

Bioinformatics HW1

Ershov Ivan

October 2021

Задание 1. Геном Бактерии *Micrococcus luteus*

1. Какова длина генома (файл .fna)

Для начала посчитаем количество строк в файле и вычтем из этого числа 1, то есть не будем учитывать первую строку.

```
(base) ershoff@MacBook-Pro-Ivan bioinfo % wc -l genomic.fna  
32370 genomic.fna
```

Получаем $32370 - 1 = 32369$ строк.

Теперь посчитаем количество вхождений букв A, T, G, C и N.

```
(base) ershoff@MacBook-Pro-Ivan bioinfo % tail -32369 genomic.fna | grep -o "A" | wc -l  
353304  
(base) ershoff@MacBook-Pro-Ivan bioinfo % tail -32369 genomic.fna | grep -o "T" | wc -l  
356478  
(base) ershoff@MacBook-Pro-Ivan bioinfo % tail -32369 genomic.fna | grep -o "G" | wc -l  
930048  
(base) ershoff@MacBook-Pro-Ivan bioinfo % tail -32369 genomic.fna | grep -o "C" | wc -l  
925346
```

Полученные числа сложим и получим ответ:

$$353304 + 356478 + 930048 + 925346 = 2'565'176$$

Ответ: 2'565'176

2. Сколько генов, кодирующих белки?

Просто посчитаем в терминале:

```
(base) ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "protein_coding" | wc -l  
2331
```

Ответ: 2331

3. Сколько рнк-генов?

Просто посчитаем в терминале:

```
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "mRNA" | wc -l
      2
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "tRNA" | wc -l
    148
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "rRNA" | wc -l
     38
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "ncRNA" | wc -l
      2
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "tmRNA" | wc -l
      2
```

И просуммируем:

$$2 + 148 + 38 + 2 + 2 = 192$$

Ответ: 192

4. Сколько транскрипционных факторов?

Просто посчитаем в терминале:

```
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "transcriptional regulator" | wc -l
      74
(base)ershoff@MacBook-Pro-Ivan bioinfo %
```

Ответ: 74

5. Сколько транспортных белков (ABC transporters)?

Просто посчитаем в терминале:

```
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "ABC transporter" | wc -l
      98
(base)ershoff@MacBook-Pro-Ivan bioinfo %
```

Ответ: 98

6. Сколько tRNA?

Просто посчитаем в терминале:

```
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "tRNA" | wc -l
    148
(base)ershoff@MacBook-Pro-Ivan bioinfo %
```

Ответ: 148

7. Сколько закодировано субъединиц ATP-synthase?

Просто посчитаем в терминале:

```
(base)ershoff@MacBook-Pro-Ivan bioinfo % 7. Сколько закодировано субъединиц ATP-synthase  
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "ATP synthase" | wc -l  
9  
(base)ershoff@MacBook-Pro-Ivan bioinfo %
```

Ответ: 9

8. Сколько генов закодировано на положительном, а сколько на отрицательном стренде?

Посчитаем количество строк, начинающихся на "gene" и заканчивающихся на '+' и '-' и вычтем из полученных чисел количество псевдогенов для '+' и '-' генов соответственно:

Начнем с положительных стрендов:

```
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "gene.*+" | wc -l  
1188  
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "gene.*pseudogen.*+" | wc -l  
38  
(base)ershoff@MacBook-Pro-Ivan bioinfo %
```

В итоге получаем $1188 - 38 = 1150$ генов на положительном стренде.

Теперь для отрицательных стрендов:

```
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "gene.*-" | wc -l  
1300  
(base)ershoff@MacBook-Pro-Ivan bioinfo % cat feature_table.txt | grep "gene.*pseudogen.*-" | wc -l  
52  
(base)ershoff@MacBook-Pro-Ivan bioinfo %
```

В итоге получаем $1300 - 52 = 1248$ генов на отрицательном стренде.

