFF051ZKJ.hg38

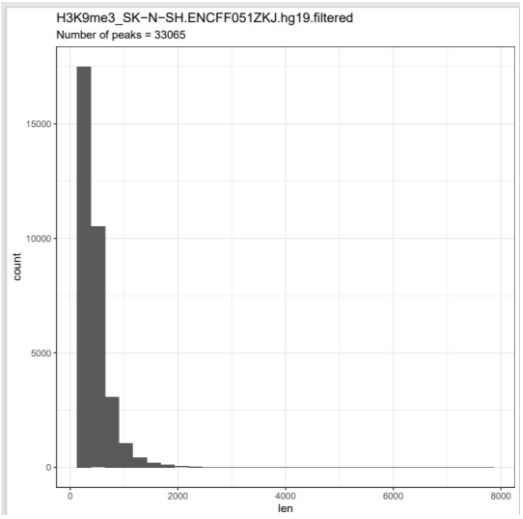


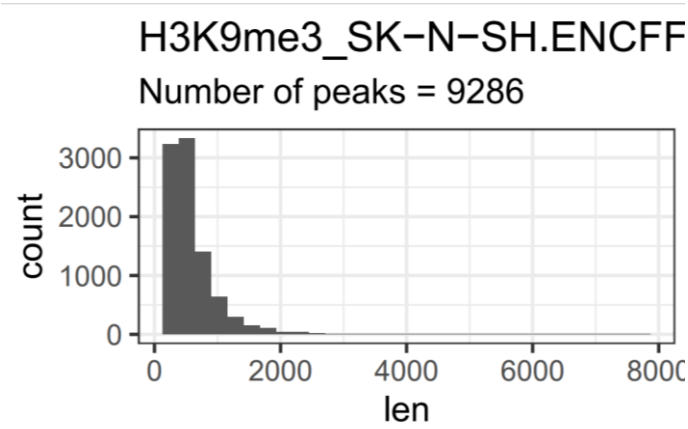
len_hist.H3K9me3_SK-N-SH.ENCFF231PXT.hg19



len_hist.H3K9me3_SK-N-SH.ENCFF231PXT.hg38

По ним видно, что в более коротком файле, в котором 9303 и 9286 пиков соответственно, распределения длин пиков совпадают(оба доходят до 8000, самое большое число до длины 500), однако в более длинном файле, 33188 и 33067 пиков соответственно, данные после конвертации сильно ухудшились. Появились сильно длинные пики, длиной до 10^5. Как видно из гистограмм до перевода в hg19 длина пиков ограничивалась длиной 8000. Поэтому мы удалим все пики, длина которых превосходит 8000 нуклеотидов. Это сделаем с помощью R, код в scr/filter_peaks.



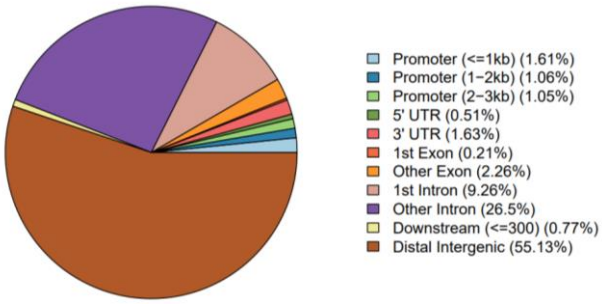


len_hist.H3K9me3_SK-N-SH.ENCFF231PXT.hg19.filtered

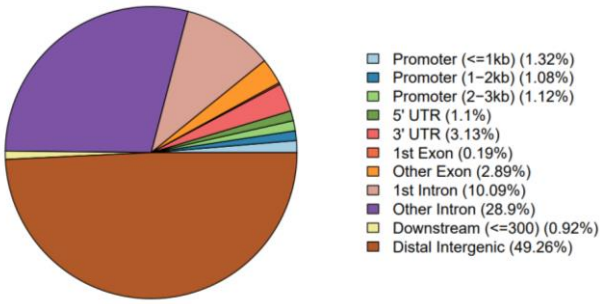
Как можем увидеть, после удаления слишком длинных пиков гистограммы стали совпадать.

