

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы физики
и исследований им. Ландау
А.В. Рогачев**

	Рабочая программа дисциплины (модуля)
по дисциплине:	Методы анализа данных NGS
по направлению:	Прикладные математика и физика
профиль подготовки:	Вычислительная биоинформатика Физтех-школа физики и исследований им. Ландау кафедра биофизики
курс:	1
квалификация:	магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Экзамен

Аудиторных часов: 30 всего, в том числе:

лекции: 30 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 75 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Программу составил: В.В. Чупин, д-р хим. наук, доцент

Программа обсуждена на заседании кафедры биофизики 16.04.2023

Аннотация

Курс сфокусирован на особенностях анализа данных высокопроизводительного секвенирования (NGS) с использованием основных методов биоинформатики и системной биологии. На курсе будут подробно разобраны особенности платформ секвенирования, их сильные стороны и лимитирующие факторы. Слушатели познакомятся с основными данными результатов секвенирования NGS и инструментами для анализа их качества и оценки успешности проведенного эксперимента. Также будут раскрыты основные подходы, используемые для анализа разных типов данных. Курс рассчитан на слушателей, знакомых с базовыми понятиями биоинформатики.

1. Цели и задачи

Цель дисциплины

Дать студентам формирование базовых знаний об особенностях данных и статистического анализа результатов, получаемых с помощью платформ высокопроизводительного секвенирования. Практическое освоение студентами методов для анализа биологических данных и компьютерных методов, разработки методов для анализа данных и приобретение ими практического опыта.

Задачи дисциплины

- Получение основных вычислительных навыков и приобретение ими практического опыта, необходимого для проведения самостоятельных научных исследований в области обработки биологических данных, полученных с помощью технологий высокопроизводительного секвенирования.
- Освоение специфических методов статистической обработки биологических, генетических, медицинских и эпидемиологических данных.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
	УК-1.3 Разрабатывает стратегию достижения поставленной цели как последовательность шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности
УК-2 Способен управлять проектом на всех этапах его реализации	УК-2.1 Формулирует в рамках обозначенной проблемы, цель, задачи, актуальность, значимость (научную, практическую, методическую и иную в зависимости от типа проекта), ожидаемые результаты и возможные сферы их применения
	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
ОПК-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)

ограничения различных методов решения	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, понимает и учитывает на практике границы применимости получаемых решений
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Способен планировать и проводить научные исследования самостоятельно или в составе научного коллектива
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные физические принципы, лежащие в основе технологий высокопроизводительного секвенирования;
- основные алгоритмы и структуры данных, применяемые при сборке de novo геномов и транскриптомов, структурной аннотации геномных последовательностей, картировании чтений;
- статистические методы, применяющиеся при анализе данных, полученных с помощью высокопроизводительного секвенирования;
- вычислительные задачи, возникающие при обработке данных, полученных с использованием высокопроизводительного секвенирования;
- основные методы оценки статистической значимости;
- методы учета множественности сравнений;
- методы мета-анализа;
- статистические характеристики ассоциативных тестов;
- ROC-анализ;
- методы оценки наследуемости и генетических рисков;
- методы сокращения числа переменных при анализе больших массивов данных;
- методы классификации данных;
- основы байесовского анализа данных.

уметь:

- применять основные программные средства, предназначенные для обработки данных, полученных с использованием высокопроизводительного секвенирования;
- применять основные алгоритмические идеи для разработки новых методов и алгоритмов для обработки данных, полученных с использованием высокопроизводительного секвенирования;

владеть:

- навыками освоения и обработки большого объема информации;
- культурой постановки и моделирования вычислительных задач обработки биологических данных, полученных с использованием технологий высокопроизводительного секвенирования и медико-биологических экспериментов.

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	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Структура биологических данных и описательные статистики	3			6
2	Анализ сопряженности признаков	3			6
3	Многомерные методы статистического анализа	2			6
4	Байесовская статистика	2			6
5	Технологии высокопроизводительного секвенирования	3			6
6	Основы работы с командной строкой Linux	2			6
7	Предобработка результатов секвенирования	2			6
8	de novo сборка геномов и транскриптомов	3			6
9	Аннотация геномных последовательностей	2			6
10	Ресеквенирование	2			6
11	RNA-seq	2			5
12	Метагеномика	2			5
13	ChIP-seq	2			5
Итого часов		30			75
Подготовка к экзамену		30 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 2 (Весенний)

1. Структура биологических данных и описательные статистики

Организация файлов и управление данными в программах EXCEL, SPSS и STATISTICA. Описательные статистики. Некоторые приемы быстрых статистических вычислений. Проверка статистических гипотез. Точные и опосредованные критерии. Ошибки I и II рода. Мощность теста. Множественные сравнения. Контроль ошибок I рода. Страты и парадокс Симпсона. Параметрические и непараметрические критерии сравнения. Дисперсионный анализ.

2. Анализ сопряженности признаков

Регрессионный анализ. Анализ остатков. Частные корреляции и конфаундеры. Сопряженность качественных признаков. Отношение шансов и относительный риск. Статистика биомаркеров. Оценки чувствительности и специфичности теста. ROC-анализ.

