Поиск гена COI в траснкриптомных сборках

Шаг 1. Выравнивание (blast) транскриптомов на референс -> множественное выравнивание топ5 на референс -> выбор лучшего хита -> тримминг по границам гена (отдельно для сборок rnaspades/trinity)

Заполнение config.py

```
python main_pipeline_step1.py
python main_pipeline_step2.py
python main_pipeline_step3.py
python main_pipeline_step4.py
python main_pipeline_step5.py
```

Если не работает mafft (см. файл для логов)-> использование онлайн выравнивания https://mafft.cbrc.jp/alignment/server/spool с настройками::

```
mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair
```

Шаг 2. Фильтрация полученных последовательностей по длине + выбор лучшего хита вручную (rnaspades/trinity)

```
python main_pipeline_check1.py
```

Рассмотрим обрезанные гены, не прошедшие проверку по длине (зеленые - удалось исправить, желтые - последовательность из другой сборки прошла проверку):

№	Вид	Решение / Комментарий	Результат
rnaspades			
1	*Boeckaxelia_carpenterii*	выбор подходящего хита (без вставки) для обрезки	проверку по длине прошел
2	*Brachyuropus_grewingkii*	н/п	берем trinity
3	*Caprella_sp*	выбор хита с лучшим покрытием для обрезки не помог	не рассматриваем далее, вид не байкальский
4	*Echiuropus_macronychus*	перемещение последнего нуклеотида в best_to_ref (выравнивание лучшего хита с референсом) вручную	проверку по длине прошел
5	*Eulimnogammarus_sp._gam2 quest*	есть гэпы в конце	берем trinity
6	*Gammarus_lacustris*	выбор подходящего хита (без вставки) для обрезки	проверку по длине прошел
7	*Gammarus_minus*	имеет гэпы в начале, выравнивание хита с лучшим покрытием `mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair` не помогло	берем trinity
8	*Gammarus_pisinnus*	1. гэпы в начале и конце	берем trinity
9	*Hyalella_azteca*	2. гэпы в середине	берем trinity
10	*Hirondellea_gigas*	выбор подходящего хита (без вставки) для обрезки	проверку по длине прошел
11	*Hyalellopsis_grisea*	выбор подходящего хита (без вставки) для обрезки	проверку по длине прошел

№	Вид	Решение / Комментарий	Результат
12	*Hyalellopsis_stebbingi*	перемещение первого нуклеотида в best_to_ref (выравнивание лучшего хита с референсом) вручную, гэпы в конце	берем trinity
13	*Macrohectopus_branickii*	выравнивание хита с лучшим покрытием `mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair`	проверку по длине прошел
14	*Macropereiopus_parvus*	выравнивание 2х лучших хитов по отдельности `mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair`, trimmed_best с гэпами	берем trinity
15	*Marinogammarus_marinus*	выравнивание без хита с худшим покрытием `mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair`	проверку по длине прошел
16	*Micruropus_wahlii*	выравнивание на хит с худшим покрытием, но без nnn `mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair` помогло, но ок ли? так как есть различия между ним и остальными в топе	берем trinity
17	*Ommatogammarus_albinus*	выравнивание на топ 20 хитов без nnn	после удаления гэпов проверку по длине прошел
18	*Oxyacanthus_flavus*	выравнивание без хита с худшим покрытием `mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair`	проверку по длине прошел
19	*Pallasea_cancelloides*	гэп ближе к концу	берем trinity

№	Вид	Решение / Комментарий	Результат
20	*Pallasea_cancellus*	выравнивание без хитов с nnn `mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair`	проверку по длине прошел
21	*Parapallasea_borowskii*	выравнивание без хитов с nnn `mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair не помогло`	берем trinity
22	*Parapallasea_wosnessenskii*	выравнивание хита с лучшим покрытием `mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair` не помогло	берем trinity
23	*Parhyale_hawaiensis*	перемещение первого нуклеотида в best_to_ref (выравнивание лучшего хита с референсом) вручную	проверку по длине прошел
trinity			
1	*Baikalogammarus_pullus*	выбор подходящего по длине хита	проверку по длине прошел
2	*Eulimnogammarus_violaceus*	гэпы в начале	берем rnaspades
3	*Gammarus_lacustris*	всего 2 хита, второй лучше по длине	проверку по длине прошел
4	*Gammarus_pisinnus*	выбор подходящего по длине хита	проверку по длине прошел
5	*Gmelinoides_fasciatus*	выбор подходящего по длине хита	проверку по длине прошел
6	*Heterogammarus_sophianosii*	выбор подходящего по длине хита	проверку по длине прошел
7	*Macrohectopus_branickii*	гэпы в начале	берем rnaspades

