1 Форматы файлов

1.1 Task 1

samtools view -S -b /srv/common/midterm/task_adh1b.sam > ~/test1/task1.bam
samtools sort task1.bam -o sorted_task1.bam
samtools index sorted_task1.bam

1.2 Task 2

samtools view -H sorted_task1.bam | grep 'LN:249250621

Результат: hg19

1.3 Task 3

Mutation	REF Sequence	Sample Variants
rs4988235	G/G	G/G
rs41380347	A/A	A/A
rs145946881	C/C	C/C
rs41525747	G/G	G/G
rs121908937	C/C	C/C
rs121908936	-	-

Непереносимости нет

1.4 Task 4

samtools view /srv/common/midterm/task_adh1b.bam chr4 | wc -l

Результат: 78105

1.5 Task 5

samtools view /srv/common/midterm/task_gender.bam chrX | wc -l

Результат: 592941

samtools view /srv/common/midterm/task_gender.bam chrY | wc -l

Результат: 380490

1.6 Task 7

view -H /srv/common/midterm/chip.vcf.gz | wc -l

Результат: 1001385

1.7 Task 8

```
bcftools index chip.vcf.gz
bcftools view -r chr21:5215000-5233000 chip.vcf.gz | grep -v "^#" | wc -l
```

Результат: 97

1.8 Task 9

```
bcftools view -r chr21:5215000-5233000 chip.vcf.gz | bcftools query -s NA21135 -f '%CHROM\t%POS\t
bcftools view -r chr21:5215000-5233000 chip.vcf.gz | bcftools query -s NA21135 -f '%CHROM\t%POS\t
```

Chromosome	Position	Reference Allele	Alternate Allele	Genotype
chr21	5219624	C	A	0|1
chr21	5231730	C	G	0|1
chr21	5225197	G	T	1|1

1.9 Task 10

```
bcftools view -r chr21:5215000-5233000 -i 'INFO/AF>=0.05' chip.vcf.gz | grep -v "^#" | wc -l
```

Результат: 5

2 Глобальные и локальные выравнивания

2.1 Task 1

Выровняйте следующие последовательности с помощью алгоритма Нидлмана-Вунша:
ATGCCCGA
GTCACCC
Используйте следующие параметры для выравнивания: награда за совпадение: +1, штраф за замену: -1, штраф за вставку или удаление: -2.

Формула для заполнения:

$$\text{Score}(i,j) = \max \begin{cases} \text{Score}(i-1,j-1) + \text{match/mismatch}, \\ \text{Score}(i-1,j) + \text{gap penalty}, \\ \text{Score}(i,j-1) + \text{gap penalty} \end{cases}$$

Получим таблицу:

		G	T	C	A	C	C	C
	A	-1	-3	-5	-7	-9	-11	-13
	T	-3	-1	-2	-4	-6	-8	-10
	G	-5	-3	-2	-4	-6	-8	-10
	C	-7	-5	-3	-2	-4	-6	-8
	C	-9	-7	-5	-4	-2	-4	-6
	C	-11	-9	-7	-6	-4	-2	-4
	G	-13	-11	-9	-8	-6	-4	-2
	A	-15	-13	-11	-7	-8	-6	-4

то есть выравнивание выглядит так:
ATGCCCGA
—TC—ACCC

2.2 Task 2

нуклеотидная последовательность белка эндонуклеазы III (Nth) из бактерии Escherichia coli, штамм K-12, субштамм MG1655, бластнув получим Shigella flexneri strain STIN_92 chromosome

