Построение филогении байкальских бокоплавов по полным транскриптомам

Попова Н.В.
ФББ МГУ
nina.tolmacheva@gmail.com

Науменко С.А.
ИППИ РАН, ФББ МГУ, РХТУ
sergey.naumenko@yahoo.com

Аннотация

В рамках проекта по изучению молекулярной эволюции и симпатрического видообразования байкальских боковплавов произведена сборка транскриптомов и простейшая аннотация генов 31 из 35 образцов. Для 11 образцов выделены консервативные блоки множественного выравнивания и построено филогенетическое дерево. Полученный результат является важной вехой в проекте изучения эволюции бокоплавов.

1. Введение

Байкальские бокоплавы (гаммариды) чрезвычайно разнообразны (более 270 видов) и силу своей всеядности являются важнейшими санитарами Байкала [1].

Возраст Байкала оценивается в 25-35 млн. лет (по другой теории, 8-150 тыс. лет) [2].

Известный возраст и большое видовое разнообразие делает бокоплавов чрезвычайно интересным эволюционным объектом для изучения.

2. Получение белок-кодирующих генов

В большинстве геномных проектов *de novo* для получения белок-кодирующих последовательностей применяют секвенирование генома с последующей сборкой и аннотацией. Таким образом получается референсный геном данного вида. Для улучшения и верификации аннотации применяется дополнительное секвенирование транскриптомов.

Для изучения внутрипопуляционной изменчивости или изменчивости для близких видов секвенируют полные геномы с меньшим покрытием и картируют полученные чтения на референсный геном.

Такой подход позволяет получить для популяционного анализа однонуклеотидные полиморфизмы и короткие выпадения-вставки.

Для изучения длинных выпадений-вставок, инверсий, геномных перестроек каждый секвенированный геном собирают *de novo* и строят полногеномные множественные выравнивания.

В нашем проекте применяется последний подход: в молекулярной лаборатории из образцов выделена РНК, подготовклен библиотеки для секвенирования и произведено секвенирование на приборе HiSeq 2000. Для каждого из 35 образцов получено в среднем 19,5 миллионов парных чтений длиной 2x101 нуклеотид.

3. Близкородственные геномы и транскриптомы

Референского генома бокоплава на данный момент не опубликовано. Ближайший имеющийся в наличии геном – это геном дафнии *Daphia pulex* [3].

Существуют работы по транскриптомам гавайского бокоплава *Parhyale hawaiensis* [4,5]. Гавайский вид является важной лабораторной моделью для изучения биологии развития ракообразных.

В первой работе [4] транскриптом был отсеквенирован на платформе Roche 454, получено более 1.1 миллиарда оснований кДНК, после сборки и аннотации получено 19,067 уникальных бласт-хитов против базы NR (evalue=e-10).

Во второй работе [5] было получено 1.6 млн. чтений на платформе Roche 454 длиной 200-600 нуклеотидов. В результате сборки транскриптома получено 41,013 групп изоформ, средняя длина транскрипта составила 1099 нуклеотидов. Аннотирование посредством blastx против 17 протеомов дало 12,271 хитов (evalue=e-10).

4. Пакет программ agalma

В настоящее время область обработки геномных данных бурно развивается, поэтому её программное обеспечение в большинстве случаев нельзя назвать зрелым. В основном программы разрабатываются в рамках исследовательских проектов, имеют недостаточно дружественные интерфейсы, большой набор параметров, сырую документацию и плохо стыкуются друг с другом. Как правило, чаще используются наиболее стабильные и простые в использовании программы, которые активно поддерживаются авторами на протяжении нескольких лет, например, в области сборки геномов лидирует пакет velvet [6], ссылки на эту программу имеются в данный момент в 1621 статье.

Для обработки больших наборов данных, когда нужно применить одни и те же операции к большому количеству образцов, процесс отлаживается и автоматизируется в виде конвееров программ (пайплайнов). В данный момент не существует общепринятого стандарта разработки такого ПО, большинство исследователей автоматизируют процесс доступными им методами на скриптовых языках.

В нашем проекте используется пакет agalma [7,8], построенный на основе библиотеки biolite на языке python. Он представляет собой набор оболочек (wrappers) для вызова различных программ обработки геномных данных [6,9-13], которые можно запускать в нужной последовательности, контролируя объем ресурсов (процессорного времени, оперативной памяти), которые потребляются на каждом шаге.

Agalma содержит 10 конвееров, которые позволяют фильтровать чтения по качеству (sanitize), оценивать размер вставки (insert_size), находить и удалять рибосомальную РНК (remove_rrna), собирать транскриптомы посредством oases и trinity (assemble), генерировать отчеты по результатам сборки (postassemble), загружать геномные данные и аннотации из других источников (load), находить кластеры гомологичных последовательностей (homologize), строить множественные выравнивния (multalign), удалять паралоги (treeprune) и строить филогенетические деревья, вызывая программу RaXML [14].