№	Вид	Решение / Комментарий	Результат
8	*Micruropus_glaber*	выбор подходящего по длине хита	проверку по длине прошел
9	*Oxyacanthus_flavus*	выбор подходящего по длине хита	проверку по длине прошел
10	*Palicarinus_puzyllii*	делеция?	берем rnaspades
11	*Pallasea_cancelloides_2*	выбор подходящего по длине хита	проверку по длине прошел
12	*Parapallasea_borowskii*	выравнивание на подходящий по длине хит	проверку по длине прошел

Шаг 3. Проверка полученных последовательностей в blast (rnaspades/trinity)

```
python main_pipeline_check2.py
```

Обсудим лучшие хиты по e-value, вызывающие вопросы:

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Comment
Eulimnogammarus_cyaneus_rnaspades	NODE_25_length_11911_cov_9730.845389_g4_i1	NC_033360.1	Eulimnogammarus cyaneus mitochondrion	0.0	98%	1534	Внутривидовое разнообразие
Eulimnogammarus_vittatus_rnaspades	NODE_34_length_9586_cov_5778.005243_g10_i1	NC_025564.1	Eulimnogammarus vittatus mitochondrion	0.0	91%	1532	Внутривидовое разнообразие
Gammarus_lacustris_rnaspades	_R_NODE_80_length_6634_cov_2707.503113_g4	NC_044469.1	Gammarus lacustris mitochondrion,	0.0	98%	1534	indel?

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Comment
	8_i1		complete genome				
Hirondellea_gigas_rnaspades	NODE_1517_length_4709_cov_11508.664508_g833_i1	KU558990.1	Hirondellea gigas clone mito1 mitochondr	0.0	98%	1534	Не байкальский вид, не интересует
Eulimnogammarus_cyaneus_T285	TRINITY_DN1663_c0_g1_i1	NC_033360.1	Eulimnogammarus cyaneus mitochondrion	0.0	98%	1534	Внутривидовое разнообразие
Eulimnogammarus_vittatus_T285	TRINITY_DN7079_c0_g1_i1	NC_025564.1	Eulimnogammarus vittatus mitochondrion	0.0	91%	1532	Внутривидовое разнообразие
Gammarus_lacustris_T285	_R_TRINITY_DN2922_c0_g1_i1	NC_044469.1	Gammarus lacustris mitochondrion, complete genome	0.0	98%	1533	Внутривидовое разнообразие

Шаг 4. Промежуточное сравнение последовательностей из разных сборок

```
python main_pipeline_check3.py
```

Сравнение результатов сборок (пока без 12 желтых для rnaspades). Рассмотрим результаты выравнивания для видов с низкой идентичностью между сборками :

№	Sequence ID	Length Seq rnaspades	Length Seq trinity	Identity Percentage	Alignment Length	Comment

1	*Boeckaxelia_carpenterii*	1534	1461	95.241199	1534	Для trinity remafft без наиболее подходящей по длине последовательности, trimmed_best с гэпами в начале - берем последовательность из сборки rnaspades
2	*Cornugammarus_maximus*	1534	1535	83.061889	1535	Для trinity в топ5 2 пула похожих последовательностей, remafft на другой пул помог
3	*Eulimnogammarus_violaceus*	1534	1523	82.398957	1536	Для trinity в топ5 есть пул последовательностей, remafft на него помог, но большой кусок с гэпами в конце - берем последовательность из сборки rnaspades
4	*Garjajewia_dershawini*	1537	1534	97.722837	1554	У trinity всего 1 последовательность в топе - берем последовательность из сборки rnaspades
5	*Hyalellopsis_costata*	1534	1537	66.428107	1609	Для trinity в топ5 есть разные последовательности, remaffты не помогли - берем последовательность из сборки rnaspades
6	*Hyalellopsis_setosa*	1538	1534	75.357607	1559	Для trinity remafft без изначально выбранной последовательности
7	*Macrohectopus_branickii*	1534	1520	74.576271	1554	Для trinity сильно отличаются все последовательности в топ5 - берем последовательность из сборки rnaspades
8	*Palicarinus_puzyllii*	1534	1534	78.09648	1544	У trinity всего 1 последовательность в топе - берем последовательность из сборки rnaspades

