ФАНО РОССИИ

Федеральное государственное бюджетное учреждение науки
Институт общей генетики им. Н.И. Вавилова
Российской академии наук
(ИОГен РАН)

«ПРИНЯТО»	**«УТВЕРЖДАЮ»**
На заседании Ученого совета ИОГен РАН	Зам. директора ИОГен РАН
Протокол № 2 от «20» февраля 2018 г.	д.б.н. _____ Ю.А. Столповский

РАБОЧАЯ ПРОГРАММА
ФАКУЛЬТАТИВНОЙ ДИСЦИПЛИНЫ
«Основы R»

Направление подготовки: 06.06.01 Биологические науки

Уровень образования: высшее образование – подготовка кадров высшей квалификации

Квалификация выпускника: Исследователь. Преподаватель-исследователь.

Разработчик:
к. б. н., С.А. Брускин
Е.В. Чекалин

Москва, 2018 г.

Рабочая программа составлена на основании федерального государственного образовательного стандарта, разработанного для реализации основных профессиональных образовательных программ высшего образования - программ подготовки научно-педагогических кадров в аспирантуре по направлению подготовки кадров высшей квалификации 06.06.01«Биологические науки».

Согласно Федеральному государственному образовательному стандарту высшего образования по направлению подготовки 06.06.01 Биологические науки (уровень подготовки кадров высшей квалификации), утвержденному приказом Минобрнауки РФ № 871 от 30 июля 2014 г., и учебному плану аспирантов, разработанного на основе этого стандарта, дисциплина «Основы R» является факультативной дисциплиной.

Объём курса составляет 2 зачетные единицы или 72 академических часа, из них 30 академических часов лекций, 20 академических часов семинаров и 22 академических часов самостоятельной внеаудиторной работы аспирантов, включая подготовку к зачету.

1. Цели и задачи

Цель дисциплины

Дать практические навыки применения языка R, IDE R-studio и основных пакетов для анализа данных.

Задачи дисциплины

• обучить основам языка программирования R;

• научить пользоваться IDE R-studio;

• ознакомить с основными пакетами для анализа данных с помощью R;

• формирование у аспирантов основных биоинформатических навыков и приобретение ими практического опыта, необходимого для проведения самостоятельных научных исследований в области системной биологии.

2. Место дисциплины (модуля) в структуре образовательной программы бакалавриата (магистратуры)

Дисциплина «Основы языка R» включает в себя разделы, которые могут быть отнесены к вариативной части цикла Б.3.

Дисциплина «Задачи биоинформатики» базируется на дисциплинах:

Информатика.

Основы биологии;

Молекулярная биология;

Генетика;

Эволюционная биология;

Дисциплина «Основы языка R» предшествует изучению дисциплин:

Основные алгоритмы биоинформатики;

Дополнительные главы биоинформатики;

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

Универсальные компетенции

- способность к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях (УК-1);

- способность проектировать и осуществлять комплексные исследования, в том числе междисциплинарные, на основе целостного системного научного мировоззрения с использованием знаний в области истории и философии науки (УК-2);

- готовность участвовать в работе российских и международных исследовательских коллективов по решению научных и научно-образовательных задач (УК-3);

- способность планировать и решать задачи собственного профессионального и личностного развития (УК-5).

Общепрофессиональные компетенции

- способность самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий (ОПК-1);

- готовность к преподавательской деятельности по основным образовательным программам высшего образования (ОПК-2).

Профессиональные компетенции

- способность к самостоятельному проведению научно-исследовательской работы и получению научных результатов, удовлетворяющих установленным требованиям к содержанию диссертаций на соискание ученой степени кандидата наук по направленной специальности (ПК-1);

- обладание представлениями о системе фундаментальных понятий и методологических аспектов биологии, форм и методов научного познания (ПК-2);

- способность приобретать новые знания с использованием современных научных методов и владение ими на уровне, необходимом для решения задач, возникающих при выполнении профессиональных функций (ПК-3).

В результате освоения дисциплины обучающиеся должны

знать:

-Основные пакеты программной среды R;

-Основы синтаксиса R;

уметь:

- Программировать на языке R

- Имплементировать и отлаживать биоинформатические алгоритмы;

- Реализовывать статистический анализ в программной среде R.

