

## Поиск гена COI в транскриптомных 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)

```
python scripts/main_pipeline.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 scripts/check_len_script.py
```

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

№	Вид	Решение / Комментарий	Результат
rnaspades			

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

№	Вид	Решение / Комментарий	Результат
13	<i>Macrohectopus_branickii</i>	выравнивание хита с лучшим покрытием mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair	проверку по длине прошел
14	<i>Macropereiopus_parvus</i>	выравнивание 2х лучших хитов по отдельности mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair, trimmed_best с гэпами	берем trinity
15	<i>Marinogammarus_marinus</i>	выравнивание без хита с худшим покрытием mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair	проверку по длине прошел
16	<i>Micruropus_wahlii</i>	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	<i>Ommatogammarus_albinus</i>	выравнивание на топ 20 хитов без nnn	после удаления гэпов проверку по длине прошел
18	<i>Oxyacanthus_flavus</i>	выравнивание без хита с худшим покрытием mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair	проверку по длине прошел
19	<i>Pallasea_cancelloides</i>	гэп ближе к концу	берем trinity
20	<i>Pallasea_cancellus</i>	выравнивание без хитов с nnn mafft --reorder --adjustdirection --maxiterate 2 --retree 1 --globalpair	проверку по длине прошел

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

№	Вид	Решение / Комментарий	Результат
10	<i>Palicarinus_puzyllii</i>	делеция?	берем rnaspades
11	<i>Pallasea_cancelloides_2</i>	выбор подходящего по длине хита	проверку по длине прошел
12	<i>Parapallasea_borowskii</i>	выравнивание на подходящий по длине хит	проверку по длине прошел

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

```
python scripts/check_blast_both_script.py
```

Обсудим лучшие хиты по e-value, вызывающие вопросы:

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Comment
<i>Eulimnogammarus_cyaneus_rnaspades</i>	NODE_25_length_11911_cov_9730.845389_g4_i1	NC_033360.1	Eulimnogammarus cyaneus mitochondrion	0.0	98%	1534	Внутривидовое разнообразие
<i>Eulimnogammarus_vittatus_rnaspades</i>	NODE_34_length_9586_cov_5778.005243_g10_i1	NC_025564.1	Eulimnogammarus vittatus mitochondrion	0.0	91%	1532	Внутривидовое разнообразие
<i>Gammarus_lacustris_rnaspades</i>	_R_NODE_80_length_6634_cov_2707.503113_g48_i1	NC_044469.1	Gammarus lacustris mitochondrion, complete genome	0.0	98%	1534	indel?

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Comment
<i>Hirondellea_gigas_rnaspades</i>	NODE_1517_length_4709_cov_11508.664508_g833_i1	KU558990.1	Hirondellea gigas clone mito1 mitochondrion	0.0	98%	1534	Не байкальский вид, не интересует
<i>Eulimnogammarus_cyaneus_T285</i>	TRINITY_DN1663_c0_g1_i1	NC_033360.1	Eulimnogammarus cyaneus mitochondrion	0.0	98%	1534	Внутривидовое разнообразие
<i>Eulimnogammarus_vittatus_T285</i>	TRINITY_DN7079_c0_g1_i1	NC_025564.1	Eulimnogammarus vittatus mitochondrion	0.0	91%	1532	Внутривидовое разнообразие
<i>Gammarus_lacustris_T285</i>	_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 scripts/comparison_script.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	<i>Boeckxelia_carpenterii</i>	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	<i>Cornugammarus_maximus</i>	1534	1535	83.061889	1535	Для trinity в топ5 2 пула похожих последовательностей, remafft на другой пул помог
3	<i>Eulimnogammarus_violaceus</i>	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 maspades
4	<i>Garjajewia_dershawini</i>	1537	1534	97.722837	1554	У trinity всего 1 последовательность в топе - берем последовательность из сборки maspades
5	<i>Hyalelloopsis_costata</i>	1534	1537	66.428107	1609	Для trinity в топ5 есть разные последовательности, remafftы не помогли - берем последовательность из сборки maspades
6	<i>Hyalelloopsis_setosa</i>	1538	1534	75.357607	1559	Для trinity remafft без изначально выбранной последовательности
7	<i>Macrohectopus_branickii</i>	1534	1520	74.576271	1554	Для trinity сильно отличаются все последовательности в топ5 - берем последовательность из сборки maspades
8	<i>Palicarinus_puzyllii</i>	1534	1534	78.09648	1544	У trinity всего 1 последовательность в топе - берем последовательность из сборки maspades