Шаг 5. Финальное сравнение последовательностей из разных сборок

```
python main_pipeline_check3.py
```

Запустим сравнение теперь для всех последовательностей. Итого 69 видов имеют обе сборки, из них:

- 26 видов имеют на 100% одинаковые последовательности в разных сборках
- 26 видов имеют идентичность > 99%, < 100% между сборками

- 17 видов имеют низкую идентичность между сборками

Из 52 видов с идентичностью > 99% только 38 имеют одинаковую длину последовательностей и выравнивания. Для 14 остальных выберем финальную сборку:

№	Sequence_ID	Len_Seq_rnaspades	Len_Seq_trinity	Identity	Alignment_Len	Justification	Final_seq
1	*Acanthogammarus_godlewskii*	1535	1534	99.93	1535	indel в 1141 позиции у rnaspades	trinity
2	*Baikalogammarus_pullus*	1535	1534	99.93	1535	indel в 1266 позиции у rnaspades	trinity
3	*Eulimnogammarus_messerschmidtii*	1534	1534	99.35	1540	trinity лучше выровнялась на проверке (шаг 6)	trinity
4	*Eulimnogammarus_similis*	1535	1535	99.87	1536	у trinity indel в 434 позиции, у rnaspades лишний нуклеотид на конце - удалим его	rnaspades
5	*Eulimnogammarus_verrucosus*	1534	1536	99.87	1536	indel в 434 позиции у trinity	rnaspades
6	*Gammarus_lacustris*	1535	1534	99.87	1535	indel в 191 позиции у rnaspades	trinity
7	*Hyalellopsis_setosa*	1538	1534	99.74	1538	indels в 191-191, 786-787 позициях у rnaspades	trinity
8	*Macropereiopus_wagneri*	1535	1535	99.74	1536	indel в 434 позиции у rnaspades, в 1266 у trinity	trinity
9	*Ommatogammarus_flavus*	1535	1534	99.93	1535	indel в 81 позиции у rnaspades	trinity
10	*Oxyacanthus_sowinskii*	1534	1534	99.48	1536	у trinity выровнялся более длинный участок на проверке (шаг 6)	trinity

11	*Pallasea_cancelloides_2*	1534	1535	99.93	1535	indel в 1330 позиции у trinity (согласно Pallasea_cancelloides)	rnaspades
12	*Pandorites_podoceroides*	1535	1534	99.87	1535	trinity лучше выровнялась на проверке (шаг 6)	trinity
13	*Pentagonurus_dawydowi*	1536	1534	99.87	1536	trinity лучше выровнялась на проверке (шаг 6)	trinity
14	*Sluginella_kietlinskii*	1535	1534	99.93	1535	indel в 191 позиции у rnaspades	trinity

Выбор сборки для случаев выравнивания с низкой идентичностью (сравнение результатов mafft топ5 хитов):

№	Sequence_ID	Len_Seq_rnaspades	Len_Seq_trinity	Identity	Alignment_Len	Len_Seq_wo_gaps_rnaspades	Len_Seq_wo_gaps_trinity	Identity_wo_gaps	Final_seq
1	*Boeckaxelia_carpenterii*	1534	1534	95.24	1534	1534	1461	100.0	rnaspades
2	*Brachyuropus_grewingkii*	1680	1534	91.31	1680	1680	1534	n/a	trinity
3	*Eulimnogammarus_sp._gam2quest*	1534	1534	97.46	1536	1502	1534	99.53	trinity
4	*Eulimnogammarus_violaceus*	1534	1534	65.19	1534	1534	1001	99.9	rnaspades
5	*Gammarus_minus*	1534	1534	93.68	1534	1437	1534	100.0	trinity
6	*Gammarus_pisinnus*	1538	1534	70.29	1642	1496	1534	72.26	trinity
7	*Garjajewia_dershawini*	1537	1534	97.72	1554	1537	1534	n/a	rnaspades
8	*Hyalellopsis_costata*	1534	1537	66.56	1609	1534	1537	n/a	rnaspades

