Санкт-Петербургский государственный университет

Васильев Артем Викторович

Выпускная квалификационная работа

"Эволюционный анализ последовательностей гена
Nxf1 (nuclear export factor) у животных"

Научный руководитель:
к.б.н., доцент, кафедра генетики и биотехнологии,
Голубкова Елена Валерьевна

Рецензент:
заведующая лабораторией, ведущий научный сотрудник,
лаборатория эволюционной геномики и палеогеномики, ЗИН,
к.б.н., с.н.с.,
Абрамсон Наталья Иосифовна

Санкт-Петербург
2023

Аннотация

Выпускная квалификационная работа по теме: "Эволюционный анализ последовательностей гена *Nxf1* (nuclear export factor) у животных".

Выполнил: Васильев Артем Викторович, студент 4 курса бакалавриата по основной образовательной программе 06.03.01 "Биология", лаборатория генетики животных кафедры генетики и биотехнологии СПбГУ.

Научный руководитель: к.б.н., доцент Голубкова Елена Валерьевна.

Продукт гена *Nxf1* является основным переносчиком мРНК из ядра в цитоплазму у всех описанных *Opisthokonta*. В составе последовательности данного гена присутствует эволюционно-консервативный блок, образованный интроном и окружающими его двумя экзонами. В результате альтернативного сплайсинга возможно образование транскрипта с сохраненным интроном, который избегает нонсенс-опосредованного распада, хотя последовательность интрона содержит преждевременный стоп-кодон.

Работа посвящена изучению структуры гена *Nxf1* у представителей разных филогенетических групп в составе таксономической группы *Arthropoda* для выявления закономерностей эволюции генов семейства *Nxf*.

Было показано, что структура консервативного блока гена *Nxf1* является специфичной для всех видов *Arthropoda*, исследованных в данной работе, и для разных клад характерно наличие специфических консервативных последовательностей. Также было продемонстрировано, что интрон-содержащие транскрипты всех изученных видов членистоногих в данной работе образуют специфические вторичные структуры.

Ключевые слова: nuclear export factor (*nxf*), small bristles (*sbr*), интрон-содержащие транскрипты, эволюционно-консервативные последовательности, вторичные структуры транскриптов, *Drosophilidae*, членистоногие, *Arthropoda*.

Abstract

Graduate qualification work on the topic: "Evolutionary analysis of *Nxf1* (nuclear export factor) gene in animals".

Performed by: Vasilev Artem, 4th year undergraduate student in the main educational program 06.03.01 "Biology", Laboratory of Animal Genetics, Department of Genetics and Biotechnology, St. Petersburg State University.

Scientific adviser: candidate of biological sciences, associate professor Golubkova Elena.

The *Nxf1* gene product is the main mRNA carrier from the nucleus to the cytoplasm in all described *Opisthokonta*. The sequence of this gene contains an evolutionarily conservative block formed by an intron and two surrounding exons. As a result of alternative splicing, it is possible to generate an intron-sparing transcript that avoids nonsense-mediated decay, although the intron sequence contains a premature stop codon.

The work is devoted to the study of the structure of the *Nxf1* gene in representatives of different phylogenetic groups within the taxonomic group *Arthropoda* in order to identify patterns of evolution of the *Nxf* family genes.

It was shown that the structure of the conserved block of the *Nxf1* gene is specific for all *Arthropoda* species studied in this work, and the presence of specific conserved sequences is typical for different clades. It was also demonstrated that intron-containing transcripts of all arthropod species studied in this work form specific secondary structures.

Keywords: nuclear export factor (*nxf*), small bristles (*sbr*), intron-containing transcripts, evolutionarily conserved sequences, transcript secondary structures, *Drosophilidae*, arthropods, *Arthropoda*.

Оглавление

Введение

Для большинства генов высших эукариот характерна мозаичная структура, в составе которой выделяют экзоны и интроны. Интроны в процессе созревания транскрипта, как правило, вырезаются путем сплайсинга, и из ядра выходят транскрипты, не содержащие интронов. Благодаря альтернативному сплайсингу сложность протеома увеличивается без изменения числа генов, в результате чего одному гену может соответствовать до нескольких тысяч разных белков.

Особый интерес представляют транскрипты, которые сохраняют интрон. Обычно в сохраненном интроне присутствует преждевременный стоп-кодон, из-за которого происходит прерывание трансляции и появление укороченного белка. Как правило, укороченные или по-другому измененные белки могут быть токсичными для клетки, поэтому транскрипты перед выходом из ядра проходят проверку качества с помощью механизма NMD (nonsense mediated mRNA decay) – нонсенс-опосредованного распада мРНК, который препятствует выходу из ядра транскриптов, содержащих преждевременный стоп-кодон. Такая проверка также осуществляется в цитоплазме во время трансляции благодаря цитоплазматической системе NMD.

Однако, несмотря на наличие специфического механизма, среди совершенно разных эволюционных групп описаны случаи существования транскриптов с сохраненным интроном. Отдельно можно выделить модельный объект *Drosophila melanogaster*, для которого известно семейство генов *Nxf* (nuclear export factor), в котором нас заинтересовал ген *Nxf1*. Данный ген кодирует белок, являющийся основным транспортером мРНК из ядра в цитоплазму.

В составе последовательности гена *Nxf1* присутствует эволюционно-консервативный блок, образованный интроном и окружающими его двумя экзонами. В результате альтернативного сплайсинга возможно образование транскрипта, где этот интрон сохраняется. Транскрипт с сохраненным интроном избегает нонсенс-опосредованного распада, хотя его последовательность содержит преждевременный стоп-кодон. Мы предполагаем, что сохранению интрона способствует образуемая им специфическая вторичная структура. Интрон-сохраняющий транскрипт кодирует белок, однако из-за присутствия преждевременного стоп-кодона, белок образуется укороченным, но при этом функциональным и необходимым для нормального развития организма.

Таким образом, можно сделать вывод о важности сохранения интрона в транскрипте и на основании этого подчеркнуть функциональную и эволюционную значимость интронов как структурных единиц гена.

Целью работы является изучение структуры гена *Nxf1* у представителей разных филогенетических групп в составе *Arthropoda* для выявления закономерностей эволюции генов семейства *Nxf*.

Были поставлены следующие задачи:

1) Поиск последовательностей гена *Nxf1* представителей семейства *Drosophilidae* в открытых базах данных с помощью BLAST и осуществление последующего выравнивания на известную последовательность *Drosophila melanogaster*.

2) Поиск последовательностей гена *Nxf1* других представителей *Arthropoda*, анализ его структуры и сравнение последовательностей между собой, а также с геном *Nxf1 Drosophila melanogaster*.

3) Поиск и изучение консервативных участков гена.

4) Анализ вторичной структуры транскрипта с сохраненным кассетным интроном.

Обзор литературы

Общая структура генома

Несмотря на очень высокий уровень развития науки в наше время, все еще существует много неразрешенных вопросов. Определение гена, как одного из основных и самых важных объектов изучения биологии, относится к данной категории. До сих пор нет четкого определения для этого понятия, но в качестве примера можно привести один из возможных вариантов. Ген – это нуклеотидная последовательность в геноме, окруженная регуляторными участками и кодирующая полипептид или РНК, выполняющие определенную функцию в организме.

В самом простом варианте ген является участком ДНК, полностью колинеарным белковому продукту. Такая структура характерна практически для всех генов бактерий. Гены эукариот же гораздо длиннее, чем образующиеся в итоге функциональные транскрипты. Это связано, во-первых, с наличием регуляторных последовательностей, которые в процессе созревания мРНК выщепляются с 5'- или 3'-конца первичного транскрипта, а во-вторых, с имеющимися вставками, которые перемежают кодирующие участки и не кодируют белки. Они получили название интроны. Элементы, несущие генетическую информацию, называют экзонами (Кребс, Голдштейн и Килпатрик, 2017). Так, для большинства эукариотических генов характерна экзон-интронная, или мозаичная, структура (Burge *et al.*, 1999).

В отличие от генов, не содержащих интроны, экспрессия т. н. прерывистых генов для полного созревания матричной РНК (мРНК), или процессинга, требует дополнительного этапа. Первичный транскрипт, или пре-мРНК, являющийся копией геномной последовательности, не может быть использован для синтеза белка, т. к. содержит интроны, которые необходимо вырезать. После этого образуется зрелая мРНК, состоящая только из экзонов. Процесс удаления интронов и последующего ковалентного соединения концов РНК с образованием цельной молекулы называется сплайсингом РНК (Кребс, Голдштейн и Килпатрик, 2017).

Путь формирования зрелой мРНК из ДНК впервые был описан Фрэнсисом Криком в 1958 году и обозначен, как "Центральная догма молекулярной биологии" – генетическая информация течет только в одном направлении: от ДНК к РНК, после чего к белку или от РНК непосредственно к белку (Crick, 1970). В наше время используется расширенная версия, которая включает, например, прионы (Prusiner, 1998).

Механизмы усложнения организации генома

Существуют механизмы, обеспечивающие общее усложнение организации в ходе эволюции не только за счет увеличения числа белок-кодирующих генов. Альтернативный

сплайсинг является основным источником транскрипционной изменчивости, когда одному гену может соответствовать несколько транскриптов. Одним из вариантов альтернативного сплайсинга является удержание интрона. Помимо него важно обозначить такой процесс, как альтернативное полиаденилирование, т. к. оно также оказывает прямое влияние на вариативность транскриптов (Mamon *et al.*, 2019).

Альтернативный сплайсинг (AS – alternative splicing) достаточно распространен у эукариот. Данный механизм способствует увеличению белкового разнообразия у организма. Выделяют следующие типы альтернативного сплайсинга: т. н. "пропуск экзонов"; вариант, когда используются наборы из нескольких экзонов, расположенных рядом друг с другом и формирующих кассеты, при этом во время сплайсинга выбирается только один из нескольких экзонов в каждой из кассет; также могут использоваться различные 5'- и 3'-сайты сплайсинга в интронах или экзонах, выбор которых зависит от цис- и транс-действующих факторов, необходимых для формирования сплайсосомы (Mamon *et al.*, 2019).

Удержание интрона (IR – intron retention) является одним из вариантов альтернативного сплайсинга, который распространен у млекопитающих. Данный механизм увеличивает сложность транскриптома (Schmitz *et al.*, 2017), но часто в интронах присутствуют кодоны преждевременной терминации (PTCs – premature termination codons) внутри основной открытой рамки считывания. В результате возможно 2 исхода: нонсенс-опосредованный распад мРНК с сохраненным интроном или продукция укороченных белков (Mamon *et al.*, 2019). Именно эти варианты интересуют нас больше всего. Такие мРНК с оставшимся интроном дают альтернативные формы белков, а значит являются источником функционального разнообразия генных продуктов, что демонстрирует важность интронов как структурной единицы гена. Например, для разных организмов показано, что наличие такой мРНК с интроном для гена *Nxf1* является консервативным признаком (Mamon, Kliver and Golubkova, 2013), о чем речь пойдет далее.

Значимость интронов

Интроны делят на 3 группы: I, II, III. Интроны первых двух групп обнаруживаются в геномах некоторых бактерий и органелл. Интроны I группы также обнаружены в рибосомных РНК (рРНК) ядер протистов и грибов. Эти две группы имеют разные структуры РНК, которые облегчают их активность самосплайсинга. Они также содержат внутренние открытые рамки считывания (ORF – open reading frame), которые облегчают как удаление интронов из транскриптов РНК, так и распространение интронов на безинтронные сайты посредством обратной транскрипции (Roy and Gilbert, 2006). Третья же группа интронов – сплайсосомные интроны – найдены в ядерных геномах всех охарактеризованных эукариот. Обычно они не

имеют открытых рамок считывания и их удаление происходит с помощью специального комплекса, состоящего из 5-ти РНК и сотен белков, называемого сплайсосомой (Jurica and Roybal, 2013).

К непосредственным функциям интронов можно отнести следующее.

Во-первых, это положительная регуляция экспрессии генов. Впервые данное явление было продемонстрировано в эксперименте с использованием конструкций вируса SV40, где показали, что количество белка было значительно снижено без его интронов (Gruss *et al.*, 1979). В последующем это было также показано на дрожжах и млекопитающих (Juneau *et al.*, 2006).

Во-вторых, можно отнести регуляцию нонсенс-опосредованного распада мРНК, который был упомянут при описании удержания интрона. Обычно NMD описывают, как механизм надзора у эукариот, который избирательно удаляет мРНК, содержащие ошибочно сгенерированные PTCs (Jo and Choi, 2015). Однако, есть доказательства в пользу того, что интроны, расположенные в 5'- или 3'-нетранслируемых областях (UTR – untranslated region), играют важную роль в контроле NMD-чувствительности транскриптов (Kalyna *et al.*, 2012).

В-третьих, интроны могут быть связаны с транспортом мРНК. Раньше считалось, что сплайсированные транскрипты быстрее экспортируются из ядра в цитоплазму, однако были и противоречивые исследования (Jo and Choi, 2015). Один из относительно недавних экспериментов с использованием флуоресцентной гибридизацией *in situ* (FISH) показал, что несущие интроны транскрипты преимущественно расположены в цитоплазме (Valencia, Dias and Reed, 2008), что сильнее подвергает сомнению высказанные ранее гипотезы и требует их пересмотра.

Помимо перечисленных функций есть также предположение о том, что часть последовательностей концов интронов может быть ответственна за истощение нуклеосом в интронах путем отталкивания нуклеосом в сторону экзонов (Schwartz, Meshorer and Ast, 2009). Это позволяет задуматься о роли интронов в сборке хроматина.

Стоит отметить порядковое положение интронов в гене. Больше всего внимание привлекает первый интрон среди всех интронов гена, обычно именно он обладает особыми функциональными характеристиками (Jo and Choi, 2015). Например, первый интрон гена *oskar Drosophila* играет важную роль в правильной цитоплазматической локализации некоторых мРНК (Siemens *et al.*, 2004).

Длина интронов также является важной характеристикой. Есть предположение о том, что длинные интроны повышают эффективность естественного отбора, устраняя интерференцию Хилла-Робертсона (Comeron, Williford and Kliman, 2008).

В одной из работ десятилетней давности на дрожжах *Saccharomyces cerevisiae* наблюдали, как короткие ORF могут эволюционировать в настоящие функциональные гены посредством своего рода непрерывного эволюционного процесса (Carvunis *et al.*, 2012). Это позволяет предположить, что интроны могут стать источником новых генов.

В наше время активно используется метод полногеномного поиска ассоциаций, сокращенно GWAS – genome-wide association study. В основе метода лежит сравнительный анализ однонуклеотидных вариантов (SNV – single-nucleotide variant), а точнее участков в геноме, которые схожи между собой по этим однонуклеотидным заменам, между индивидуумами (Smith, Steinbrekera and Dagle, 2019). Примечательно, что при картировании аллелей, ассоциированных с какими-то заболеваниями, место картирования соответствует именно интронным областям (Welter *et al.*, 2014). Данный факт является очередным подтверждением важности изучения интронов.

Систематика, общие проблемы

Систематическая биология, или филогенетический анализ, используется для восстановления эволюционной истории видов, генов или белков (Som, 2014). Она имеет фундаментальное значение для предоставления систематизированной информации о живом мире и проясняет, какие организмы населяют нашу планету, а также раскрывает основные краткосрочные и долгосрочные процессы, породившие все наблюдаемое разнообразие. За последние 50 лет систематика добилась значительного прогресса. За этот период были разработаны количественные способы построения классификаций, точные методы реконструкции филогении, а также появилось глубокое понимание эволюционных процессов на популяционном уровне (Stuessy, 2020).

Принимая во внимание весь вклад, внесенный систематиками и эволюционными биологами в познание филогенеза, почти все исследователи согласны с тем, что понимание этого процесса играет важную роль в классификации, однако вопрос о том, как именно филогенетические данные следует интегрировать в эту классификацию, до сих пор остается спорным (Stuessy, 2013).

Несмотря на огромный прогресс, достигнутый в последние годы, филогенетическая реконструкция сопряжена со многими проблемами, которые создают неопределенность в отношении истинных эволюционных взаимоотношений анализируемых видов или генов (Som, 2014).

Одной из наиболее заметных трудностей является широко распространенное несоответствие между методами, а также между отдельными генами или различными участками генома (Jeffroy *et al.*, 2006). Наличие широко распространенной неконгруэнтности

препятствует успешному выявлению эволюционных связей и применению филогенетического анализа (Galtier and Daubin, 2008).

Ожидалось, что появление и совершенствование молекулярных и геномных данных уменьшит проблему несоответствий, но в итоге увеличилось разнообразие классификаций, а проблема никак не уменьшилась (Jeffroy *et al.*, 2006). К сожалению, все еще остаются многие важные неразрешенные узлы (Galtier and Daubin, 2008).

На сегодняшний день с помощью филогеномного подхода уже устранены некоторые из негативных факторов, которые вызывают филогенетическое несоответствие, но это все еще остается непростой задачей. Например, при попытке устранения одного негативного фактора, можно ввести другой негативный фактор, вызывающий несоответствие. Помимо этого, допустимо существование неизвестных скрытых факторов, которые тоже оказывают негативное влияние (Som, 2014).

Для достижения основной цели восстановления дерева жизни с высоким разрешением одним из важных факторов является точное установление гомологии признаков, т.е. множественное выравнивание последовательностей (Som, 2014). Этот подход будет активно использоваться в данной работе.

В настоящее время особое внимание уделяется поиску лучших локусов, которые будут нести информацию о разных уровнях эволюционной истории. Желательны одиночные или малокопийные ядерные гены, потому что важным пунктом для филогенетических маркеров считается ортологичность локусов (Tekaia, 2016). Неоспоримо, поиск таких маркеров является актуальной задачей.

Семейство генов *Nxf*, общая характеристика

Принимая во внимание все выше сказанное, можно перейти к описанию основного объекта, на который направлено данное исследование. Перед этим стоит сказать несколько вводных слов про само семейство.

