Работа со вторичной структурой ДНК. Отчёт. (Пахалко Илья Александрович)

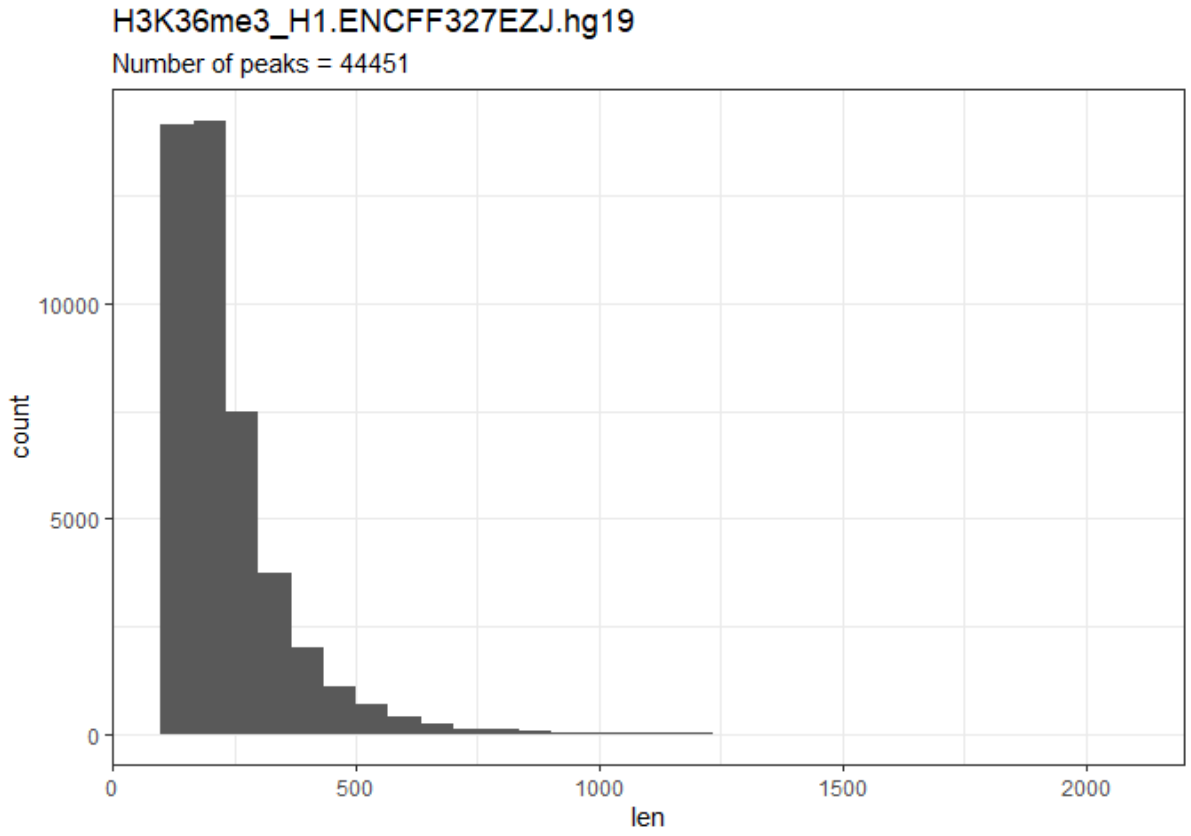
1. Скачиваем наши эксперименты и распаковываем, оставляя всего 5 колонок