## Шаг 5. Финальное сравнение последовательностей из разных сборок

```
python scripts/comparison_script.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	<i>Acanthogammarus_godlewskii</i>	1535	1534	99.93	1535	indel в 1141 позиции у rnaspades	trinity
2	<i>Baikalogammarus_pullus</i>	1535	1534	99.93	1535	indel в 1266 позиции у rnaspades	trinity
3	<i>Eulimnogammarus_messerschmidtii</i>	1534	1534	99.35	1540	trinity лучше выровнялась на проверке (шаг 6)	trinity
4	<i>Eulimnogammarus_similis</i>	1535	1535	99.87	1536	у trinity indel в 434 позиции, у rnaspades лишний нуклеотид на конце - удалим его	rnaspades
5	<i>Eulimnogammarus_verrucosus</i>	1534	1536	99.87	1536	indel в 434 позиции у trinity	rnaspades
6	<i>Gammarus_lacustris</i>	1535	1534	99.87	1535	indel в 191 позиции у rnaspades	trinity
7	<i>Hyalelloopsis_setosa</i>	1538	1534	99.74	1538	indels в 191-191, 786-787 позициях у rnaspades	trinity
8	<i>Macropereiopus_wagneri</i>	1535	1535	99.74	1536	indel в 434 позиции у rnaspades, в 1266 у trinity	trinity
9	<i>Ommatogammarus_flavus</i>	1535	1534	99.93	1535	indel в 81 позиции у rnaspades	trinity
10	<i>Oxyacanthus_sowinskii</i>	1534	1534	99.48	1536	у trinity выровнялся более длинный участок на проверке (шаг 6)	trinity
11	<i>Pallasea_cancelloides_2</i>	1534	1535	99.93	1535	indel в 1330 позиции у trinity (согласно <i>Pallasea_cancelloides</i> )	rnaspades
12	<i>Pandorites_podoceroideis</i>	1535	1534	99.87	1535	trinity лучше выровнялась на проверке (шаг 6)	trinity



13	<i>Pentagonurus_dawydowi</i>	1536	1534	99.87	1536	trinity лучше выровнялась на проверке (шаг 6)	trinity
14	<i>Sluginella_kietlinskii</i>	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	<i>Boeckxelia_carpenterii</i>	1534	1534	95.24	1534	1534	1461	100.0	rnaspades
2	<i>Brachyuropus_grewingkii</i>	1680	1534	91.31	1680	1680	1534	n/a	trinity
3	<i>Eulimnogammarus_sp._gam2quest</i>	1534	1534	97.46	1536	1502	1534	99.53	trinity
4	<i>Eulimnogammarus_violaceus</i>	1534	1534	65.19	1534	1534	1001	99.9	rnaspades
5	<i>Gammarus_minus</i>	1534	1534	93.68	1534	1437	1534	100.0	trinity
6	<i>Gammarus_pisinnus</i>	1538	1534	70.29	1642	1496	1534	72.26	trinity
7	<i>Garjajewia_dershawini</i>	1537	1534	97.72	1554	1537	1534	n/a	rnaspades
8	<i>Hyalelloopsis_costata</i>	1534	1537	66.56	1609	1534	1537	n/a	rnaspades
9	<i>Hyalelloopsis_stebbingi</i>	1535	1534	98.5	1535	1514	1534	99.87	trinity
10	<i>Macrohectopus_branickii</i>	1534	1534	74.58	1554	1534	1520	75.39	rnaspades

11	<i>Macropereiopus_parvus</i>	1536	1534	96.55	1565	1536	1534	n/a	trinity
12	<i>Micruropus_wahlilii</i>	1534	1534	79.66	1536	1534	1534	n/a	trinity
13	<i>Ommatogammarus_albinus</i>	1534	1534	83.57	1539	1534	1534	n/a	rnapad es
14	<i>Palicarinus_puzyllilii</i>	1534	1534	78.1	1544	1534	1534	n/a	rnapad es
15	<i>Pallasea_cancelloides</i>	1541	1534	95.46	1587	1541	1534	n/a	trinity
16	<i>Parapallasea_borowskii</i>	3036	1534	50.26	3039	3036	1534	n/a	trinity
17	<i>Parapallasea_wosnessenskii</i>	1549	1534	94.06	1563	1520	1534	95.86	trinity

## Шаг 6. Финальная проверка последовательностей в blast

