Научно-исследовательский университет
“Высшая школа экономики”
+
Московский институт электроники и
математики им. А.Н.Тихонова

Майнор
“Эпигеномика”

Москва
2025

Проект по биоинформатике

Исследуемый таксон:
Amoebozoa (амёбозои)

Группа: 5

Участники

Беренштейн Аркадий, Аладышев Дмитрий,
Фадеева Анна, Харламов Вадим,
Пашенцев Павел, Шиверских Елизавета,
Садковская Маргарита, Голованкова Светлана,
Гульев Алексей, Ивонинская Алина

Участник	Организм	Среда обитания	GC
Садковская Маргарита	Entamoeba nuttalli P19	Кишечник макак	25%
Беренштейн Аркадий	Acanthamoeba castellanii	Почва, пресная вода, и т. д.	58,5%
Пашенцев Павел	Dictyostelium discoideum AX4	Почва лесов Северной Америки	22,5%
Шиверских Елизавета	Entamoeba invadens	Кишечник рептилий	30%
Гульев Алексей	Entamoeba histolytica HM-1:IMSS	Кишечник человека	24,5%
Аладышев Дмитрий	Entamoeba dispar SAW760	Кишечник человека	24%
Ивонинская Алина	Naegleria fowleri	Вода, почва, кишечник человека и животных	29,5%
Фадеева Анна	Dictyostelium purpureum	Почва и подстилка из опавших листьев в лесах и полях	24,5%
Голованкова Светлана	Acytostelium subglobosum LB1	Влажная почва, листовой опад, гниющая древесина	44,2%
Харламов Вадим	Dictyostelium firmibasis	Почвенная среда Японии	24%

Индивидуальная часть

Acanthamoeba castellanii

Свободноживущий амебоидный протозой

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly Acastellanii.strNEFF v1
Идентификатор в базе NCBI	GCF_000313135.1
Длина генома	42 Mb
Число генов	15 650

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	`model:` HG kouzine `model_confidence_threshold:` 0.5 `minimum_sequence_length:` 10

Acanthamoeba castellanii

Распределение вторичных структур по участкам генома

Участок	Число G4	Доля G4	Доля G4 по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка от всего генома
Exons	787	11.87%	0.11%	58356	58.01%	9.48%	43510	45.47%	3.72%	39.02%
Introns	471	7.1%	0.12%	45224	44.96%	17.78%	47715	49.87%	11.5%	33.59%
Promoters (1000 up from TSS)	891	13.43%	0.15%	25246	25.1%	5.01%	26310	27.5%	3.16%	5.77%
Downstream (200 bp)	217	3.27%	0.2%	4903	4.87%	4.61%	4654	4.86%	2.65%	5.77%
Intergenic	968	14.6%	0.28%	10668	10.61%	2.91%	9258	9.68%	1.52%	5.87%

Acanthamoeba castellanii

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	787	0.74%	≈3%	41270	39.02%	14.9%	43510	41.14%	≈10%
Introns	471	0.52%	≈23%	37503	41.19%	47.1%	47715	52.4%	≈30%
Promoters	891	5.7%	≈4%	12077	77.21%	35%	26310	68.21%	≈30%
Downstream	217	1.39%	≈0%	4663	29.82%	3%	4654	29.76%	≈0%
Intergenic	845	5.31%	≈70%	6180	38.86%	0%	10386	65.31%	≈30%

Entamoeba histolytica HM-1:IMSS

Вид паразитических простейших типа амёбозои

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly JCVI-ESG2-1.0
Идентификатор в базе NCBI	GCF_000208925.1
Длина генома	20.8 Mb
Число генов	8 327

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	`model:` HG kouzine `model_confidence_threshold:` 0.5 `minimum_sequence_length:` 10

Entamoeba histolytica HM-1:IMSS

Распределение вторичных структур по участкам генома

Участок	Число G4*	Доля G4	Доля G4 по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка от всего генома
Exons	2611	84.44%	0.58%	44	52.38%	0.01%	0	0.0%	0.0%	27.41%
Introns	19	0.61%	0.21%	0	0.0%	0.0%	0	0.0%	0.0%	6.46%
Promoters	577	18.66%	0.22%	14	16.67%	0.01%	3	27.27%	0.0%	20.85%
Downstream	113	3.65%	0.17%	2	2.38%	0.0%	0	0.0%	0.0%	20.77%
Intergenic	494	15.98%	0.1%	40	47.62%	0.01%	11	100.0%	0.0%	24.5%

*поиск производился по ослабленному паттерну

Entamoeba histolytica HM-1:IMSS

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	1942	24.25%	3%	44	0.41%	14.9%	0	0.0%	10%
Introns	19	0.75%	23%	0	0.0%	47.1%	0	0.0%	30%
Promoters	510	7.05%	4%	15	0.18%	35%	3	0.04%	30%
Downstream	113	1.38%	0%	2	0.02%	3%	0	0.0%	0%
Intergenic	389	5.13%	70%	26	0.27%	0%	11	0.11%	30%

Entamoeba dispar SAW760

Амёба, колонизирующая кишечник человека без инвазивных последствий

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly JCVI_EDISG_1.0
Идентификатор в базе NCBI	GCF_000209125.1
Длина генома	30.6 Mb
Число генов	8 814

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	`model:` HG kouzine `model_confidence_threshold:` 0.5 `minimum_sequence_length:` 10

Entamoeba dispar SAW760

Распределение вторичных структур по участкам генома

Участок	Число G4*	Доля G4	Доля G4 по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка от всего генома
Exons	2566	38.46%	0.54%	89	14.35%	0.02%	19	4.74%	0.003%	20.56%
Introns	19	0.28%	0.2%	1	0.16%	0.01%	1	0.25%	0.008%	5.67%
Promoters (1000 up from TSS)	625	9.37%	0.23%	28	4.52%	0.01%	11	2.74%	0.003%	14.82%
Downstream (200 bp)	104	1.56%	0.15%	6	0.97%	0.01%	6	1.5%	0.005%	14.38%
Intergenic	4117	61.71%	0.53%	532	85.81%	0.08%	382	95.26%	0.028%	35.03%

*поиск производился по ослабленному паттерну

Entamoeba dispar SAW760

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	1926	15.8%	≈3%	76	0.62%	14.9%	12	0.1%	≈10%
Introns	17	0.5%	≈23%	1	0.03%	47.1%	1	0.03%	≈30%
Promoters	570	6.5%	≈4%	28	0.32%	35%	10	0.11%	≈30%
Downstream	102	1.2%	≈0%	6	0.07%	3%	6	0.07%	≈0%
Intergenic	2441	11.8%	≈70%	495	2.39%	0%	356	1.72%	≈30%

Entamoeba invadens

Паразит, обитающий в кишечнике рептилий

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly Acastellanii.strNEFF v1
Идентификатор в базе NCBI	GCF_000330505.1
Длина генома	40.9 Mb
Число генов	12.007

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	`model:` HG kouzine `model_confidence_threshold:` 0.5 `minimum_sequence_length:` 10

Entamoeba invadens

Распределение вторичных структур по участкам генома

Участок	Число G4	Доля G4	Доля G4 по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка от всего генома
Exons	19	19.58%	0.17%	12498	75.35%	0.04%	489	88.9%	0.01%	21.65%
Introns	0	0.0%	0.0%	113	0.7%	1.52%	1	0.18%	1.04%	6.75%
Promoters (1000 up from TSS)	13	13.4%	0.29%	4460	26.89%	0.01%	153	27.8%	0.01%	14.88%
Downstream (200 bp)	0	0.0%	1.44%	660	3.97%	0.02%	16	2.9%	0.01%	14.88%
Intergenic	78	80.41%	0.19%	4146	24.99%	0.001%	53	9.63%	0.001%	16.28%

Entamoeba invadens

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	18	15.8%	≈3%	11499	70.2%	14.9%	420	70.1%	≈10%
Introns	2	0.7%	≈23%	132	0.9%	47.1%	1	0.18%	≈30%
Promoters	12	11.2%	≈4%	4301	25.8%	35%	153	27.8%	≈30%
Downstream	1	0.39%	≈0%	619	3.6%	3%	15	2.6%	≈0%
Intergenic	73	71.9%	≈70%	3981	23.5%	0%	60	11.2%	≈30%

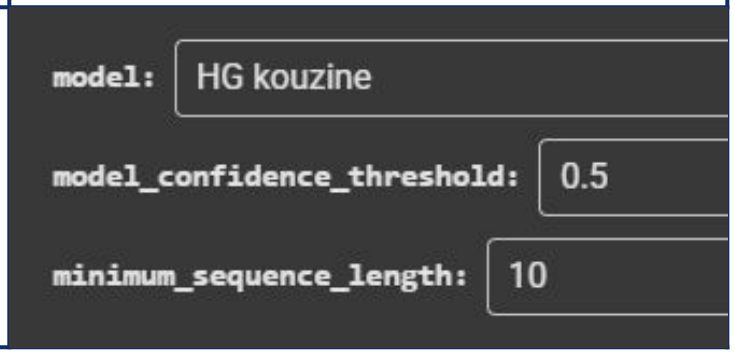
Entamoeba nuttalli P19

Кишечный паразит диких резус-макак Катманду, Непал

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly ENU1 v1
Идентификатор в базе NCBI	GCF_000257125.1
Длина генома	14.4 Mb
Число генов	6,193

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	

Entamoeba nuttalli P19

Распределение вторичных структур по участкам генома

Участок	Число G4	Доля G4	Доля G4 по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка от всего генома
Exons	1144	47.57 %	0.32%	46	0.01%	0.04%	1	3.7%	0.0%	23.60%
Introns	4	0.116 %	0.08%	0	0.0%	0.0%	0	0.0%	0.0%	5.03%
Promoters (1000 up from TSS)	406	16.9%	0.17%	15	22.72%	0.01%	1	3.7%	0.0%	18.45%
Downstream (200 bp)	60	2.5%	0.12%	3	4.5%	0.01%	0	0.0%	0.0%	18.42%
Intergenic	541	22.5%	0.19%	20	30.3%	0.01%	26	96.3%	0.01%	34.02%

Entamoeba nuttalli P19

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	929	11.8	≈3%	46	0.585%	14.9%	1	0.013%	≈10%
Introns	4	0.2%	≈23%	0	0.0%	47.1%	0	0.0%	≈30%
Promoters	368	5.98%	≈4%	15	0.244%	35%	1	0.016%	≈30%
Downstream	56	0.91%	≈0%	3	0.049%	3%	0	0.0%	≈0%
Intergenic	431	3.8%	≈70%	18	0.159%	0%	7	0.062%	≈30%