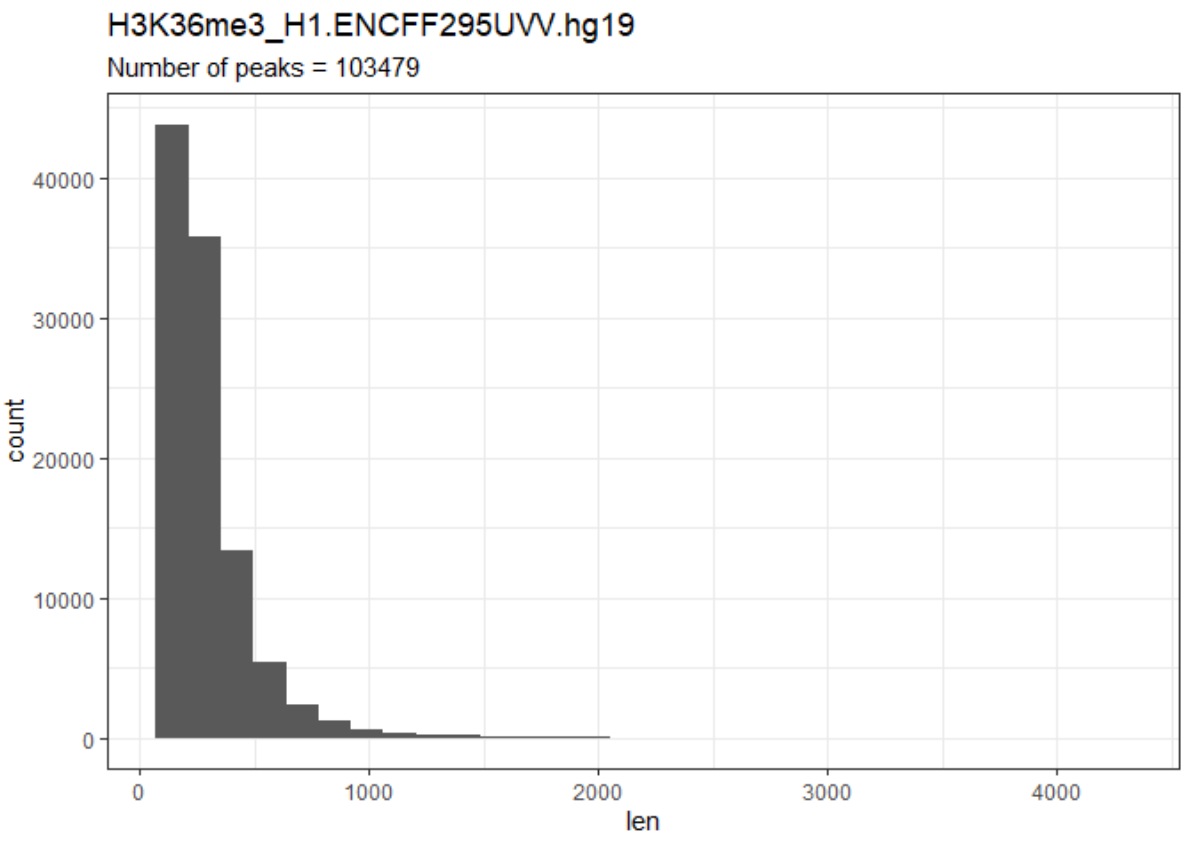
```
zcat ENCFF295UVV.bed.gz  |  cut -f1-5 >
H3K36me3_H1.ENCFF295UVV.hg19.bed

zcat ENCFF327EZJ.bed.gz  |  cut -f1-5 >
H3K36me3_H1.ENCFF327EZJ.hg19.bed
```