Семейство генов *Nxf* (nuclear export factor) было названо в честь функции универсального гена *Nxf1*, продукт которого ответственен за ядерно-цитоплазматический транспорт большинства мРНК. Гены данного семейства обнаружены у всех эукариотических организмов группы *Opisthokonta* и характеризуются эволюционной консервативностью (Mamon, Kliver and Golubkova, 2013).

Геномы различных грибов имеют лишь один ген-представитель этого семейства, растения и некоторые простейшие вовсе их лишены. Животные, как правило, имеют от двух до пяти паралогичных генов (Mamon, Kliver and Golubkova, 2013).

Характерной особенностью для гена *Nxf1* является существование транскриптов с невырезанным гомологичным интроном. Особое внимание выделяется консервативному блоку, состоящему из последовательно идущих экзонов размером 110 п. н. и 37 п. н., а также интрону, который находится между ними и который принято называть "кассетным" (Мамон *et al.*, 2013). Соответственно, сам блок из трех единиц мы будем называть "консервативной кассетой".

Среди животных, для которых есть данные про семейство генов *Nxf,* были выявлены 3 таксономические группы, и в каждой из них кассетный интрон имеет определенные характеристики (Мамон *et al.*, 2013).

К первой группе относятся позвоночные, для них характерен интрон, находящийся между 10-м и 11-м экзонами в гене *Nxf1*. В интроне позвоночных выявили 4 участка протяженной гомологии: 1) 5′-концевая последовательность; 2) фрагмент, содержащий конститутивный транспортный элемент — CTE (constitutive transport element), характерный для ретровирусов; 3) третий консервативный участок; 4) 3′-концевая последовательность (Мамон *et al.*, 2013).

Ко второй группе принадлежат дрозофилиды с кассетным интроном 5-м в генах *Nxf1*. Несмотря на отсутствие участков протяженной гомологии, что характерно для позвоночных, у разных видов *Drosophilidae* есть схожая черта – две протяженные последовательности поли(A). Предполагается, что вторичная структура транскриптов с сохраненным интроном, образовавшаяся за счет наличия этих последовательностей, способствует их сохранению (Мамон *et al.*, 2013).

В последнюю группу включены нематоды с интроном 5-м или 6-м. Его размер на порядок меньше, а участки протяженной гомологии не обнаружены. Однако, внутри группы есть общее сходство – это повышенное содержание тимина (T), что отличает данную группу организмов от остальных (Мамон *et al.*, 2013).

Как уже упоминалось ранее, важной особенностью для генов *Nxf1* у позвоночных и дрозофилид является способность к образованию сложных вторичных структур, которая может играть существенную роль в пост-транскрипционной судьбе транскрипта (Golubkova *et al.*, 2012). Помимо этого, все же главной особенностью, кроме сохраненного интрона в транскрипте, становится наличие в этом сохранившемся интроне преждевременного стоп-кодона, но мРНК избегает нонсенс-опосредованного распада, и в итоге возможно образование укороченного белка. Перейдем к подробному рассмотрению гена *Nxf1*.

Структура и функции Nxf1 (гена и белка)

Первоначально белок Nxf1 у людей был идентифицирован как потенциальный цитоплазматический кофактор для Tip (белок, взаимодействующий с тирозинкиназой), кодируемый вирусом герпеса saimiri, и был назван TAP (Tip-associated protein, или белок, ассоциированный с Tip) (Yoon *et al.*, 1997). Позже было показано, что TAP участвует в ядерно-цитоплазматическом транспорте несплайсированной или частично сплайсированной РНК ретровирусов. TAP напрямую распознает только одну последовательность – CTE, которая была первоначально обнаружена в РНК ретровирусов (Zolotukhin *et al.*, 2001). Мы предполагаем, что функцию CTE у дрозофилид выполняет образуемая благодаря сохранению интрона вторичная структура. С этого момента речь пойдет о гене *Nxf1 Drosophila melanogaster*, если не указывается иное, который также принято обозначать как *sbr* (small bristles).

Nxf1 имеет модульную доменную организацию (на Рис. 1 представлена структура гена, последовательности в котором кодируют белковую структуру), состоящую из РНК-связывающего домена (RBD), четырех богатых лейцином повторов (LRR), домена, проявляющего сходство с ядерным транспортным фактором 2 (NTF2-подобный домен), и С-концевой убиквитин-ассоциированный, или UBA-подобный, домен. Также в начале гена присутствуют сигналы ядерной локализации (NLS) (Mamon, Kliver and Golubkova, 2013).

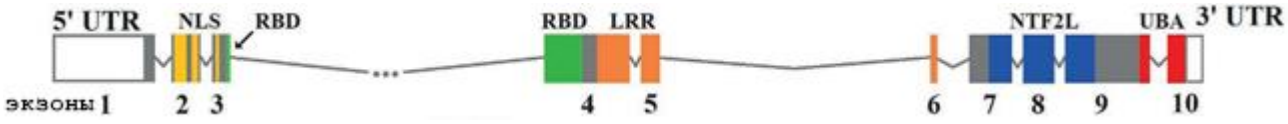


Рис. 1. Структура гена *sbr Drosophila melanogaster* (Mamon *et al.*, 2019).

RBD – RNA-Binding Domain, LRR – Leucine-Rich Repeats, NTF2L – Nuclear Transport Factor 2 Like, UBA – Ubiquitin Associated domain, NLS – Nuclear Localization Signal.

Цветом выделены области экзонов, соответствующие конкретным белковым доменам.

Основной и общеизвестной функцией продукта гена *sbr* является транспорт всех типов мРНК из ядра в цитоплазму (Herold *et al.*, 2000). Кассетный интрон условно разделяет *sbr* на две функциональные половины – рецепторную, куда относятся RBD и LRR, отвечающую за взаимодействие с РНК, и транспортную, куда входят NTF2L и UBA, позволяющую, взаимодействуя с другими белками, обеспечивать транспорт комплекса макромолекул через ядерные поры (Мамон *et al.*, 2013).

Помимо классической функции нашей группой выявлены дополнительные, жизненно важные функции продукта гена *sbr*. Во-первых, доказано, что sbr выполняет семенниково-специфические функции. Показано, что существуют специфичные для семенников альтернативные транскрипты, полученные за счет альтернативного промотора, этого гена и укороченный белок sbr, обнаруженный только в семенниках (Ginanova *et al.*, 2016).

Во-вторых, установлено, что белок sbr необходим для формирования внутренней структуры и установления границ в мозговом веществе зрительной системы дрозофилы. Распределение sbr в ядре и цитоплазме специфических нейронов и глиальных клеток свидетельствует о специализированных функциях этого белка (Mamon *et al.*, 2021).

В-третьих, в одной из самых последних работ была продемонстрирована значимость кассетного интрона в эволюции гена *Nxf1* у представителей *Chiroptera* (Bondaruk, Golubkova and Mamon, 2022).

Таким образом, уже на данном этапе изучения становится ясна важная роль гена *Nxf1*, а также кассетного интрона в нем. Тем не менее, очень интересно было бы проанализировать и описать данный ген у представителей других таксономических групп, о чем и пойдет речь далее.

Материалы и методы исследования

Использованные последовательности

В качестве материалов для исследования использовались нуклеотидные последовательности 89 видов *Arthropoda*, среди которых 38 принадлежат семейству *Drosophilidae*.

В таблице 1 представлен перечень латинских названий видов, идентификаторы последовательностей, а также координаты скачанных участков последовательностей, использованных в работе, для соответствующих видов, принадлежащих *Arthropoda*, но не входящих в состав семейства *Drosophilidae*. Те же виды данных для *Drosophilidae* представлены в таблице 2.

Таблица 1. Перечень видов из таксономической группы *Arthropoda*, последовательности гена *Nxf1* которых были взяты для исследования.

Название вида	**Идентификатор**	**Координаты**
Bactrocera dorsalis	NC_064306.1	c455567-438736
Ceratitis capitata	NW_019378534.1	c257647-230560
Anoplophora glabripennis	NW_019416348.1	c124436-99701
Aethina tumida	NC_065435.1	72246447-72252173
Pararge aegeria	NC_053197.1	c12908023-12892754
Maniola hyperantus	NC_048551.1	3773258-3780611
Papilio xuthus	NW_013530472.1	c3742132-3736391
Papilio polytes	NW_013525224.1	1911103-1918289
Zerene cesonia	NC_052117.1	2387943-2398756
Danaus plexippus plexippus	NC_045821.1	c6179130-6173317
Vanessa tameamea	NW_020663538.1	678406-687971
Ostrinia furnacalis	NW_021132278.1	c100539-87205
Manduca sexta	NC_051125.1	3598358-3609995
Orussus abietinus	NW_019394833.1	c12741-9931
Cephus cinctus	NW_014333043.1	49745-52556
Leptopilina heterotoma	NW_025111133.1	c2216556-2209975
Venturia canescens	NC_057428.1	18815406-18822855
Vespa mandarinia	NW_023395839.1	c2278118-2272556
Vespa crabro	NC_060964.1	c2569260-2562569
Vespa velutina	NC_062195.1	6256342-6263061

Vespula pensylvanica	NC_057696.1	2883545-2887115
Polistes fuscatus	NW_025113124.1	c231111-226443
Megalopta genalis	NW_022900912.1	426827-434757
Apis laboriosa	NW_025221248.1	c1030737-1026398
Bombus pyrosoma	NC_057770.1	847473-850748
Colletes gigas	NW_025106644.1	c5429460-5423735
Odontomachus brunneus	NW_022639455.1	2568099-2573031
Acromyrmex echinatior	NW_011627193.1	c742711-736930
Aphidius gifuensis	NC_057790.1	c27180051-27176323
Cotesia glomerata	NC_058164.1	c3430608-3423502
Cimex lectularius	NW_019392632.1	1292564-1308188
Halyomorpha halys	NW_020110579.1	449164-482789
Diuraphis noxia	NW_015368779.1	7569-18733
Myzus persicae	NW_019104011.1	510754-526117
Acyrthosiphon pisum	NC_042496.1	20015402-20023198
Melanaphis sacchari	NW_020271618.1	358729-383222
Rhopalosiphum maidis	NC_040878.1	7526748-7532514
Sipha flava	NW_020272985.1	c149705-143117
Cryptotermes secundus	NW_019726235.1	c79073-52417
Lepeophtheirus salmonis	NC_052135.1	c25814317-25810340
Homarus americanus	NW_024734940.1	12286796-12308559
Pollicipes pollicipes	NW_023596420.1	28951189-28965644
Amphibalanus amphitrite	NW_024882762.1	31093-46408
Limulus polyphemus	NW_013667881.1	c78152-26836
Centruroides sculpturatus	NW_019385444.1	47508-71193
Parasteatoda tepidariorum	NW_024970569.1	c29721-13525
Tetranychus urticae	NW_015449956.1	464842-468488
Ixodes scapularis	NW_024609836.1	29687589-29774934
Metaseiulus occidentalis	NW_003805473.1	c2121969-2117393
Varroa destructor	NW_019211457.1	c28702434-28693892
Varroa jacobsoni	NW_019213089.1	c120889-110531

В первом столбце указано латинское название вида. Во втором столбце указан идентификатор (ID). В третьем столбце указаны координаты участка последовательности,

использованного для работы. Жирным шрифтом указаны виды, для которых была построена модель вторичной структуры.

Таблица 2. Перечень видов из семейства *Drosophilidae*, последовательности гена *Nxf1* которых были взяты для исследования.

Название вида	**Идентификатор**	**Скачанный участок**	**QC**
Drosophila melanogaster	NC_004354.4	c10847092-10832752	100%
Drosophila simulans	JPYS01000001.1	c10187106-10173139	88 – 97 %
Drosophila mauritiana	NIGA01000006.1	c10628136-10614218	
Drosophila sechellia	NIFZ01000403.1	c10629007-10615117	
Drosophila erecta	QMER02000023.1	1390117-1404493	
Drosophila teissieri	JAEDAA020000001.1	c15106748-15091720	
Drosophila santomea	JAECZZ020000001.1	c19639894-19624893	
Drosophila orena	VCKV01000288.1	5874960-5889472	
Drosophila yakuba	JAEDAC020000001.1	c19885222-19870338	
Drosophila subpulchrella	JACOEE010000252.1	12811587-12827899	47 – 68 %
Drosophila pseudotakahashii	JAJJHV010000220.1	2192016-2208740	
Drosophila suzukii	WWNF01000007.1	2159428-2174900	
Drosophila fuyamai	JAECXW010000062.1	c3434513-3417461	
Drosophila eugracilis	JAECYE010002139.1	590361-605445	
Drosophila kurseongensis	JAECXX010000108.1	3042160-3058047	
Drosophila carrolli	JAECXJ010000002.1	c17409233-17393947	
Drosophila biarmipes	JANEAA010000003.1	c14630219-14615484	
Drosophila takahashii	AFFI02007482.1	c285093-268069	
Drosophila rhopaloa	JAECXI010000003.1	c6622730-6589663	
Drosophila gunungcola	JAMKOV010000047.1	238054-254586	
Drosophila elegans	JAECXL010000098.1	c18129033-18112363	
Drosophila oshimai	JAECWM010000001.1	c5534214-5519272	
Drosophila ficusphila	JAECXK010000196.1	c155120-137733	
Drosophila tani	VNJO01009917.1	76103-93049	
Drosophila pectinifera	VNKC01002694.1	c34231-19489	
Drosophila aff. chauvacae	JAEIGO010000756.1	3652980-3667471	
Drosophila punjabiensis	VNJR01004209.1	83836-98011	
Drosophila jambulina	JAJJHR010003625.1	8530604-8544902	

Drosophila burlai	JAEIIQ010030601.1	58835-73625	20 – 36 %
Drosophila bakoue	VNJL01005698.1	c32167-17808	
Drosophila mayri	VNJN01005011.1	c128161-115374	
Drosophila bocki	VNJY01000469.1	7073-21268	
Drosophila nikananu	VNJV01009601.1	7042-20752	
Drosophila birchii	VNKA01003542.1	588928-602755	
Drosophila vulcana	VNJP01007434.1	8888-23912	
Drosophila rufa	JAECXS010000111.1	c8932028-8916088	
Drosophila asahinai	VNJZ01004460.1	c63600-46922	
Drosophila lacteicornis	VNKF01011471.1	c36720-20762	

В первом столбце указано латинское название вида. Во втором столбце указан идентификатор (ID). В третьем столбце указаны координаты участка последовательности, использованного для работы. В четвертом столбце указан процент покрытия для исследуемых групп (Query Coverage – QC). Жирным шрифтом указаны виды, для которых была построена модель вторичной структуры.

Подготовка референсной последовательности гена *Nxf1*

В качестве референсной последовательности использовалась последовательность гена *Nxf1 (sbr)* модельного объекта *Drosophila melanogaster*, как наиболее полно изученная. Последовательность была взята из базы данных Genbank (https://www.ncbi.nlm.nih.gov/gene/).

Далее полная последовательность гена *sbr* была размечена в программе Microsoft Word (Рис. 2):

• оранжевым цветом жирным шрифтом выделены экзоны

• зеленым цветом жирным шрифтом подчеркнутым курсивом выделены экзоны из консервативной кассеты

• красным цветом жирным шрифтом курсивом выделен кассетный интрон

• желтой заливкой выделены поли(А)-последовательности внутри кассетного интрона

• нуклеотиды без форматирования соответствуют последовательностям других интронов

Этап разметки необходим для дальнейшего получения последовательности других представителей дрозофилид, содержащей только экзоны, а также получения участка консервативной кассеты.

Рис. 2. Пример разметки фрагмента последовательности гена *Nxf1* у *Drosophila melanogaster* в программе Microsoft Word.

Разметка, как в данном случае, так и в последующих, осуществлялась с помощью поиска и замены текста с использованием разработанных макросов для ускорения процесса обработки.

BLAST

BLAST проводился на базе NCBI.

В качестве входных данных (Query Sequence) использовалась полная последовательность гена *Nxf1 Drosophila melanogaster*.

Измененные параметры:

• Database – Whole-genome shotgun contigs (wgs)

• Limit by – Organism

Drosophilidae (taxid:7214)

Drosophila melanogaster (taxid:7227) exclude

• Word size – 16

Остальные настройки по умолчанию.

Последовательности, имеющие значение параметра покрытия запроса (**Q**uery **C**overage) **QC** < 10% не были включены в дальнейший анализ.

Сортировка результатов BLAST с помощью кластерного анализа

Кластерный анализ проводился в программе PAleontological STatistics (PAST) с настройками по умолчанию (Hammer, Harper and Ryan, 2001). По его результатам (Рис. 3), основываясь на параметре "Query Coverage", последовательности были разбиты на три группы: 1) 88-97%, 2) 46-68%, 3) 11-36%.

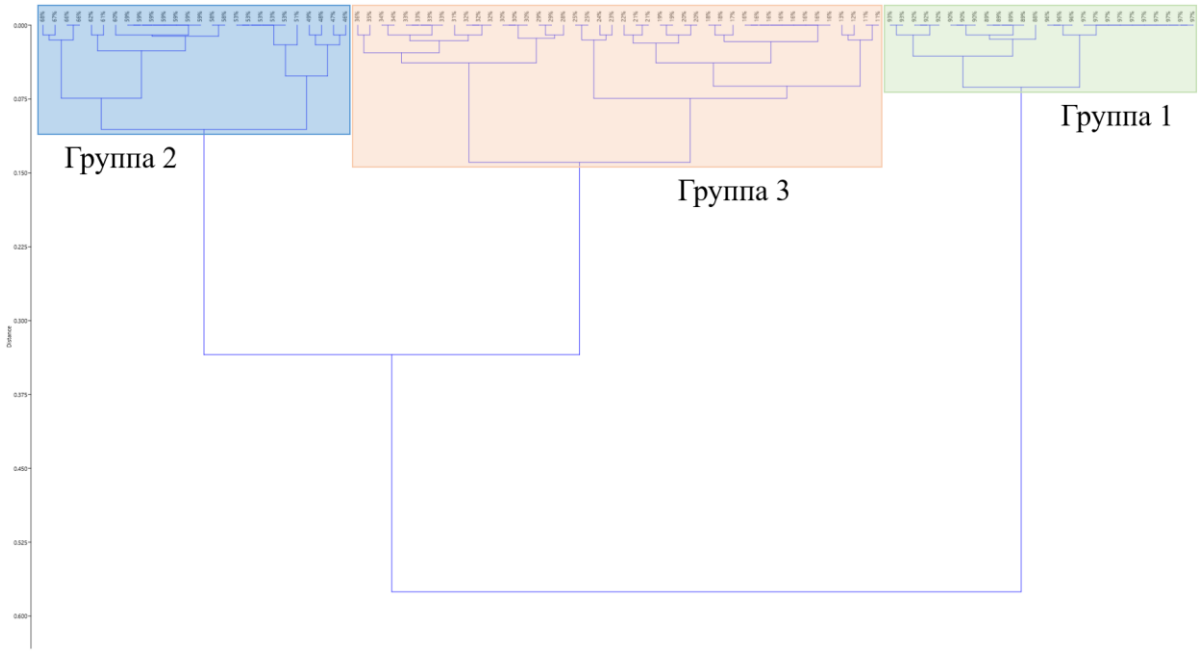


Рис. 3. Результаты кластерного анализа. Прямоугольниками разного цвета выделены значения для 3-х групп последовательностей.