```
python scripts/check_blast_both_script.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
rnapades								
<i>Acanthogamm</i>	NODE_41_le	JN39375	Acanthogammaru	0.0e+00	99.66	585	высокая	последовательность

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Вопрос	Решение
<i>arus_godlewskii</i>	ngth_9587_cov_15664.560390_g2_i2	5.1	s cf. maculosus MED-2011 h		%		идентичность для короткой последовательности и соседнего вида	не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Brandtia latissima</i>	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 и того же вида	внутривидовое разнообразие
<i>Eulimnogammarus_sp._gam2quest</i>	NODE_2643_length_4669_cov_13692.467316_g13_i13	AY926663.1	Eulimnogammarus maackii cytochrome oxidase	0.0e+00	99%	450	можем ли утверждать, что это вид мааски, берем trinity	найденный ген 18S в сборке выравнивается на E. maackii с высокой идентичностью
<i>Eulimnogammarus_viridulus</i>	_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 и соседнего вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Micruropus glaber</i>	NODE_299_length_7239_cov_6437.71	AY926682.1	Micruropus glaber cytochrome	0.0e+00	91.79%	694	низкая идентичность для короткой	последовательность не включаем в базу данных,

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Alignment	Вопрос	Решение
	7524_g7_i3		oxidase sub				последовательность и того же вида	откладываем в папку "no_pass_blast_check"
<i>Micruropus_wahl</i> <i>hlii</i>	NODE_22247_length_1635_cov_1417.979219_g11519_i0	FJ756340.1	Micruropus wahl voucher MWAH2 cytochro	0.0e+00	90.26%	626	низкая идентичность для короткой последовательности и того же вида	берем trinity
<i>Odontogammarus_calcaratus</i>	_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 и того же вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Oxyacanthus_curtus</i>	_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 и соседнего вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Pachyschesis_branchialis_2</i>	NODE_1_length_14059_cov_26894.746966_g0_i0	MN148359.1	Pachyschesis branchialis isolate 35-6 cy	0.0e+00	98.44%	705	низкая идентичность для короткой последовательности	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	папку "no_pass_blast_check"
<i>Pachyschesis_branchialis</i>	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 и того же вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Pallasea_grubei</i>	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 и того же вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Pallaseopsis_kessleri</i>	_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 и соседнего вида, на сам <a href="#">ген</a> из NCBI выравнился очень плохо - 78%	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
trinity								

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Вопрос	Решение
<i>Acanthogammarus_godlewskii</i>	_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 и соседнего вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Brandtia_latissima</i>	_R_TRINITY_DN94_c0_g2_i4	FJ756302.1	Brandtia latissima lata voucher BLATA66	0.0e+00	95.85 %	626	низкая идентичность для короткой последовательности и того же вида	внутривидовое разнообразие
<i>Eulimnogammarus_sp._gam2quest</i>	_R_TRINITY_DN531_c0_g1_i4	AY926663.1	Eulimnogammarus maackii cytochrome oxidase	0.0e+00	99%	450	можем ли утверждать, что это вид мааски	найденный ген 18S в сборке выравнивается на E. maackii с высокой идентичностью
<i>Eulimnogammarus_viridulus</i>	TRINITY_DN93_c0_g3_i2	MK887742.1	Eulimnogammarus vittatus voucher Evi_Lis	0.0e+00	99.84 %	640	высокая идентичность для короткой последовательности и соседнего вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Micruropus_glaber</i>	_R_TRINITY_DN7569_c0	AY926682.1	Micruropus glaber	0.0e+00	91.79 %	694	низкая идентичность для	последовательность не включаем в