Download GenBank Graphics

Shigella flexneri strain STIN_92 chromosome, complete genome

Sequence ID: CP054977.1 Length: 4813336 Number of Matches: 1

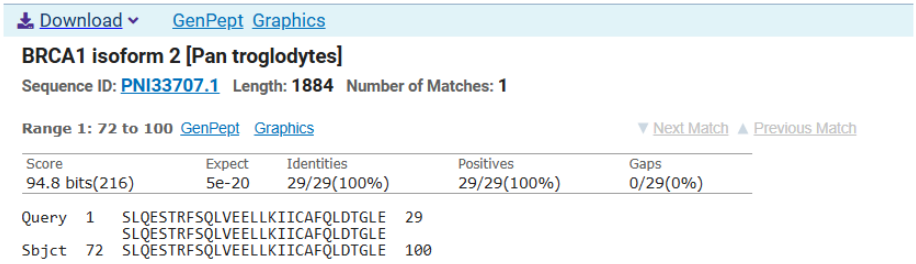
Range 1: 2260016 to 2260651 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1175 bits(636)	0.0	636/636(100%)	0/636(0%)	Plus/Minus

```
Query  1        ATGAATAAAGCAAAACGCCTGGAGATCCTCACTCGCCTGCGTGAGAACAATCCTCATCCC  60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260651  ATGAATAAAGCAAAACGCCTGGAGATCCTCACTCGCCTGCGTGAGAACAATCCTCATCCC  2260592

Query  61       ACCACCGAGCTTAATTTCAGTTCGCCTTTTGAATTGCTGATTGCCGTACTGCTTTCCGCT  120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260591  ACCACCGAGCTTAATTTCAGTTCGCCTTTTGAATTGCTGATTGCCGTACTGCTTTCCGCT  2260532

Query  121      CAGGCGACCGATGTCAGTGTTAATAAGGCGACGGCGAAACTCTACCCGGTGGCGAATACG  180
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260531  CAGGCGACCGATGTCAGTGTTAATAAGGCGACGGCGAAACTCTACCCGGTGGCGAATACG  2260472

Query  181      CCTGCAGCGATGCTTGAACTGGGCGTTGAAGGGGTGAAAACCTATATCAAAACGATTGGG  240
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260471  CCTGCAGCGATGCTTGAACTGGGCGTTGAAGGGGTGAAAACCTATATCAAAACGATTGGG  2260412

Query  241      CTTTATAACAGCAAAGCAGAAAATATCATCAAAACCTGCCGTATCTTGCTGGAGCAGCAT  300
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260411  CTTTATAACAGCAAAGCAGAAAATATCATCAAAACCTGCCGTATCTTGCTGGAGCAGCAT  2260352

Query  301      AATGGCGAGGTTCCGGAAGATCGTGCTGCGCTTGAAGCCCTGCCCGGCGTAGGTCGTAAA  360
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260351  AATGGCGAGGTTCCGGAAGATCGTGCTGCGCTTGAAGCCCTGCCCGGCGTAGGTCGTAAA  2260292

Query  361      ACAGCCAACGTCGTATTAAACACTGCATTCGGCTGGCCGACTATTGCTGTCGACACGCAC  420
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260291  ACAGCCAACGTCGTATTAAACACTGCATTCGGCTGGCCGACTATTGCTGTCGACACGCAC  2260232

Query  421      ATTTTCCGCGTTTGTAATCGTACTCAATTTGCGCCGGGGAAAAACGTCGAACAGGTAGAA  480
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260231  ATTTTCCGCGTTTGTAATCGTACTCAATTTGCGCCGGGGAAAAACGTCGAACAGGTAGAA  2260172

Query  481      GAAAAGCTACTGAAAGTGGTTCCAGCAGAGTTTAAAGTCGACTGCCACCATTGGTTGATC  540
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260171  GAAAAGCTACTGAAAGTGGTTCCAGCAGAGTTTAAAGTCGACTGCCACCATTGGTTGATC  2260112

Query  541      CTGCACGGGCGTTATACCTGCATTGCCCGCAAGCCCCGCTGTGGCTCTTGTATTATTGAA  600
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260111  CTGCACGGGCGTTATACCTGCATTGCCCGCAAGCCCCGCTGTGGCTCTTGTATTATTGAA  2260052

Query  601      GATCTTTGTGAATACAAAGAGAAAGTTGACATCTGA  636
                ||||||||||||||||||||||||||||||||||||
Sbjct  2260051  GATCTTTGTGAATACAAAGAGAAAGTTGACATCTGA  2260016
```