Dictyostelium purpureum

Социальная амёба обитает в почве и подстилке из опавших листьев в лесах и полях

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly v1.0
Идентификатор в базе NCBI	GCF_000190715.1
Длина генома	33 Mb
Число генов	12 399

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	`model:` HG kouzine `model_confidence_threshold:` 0.5 `minimum_sequence_length:` 10

Dictyostelium purpureum

Распределение вторичных структур по участкам генома

Участок	Число G4	Доля G4	Доля G4 по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка от всего генома
Exons	10	8.3%	0.0023%	800	8.4%	0.0802%	3	30%	0.000259%	35.53%
Introns	14	6%	0.0077%	965	10.1%	0.5813%	0	0%	0%	21.39%
Promoters (1000 up from TSS)	34	20.2%	0.007%	4836	50.8%	0.9632%	0	0%	0%	14.07%
Downstream (200 bp)	17	10.1%	0.0141%	377	4%	0.3125%	1	10%	0.000529%	14.13%
Intergenic	115	68.5%	0.0251%	7868	82.6%	1.426%	7	70%	0.000961%	14.88%

Dictyostelium purpureum

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	13	0.04%	≈3%	137	0.44%	14.9%	3	0.009%	≈10%
Introns	10	0.05%	≈23%	39	0.47%	47.1%	0	0%	≈30%
Promoters	33	0.27%	≈4%	520	4.22%	35%	0	0%	≈30%
Downstream	17	0.14%	≈0%	41	0.33%	3%	1	0.008%	≈0%
Intergenic	108	0.83%	≈70%	523	4.01%	0%	7	0.05%	≈30%

Dictyostelium firmibasis

Амеба обитающая в почве во влажных лесах

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly ASM3616959v1
Идентификатор в базе NCBI	GCA_036169595.1
Длина генома	31.4 Mb
Число генов	11044

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	`model:` HG kouzine `model_confidence_threshold:` 0.5 `minimum_sequence_length:` 10

Dictyostelium firmibasis

Распределение вторичных структур по участкам генома

Участок	Число G4	Доля G4	Доля G4 по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка в геноме
Exons	10	4.63%	0.0023%	1645	15.52%	0.0802%	3	30%	0.0003%	20.5%
Introns	2	0.93%	0.0077%	1134	10.7%	0.5813%	0	0%	0%	11%
Promoters (1000 up from TSS)	25	11.57 %	0.007%	7494	70.71%	0.9632%	2	20%	0.0001%	10.5%
Downstream (200 bp)	4	1.85%	0.0141%	352	3.32%	0.3125%	1	10%	0.00006%	26.5%
Intergenic	204	94.44 %	0.0251%	7886	74.41%	1.426%	4	40%	0.0009%	34%

Dictyostelium firmibasis

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	12	0.13%	≈4%	120	0.36%	19.9%	4	0.02%	≈10%
Introns	9	0.1%	≈4%	33	0.47%	45.4%	0	0%	≈30%
Promoters	38	0.5%	≈11%	499	5.18%	15.1%	0	0%	≈33%
Downstream	7	0.1%	≈0%	27	0.33%	5.9%	3	0.01%	≈0%
Intergenic	180	0.75%	≈81%	453	5.19%	0%	6	0.03%	≈27%

Dictyostelium discoideum AX4

Социальная амёба из почв лесов Северной Америки

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly dicty_2.7
Идентификатор в базе NCBI	GCF_000004695.1
Длина генома	34.1 Mb
Число генов	11044

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	`model:` HG kouzine `model_confidence_threshold:` 0.5 `minimum_sequence_length:` 10

Dictyostelium discoideum AX4

Распределение вторичных структур по участкам генома

Участок	Число квадруплексов	Доля квадруплексов	Доля квадруплексов по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка в геноме
Exons	13	9.35	0.11%	135	13.35	9.48%	1	7.69	3.72%	26.65
Introns	3	2.16	0.12%	170	16.82	17.78%	0	0.00	11.5%	14.68
Promoters (1000 up)	37	26.62	0.15%	581	57.47	5.01%	3	23.08	3.16%	12.14
Downstream (200 bp)	10	7.19	0.20%	87	8.61	4.61%	0	0.00	2.65%	12.12
Intergenic	88	63.31	0.28%	739	73.10	2.91%	12	92.31	1.52%	12.08

Dictyostelium discoideum AX4

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	13	0.04	≈3%	132	0.43	14.9%	1	0.003	≈10%
Introns	3	0.02	≈23%	167	0.99	47.1%	0	0.000	≈30%
Promoters (1000 up)	37	0.27	≈4%	690	4.94	35%	3	0.020	≈30%
Downstream (200 bp)	10	0.07	≈0%	101	0.72	3%	0	0.000	≈0%
Intergenic	83	0.60	≈70%	638	4.59	0%	13889	99.950	≈30%

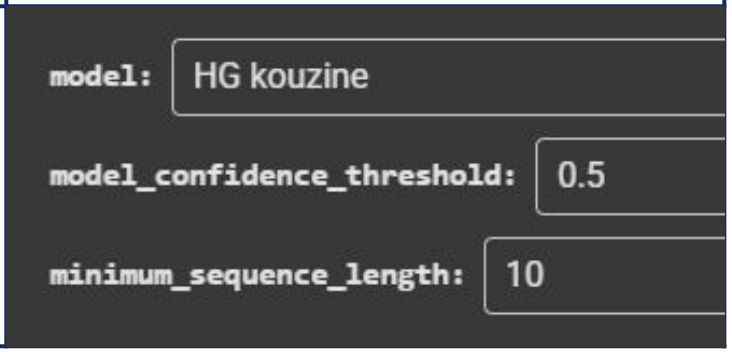
Naegleria fowleri

Смертельная термофильная амёба

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly ASM840351v1
Идентификатор в базе NCBI	GCF_008403515.1
Длина генома	29.5 Mb
Число генов	13 854

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	

Naegleria fowleri

Распределение вторичных структур по участкам генома

Участок	Число G4	Доля G4	Доля G4 по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка от всего генома
Exons	73	45.06%	0.27%	675	2.53%	13.46%	12	10.91%	0.04%	24.56%
Introns	12	7.1%	0.09%	304	2.37%	21.95%	5	4.55%	0.04%	11.81%
Promoters (1000 up from TSS)	13	8.02%	0.09%	1029	7.44%	60.67%	91	82.73%	0.66%	12.73%
Downstream (200 bp)	12	7.41%	0.09%	263	1.91%	14.97%	18	16.36%	0.13%	12.69%
Intergenic	0	0.00%	0.00%	881	6.36%	53.52%	94	85.45%	0.68%	12.74%

Naegleria fowleri

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	61	0.23%	3%	643	2.41%	14.9%	12	0.04%	10%
Introns	11	0.09%	23%	282	2.20%	47.1%	5	0.04%	30%
Promoters (1000 up from TSS)	14	0.10%	4%	1246	9.00%	35%	117	0.85%	30%
Downstream (200 bp)	9	0.07%	0%	288	2.09%	3%	18	0.13%	0%
Intergenic	0	0.00%	70%	779	5.62%	0%	87	0.63%	30%

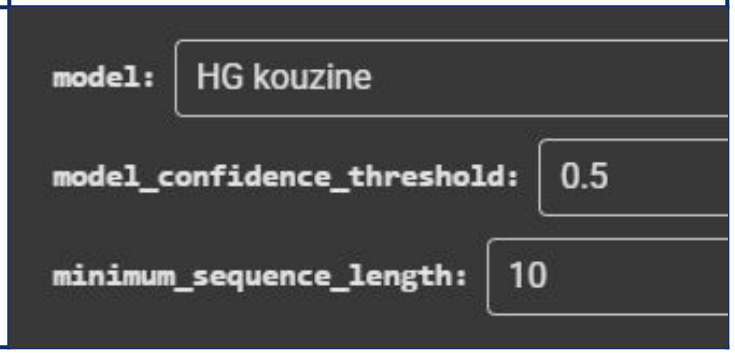
Acytostelium subglobosum LB1

редкий вид слизевиков (миксомицетов), относящийся к порядку Dictyosteliida

1. Для исследования использовалась сборка из базы NCBI.
Информация о сборке представлена ниже:

Название генома	Genome assembly Asub_2.0
Идентификатор в базе NCBI	GCF_000787575.1
Длина генома	31 Mb
Число генов	12 682

2. Далее геном обрабатывался программами Z-Hunt и ZDNAbert.
Параметры запуска представлены ниже:

Параметры запуска Z-Hunt	`dinucleotides = 12;` `min, max = 8, 12`
Параметры запуска ZDNAbert	

Acytostelium subglobosum LB1

Распределение вторичных структур по участкам генома

Участок	Число G4	Доля G4	Доля G4 по длине	Число предсказаний Z-hunt	Доля предсказаний Z-hunt	Доля Z-ДНК по длине	Число предсказаний ZDNAbert	Доля предсказаний ZDNAbert	Доля предсказаний ZDNAbert по длине	Доля участка от всего генома
Exons	30	2,58%	0.005%	42199	57%	4.05%	5066	87,35%	0.38%	30.15%
Introns	10	0,86%	0.014%	4152	5,6%	3.82%	202	3,48%	0.15%	19.32%
Promoters (1000 up from TSS)	217	18,67%	0.056%	33148	44,78%	5.10%	910	15,69%	0.10%	10.82%
Downstream (200 bp)	111	9,55%	0.13%	6668	9%	4.98%	276	4,75%	0.16%	10.82%
Intergenic	1123	96,64%	0.33%	29483	55.14%	5.82%	562	9,69%	0.09%	10.77%

Acytostelium subglobosum LB1

Распределение участков генома, содержащих вторичные структуры

Участок	Число участков с G4	Доля участков с G4	Доля участков с предсказанным G4 у человека	Число участков предсказаний Z-hunt	Доля участков с предсказанным Z-hunt	Доля участков с предсказанным Z-hunt у человека	Число участков предсказаний ZDNAbert	Доля участков с предсказанным ZDNAbert	Доля участков с предсказанным ZDNAbert у человека
Exons	30	0.09%	≈3%	4138	43,4%	14.9%	3267	56,33%	≈10%
Introns	10	0.04%	≈23%	492	5,16%	47.1%	191	3,29%	≈30%
Promoters (1000 up from TSS)	217	2.44%	≈4%	2409	25,26%	35%	788	13,58%	≈30%
Downstream (200 bp)	111	1.07%	≈0%	672	7,04%	3%	266	4,58%	≈0%
Intergenic	596	4.72%	≈70%	2247	23,56%	0%	476	8,2%	≈30%