9	*Hyalellopsis_stebbingi*	1535	1534	98.5	1535	1514	1534	99.87	trinity
10	*Macrohectopus_branickii*	1534	1534	74.58	1554	1534	1520	75.39	rnaspades
11	*Macropereiopus_parvus*	1536	1534	96.55	1565	1536	1534	n/a	trinity
12	*Micruropus_wahlii*	1534	1534	79.66	1536	1534	1534	n/a	trinity
13	*Ommatogammarus_albinus*	1534	1534	83.57	1539	1534	1534	n/a	rnaspades
14	*Palicarinus_puzyllii*	1534	1534	78.1	1544	1534	1534	n/a	rnaspades
15	*Pallasea_cancelloides*	1541	1534	95.46	1587	1541	1534	n/a	trinity
16	*Parapallasea_borowskii*	3036	1534	50.26	3039	3036	1534	n/a	trinity
17	*Parapallasea_wosnessenskii*	1549	1534	94.06	1563	1520	1534	95.86	trinity

Шаг 6. Финальная проверка последовательностей в blast

```
python main_pipeline_check2.py
```

Финальная проверка в blast (2 лучших хита: по e-value и по идентичности). Ниже приведены последовательности, вызывающие вопросы. Дополнительно запустили для них main_pipeline с топ20 - не помогло или стало хуже.

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Вопрос	Решение
rnaspades								
Acanthogammarus_godlewskii	NODE_41_length_9587_cov_15664.560390_g2_i2	JN393755.1	Acanthogammarus cf. maculosus MED-2011 h	0.0e+00	99.66 %	585	высокая идентичность для короткой последовательности соседнего вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Brandtia_latissima	NODE_6_length_14004_cov_4175.113149_g1_i2	FJ756302.1	Brandtia latissima lata voucher BLATA66	0.0e+00	95.85 %	626	низкая идентичность для короткой последовательности того же вида	внутривидовое разнообразие
Eulimnogammarus_sp._gam2quest	NODE_2643_length_4669_cov_13692.467316_g13_i13	AY926663.1	Eulimnogammarus maacki cytochrome oxidas	0.0e+00	99%	450	можем ли утверждать, что это вид maacki, берем trinity	найденный ген 18S в сборке выравнивается на E. maackii с высокой идентичностью
Eulimnogammarus_viridulus	_R_NODE_21_length_14373_cov_11576.668389_g1_i2	MK887742.1	Eulimnogammarus vittatus voucher Evi_Lis	0.0e+00	99.84 %	640	высокая идентичность для короткой последовательности соседнего вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Вопрос	Решение
Micruropus_glaber	NODE_299_length_7239_cov_6437.717524_g7_i3	AY926682.1	Micruropus glaber cytochrome oxidase sub	0.0e+00	91.79 %	694	низкая идентичность для короткой последовательности того же вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Micruropus_wahlii	NODE_22247_length_1635_cov_1417.979219_g11519_i0	FJ756340.1	Micruropus wahlii voucher MWAH2 cytochro	0.0e+00	90.26 %	626	низкая идентичность для короткой последовательности того же вида	берем trinity
Odontogammarus_calcaratus	_R_NODE_235_length_8168_cov_13888.785442_g5_i10	FJ756341.1	Odontogammarus calcaratus voucher OCAL20	0.0e+00	92.82 %	627	низкая идентичность для короткой последовательности того же вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Oxyacanthus_curtus	_R_NODE_17_length_17753_cov_5185.770843_g0_i10	JN393847.1	Oxyacanthus flavus cytochrome oxidase su	0.0e+00	99.66 %	585	высокая идентичность для короткой последовательности соседнего вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
*Pachyschesis_*	NODE_1_len	MN14835	Pachyschesis	0.0e+00	98.44	705	низкая	последовательност