Ответ: 1150 и 1248 соответственно

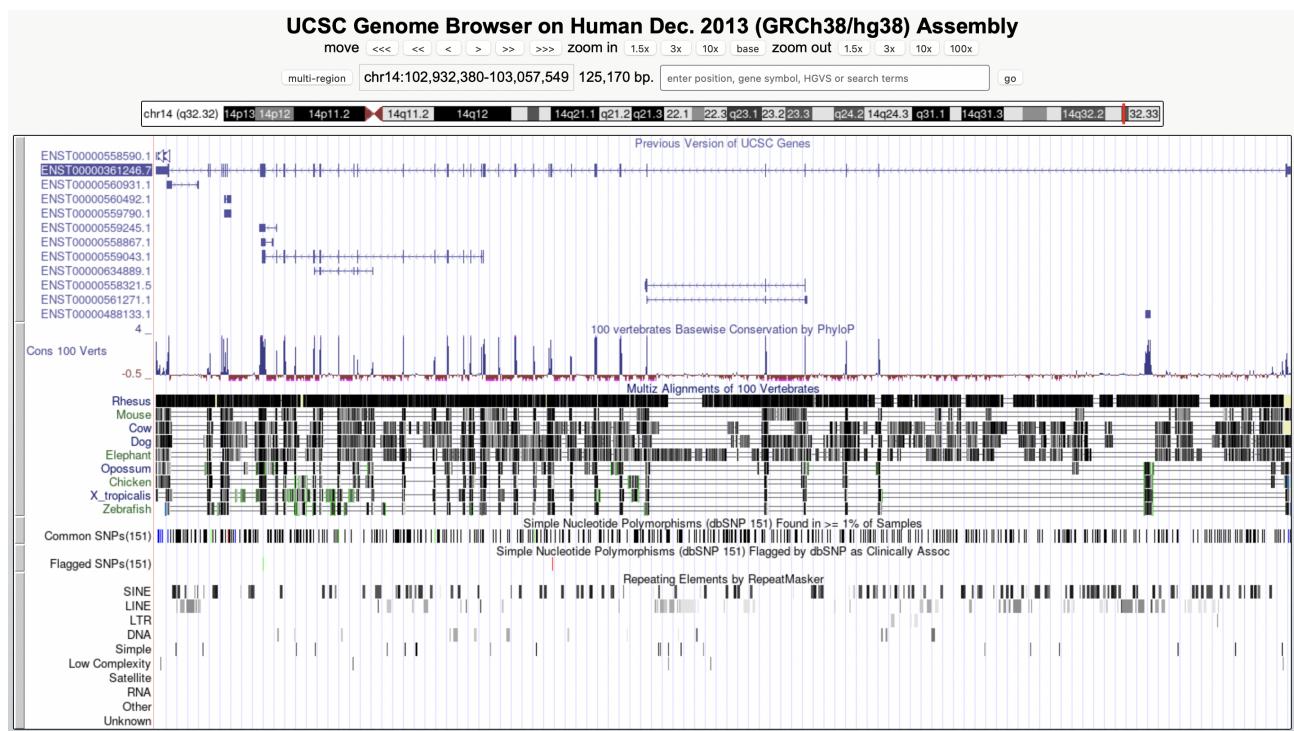
Задание 2. Ген Человека CDC42BPB

1. В геномном браузере UCSC отобразить все изоформы гена, а также SNPs (common and clinically relevant), участки консервативности среди позвоночных и транспозоны. Сохранить скриншот в виде графического файла.