2.3 Task 3

BRCA1



3 Множественные выравнивания

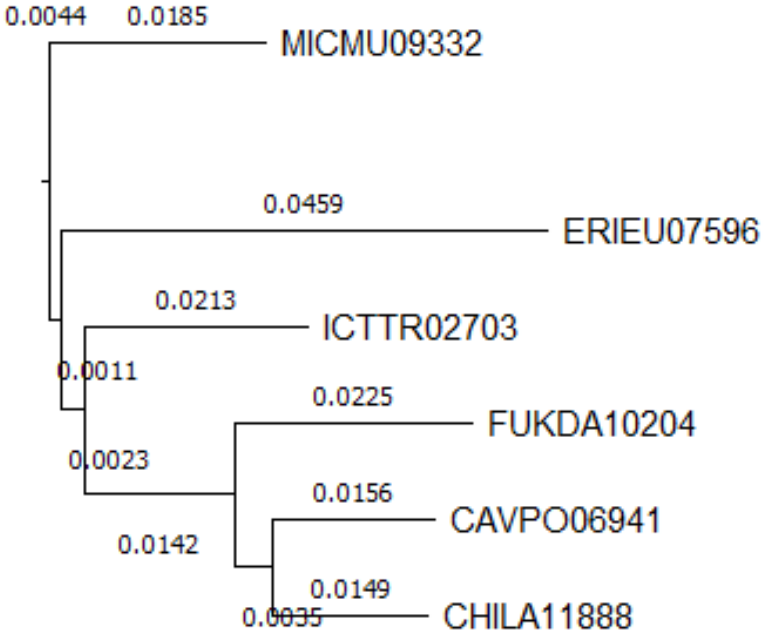
3.1 Task 1

Возьмем:
BRCA1 isoform 2 [Pan troglodytes] Sequence ID: PNI33707.1,
breast cancer type 1 susceptibility protein homolog isoform X1 [Gorilla gorilla gorilla] Sequence ID: XP_030867412.3,
breast cancer type 1 susceptibility protein [Pan paniscus] Sequence ID: NP_001288687.1,
breast cancer type 1 [Gorilla gorilla] Sequence ID: AAT44835.1

```
>BRCA1 Pan troglodytes
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKG
>BRCA1 Gorilla gorilla gorilla
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKG
>BRCA1 Pan paniscus
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKG
>BRCA1 Gorilla gorilla
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKG

```

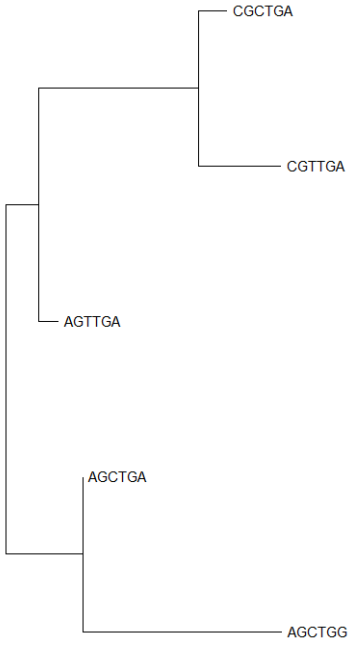
4 Филогенетика

4.1 Task 1



4.2 Task 2

	AGCTGA	AGTTGA	CGCTGA	AGCTGG	CGTTGA
AGCTGA	0	1	1	1	2
AGTTGA	1	0	2	2	1
CGCTGA	1	2	0	2	1
AGCTGG	1	2	2	0	2
CGTTGA	2	1	1	2	0



4.3 Task 3

4.4 Task 4

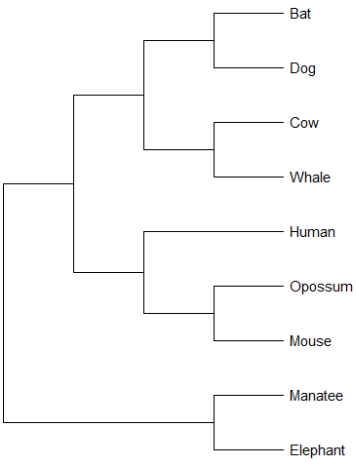


Рис. 1: MP

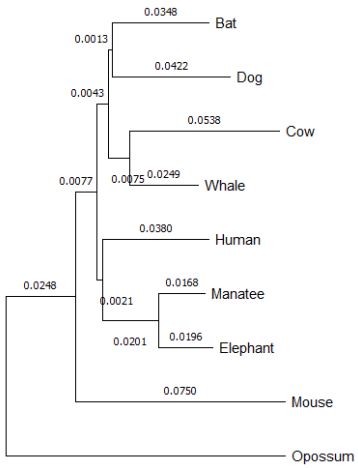


Рис. 2: NJ

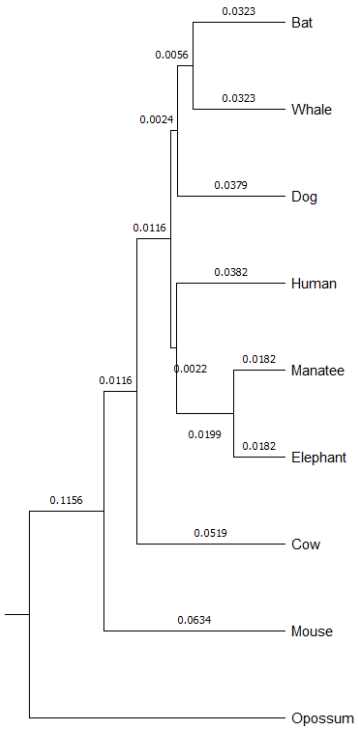


Рис. 3: UPGMA

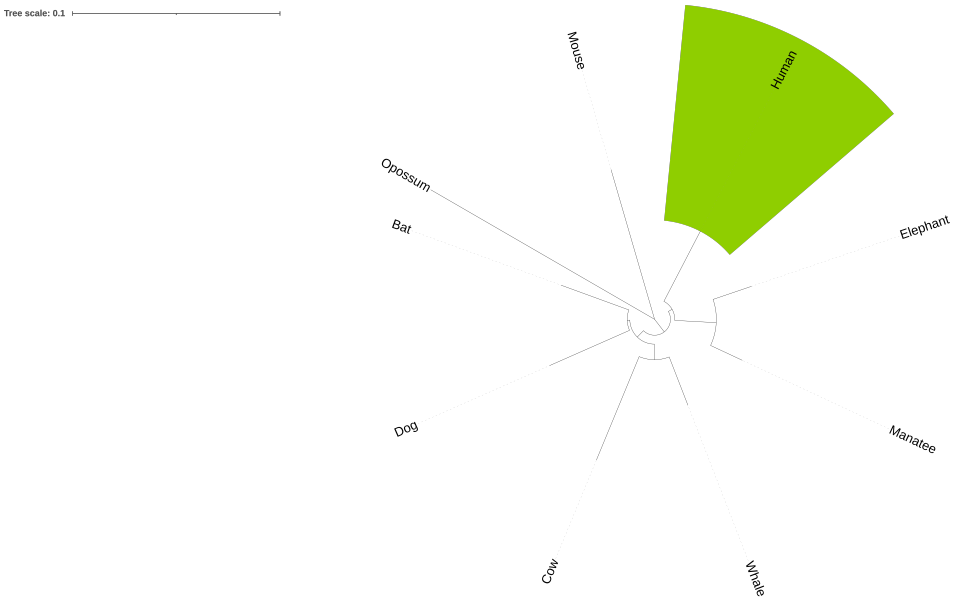
Кто ближайший сосед человеку – мышь или собака? Можем ли мы доказать независимое происхождение ламантинов и китообразных? Кто ближайший сосед летучим мышам – собака или человек?

MP
- Мышь
- Да
- Человек

NJ
- Собака
- Нет
- Человек

UPGMA
- Собака
- Нет
- Собака

4.5 Task 5



4.6 Task 6