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Вопрос	Решение
branchialis_2	gth_14059_cov_26894.746966_g0_i0	9.1	branchialis isolate 35-6 cy		%		идентичность для короткой последовательности того же вида	ь не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Pachyschesis_branchialis	NODE_4411_length_3344_cov_10806.215175_g4_i14	MN148359.1	Pachyschesis branchialis isolate 35-6 cy	0.0e+00	87.68 %	706	низкая идентичность для короткой последовательности того же вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Pallasea_grubei	NODE_1153_length_5420_cov_10579.713461_g494_i1	AY926688.1	Pallasea grubei cytochrome oxidase subun	0.0e+00	92.17 %	498	низкая идентичность для короткой последовательности того же вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Pallaseopsis_kessleri	_R_NODE_14_length_19711_cov_8846.089971_g0_i8	GQ919202.1	Babr nigromaculatus isolate A60 cytochro	0.0e+00	98.56 %	627	высокая идентичность для короткой последовательности соседнего вида, на сам ген из NCBI	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Вопрос	Решение
							выравнился очень плохо - 78%	
trinity								
Acanthogammarus_godlewskii	_R_TRINITY_DN65_c0_g2_i1	JN393755.1	Acanthogammarus cf. maculosus MED-2011 h	0.0e+00	99.66 %	585	высокая идентичность для короткой последовательности соседнего вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Brandtia_latissima	_R_TRINITY_DN94_c0_g2_i4	FJ756302.1	Brandtia latissima lata voucher BLATA66	0.0e+00	95.85 %	626	низкая идентичность для короткой последовательности того же вида	внутривидовое разнообразие
Eulimnogammarus_sp._gam2quest	_R_TRINITY_DN531_c0_g1_i4	AY926663.1	Eulimnogammarus maacki cytochrome oxidas	0.0e+00	99%	450	можем ли утверждать, что это вид maacki	найденный ген 18S в сборке выравнивается на E. maackii с высокой идентичностью
Eulimnogammarus_viridulus	TRINITY_DN93_c0_g3_i2	MK887742.1	Eulimnogammarus vittatus voucher Evi_Lis	0.0e+00	99.84 %	640	высокая идентичность для короткой последовательности соседнего вида	последовательность не включаем в базу данных, откладываем в папку

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Вопрос	Решение
								“no_pass_blast_check”
Micruropus_glaber	_R_TRINITY_DN7569_c0_g1_i1	AY926682.1	Micruropus glaber cytochrome oxidase sub	0.0e+00	91.79 %	694	низкая идентичность для короткой последовательности того же вида (но для 18s все хорошо)	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Odontogammarus_calcaratus	_R_TRINITY_DN1782_c0_g1_i4	FJ756341.1	Odontogammarus calcaratus voucher OCAL20	0.0e+00	92.82 %	627	низкая идентичность для короткой последовательности того же вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Oxyacanthus_curtus	TRINITY_DN1_c0_g1_i4	JN393847.1	Oxyacanthus flavus cytochrome oxidase su	0.0e+00	99.66 %	585	высокая идентичность для короткой последовательности соседнего вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Pachyschesis_branchialis_2	TRINITY_DN40_c0_g1_i4	MN148359.1	Pachyschesis branchialis isolate 35-6 cy	0.0e+00	98.44 %	705	низкая идентичность для короткой последовательност	последовательность не включаем в базу данных, откладываем в

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Вопрос	Решение
							и того же вида (но для 18s все хорошо)	папку “no_pass_blast_check”
Pachyschesis_branchialis	_R_TRINITY_DN387_c0_g1_i3	MN148359.1	Pachyschesis branchialis isolate 35-6 cy	0.0e+00	87.68 %	706	низкая идентичность для короткой последовательности того же вида (но для 18s все хорошо)	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Pallasea_grubei	TRINITY_DN56_c0_g1_i1	AY926688.1	Pallasea grubei cytochrome oxidase subun	0.0e+00	92.17 %	498	низкая идентичность для короткой последовательности того же вида	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”
Pallaseopsis_kessleri	TRINITY_DN670_c0_g1_i5	GQ919202.1	Babr nigromaculatus isolate A60 cytochro	0.0e+00	98.56 %	627	высокая идентичность для короткой последовательности соседнего вида, на сам ген из NCBI выравнился плохо - 78%	последовательность не включаем в базу данных, откладываем в папку “no_pass_blast_check”

Поиск гена 18S в траснкриптомных сборках

Шаг 1. Выравнивание (blast) транскриптомов на референс -> множественное выравнивание топ5 на референс -> выбор лучшего хита -> тримминг по границам гена (rnaspades/trinity)

Заполнение config.py

```
python main_pipeline_step1.py
python main_pipeline_step2.py
python main_pipeline_step3.py
python main_pipeline_step4.py
python main_pipeline_step5.py
```

Шаг 2. Фильтрация полученных последовательностей по длине + выбор лучшего хита (rnaspades/trinity)

```
python main_pipeline_check1.py
```