В UCSC установим следующие расширения:

- Gencode v36
- Old UCSC Genes(full) (для изоформ)
- Common SNPs(151) dense (для common SNP)
- Flagged SNPs(151) dense (для clinically relevant SNP)
- Conservation full (для участков консервативности среди позвоночных)
- Repeatmasker full (для транспозонов)
- остальные расширения уберем (hide)

В итоге получаем:



2. Отобрать в табличном виде и сохранить в текстовый файл все изоформы генов, попавших в заданный участок

Чтобы получить таблицу с изоформами, в Genome Browser выберем только Gencode v36 и Old UCSC Genes(full), а все остальное hide. Теперь зайдем во вкладку Tools → Table Browser и выберем следующее:

The screenshot shows the 'Table Browser' interface. At the top, there's a brief description: 'Use this tool to retrieve and export data from the Genome Browser annotation track database. You can retrieve DNA sequence covered by a track. More...'. Below this is a 'Select dataset' section with dropdown menus for 'clade' (Mammal), 'genome' (Human), and 'assembly' (Dec. 2013 (GRCh38/hg38)). There are also dropdowns for 'group' (Custom Tracks) and 'track' (tb_knownGene). A 'table' dropdown shows 'ct_tbknownGene_5624' with a 'describe table schema' button. The next section is 'Define region of interest', containing a 'region' dropdown set to 'genome' with a value of 'chr14:102,932,380-103,057,549', and buttons for 'lookup' and 'define regions'. Below it is a 'identifiers (names/accessions)' field with 'paste list' and 'upload list' buttons. The third section is 'Optional: Subset, combine, compare with another track', with 'filter', 'intersection', and 'correlation' buttons. The final section is 'Retrieve and display data', with 'output format' set to 'all fields from selected table', 'Send output to' checkboxes for 'Galaxy' and 'GREAT', 'output filename' set to 'output' (with a note '(leave blank to keep output in browser)'), 'file type returned' set to 'plain text', and buttons for 'get output' and 'summary/statistics'.

Чтобы сохранить таблицу в виде текстового файла, укажем output file-name и plain text

В итоге получаем таблицу:

#chrom	chromStart	chromEnd	name	score	strand	thickStart	thickEnd	itemRgb	blockCount	blockSizes
chromStarts										
chr14	102928644	102933596	ENST00000558590.1	0	+	102928644	102928644	254,0,0	9	
614,109,83,163,163,88,827,141,584,		0,783,1010,1279,1520,1761,2943,3959,4368,								
chr14	102932379	103057549	ENST00000361246.7	0	-	102933711	103057173	12,12,120	37	
1464,71,106,118,118,85,98,597,63,217,82,140,137,106,78,87,80,95,149,106,125,106,245,111,243,134,120,167,80,249,201,94,149,96,84,92,551,		0,5724,5926,7230,7450,7666,7847,11511,13282,14088,15341,17385,18086,20118,21818,22222,27251,30681,32122,33902,34666,35873,36092,37771,3								
9539,41636,43304,43503,45746,48393,51176,54107,67185,71548,76092,79717,124619,										
chr14	102933573	102937177	ENST00000560931.1	0	+	102933573	102933573	0,100,0	2	
592,236,	0,3368,									
chr14	102939907	102940747	ENST00000560492.1	0	-	102939907	102939907	254,0,0	3	
40,85,521,	0,138,319,									
chr14	102939915	102940726	ENST00000559790.1	0	-	102939915	102939915	254,0,0	2	32,681,
0,130,										
chr14	102943812	102945724	ENST00000559245.1	0	-	102943812	102943812	254,0,0	2	675,63,
0,1849,										
chr14	102944010	102945331	ENST00000558867.1	0	-	102944010	102944010	0,100,0	2	
477,158,	0,1163,									
chr14	102944042	102968532	ENST00000559043.1	0	-	102944042	102968532	12,12,120	15	
445,63,217,82,140,137,106,87,80,95,149,106,125,106,61,										
0,1619,2425,3678,5722,6423,8455,10559,15588,19018,20459,22239,23003,24210,24429,										
chr14	102949856	102956290	ENST00000634889.1	0	-	102949856	102949856	0,100,0	6	
48,137,106,78,87,51,	0,609,2641,4341,4745,6383,									
chr14	102986297	103004023	ENST00000558321.5	0	-	102986435	103004023	12,12,120	3	
283,149,96,	0,13267,17630,									
chr14	102986509	103004279	ENST00000561271.1	0	-	102986509	102986509	254,0,0	3	
71,149,352,	0,13055,17418,									
chr14	103041468	103042094	ENST00000488133.1	0	+	103041468	103041468	255,51,255	1	
626,	0,									

3. Отобрать координаты только экзонов и только инtronов для заданного участка.