Новое количество пиков – 33065 и 9286 соответственно.

Далее посмотрим на то, где наши пики находятся в геноме. Для этого воспользуемся кодом на R, который лежит в src/chip_seeker. Получаем пай-чарты:



chip_seeker.H3K9me3_SK-N-SH.ENCFF051ZKJ.hg19.filtered.plotAnnoPie



chip_seeker.H3K9me3_SK-N-SH.ENCFF231PXT.hg19.filtered.plotAnnoPie

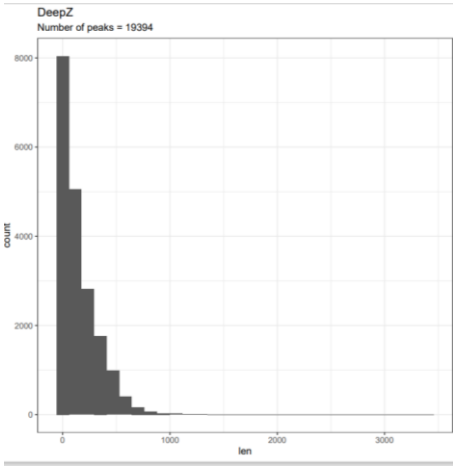
Как мы можем увидеть, то около или больше половины в обоих случаях попадает на межгенные регионы, и около четверти на интроны, кроме первого. Около 10% попадаем на первый интрон. Остальные части распределены более менее одинаково (0.5% – 3%).

Далее объединяем наши .bed файлы в один на сервере с помощью команды:

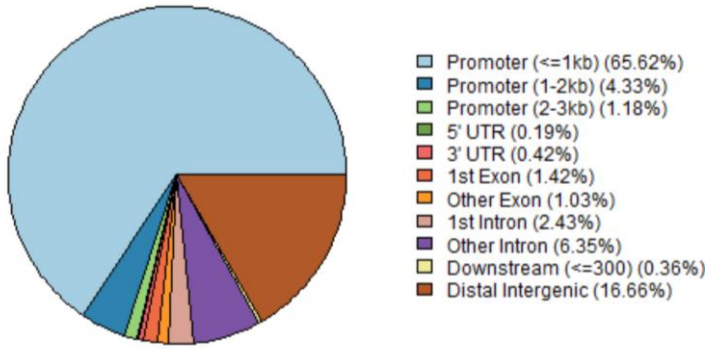
cat *.filtered.bed | sort -k1,1 -k2,2n | bedtools merge > H3K9me3_SK-N-SH.merge.hg19.bed

Вся визуализация в геномном браузере по ссылке: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr16%3A68595206%2D68602891&hgsid=1124144643_P0K2W3jaGi1u1DtgDfXjAZ7uutpB

Теперь смотрим на нашу вторичную структуру из DeepZ. Она имеет такую гистограмму распределения длин:



Также посмотрим на распределение аннотированных генов в DeepZ, с помощью R с кодом scr/chip_seeker