Все последовательности содержат гэпы, рассмотрим результаты mafft для тех, что имеют большое количество гэпов или длину >> 2267 (зеленые - удалось исправить, желтые - последовательность из другой сборки прошла проверку, кроме Pachyschesis_branchialis_2):

#	Species	Seq_ID	Seq_len	Cleaned_seq_len (without gaps)	Without 'NNN'	Decision/comment	Result
rnaspades							
1	*Baikalogammarus_pullus*	NODE_159_length_5935_cov_26498.83	2347	2333	True	Замена в best_to_ref на хит без вставки	проверку по длине прошел

		4523_g1_i4					
2	*Caprella_sp*	NODE_8049_length_2019_cov_1517.794027_g4408_i0	3508	2019	True	Не байкальский вид	Не интересует
3	*Echinogammarus_berilloni*	_R_NODE_675_length_9538_cov_2297.755945_g503_i0	2389	2377	True	Не байкальский вид	Не интересует
4	*Eogammarus_possjeticus*	_R_NODE_272_length_9513_cov_2210.622192_g140_i0	2338	2288	True	Не байкальский вид	Не интересует
5	*Eulimnogammarus_testaceus*	NODE_3500_length_3924_cov_14024.027097_g747_i10	2380	2371	True	топ 20	проверку по длине прошел
6	*Eulimnogammarus_ussolzewii*	NODE_1759_length_4552_cov_29029.299134_g1279_i0	5365	4552	True	топ20 - не помогло	trinity
7	*Eulimnogammarus_viridulus*	_R_NODE_1211_length_5584_cov_34217.632159_g42_i9	2277	2214	True	топ20	проверку по длине прошел
8	*Gammarus_chevreuxi*	NODE_479_length_9075_cov_495.103922_g118_i0	2291	2275	True	Не байкальский вид	Не интересует
9	*Gammarus_lacustris*	_R_NODE_16_length_8577_cov_613.017237_g9_i0	2306	2287	True	топ20 - не помогло, мало контигов, низкие покрытия	trinity

10	*Gammarus_pisinnus*	_R_NODE_522703_length_4619_cov_26.735526	4692	4619	True	Не байкальский вид	Не интересует
11	*Hirondellea_gigas*	_R_NODE_811_length_5757_cov_5491.491197_g338_i1	2556	2534	True	Не байкальский вид	Не интересует
12	*Hyalella_azteca*	NODE_26_length_10142_cov_641.959327_g11_i1	2441	2370	True	Не байкальский вид	Не интересует
13	*Hyalellopsis_setosa*	_R_NODE_3972_length_2187_cov_3188.257250_g547_i3	2267	2187	True	топ20 - не помогло	trinity
14	*Linevichella_vortex*	_R_NODE_14_length_11798_cov_4941.989957_g9_i0	2270	2249	True	топ20 - не помогло	trinity
15	*Macropereiopus_parvus*	_R_NODE_9621_length_2238_cov_9420.869804_g15_i26	2268	2238	True	топ20 - не помогло	trinity
16	*Marinogammarus_marinus*	_R_NODE_429_length_10268_cov_2586.597979_g246_i1	2379	2361	True	Не байкальский вид	Не интересует
17	*Ommatogammarus_albinus*	NODE_8224_length_3017_cov_15755.705526_g5424_i0	2364	2292	True	топ20	проверку по длине прошел

18	*Pachyschesis_branchialis_2*	_R_NODE_23_length_6446_cov_20621.238706_g3_i3	2294	2292	True	remafft без лучшего контига (немного отличается от остальных последовательностей) - много гэпов, топ 20 - не помогло	trinity
19	*Palicarinus_puzyllii*	NODE_2889_length_4280_cov_11925.144174_g830_i5	2277	2215	True	топ20	проверку по длине прошел
20	*Pallasea_cancelloides*	_R_NODE_1545_length_2819_cov_6883.720217_g1047_i0	2267	2211	True	топ 20 - не помогло	trinity
21	*Pallasea_grubei*	NODE_12_length_15903_cov_8094.016021_g7_i1	2319	2257	True	топ20	проверку по длине прошел
22	*Pandorites_podoceroides*	NODE_4_length_9829_cov_950.867280_g2_i0	2273	2224	True	Не байкальский вид	Не интересует
23	*Parhyale_hawaiensis*	NODE_54_length_10241_cov_1174.547913_g31_i0	2533	2490	True	Не байкальский вид	Не интересует
24	*Sluginella_kietlinskii*	NODE_309_length_3524_cov_3241.044291_g220_i0	2268	2242	True	топ 20 - не помогло	trinity
25	*Talitrus_saltator*	NODE_36_length_9591_cov_379.38189	2368	2290	True	Не байкальский вид	Не интересует