Групповая часть

Доли участков с квадруплексами

Участок	E. nuttalli	A. castellanii	D. discoideum	E. invadens	E. histolytica	E. dispar	N. fowleri	D. purpureum	A. subglobosum	D. firmibasis
Exons	5.981%	0.74%	0.04%	15.8%	24.25%	15.83%	0.23%	0.04%	0.09%	0.13%
Introns	0.239%	0.52%	0.02%	0.7%	0.75%	0.5%	0.09%	0.05%	0.04%	0.1%
Promoters (1000 up from TSS)	5.981%	5.7%	0.27%	11.2%	7.05%	6.5%	0.10%	0.27%	2.44%	0.5%
Downstream (200 bp)	0.912%	1.39%	0.07%	0.39%	1.38%	1.2%	0.07%	0.14%	1.07%	0.1%
Intergenic	3.80%	5.31%	0.60%	71.9%	5.13%	11.77%	0.00%	0.83%	4.72%	0.75%

Доли участков с предсказанями Z-Hunt

Участок	E. nuttalli	A. castellanii	D. discoideum	E. invadens	E. histolytica	E. dispar	N. fowleri	D. purpureum	A. subglobosum	D. firmibasis
Exons	0.585%	39.02%	0.43%	70.2%	0.41%	0.62%	2.41%	0.44%	43,4%	0.36%
Introns	0	41.19%	0.99%	0.9%	0.0%	0.03%	2.20%	0.47%	5,16%	0.47%
Promoters (1000 up from TSS)	0.244%	77.21%	4.94%	25.8%	0.18%	0.32%	9.00%	4.22%	25,26%	5.18%
Downstream (200 bp)	0.049%	29.82%	0.72%	3.6%	0.02%	0.07%	2.09%	0.33%	7,04%	0.33%
Intergenic	0.159%	38.86%	4.59%	23.5%	0.27%	2.39%	5.62%	4.01%	23,56%	5.19%

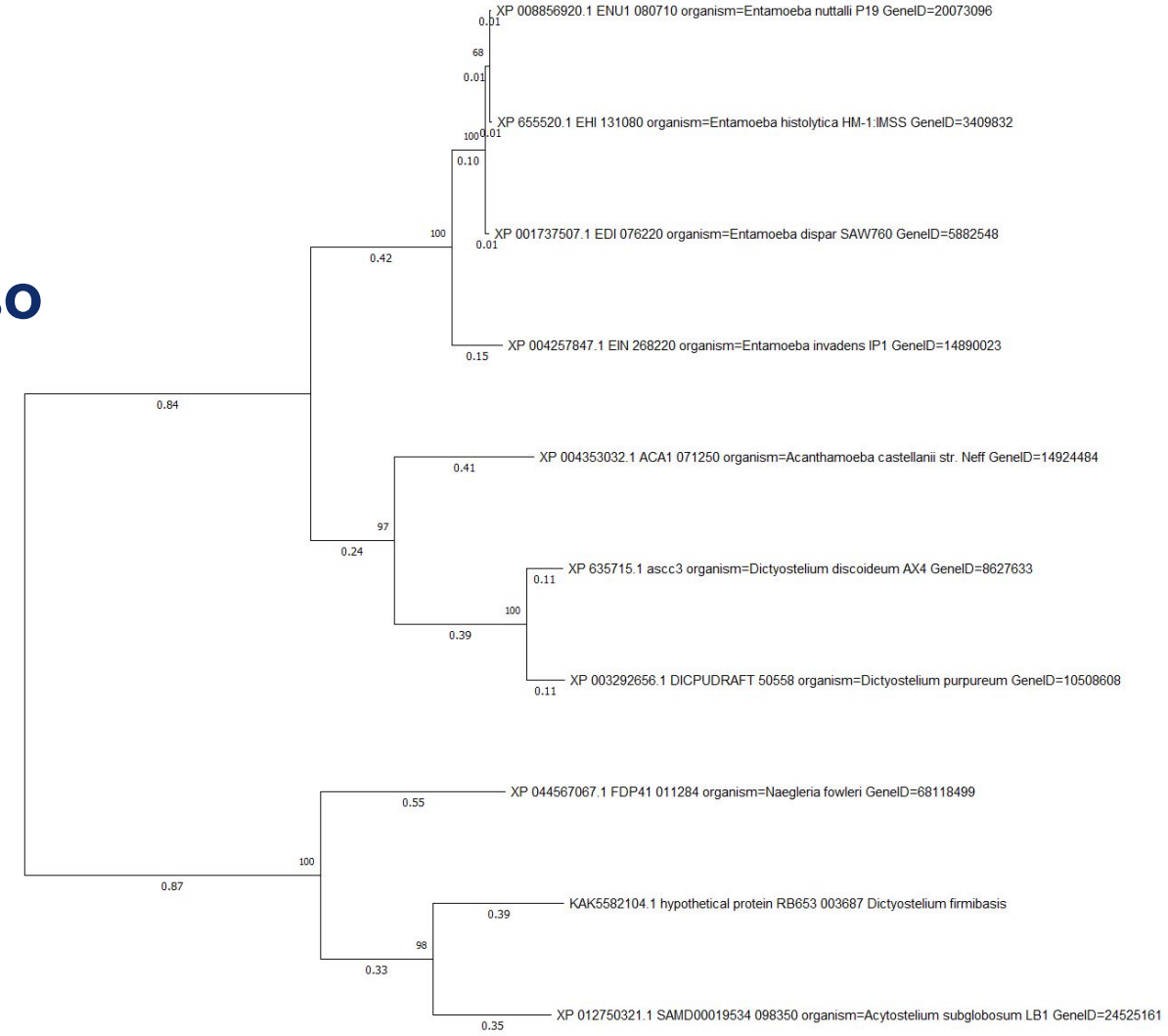
Доли участков с предсказаниями ZDNAbert

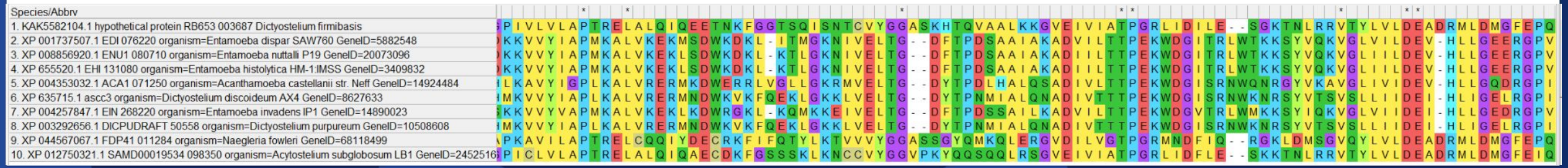
Участок	E. nuttalli	A. castellanii	D. discoideum	E. invadens	E. histolytica	E. dispar	N. fowleri	D. purpureum	A. subglobosum	D. firmibasis
Exons	0.013%	41.14%	0.003%	70.1%	0.0%	0.1%	0.04%	0.009%	56,33%	0.02%
Introns	0%	52.4%	0.000%	0.18%	0.0%	0.03%	0.04%	0.0%	3,29%	0%
Promoters (1000 up from TSS)	0.016%	68.21%	0.020%	27.8%	0.04%	0.11%	0.85%	0.0%	13,58%	0%
Downstream (200 bp)	0%	29.76%	0.000%	2.6%	0.0%	0.07%	0.13%	0.008%	4,58%	0.01%
Intergenic	0.062%	65.31%	99.950%	11.2%	0.11%	1.72%	0.63%	0.05%	8,2%	0.03%

