Министерство науки и высшего образования РФ
Федеральное государственное автономное образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Б1.В.07 Языки программирования для биоинформатиков

наименование дисциплины (модуля) в соответствии с учебным планом

Направление подготовки / специальность

06.04.01 Биология

Направленность (профиль)

06.04.01.06 Геномика и биоинформатика

Форма обучения очная

Год набора 2022

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Программу составили _____

к.тех.н., Доцент, Кузьмин Дмитрий Александрович

должность, инициалы, фамилия

1 Цели и задачи изучения дисциплины

1.1 Цель преподавания дисциплины

Целью изучения дисциплины является формирование у магистров знаний об организации и методах проведения исследований в области практической биоинформатики, освоение современного программного обеспечения для решения биоинформатических задач и получение навыков его практического применения.

1.2 Задачи изучения дисциплины

В задачи курса входит:

- освоение навыков работы с unix-подобными операционными системами,
- освоение навыков программирования командных файлов на unix -shell (на примере языка bash),
- решение задач по обработке биоинформатических данных с использованием утилит grep, awk и sed, изучение особенностей программирование на awk,
- изучение основ языка программирования высокого уровня (на примере Python) и формирование навыков использования скриптовых языков программирования,
- обзор специализированных языков программирования используемых в решении задач биоинформатики,
- знакомство с алгоритмами сборки геномов de novo и на основе референсной последовательности, знакомство с методами аннотирования геномов,
- знакомство с основами структурной биоинформатики.

1.3 Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Код и наименование индикатора достижения компетенции	Запланированные результаты обучения по дисциплине
ПК-1: Способен осуществлять выбор форм и методов научно-исследовательской деятельности в соответствии с профилем научного исследования	
ПК-1.2: Способен: - решать задачи, связанные с проведением исследований с использованием современных методических подходов и специализированного оборудования	
ПК-3: Способен выполнять работы, связанные с исследованием и анализом генома и протеома живых организмов в т. ч. в областях здравоохранения, лесного хозяйства и охраны природы.	

ПК-3.1: Умеет: - в полном объеме	
планировать и реализовывать проведение лабораторных молекулярно-генетических исследований живых организмов; - планировать и реализовывать проведение работ с биоинформационными ресурсами.	
ПК-3.2: Владеет: - современными методами обработки и интерпретации генетической информации при проведении научных исследований; - методами обработки данных геномного секвенирования, полученных с разных платформ; способностью извлекать необходимые данные из банков генетических данных; - знаниями для обработки полученных результатов, анализа и осмысливания их с учетом имеющихся литературных данных.	
ПК-3.3: Способен: - использовать знания геномики и биоинформатики для объяснения важнейших биохимических процессов, протекающих в живых организмах, как в норме, так и при возникновении патологий; ориентироваться в вопросах, связанных с анализом нуклеиновых кислот и белков;	

1.4 Особенности реализации дисциплины

Язык реализации дисциплины: Русский.

Дисциплина (модуль) реализуется с применением ЭО и ДОТ

URL-адрес и название электронного обучающего курса: https://e.sfu-kras.ru/course/view.php?id=12486.

2. Объем дисциплины (модуля)

Вид учебной работы	Всего, зачетных единиц (акад.час)	е
		1
Контактная работа с преподавателем:	**1,33 (48)**	
занятия лекционного типа	0,44 (16)	
практические занятия	0,89 (32)	
Самостоятельная работа обучающихся:	**1,67 (60)**	
курсовое проектирование (КП)	Нет	
курсовая работа (КР)	Нет	

3 Содержание дисциплины (модуля)

3.1 Разделы дисциплины и виды занятий (тематический план занятий)

№ п/п	Модули, темы (разделы) дисциплины	Контактная работа, ак. час.							
		Занятия лекционного типа		Занятия семинарского типа				Самостоятельная работа, ак. час.	
				Семинары и/или Практические занятия		Лабораторные работы и/или Практикумы			
		Всего	В том числе в ЭИОС	Всего	В том числе в ЭИОС	Всего	В том числе в ЭИОС	Всего	В том числе в ЭИОС
1.									
	1. Введение в Linux для биоинформатиков Основы работы в Linux. Особенности системы. Основные команды. Работа в терминале. Создание командных файлов. Ввод/вывод. Работа с файлами в сети. Работа с архивами. Обработка биоинформатических данных с использованием утилит grep, awk и sed. Использование регулярных выражений Программирование командных файлов на языке bash. Работа с удаленным сервером. Знакомство с сервером. Обмен файлами. Запуск приложений. Контроль запускаемых программ. Многопоточные приложения. PBS Troque.	4							