Tools → Table Browser → выберем BED в output format → get output → Exon plus / Intron plus → get BED

Получаем таблицы, где во втором столбце координата начала, а в третьем столбце координата конца экзона/интрана.

(Скриншоты полностью не поместились)

ЭКЗОНЫ:

track	name="tb_knownGene"	description="table browser query on knownGene"	visibility=2	url=		
chr14	102928644	102929258	ENST00000558590.1_exon_0_0_chr14_102928645_f	0	+	
chr14	102929427	102929536	ENST00000558590.1_exon_1_0_chr14_102929428_f	0	+	
chr14	102929654	102929737	ENST00000558590.1_exon_2_0_chr14_102929655_f	0	+	
chr14	102929923	102930086	ENST00000558590.1_exon_3_0_chr14_102929924_f	0	+	
chr14	102930164	102930327	ENST00000558590.1_exon_4_0_chr14_102930165_f	0	+	
chr14	102930405	102930493	ENST00000558590.1_exon_5_0_chr14_102930406_f	0	+	
chr14	102931587	102932414	ENST00000558590.1_exon_6_0_chr14_102931588_f	0	+	
chr14	102932603	102932744	ENST00000558590.1_exon_7_0_chr14_102932604_f	0	+	
chr14	102933012	102933596	ENST00000558590.1_exon_8_0_chr14_102933013_f	0	+	
chr14	1029332379	102933843	ENST00000361246.7_exon_0_0_chr14_102932380_r	0	-	
chr14	102938103	102938174	ENST00000361246.7_exon_1_0_chr14_102938104_r	0	-	
chr14	102938305	102938411	ENST00000361246.7_exon_2_0_chr14_102938306_r	0	-	
chr14	102939609	102939727	ENST00000361246.7_exon_3_0_chr14_102939610_r	0	-	
chr14	102939829	102939947	ENST00000361246.7_exon_4_0_chr14_102939830_r	0	-	
chr14	102940045	102940130	ENST00000361246.7_exon_5_0_chr14_102940046_r	0	-	
chr14	102940226	102940324	ENST00000361246.7_exon_6_0_chr14_102940227_r	0	-	
chr14	102943890	102944487	ENST00000361246.7_exon_7_0_chr14_102943891_r	0	-	
chr14	102945661	102945724	ENST00000361246.7_exon_8_0_chr14_102945662_r	0	-	
chr14	102946467	102946684	ENST00000361246.7_exon_9_0_chr14_102946468_r	0	-	
chr14	102947720	102947802	ENST00000361246.7_exon_10_0_chr14_102947721_r	0	-	
chr14	102949764	102949904	ENST00000361246.7_exon_11_0_chr14_102949765_r	0	-	
chr14	102950465	102950602	ENST00000361246.7_exon_12_0_chr14_102950466_r	0	-	
chr14	102952497	102952603	ENST00000361246.7_exon_13_0_chr14_102952498_r	0	-	
chr14	102954197	102954275	ENST00000361246.7_exon_14_0_chr14_102954198_r	0	-	
chr14	102954601	102954688	ENST00000361246.7_exon_15_0_chr14_102954602_r	0	-	
chr14	102959630	102959710	ENST00000361246.7_exon_16_0_chr14_102959631_r	0	-	
chr14	102963060	102963155	ENST00000361246.7_exon_17_0_chr14_102963061_r	0	-	
chr14	102964501	102964650	ENST00000361246.7_exon_18_0_chr14_102964502_r	0	-	
chr14	102966281	102966387	ENST00000361246.7_exon_19_0_chr14_102966282_r	0	-	
chr14	102967045	102967170	ENST00000361246.7_exon_20_0_chr14_102967046_r	0	-	
chr14	102968252	102968358	ENST00000361246.7_exon_21_0_chr14_102968253_r	0	-	
chr14	102968471	102968716	ENST00000361246.7_exon_22_0_chr14_102968472_r	0	-	
chr14	102970150	102970261	ENST00000361246.7_exon_23_0_chr14_102970151_r	0	-	
chr14	102971918	102972161	ENST00000361246.7_exon_24_0_chr14_102971919_r	0	-	
chr14	102974015	102974149	ENST00000361246.7_exon_25_0_chr14_102974016_r	0	-	
chr14	102975683	102975803	ENST00000361246.7_exon_26_0_chr14_102975684_r	0	-	
chr14	102975882	102976049	ENST00000361246.7_exon_27_0_chr14_102975883_r	0	-	
chr14	102978125	102978205	ENST00000361246.7_exon_28_0_chr14_102978126_r	0	-	
chr14	102980772	102981021	ENST00000361246.7_exon_29_0_chr14_102980773_r	0	-	
chr14	102983555	102983756	ENST00000361246.7_exon_30_0_chr14_102983556_r	0	-	
chr14	102986486	102986580	ENST00000361246.7_exon_31_0_chr14_102986487_r	0	-	
chr14	102999564	102999713	ENST00000361246.7_exon_32_0_chr14_102999565_r	0	-	
chr14	103003927	103004023	ENST00000361246.7_exon_33_0_chr14_103003928_r	0	-	
chr14	103008471	103008555	ENST00000361246.7_exon_34_0_chr14_103008472_r	0	-	
chr14	103012096	103012188	ENST00000361246.7_exon_35_0_chr14_103012097_r	0	-	
chr14	103056998	103057549	ENST00000361246.7_exon_36_0_chr14_103056999_r	0	-	
chr14	102933573	102934165	ENST00000560931.1_exon_0_0_chr14_102933574_f	0	+	
chr14	102936941	102937177	ENST00000560931.1_exon_1_0_chr14_102936942_f	0	+	
chr14	102939907	102939947	ENST00000560492.1_exon_0_0_chr14_102939908_r	0	-	
chr14	102940045	102940130	ENST00000560492.1_exon_1_0_chr14_102940046_r	0	-	
chr14	102940226	102940747	ENST00000560492.1_exon_2_0_chr14_102940227_r	0	-	
chr14	102939915	102939947	ENST00000559790.1_exon_0_0_chr14_102939916_r	0	-	