Пункт 1
Гены, отвечающие за эпигенетику

Выравнивание и филогенетическое дерево по семейству DEAD

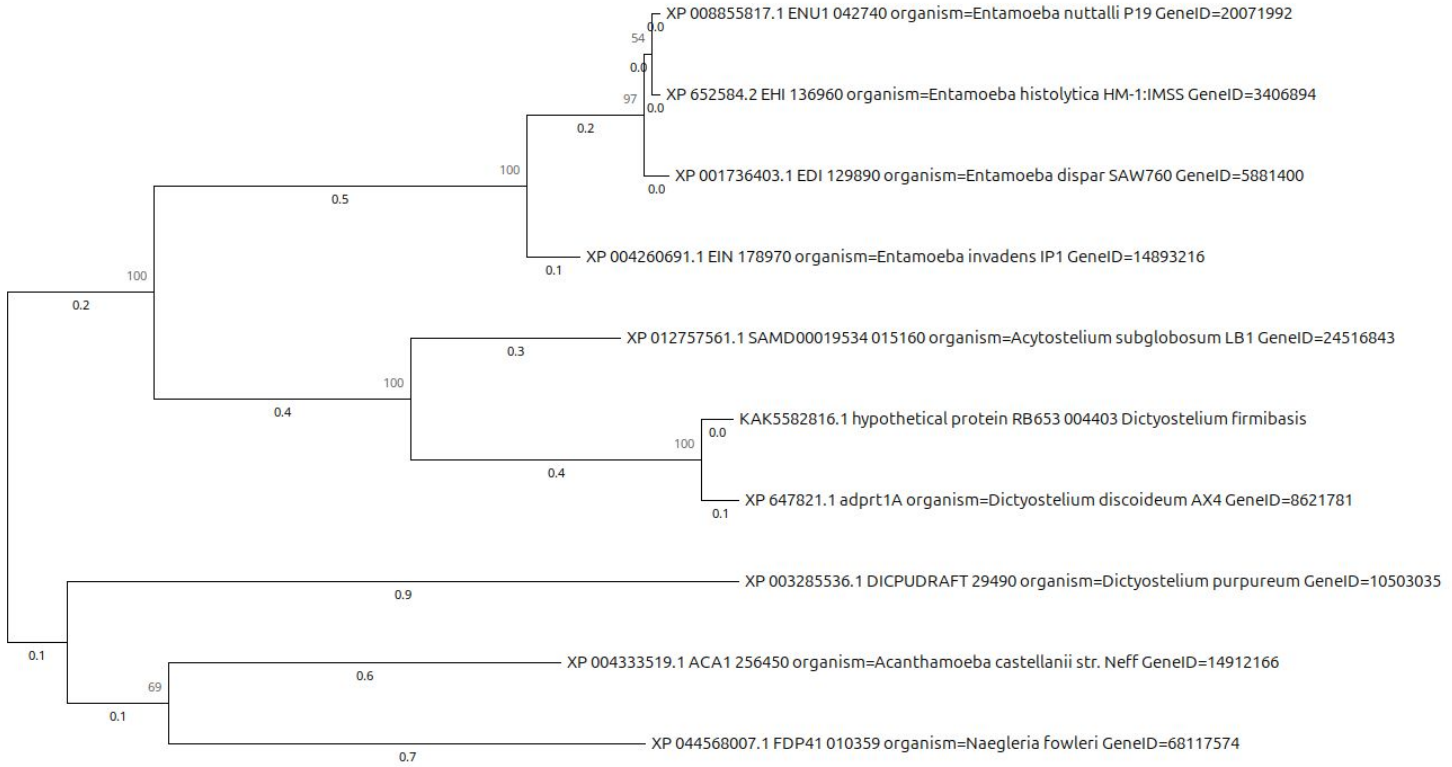
Семейство, влияющее на модификацию РНК

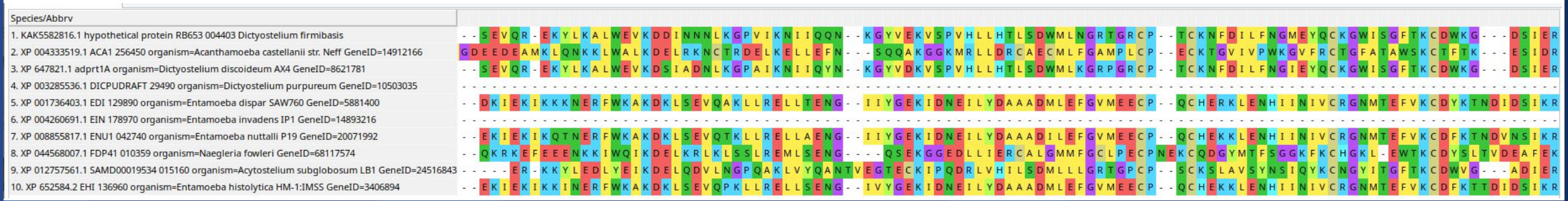




Выравнивание и филогенетическое дерево по семейству PARP

Семейство, влияющее на ремоделирование хроматина

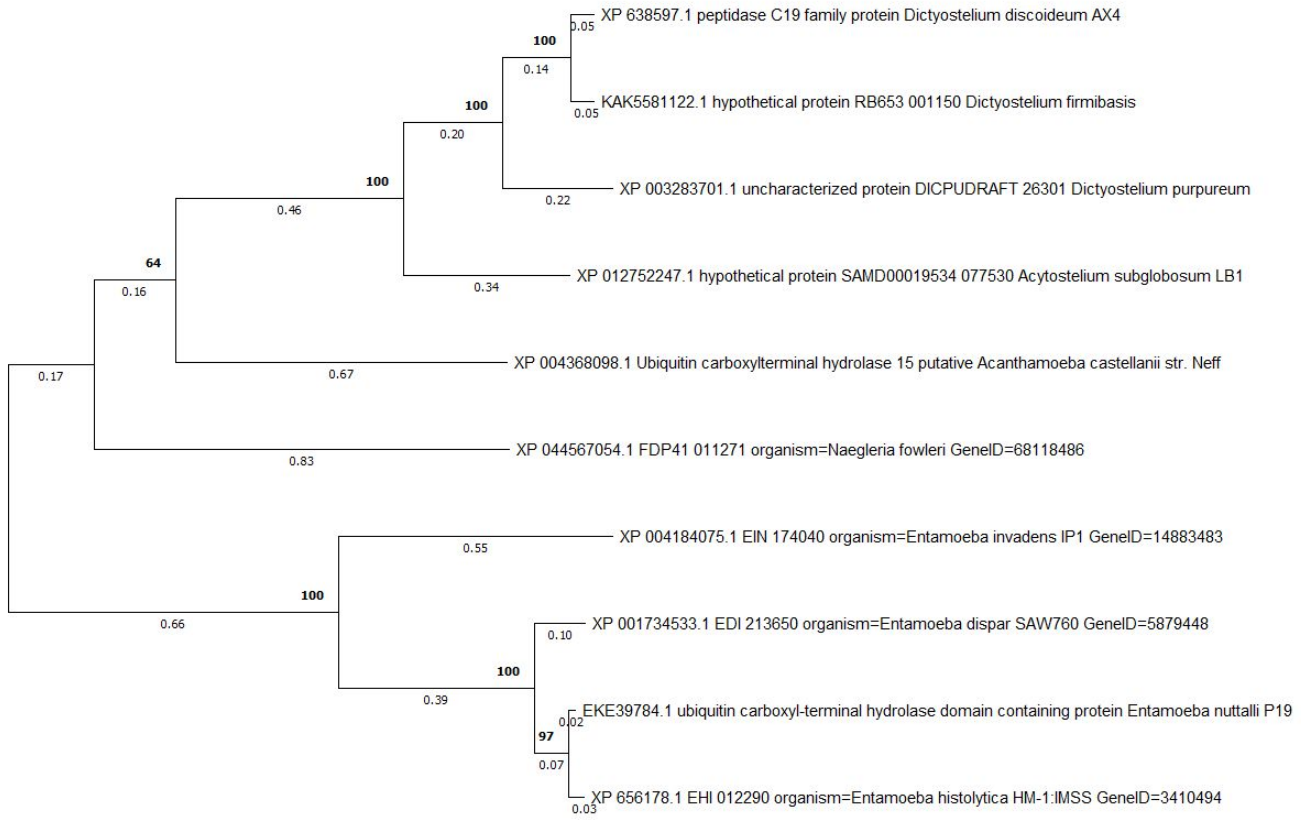


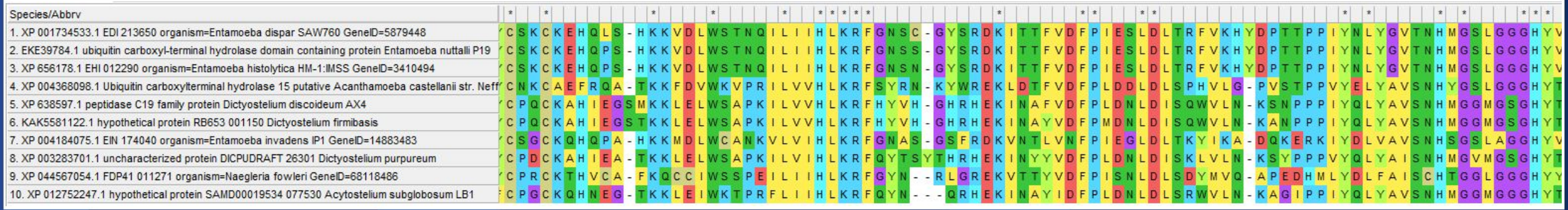


Выравнивание и филогенетическое дерево по семейству UCH

Семейство, отвечающее за стирание гистоновых модификаций

Влияет на следующие гистоновые метки: H2Aub, H2Bub

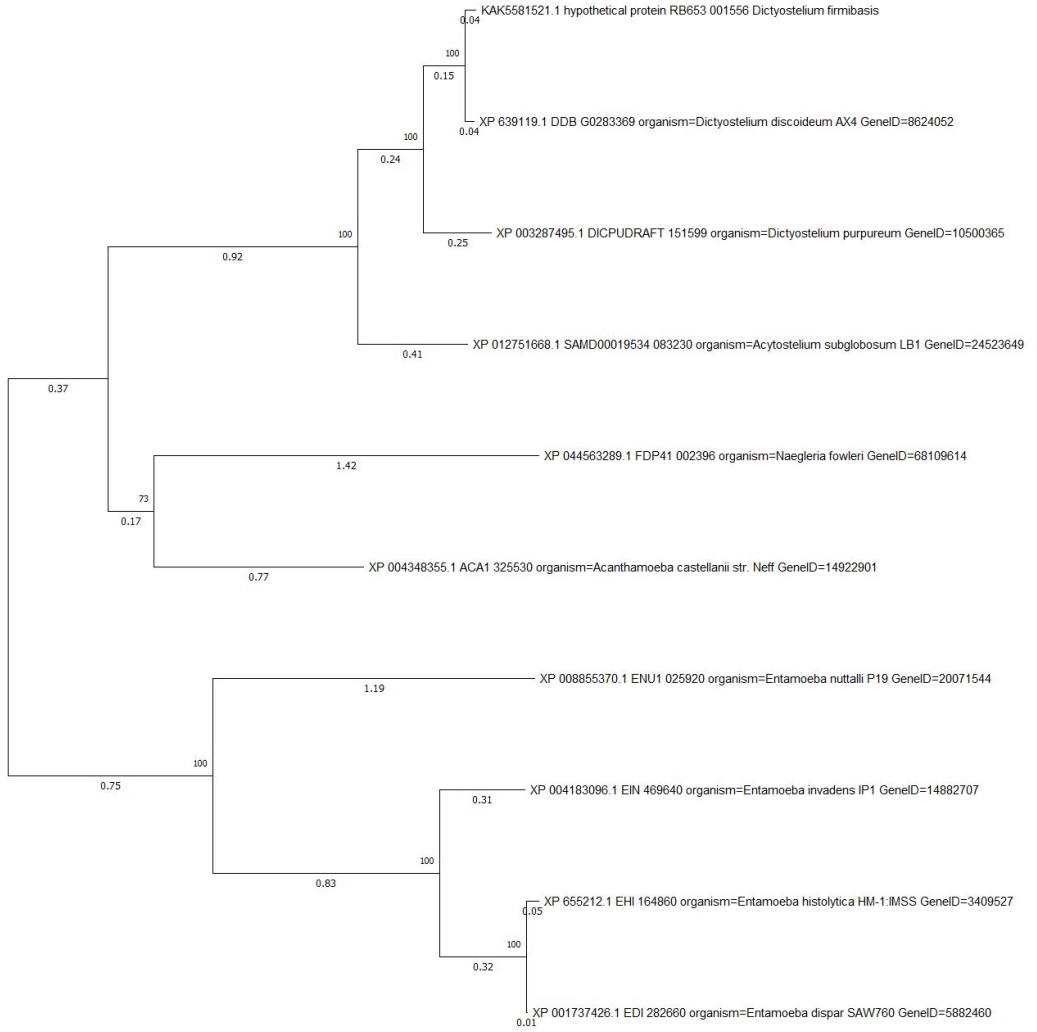


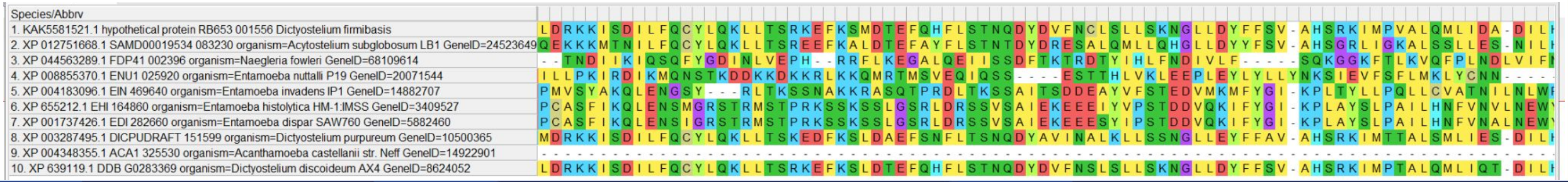


Выравнивание и филогенетическое дерево по семейству RCC1

