УДК 577.29

СБОРКА И АННОТАЦИЯ ГЕНОМА *ACYRTHOSIPHON CARAGANAE*

Н.В. Воронова-Барте, Р.С. Шулинский, Ю.В. Бондаренко, М.Д. Ракова
Белорусский государственный университет, Минск, rkvmry@gmail.com

При благоприятных условиях окружающей среды тли способны за короткие сроки приобрести высокую плодовитость, миграционную активность, трофическую лабильность, устойчивость к ксенобиотикам и захватить все доступные экологические пространства, что приводит к быстрой смене генетической структуры популяции и вида в целом [1]. Таким образом, понимание молекулярных механизмов приобретения устойчивости к токсинам необходимо для разработки качественных программ мониторинга численности популяций фитофагов, понимания популяционной генетики и эволюции резистентности. По этой причине в последнее время внимание исследователей привлечено к изучению системы детоксикации тлей, поскольку без понимания особенностей функционирования этой системы контроль процессов, связанных с формированием устойчивости и преодоления защитных механизмов растений, не представляется возможным. Для изучения системы детоксикации тлей необходимо иметь качественно собранные и проаннотированные геномы.

Acyrthosiphon caraganae Chol. (большая акациевая тля) является инвазивным видом, который способен вредить культивируемым и иным хозяйственно ценным растениям, включая интродуценты. Агрегации *A. caraganae* размещаются на вершинах растущих побегов, молодых листьях и зеленых плодах *Caragana sp.* Lam. (карагана) из семейства бобовых (Fabaceae) [2], а также на *Colutea arborescens* (пузырнике обыкновенном) [3]. Очевидно, что более узкий круг растений-хозяев вынуждает насекомых вырабатывать соответствующие узкоспециализированные адаптации, а эффективность данных адаптаций будет максимальной, благодаря тщательному отбору особей в ряду поколений [4, 5]. Важно отметить, что именно способность нейтрализовать вторичные метаболиты является краеугольным камнем способности насекомых к утилизации того или иного вида растения. Была выявлена прямая зависимость между устойчивостью насекомых к инсектицидам и их способностью адаптироваться к питанию на токсичных растениях, таким образом, чем более развита система детоксикации у насекомых, тем легче последним адаптироваться к питанию на токсичных растениях [6], что делает изучение системы детоксикации чрезвычайно важной задачей.

Пробоподготовка проводилась сотрудниками СНИЛ биоинформатики и молекулярной эволюции животных, коллектирование имаго *A. caraganae* с *C. arborescens* было осуществлено на территории Республики Беларусь. Для выделения ДНК был использован набор «Blood-Animal-Plant DNA Preparation Kit» (JenaBioscience, ФРГ) в соответствии с протоколом производителя. Геном *A. caraganae* был секвенирован компанией Macrogen (Республика Корея) и передан в виде не подвергнутых какой-либо обработке данных. Полногеномное секвенирование *A. caraganae* проводилось на MiSeq с использованием библиотеки TruSeq 350. Объем полученных данных составил 45 GB, что соответствует примерно 100-кратному покрытию генома.

При обработке прочтений (ридов) были проведены оценка качества прочтений при помощи программы FastQC и удаление некачественных оснований в программе Trimmomatic. Для дальнейшей работы отобрали эукариотические прочтения за счет выравнивания ридов по всем бактериальным геномам, доступных в RefSeq, а также путем выравнивания прочтений на митохондриальные геномы тлей в программе Bowtie2 с последующим отбором прочтений, которые не выровнялись на бактериальный и митохондриальный геномы соответственно. Для работы с форматами нуклеотидных данных использовались Samtools и Bedtools. После отобранные прочтения были использованы для сборки в ассемблере AbySS2.0 (алгоритм нахождения Эйлерового пути в графах де Брейна для генерации контиг). Подбор k-мера осуществлялся эвристически. Оценка и выбор сборки с наилучшими статистическими показателями проводилась в программе Quast. В процессе сборки вставки могут не перекрываться с вершинами графа по причине достаточно большого значения k-мера, поэтому для закрытия пробелов в сборке была использована программа Pilon, за счет проведения выравнивания прочтений на полученную сборку. Также проводилась процедура удаления контаминаций на уровне контиг в программе Diamond.

Для осуществления структурной аннотации генома было построено несколько предсказаний моделей генов с использованием РНК-последовательностей из NCBI, белковой гомологии и скрытых марковских моделей с последующей финальной генерацией консенсусных вариантов моделей

при помощи утилиты EVidenceModeller. Основные действия были проведены в пайплане Maker. В качестве *ab initio* предсказаний были использованы SNAP, обученный на РНК-последовательностях, AUGUSTUS обученный на моделях, предсказанных SNAP и самообучающийся GeneMark. Для предсказания генных моделей по белковой гомологии были использованы транслированные CDS геномов тлей из базы данных RefSeq представленные в таблице 1.