Пакет agalma выпущен не так давно, он поддерживает только парные чтения, полученные на платформе Illumina. Документацию по пакету нельзя назвать исчерпывающей, сообщество накопило еще очень мало опыта его использования, однако авторы оперативно комментируют возникающие проблемы.

5. Сборка транскриптомов

Результаты сборки транскриптомов отражены в таблице 1. Сборщик oases генерирует более длинные контиги, однако по числу контигов, имеющих совпадения в базе белков SwissProt предпочтительнее сборки Trinity. Полная обработка одного образца конвеером agalma transcriptome занимает около 12 часов (10 ядер, 100G RAM). Расчеты проводились на вычислительном кластере лаборатории эволюционной геномики ФББ МГУ [15].

6. Филогенетическое дерево

Для построения филогенетического дерева (рис. 1) были методом двустороннего совпадения (Best Blast Hit) с белками дафнии были извлечены ортологичные последовательности 11 бокоплавов, из них сформированы группы ортологов.

По выравниваниям с белками дафнии в последовательностях транскриптов была восстановлена рамка чтения, и нуклеотидные последовательности были транслированы в аминокислотные.

Затем были построены множественные выравнивания и извлечены консервативные сайты (имеющие лишь 2 аллеля в бокоплавах).

Дерево построено методом наибольшей экономии при помощи пакета phylip [16] с примененем бутстрэп-анализа: на ветвях отражен процент поддержки топологии дерева при построении 100 деревьев с перемешиваением входных последовательностей.

Данное дерево построено без применения пакета agalma.

7. Благодарности

Проект осуществляется в широкой коллаборации, авторы благодарны инициаторам

проекта А.С. Кондрашову, Л. Ю Ямпольскому и Г.А. Базыкину за плодотворные обсуждения,
Л. Ю. Ямпольскому и Н. Мюге – за сбор материала, М. Логачевой, А. Пенину – за проведеные работы по секвенированию, М. Щелкунову – за первичную обработку данных, М.С. Гельфанду – за предоставленную ссылку на программу agalma.

Работа осуществлялась при поддержке Министерства образования и науки РФ (гранты 11.G34.31.0008 and 8814) и РФФИ (грант 12-07-31261).

Литература

[1] Научно-образовательный центр “Байкал”, доступно по адресуhttp://lake.baikal.ru/ru/baikalinfo/excurs/crustacea3.html

[2]Байкал, доступно по адресу http://ru.wikipedia.org/wiki/%D0%91%D0%B0%D0%B9%D0%BA%D0%B0%D0%BB

[3] Colbourne JK et al. The ecoresponsive genome of Daphnia pulex. Science. 2011 Feb 4;331(6017):555-61.

[4] V. Zeng, K.E. Villanueva, B.S. Ewen-Campen, F.Alwes, W.E. Browne and C.G. Extavour. *De novo* assembly and characterization of a maternal and developmental transcriptome for the emergin model crustacean *Parhyale hawaiensis*. *BMC Genomics* 2011, **12**:581 doi:10.1186/1471-2164-12-581.

[5] M.J. Blythe, S. Malla, R. Everall, Y. Shih, V. Lemay, J. Moreton, R. Wilson, A.A. Aboobaker. High through-put sequnecing of the *Parhyale hawaiensis* mRNAs and microRNAs to aid comparative developmental studies. PLoS ONE 2012, 7(3): e33784. doi:10.1371/journal.pone.0033784.

[6] D.R. Zerbino and E. Birney. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 18:821-829.

[7] M. Howison, N. Sinnott-Armstrong, & C.W. Dunn (2012). BioLite, a lightweight bioinformatics framework with automated tracking of diagnostics and provenance. In Proceedings of the 4th USENIX Workshop on the Theory and Practice of Provenance (TaPP '12).

[8] Agalma, доступно по адресу https://bitbucket.org/caseywdunn/agalma

[9] Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. doi:10.1093/bioinformatics/bts094

[10]Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. doi:10.1186/gb-2009-10-3-r25

[11]Andrews, S. FastQC: A quality control tool for high throughput sequence data. [http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/]

[12]Wasmuth JD, Blaxter ML (2004) prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics 5:187. doi: 10.1186/1471-2105-5-187

[13]Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. doi:10.1093/nar/25.17.3389

[14]Stamatakis A. RAxML-VI-HPC:Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models”, Bioinformatics 22(21):2688–2690, 2006.

[15] Арифулов Р.Н., Науменко С.А. Опыт создания центра обработки данных и вычислительного кластера для лаборатории эволюционной геномики. //Сборник трудов конференции «Информационные технологии и системы» (ИТиС'12). Петрозаводск, 19-25 августа 2012г. ISBN 978-5-901158-19-7.

[16] Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

Таблица 1. Результаты сборки транскриптомов.