		1_g28_i0					
trinity							
1	*Asprogammarus_rhodophthalmus*	_R_TRINITY_DN1125_c0_g1_i2	2270	2085	True	топ20 - не помогло	rnaspades
2	*Dorogostaiskia_parasitica*	_R_TRINITY_DN1430_c0_g2_i6	2267	2245	True	только 2 последовательности длины > 2000	rnaspades
3	*Eulimnogammarus_cruentus_2*	TRINITY_DN1113_c0_g1_i6	3536	3533	True	в best_to_ref перенос последнего нуклеотида вручную	проверку по длине прошел
4	*Eulimnogammarus_testaceus*	_R_TRINITY_DN138_c0_g1_i15	2321	2318	True	топ20 - не помогло	rnaspades
5	*Eulimnogammarus_vittatus*	TRINITY_DN10693_c0_g2_i3	2304	2299	True	только 1 контиг длинный	rnaspades
6	*Hyalella_azteca*	_R_TRINITY_DN5596_c0_g1_i1	3449	3370	True	Не байкальский, не нужен	
7	*Hyalellopsis_costata*	_R_TRINITY_DN11_c0_g1_i11	2267	1686	True	топ20 - не помогло	rnaspades
8	*Hyalellopsis_grisea*	_R_TRINITY_DN11_c0_g1_i6	2267	1644	True	топ20 - не помогло	rnaspades
9	*Hyalellopsis_stebbingi*	_R_TRINITY_DN70_c0_g1_i24	2269	1636	True	топ20 - не помогло	rnaspades
10	*Micruropus_parv*	_R_TRINITY_DN14	2270	1980	True	топ20 - не помогло	rnaspades

	ulus	_c0_g2_i3					
11	*Pachyschesis_branchialis_2*	_R_TRINITY_DN12_c0_g1_i5	2294	2292	True	топ20 - не помогло	rnaspades
12	*Pandorites_podoceroides*	_R_TRINITY_DN1859_c0_g1_i11	2273	2224	True	Не байкальский, не нужен	

Шаг 3. Проверка последовательностей в blast (rnaspades/trinity)

```
python main_pipeline_check2.py
```

Большинство последовательностей выровнялись с максимальной идентичностью/e-value на референсную последовательность - проверка малоинформативна.

Шаг 4. Сравнение последовательностей из разных сборок

```
python main_pipeline_check3.py
```

Итого 65 байкальских видов имеют обе сборки, из них:

- 14 видов имеют на 100.000% одинаковые последовательности в разных сборках
- 37 видов имеют идентичность > 99%, < 100% между сборками
- 14 видов имеют низкую идентичность между сборками

Из 51 вида с идентичностью > 99% только 36 имеют одинаковую длину последовательностей и выравнивания (+ не имеют гэпов). Для 15 остальных выберем финальную сборку:

№	Sequence_ID	Len_Seq_rnaspades	Len_Seq_trinity	Identity	Alignment_Len	Len_Seq_wo_gaps_rnaspades	Len_Seq_wo_gaps_trinity	Identity_wo_gaps	Justification	Final_seq

1	*Baikalogammarus_pullus*	2272	2270.0	99.428	2274.0	2272	2270.0	n/a	выше идентичность в бласте	trinity
2	*Boeckaxelia_carpenterii*	2270	2271.0	99.119	2271.0	2259	2271.0	99.65	нет гэпов в начале	trinity
3	*Brandtia_latissima*	2268	2268.0	99.471	2268.0	2257	2268.0	99.96	нет гэпов в начале	trinity
4	*Cornugammarus_maximus*	2268	2268.0	99.515	2268.0	2257	2268.0	100.0	нет гэпов в начале	trinity
5	*Echiuropus_macronychus*	2269	2268.0	99.956	2269.0	2269	2268.0	n/a	выше идентичность в бласте	trinity
6	*Eucarinogammarus_wagii*	2268	2272.0	99.604	2272.0	2268	2272.0	n/a	выше идентичность с референсом	rnaspades
7	*Eulimnogammarus_czerskii*	2267	2268.0	99.162	2268.0	2249	2268.0	100.0	нет гэпов в конце	trinity
8	*Eulimnogammarus_sp._gam16.4*	2268	2268.0	99.868	2270.0	2268	2268.0	n/a	выше идентичность с референсом	rnaspades
9	*Hyalellopsis_carinata*	2268	2268.0	99.735	2270.0	2268	2268.0	n/a	выше идентичность с референсом	rnaspades
10	*Linevichella_vortex*	2270	2271.0	99.648	2271.0	2270	2271.0	n/a	выше идентичность в бласте	rnaspades
11	*Macrohectopus_branickii*	2269	2270.0	99.031	2278.0	2269	2270.0	n/a	выше длина выравнивания на реф	rnaspades
12	*Odontogammarus_calcaratus*	2268	2268.0	99.868	2270.0	2268	2268.0	n/a	выше идентичность с референсом	trinity
13	*Ommatogammarus_albinus*	2268	2268.0	99.515	2268.0	2257	2268.0	100.0	нет гэпов в начале	trinity