Семейство, влияющее на ремоделирование хроматина

Влияет на следующие гистоновые метки: H2A, H2B

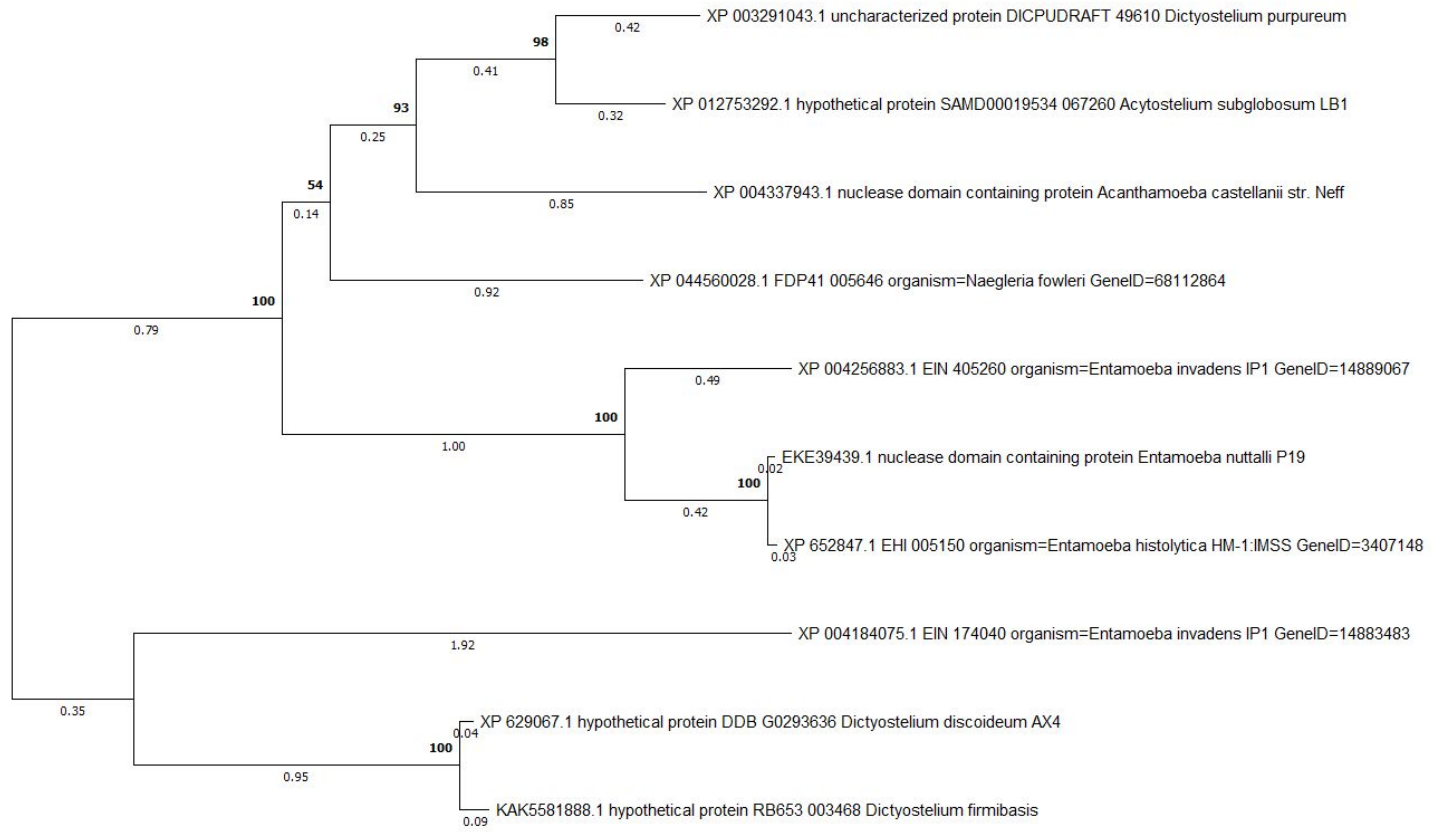


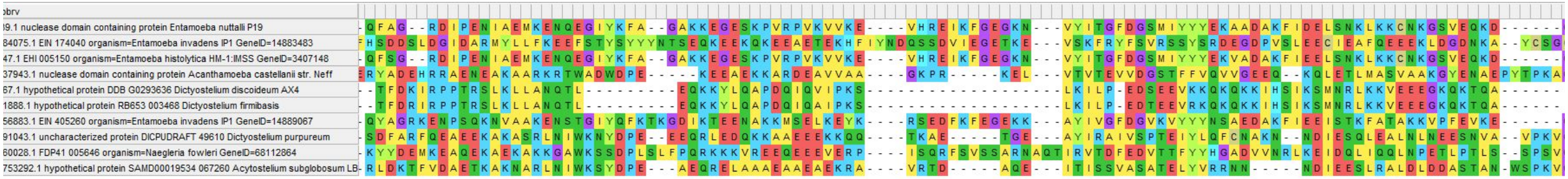


Выравнивание и филогенетическое дерево по семейству TUDOR

Семейство, отвечающее за чтение гистоновой модификации

Влияет на следующие гистоновые метки: H3R17me2a, H4R3me2a



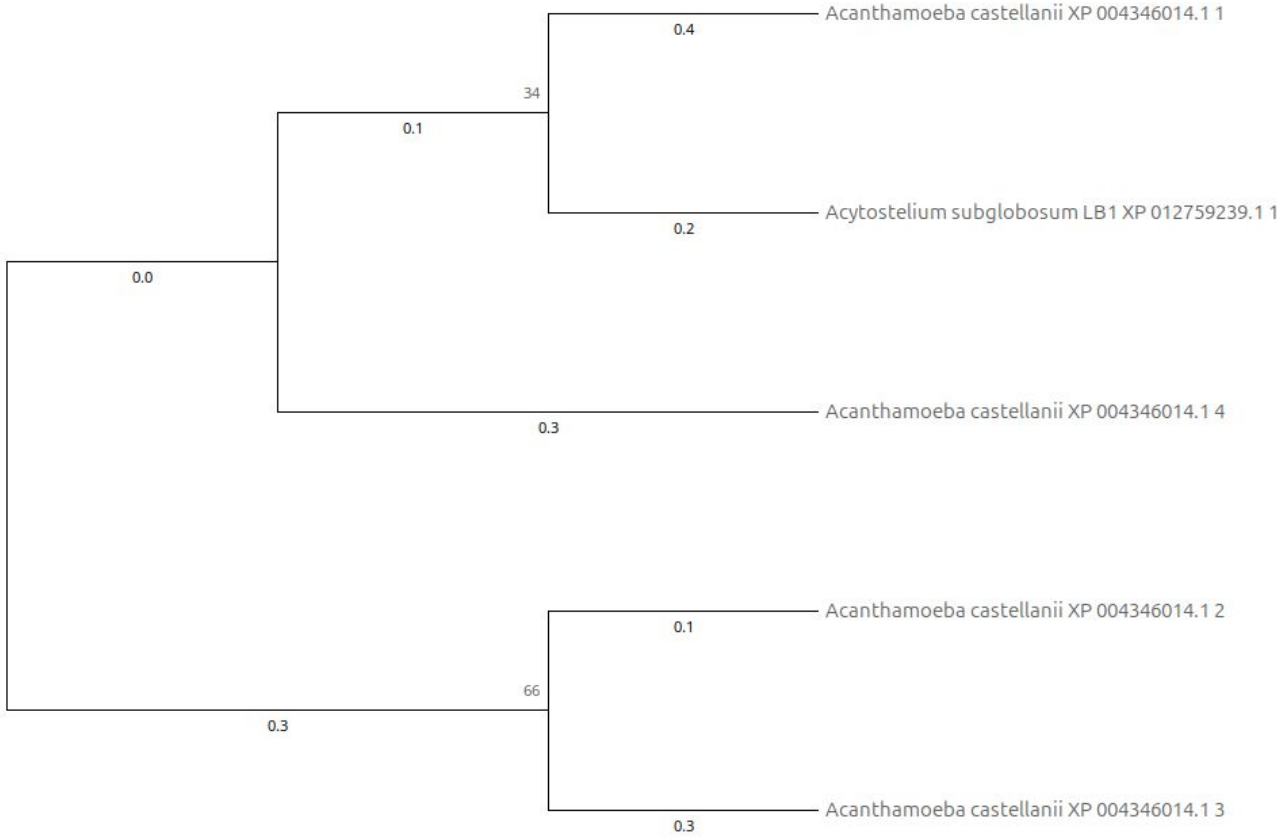


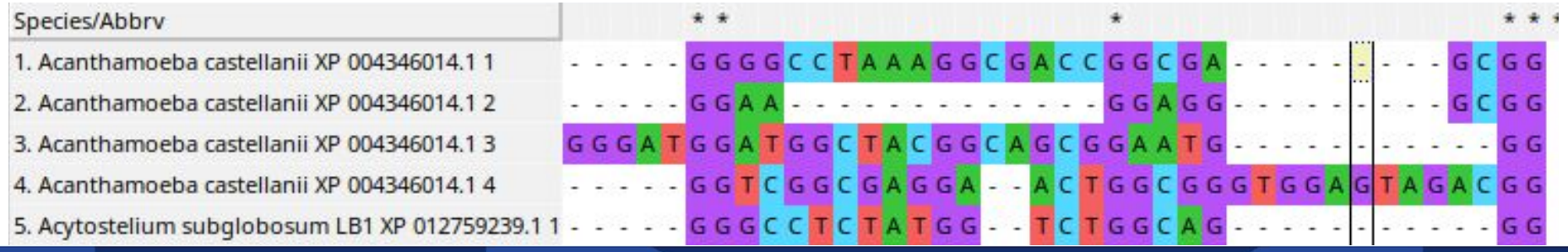
Пункт 2
Гены, отвечающие за квадруплексы

Выравнивание и филогенетическое дерево ортологичных генов

Ортогруппа субъединиц S10B регуляторной части 26S протеасомы (AAA-ATPase);

в некоторых организмах аннотирована как гипотетический / неохарактеризованный белок.