Интроны:

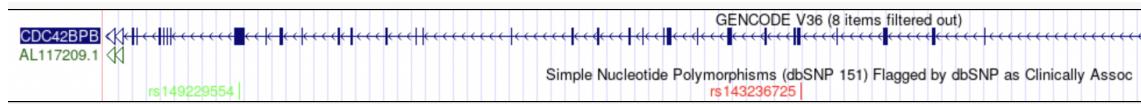
```

track name="tb_knownGene" description="table browser query on knownGene" visibility=2 url=
chr14 102929258 102929427 ENST00000558590.1_intron_0_0_chr14_102929259_f 0 +
chr14 102929536 102929654 ENST00000558590.1_intron_1_0_chr14_102929537_f 0 +
chr14 102929737 102929923 ENST00000558590.1_intron_2_0_chr14_102929738_f 0 +
chr14 102930086 102930164 ENST00000558590.1_intron_3_0_chr14_102930087_f 0 +
chr14 102930327 102930405 ENST00000558590.1_intron_4_0_chr14_102930328_f 0 +
chr14 102930493 102931587 ENST00000558590.1_intron_5_0_chr14_102930494_f 0 +
chr14 102932414 102932603 ENST00000558590.1_intron_6_0_chr14_102932415_f 0 +
chr14 102932744 102933012 ENST00000558590.1_intron_7_0_chr14_102932745_f 0 +
chr14 102933843 102938103 ENST00000361246.7_intron_0_0_chr14_102933844_r 0 -
chr14 102938174 102938305 ENST00000361246.7_intron_1_0_chr14_102938175_r 0 -
chr14 102938411 102939609 ENST00000361246.7_intron_2_0_chr14_102938412_r 0 -
chr14 102939727 102939829 ENST00000361246.7_intron_3_0_chr14_102939728_r 0 -
chr14 102939947 102940045 ENST00000361246.7_intron_4_0_chr14_102939948_r 0 -
chr14 102940130 102940226 ENST00000361246.7_intron_5_0_chr14_102940131_r 0 -
chr14 102940324 102943890 ENST00000361246.7_intron_6_0_chr14_102940325_r 0 -
chr14 102944487 102945661 ENST00000361246.7_intron_7_0_chr14_102944488_r 0 -
chr14 102945724 102946467 ENST00000361246.7_intron_8_0_chr14_102945725_r 0 -
chr14 102946684 102947720 ENST00000361246.7_intron_9_0_chr14_102946685_r 0 -
chr14 102947802 102949764 ENST00000361246.7_intron_10_0_chr14_102947803_r 0 -
chr14 102949904 102950465 ENST00000361246.7_intron_11_0_chr14_102949905_r 0 -
chr14 102950602 102952497 ENST00000361246.7_intron_12_0_chr14_102950603_r 0 -
chr14 102952603 102954197 ENST00000361246.7_intron_13_0_chr14_102952604_r 0 -
chr14 102954275 102954601 ENST00000361246.7_intron_14_0_chr14_102954276_r 0 -
chr14 102954688 102959630 ENST00000361246.7_intron_15_0_chr14_102954689_r 0 -
chr14 102959710 102963060 ENST00000361246.7_intron_16_0_chr14_102959711_r 0 -