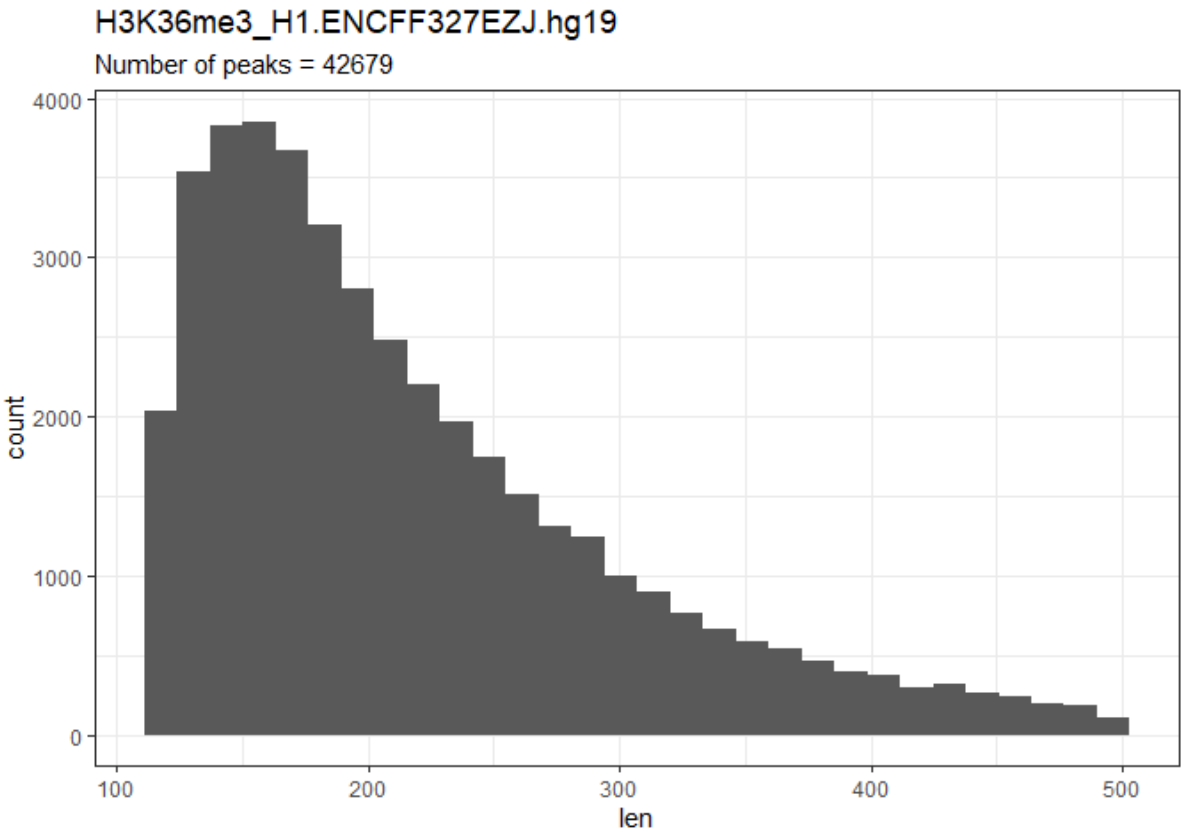
Обратим внимание, что файлы уже в нужной версии генома, конвертация не требуется.

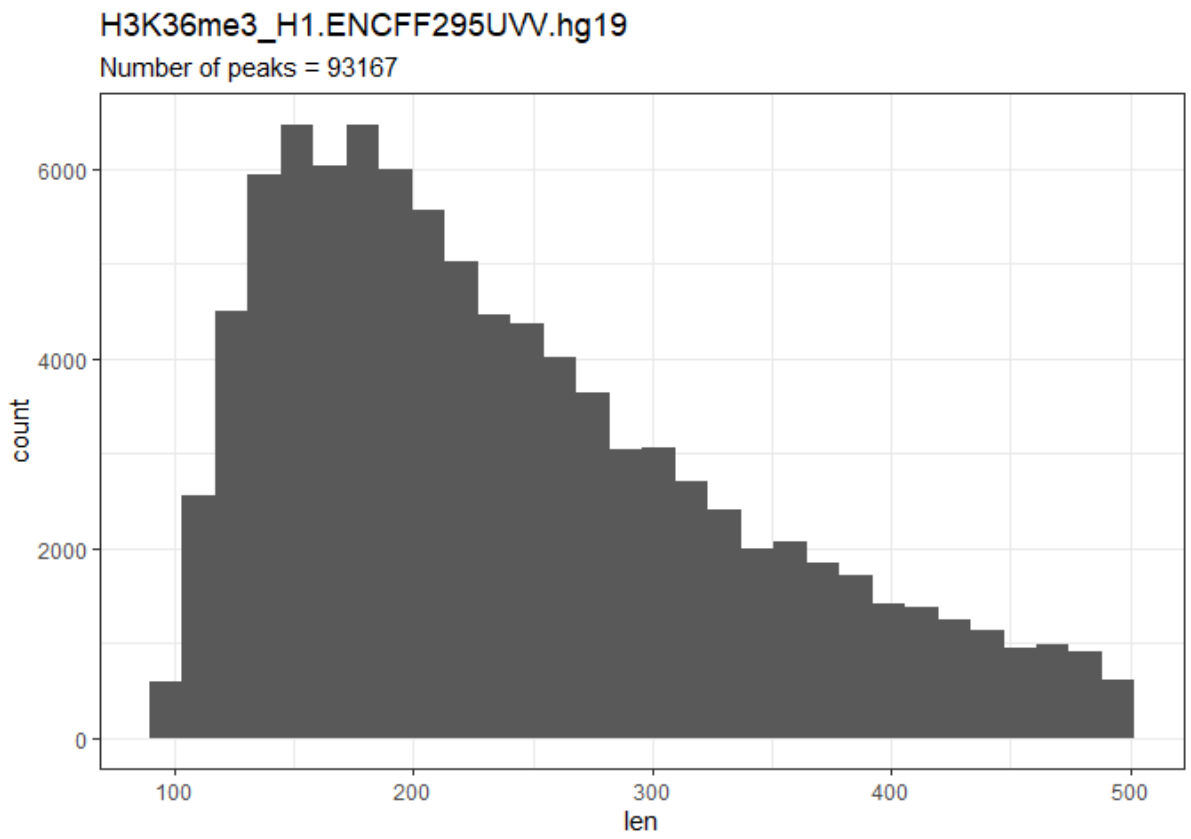
2. Строим гистограммы с помощью len_hist.R