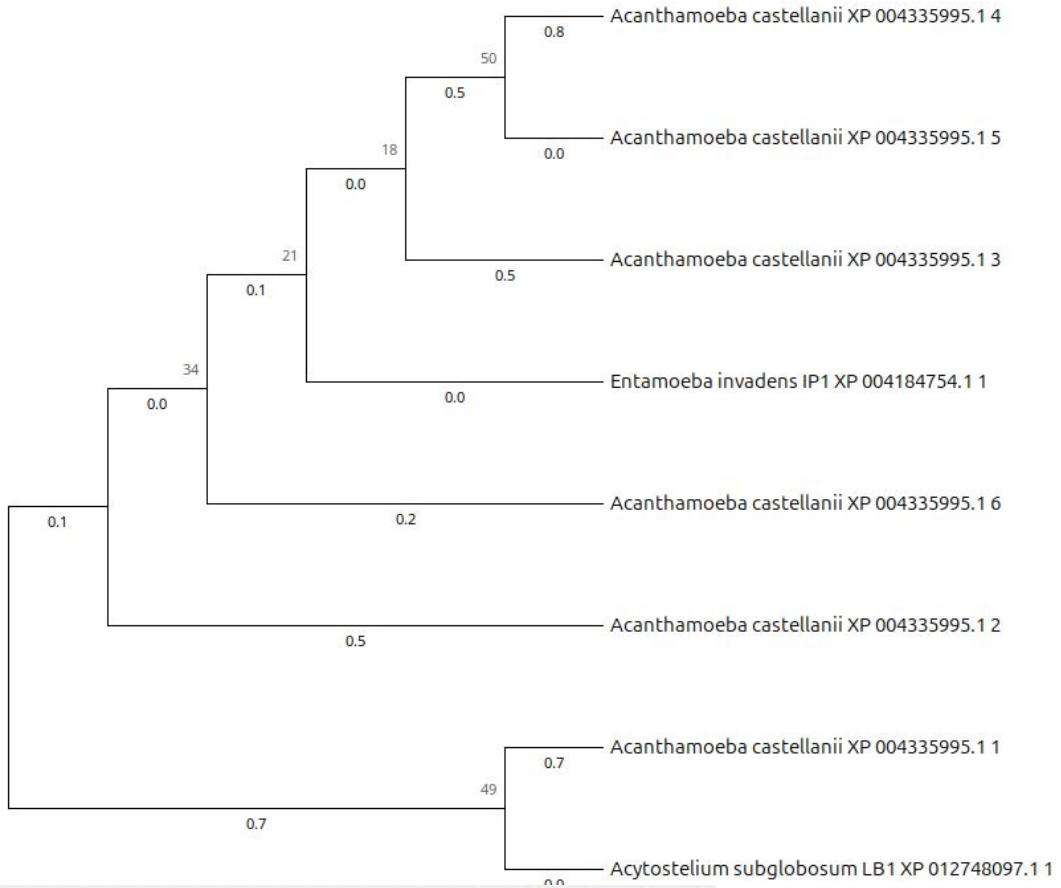


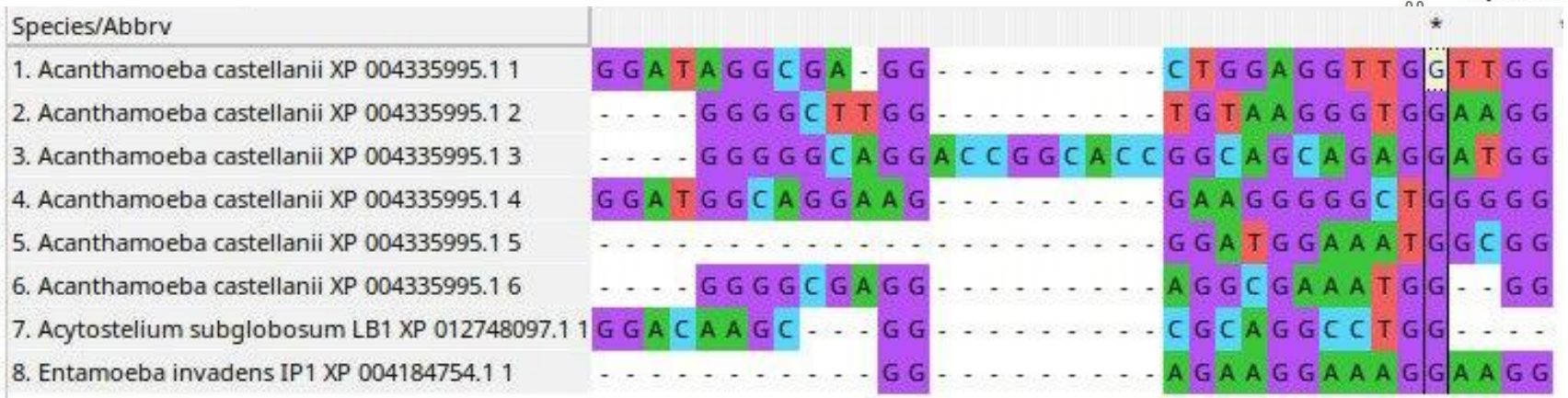


Выравнивание и филогенетическое дерево ортологичных генов

Ортогруппа субъединиц 6B (P45) регуляторной части 26S протеасомы.

В ряде организмов аннотирована как гипотетический или неохарактеризованный белок.

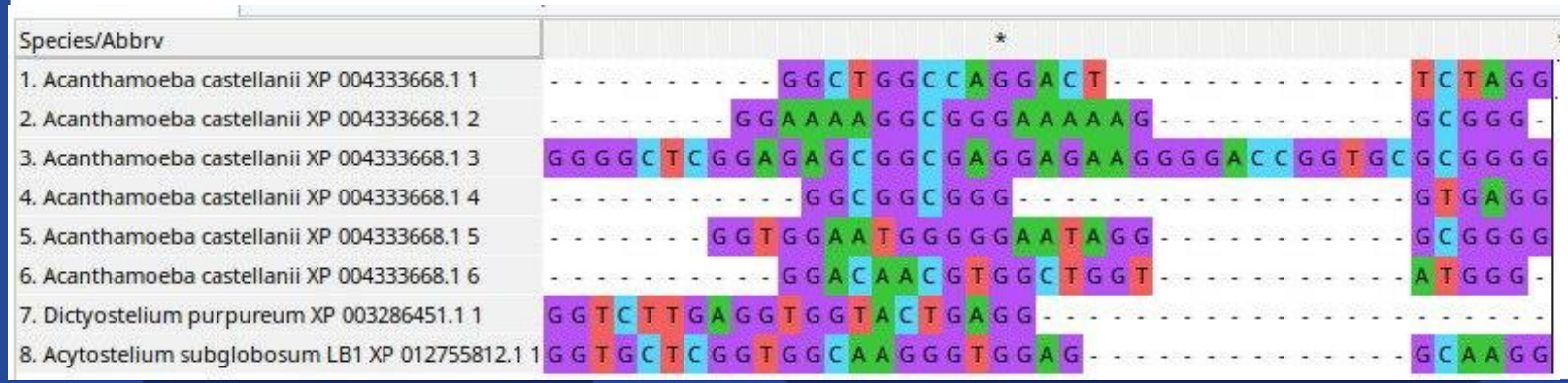


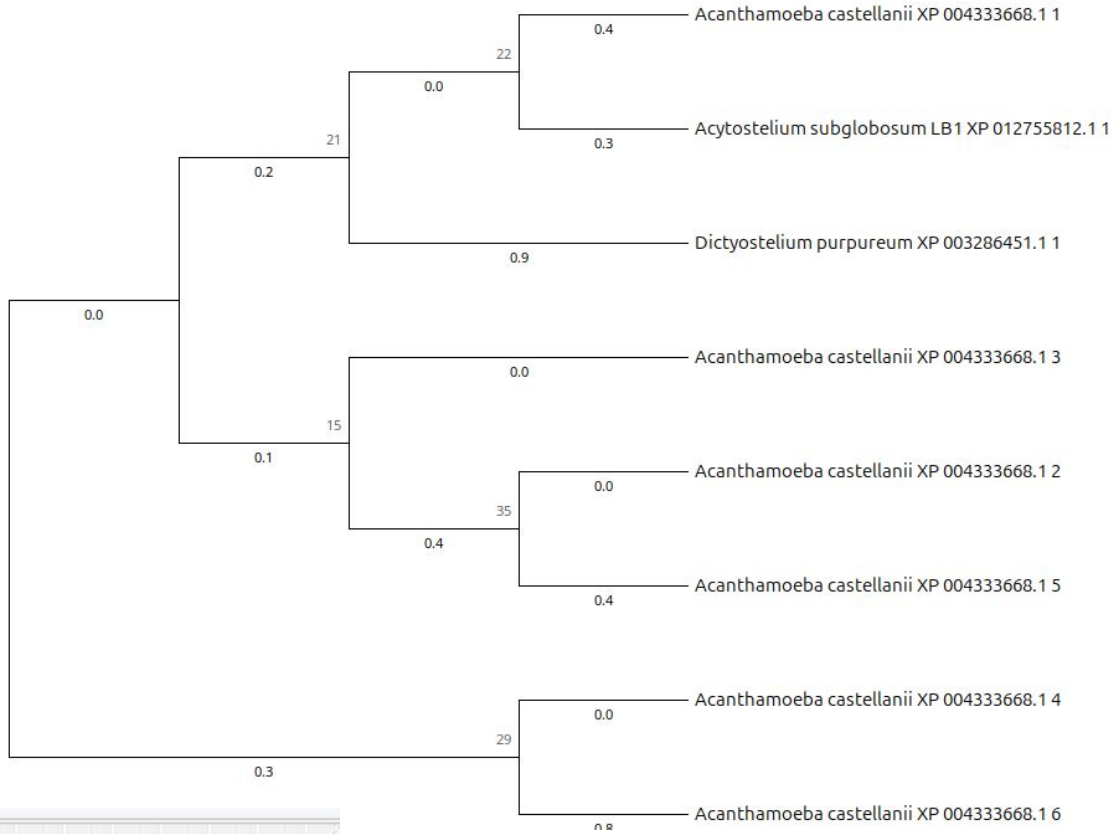


Выравнивание и филогенетическое дерево ортологичных генов

Ортогруппа субъединиц AAA-ATPase регуляторной части 26S протеасомы (например, RPT2, ATPase 1 subunit, regulatory subunit 4).

В некоторых организмах аннотирована как гипотетический или неохарактеризованный белок.