Для дальнейшей обработки результатов было написано 2 скрипта на языке программирования Python 3. Для работы второго скрипта необходимы следующие

библиотеки: os, Entrez из Biopython. Исходный код скриптов с комментариями и протоколом находится в Приложении 1.

Результатом работы скрипта № 1 является полуавтоматическое формирования списка из названия вида, его идентификатора, предполагаемых координат гена *Nxf1*, которые были найдены с помощью BLAST, а также информации о том, какая цепь была использована для выравнивания в процессе BLAST. Например, "1. Drosophila tani isolate 14020-0011.00 (Sequence ID: VNJO01009917.1): 76103..93049 - Strand: Plus/Plus". Полуавтоматическое, потому что 1) можно выбрать ручной вариант ввода данных, 2) в определенных условиях скрипт для дальнейшей работы потребует ввода данных от пользователя.

Ген *Nxf1* у *Drosophila melanogaster*, как и у многих других дрозофилид (возможно, у всех, но проверка этого не проводилась) находится на X-хромосоме. Последовательность, взятая в качестве референса в базе данных NCBI обозначена как "inverted, complemented". Это означает, что на хромосоме она располагается комплементарно в обратном порядке. Однако, такое расположение гена может быть не характерно для анализируемых последовательностей, поэтому важно учитывать параметр "Strand" при их скачивании для последующего выравнивания.

Скрипт № 2 на основании данных, скомпилированных скриптом №1, извлекает из базы данных NCBI последовательности, соответствующие каждому виду из списка.

Далее необходимо было провести анализ скачанных последовательностей. Для одного и того же вида могут быть дублирующиеся результаты (это связано с выбором Database в настройках BLAST), поэтому все дубликаты необходимо было сравнить между собой и выбрать наилучшие. Выбор был сделан на основе выравниваний дубликатов между собой, а также с референсом. Благодаря этому список последовательностей значительно сократился, а также изменились значения QC для некоторых из сформированных групп. Теперь обозначение групп представлено следующим образом: 1) 88-97%, 2) 47-68%, 3) 20-36%.

После данной процедуры можно переходить к множественному выравниванию последовательностей.

Множественное выравнивание

Множественное выравнивание нуклеотидных последовательностей осуществлялось с помощью алгоритма MUSCLE (Manuel, 2013) в программе UNIPRO UGENE (Okonechnikov *et al.*, 2012) с использованием настроек по умолчанию (GAP Open 400, 16 итераций). В качестве метода кластеризации использовался UPGMA (Michener and Sokal, 1957).

Анализ качества выравниваний

Оценка степени выравнивания проводилась в программе MEGA-X (Kumar *et al.*, 2018). Анализировалось количество консервативных нуклеотидов на выравнивание. Каждому присвоено обозначение римской цифрой, сводная таблица представлена в Приложении 4.

Разметка генов *Nxf1* для представителей семейства *Drosophilidae*

По результатам выравниваний стало возможно разметить гены у других представителей дрозофилид, т. к. ранее для них не были обозначены экзоны. Помимо разметки экзонов, также была произведена разметка кассетного интрона и поли(A)-последовательностей в нем.

Разметка была произведена также с помощью макросов, как и для гена *Nxf1 Drosophila melanogaster*, с использованием программы Microsoft Word и замены текста на форматированную версию. Однако, ввиду использования данных из сторонней программы для большей экономии времени была использована программа Key Manager, которая специализируется на выполнении макросов.

Построение вторичных структур

Для построения вторичных структур была использована онлайн-программа RNAfold (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi) (Gruber *et al.*, 2008), настройки по умолчанию. Выделение нужных участков на вторичной структуре было реализовано также внутри программы. Поиск координат для выделения кассетного интрона или поли(A)-последовательностей внутри него осуществлялся, основываясь на результатах разметки (см. предыдущий раздел).

Методы, примененные для представителей *Arthropoda*, не включая семейство *Drosophilidae*

Для других представителей *Arthropoda* не проводился BLAST для поиска гена *Nxf1* внутри группы. Были взяты полные последовательности гена *Nxf1*, уже находящиеся в базе данных NCBI. Экзонные последовательности также были взяты оттуда.

Разметка, выравнивание и моделирование вторичных структур осуществлялось также, как и для семейства *Drosophilidae* (см. предыдущие разделы).

Результаты

Были проанализированы 89 нуклеотидных последовательностей гена *Nxf1* из таксономической группы *Arthropoda*, членистоногие, не считая исключенных из анализа.

Размечена нуклеотидная последовательность гена *Nxf1 Drosophila melanogaster*, которую можно использовать в качестве референса. Полная последовательность находится в Приложении 2. Последовательность, содержащая только экзоны и кассетный интрон между 5-м и 6-м экзонами, находится в Приложении 3.

Проведен BLAST и найдены последовательности 37 видов из семейства *Drosophilidae*, которые также были размечены (предоставляются по запросу).

Разработаны специализированные скрипты для обработки результатов BLAST.

Осуществлены множественные выравнивания для представителей семейства *Drosophilidae* (Рис. 4, Рис. 5, Рис. 6).

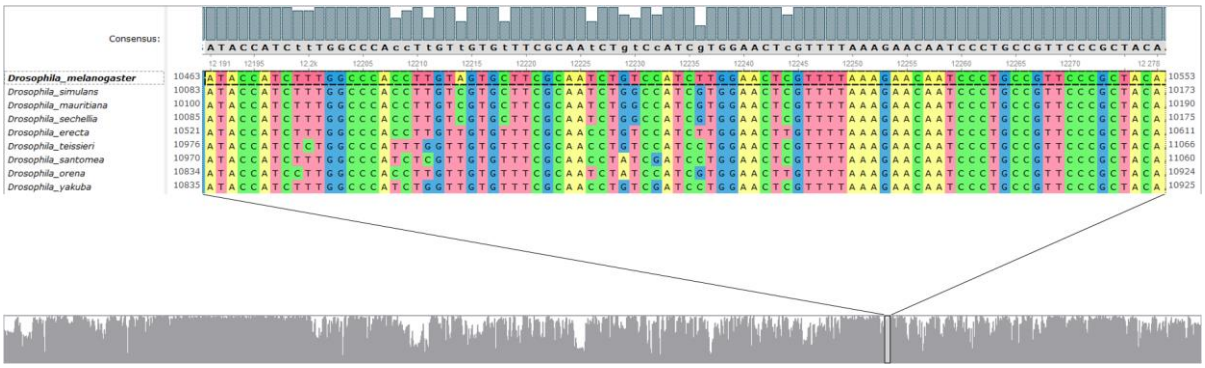


Рис. 4. Результат выравнивания последовательностей группы 1 (диапазон Query Coverage 88-97%). Полная последовательность гена *Nxf1*. Выравнивание I (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 16570 п. н. Серый прямоугольник соответствует началу 5-го экзона из консервативной кассеты.

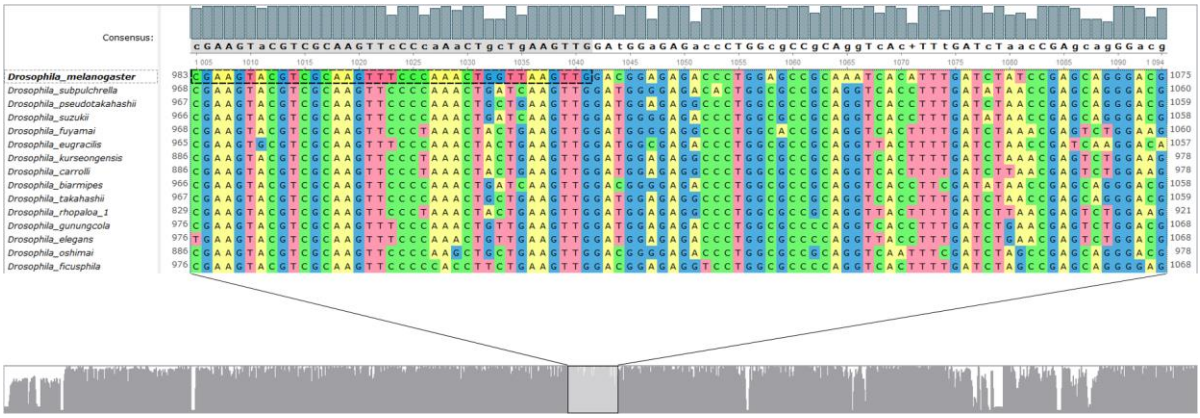


Рис. 5. Результат выравнивания последовательностей группы 2 (диапазон Query Coverage 47-68%). Вариант последовательности гена *Nxf1*, содержащей только экзоны. Выравнивание V (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 2124 п. н. Пунктиром обозначен шестой экзон. Серый прямоугольник – шестой экзон размером 37 п. н., а также частично седьмой экзон.

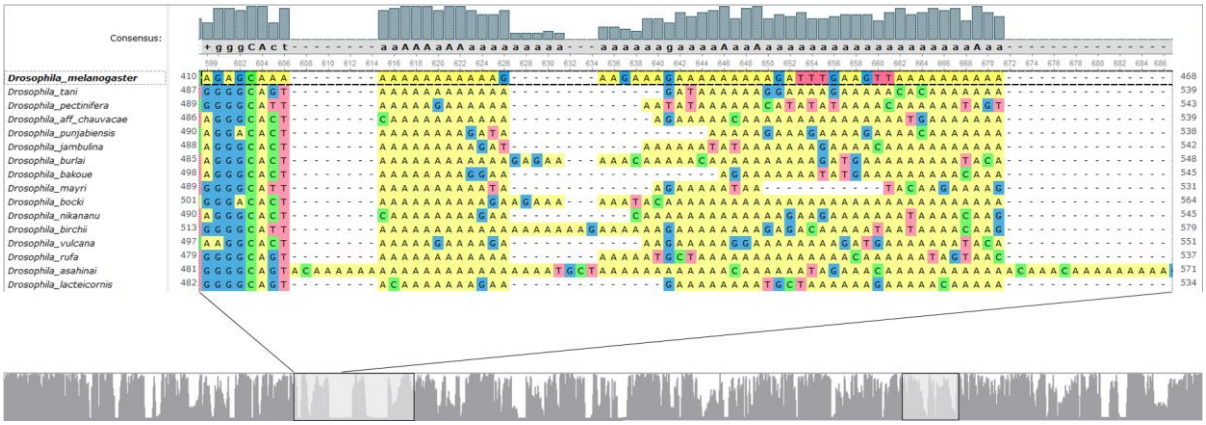


Рис. 6. Результат выравнивания последовательностей группы 3 (диапазон Query Coverage 20-36%). Консервативная кассета гена *Nxf1*. Выравнивание IX (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 2471 п. н. Серые прямоугольники соответствуют 2-м поли(А)-последовательностям.

Выравнивания для тех же участков других диапазонов представлены в Приложении 5.

Также проведены множественные выравнивания для других представителей *Arthropoda* (Рис. 7, Рис. 8).

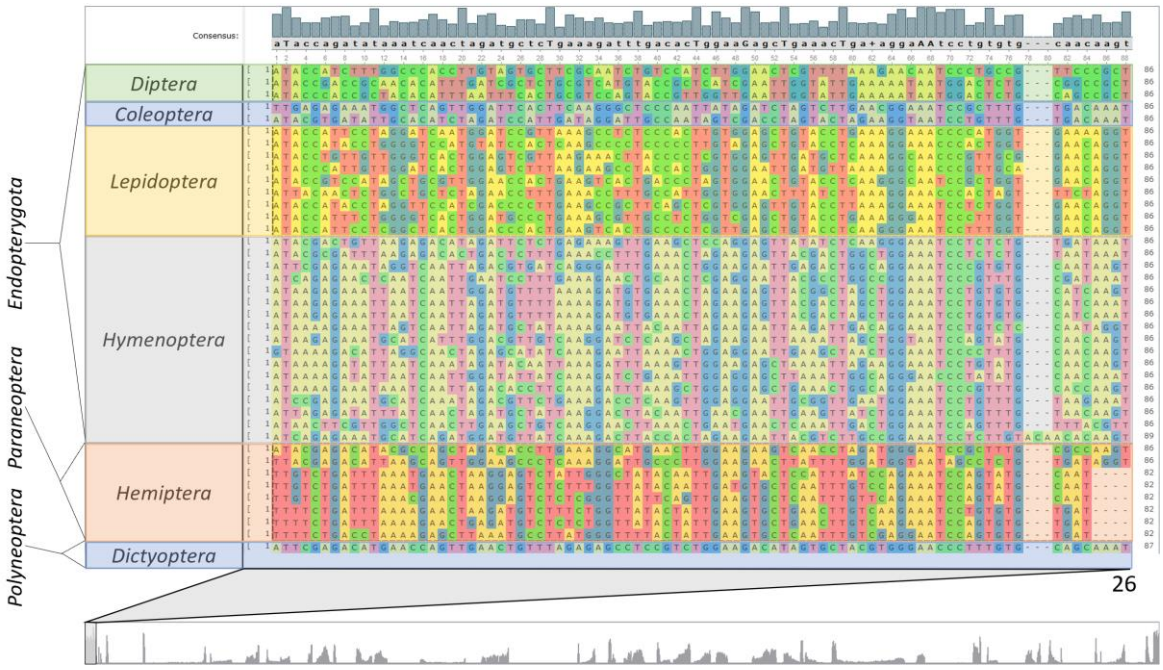


Рис. 7. Результат выравнивания последовательностей консервативной кассеты для представителей *Mandibulata – Hexapoda – Insecta*. Выравнивание X (см. Приложение 4).

Отдельные строки соответствуют отдельным видам. Первая строка соответствует нуклеотидной последовательности референса. Необработанный результат находится в Приложении 6.

Диаграмма для множественного выравнивания общей размерностью 8131 п. н. Серый прямоугольник на диаграмме соответствуют началу первого консервативного экзона из кассеты. Разным цветом выделены блоки последовательностей, характерные для разных отрядов. Деление на более высокие таксоны соответствует информации из базы данных NCBI (Taxonomy).

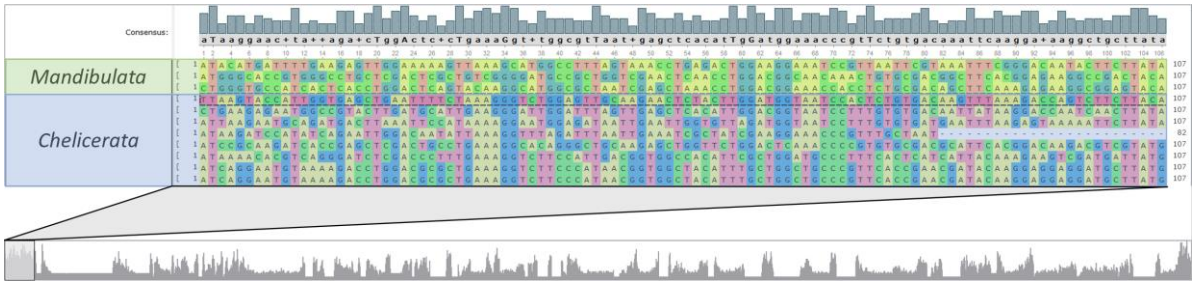


Рис. 8. Результат выравнивания последовательностей консервативной кассеты для других представителей *Arthropoda*. Выравнивание XI (см. Приложение 4).

Отдельные строки соответствуют отдельным видам. Необработанный результат находится в Приложении 7.

Диаграмма для множественного выравнивания общей размерностью 4124 п. н. Серый прямоугольник на диаграмме соответствуют началу первого консервативного экзона из кассеты. Разным цветом выделены блоки последовательностей, характерные для разных таксономических групп (деление соответствует информации из базы данных NCBI (Taxonomy)).

Были построены модели вторичных структур для 4-х представителей семейства *Drosophilidae* (Рис. 9): *D. melanogaster*, *D. birchii*, *D. biarmipes*, *D. yakuba*.