chr14 102963155 102964501 ENST00000361246.7_intron_17_0_chr14_102963156_r 0 -
chr14 102964650 102966281 ENST00000361246.7_intron_18_0_chr14_102964651_r 0 -
chr14 102966387 102967045 ENST00000361246.7_intron_19_0_chr14_102966388_r 0 -
chr14 102967170 102968252 ENST00000361246.7_intron_20_0_chr14_102967171_r 0 -
chr14 102968358 102968471 ENST00000361246.7_intron_21_0_chr14_102968359_r 0 -
chr14 102968716 102970150 ENST00000361246.7_intron_22_0_chr14_102968717_r 0 -
chr14 102970261 102971918 ENST00000361246.7_intron_23_0_chr14_102970262_r 0 -
chr14 102972161 102974015 ENST00000361246.7_intron_24_0_chr14_102972162_r 0 -
chr14 102974149 102975683 ENST00000361246.7_intron_25_0_chr14_102974150_r 0 -
chr14 102975803 102975882 ENST00000361246.7_intron_26_0_chr14_102975804_r 0 -
chr14 102976049 102978125 ENST00000361246.7_intron_27_0_chr14_102976050_r 0 -
chr14 102978205 102980772 ENST00000361246.7_intron_28_0_chr14_102978206_r 0 -
chr14 102981021 102983555 ENST00000361246.7_intron_29_0_chr14_102981022_r 0 -
chr14 102983756 102986486 ENST00000361246.7_intron_30_0_chr14_102983757_r 0 -
chr14 102986580 102999564 ENST00000361246.7_intron_31_0_chr14_102986581_r 0 -
chr14 102999713 103003927 ENST00000361246.7_intron_32_0_chr14_102999714_r 0 -
chr14 103004023 103008471 ENST00000361246.7_intron_33_0_chr14_103004024_r 0 -
chr14 103008555 103012096 ENST00000361246.7_intron_34_0_chr14_103008556_r 0 -
chr14 103012188 103056998 ENST00000361246.7_intron_35_0_chr14_103012189_r 0 -
chr14 102934165 102936941 ENST00000560931.1_intron_0_0_chr14_102934166_f 0 +
chr14 102939947 102940045 ENST00000560492.1_intron_0_0_chr14_102939948_r 0 -
chr14 102940130 102940226 ENST00000560492.1_intron_1_0_chr14_102940131_r 0 -
chr14 102939947 102940045 ENST00000559790.1_intron_0_0_chr14_102939948_r 0 -
chr14 102944487 102945661 ENST00000559245.1_intron_0_0_chr14_102944488_r 0 -
chr14 102944487 102945173 ENST00000558867.1_intron_0_0_chr14_102944488_r 0 -
chr14 102944487 102945661 ENST00000559043.1_intron_0_0_chr14_102944488_r 0 -
chr14 102945724 102946467 ENST00000559043.1_intron_1_0_chr14_102945725_r 0 -

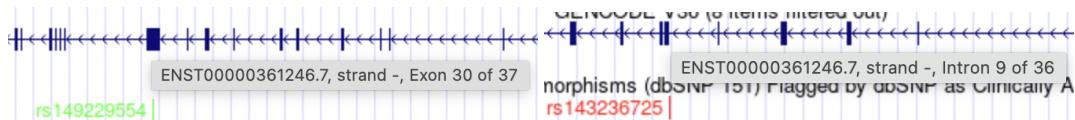
```