3. (На глаз) было решено отфильтровать пики более 500 единиц длиной.

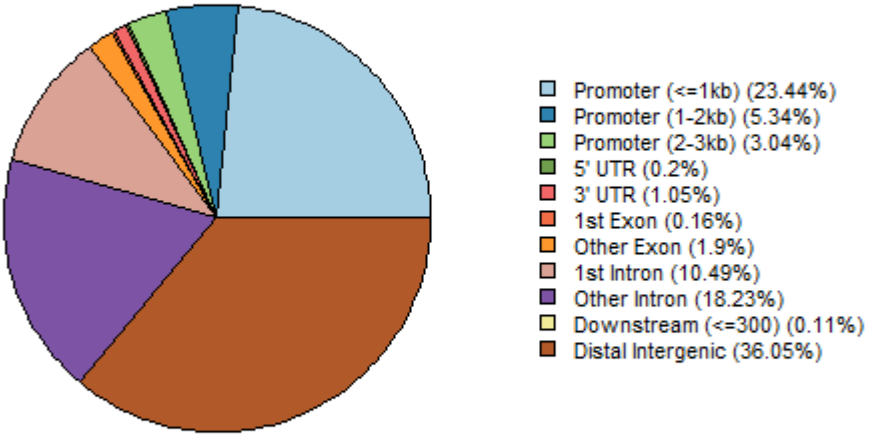


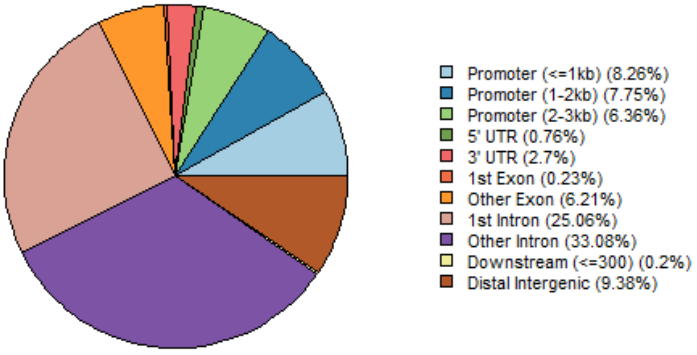


4.

Итого сократили число пиков: с 44451 до 42679, и с 103479 до 93167.

5.
Смотрим расположение пиков относительно аннотированных генов. В обоих случаях существенная часть располагается вблизи экзонов (расстояние в 1-2 килобаз).