	2. Python для биоинформатиков. Введение в Python. Функции в языке Python. PEP8. Синтаксис Python. Модель данных, объекты, типы и классы. Области видимости, пространства имен и классы. Работа с файлами в Python. Модули, импорт, модели. Регулярные выражения. Работа с данными в Интернет. Итераторы и генераторы . Структурное программирование, numpy, matplotlib	6							
	3. Специализированные языки программирования для решения задач Языки R и Julia - особенности, обзор возможностей.	2							

	4. Сборка и аннотирование геномов. Ассемблирование геномов. Сборка геномов de novo. Алгоритмы ассемблирования геномов. Графы Де Брёйна. Сборка геномов по референсу. Ресеквенирование геномов и сборка по референсу. Выравнивание коротких прочтений. Ассемблирование транскриптомов. Особенности сборки транскриптомов. Методы оценки качества транскриптомных и геномных сборок. Анализ данных RNA-Seq: анализ дифференциальной экспрессии генов. Аннотирование геномов. Алгоритмы предсказания генных моделей. Особенности аннотирования геномов прокариот и эукариот. Пайплайны для автоматического аннотирования геномов. Функциональная аннотация. Gene Ontology. Rfam и InterPro. Оценка качества результатов ассемблирования и аннотации. Анализ генных моделей эукариот.	4							
2.									

1. Введение в Linux для биоинформатиков. 1. Основные команды. Работа в терминале. Исполняемые файлы. 2. Ввод/вывод. Работа с файлами в сети. Работа с архивами. Поиск файлов и слов в строках. 3. Знакомство с сервером. Обмен файлами. Запуск приложений. Установка приложений. Контроль запускаемых программ. Многопоточные приложения. PBS Troque. 4. Текстовый редактор vim. Bash: основные команды, ветвления и циклы. Часто используемые в биоинформатике команды.			8					
2. Программирование на Python. 1.Функции в языке Python. 2.Модель данных, объекты, типы и классы. Области видимости, пространства имен и классы. 3.Работа с файлами в Python 4.Модули, импорт, модели. 5.Регулярные выражения 6.Работа с данными в Интернет			12					
3. Специализированные языки программирования для решения задач Языки R и Julia – знакомство с базовым синтаксисом и возможностями.			4					

4. Сборка и аннотирование геномов. 1. Ассемблирование геномов. Сборка геномов de novo различными ассемблерами (SPAdes, SOAPdenovo, CLC Assembly Cell, Abyss и др.). Оценка качества геномной сборки. 2. Сборка геномов по референсу. Ресеквенирование геномов и сборка по референсу. Выравнивание коротких прочтений при помощи BWA, Bowtie2. Работа с SAM-файлами, поиск вариантов (variant calling). 3. Ассемблирование транскриптомов. Сборка транскриптомов de novo при помощи Trinity. Оценка качества транскриптомной сборки. 4. Анализ данных RNA-Seq: анализ дифференциальной экспрессии генов. 5. Аннотирование геномов прокариот. Пайплайны для автоматического аннотирования геномов прокариот. 6. Аннотирование геномов эукариот. Пайплайны для автоматического аннотирования геномов эукариот. 7. Функциональная аннотация. Gene Ontology. Rfam и InterPro. Оценка качества результатов ассемблирования и аннотации. 8. Анализ генных моделей эукариот.			8					
3.								
1. Введение в Linux для биоинформатиков.							15	
2. Программирование на Python.							15	
3. Специализированные языки программирования для решения задач.							15	
4. Сборка и аннотирование геномов.							15	

Всего	16		32				60	

4 Учебно-методическое обеспечение дисциплины

4.1 Печатные и электронные издания:

1. Колисниченко Д. Н. Linux. От новичка к профессионалу: наиболее полное руководство(Санкт-Петербург: БХВ-Петербург).
2. Лав Р., Сивченко О. Linux. Системное программирование(Санкт-Петербург: Питер).
3. Игнасимуту С. Основы биоинформатики: перевод с английского (МоскваМосква: [R&C Dynamics] Регулярная и хаотическая динамика [РХД]).
4. Глик Б., Пастернак Д., Янковский Н. К. Молекулярная биотехнология: принципы и применение: перевод с английского(Москва: Мир).
5. Леск А., Миронов А. А., Швядас В. К. Введение в биоинформатику: учеб. пособие: пер. с англ.(Москва: БИНОМ, Лаборатория знаний).
6. Хаубольд Б., Вие Т., Чудов С. В., Артамонова И. И. Введение в вычислительную биологию. Эволюционный подход(Москва: Регулярная и хаотическая динамика).
7. Колесниченко Д. Н. Самоучитель Linux. Установка, настройка, использование: [самоучитель](Санкт-Петербург: Наука и техника).
8. Кузьмин Д. А., Удалова Ю. В. Разработка компонентов системного программного обеспечения. Процессы в Linux: учеб.-метод. пособие для студентов спец. 010501, 090102, 230100(Красноярск: СФУ).

4.2 Лицензионное и свободно распространяемое программное обеспечение, в том числе отечественного производства (программное обеспечение, на которое университет имеет лицензию, а также свободно распространяемое программное обеспечение):

1. Современные биоинформатические исследования требуют умения решать поставленные задачи с использованием самого разнообразного программного обеспечения, от пользовательских скриптов, размещенных в репозиториях, до дорогостоящего проприетарного ПО, такого как CLCbio. Философия современного биоинформатического сообщества заключается в том, что любую задачу можно решить несколькими способами: с использованием бесплатно распространяемого ПО, при помощи онлайн-сервисов (пайплайнов) и проприетарного ПО, или самостоятельно создать новый программный продукт для решения конкретной пользовательской задачи. В рамках данного курса используется только свободно распространяемое ПО: BLAST, FastQC, Trimmomatic, ABySS, MaSuRCA, SPAdes, Bowtie2, BWA, Samtools, GATK, SSPACE , MAKER , Trinity, Trinotate, Blast2GO, QUAST, UGENE, MEGA, BioEdit.

4.3 Интернет-ресурсы, включая профессиональные базы данных и информационные справочные системы:

1. Биоинформатика – та область знаний, в которой ресурсы Интернет используются практически для решения любой задачи. Вся биоинформатика основана на создании баз данных, наполнении их результатами научных работ исследователями со всего мира, открытости доступа к этим данным и сравнении новых результатов с уже опубликованными.
2. В рамках освоения дисциплины используется одна из крупнейших информационных систем в области биологии медицины, биофизики Национального центра биотехнологической информации (National Center for Biotechnology Information (NCBI)), США (www.NCBI.nlm.nih.gov).
3. БД NCBI являются достаточно сложным инструментарием с разнообразным функционалом.
4. Ниже приведено краткое описание основных БД NCBI, которые могут быть полезны при освоении тем дисциплины.
5. БД Nucleotide (http://www.NCBI.nlm.nih.gov/sites/Entrez?db=nucleotide) объединяет данные последовательностей нуклеиновых кислот из нескольких исходных БД, в том числе GenBank, RefSeq и др. Данные могут быть найдены по регистрационному номеру, имени автора, наименованию организма, генома/белка, а также ряду других параметров.
6. БД Protein (http://www.NCBI.nlm.nih.gov/sites/Entrez?db=protein) является коллекцией аминокислотных последовательностей из нескольких источников, в том числе из GenBank, RefSeq и TPA, а также SwissProt, PIR, PRF и PDB.
7. БД Structure (http://www.NCBI.nlm.nih.gov/Structure/index.shtml) организуют доступ к результатам молекулярного моделирования макромолекул и связанным с ними БД: трехмерных биомолекулярных структур полученных с помощью рентгеновской кристаллографии и ЯМР-спектроскопии; БД химических структур небольших органических молекул; к информации об их биологической активности и т. д.
8. БД Gene (http://www.NCBI.nlm.nih.gov/sites/Entrez?db=gene) представляет собой инструмент для просмотра данных из широкого спектра геномов. Каждая запись – это один из генов определенного организма. Минимальный набор данных в гене запись включает уникальный идентификатор, т. н. Gene-ID.
9. БД dbMHC (http://www.NCBI.nlm.nih.gov/gv/mhc/main.cgi?cmd=init) предоставляет открытую платформу, где научное сообщество может размещать, просматривать и редактировать данные MajorHistocompatibilityComplex (MHC) для человека. БД dbMHCполностью интегрирована с другими ресурсами NCBI, а также с Международной рабочей группой гистосовместимости (IHWG).
10. DbSNP (http://www.NCBI.nlm.nih.gov/SNP/) – БД одиночных нуклеотидных полиморфизмов, полиморфных повторяющихся элементов, включающая как гибридные данные, так и полученные только экспериментальным путем.