4. Ответить письменно, в какие участки гена попадают clinically relevant SNPs.

Для удобства выведем только clinically relevant SNPs.



(зеленым и красным отмечены clinically relevant SNPs)

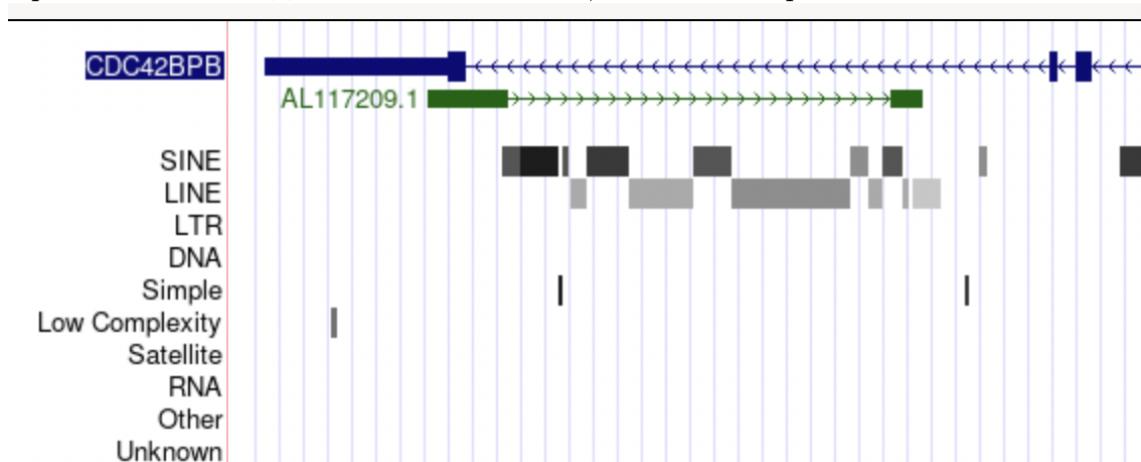


Мы видим, что всего 2 clinically relevant SNPs попадают на ген: один попадает на инtron, другой на экзон.

5. Ответить письменно, в какие участки гена попадают транспозоны.

Чтобы отобразить транспозоны, выберем только Gencode v36 pack и RepeatMasker full.

Транспозоны попадают как и на экзоны, так и на интроны:



(Из скриншота это явно видно)