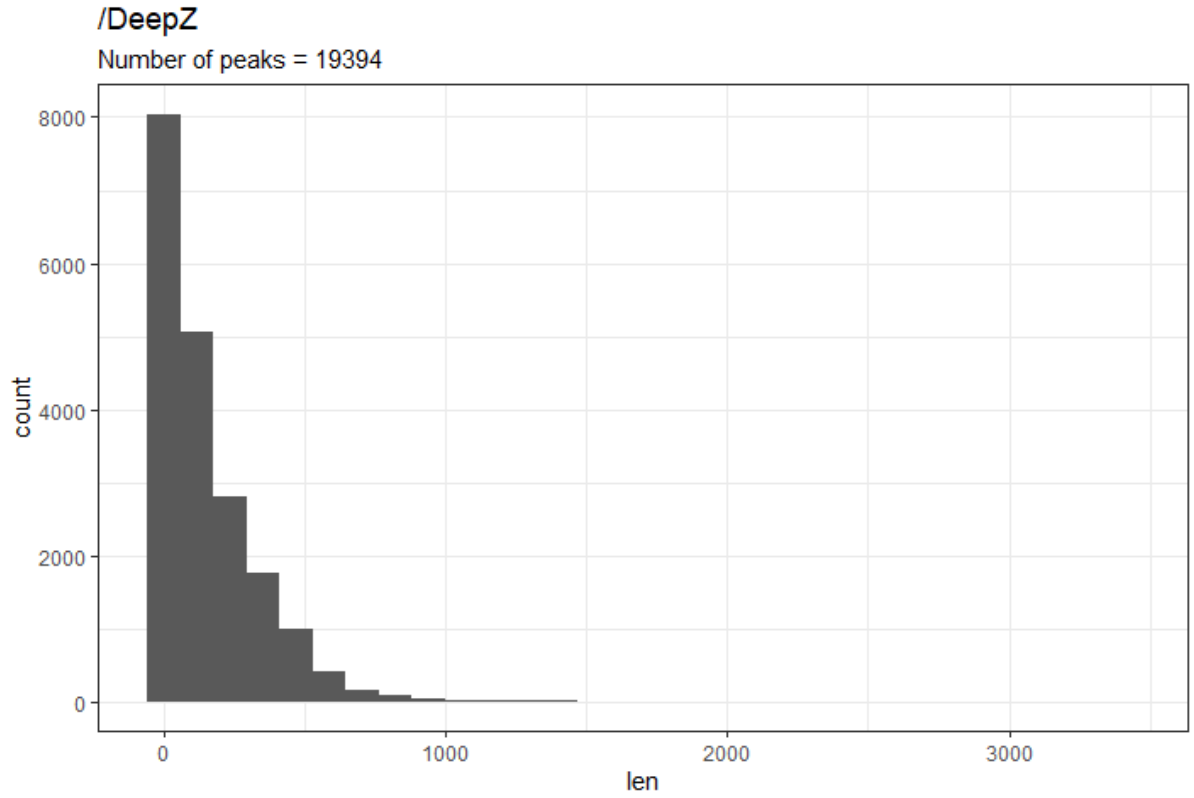




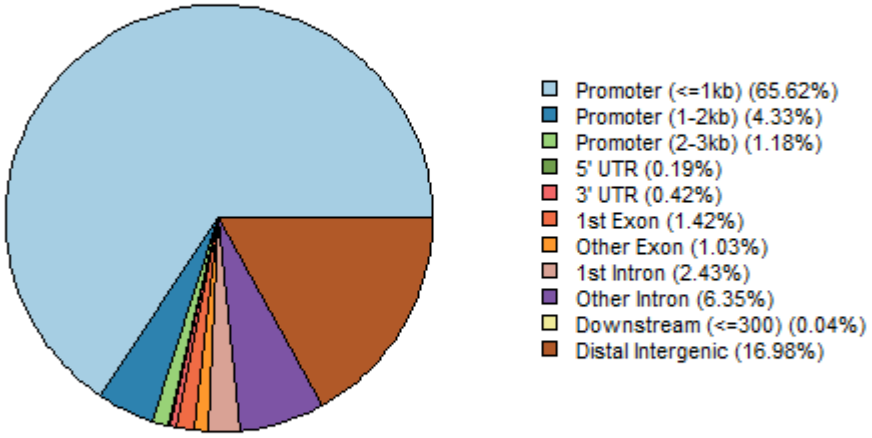
6. Объединяем наши bed-файлы командой merge:

```
cat  *.filtered.bed  |  sort -k1,1 -k2,2n  |  bedtools merge  >
H3K36me3_H1.merge.hg19.bed
```

7. Число пиков нашей вторичной структуры (Z-ДНК) = 19394. Поскольку дальше мы с ней будем пересекать пики из экспериментов, то фильтровать по длине особого смысла нет.



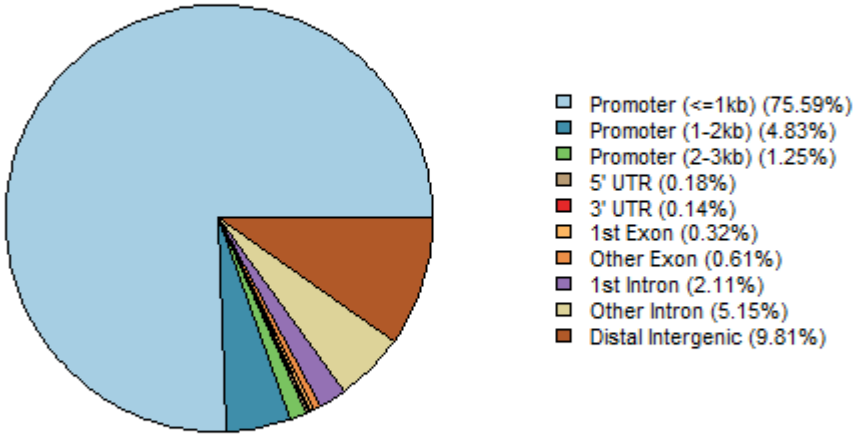
Расположение в почти трех четвертях случаев - очень близко к гену:



8. Пересекаем наш merged файл со вторичной структурой командой:

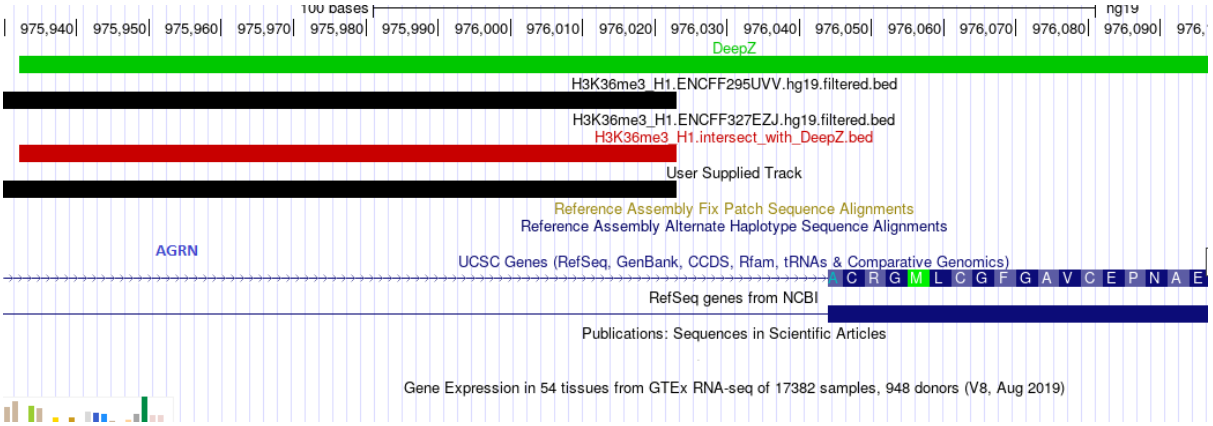
```
bedtools intersect  -a DeepZ.bed   -b  *.merge.hg19.bed  >
H3K36me3_H1.intersect_w_DeepZ.bed
```

9. Строим пайчарт пересечения. Теперь подавляющее большинство пиков находятся очень близко к гену, больше 75%. Получается, что очень большая часть пиков вторичной структуры так или иначе приходится на участки днк, подверженные метилированию (?)



10.
Визуализация треков в геномном браузере:
http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr1%3A975894-976169&hgsid=1124209057_VHRSFA8ARQQ78ACNFMuy0kpaucap

Наблюдаем пересечение пиков гист.меток с Z-ДНК вблизи гена AGRN (chr1:975932-976023)



11. Полный список ассоциаций пиков с генами получаем с помощью скрипта anno.r.

В списке уникальных генов - 1482 строки (genes_uniq.txt можно проверить редактором, можно wc -l)
В списке пиков, ассоциированных с аннотацией - 1805 строк (и, соответственно, пиков)

12. Статистический анализ по списку уникальных генов показывает, что больше всего затронутые гены ассоциируются с процессами метаболизма клетки (и метаболизма в целом).

GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P-value	FDR
cellular metabolic process	7542	766	486.32	1.58	+	3.32E-51	5.25E-47
cellular nitrogen compound metabolic process	3395	441	218.92	2.01	+	1.41E-46	1.12E-42
primary metabolic process	7352	734	474.07	1.55	+	8.93E-45	4.71E-41
nitrogen compound metabolic process	6886	698	444.02	1.57	+	1.85E-43	7.33E-40
metabolic process	8313	794	536.04	1.48	+	2.82E-43	8.92E-40
nucleobase-containing compound metabolic process	2705	365	174.42	2.09	+	2.68E-40	7.07E-37
organic substance metabolic process	7840	752	505.54	1.49	+	7.37E-40	1.66E-36
heterocycle metabolic process	2891	377	186.42	2.02	+	6.94E-39	1.37E-35