Вид	Кол-во парных чтений в млн	Сборка oases				Сборка trinity			
		Локусов	Средняя длина транскрипта	N50	Blast hits	Локусов	Средняя длина транскрипта	N50	Blast hits
Pallassea cancelloides	37	18283	862	1117	3758	25094	750	912	5474
Eucarinogammarus wagii	17	-	-	-	-	-	-	-	-
Gmelinoides fasciatus	20	14273	1192	1725	3367	24525	871	1159	5141

Вид	Кол-во парных чтений в млн	Сборка oases				Сборка trinity			
Acanthogammarus maximus	14	19828	953	1279	6192	37290	722	833	10493
Unknown gammarus	19	18348	1192	1772	5329	34735	848	1105	8302
Acanthogammarus godlewskii	15	18183	1031	1461	4953	34303	766	936	8551
Gammarus lacustris	21	20243	1242	1854	4954	35685	905	1261	7551
Eulimnogammarus cruentus	14	23510	1091	1601	6,547	42289	814	1048	10238
Gmelinoides fasciatus	20	20347	1149	1652	5658	37462	776	958	9059
Eulimnogammarus sp.	35	20671	1152	1699	4731	29065	858	1126	6318
Acanthogammarus (Brachiuropus) flavus sowinskii	12	19374	1009	1419	5366	34930	777	963	8471
Pallasea (Homalogammarus) brandtii	20	13799	1178	1685	3297	23903	850	1108	5014
Pallasea kessleri	13	15226	1007	1403	4191	27543	775	949	6647
Brandtia latissima	11	19794	990	1388	6398	36173	745	895	10631
Pallasea sp.	20	16682	1143	1627	5441	30922	843	1094	8642
Acanthogammarus (Brachiuropus) flavus curtus	11	18145	952	1278	5381	33374	741	871	8964
Eulimnogammarus sp.	22	12246	1016	1290	2323	16718	791	981	3128
Pallasea cancellus	23	17886	1192	1779	5533	35360	801	1001	10315
Eulimnogammarus cruentus	22	23281	1154	1691	6256	42424	859	1146	10862
Heterogammarus sophi-anosi	17	20199	1154	1672	6338	40441	835	1085	11045
Crypturopus inflatus	8	12685	862	1076	3644	23456	649	697	6451
Acanthogammarus (Brachiuropus) flavus flavus	20	20801	1116	1624	6011	38763	853	1106	10142
Pallasea (Homalogammarus) dawydowi	34	21632	1129	1683	5145	32364	843	1085	7208
Echiuropus macronychus	19	-	-	-	-	-	-	-	-
Macropereiopus sp.	8	15472	899	1137	4419	25655	693	797	6798
Garjajevia calanisi dershavini	14	15198	1141	1623	4049	27403	831	1058	6455
Coniurus sp.	25	18278	1148	1649	3842	26881	844	1097	5378

Вид	Кол-во парных чтений в млн	Сборка oases				Сборка trinity			
Ommatogammarus flavus	21								
Ommatogammarus albus	22								
Micruropus wahli	22	-	-	-	-	-	-	-	-
Eulimnogammarus marituji	22	22343	1142	1739	7147	43927	810	1026	12155
Spinachanthus parasiticus	20	17698	1034	1427	4541	27599	787	979	6554
Macrohectopus branickii	16	-	-	-	-	-	-	-	-
Eulimnogammarus sp.	18	16057	1013	1312	3983	24762	754	919	5848

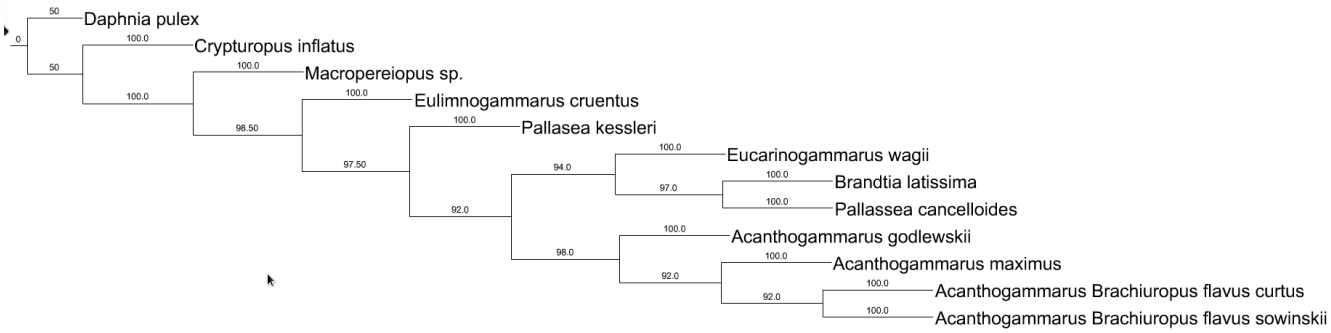

Рис.1. Филогенетическое дерево 11 гаммарусов с дафнией в качестве ветви-кузины.