Name	Sequence_ID	Subject_ID	Organism	E-value	Identity	Align_len	Вопрос	Решение
	_g1_i1		cytochrome oxidase sub				короткой последовательность и того же вида (но для 18s все хорошо)	базу данных, откладываем в папку "no_pass_blast_check"
<i>Odontogammarus calcaratus</i>	_R_TRINITY_DN1782_c0_g1_i4	FJ756341.1	Odontogammarus calcaratus voucher OCAL20	0.0e+00	92.82 %	627	низкая идентичность для короткой последовательность и того же вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Oxyacanthus curtus</i>	TRINITY_DN1_c0_g1_i4	JN393847.1	Oxyacanthus flavus cytochrome oxidase su	0.0e+00	99.66 %	585	высокая идентичность для короткой последовательность и соседнего вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Pachyschesis branchialis_2</i>	TRINITY_DN40_c0_g1_i4	MN148359.1	Pachyschesis branchialis isolate 35-6 cy	0.0e+00	98.44 %	705	низкая идентичность для короткой последовательность и того же вида (но для 18s все хорошо)	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
<i>Pachyschesis_branchialis</i>	_R_TRINITY_DN387_c0_g1_i3	MN14835 9.1	Pachyschesis branchialis isolate 35-6 cy	0.0e+00	87.68 %	706	низкая идентичность для короткой последовательности и того же вида (но для 18s все хорошо)	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Pallasea_grubei</i>	TRINITY_DN56_c0_g1_i1	AY92668 8.1	Pallasea grubei cytochrome oxidase subun	0.0e+00	92.17 %	498	низкая идентичность для короткой последовательности и того же вида	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"
<i>Pallaseopsis_kessleri</i>	TRINITY_DN670_c0_g1_i5	GQ91920 2.1	Babr nigromaculatus isolate A60 cytochro	0.0e+00	98.56 %	627	высокая идентичность для короткой последовательности и соседнего вида, на сам <a href="#">ген</a> из NCBI выравнился плохо - 78%	последовательность не включаем в базу данных, откладываем в папку "no_pass_blast_check"



## Поиск гена 18S в транскриптомных сборках

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

```
python scripts/main_pipeline.py
```