владеть:

- навыками работы с большими объемами биологических данных;

- культурой планирования и осуществления многоступенчатого биоинформатического анализа.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Виды учебных занятий, включая самостоятельную работу		
		Лекции	**Практич. (семинар.) задания**	**Самост. работа**
1	Введение в язык программирования R и IDE R-studio.	2	2	2
2	Синтаксис R	2	2	2
3	Операторы и функции	2	2	2
4	Базовая графика в R	2	2	2
5	Реализация графиков с помощью пакета ggplot2	2	2	2
6	Корреляция и линейная регрессия	4	2	2
7	PCA и Heatmaps	2	2	2
8	Кластеризация в R;	4	2	2
9	NGS и поиск дифференциально экспрессирующихся генов;	4	2	2
10	Пакеты для работы с сиквенсами. Завершающее занятие.	4	2	4
	Итого часов	**30**	**20**	**22**
	Общая трудоёмкость	**72 час., 2 зач.ед.**		

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

1. Введение в язык программирования R и IDE R-studio.

Ознакомительное занятие. Основные команды в R. Настройка рабочего пространства R-studio. Установка пакетов и обновлений.

2. Синтаксис R;

Реализация кода на R. Синтаксис R. Основные объекты в R, понятие переменных, массива, матрицы, data.frame, list, array.

3. Операторы и функции;

Операторы в R. Реализация функций на языке R, функции первого уровня. Функции n-ного уровня. Предполагается, что по результатам модуля аспиранты смогут реализовывать собственные функции на языке R, а также манипулировать переменными с помощью основных операторов в R.

4. Базовая графика в R;

Основы графики в R. Построение двумерных графиков на плоскости с помощью пакета base(). Boxplot, barplot, pie plot, dot plot, histogram. Регулируемые параметры объекта plot и par(). Легенда на графиках base(). Параметры осей графика в base plot.

5. Реализация графиков с помощью пакета ggplot2;

Реализация графиков с помощью пакета ggplot2. Слои в объектах ggplot2. Дополнительные функции пакета ggplot2: geom_point, geom_abline, geom_polygon, geom_rect. Facet и grid на графиках ggplot2. Регулируемые параметры окружения графика. Main(), axis labesl. Легенда в ggplot2

6. Корреляция и линейная регрессия.

Понятие о корреляции. Корреляция по Пирсону и Спирену. Визуализация данных. Линейная регрессия. Множественная линейная регрессия. R2 и F-статистика. Тестирование моделей. Тестовая и обучающая выборка. ANOVA. Glm, generalized linear model. Logit-регрессия и AIC. Поправка на множественное сравнение. FDR, поправка Бонферрони.

7. PCA и Heatmaps;

Анализ главных компонент. Общие принципы реализации PCA. Функции prcomp пакета stats. Пакет pca3d. Теплокарты. Пакеты heatmap и heatmap.2. Визуализация данных с помощью PCA и heatmap.

8. Кластеризация в R;

Основные алгоритмы кластеризации. Euclidean, manhattan, maximum, canberra. Иерархическая кластеризация. Расстояние между кластерами. Comlete, Single, Average linkage. Дендрограмма как объект R. K-means, k-medoids. Self-organizing map. Silhouette.

9. NGS и поиск дифференциально экспрессирующихся генов;

Общие принципы секвенирования следующего поколения. Тримминг и QC ридов. FastQC и trimmomatic. Обзор алгоритмов выравнивания. BWA, Bowtie, STAR. Выравнивание ридов на геном. Детекция альтернативного сплайсинга. Cufflinks. Cuffdiff. Нормализация ридов. DEXseq, EdgeR, limma. Log fold-change. Поиск дифференциально экспрессирующихся генов и идентификация достоверных DEGов.

10. Пакеты для работы с сиквенсами. Завершающее занятие.

Основные функции пакетов Seqinr, Ape, аннотирование генов с помощью пакета GenomicRanges, GenomicAlignments. Go-enrichent. Гипергеометрическое распределение.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Оборудование, необходимое для лекций и семинаров: компьютер, Windows, MS Office и мультимедийное оборудование (проектор, звуковая система)

6. Перечень основной и дополнительной литературы, необходимой для освоения дисциплины (модуля)

Основная литература

Team, R. C. (2013). R: A language and environment for statistical computing.