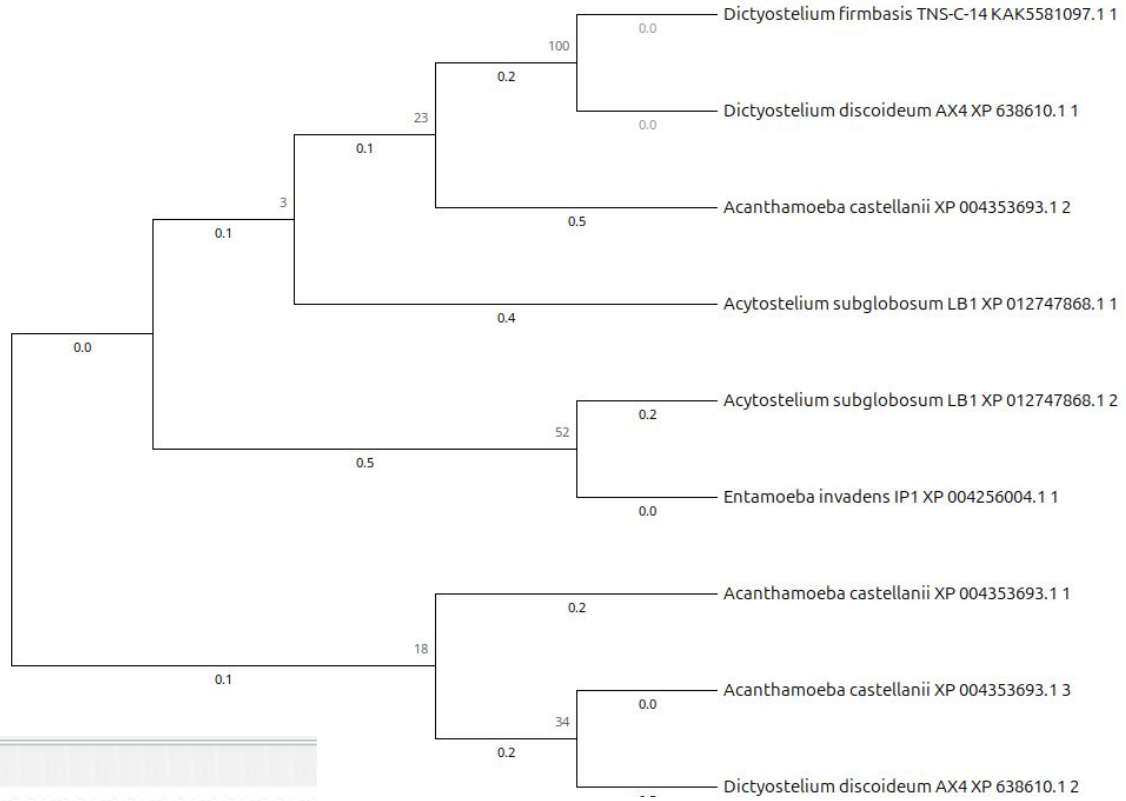


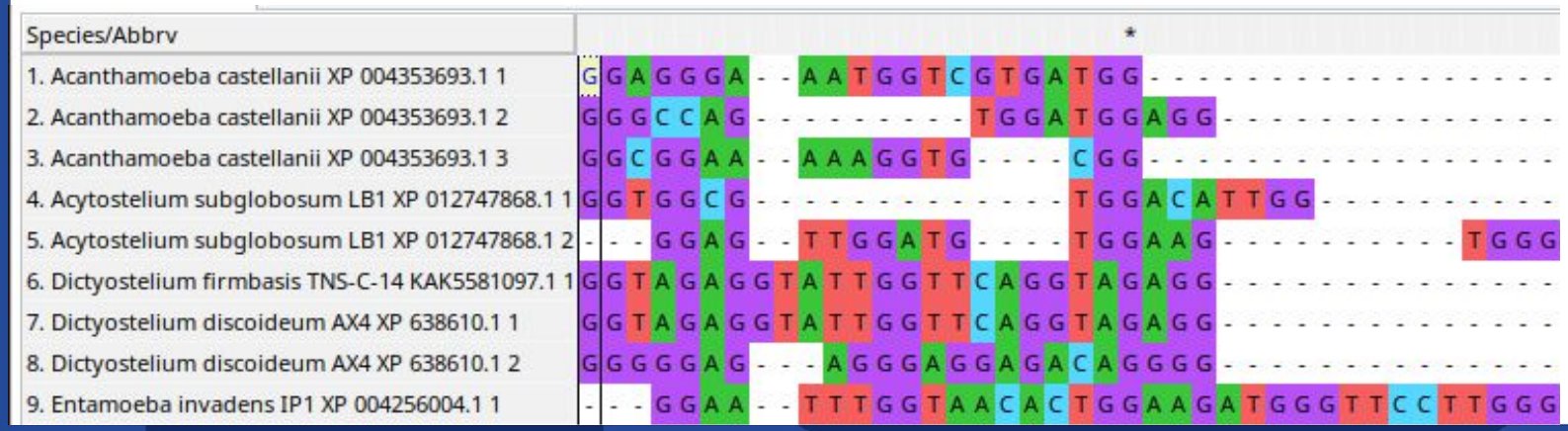


Выравнивание и филогенетическое дерево ортологичных генов

Ортогруппа субъединиц B вакуолярной (V-тип) АТФазы (ATP-синтазы).

В ряде организмов аннотирована как гипотетический или неохарактеризованный белок.



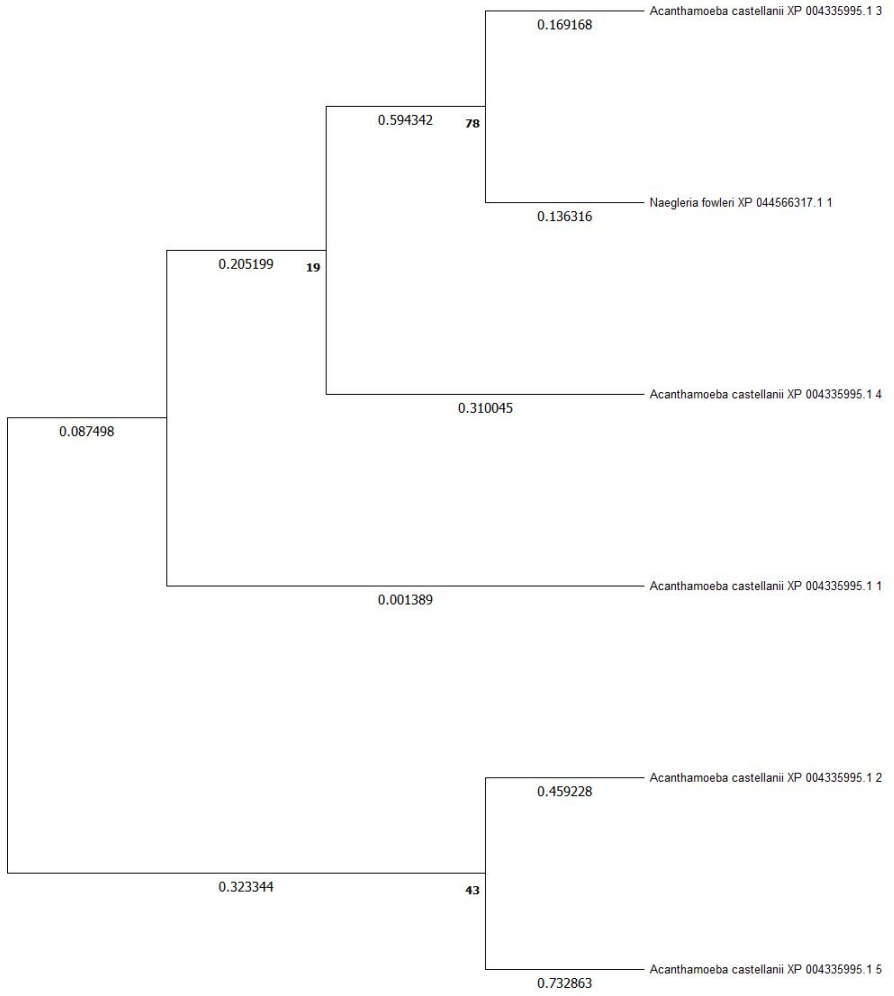


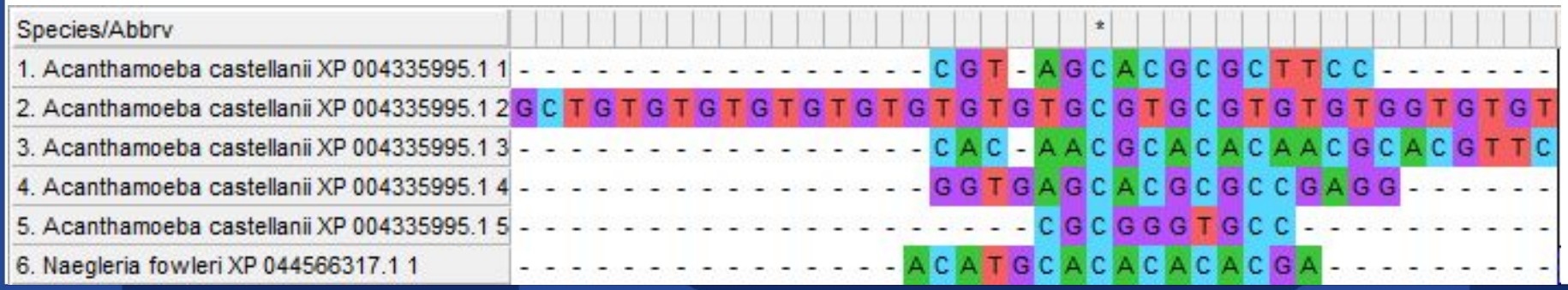
Пункт 3
Гены, отвечающие за Z-ДНК

Выравнивание и филогенетическое дерево ортологичных генов

Ортогруппа субъединиц AAA-ATPase регуляторной части 26S протеасомы (например, RPT2, ATPase 1 subunit, regulatory subunit 4).

В некоторых организмах аннотирована как гипотетический или неохарактеризованный белок.