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

```
python scripts/check_len_script.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	<i>Baikalogrammarus pullus</i>	NODE_159_length_5935_cov_26498.83_4523_g1_i4	2347	2333	True	Замена в best_to_ref на хит без вставки	проверку по длине прошел
2	<i>Caprella_sp</i>	NODE_8049_length_2019_cov_1517.79_4027_g4408_i0	3508	2019	True	Не байкальский вид	Не интересует
3	<i>Echinogammarus</i>	_R_NODE_675_len	2389	2377	True	Не байкальский вид	Не интересует

	<i>_berilloni</i>	gth_9538_cov_2297.755945_g503_i0					
4	<i>Eogammarus_posseticus</i>	_R_NODE_272_len_gth_9513_cov_2210.622192_g140_i0	2338	2288	True	Не байкальский вид	Не интересует
5	<i>Eulimnogammarus_testaceus</i>	NODE_3500_length_3924_cov_14024.027097_g747_i10	2380	2371	True	топ 20	проверку по длине прошел
6	<i>Eulimnogammarus_ussolzevii</i>	NODE_1759_length_4552_cov_29029.299134_g1279_i0	5365	4552	True	топ20 - не помогло	trinity
7	<i>Eulimnogammarus_viridulus</i>	_R_NODE_1211_len_gth_5584_cov_34217.632159_g42_i9	2277	2214	True	топ20	проверку по длине прошел
8	<i>Gammarus_chevreuxi</i>	NODE_479_length_9075_cov_495.103922_g118_i0	2291	2275	True	Не байкальский вид	Не интересует
9	<i>Gammarus_lacustris</i>	_R_NODE_16_length_8577_cov_613.017237_g9_i0	2306	2287	True	топ20 - не помогло, мало контигов, низкие покрытия	trinity
10	<i>Gammarus_pisinnus</i>	_R_NODE_522703_length_4619_cov_26.735526	4692	4619	True	Не байкальский вид	Не интересует
11	<i>Hirondellea_gigas</i>	_R_NODE_811_length_5757_cov_5491.491197_g338_i1	2556	2534	True	Не байкальский вид	Не интересует

12	<i>Hyalella_azteca</i>	NODE_26_length_1 0142_cov_641.9593 27_g11_i1	2441	2370	True	Не байкальский вид	Не интересует
13	<i>Hyalellopsis_setosa</i>	_R_NODE_3972_length_2187_cov_318 8.257250_g547_i3	2267	2187	True	топ20 - не помогло	trinity
14	<i>Linevichella_vortex</i>	_R_NODE_14_length_11798_cov_4941. 989957_g9_i0	2270	2249	True	топ20 - не помогло	trinity
15	<i>Macropereiopus_parvus</i>	_R_NODE_9621_length_2238_cov_942 0.869804_g15_i26	2268	2238	True	топ20 - не помогло	trinity
16	<i>Marinogammarus_marinus</i>	_R_NODE_429_length_10268_cov_258 6.597979_g246_i1	2379	2361	True	Не байкальский вид	Не интересует
17	<i>Ommatogammarus_albinus</i>	NODE_8224_length_3017_cov_15755.7 05526_g5424_i0	2364	2292	True	топ20	проверку по длине прошел
18	<i>Pachyschesis_branchialis_2</i>	_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	<i>Palicarinus_puzyl lii</i>	NODE_2889_length_4280_cov_11925.144174_g830_i5	2277	2215	True	топ20	проверку по длине прошел
20	<i>Pallasea_cancell oides</i>	_R_NODE_1545_length_2819_cov_6883.720217_g1047_i0	2267	2211	True	топ 20 - не помогло	trinity
21	<i>Pallasea_grubei</i>	NODE_12_length_15903_cov_8094.016021_g7_i1	2319	2257	True	топ20	проверку по длине прошел
22	<i>Pandorites_podo ceroides</i>	NODE_4_length_9829_cov_950.867280_g2_i0	2273	2224	True	Не байкальский вид	Не интересуется
23	<i>Parhyale_hawaie nsis</i>	NODE_54_length_10241_cov_1174.547913_g31_i0	2533	2490	True	Не байкальский вид	Не интересуется
24	<i>Sluginella_kietlin skii</i>	NODE_309_length_3524_cov_3241.044291_g220_i0	2268	2242	True	топ 20 - не помогло	trinity
25	<i>Talitrus_saltator</i>	NODE_36_length_9591_cov_379.381891_g28_i0	2368	2290	True	Не байкальский вид	Не интересуется
trinity							
1	<i>Asprogammarus_rhodophthalmus</i>	_R_TRINITY_DN1125_c0_g1_i2	2270	2085	True	топ20 - не помогло	rnaspades

2	<i>Dorogostaiskia_p arasitica</i>	_R_TRINITY_DN14 30_c0_g2_i6	2267	2245	True	только 2 последовательности длины > 2000	rnapades
3	<i>Eulimnogammaru s_cruentus_2</i>	TRINITY_DN1113_c 0_g1_i6	3536	3533	True	в best_to_ref перенос последнего нуклеотида вручную	проверку по длине прошел
4	<i>Eulimnogammaru s_testaceus</i>	_R_TRINITY_DN13 8_c0_g1_i15	2321	2318	True	топ20 - не помогло	rnapades
5	<i>Eulimnogammaru s_vittatus</i>	TRINITY_DN10693_ c0_g2_i3	2304	2299	True	только 1 контиг длинный	rnapades
6	<i>Hyalella_azteca</i>	_R_TRINITY_DN55 96_c0_g1_i1	3449	3370	True	Не байкальский, не нужен	
7	<i>Hyalelloopsis_cost ata</i>	_R_TRINITY_DN11_ c0_g1_i11	2267	1686	True	топ20 - не помогло	rnapades
8	<i>Hyalelloopsis_gris ea</i>	_R_TRINITY_DN11_ c0_g1_i6	2267	1644	True	топ20 - не помогло	rnapades
9	<i>Hyalelloopsis_steb bingi</i>	_R_TRINITY_DN70 _c0_g1_i24	2269	1636	True	топ20 - не помогло	rnapades
10	<i>Micruropus_parv ulus</i>	_R_TRINITY_DN14 _c0_g2_i3	2270	1980	True	топ20 - не помогло	rnapades
11	<i>Pachyschysis_br anchialis_2</i>	_R_TRINITY_DN12 _c0_g1_i5	2294	2292	True	топ20 - не помогло	rnapades
12	<i>Pandorites_podo ceroides</i>	_R_TRINITY_DN18 59_c0_g1_i11	2273	2224	True	Не байкальский, не нужен	

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