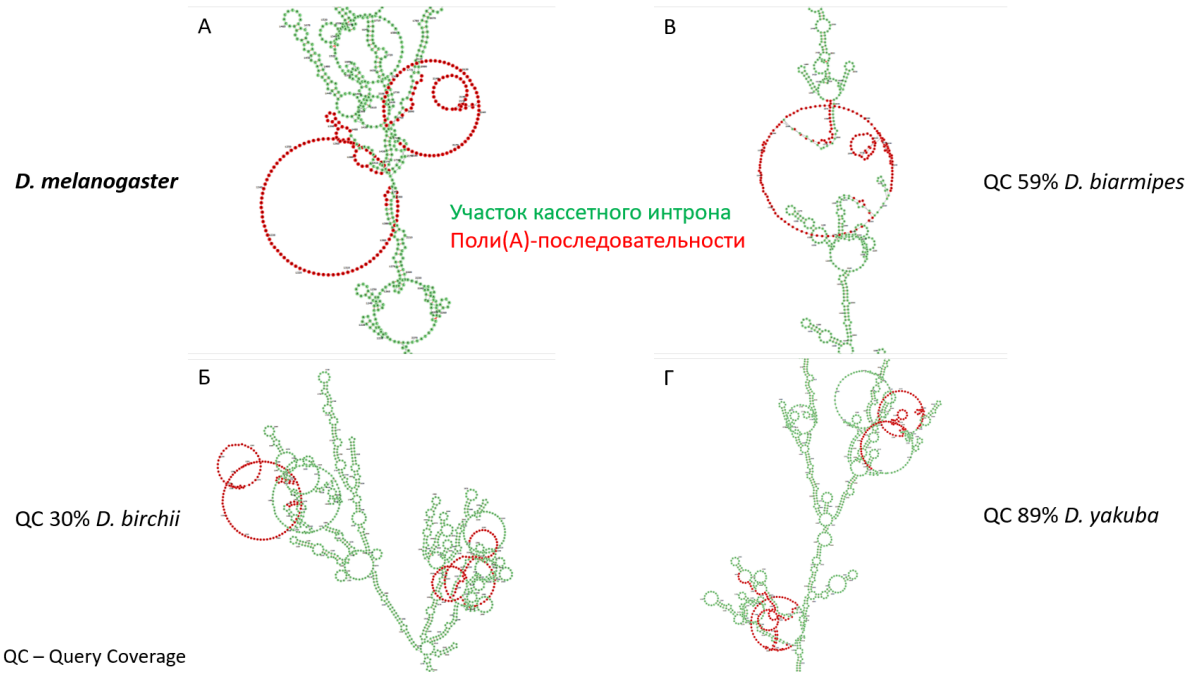


Рис. 9. Визуализация вторичной структуры интрон-содержащего транскрипта гена *Nxf1* дрозофилид.

Зеленым цветом обозначен участок кассетного интрона. Красным цветом выделены поли(А)-последовательности внутри кассетного интрона.

Выбор видов для построения вторичных структур осуществлялся случайным образом по одному виду из каждой полученной группы.

Построены модели вторичных структур для всех проанализированных представителей отряда *Diptera*, как наиболее эволюционно близких для референса (Приложение 8).

Кроме этого, модели вторичных структур также построены для других 16-ти представителей *Arthropoda* (Приложение 9).

Обсуждение

Интересно, что у всех проанализированных видов размер второго экзона из консервативной кассеты равен 37 нуклеотидов, в то время как размер первого экзона из кассеты варьирует в основном от 107 п. н. (отряд *Hemiptera*) до 113 п. н. (отряд *Hymenoptera*). Самым распространенным размером является 110 п. н., что характерно для всех представителей семейства *Drosophilidae*, а также для остальных упомянутых видов из разных групп. Такой же размер экзона характерен и для других животных, например, млекопитающих.

Также встречались уникальные варианты размера первого экзона из кассеты. Например, для *Tetranychus urticae* размер экзона составляет 122 нуклеотида. Т. к. он принадлежит группе *Scorpiones* и является единственным представителем данной группы в работе, то сделать выводов по поводу структуры последовательности не представляется возможным.

Множественные выравнивания для представителей *Arthropoda*, не включая семейство *Drosophilidae*, были проведены только для участка консервативной кассеты, т. к. количество и качество сборок экзонов для представителей разных групп отличается.

Также множественное выравнивание видов, не принадлежащих к насекомым, было проведено без вида *Homarus americanus*, ввиду размера интрона у данного вида, из-за чего качество выравнивания сильно ухудшается.

Для вида *Drosophila rhopaloa* характерна дупликация некоторых участков гена, что отображено в Приложении 10. По результатам анализа было принято решение взять вторую половину полной последовательности для множественного выравнивания дрозофилид, т. к. при сравнении с первой половиной у нее наблюдается выравнивание большего сходства на референсную последовательность *Drosophila melanogaster*.

По результатам множественных выравниваний можно заключить, что при рассмотрении семейства *Drosophilidae*, внутри него наблюдается высокая степень сходства полной последовательности гена *Nxf1*, которая включает все экзоны и интроны. Особая консервативность характерна для варианта гена *Nxf1*, содержащего только экзоны, что логично. Однако, помимо этого высокая степень сходства также характерна и для кассетного интрона, что представляет собой очень интересный факт, потому что интроны, как известно, эволюционируют с гораздо большей скоростью, нежели экзоны, а мы можем наблюдать не только хорошее выравнивание (Рис. 6), но и наличие у всех видов протяженных поли(А)-

последовательностей, которые, как мы предполагаем, играют ключевую роль в избегании транскриптом с сохраненным интроном нонсенс-опосредованного распада за счет формирования особых вторичных структур. Данные структуры можно наблюдать на Рис. 9.

При рассмотрении остальных представителей группы *Arthropoda* можно также наблюдать высокую степень сходства экзонов из консервативной кассеты. Несмотря на то, что рассматриваемые виды на эволюционном древе находятся на большом расстоянии друг от друга, все равно можно видеть схожую структуру последовательностей экзонов. Однако, также можно видеть, что последовательности внутри более мелких таксономических групп проявляют большую степень сходства. На основании этого можно заключить, что в определенных таксономических группах есть особый паттерн строения экзонов из консервативной кассеты. Можно предположить использование участков кассеты в качестве филогенетического маркера, но перед этим требуется тщательный анализ структур паралогичных генов.

Если рассматривать кассетный интрон у артропод, не относящихся к дрозофилидам, также можно наблюдать выровненные участки (Рис. 7, Рис. 8). У некоторых видов в составе этих участков находятся поли(А)-повторы, которые характерны для дрозофилид (например, виды из отряда *Diptera*), но в основном для этих участков характерна структура в виде АТ-повторов (аденин, тимин). Удивительно, что данные участки в большинстве своем соответствуют характерным выпетливаниям на моделях вторичных структур (Рис. 10), что является еще одним подтверждением значимости образования вторичных структур для транскриптов исследуемого гена.

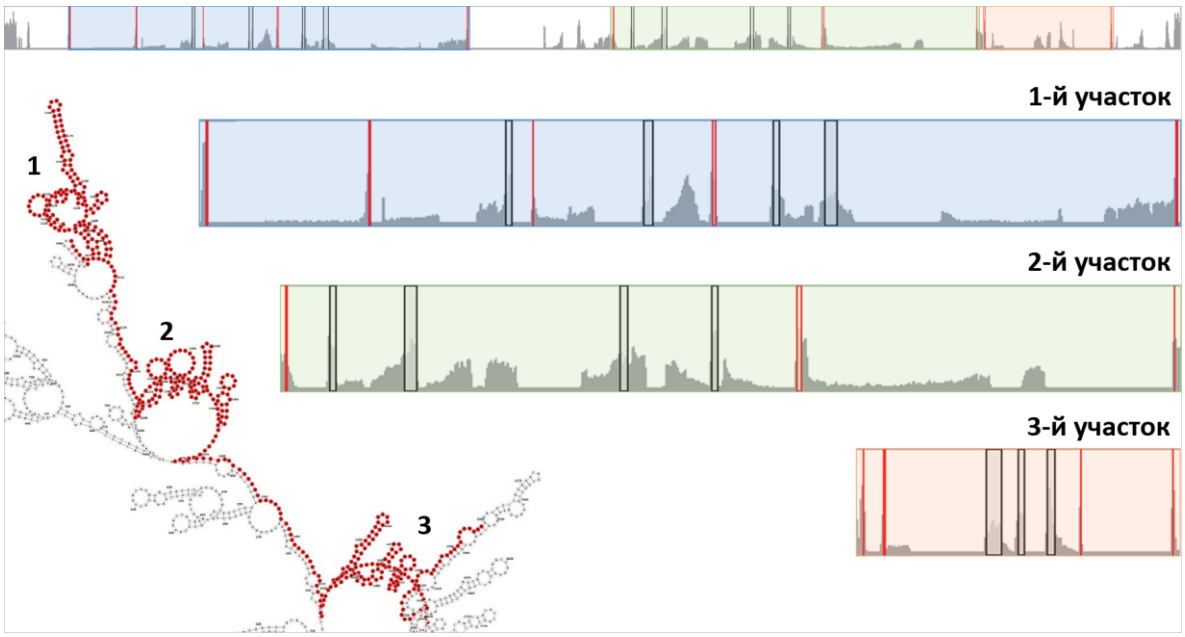


Рис. 10. Проекция специфических участков вторичной структуры внутри кассетного интрона на множественное выравнивание X.

В качестве примера была использована вторичная структура *Odontomachus brunneus*. Залитые цветом участки на диаграмме соответствуют трем специфическим участкам. Ниже представлены те же участки, но в увеличенном масштабе. Красные и черные прямоугольники внутри залитых участков соответствуют частям нуклеотидной последовательности внутри кассетного интрона *Odontomachus brunneus*.

Можно видеть, что участки кассетного интрона для рассматриваемого вида соответствуют выровненным участкам с наибольшей степенью сходства.

Выводы

По результатам проведенной работы были сформулированы следующие выводы:

1) Структура "консервативной кассеты" является специфичной для всех исследованных видов в данной работе, относящихся к таксономической группе *Arthropoda*. Для разных клад характерно наличие специфических консервативных последовательностей.

2) Интрон-содержащие транскрипты всех изученных видов членистоногих в данной работе образуют специфические вторичные структуры.

Список использованной литературы

Bondaruk, D. D., Golubkova, E. V. and Mamon, L. A. (2022) 'Contribution of the intron retained in the Nxf1 gene transcript to the phylogeny of the order Chiroptera', *Ecological Genetics*, 20(2), pp. 73–88. doi: 10.17816/ECOGEN90940.

Burge C.B., Tuschl T., Sharp P.A., 1999. Splicing of precursors of mRNAs by the spliceosomes. The RNA World / Eds. Gesteland R. F. et al. Cold Spring Harbor Lab. Press. Cold Spring Harbor., NY. P. 525–560.

Carvunis, A. *et al.* (2012) 'Proto-gene and de novo gene brith', *Nature*, 487(7407), pp. 370–374. doi: 10.1038/nature11184.Proto-genes.

Comeron, J. M., Williford, A. and Kliman, R. M. (2008) 'The Hill-Robertson effect: Evolutionary consequences of weak selection and linkage in finite populations', *Heredity*, 100(1), pp. 19–31. doi: 10.1038/sj.hdy.6801059.

Crick, F. (1970) '© 1970 Nature Publishing Group', *Nature Publishing Group*, 228, pp. 726–734.

Galtier, N. and Daubin, V. (2008) 'Dealing with incongruence in phylogenomic analyses', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512), pp. 4023–4029. doi: 10.1098/rstb.2008.0144.

Ginanova, V. *et al.* (2016) 'Testis-specific products of the Drosophila melanogaster sbr gene, encoding nuclear export factor 1, are necessary for male fertility', *Gene*, 577(2), pp. 153–160. doi: 10.1016/j.gene.2015.11.030.

Golubkova, E. *et al.* (2012) 'The evolutionarily conserved family of nuclear export factor (NXF) in drosophila melanogaster', *Drosophila Melanogaster: Life Cycle, Genetics and Development*, (December 2015), pp. 63–82.

Gruber, A. R. *et al.* (2008) 'The Vienna RNA websuite.', *Nucleic acids research*, 36(Web Server issue), pp. 70–74. doi: 10.1093/nar/gkn188.

Gruss, P. *et al.* (1979) 'Splicing as a requirement for biogenesis of functional 16S mRNA of simian virus 40', *Proceedings of the National Academy of Sciences of the United States of America*, 76(9), pp. 4317–4321. doi: 10.1073/pnas.76.9.4317.

Hammer, Ø., Harper, D. A. T. and Ryan, P. D. (2001) 'PAST : Paleontological Statistics Software Package for Education and Data Analysis, *Palaeontologia Electronica*, 4(1), pp. 1–9.

Herold, A. *et al.* (2000) 'TAP (NXF1) Belongs to a Multigene Family of Putative RNA Export Factors with a Conserved Modular Architecture', *Molecular and Cellular Biology*, 20(23), pp. 8996–9008. doi: 10.1128/mcb.20.23.8996-9008.2000.

Jeffroy, O. *et al.* (2006) 'Phylogenomics: the beginning of incongruence?', *Trends in Genetics*, 22(4), pp. 225–231. doi: 10.1016/j.tig.2006.02.003.

Jo, B.-S. and Choi, S. S. (2015) 'Introns: The Functional Benefits of Introns in Genomes', *Genomics & Informatics*, 13(4), p. 112. doi: 10.5808/gi.2015.13.4.112.

Juneau, K. *et al.* (2006) 'Introns regulate RNA and protein abundance in yeast', *Genetics*, 174(1), pp. 511–518. doi: 10.1534/genetics.106.058560.

Jurica, M. S. and Roybal, G. A. (2013) 'RNA splicing', *Encyclopedia of Biological Chemistry: Second Edition*, pp. 185–190. doi: 10.1016/B978-0-12-378630-2.00674-5.

Kalyna, M. *et al.* (2012) 'Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis', *Nucleic Acids Research*, 40(6), pp. 2454–2469. doi: 10.1093/nar/gkr932.

Kumar, S. *et al.* (2018) 'MEGA X: Molecular evolutionary genetics analysis across computing platforms', *Molecular Biology and Evolution*, 35(6), pp. 1547–1549. doi: 10.1093/molbev/msy096.

Mamon, L. *et al.* (2019) 'Organ-specific transcripts as a source of gene multifunctionality: Lessons learned from the Drosophila melanogaster sbr (Dm nxf1) gene', *Biological Communications*, 64(2), pp. 146–157. doi: 10.21638/spbu03.2019.206.

Mamon, L. *et al.* (2021) 'The RNA-binding protein SBR (Dm NXF1) is required for the constitution of medulla boundaries in Drosophila melanogaster optic lobes', *Cells*, 10(5). doi: 10.3390/cells10051144.

Mamon, L. A., Kliver, S. F. and Golubkova, E. V. (2013) 'Evolutionarily conserved features of the retained intron in alternative transcripts of the nxf1 (nuclear export factor) genes in different organisms', *Open Journal of Genetics*, 03(03), pp. 159–170. doi: 10.4236/ojgen.2013.33018.

Manuel, M. (2013) 'A new semi-subterranean diving beetle of the Hydroporus normandi-complex from south-eastern France, with notes on other taxa of the complex (Coleoptera: Dytiscidae)', *Zootaxa*, 3652(4), pp. 453–474. doi: 10.11646/zootaxa.3652.4.4.

Michener, C. D. and Sokal, R. R. (1957) 'A Quantitative Approach to a Problem in Classification', *Evolution*, 11(2), p. 130. doi: 10.2307/2406046.

Okonechnikov, K. *et al.* (2012) 'Unipro UGENE: A unified bioinformatics toolkit', *Bioinformatics*, 28(8), pp. 1166–1167. doi: 10.1093/bioinformatics/bts091.

Prusiner, S. B. (1998) 'Prions', *Proceedings of the National Academy of Sciences of the United States of America*, 95(23), pp. 13363–13383. doi: 10.1073/pnas.95.23.13363.

Roy, S. W. and Gilbert, W. (2006) 'The evolution of spliceosomal introns: Patterns, puzzles and progress', *Nature Reviews Genetics*, 7(3), pp. 211–221. doi: 10.1038/nrg1807.

Schmitz, U. *et al.* (2017) 'Intron retention enhances gene regulatory complexity in vertebrates', *Genome Biology*, 18(1), pp. 1–15. doi: 10.1186/s13059-017-1339-3.

Schwartz, S., Meshorer, E. and Ast, G. (2009) 'Chromatin organization marks exon-intron structure', *Nature Structural and Molecular Biology*, 16(9), pp. 990–995. doi: 10.1038/nsmb.1659.

Siemens, J. *et al.* (2004) ‘Cadherin 23 Is a component of the tip link in hair-cell stereocilla’, *Nature*, 428(6986), pp. 950–955. doi: 10.1038/nature02483.

Smith, C. J., Steinbrekera, B. and Dagle, J. M. (2019) *Genetic Basis of Patent Ductus Arteriosus*. Third Edit, *Hematology, Immunology and Genetics*. Third Edit. Elsevier. doi: 10.1016/b978-0-323-54400-9.00012-6.

Som, A. (2014) ‘Causes, consequences and solutions of phylogenetic incongruence’, *Briefings in Bioinformatics*, 16(3), pp. 536–548. doi: 10.1093/bib/bbu015.

Stuessy, T. F. (2013) ‘Schools of data analysis in systematics are converging, but differences remain with formal classification’, *Taxon*, 62(5), pp. 876–885. doi: 10.12705/625.12.

Stuessy, T. F. (2020) ‘Challenges facing systematic biology’, *Taxon*, 69(4), pp. 655–667. doi: 10.1002/tax.12279.

Tekaia, F. (2016) ‘Inferring orthologs: Open questions and perspectives’, *Genomics Insights*, 9, pp. 17–28. doi: 10.4137/GEI.S37925.

Valencia, P., Dias, A. P. and Reed, R. (2008) ‘Splicing promotes rapid and efficient mRNA export in mammalian cells’, *Proceedings of the National Academy of Sciences of the United States of America*, 105(9), pp. 3386–3391. doi: 10.1073/pnas.0800250105.

Welter, D. *et al.* (2014) ‘The NHGRI GWAS Catalog, a curated resource of SNP-trait associations’, *Nucleic Acids Research*, 42(D1), pp. 1001–1006. doi: 10.1093/nar/gkt1229.

Yoon, D. W. *et al.* (1997) ‘Tap: A novel cellular protein that interacts with tip of herpesvirus saimiri and induces lymphocyte aggregation’, *Immunity*, 6(5), pp. 571–582. doi: 10.1016/S1074-7613(00)80345-3.

