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

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Б1.В.06 Статистика в биоинформатике

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.ф.-м.н., Доцент, Шуваев Андрей Николаевич

должность, инициалы, фамилия

1 Цели и задачи изучения дисциплины

1.1 Цель преподавания дисциплины

Цель изучения дисциплины – формирование у магистров навыков, необходимых для статистического анализа биологических данных на примере биоинформационных данных.

1.2 Задачи изучения дисциплины

В задачи курса входит:

- изучение основных принципов статистической обработки биологических данных;
- изучение основ языка программирования R;
- освоение навыков статистической обработки биологических данных с использованием пакетов статистической обработки и языков программирования (на примере R).

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	
Промежуточная аттестация (Экзамен)	**1 (36)**	

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. 1. Основы теории вероятности. Основные понятия и классификации. Предмет и методы. Этапы статистического исследования. Статистическое наблюдение	1							
	2. 2. Знакомство и работа в R. Установка R, работа с консолью. Типичные ошибки, возникающие при вычислениях. Список базовых команд. Визуализация данных.	1							

	3. 3. Описательная статистика: пределы и распределения. Генеральная совокупность и выборка. Характер распределения признаков. Закон нормального распределения. Показатели описательной статистики. Среднее значение. Ошибка среднего, стан-дартное отклонение, доверительный интервал, медиана, ассиметрия, эксцесс. Коэффициент вариации и уровень изменчивости признаков.	2							
	4. 4. Метод ресэмплинга. Пермутации в R. Формирование случайных наборов данных из имеющегося. Сходимость при увеличении числа случайных наборов.	2							
	5. 5. Методы получения оценок. Доверительные интервалы. Метод центральной статистики, Асимптотические доверительные интервалы. Построение асимптотических доверительных интервалов. Метод максимального правдоподобия. Экстремальное свойство функции правдоподобия. Состоятельность оценки максимального правдоподобия. Байесовские оценки, их оптимальность в байесовском подходе к сравнению оценок. Сопряженное распределение. Неравенство Крамера-Рао и эффективные оценки. Критерий эффективности оценки.	2							

	6. 6. Проверки статистических гипотез. Общие принципы и основные понятия, ошибки I и II родов, уровень значимости, мощность критерия. Распределения хи квадрат, Стьюдента, Фишера, их свойства. Метод ANOVA. Несмещенность и состоятельность статистического критерия, p-значение.	2							
	7. 7. Теория корреляции. Оценка наименьших квадратов, ее основные свойства. Линейная регрессионная модель. Коэффициенты корреляции Пирсона, Спирмена, Кендалла. Обобщенный коэффициент корреляции.	2							
	8. 8. Непараметрические методы. Статистика Манна-Уитни, критерий Краскела-Уоллиса, тест серий (Wald–Wolfowitz test), ранговая корреляция Спирмана	2							
	9. 9. Классификация. Кластеризация. Иерархические и неиерархические методы кластерного анализа. Метод к-средних. Общие представления о классификации. Формальные основания классификации. Методы кластер-анализа. Результаты иерархической кластеризации. Дендрограммы. Правила и методы построения кластеров. Правила объяснения результатов. Метод главных компонент. Отбор переменных для анализа главных компонент. Интерпретация.	2							
2.									
	1. Основы теории вероятности. Решение задач по теории вероятности.			2					

	2. Подходы к проектированию эксперимента. Работа с общедоступными данными. Знакомство и изучение синтаксиса R.			2					
	3. Визуализирование данных Работа с пакетами в R для визуализирования данных.			4					
	4. Описательная статистика: пределы и распределения. Проверка гипотез о различии распределений (двухвыборочные и одновыборочные тесты): Базовый анализ номинативных данных. Визуализация результатов.			4					
	5. Методы получения оценок. Применение метода максимального правдоподобия, байесовского метода для анализа биологических данных в R.			4					
	6. Проверки статистических гипотез. Применение различных статистических методов (критерий Стьюдента, Фишера, хи-квадрат) на биологических данных в R. Визуализация результатов. Дисперсионный анализ (ANOVA). Однофакторный дисперсионный анализ, многофакторный дисперсионный анализ.			4					
	7. Теория корреляции. Применение на биологических данных метода наименьших квадратов, линейной регрессии. Изучение корреляции Пирсона, Спирмена, Кендалла на биологических данных. Визуализация данных корреляционного анализа и сравнения групп.			4					