14	*Pallasea_sp._gam7.3*	2268	2268.0	99.427	2268.0	2268	2257.0	99.91	нет гэпов в начале	rnaspades
15	*Pentagonurus_dawydowi*	2268	2268.0	99.912	2269.0	2268	2268.0	n/a	cleaned одинаковые	rnaspades/trinity

Выбираем, какую сборку будем брать для случаев выравнивания с низкой идентичностью:

№	Sequence_ID	Len_Seq_rnaspades	Len_Seq_trinity	Identity	Alignment_Len	Len_Seq_wo_gaps_rnaspades	Len_Seq_wo_gaps_trinity	Identity_wo_gaps	Justification	Final_seq
1	*Asprogammarus_rhodophthalmus*	2270	2270.0	87.445	2387.0	2270	2210.0	89.82	нет гэпов в начале	rnaspades
2	*Dorogostaiskia_parasitica*	2268	2267.0	98.589	2281.0	2257	2267.0	99.07	чуть длиннее, но вместе - целый ген	rnaspades
3	*Eulimnogammarus_testaceus*	2285	2321.0	94.313	2355.0	2285	2321.0	n/a	по длине	rnaspades
4	*Eulimnogammarus_ussolzewii*	5365	2268.0	33.439	5414.0	4579	2268.0	79.1	по длине	trinity
5	*Eulimnogammarus_vittatus*	2268	2304.0	98.264	2304.0	2268	2304.0	n/a	по длине	rnaspades
6	*Gammarus_lacustris*	2306	2269.0	98.395	2306.0	2306	2269.0	n/a	по длине	trinity
7	*Hyalellopsis_costata*	2268	2267.0	73.545	2695.0	2268	1691.0	98.64	нет гэпов	rnaspades

8	*Hyalellopsis_grisea*	2268	2267.0	71.825	2336.0	2268	1648.0	99.03	нет гэпов	rnaspades
9	*Hyalellopsis_setosa*	2267	2268.0	96.649	2268.0	2192	2268.0	100.0	нет гэпов	trinity
10	*Hyalellopsis_stebbingi*	2268	2269.0	71.177	2309.0	2268	1638.0	98.29	нет гэпов	rnaspades
11	*Macropereiopus_parvus*	2268	2268.0	98.589	2268.0	2241	2268.0	99.78	нет гэпов	trinity
12	*Micruropus_parvulus*	2271	2338.0	73.524	2490.0	2271	2338.0	n/a	по длине	rnaspades
13	*Pallasea_cancelloides*	2267	2268.0	97.531	2268.0	2214	2268.0	99.91	нет гэпов	trinity
14	*Sluginella_kietlinskii*	2268	2268.0	98.765	2277.0	2255	2268.0	99.33	нет гэпов	trinity

Добавление последовательностей COI и 18S в базу данных

Из всех последовательностей 18S были исключены “-”.
В файле seqs_to_database.csv сведена информация по добавленным из транскриптомных сборок последовательностям. Также для COI отмечены последовательности, которые не прошли проверку в NCBI, они собраны в отдельном файле final_seqs/COI_no_pass_NCBI_check.fasta.

Создание бд для приложения:

```
makeblastdb -in database_specoident999.fa -parse_seqids -blastdb_version 5 -taxid_map table_lang_and_latit.csv
-title "bestdb" -dbtype nucl -out specoident99.blastdb
```