Zolotukhin, A. S. *et al.* (2001) ‘Retroviral Constitutive Transport Element Evolved from Cellular TAP(NXF1)-Binding Sequences’, *Journal of Virology*, 75(12), pp. 5567–5575. doi: 10.1128/jvi.75.12.5567-5575.2001.

Кребс, Д., Голдштейн, Э. и Килпатрик, С. *Гены по Льюину*. Лаборатория знаний, 2017, 919 стр.

Мамон, Л. А. *et al.* (2013) ‘Интрон-содержащий транскрипт — эволюционно-консервативная особенность генов-ортологов nxf1 (nuclear export factor)’, *Генетика популяций и эволюция*.

Благодарности

Я хотел бы поблагодарить моего научного руководителя, к. б. н. Голубкову Елену Валерьевну, за предоставление возможности работы в ее научной группе, несмотря на чрезвычайно малое количество времени, оставшееся на выполнение и написание работы с момента вступления в нее, а также за постоянную поддержку и помощь в обсуждении результатов работы.

Отдельно я хотел бы поблагодарить своего куратора, Бондарука Дмитрия Денисовича, за круглосуточное предоставление советов в процессе выполнения поставленных задач.

Также хочу выразить благодарность коллективу кафедры генетики и биотехнологии СПбГУ за объективные советы на лабораторных семинарах и предоставленные знания за все годы обучения.

Приложения

Приложение 1

```
# СКРИПТ №1

# Функция добавления элементов в список
def listed_list(l, line):
    ap_line = line[:len(line) - 1]
    l.append([ap_line[:ap_line.find(':') + 1],
                        ap_line[ap_line.find(':') + 2:ap_line.find('to') - 1],
                        ap_line[ap_line.find('to') + 3:]])

# Функция сортировки по второму элементу в списке
def custom_key(range_list):
    return int(range_list[1])

# Ввод данных для анализа вручную / автоматически
input_file = input('Введите название файла без расширения: ') + '.txt'
Automate = input('Проанализировать результаты BLAST`а автоматически или ввести вручную? (1/2): ')
if int(Automate) == 2:
    species_list = input('Введите названия видов через точку с запятой: ').split(';')
else:
    species_list = []
    r1, r2 = input('Введите диапазон Query Coverage через пробел (от 0 до 101): ').split()
    with open(input_file) as inf:
        for line in inf:
            if line.find('Alignments:') != -1:
                break
            elif line.find('%') != -1:
                if int(line[line.find('%') - 2:line.find('%')]) in range(int(r1), int(r2)):
                    if line.find(',') > -1:
                        species_list.append(line[:line.find(',')])
                    else:
                        species_list.append(line[:line.find('...')])

# Запись в список в виде списков с помощью функции сразу нескольких видов
UGene = []
ID_list = []
Range_list_pairs = []
Strand_list = []

for species in species_list:
    record = False
    cont = True
    range_list = []
    ID = ''
    Strand = ''

    with open(input_file) as inf:
        for line in inf:
            if record == True:
                if ID == '':
                    ID = line[:line.find('Length') - 1]
                if Strand == '':
                    pos = line.find('Strand')
                    if pos != -1:
                        Strand = line[pos:len(line) - 1]
                        # range_list[0].insert(3, Strand)
                if line[:8] == 'Range 1:' and cont == False:
                    break
                elif line[:8] == 'Range 1:' and cont == True:
                    listed_list(range_list, line)
                    cont = False
                    continue
                elif line[:5] == 'Range':
                    listed_list(range_list, line)
            if line[:len(species) + 1] == '>' + species:
                record = True

    # Сортировка списка по возрастанию по второму элементу (первому числовому значению)
    range_list.sort(key = custom_key)
```

```
    # Автоматическое заполнение координат, если разница между 1-и и 2-м значением в диапазоне от 10.000 до 20.000
    n1 = range_list[0][1]
    n2 = range_list[-1][2]
    if int(n2) - int(n1) not in range(10000, 20000):

        # Вывод результата сортировки, если разница между координатами меньше 10.000 или больше 20.000 нуклеотидов
        print('\nОтсортированные значения для ' + species + ':')
        for Range in range_list:
            print(*Range)

        # Создание списка, в котором будут пары (вид, координаты)
        n1, n2 = input('Введите 1-ю и 2-ю координату через пробел для ' + species + ': ').split()

    UGene.append([str(len(UGene) + 1) + '. ' + species + ' (' + ID + ')' + ': ' + n1 + '..' + n2 + ' - ' + Strand])
    ID_list.append(ID[ID.find(':') + 2:])
    Range_list_pairs.append([n1, n2])
    Strand_list.append(Strand[Strand.find(':') + 2:])

# Вывод данных на экран
print('\nКоординаты, ID и направление цепи для исследуемых видов:')
for coord in UGene:
    print(*coord)

print('\nПоследовательный список ID: ')
print(*ID_list)

# Сохранение результатов работы скрипта в файл
save = input('\nСохранить эти результаты в отдельный файл формата.txt? (да/нет): ')
if save == 'да':
    output_file = input('Введите название файла без расширения: ')

    with open(output_file + '.txt', 'w') as ouf:
        ouf.write('Координаты, ID и направление цепи для исследуемых видов:\n')
        for coord in UGene:
            ouf.write(*coord)
            ouf.write('\n')

        ouf.write('\nПоследовательный список ID:\n')
        for i in ID_list:
            ouf.write(i + ' ')

        ouf.write('\n')
        ouf.write('\nНазвания для экспортированных участков:\n')
        q = 0
        while q < len(UGene):
            name_for_UGene = UGene[q][0].split()[:3]
            ouf.write(name_for_UGene[1] + ' ' + name_for_UGene[2] + ' ' + ID_list[q] + '\n')
            q += 1
```

```
# СКРИПТ №2

# Создание файлов по полученным координатам

import os
from Bio import Entrez
Entrez.email = 'artemvaskaa@gmail.com' # для работы библиотеки необходимо указывать e-mail

# Загрузка файлов на локальный компьютер
dir_name = input('В какую директорию загрузить файлы? (1 - в текущую директорию, 2 - создать новую папку): ')
if int(dir_name) == 1:
    dir_name = os.getcwd()
else:
    name_of_folder = 'Results for ' + r1 + '-' + str(int(r2) - 1) # название новой директории будет в виде "Result for x-y"
    dir_name = os.getcwd() + '\\' + name_of_folder
    if not os.path.exists(dir_name):
        os.mkdir(name_of_folder)

# Загрузка файлов из NCBI по переданным данным
q = 0
while q < len(ID_list):
    if Strand_list[q] == 'Plus/Plus':
        Strand = 1
    else:
        Strand = 2
    filename = UGene[q][0].split()[:3][1] + ' ' + UGene[q][0].split()[:3][2] + ' ' + ID_list[q] + ' nxf1.fas'

    if not os.path.isfile(dir_name + '\\' + filename):
        # название файлов будут в виде "название вида_номер депонирования_nxf1.fas", где _ является пробелом
        # загрузка...
        net_handle = Entrez.efetch(db = 'nucleotide',
                                   id = ID_list[q],
                                   strand = Strand,
                                   seq_start = Range_list_pairs[q][0],
                                   seq_stop = Range_list_pairs[q][1],
                                   rettype = 'fasta',
                                   retmode = 'text')
        out_handle = open(dir_name + '\\' + filename, 'w')
        out_handle.write(net_handle.read())
        out_handle.close()
        net_handle.close()
        print(filename + ' сохранен')
    else:
        print(filename + ' уже существует')
    q += 1

print('Работа выполнена!')
```

Изначально первый скрипт разрабатывался для последующей обработки в программе UNIPRO UGENE, поэтому в переменных присутствует название данной программы.

Для правильной работы скрипты должны запускаться последовательно. Перед запуском скриптов необходимо провести подготовку:

1) Таблицу результатов BLAST отсортировать по параметру "Query Cover" (покрытие запроса, Query Coverage, QC). Виды с QC < 10% не включаются в дальнейший анализ (необходимо снять галочки с таких значений перед скачиванием)
2) Результаты сохранить в текстовый файл (необходимо нажать Download → Text в верхней строке таблицы, при использовании ссылки Download All в результаты также будут включены значения с QC < 10%, что негативно скажется на дальнейшей обработке)
3) Скачанный текстовый файл формата .txt можно переименовать при желании (название файла будет использовано в дальнейшей работе скрипта)
4) Переместить файл формата .txt в директорию со скриптом (важно для работы скрипта)

5) Просмотреть результаты BLAST, и если в столбце Description есть одинаковые названия – сделать так, чтобы они отличались во всех частях файла (не использовать знаки "," или "..." – важно для работы скрипта)
6) Вернуть структуру таблицы в исходный вид с помощью удаления / добавления пробелов (важно для работы скрипта)

После подготовки необходимо запустить скрипт № 1 и следовать его запросам. После завершения его работы необходимо запустить скрипт № 2 и также следовать его запросам.

Приложение 2

Размеченная последовательность гена *Nxf1 Drosophila melanogaster*:

• оранжевым цветом жирным шрифтом выделены экзоны

• зеленым цветом жирным шрифтом подчеркнутым курсивом выделены экзоны из консервативной кассеты

• красным цветом жирным шрифтом курсивом выделен кассетный интрон

• желтой заливкой выделены поли(А)-последовательности внутри кассетного интрона

• нуклеотиды без форматирования соответствуют последовательностям других интронов

>NC_004354.4:c10847092-10832752 Drosophila melanogaster chromosome X
GATAGACCTGTCAGAAATTAAGAAAACAATATCAACATCAAGCGAATATACTTCGCTCGCTACATATTTCTAGT
TAGTTTTCGTAACTTTTTCGCTTTGGTGGCAATTAAATCAAATTTCCTCGAAGAATAACTTGTAAAAAGGTCGC
AAACATCCGTCCAGTTTCTGCCAACCATCGGCCACTGGCTATCGAAAGAAGATCGCTAGTGAATTAAATTTCAC
AAAATTGCGTTGTACTTTGGCCGTGGAAATTTGCCAAAAAACAGCAGCTACTTTGCTGCTATATTAGCAAATAG
GTTCAGGAAGTCGAGAAAAATTTTTCCAGAAGTTGGCAGCAGTTTGTGTGAATGCAAAAAATACATCGCACATA
GTTTTTTGCCCTCAAGACAGTTTTCTCAGTAATAACAACAGCACCCACACAAACACACAACACCACTCAGCTAA
AAACGTGCTGAAATTCGCATATCTTTTTGTTGCATACGTTTTTGAGTGAATATTAGTAAG**ATGCCCAAACGCGG**
CGGTGGCAGTAGCCAGCGGTACAACAACAACGTTGGAAATGGCGGCGGACGTTACAACGCTCCCGAGGATTTCG
ATGATTTTGGTCAGTACACGGGAGTGCAAGAAAATGAACCCTCCAGTAACCTAAATATAATGTATTCCTATAG**A**
TGTGGAGGATCGCCAGCGACGCAAGGATCGAAACAAGCGGCGCGTCAGCTTTAAGCCCTCCCAATGTCTACATA
ACAAAAAGGACATCAAGCTGCGACCCGAAGATTTGCGTCGATGGGACGAGGATGATGACATGAGCGACATGACC
ACGGCCGTTAAGGTGCCCACCAACCCCACATGTTATCCCTATGCACATTTCTAATCTAATGTATAATTAATTCC
AACAG**GATAGACCCACCTCCCGACGTCGGGGATCGCCCATTCCGCGCGGCAAGTTCGGCAAACTGATGCCCAAC**
AGCTTTGGCTGGTACCAAGTCACGGTGAGTTTGGACTACCATGTCATATAGATATCATAGCCGCAATGGCCCAG
ACCAAAATCATGTGATTTAATAGTAGTTTTATTACTAAGCTAAAGAAGCAAAAGCATTCAGTCCTCCAGCACCT
GTTAGATTTATAACGTTCGCCATTATACAAATTAATCGACTGTTCTACTGGATGGCTCTTCTTATGCGAAGTAC
ATATATGTTAAATGTAGTTTTAAGTATTTAGATTACTAACGTCACATAAAACTTCCAGAAGTTTGCTCCAGTGC
TCTACTAATTTCCCGCCCCTACATAGAACTACAAACCCCAGACTGGTCCATATAGATACCAGATTTACGACAGC
AGTCATAGAACTATTTTTGGTCTCCAAGAACATCATGTGATTCGTCTGATGGCTTTTTTTTTTTGTTATTGGCA
AAAGATCGCGTTGAGGGAGCAAAGTATCCGGCAAGAATCGCATAAGTTACAGCGAAATGTTGAGACTTATTTGT
TTATTTGTACACTGCTTCCAAAGCTCCAATGCCGACTTGCCGACTGGGCTCTTTGTGTTTATTAATTAATTTCC
CTCGTAAAAATCGTAAAGCAAACAAGCCAAAGTTGGGCCGGGGGGATACCAGCTCTACACAGCTAACGACAATC
GGGTACTACAATTACAATTGCAATCCGACGAGCGGGCAAATAAATATTTATGGATGGCTAAACACAGTCGTCCA
CAGTACCAGCCACAAGTGTTTGTCGGCACGAGCTAAAAGTGCAATGAAGTCGAGCCTTTGGTGGTTTGACAGTT
GCCATGACCCTCACCTAAATGGTTTCCCATTTTAAGGGCTTGCGGGCCTGCGGTTGGGTTATCCTTCGCCTGGT
TGTCACCCTATCCAGCCGCACGTGGTGCAACACAACAGGTCGCTATCTCTCGACGGGGTCGTAAATGCCATAGT
AAAATGCCCAATGCTGACTCTCCACCAACAAAAGCTTGCGGTTTCATTCCGTTTGGGCCTTTATTTCATCGACG
GCCGGCAGACGGTGAAGACGTGTTCCAACGTCCGACGCCCCATAGAATATATCGCTTGAATCTCAAGAGGCCCG
ATTAGGGGCAACTTATCAGTTTGGAGTGATTAGATGTGCGACTAGGGTTCCGTCGGGAAAATCAGCCTAACACG
TGCCCCACTTAAAACCCGGCATCTGCAATGGCTACTTTGGGCGCCTGGGACTACACGATCCTGGCGGTGGTGCT
GATCATCTCCGTTGCGATTGGCATCTACTACCGTTTCGTGGGCGGCAAGCAGAGCACCACCACCGAATATCTTC