11. БД ReferenceSequence (RefSeq) (http://www.NCBI.nlm.nih.gov/RefSeq/), содержащая последовательности, в том числе геномных ДНК, белков и т. д., является основой для проведения функциональных исследований, генной идентификации, сравнительного анализа и т. п. В частности, релиз от 11.07.2012 включал в себя описания 16 393 342 белков и 17 605 организмов.
12. БД Genomic Biology представляет собой объединение нескольких ресурсов и инструментов геномной биологии, в том числе геномных карт для Fruitfly, Human, Malariaparasite, Mouse, Rat, Retroviruses, Zebrafish и т. д., которые дополнительно содержат ссылки на интернет-ресурсы и БД, касающиеся рассматриваемых видов.
13. В БД UniGene (http://www.NCBI.nlm.nih.gov/unigene/) полноразмерные mRNA последовательности организованы в уникальные кластеры, представляющие известные или предполагаемые гены. Для кластеров доступна информация по картированию, экспрессии и другие ресурсы.
14. HomoloGene (http://www.NCBI.nlm.nih.gov/homologene) – инструмент для автоматизированного выявления гомологов среди аннотированных генов, который сравнивает нуклеотидные последовательности между парами организмов в целях выявления предполагаемых ортологов.
15. Basic Local Alignment Search Tool (http://www.NCBI.nlm.nih.gov/BLAST/) - основной метод поиска гомологичных последовательностей на основе локального выравнивания.
16. Public repository Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) - публичная электронная библиотека данных экспрессии генов «Омнибус Экспрессии Генов»
17. GenBank (http://www.NCBI.nlm.nih.gov/genbank/index.html) – БД, содержащая доступные последовательности нуклеотидов для более чем 260 000 организмов, вся информация в генетическом банке данных сопровождается библиографическими ссылками и биологическими аннотациями. GenBank автоматически интегрирует информацию о геноме и БД белковых последовательностей для изучения, учитывая таксономию, геном, белковую структуру и другую информацию.
18. Для представления последовательностей в GenBank предложено два инструмента:
19. • BankIt – интернет-представление одной или нескольких последовательностей;
20. • Sequin – интернет-представление для длинных последовательностей, полных геномов, результатов популяционных и филогенетических исследований.

21. Объединяющим фактором и при этом крайне удобным инструментом поиска в NCBI является поисковая система Search NCBI databases (http://www.NCBI.nlm.nih.gov/sites/gquery). Она обеспечивает одновременный доступ как к нуклеотидным и белковым последовательностям (GenBank, EMBL, DDBJ, PIR-International, PRF, Swiss-Prot и PDB, GenPept, RPF), 3-мерным структурам и популяционным данным, так и к библиографическим БД (PubMed, PubMed Centralи т. д.). Доступ к поисковой системе Search NCBI databases может быть легко получен с помощью прямого интернет-адреса (http://www.NCBI.nlm.nih.gov/gquery/) либо посредством использования стартовой страницы NCBI (http://www.NCBI.nlm.nih.gov/). На этой странице приведен полный перечень инструментария и БД NCBI и существует возможность получить доступ к любой из перечисленных БД.
22. Крайне полезным инструментом, который сохраняет информацию о пользователе, используется для более точной настройки поисковых запросов в NCBI (http://www.NCBI.nlm.nih.gov/index.html) и т. д., является сервис «My NCBI» (http://www.NCBI.nlm.nih.gov/sites/My NCBI/). Этот инструмент позволяет сохранять результаты поиска, выбирать форматы отображения, фильтрации, настраивать автоматический поиск и отправлять его результаты по электронной почте. Пользователи «My NCBI» могут сохранять свои БД, построенные на основе поисковых запросов в NCBI, и управлять политикой общественного доступа.

5 Фонд оценочных средств

Оценочные средства находятся в приложении к рабочим программам дисциплин.

6 Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине (модулю)

Аудиторный класс, наличие проектора для демонстрации наглядных пособий и экрана. Компьютерный класс, лицензионное программное обеспечение, Internet.