3. Многомерные методы статистического анализа

Множественный регрессионный анализ. Методы сокращения числа предикторов. Парадокс Фридмана. Оценки наследуемости и генетического риска. Проблема «missing heritability». Факторный анализ. Метод главных компонент. Методы классификации. Кластерный анализ. Дискриминантный анализ.

4. Байесовская статистика

Ограниченность концепции p-value. Анализ воспроизводимости результатов экспериментов. Байесовский фактор. Приоры. Статистика в эпидемиологии. Анализ больших выборок. Байесовские оценки частот редких событий.

5. Технологии высокопроизводительного секвенирования

Физические принципы и технологические решения, используемые в технологиях высокопроизводительного секвенирования. Характеристики основных платформ высокопроизводительного секвенирования.

6. Основы работы с командной строкой Linux

Командная оболочка Bash. Устройство файловой системы в операционных системах семейства Linux. Команды cd, ls, pwd, cp, mv, rm, more, head, tail, grep. Редактор vi.

7. Предобработка результатов секвенирования

Основные типы ошибок, свойственные технологиям высокопроизводительного секвенирования. Основные форматы данных. Оценка качества чтений. Тримминг.

8. de novo сборка геномов и транскриптомов

Алгоритмы de novo сборки, основанные на графа де Брейна и графах перекрытий. Особенности геномных последовательностей, затрудняющих сборку. Оценка качества сборки. Практические аспекты больших геномных проектов. Особенности сборки транскриптомов de novo.

9. Аннотация геномных последовательностей

Основные принципы построения алгоритмов аннотации. Оценка качества аннотации. Практические аспекты применения алгоритмов аннотации для эукариотических геномов.

10. Ресеквенирование

Картирование чтений на референсный геном. Преобразование Барроуза-Уилера для картирования ридов при секвенировании ДНК. Оценка качества картирования. SNP calling. Особенности, возникающие при детекции соматических мутаций.

11. RNA-seq

Особенности картирования чтений, полученных в результате RNA-seq эксперимента на референсный геном. Методы нормализации и анализ экспрессии генов.

12. Метагеномика

Таргетное секвенирование 16S рРНК. Таксономический анализ и анализ биоразнообразия. Полнометагеномное секвенирование. De novo сборка и аннотация генов.

13. ChIP-seq

Взаимодействие ДНК и белка. Методы для изучения ДНК-белкового взаимодействия, применяющиеся до появления высокпроизводительного секвенирования. ChIP – seq протокол. Основные методы анализа ChIP-seq данных.

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

Занятия по учебной дисциплине проводятся с использованием дистанционных образовательных технологий. Каждый обучающийся обеспечен доступом к образовательной платформе bostongene.com

6. Перечень рекомендуемой литературы

Основная литература

1. Biswas, A., Datta, S., Fine, J. P., Segal, M. R. (eds.) Statistical Advances in the Biomedical Sciences Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics, 2008, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany
2. Statistical Human Genetics. Edited by Robert C. Elston. Springer Science+Business Media, LLC 2012
3. Phillip Compeau, Pavel Pevzner, Bioinformatics Algorithms: An Active Learning Approach 2014 Book
4. Xinkun Wang Next-Generation Sequencing Data Analysis 2016 Book
5. Ion Mandoiu, Alexander Zelikovsky. Computational Methods for Next Generation Sequencing Data Analysis 2016 Book

Дополнительная литература

1. Eija Korpelainen, Jarno Tuimala, Panu Somervuo, Mikael Huss, Garry Wong RNA-seq Data Analysis: A Practical Approach. 2014 Book.
2. Topics in Biostatistics. Edited by Walter T. Ambrosius 2007 Humana Press Inc. 999 Riverview Drive, Suite 208, Totowa, New Jersey 07512
3. Agostino Di Ciaccio, Mauro Coli, Jose Miguel Angulo Ibañez. Advanced Statistical Methods for the Analysis of Large Data-Sets. Springer-Verlag Berlin Heidelberg, 2012

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

База данных Национального института стандартизации и технологии США по свойствам соединений: <http://webbook.nist.gov/chemistry/>

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

Аккаунт в Zoom.

Персональные компьютеры с UNIX системой, Python.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, 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 студент должен знать основные определения и понятия, 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 студента.

Самостоятельная работа включает в себя:

- чтение рекомендованной литературы;
- проработку учебного материала, подготовку ответов на вопросы, предназначенных для самостоятельного изучения;
- решение задач, предлагаемых студентам на занятиях;
- 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 к преподавателю.