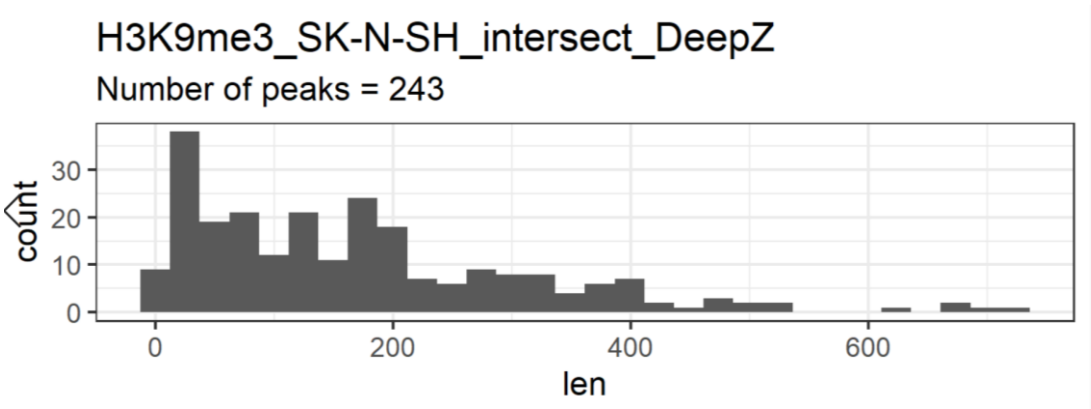


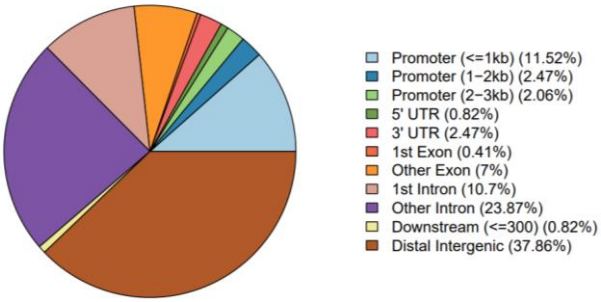
Здесь у нас основная часть лежит в промоутере, причем в начальной его части. В наших экспериментальных файлах большую часть занимала межгенная ДНК, тут она составляет всего 16,6%.

Пересечем наши объединенные пики гистоновых меток с метками Z-DNA:

bedtools intersect -a DeepZ.bed -b H3K9me3_SK-N-SH.merge.hg19.bed > H3K9me3_SK-N-SH..intersect_with_DeepZ.bed

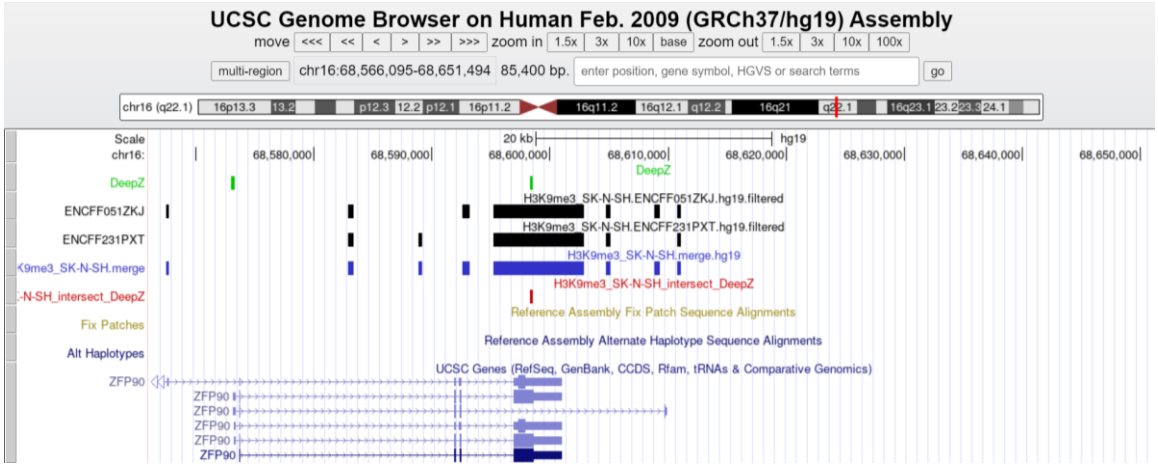
Распределение длин и их количество в пересечении:





При пересечении большую часть опять стали занимать части на межгенном пространстве, а так же промотеры, и не первые интроны.

Визуализация в геномном браузере все по той же ссылке: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr16%3A68595206%2D68602891&hgsid=1124144643_P0K2W3jaGi1u1DtgDfXjAZ7uutpB



chr16:68,566,095-68,651,494

Проаннотируем гены с получившимися при пересечении пиками. Сделаем это опять на R, код в scr/chip_anno

Удалось проаннотировать 28 участков, из них получилось 17 генов.

При GO анализе статистически значимых генов найти не удалось.