TGGCGGATCGTTCCATGAACGTGGCGCCTGTGGCCTTTAGCCTGATGGCCAGTTTCATGTCGGCCGTCACCATA
TTGGGCGTATCCATGGAGAACTACCAGTACGGTACAATGTTCGTCGTGATCAATCTGTCGTATGTGTTGTCCAC
GCCGGTGGCCGCCTACCTCATCATTCCGGTCTTCTATCGCCTAAAGACGGCCAGTGTGTACGAGTATCTGGAAC
TGCGGTTTGGCTATGCCACCCGATTGGCCGCCTCGCTGAGCTTCTCGCTCCAGATGGTCCTATACATGGGCATT
GTGGTCTATGCGCCGGCTCTGGCCCTGGAAGCTGTCACTGGACTGAGCCAGGTCTTCTCCATCGTAATTGTTGG
AGTCGTGTGCACGTTCTATGCGACGCTCGGCGGAATGAAGGCCGTACTCATCACGGACATCTATCAGTCGCTCC
TCATGTTCGCCGCTGTCTTCAGTGTAATCATCTGTGCCTGGGTGAAGGCGGGCAGCTTGGAGGTGATATGGCGC
GTGGCGCAGGAAAACGGACGCATCAATCTGACCAACTTCAGTGTGGATCCAACGGAACGTCACACTTGGTTCAC
CCAGATTTTAGGCGGATGTGCCACCTATTTGGCCATCTATGGTGTGAATCAAACGCAGGTCCAGCGCCTGATGG
CGGTGAAGAGTTTGAGTGCCGCTCGAGCGGCCCTGTGGTGGTGCCTGCCCATTCTGTGCCTGCTTAGTCTGAGC
ACCTGTTTCTCCGGTCTGTGCATCTACTGGTACTACCGCGACTGTGACCCGCTGCTGGAGGGTAGAGTCAACTC
GCGCGATCAAGTGATGCCGCTCTTTGTCGTGGACACGATGGGCGAATACACCGGTTTGGCTGGCCTCTTCGTTT
CGGGAATCTTTTGCGCCAGCCTCTCCACGATTAGTTCGATAATCAGCTCGCTGGCAGCCGTCACCCTGGAGGAT
TACCTCAAGCCCTTGGTCAGCTGCTGTGCCAAACGCACGCTCACCGACCGTCAGACCTTGTGGTACTCCAAGCT
GTTGTCCCTATTCTTCGGAGCCCTCTGCATTGGCATGGCCTTCATGGCGGGATCAATTGGCGGACTACTGCAGG
CGGCACTGTCAATCTTCGGCATCATTGGAGGCCCTCTGCTGGGCCTCTTCACGCTGGGCATGTATGTGACCAAG
GCGAATGAAAAGGGAGCCATTGGTGGGCTGCTCATATCGCTGGCGTTCTGCTTTTGGATCGGTTTCGGACAACC
GAAGCCACCGCTGGTCAGCTTGGATATGTCCACGGCTGGATGTCCGGTGGACAGGAGCTTTGCCCGCGACATTT
TCCTCAAGACTGTGATAGCGGTGGCCGAGGAAGAGCATTACTTTTACCTGTACAGGATTAGCTACATGTGGTAC
GCTGCTTTGGGCTTCCTGATAACCTTCTTTGGCGGATGGCTGTTGTCCTGGCTGTTTGCCCTGCTGAAATGGGA
CAACAATCGACGCATCTACCAGGATGCGGACTGCACGCTAATCAAGCACGATCTCTTCGTGCCGCCGATAGCCA
AACGTTTGCAACGGCGACAAATGCCGCTCCTGGTGGTCACCGGCACCAGTTCGGAGATTGGCGGCATCACGACA
GAGAGCGCTACGGCGCCGCCGCGGCTGGACGAGATCGAGTGGGAGAAACACAAGGCAGTGGCGTAACTGAACAC
ATCACCTTAGTAGTTTTTTCTTTTGTTTTTTTTTTTGTAGCGCCGTTTTGACCTTAGTTACTCTCCATTAGCTAT
CACTGCAAGAAATAATAACTTTGCTGTACAATAATACGGTTTAGTTTATTTGCATTTCTATACGAGTCTAAAAA
TGAGGCAAACTATTATTTGAAAGAATTGCTGCTGGAATTACAATGCTCCTTATAATGTTCTAAAGAAAGTCAGG
TCGGAGAATATCATATAGTTGCCATAGTATGTATATCTGCATGGGTCATCCCACTTCGTTCTCCGTGGAGCGTA
CTACTAGCCTTTTTCTCTTGTCACACAAAATATATATATATATATATATGTATATATATATTAAGTGAACTTTG
GCTTTAAATAAACGAGCTTACTCATTAAGCAACTAAATTAACGCTCATCGATGGCCTGATTGGGCTGCTCCTCC
GGTCCTTTGGCATTGTAAACATTGCGATTCTGGGACAGATGCCATAGGGCTAGCAGGCGGGCCAGCTCCGTGTA
GGTTAGCTTGCGCAGATCACGTCGATCGGCCAGCTTAACTCCGGCGGCCTCTAGCTCCTCCACCAATGTGACGG
CATCGTATTCATCCTCCGGATTCTGCTCATTGCGTGGCGCAATTTTTCCGACGAACTCCGACTCTGGCACCGCT
TCGCTCTTCACCATGCGCTGCATCTGCTGCACATGCTGCTGTTTATGCTGACGTTGCTGACCCAATGGCTCGTC
GACATCTTTTAGCTCCTGTTCGGGATTCATGGGCAGCGAATTGGCCAGGCTGACCATGCCGAGGATCACCAGGA
AGATGACCGGAATATTGTTGCACCCAATGGGTTTCATGTTTACGTCCGACTGCTAGTAGCTTTGCTGTGGCACT
GCTGCCCGCCTGTCGCATGTCGATCGATCGTGGCCTGGACTCGACTCTCGGGGTGCGCTCCAATTTCTATCTGC
CTTCCAACCCGCTCGTACGCTCGCTCGGCTCCCTATCGCTCACTCGAACGCCCCGTGTGGCAGTGCTGGCTGGC
GAAGACATTCGCTTTCTTCTTCCAGCATTTGGCTATTATTTTTTTTGGTTACTTTCGAATGCCATGGATTTTGG
CAAACAATGGTGCCATTAAACTGCAAGTAAGGCGGGGATATAGATTCTTTTACTTCTAAAACTATATAAATTTG
TAATAGATGAGCAGTGAGTGTTTTTAAAATGTATATTTTATTGTATGTTATGCAAAAAATGTTCCATTTAATTC
AATACGCTGGAAACGAAGCGCAGCACTGTCGTTTGGGGGGCTGCTTTTGATTGAAAAGTCAGTGTGTAGGGTCG
GTTAAATTATTTTATCTGCATATGCAGCACATACACTATACTATACACTATACAACATCGACCCCTGGTACATA
TGTTATTTAATTGTCTGGCTGGTGCCTTTTGTATTTTGTATTTTGTATGTAGTAGATCCCGCATTTCGTCGTTT
TTTATTGTGTATTTTTGTTGGCTACGTTTTTTTTTTATTTTTTCTGTCCGCCTCCAATTCAATTTATCTGCGCA
GTTTGTGCTCACTCGGTTCGCCGTCTTCTATTTTCTGCATGTTGGTTTGGCCTAATCTTATCAGCTCTTATTCA
TAGCCAATTGAACGAGTTGTAATGGTCGAGAGTTTTTAATAATTTCTAGTAGATATGGGTTAATGGACGACGCG
CCCCGCGTTAATCAGTCGCAATGTTTCCTTGTGTACCGAACCAATCGGTCAGGCAATCTTAGGTCTAAGAACTT
TGTTTGATTGGTCAAAATGTGAATGATATTATTCTACATTTACATGTTCATATTTCACACAAACAATTATTGCT
CTATTTGCAACAATTCTAATGGAATACAGTTGAAAAGCAACTAGAGATACCTCGTCTAGAGTTTTAAATTCTTA
TCGGTAATTGGGCTTATTTAATTACTCACTGCTACAATTACTTTGCCTGCCTTCAGGCGTTGCTAAAAGCTTTT
TTTTATAAACAAATTGCAATAGCATAGTCTAACTTGTAGTGCATGCCGAATTAGTGTCAACAAAGAAATTATTG
TTAAATATATTGCGCTATTATGTGTTGAATTATAACGATTTGACAAATACACGTAACGACGTTCTTGAGTGACA
ATGGAAATTGAAAAAAAGTCGAGATTGTCAAGTTGCATGTAACGTGCGACGAAGTAAGTATTGATAAATGTAAA
TTCTAAATGGAAAATCGCAATTGTCAGTCTTATTTGCATTTTCCTAATGGTTTCACATCTGGCATGCGAGCAAC
CTACATACATATGTATGTGCTATGCTATATATACATACATATGTATATAAAGTGTTTATAAAATAAATGTACGT
GCGAATGGAATGGCCAAAACATTCCACTGACTGATCGTTGTTAGCTTTTATTAGACTGCAAACACGTGCGCCAG
CCCGCTTTGGTTAGTTAGTCATTTTTGGTTGAGAATGAAAAGAAGAAAAAAAAAAAACCCAGAACTGTCAGCAG
CGTTGATAACCGTGTTGTGGGCCAAACACAACTGACTGACTCACTGTATAGGATAGCTCCATAGCCAGCTATGG
GAGGTACCCTATCAAGGGCACCCAATAGCAGCAACCAATCCACTTGACTCTCTTTTGTGTGCATCTCCTGATGT
CATCGACCAATAATGCGTTGGATCGTTTCCCTTCTTATCGCTAGCCAGTTGAACTGGCAGATTCCGCCGATTCC
AACCTGCGCTCCTGCGGATTCCCACCTTGCCAATCTCATATCCCACATTTATACAGTAGATATGGGATGAGCGA
ATCTCGTTGAGTACTAGTTTCTACTCAGAACTATCACCGATCACAGAACAGTTAAGAACTTCTATTTCTAACTA
TGATTTCGTTTCTCAATACATATTTTGAAACAAATACAAGTTTGATGCAACAATTGTTAAACATTATTCATAAA

AATTGCATTGAACAGCGAATATTATAATCAAAAGCAAAAGCAGCACTTATTTAATGATATTTGTAGATAGTTAA
ATAGTTATAGATGCATAGCTTTAAGGTTGTTCTATATGAAATTATCCCATTGAATGCAAAAAATGCAAAAGTGA
TTGATGATTTCCTGAATCGAAGCCACTTTATAAGACTGTATTTGATTTGAATGAAAGCCCGAGTTTGCCTTTTA
ATTTAAAGCCAAGTGCAATATTTAGCGATTCAACAATTTATTACGATATTGAAATTCAGGCGCATAACAATTAA
TAATTACTCGTTAACAGCTTAACTTGATTCATATTCATATTCACATTGTTCGTATGTACATTATCTTTTTATTA
TTTATTCACACACTCACAACAGACTCACGCACACACTCACGCACACATTGGCATTGAAATGGTTTCAAAAAAAA
AGGTTTTGGTTCTTTTCACGTCAGTTAATCTGATGGCAGCTAATCCTGCTCCTGCTCCTCGCTGGTGGTGGTCA
CCGTCTTCTCCTGCCGCGTGGTGCTAGTGGTGGTGCGTGTCAATGTGAAGACACCGGTGGTGCCCACCTCATCC
AGCGGCTGGGCAATGATTGTTGATGTGGTGCTCTGGAAACGAGGTGTTAGACATTTGATACTACATATATTATG
AGGAGGGTGTCGTACGCCATTCTTGGTCTCCTCGGGACTCCACGAGCGTTTGTCATCCAAATCCTTTGAGCAAC
AGCCTCCCATTCTTTGGTTTTGTGCGTATATAGAAATACGTATTCGTTGTATATATGGTATAAATGGTTTGGTT
CGGAATGCCTTTATTTTTCGTTCTATATTAATGTATATCGGGGGCCAGTAACTTGTTGCTGCTTCATCGGCGCA
CGTCTGTCAGCGAATGCCGGTGTTCGATGTGGTGTTCACTTGTTAGCCAAACGAACGAACCGAAAGCCCAACCC
AACCCAACCCAACCCAAGCTGCGCCGCCAGCTGTGACTGCGACTGCGACGCCGGAGTTGGGTTGTCATAGCTAT
ATAGCTCCATATCAATTCTAGAAATATGTCTCAGCCAGCGAGCCAGCCAGCCAGCCAGTACTCGCATTCGGTGT
ATATTGTATTGTGTTGAATACAATAAGTATATGTATAGCTCCATGTATGCCTGTGAGCTGTACAAACAAACCAT
TCACATTCCAAACAACCAACCAGCCAGCCACCCACCTCAATGCCCCCGTTCGAAACTCTATGTTCCCGAGAATG
TCAAATCCGATCTGATTCCCGCTTTGAAAACAGTCACAGTCATCATCTTTATTTTCAAATGGGAACCACTGGCC
TTGTCGCCTTAACGGTTCCATTTGCCCATCGAGCAGGTGCAGCGTGTCAACTCAAGGCTGCCATTGTTGTCTGT
GGCCTATTTGGTGGCACTATTCAAAGTTGCAACGTGACAACTAATGTCAAAATCCTTCCAAGAGGATTGTGTGC
TATGATCTTCCTTTGGTTCATTTAAAAAAACCAAATGGGCCATTAACAGCTTAAATTGGGCACCCTTATATAGA
GCTTTTAACTAATTAGCAGAATTAGTCCCAAAAAAAATATAAATAAATACCTTTGTAGTTAGGACCAAACTTTA
TTTAAAGATCCTATGACACATCAACTTTGCGGGGAAATTCAAATACACAATGGTTGTAAATTAATAAAAACACA
ACAATTTGTCATTTTTGAATGCGATATCACGAGGGGGTTTATGTTGTTCGTATGTTTTAGTCAATGGAAAGTTA
GGGTTAACCATATCGAGGCAACCAATAGACGCAAAAGAGCCAAAATATACATATAATCGTATACAGCTAATGCC
CGTAATAAATAAATTTATCGCATTAATGCGCAATGAATGAGTGGGCTAGATGTGTGCTGTGACTGTTCGCGGAC
TTATCTGCTTGATACTGTTGATATGAATAGGGTCGAGTGGATTGGCATGGTTTGTGGCATCATCGGGATGGATA
AGTGGGTAGTGGGTGCTGGGAACTATGAGCTGGTAGCTGGGCAGCTAGGCAGCATTCACACGATCATTATACGC
ATACGACCCCCCTATCACTGTGCCCGAGTGTAACCGAAGTGATCCTATAAATTGAATTACCTACCACTTATTCC
AGTGAATCTCTGAAGCACATAGCACCATGGTCATATATATGGCGCACATGTGTACGTACATATATATGTATGTA
CGTCTGGATGCGATGCAGTCAGTCCGATTAGATAGCCATCGACCAATATTAATAATAATAATAATAACATTAAT
AGCCAAGTGACCGATATTGCGTATCGCTGACCTTCGGTGTCAGAGGAAGTGGTAACTCGTTTCCACTCTCTCCT
CGTTTTGCATCCACCTCTGTTTGCCGTACACTCACTCATATCCGGATATAAGTGCACGTAATTAAATCACTGAA
GGCTTGAATGATAAGACTGCTCCGCCCACAAAAAAAAAAAAAACGTTGCAATTAGGTGATGAGTTCCACGCATAG
CTCTGCTCAATGGCATGCCAATCAGTAAGGTCATATGTGTACACATTGATAGAACAACAATATTATACATATTA
TATAATGTATATAGTAGAAAGTGATTTAAGTTCAGAGAAGATTTCAGTAAGCTAACTGAATAAAGTCAAATATT
ATTGCTTACATCTACAGACGTGTACGTGATTCTTCCTATTTTTAGACATTTAGAGCTATCTCTCCGTTTTACAC
ACTCCATTGGATGATCTATCGCAGAACCAACACGAAATATGGAGCACTGCTGGCAAAACATGAGTTCATGAGTT
GGCCGATGAGATAATTTCGGTCGCTGCTATCTGTGCACCATATAAGTTTATATAATAATCGACAGACGACGTCT
AATGGTTTCCCTTTTTCCGCATTTTTCTTTTTGCAG**TTACAAAACGCCCAGATATACGAAAAGGAAACACTCTT**
GAGTGCTCTATTGGCAGCGATGTCGCCACATGTCTTTATTCCTCAATATTGGCGAGTGGAGCGAAACTGCGTAA
TCTTCTTTACGGACGACTACGAGGCAGCCGAACGCATTCAACATCTGGGCAAGAATGGCCATCTTCCAGATGGC
TATCGTCTGATGCCACGAGTACGCAGCGGTATACCACTAGTGGCCATCGACGATGCCTTCAAGGAGAAGATGAA
GGTCACAATGGCCAAGCGTTACAATATTCAAACCAAGGCGCTGGATCTTTCCCGTTTTCATGCAGATCCGGATC
TTAAGCAAGTTTTCTGCCCACTCTTTCGTCAGAATGTGATGGGCGCTGCCATTGACATTATGTGCGACAATATA
CCCGATTTGGAGGCACTTAACCTGAATGACAACTCCATTAGCAGCATGGAGGCGTTTAAGGGTGTGGAGAAACG
CTTACCGAACCTCAAGATTCTCTATTTGGGGGATAACAAGGTTAGTGTTGATTGAAGATAGTCACTTTATCAAT
AACTAACTGACGACCATTACCTTTTCCAG***ATACCATCTTTGGCCCACCTTGTAGTGCTTCGCAATCTGTCCATC***
TTGGAACTCGTTTTAAAGAACAATCCCTGCCGTTCCCGCTACAAGGATTCCCAGCAGTTTATCAG*GTATACTAT*
GGTGTGATTCGTTGATCTTCTAACTTTTCTAGTGGATGCTGCCGCATATTCGACGTGGGATTTGTCGCGTCTGG
AGTTACAGAGGAATTAGGATTACCACATTTCCATTGTTCTGTTGTCTCCATTCCTTTTCTTTGAAGCAAATGCG
CTGCAGAATTGGGGACCAAGGTCAGGTGTTGCCTTCGTCCAAAGTTTAAAGCAGCTGAGCAGGCAAAGGCATTT
GGGAAATATTTTCCTTATCAGCTGTCTTAAAGCCAGGACAGATCGTATGCAGCGGAATGGATTTTGGGC*AGAGC*
AAAAAAAAAAAAAAAGAAGAAAGAAAAAAAAAGATTTGAAGTTAAAAAAAAAAAAGAAAAAAAAAGAAGAAAAAA
AAAGAAAAAAAAGTACAAGCAAAAAAGGCCCAAGCCAA*TTGCAGTAAACTTTCGACGATGTGGCAACGCTAAAC*
TAAGAACGCAACTGTACTTTTGACACACACACACCGGCACCGTTCGTACTTCCAATTGAGCAGGCGGACGTGCA
GCAGTCCCAAAGCGAATCACATTAACGCATTAACATCGACAAATATAAATTATTTTACGAATAACAAATAATCC
ACCGAACCCACCCACCGCGCACCCCGCTACAACACACACACGGGCGCACGCATGTAGCCACTCCACGCCAGCGC
ACGCACGCATGCGCATGCGCATATGCAGTTGTGTACGTGCCACATCTAATGCTGCGTGCCCTGTGCGCGCCGTT
GCGCCTTGCATTGTCGTGTCGCCGCTTGTTGATTTTGCGGTGGTTGCCAAGTAGTGGAGGCAGAGTGGAGTGCA
TAGGCTCACCATTCGGGTGCGCAGCGTCCAGCAGGAAGCAGAAACGACTACCAAGACGAAGGACACACCACTGA
CTGACTCACGTGCTAACGGACATTCTGCTGGCGCCAGCGGCCGGCGCGTTGTTTTGCCAGTCGATTACAATTTA
CAATTAGTATTTTCGGCTTTGTTGGGCGCATGAGCCGCCTGCGGATGCGACGTGCTTGAACATGACGAGACGAA

*CAGAAGTTACACAGCAGTCAGCTTTGAAAGCCACTTTTTGCTTGTGGCCACTG**AAAGAAAAAACAAAAAAAAAA**
AAATGAAAAAAAGGAAAAGAAAAAAGTGTTGATAAAAAAAAAAAAAACAAAAAATACAAAAGAGAAAATACCACA
AAAAAGATGCCCCGGCACAATGCACTGCTCAATTGTACTTTTCCGCTGTACTTTTCACATCGAACTATTGAAGG*
CACAACAATGGCCAAGCACCTAGTGCCTCTGCCTCTTCAGCTGATCAAAGCTCCGCCAATTTGTGTGTGAGCCG
GTCGAGTAACGGTGAAGCTATATCAAAGATACAAACATATATACTCCGCGTGAGAAGAAGAAGGAGTCACATCT
GGCCACTGACTAATCTGCAATTCCTGCAGCGCATACATCCTAAACCTCCCCCCCCCCCCCACCAACGCACAAATT
TAAAAAAAAAAACCCCCATTTAATTATTATTTCTCGCTATTTAACTCGCATATCTATCATCGCCTCCCTGTGCTA
*ATGACAACAAAAATACGGCCGAATGTCCACTATTTACAG**CGAAGTACGTCGCAAGTTTCCCAAACTGGTTAAGT***
***TG**GTAAGTTCTATTTGTAAGTGCCCCTTCCGTTAGCAAAAAGATGGGAATTCCATAAAAATCGTATACATATTC*
CATGGATTACATGGATCTAAGGTTTGGAATATCCTTACCGATTATTTGGCTCATCAATTCCCATCACAGCTAGG
CAATCACACGACCATAACCAAATTAAATCCCATATAG**GACGGAGAGACCCTGGAGCCGCAAATCACATTTGATC**
TATCCGAGCAGGGACGTCTTCTCGAAACGAAGGCATCCTATCTGTGCGACGTCGCTGGTGCCGAGGTGGTGCGC
CAGTTCCTGGACCAGTACTTCCGCATATTTGACTCGGGCAATCGGCAAGCTCTGCTAGATGCCTACCATGAGAA
AGCGATGCTCTCCATATCAATGCCTTCGGCCAGTCAGGCGGGCAGGTGAGCATATTAGTGCTTGATAATAAATG
AATACCACCAATGCATTAATCGCATTCATGTCCTTCAG**ATTGAACAGTTTCTGGAAGTTCAATCGCAATCTCCG**
GCGCTTGTTAAACGGCGAAGAGAATCGCACCCGAAACTTGAAGTACGGACGCCTGGCATGTGTTTCCACATTGG
ATGAATGGCCAAAAACGCAGCACGACCGACGCACCTTCACCGTCGACCTGACCATCTACAATGTAAGTAAAGAT
TTCAAATTCTATTAGCTTGAACTAAATGCATCTCTTCCCATCCAG**ACTTCAATGATGGTTTTCACCGTGACGGG**
ATTATTCAAAGAGCTGAACGACGAGACCAACAATCCCGCCTCCATGGAATTATATGACGTTCGCCACTTTGCCC
GCACCTACGTGGTGGTGCCACAGAATAATGGCTTTTGTATCCGCAACGAGACGATCTTCATCACAAACGCTACG
CACGAGCAGGTGCGAGAGTTCAAGCGATCGCAGCACCAGCCTGCTCCCGGAGCTATGCCCTCCACTTCCAGTGC
AGTGACCAGTCCTCAGGCCGGGGCAGCGGCGGGTCTGCAGGGTCGTCTGAATGCGTTGGGCGTGGCCACTGGAC
CGGTGGCTATACTATCAGGAGATCCGTTGGCGGCCACCGCACCGGTTAACAGCGGCAGTGCCGCCATATCGACA
ACAGCAGTGGCACCTGGCGCCCAGGATGAGAGCACTAAAATGCAAATGATTGAAGCCATGAGCGCCCAAAGCCA
AATGAATGTGATCTGGAGTCGGAAGTAAGATCCAAGTGGCAATCTATCAATAGCTAGATAAATAAGTAAACCCA
TTTCTTCATTTAACAG**ATGCCTGGAGGAAACGAATTGGGACTTTAACCATGCCGCCTTTGTGTTCGAGAAACTA**
TTCAAGGAAAACAAAATACCGCCTGAGGCTTTTATGAAGTAAATCGCATAGGAGTTTCCGTAGGACAGAGCCGC
GTGCCACATCCACATAATCGAATGCTGTTTTTTTTTTTTTGGTTTTGTAATTAAATTTTAAAATTATTAGAGAA
ACCTCTATATAATAATAATAATTAATATTATTAAGCTGCGAAGTTGTGTGCACATTCGGGCAGTAGCAATTATT
ATCCCAGCACTGCGGGCAATGTGCATCAACGATCACAGTTCTTCGATAGATTAGTTTAGCTCTCTTTAAGTTCC
GTCCGCAGATCCGCTGGTCTACATTGAGGCCAGGACGAGTCTGCGAACGGTAGTCCCTTTAGAGTTAAAGTTGT
TTTAGATTCCTAAGCCAAACACCTTCAAACACACACAACACGACAACACTATTAAACGTAATGCAACAATGTTG
TCGAATCGAAAGAAACCCAATTTTTATTTTAATATATACGACGATACCGAAACCAAGGCGATGGTCGAAAAGCA
TGGATCGCGCCAATAATTCTATATTCCCCGCTTTCCCAGATCCTCGACTCGCCTTACATTTTGTATACAAATAC
CATAGAATAAAAAGAAACATTTTGACCACTGTAAAAATTTTGTAACGACTCGGAAACCAAAATTACCTTTATTT
CTTAATAGAAAAAAATTATTCGATTTACAATACACGCTTAGCCGTAAATTCAATTGAATTTAAGTGTAAGATAT
ATAAAAATTATATACATTAAGACAAAAGTGTAGATTCATAAATAAATTTTTCAGAATAG

Приложение 3

Экзонная структура гена *Nxf1 Drosophila melanogaster*, с включенным кассетным интроном. Форматирование аналогично Приложению 2. 5'- и 3'-UTR не представлены в последовательности.

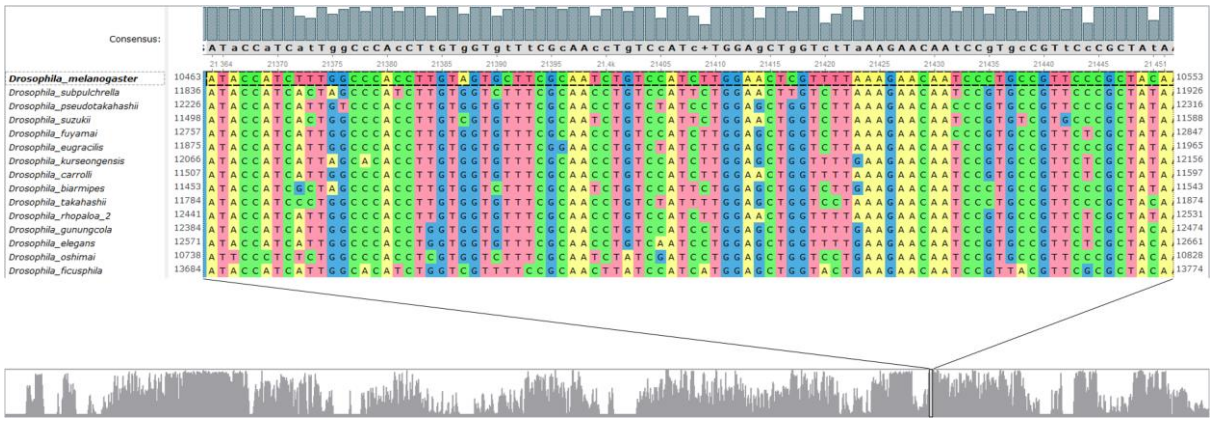
1.ATGCCCAAACGCGGCGGTGGCAGTAGCCAGCGGTACAACAACAACGTTGGAAATGGCGGCGGACGTTACAAC
GCTCCCGAGGATTTCGATGATTTTG
2.ATGTGGAGGATCGCCAGCGACGCAAGGATCGAAACAAGCGGCGCGTCAGCTTTAAGCCCTCCCAATGTCTAC
ATAACAAAAAGGACATCAAGCTGCGACCCGAAGATTTGCGTCGATGGGACGAGGATGATGACATGAGCGACATG
ACCACGGCCGTTAAG
3.GATAGACCCACCTCCCGACGTCGGGGATCGCCCATTCCGCGCGGCAAGTTCGGCAAACTGATGCCCAACAGC
TTTGGCTGGTACCAAGTCACG
4.TTACAAAACGCCCAGATATACGAAAAGGAAACACTCTTGAGTGCTCTATTGGCAGCGATGTCGCCACATGTC
TTTATTCCTCAATATTGGCGAGTGGAGCGAAACTGCGTAATCTTCTTTACGGACGACTACGAGGCAGCCGAACG
CATTCAACATCTGGGCAAGAATGGCCATCTTCCAGATGGCTATCGTCTGATGCCACGAGTACGCAGCGGTATAC
CACTAGTGGCCATCGACGATGCCTTCAAGGAGAAGATGAAGGTCACAATGGCCAAGCGTTACAATATTCAAACC
AAGGCGCTGGATCTTTCCCGTTTTCATGCAGATCCGGATCTTAAGCAAGTTTTCTGCCCACTCTTTCGTCAGAA
TGTGATGGGCGCTGCCATTGACATTATGTGCGACAATATACCCGATTTGGAGGCACTTAACCTGAATGACAACT
CCATTAGCAGCATGGAGGCGTTTAAGGGTGTGGAGAAACGCTTACCGAACCTCAAGATTCTCTATTTGGGGGAT
AACAAG
5.*ATACCATCTTTGGCCCACCTTGTAGTGCTTCGCAATCTGTCCATCTTGGAACTCGTTTTAAAGAACAATCCC*
TGCCGTTCCCGCTACAAGGATTCCCAGCAGTTTATCAG
GTATACTATGGTGTGATTCGTTGATCTTCTAACTTTTCTAGTGGATGCTGCCGCATATTCGACGTGGGATTTGT
CGCGTCTGGAGTTACAGAGGAATTAGGATTACCACATTTCCATTGTTCTGTTGTCTCCATTCCTTTTCTTTGAA
GCAAATGCGCTGCAGAATTGGGGACCAAGGTCAGGTGTTGCCTTCGTCCAAAGTTTAAAGCAGCTGAGCAGGCA
AAGGCATTTGGGAAATATTTTCCTTATCAGCTGTCTTAAAGCCAGGACAGATCGTATGCAGCGGAATGGATTTT
GGGCAGAGCAAAAAAAAAAAAAAGAAGAAAGAAAAAAAAAGATTTGAAGTTAAAAAAAAAAAAGAAAAAAAAAG
AAGAAAAAAAAAGAAAAAAAAGTACAAGCAAAAAAGGCCCAAGCCAATTGCAGTAAACTTTCGACGATGTGGCA
ACGCTAAACTAAGAACGCAACTGTACTTTTGACACACACACCGGCACCGTTCGTACTTCCAATTGAGCAGGC
GGACGTGCAGCAGTCCCAAAGCGAATCACATTAACGCATTAACATCGACAAATATAAATTATTTTACGAATAAC
AAATAATCCACCGAACCCACCCACCGCGCACCCCGCTACAACACACACACGGGCGCACGCATGTAGCCACTCCA
CGCCAGCGCACGCACGCATGCGCATGCGCATATGCAGTTGTGTACGTGCCACATCTAATGCTGCGTGCCCTGTG
CGCGCCGTTGCGCCTTGCATTGTCGTGTCGCCGCTTGTTGATTTTGCGGTGGTTGCCAAGTAGTGGAGGCAGAG
TGGAGTGCATAGGCTCACCATTCGGGTGCGCAGCGTCCAGCAGGAAGCAGAAACGACTACCAAGACGAAGGACA
CACCACTGACTGACTCACGTGCTAACGGACATTCTGCTGGCGCCAGCGGCCGGCGCGTTGTTTTGCCAGTCGAT
TACAATTTACAATTAGTATTTTCGGCTTTGTTGGGCGCATGAGCCGCCTGCGGATGCGACGTGCTTGAACATGA
CGAGACGAACAGAAGTTACACAGCAGTCAGCTTTGAAAGCCACTTTTTGCTTGTGGCCACTGAAAGAAAAAACA
AAAAAAAAAAAATGAAAAAAAGGAAAAGAAAAAAGTGTTGATAAAAAAAAAAAAAACAAAAAATACAAAAGAGAA
AATACCACAAAAAAAGATGCCCCGGCACAATGCACTGCTCAATTGTACTTTTCCGCTGTACTTTTCACATCGAAC
TATTGAAGGCACAACAATGGCCAAGCACCTAGTGCCTCTGCCTCTTCAGCTGATCAAAGCTCCGCCAATTTGTG
TGTGAGCCGGTCGAGTAACGGTGAAGCTATATCAAAGATACAAACATATATACTCCGCGTGAGAAGAAGAAGGA
GTCACATCTGGCCACTGACTAATCTGCAATTCCTGCAGCGCATACATCCTAAACCTCCCCCCCCCCCCACCAAC
GCACAAATTTAAAAAAAAAACCCCCATTTAATTATTATTTCTCGCTATTTAACTCGCATATCTATCATCGCCTC
CCTGTGCTAATGACAACAAAAATACGGCCGAATGTCCACTATTTACAG
6.*CGAAGTACGTCGCAAGTTTCCCAAACTGGTTAAGTTG*
7.GACGGAGAGACCCTGGAGCCGCAAATCACATTTGATCTATCCGAGCAGGGACGTCTTCTCGAAACGAAGGCA
TCCTATCTGTGCGACGTCGCTGGTGCCGAGGTGGTGCGCCAGTTCCTGGACCAGTACTTCCGCATATTTGACTC
GGGCAATCGGCAAGCTCTGCTAGATGCCTACCATGAGAAAGCGATGCTCTCCATATCAATGCCTTCGGCCAGTC
AGGCGGGCAG
8.ATTGAACAGTTTCTGGAAGTTCAATCGCAATCTCCGGCGCTTGTTAAACGGCGAAGAGAATCGCACCCGAAA
CTTGAAGTACGGACGCCTGGCATGTGTTTCCACATTGGATGAATGGCCAAAAACGCAGCACGACCGACGCACCT
TCACCGTCGACCTGACCATCTACAAT
9.ACTTCAATGATGGTTTTCACCGTGACGGGATTATTCAAAGAGCTGAACGACGAGACCAACAATCCCGCCTCC
ATGGAATTATATGACGTTCGCCACTTTGCCCGCACCTACGTGGTGGTGCCACAGAATAATGGCTTTTGTATCCG
CAACGAGACGATCTTCATCACAAACGCTACGCACGAGCAGGTGCGAGAGTTCAAGCGATCGCAGCACCAGCCTG
CTCCCGGAGCTATGCCCTCCACTTCCAGTGCAGTGACCAGTCCTCAGGCCGGGGCAGCGGCGGGTCTGCAGGGT
CGTCTGAATGCGTTGGGCGTGGCCACTGGACCGGTGGCTATACTATCAGGAGATCCGTTGGCGGCCACCGCACC
GGTTAACAGCGGCAGTGCCGCCATATCGACAACAGCAGTGGCACCTGGCGCCCAGGATGAGAGCACTAAAATGC
AAATGATTGAAGCCATGAGCGCCCAAAGCCAAATGAATGTGATCTGGAGTCGGAA
10.ATGCCTGGAGGAAACGAATTGGGACTTTAACCATGCCGCCTTTGTGTTCGAGAAACTATTCAAGGAAAACA
AAATACCGCCTGAGGCTTTTATGAAGTAA

Приложение 4

Сводная таблица оценки качества выравниваний на основе количества консервативных нуклеотидов.

Номер выравнивания	Количество консервативных нуклеотидов	Общая длина множественного выравнивания	Процентное соотношение консервативных участков к вариабельным
I	12476	16570	75.29 %
II	7722	27847	27.73 %
III	8137	24662	32.99 %
IV	1680	2046	82.11 %
V	1218	2124	57.35 %
VI	1216	2103	57.82 %
VII	1454	2042	71.20 %
VIII	896	2154	41.60 %
IX	919	2471	37.19 %
X	427	8131	5.25 %
XI	748	4124	18.14 %

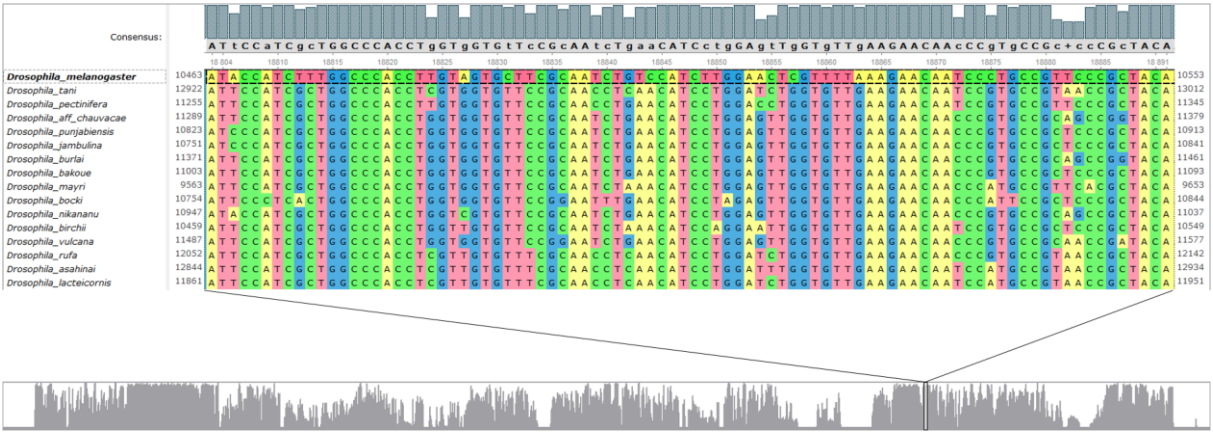
Приложение 5



Результат выравнивания последовательностей группы 2 (диапазон Query Coverage 47-68%). Полная последовательность гена *Nxf1*. Выравнивание II (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 27847 п. н.

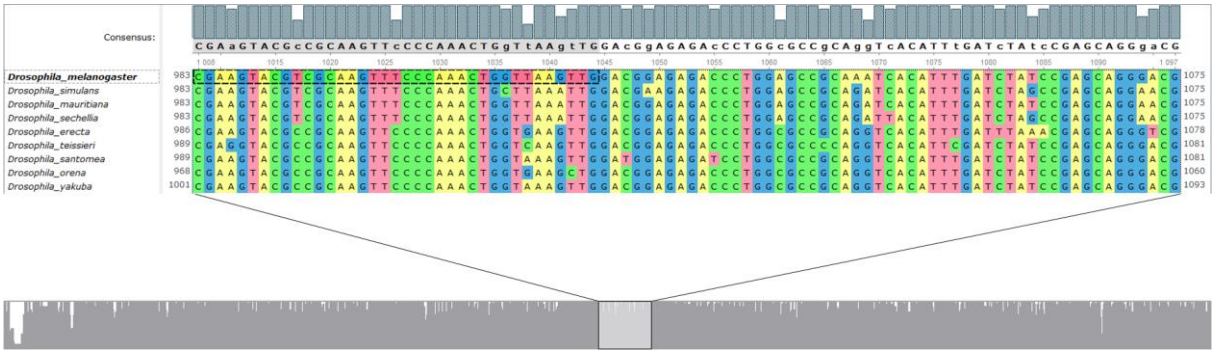
Серый прямоугольник – начало 5-го экзона.