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

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Прикладные математика и физика
профиль подготовки: Вычислительная биоинформатика
Физтех-школа физики и исследований им. Ландау
кафедра биофизики
курс: 1
квалификация: магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Экзамен

Разработчик: В.В. Чупин, д-р хим. наук, доцент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
	УК-1.3 Разрабатывает стратегию достижения поставленной цели как последовательность шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности
УК-2 Способен управлять проектом на всех этапах его реализации	УК-2.1 Формулирует в рамках обозначенной проблемы, цель, задачи, актуальность, значимость (научную, практическую, методическую и иную в зависимости от типа проекта), ожидаемые результаты и возможные сферы их применения
	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
ОПК-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и ограничения различных методов решения	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, понимает и учитывает на практике границы применимости получаемых решений
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Способен планировать и проводить научные исследования самостоятельно или в составе научного коллектива
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

2. Показатели оценивания компетенций

В результате изучения дисциплины «Методы анализа данных NGS» обучающийся должен:

знать:

- основные физические принципы, лежащие в основе технологий высокопроизводительного секвенирования;
- основные алгоритмы и структуры данных, применяемые при сборке de novo геномов и транскриптомов, структурной аннотации геномных последовательностей, картировании чтений;
- статистические методы, применяющиеся при анализе данных, полученных с помощью высокопроизводительного секвенирования;
- вычислительные задачи, возникающие при обработке данных, полученных с использованием высокопроизводительного секвенирования;
- основные методы оценки статистической значимости;
- методы учета множественности сравнений;
- методы мета-анализа;
- статистические характеристики ассоциативных тестов;
- ROC-анализ;
- методы оценки наследуемости и генетических рисков;
- методы сокращения числа переменных при анализе больших массивов данных;
- методы классификации данных;
- основы байесовского анализа данных.

уметь:

- применять основные программные средства, предназначенные для обработки данных, полученных с использованием высокопроизводительного секвенирования;
- применять основные алгоритмические идеи для разработки новых методов и алгоритмов для обработки данных, полученных с использованием высокопроизводительного секвенирования;

владеть:

- навыками освоения и обработки большого объема информации;
- культурой постановки и моделирования вычислительных задач обработки биологических данных, полученных с использованием технологий высокопроизводительного секвенирования и медико-биологических экспериментов.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале занятия по теме прошлого занятия.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Основные физические принципы, лежащие в основе технологий высокопроизводительного секвенирования
2. Поколения технологий секвенирования. Основные принципиальные отличия технологий секвенирования второго поколения от первого.
3. Основные ошибки в данных, возникающие при использовании различных платформ высокопроизводительного секвенирования
4. Алгоритмы сборки de novo геномных последовательностей.
5. Особенности геномных последовательностей, приводящие к трудностям при сборке de novo.
6. Оценка качества геномныхборок
7. Особенности сборки транскриптомов de novo.
8. Оценка качества транскриптомной сборки.
9. Основные методы, использующиеся при аннотации геномных последовательностей.
10. Оценка качества аннотации.
11. Картирование чтений на референсный геном. Преобразование Барроуза-Уилера.
12. SNP calling.
13. Особенности детекции соматических мутаций на основе данных высокопроизводительного секвенирования.
14. Дизайн RNA-seq эксперимента.
15. Основные способы нормализации экспрессионных данных.

16. Анализ диф. экспрессии.
17. Таргентное секвенирование 16s РНК в метагеномике.
18. Полнометагеномное секвенирование
19. Таксономический анализ и анализ биоразнообразия.
20. De novo сборка и аннотация данных, полученных в результате полнометагеномного секвенирования
21. Дизайн ChIP – seq эксперимента.
22. Основные элементы вычислительного конвейера, используемого для обработки данных, полученных в результате ChIP-seq эксперимента.
23. Отношение шансов и относительный риск
24. Множественная регрессия и парадокс Фридмана
25. Методы оценки публикационного сдвига. Графики-воронки.
26. FDR-метод учета множественности сравнений
27. Байесовские оценки частот редких событий

Пример экзаменационного билета:

Билет 1.

1. Алгоритмы сборки de novo геномных последовательностей.
2. Дизайн ChIP – seq эксперимента.

Билет 2.

1. Картирование чтений на референсный геном. Преобразование Барроуза-Уилера.
2. FDR-метод учета множественности сравнений.

Билет 3.

1. Таксономический анализ и анализ биоразнообразия.
2. Основные физические принципы, лежащие в основе технологий высокопроизводительного секвенирования.

Критерии оценивания

Оценка отлично 10 баллов - выставляется студенту, 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 отлично 9 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 8 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений, с некоторыми недочетами.

Оценка хорошо 7 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка хорошо 6 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности.

Оценка хорошо 5 баллов - выставляется студенту, если он в основном знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач достаточно большое количество неточностей.

Оценка удовлетворительно 4 балла - выставляется студенту, 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 удовлетворительно 3 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, допускающему ошибки в формулировках базовых понятий, нарушения логической последовательности в изложении программного материала, слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и с трудом применяет полученные знания даже в стандартной ситуации.

Оценка неудовлетворительно 2 балла - выставляется студенту, 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.

Оценка неудовлетворительно 1 балл - выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубейшие ошибки в формулировках базовых понятий дисциплины и вообще не имеет навыков решения типовых практических задач.

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

Экзамен проводится в устной форме по билетам. В каждом билете представлено два теоретических вопроса. При проведении экзамена обучающемуся предоставляется 50 минут на подготовку. Опрос обучающегося не должен превышать одного астрономического часа.