Шипунов, А. Б., Балдин, Е. М., Волкова, П. А., Коробейников, А. И., Назарова, С. А., Петров, С. В., & Суфиянов, В. Г. (2012). Наглядная статистика. Используем R!. *М.: ДМК Пресс*, *298*, 1.

7. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю) (необязательный пункт)

8. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля) (необязательный пункт)

Научно-библиографические и патентные базы данных в области физико-химической биологии, доступные по сети Интернет в бесплатном режиме - Science Citation Index (Web of Science), Medline (PubMed), Научная электронная библиотека (НЭБ), Российская патентная БД ФГУ ФИПС и американская патентная БД USPAFULL; электронные адреса крупных научных издательств, предоставляющих доступ к полным текстам текущих и архивным выпускам этих журналов.

http://stackoverflow.com/

https://stat.ethz.ch/pipermail/r-help/
https://www.biostars.org/
https://www.statmethods.net/

9. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем (при необходимости)

Доступ в Интернет, UNIX сервер с отдельным аккаунтом для каждого аспиранта

10. Методические указания для обучающихся по освоению дисциплины

Аспирант, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

В результате изучения дисциплины аспирант должен знать основные определения дисциплины, уметь применять полученные знания для решения различных задач.

Успешное освоение курса требует:

– посещения всех занятий, предусмотренных учебным планом по дисциплине;

– ведения конспекта занятий;

– напряжённой самостоятельной работы аспиранта.

Самостоятельная работа включает в себя:

– чтение рекомендованной литературы;

– проработку учебного материала, подготовку ответов на вопросы, предназначенных для самостоятельного изучения;

– решение задач, предлагаемых аспирантам на занятиях;

– решение задач, предлагаемых аспирантам в качестве домашнего задания;

– подготовку к выполнению заданий текущей и промежуточной аттестации.

Показателем владения материалом служит умение без конспекта отвечать на вопросы по темам дисциплины.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями преподавателю.

Возможен промежуточный контроль знаний аспирантов в виде решения задач в соответствии с тематикой занятий.

11. Фонд оценочных средств для проведения промежуточной аттестации по итогам обучения

Приложение А.

Приложение А

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ «Основы R»

1. Компетенции, формируемые в процессе изучения дисциплины

Освоение дисциплины направлено на формирование у обучающегося следующих универсальных (УК), общепрофессиональных (ОПК) и профессиональных (ПК) компетенций:

Универсальные компетенции

- способность к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях (УК-1);
- способность проектировать и осуществлять комплексные исследования, в том числе междисциплинарные, на основе целостного системного научного мировоззрения с использованием знаний в области истории и философии науки (УК-2);
- готовность участвовать в работе российских и международных исследовательских коллективов по решению научных и научно-образовательных задач (УК-3);
- способность планировать и решать задачи собственного профессионального и личностного развития (УК-5).

Общепрофессиональные компетенции

- способность самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий (ОПК-1);
- готовность к преподавательской деятельности по основным образовательным программам высшего образования (ОПК-2).

Профессиональные компетенции

- способность к самостоятельному проведению научно-

исследовательской работы и получению научных результатов, удовлетворяющих установленным требованиям к содержанию диссертаций на соискание ученой степени кандидата наук по направленной специальности (ПК-1);

- обладание представлениями о системе фундаментальных понятий и методологических аспектов биологии, форм и методов научного познания (ПК-2);

- способность приобретать новые знания с использованием современных научных методов и владение ими на уровне, необходимом для решения задач, возникающих при выполнении профессиональных функций (ПК-3).

2. Показатели оценивания компетенций

В результате освоения дисциплины обучающиеся должны

знать:

-Основные пакеты программной среды R;

-Основы синтаксиса R;

уметь:

- Программировать на языке R

- Имплементировать и отлаживать биоинформатические алгоритмы;

- Реализовывать статистический анализ в программной среде R.

владеть:

- навыками работы с большими объемами биологических данных;

- культурой планирования и осуществления многоступенчатого биоинформатического анализа.