Таблица 1. – Геномы тлей, использованные для аннотации

Вид	Код доступа в RefSeq
Acyrthosiphon pisum	GCF_005508785.1
Aphis gossypii	GCF_020184175.1
Rhopalosiphum maidis	GCF_003676215.2
Diuraphis noxia	GCF_001186385.1
Myzus persicae	GCF_001856785.1
Sipha flava	GCF_003268045.1
Melanaphis sacchari	GCF_002803265.2

Для оценки качества структурной аннотации были использованы программы CEGMA и BUSCO. При проведении функциональной аннотации полученный в программе EVM финальный набор генов был идентифицирован путем проведения выравнивания против баз данных RefSeq, KEGG Ontology, InterPro, PFAM, Gene Ontology.

Таблица 2. – Статистика сборки генома *A. caraganae* с разными заданными k-мерами

Результат сборки	k-мера, длина, пар нуклеотидов (п.н.)						
	82	86	90	104	110	118	124
Число полученных контигов	39311	40102	40556	48844	55654	60108	58795
Число контигов с длиной ≥ 1000 п.н.	32656	33012	33325	40127	45797	47993	44521
Число контигов с длиной ≥ 5000 п.н.	17394	17563	17863	19716	20250	19446	16325
Число контигов с длиной ≥ 10000 п.н.	10083	10141	10243	10530	10190	9451	7949
Число контигов с длиной ≥ 25000 п.н.	3554	3551	3624	3495	3144	2864	2517
Число контигов с длиной ≥ 50000 п.н.	975	1018	1040	991	876	835	697
Суммарная длина контигов с длиной ≥ 500 п.н.	362561292	369058184	375309581	398747801	403255361	389541600	341084692
Суммарная длина контигов с длиной ≥ 1000 п.н.	357963805	364166147	370329605	392751668	396467288	381202654	331219872
Суммарная длина контигов с длиной≥ 5000 п.н.	314003748	319439889	325575650	333657984	323578658	303080971	255021403
Суммарная длина контигов с длиной≥ 10000 п.н.	262160508	266866609	271462056	268919844	253172011	233128098	196779314
Суммарная длина контигов с длиной≥ 25000 п.н.	159595511	163181358	167230853	160172776	144822392	132470237	112838437
Суммарная длина контигов с длиной≥ 50000 п.н.	71120748	75609911	77441243	73924595	66453280	62713695	50231829
Размер длиннейшей контиги	292105	292113	292252	292082	292086	292119	200386
GC, %	29,96	29,96	29,97	29,96	29,93	29,88	29,78
N_{50}	21132	21260	21436	18266	15572	14192	13503
N_{75}	8946	8953	8969	7397	6223	5630	4951

Сборка генома *A. caraganae* осуществлялась *de novo*, т. е. в отсутствие референсного генома. Прочтения *A. caraganae* были выравнены на геномы *Buchnera aphidicola* (эндосимбионт тлей) и остальные геномы бактерий представленных в RefSeq для предотвращения искажения дальнейших результатов в следствии контаминации собираемого генома последовательностями из других ор-

ганизмов, а также прочтения выравнивались на митохондриальные геномы тлей, с последующим отбором невыровненных прочтений и их загрузкой в программу для сборки геномов *de novo* ABySS2.0. Далее была осуществлена *de novo* сборка с последующей ее оценкой в Quast. А также проведено закрытие пробелов в полученной сборке за счет выравнивания на нее прочтений используя программу Pilon. Для удаления контаминаций на уровне контиг была задействована программа Diamond.

При сборке геномов *de novo* необходимо правильно подобрать и задать условия для проведения данного процесса. Эмпирическим путем был проведен подбор k-мера, результаты которого представлены в таблице 2. Наилучшие результаты были получены при длине k-мера равном 90. При этом условии параметр N_{50} имел наиболее высокое значение (21436 п.н.), также как и параметр N_{75} (8969 п.н.), а сумма длин всех полученных контигов составила 370,330 Mb (млн п.н.), что соответствует среднему размеру генома тли. Контиги, полученные с использованием k-мера равном 90, были в дальнейшем использованы для структурной аннотации.

Для осуществления структурной аннотации генома было построено несколько предсказаний моделей генов с использованием РНК-последовательностей, белковой гомологии и скрытых марковских моделей с последующей финальной генерацией консенсусных вариантов моделей при помощи утилиты EVidenceModeller. Основные действия были проведены в пайплане Maker. Данная программа определяет и маскирует повторяющиеся элементы, выравнивает ESTs/РНК-последовательности на геном, выравнивает белки на геном и использует существующие программы для предсказания генов *ab initio* и интегрирует полученные результаты для создания наилучшей предполагаемой модели генов. Предсказание генов осуществляется на основе базовых математических моделей, описывающих паттерны структуры интронов и экзонов, т.к. паттерны структуры генов отличаются между организмами, то данные предсказатели необходимо обучить прежде, чем использовать. В качестве *ab initio* предсказаний были использованы SNAP, обученный на РНК-последовательностях, AUGUSTUS, обученный на предсказаниях SNAP, и самообучающийся GeneMark. Заключительным этапом структурной аннотации являлось построение консенсусной модели генома в программе EVidenceModeler (EVM). Результаты структурной аннотации представлены в таблице 3.