```
python scripts/check_blast_both_script.py
```

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

### Шаг 4. Сравнение последовательностей из разных сборок

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	<i>Baikalogrammarus_pullus</i>	2272	2270.0	99.428	2274.0	2272	2270.0	n/a	выше идентичность в бласте	trinity
2	<i>Boeckxelia_carpensterii</i>	2270	2271.0	99.119	2271.0	2259	2271.0	99.65	нет гэпов в начале	trinity
3	<i>Brandtia_latissima</i>	2268	2268.0	99.471	2268.0	2257	2268.0	99.96	нет гэпов в начале	trinity
4	<i>Cornugammarus_maximus</i>	2268	2268.0	99.515	2268.0	2257	2268.0	100.0	нет гэпов в начале	trinity
5	<i>Echiuropus_macronychus</i>	2269	2268.0	99.956	2269.0	2269	2268.0	n/a	выше идентичность в бласте	trinity

6	<i>Eucarinogammarus_wagii</i>	2268	2272.0	99.604	2272.0	2268	2272.0	n/a	выше идентичность с референсом	rnapades
7	<i>Eulimnogammarus_czerskii</i>	2267	2268.0	99.162	2268.0	2249	2268.0	100.0	нет гэпов в конце	trinity
8	<i>Eulimnogammarus_sp._gam16.4</i>	2268	2268.0	99.868	2270.0	2268	2268.0	n/a	выше идентичность с референсом	rnapades
9	<i>Hyalellopsis_carinata</i>	2268	2268.0	99.735	2270.0	2268	2268.0	n/a	выше идентичность с референсом	rnapades
10	<i>Linevichella_vortex</i>	2270	2271.0	99.648	2271.0	2270	2271.0	n/a	выше идентичность в бласте	rnapades
11	<i>Macrohectopus_branickii</i>	2269	2270.0	99.031	2278.0	2269	2270.0	n/a	выше длина выравнивания на реф	rnapades
12	<i>Odontogammarus_alcaratus</i>	2268	2268.0	99.868	2270.0	2268	2268.0	n/a	выше идентичность с референсом	trinity
13	<i>Ommatogammarus_albinus</i>	2268	2268.0	99.515	2268.0	2257	2268.0	100.0	нет гэпов в начале	trinity
14	<i>Pallasea_sp._gam7.3</i>	2268	2268.0	99.427	2268.0	2268	2257.0	99.91	нет гэпов в начале	rnapades
15	<i>Pentagonurus_dawydowi</i>	2268	2268.0	99.912	2269.0	2268	2268.0	n/a	cleaned одинаковые	rnapades/trinity

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

№	Sequence_ID	Len_Se q_rnapades	Len_S eq_trinity	Identity	Alignm ent_Le n	Len_Seq_wo _gaps_rnapades	Len_Seq_ wo_gaps_t rinity	Identity_ wo_gaps	Justification	Final_s eq
---	-------------	----------------------	---------------------	----------	-----------------------	------------------------------	---------------------------------	----------------------	---------------	---------------

1	<i>Asprogammarus_rhodophthalmus</i>	2270	2270.0	87.445	2387.0	2270	2210.0	89.82	нет гэпов в начале	maspades
2	<i>Dorogostaiskia_parasitica</i>	2268	2267.0	98.589	2281.0	2257	2267.0	99.07	чуть длиннее, но вместе - целый ген	maspades
3	<i>Eulimnogammarus_testaceus</i>	2285	2321.0	94.313	2355.0	2285	2321.0	n/a	по длине	maspades
4	<i>Eulimnogammarus_ussolzewii</i>	5365	2268.0	33.439	5414.0	4579	2268.0	79.1	по длине	trinity
5	<i>Eulimnogammarus_vittatus</i>	2268	2304.0	98.264	2304.0	2268	2304.0	n/a	по длине	maspades
6	<i>Gammarus_lacustris</i>	2306	2269.0	98.395	2306.0	2306	2269.0	n/a	по длине	trinity
7	<i>Hyalelloopsis_costata</i>	2268	2267.0	73.545	2695.0	2268	1691.0	98.64	нет гэпов	maspades
8	<i>Hyalelloopsis_grisea</i>	2268	2267.0	71.825	2336.0	2268	1648.0	99.03	нет гэпов	maspades
9	<i>Hyalelloopsis_setosa</i>	2267	2268.0	96.649	2268.0	2192	2268.0	100.0	нет гэпов	trinity
10	<i>Hyalelloopsis_stebbingi</i>	2268	2269.0	71.177	2309.0	2268	1638.0	98.29	нет гэпов	maspades
11	<i>Macropereiopus_parvus</i>	2268	2268.0	98.589	2268.0	2241	2268.0	99.78	нет гэпов	trinity
12	<i>Micruropus_parvulus</i>	2271	2338.0	73.524	2490.0	2271	2338.0	n/a	по длине	maspades
13	<i>Pallasea_cancelloides</i>	2267	2268.0	97.531	2268.0	2214	2268.0	99.91	нет гэпов	trinity



14	<i>Sluginella_kietlinskii</i>	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

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