3. Перечень типовых контрольных заданий, используемых для оценки знаний, умений, навыков

1. Сгенерируйте два случайных Пуассоновских набора по 200 чисел, один с средним 0.5, другой с средним 3. Есть ли между ними линейная зависимость?
2. Какая структура подойдёт для хранения значений температуры измеряемой ежечасно у пяти больных за сутки. Создайте такую структуру и заполните ее случайными данными. Когда у второго пациента температура была выше 40 градусов?
3. Создайте новую числовую переменную new_var в данных mtcars, которая содержит единицы в строчках, если в машине не меньше четырёх карбюраторов (переменная "carb") или больше шести цилиндров (переменная "cyl"). В строчках, в которых условие не выполняется, должны стоять нули.
4. В датафрейме mtcars создайте новую колонку (переменную) под названием even_gear, в которой будут единицы, если значение переменной (gear) четное, и нули если количество нечетное.

5. Создайте 3 линейные переменные одинаковой длинны с произвольными названиями, содержащие Проверьте, действительно ли сумма первых двух чисел строго больше, чем третье число. Результат сравнения (TRUE или FALSE) сохраните в новую переменную с именем **result**.

6. Превратите датафрейм mtcars в лист и создайте новый элемент листа под названием even_gear, в которой будут единицы, если значение переменной (gear) четное, и нули если количество нечетное.

7. Создайте структуру для хранения имени, фамилии, возраста и пола трех человек и заполните ее. Как сделать простой поиск по фамилии?

8. Пьяный идет по мосту шириной *l* шагов. Каждый шаг пьяный смещается случайно на один шаг вправо или влево. Если пьяный переступит через край моста — он падает. Напишите функцию, которая считает, сколько шагов сделает пьяный до падения. Ширина моста и положение пьяного — начальные параметры. Drunken_path(l), вывод: ‘Our boozer will fall at n-d step’.

9. Напишите функцию, которая получает на вход две экспериментально померенные зависимости (x1,y1,x2,y2), аппроксимирует их прямыми методом наименьших квадратов и возвращает координаты пересечения.

10. Напишите функцию, которая сравнивает набор слов одинаковой длины и возвращает матрицу попарных расстояний — долю несовпадающих букв.

11. Напишите функцию, которая рисует траекторию луны относительно Солнца. Считать что земля движется равномерно со скоростью V1 вокруг Солнца по орбите радиуса R1. Луна движется равномерно со скоростью V2 вокруг земли по орбите радиуса R1.

12. Написать функцию которая рисует заданный набор точек на плоскости и соединяет прямыми те из них, что находятся на расстоянии менее заданного (максимальное расстояние — параметр функции). Размер точек пропорционален числу ребер входящих в нее (максимальный размер — параметр функции). Цвет прямых зависит от расстояния (цвет соответствующий минимальному и максимальному расстоянию — параметры функции).

13. Написать функцию находящую все вхождения слова (w) в текст (t) с не более чем n ошибок. W, t и n — параметры функции. Функция возвращает позиции начала вхождений.

14. Написать функцию находящую все вхождения слова (w) в текст (t) с не более чем n ошибок. W, t и n — параметры функции. Функция возвращает позиции начала вхождений.

15. Написать функцию рисующую случайную ломанную. Длина шага и поворот определяется (от предыдущего направления) определяется случайно исходя из равномерного распределения. Диапазон значений длин и поворота, равно как и число шагов — параметры функции..

16. Напишите функцию, генерирующую все возможные последовательности данной длины из данного алфавита.

17. Для встроенных в R данных AirPassengers рассчитайте скользящее среднее с интервалом сглаживания равным 10. Напечатайте получившийся результат (первым значением в выводе должно быть среднее для элементов 1:10, во втором значении - среднее для элементов 2:11 и т.д., в последнем - среднее для элементов 135 :144). Все полученные значения средних сохраните в переменную moving_average.

4. Критерии оценивания

Оценка отлично (5 баллов) - выставляется аспиранту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины, проявляющему интерес к данной предметной области, продемонстрировавшему умение уверенно и творчески применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка хорошо (4 балла) - выставляется аспиранту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка удовлетворительно (3 балла) - выставляется аспиранту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он освоил основные разделы учебной программы, необходимые для дальнейшего обучения, и может применять полученные знания по образцу в стандартной ситуации.

Оценка неудовлетворительно (2 балла) - выставляется аспиранту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных принципов и не умеет использовать полученные знания при решении типовых задач.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

При проведении устного зачета аспиранту предоставляется 30 минут на подготовку. Опрос обучающегося по билету на устном зачете не должен превышать одного астрономического часа.