8. Непараметрические методы. Применение непараметрический статистических критериев (Статистика Манна-Уитни, критерий Краскела-Уоллиса, тест серий (Wald–Wolfowitz test), ранговая корреляция Спирмана) на биологических данных. Обсуждение условий применения того или иного критерия. Визуализация результатов.			4					
9. Классификация Применение иерархических и неиерархических методов кластеризации на биологических данных в R. Применение метода главных компонент, грамотный подбор компонент для визуализации. Визуализация кластеров.			4					
3.								
1. Основы теории вероятности.							6	
2. Подходы к проектированию эксперимента.							8	
3. Визуализирование данных.							6	
4. Описательная статистика: пределы и распределения.							6	
5. Методы получения оценок.							8	
6. Проверки статистических гипотез.							8	
7. Теория корреляции.							6	
8. Непараметрические методы.							6	
9. Классификация.							6	
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. Одной из крупнейших информационных систем в области биологии медицины, биофизики является Национальный центр биотехнологической информации (National Center for Biotechnology Information (NCBI), США (www.NCBI.nlm.nih.gov). БД NCBI являются достаточно сложным инстру-ментарием с разнообразным функционалом.
2. Ниже приведено краткое описание основных БД NCBI, которые мо-гут быть полезны при освоении тем дисциплины.
3. БД Nucleotide (http://www.NCBI.nlm.nih.gov/sites/Entrez?db=nucleotide) объединяет дан-ные последовательностей нуклеиновых кислот из нескольких исходных БД, в том числе GenBank, RefSeq и др. Данные могут быть найдены по ре-гистрационному номеру, имени автора, наименованию организма, гено-ма/белка, а также ряду других параметров.
4. БД Protein (http://www.NCBI.nlm.nih.gov/sites/Entrez?db=protein) яв-ляется коллекцией аминокислотных последовательностей из нескольких источников, в том числе из GenBank, RefSeq и TPA, а также SwissProt, PIR, PRF и PDB.
5. БД Structure (http://www.NCBI.nlm.nih.gov/Structure/index.shtml) ор-ганизуют доступ к результатам молекулярного моделирования макромо-лекул и связанным с ними БД: трехмерных биомолекулярных структур полученных с помощью рентгеновской кристаллографии и ЯМР-спектроскопии; БД химических структур небольших органических моле-кул; к информации об их биологической активности и т. д.
6. БД Gene (http://www.NCBI.nlm.nih.gov/sites/Entrez?db=gene) представ-ляет собой инструмент для просмотра данных из широкого спектра гено-мов. Каждая запись – это один из генов определенного организма. Мини-мальный набор данных в гене запись включает уникальный идентифика-тор, т. н. Gene-ID.
7. БД dbMHC (http://www.NCBI.nlm.nih.gov/gv/mhc/main.cgi?cmd=init) предоставляет открытую платформу, где научное сообщество может раз-мещать, просматривать и редактировать данные MajorHistocompatibilityComplex (MHC) для человека. БД dbMHC полно-стью интегрирована с другими ресурсами NCBI, а также с Международ-ной рабочей группой гистосовместимости (IHWG).
8. DbSNP (http://www.NCBI.nlm.nih.gov/SNP/) – БД одиночных нуклео-тидных полиморфизмов, полиморфных повторяющихся элементов, вклю -чающая как гибридные данные, так и полученные только эксперименталь-ным путем.
9. 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, ген -ной идентификации, сравнительного анализа и т. п. В частности, релиз от 11.07.2012 включал в себя описания 16 393 342 белков и 17 605 организ-мов.

10. БД Genomic Biology представляет собой объединение нескольких ре-сурсов 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, ка-сающиеся рассматриваемых видов.
11. В БД UniGene (http://www.NCBI.nlm.nih.gov/unigene/) полноразмер-ные mRNA последовательности организованы в уникальные кластеры, представляющие известные или предполагаемые гены. Для кластеров до-ступна информация по картированию, экспрессии и другие ресурсы.
12. HomoloGene (http://www.NCBI.nlm.nih.gov/homologene) – инструмент для автоматизированного выявления гомологов среди аннотированных генов, который сравнивает нуклеотидные последовательности между па-рами организмов в целях выявления предполагаемых ортологов.
13. Basic Local Alignment Search Tool (http://www.NCBI.nlm.nih.gov/BLAST/) - основной метод поиска гомологич-ных последовательностей на основе локального выравнивания.
14. Public repository Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) - публичная электронная библиотека данных экспрессии генов «Омнибус Экспрессии Генов»
15. GenBank (http://www.NCBI.nlm.nih.gov/genbank/index.html) – БД, со-держащая доступные последовательности нуклеотидов для более чем 260 000 организмов, вся информация в генетическом банке данных сопро -вождается библиографическими ссылками и биологическими аннотациями. GenBank автоматически интегрирует информацию о геноме и БД белковых последовательностей для изучения, учитывая таксономию, геном, белко-вую структуру и другую информацию.
16. Для представления последовательностей в GenBank предложено два инструмента:
17. ● BankIt – интернет-представление одной или нескольких последова-тельностей;
18. ● Sequin – интернет-представление для длинных последовательно-стей, полных геномов, результатов популяционных и филогенетических исследований.

19. 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 одновремен -ный 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 библио-графическим БД (PubMed, PubMed Centralи т. д.). Доступ к поисковой си-стеме Search NCBI databases может быть легко получен с помощью прямо-го интернет-адреса (http://www.NCBI.nlm.nih.gov/gquery/) либо посред-ством использования стартовой страницы NCBI (http://www.NCBI.nlm.nih.gov/). На этой странице приведен полный пере-чень инструментария и БД NCBI и существует возможность получить до-ступ к любой из перечисленных БД.
20. Крайне полезным инструментом, который сохраняет информацию о пользователе, используется для более точной настройки поисковых запро -сов 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 резуль-таты 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.