Таблица 3. – Результаты аннотации генома *A. Caraganae*

Метод	Число генов	Средняя длина CDS, п.н.	Средняя длина экзона, п.н.	Среднее число экзонов в гене	Средняя длина гена, п.н.
Белковая гомология					
*Acyrthosiphon pisum**	16586	786	302	3	1626
Aphis gossypii	17530	1001	231	4	3049
Rhopalosiphum maidis	8061	1039	239	4	2897
Diuraphis noxia	7979	854	278	3	1962
Myzus persicae	13011	863	308	3	1850
Sipha flava	9383	808	271	3	1869
Melanaphis sacchari	7935	948	279	3	2226
Гомология с мРНК					
Transcript2genome	18490	1026	277	4	2343
Ab initio					
Augustus**	35443	855	230	4	2467
GeneMark	59849	620	196	3	2183
Snap	45782	372	81	5	3410
Консенсусный набор генов					
EVM	28986	724	239	3	1800

Примечания – *Название референсного комплекта данных, полученного из базы данных RefSeqDatabasedata; ** – Использованный *ab initio* предиктор

Наибольшее количество совпадений по белок-кодирующим генам в собранном геноме *A. caraganae* идентифицировано между геномами *Acyrthosiphon pisum* (16586), *Aphis gossypii* (17530) и *Myzus persicae* (13011), что может быть связано с наилучшей изученностью геномов данных представителей, а также высокой степенью эволюционного родства между данными организмами. *Ab initio* предикторы предсказали в несколько раз больше потенциальных белок-

кодирующих последовательностей, от 35443 (AUGUSTUS) до 59849 (GeneMark). Финальная структурная аннотация представляет собой консенсус между аннотациями, проведенными по белковой гомологии, гомологии с мРНК и с использованием *ab initio* предикторов, и включает 28986 генов, со средней длиной гена 1800 п.н.

Оценка сборки с использованием CEGMA показала, что 93,14 % потенциально консервативных генов были полностью восстановлены. Кроме того, подход BUSCO показал, что 95,47 % универсальных однокопиийных ортологов были аннотированы. Данные показатели позволяют сделать предположение о том, что сборка генома *A. caraganae* представляет относительно полное содержание генов, что коррелирует с раннее полученными геномными сборками других тлей.

По результатам выполненной функциональной аннотации, проведенной путем выравнивания полученных консенсусных моделей генов на аминокислотные последовательности из международных баз данных, можно заключить, что наибольшее число подтвержденных генов было идентифицировано в RefSeq (26410), наименьшее в KEGG Orthology (4730), что отображено в таблице 4.

Таблица 4. – Финальное число генов в геноме *A. caraganae*, установленное по результатам использования различных баз данных

База данных	Число подтвержденных генов	Процент от общего числа обнаруженных генов
RefSeq	26410	91,11
KO	4730	16,32
InterPro	19890	68,61
PFAM	20659	71,27
GO	9498	32,77

Показатель обнаруженных генов в эталонной базе данных Pfam, отражающих качество сборки генома, составил 71,27 %, что свидетельствует о достоверности и качестве проведенной обработки и анализа исходных данных.

Таким образом, в результате проведенной сборки, а также структурной и функциональной аннотации было установлено, что размер генома *A. caraganae* составил 370,330 Mb, что соответствует среднему размеру генома тли. Консенсусный набор генов включает в себя 28986 гена со средней длиной CDS 724 п.н., средней длиной экзона в размере 239 п.н., средним количеством экзонов в гене в размере 3 и средней длиной гена в размере 1800 п.н. Показатели оценки сборки позволяют сделать предположение о том, что сборка генома *A. caraganae* представляет относительно полное содержание генов, что коррелирует с раннее полученными геномными сборками других тлей.

Список использованных источников

1. Воронова, Н.В. Цитохромы p450 у тлей: роль коэволюции с растениями в формировании устойчивости насекомых к инсектицидам / Н.В. Воронова // Труды БГУ. – 2016. – Т. 11. – С. 92–110.

2. Девяткин, А.М. Сельскохозяйственная энтомология. Электронный курс лекций / А.М. Девяткин, А.И. Белый, А.С. Замотайлов. – Краснодар, 2012. – 301 с.

3. Ежов, О.Н. Видовое разнообразие грибных болезней и вредителей ассимиляционного аппарата деревьев и кустарников в городских зеленых насаждениях Архангельской области / О.Н. Ежов. – Москва, 2016. – 224 с.

4. Cohen, M.B. A host-inducible cytochrome P450 from a host-specific caterpillar: molecular cloning and evolution / M.B. Cohen, M.A. Schuler, M.R. Berenbaum // Proc. Natl. Acad. Sci. U.S.A. – 1992. – Vol. 89, № 22. – P. 10920–10924.

5. Li, X. Jasmonate and salicylate induce expression of herbivore cytochrome P450 genes / X. Li, M.A. Schuler, M.R. Berenbaum // Nature. – 2002. – Vol. 419, № 6908. – P. 712–715.

6. Alyokhin, A. Adaptation to toxic hosts as a factor in the evolution of insecticide resistance / A. Alyokhin, Y.H. Chen // Current Opinion in Insect Science. – 2017. – Vol. 21. – P. 33–38.