Результат выравнивания последовательностей группы 3 (диапазон Query Coverage 20-36%). Полная последовательность гена *Nxf1*. Выравнивание III (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 24662 п. н.

Серый прямоугольник – начало 5-го экзона.

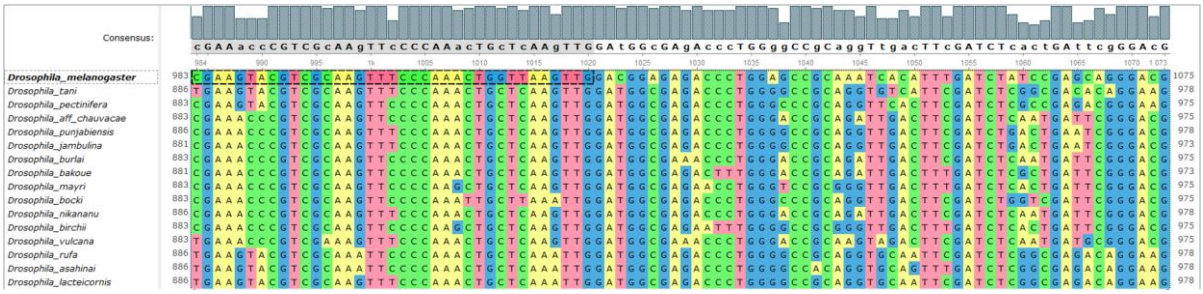


Результат выравнивания последовательностей группы 1 (диапазон Query Coverage 88-97%). Вариант последовательности гена *Nxf1*, содержащей только экзоны. Выравнивание IV (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 2046 п. н.

Пунктиром обозначен шестой экзон.

Серый прямоугольник – шестой экзон размером 37 п. н., а также частично седьмой экзон.

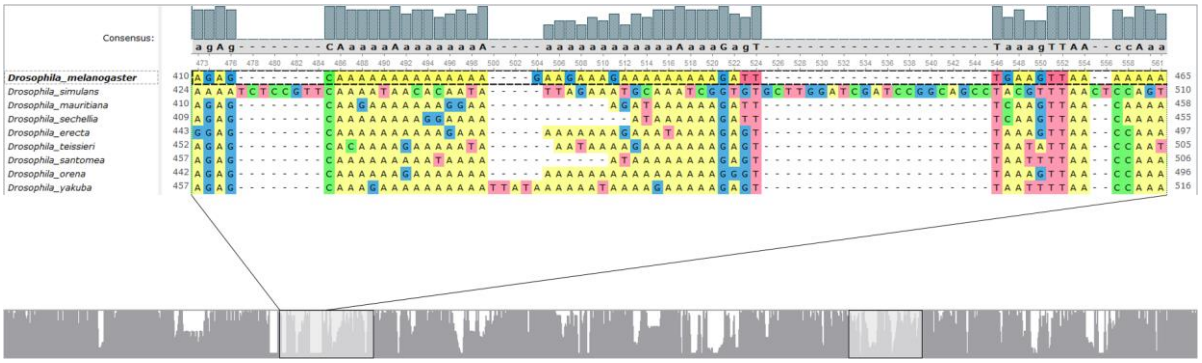


Результат выравнивания последовательностей группы 3 (диапазон Query Coverage 20-36%). Вариант последовательности гена *Nxf1*, содержащей только экзоны. Выравнивание VI (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 2103 п. н.

Пунктиром обозначен шестой экзон.

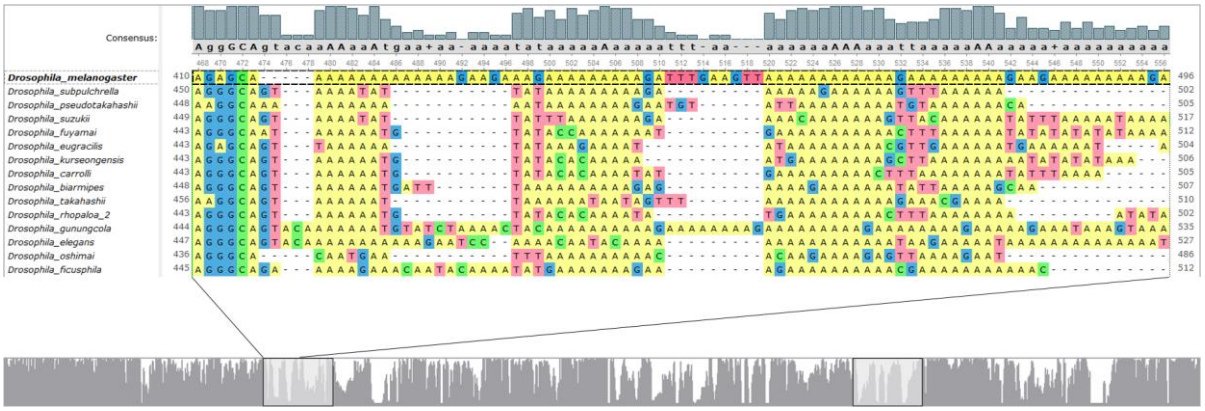
Серый прямоугольник – шестой экзон размером 37 п. н., а также частично седьмой экзон.



Результат выравнивания последовательностей группы 1 (диапазон Query Coverage 88-97%). Консервативная кассета гена *Nxf1*. Выравнивание VII (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 2042 п. н.

Серые прямоугольники соответствуют 2-м поли(А)-последовательностям.

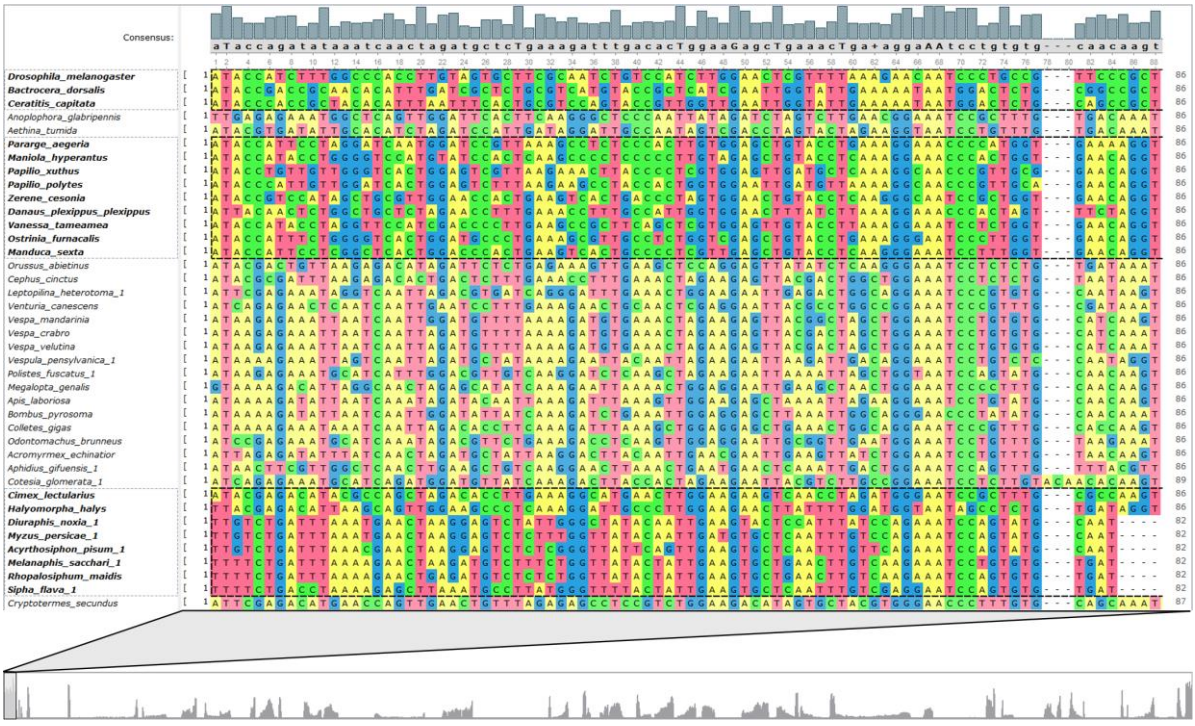


Результат выравнивания последовательностей группы 2 (диапазон Query Coverage 47-68%). Консервативная кассета гена *Nxf1*. Выравнивание VIII (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 2154 п. н.

Серые прямоугольники соответствуют 2-м поли(A)-последовательностям.

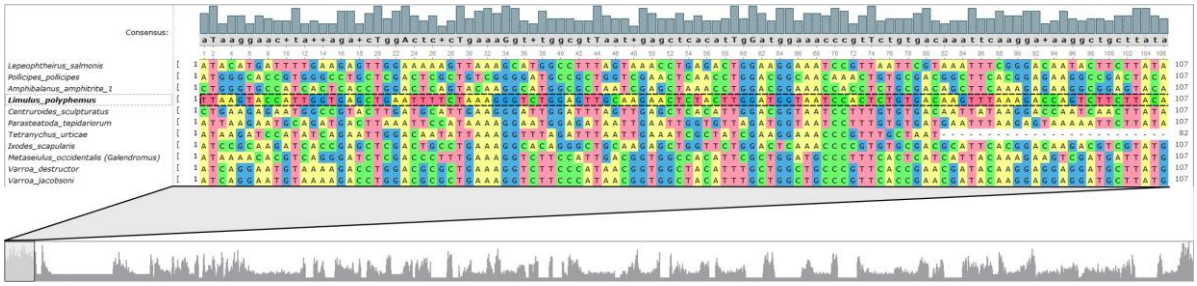
Приложение 6



Результат выравнивания последовательностей консервативной кассеты для представителей *Mandibulata – Hexapoda – Insecta*. Выравнивание X (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 8131 п. н. Серый прямоугольник соответствуют началу первого консервативного экзона из кассеты.

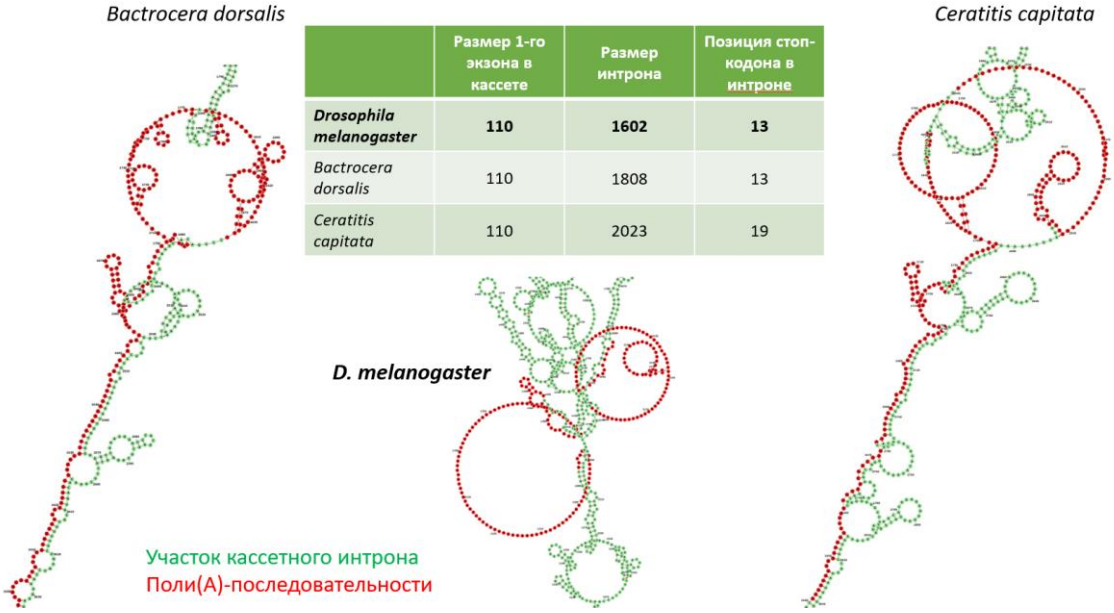
Приложение 7



Результат выравнивания для других представителей *Arthropoda*. *Mandibulata* – *Crustacea* (3 первых вида). *Chelicerata* (остальные виды). Выравнивание XI (см. Приложение 4).

Диаграмма для множественного выравнивания общей размерностью 4124 п. н. Серый прямоугольник соответствуют началу первого консервативного экзона из кассеты.

Приложение 8

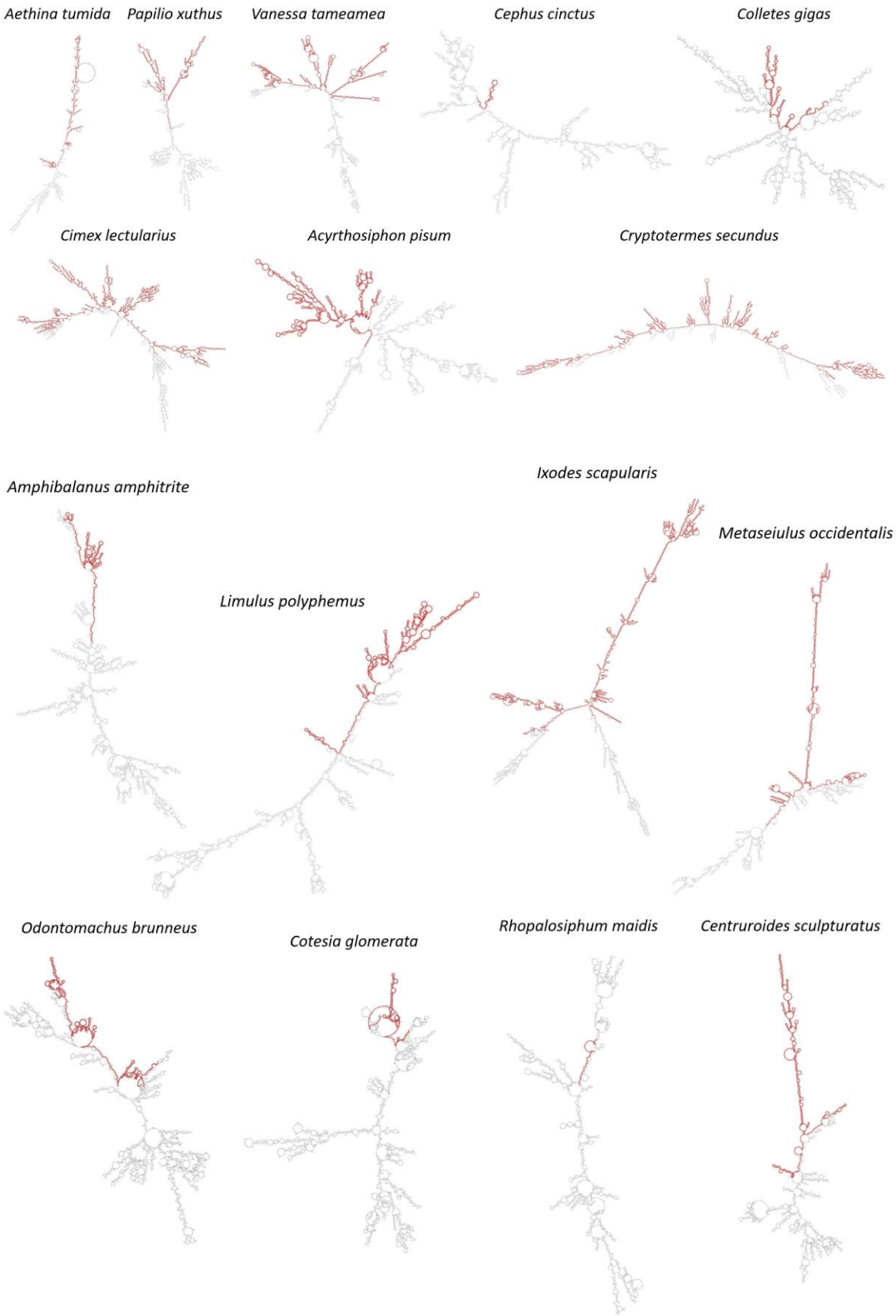


	Размер 1-го экзона в кассете	Размер интрона	Позиция стоп-кодона в интроне
Drosophila melanogaster	**110**	**1602**	**13**
Bactrocera dorsalis	110	1808	13
Ceratitis capitata	110	2023	19

Визуализация вторичной структуры кассетного интрона гена *Nxf1* у представителей отряда *Diptera*.

Зеленым цветом обозначен участок кассетного интрона. Красным цветом выделены поли(А)-последовательности внутри кассетного интрона.

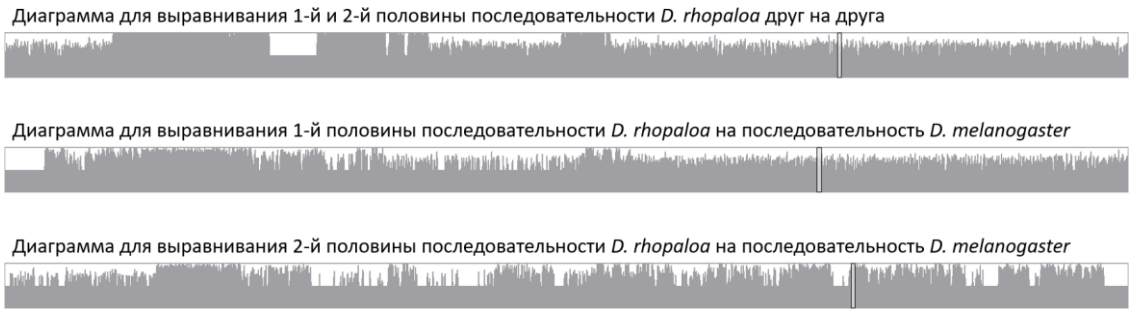
Приложение 9



Визуализация вторичной структуры для кассетного интрона гена *Nxf1* у разных представителей *Arthropoda*.

Красным цветом обозначен кассетный интрон.

Приложение 10



Сравнение двух участков последовательности *Drosophila rhopaloa*.