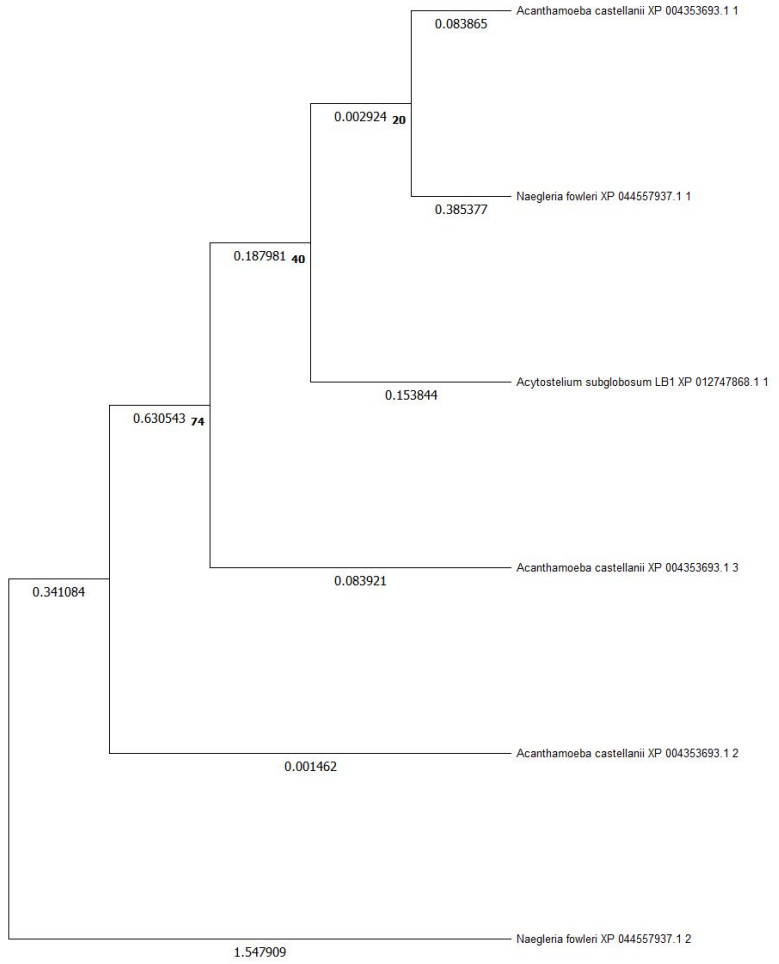


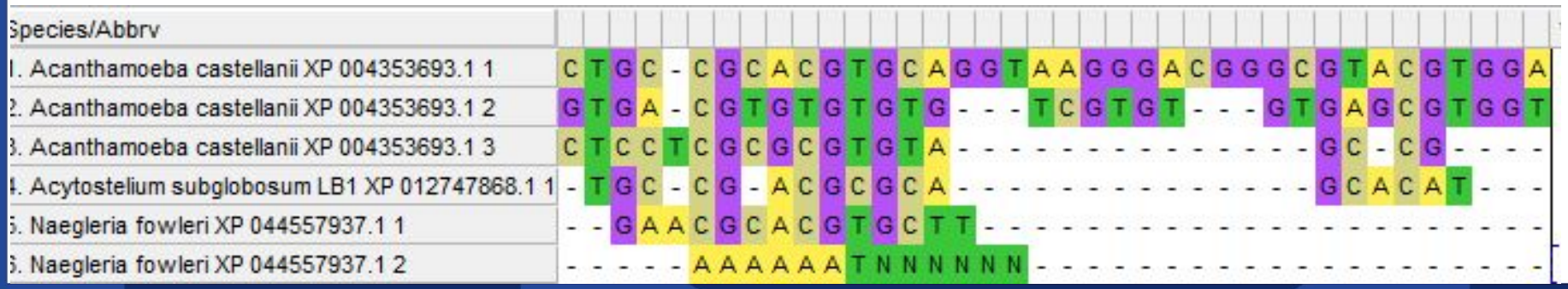


Выравнивание и филогенетическое дерево ортологичных генов

Ортогруппа белков с доменом PHD finger (семейство PHF5, включая PHD finger-like domain-containing protein 5A).

В ряде организмов аннотирована как гипотетический или неохарактеризованный белок.

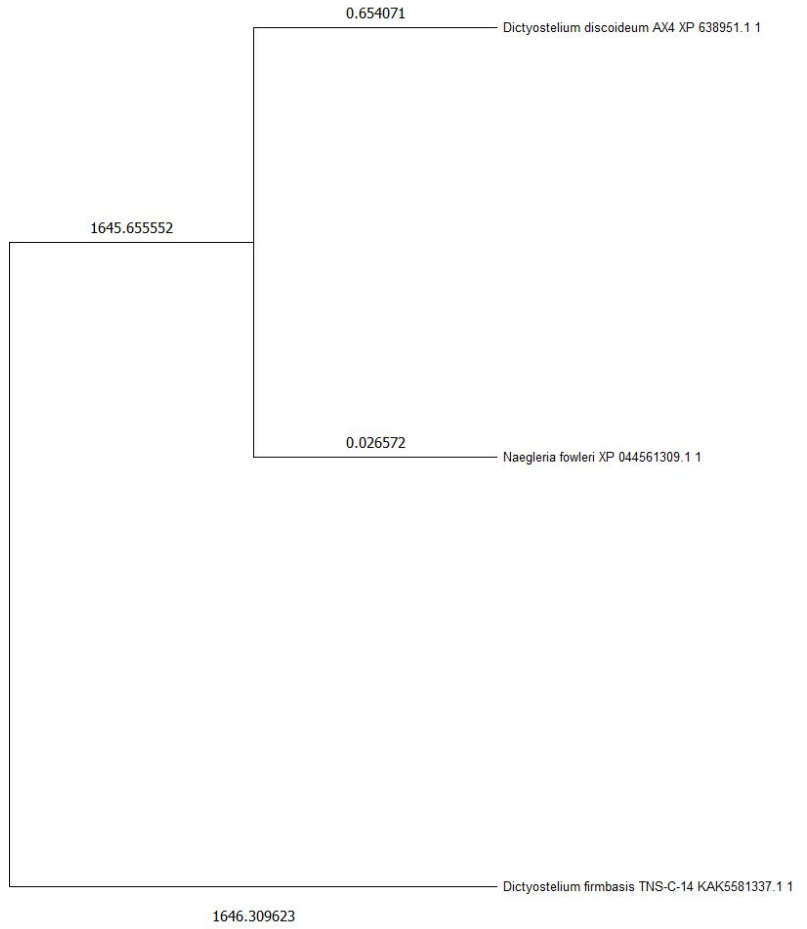


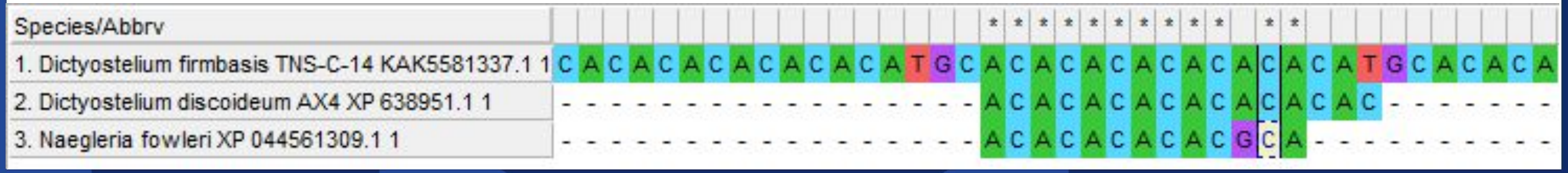


Выравнивание и филогенетическое дерево ортологичных генов

Ортогруппа пептидилпролил цис-транс изомераз (peptidylprolyl cis-trans isomerase, включая циклофилины).

В ряде организмов аннотирована как гипотетический или неохарактеризованный белок.





Код

помогший нам организовать данные между собой

Код для составления таблицы семейство-ген

Этот код берёт из таблицы с генными доменами мыши, по каждому домену смотрит, есть ли в данном протеоме ген, кодирующий белок из этого семейства и записывает самый правдоподобный.

Получается таблица, в которой каждому семейству сопоставляется тэг гена, который кодирует белок этого семейства.

```
import pandas as pd
import os
from os import walk
import requests


domains = set([s.split()[0] for s in data['Pfam domains'].to_list()])

for domain in domains:
    !hmmfetch Pfam-A.hmm.h3m "{domain}" > "hmm/{domain}.hmm"
    with open(f'hmm/{domain}.hmm') as f:
        if f.readlines() == []:
            os.remove(f'hmm/{domain}.hmm')

table = []
files = next(walk("hmm"), (None, None, []))[2]
for file in files:
    output = !hmmsearch --notextw --noali "hmm/{file}" proteome.faa
    if len(output) > 18 and len(output[18].split()) > 8:
        link = f'https://www.ncbi.nlm.nih.gov/gene/?term={output[18].split()[8]}'
        f = requests.get(link)
        t = str(f.content)
        start = t.find("<title>") + 7
        end = t[start:].find("</title>")
        table.append([file[:-4], t[start:start + end].split()[0]])

df = pd.DataFrame(table, columns=["Проверяемое семейство", "Название гена"])
df = df.set_index("Проверяемое семейство")

df.to_csv("hmmer.csv")
```

Код для объединения таблиц с генами

Этот код делает разметку семейств для общей таблицы. Domains берётся из таблицы с семействами, а hmmer.csv – файл с генами от участников.

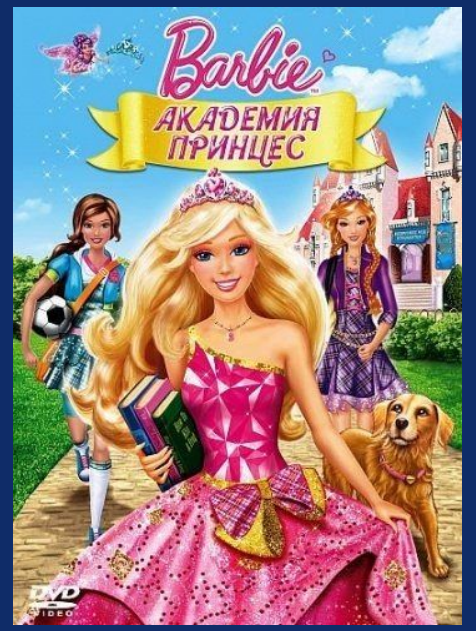
Это делается для того, чтобы в общей таблице семействам сопоставлялись соответствующим генам.

```python
import pandas as pd

table = []
for domain in domains:
    table.append([domain, None])

df2 = pd.read_csv('hmmer.csv', encoding='utf-8', sep=',')
df2.columns = ["Проверяемое семейство", "Название гена"]
df1 = pd.DataFrame(table, columns=["Проверяемое семейство", "Название гена"])

mapping = dict(zip(df2['Проверяемое семейство'], df2['Название гена']))

families_df1 = set(df1['Проверяемое семейство'])
families_df2 = set(df2['Проверяемое семейство'])
common = families_df1 & families_df2
print(f"Всего семейств в df1: {len(families_df1)}")
print(f"Всего семейств в df2: {len(families_df2)}")
print(f"Пересекаются: {len(common)}")

diff = list(families_df1 - families_df2)
print("Примеры семейств из df1, которых нет в df2:", diff[:10])

df1['gene_from_df2'] = df1['Проверяемое семейство'].map(mapping)
df1['Название гена'] = df1['Название гена'].combine_first(df1['gene_from_df2'])
```

Выводы

Судя по таблице распределения вторичных структур и филогенетическим деревьям, можно сделать следующие выводы:

1. По эпигенетике: по семействам, отвечающим за модификацию РНК и стирание гистоновых модификаций, удачно построилось выравнивание, что свидетельствует о наличии этих эпигенетических механизмов у всех организмов
2. По квадруплексам: нашлись преимущественно у трёх организмов: A. Castellani, E. Invadens, A. Subglobusum.
3. По Z-ДНК: нашлись преимущественно у A. Castellani и E. Invadens
4. Это неудивительно, так как GC-содержание у данных организмов высокое



ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •