

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

УТВЕРЖДЕНО

**Директор физтех-школы
биологической и медицинской
физики**

Д.В. Кузьмин

	Рабочая программа дисциплины (модуля)
по дисциплине:	Анализ NGS данных человека
по направлению:	Прикладные математика и физика
профиль подготовки:	Алгоритмическая биология
	Физтех-школа Биологической и Медицинской Физики
	центр образовательных программ Физтех-школы биологической и медицинской физики
курс:	1
квалификация:	магистр

Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен

Аудиторных часов: 30 всего, в том числе:

лекции: 15 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Подготовка к экзамену: 30 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 1

Программу составил: Ф.Е. Френкель, канд. биол. наук

Программа обсуждена на заседании центра образовательных программ Физтех-школы биологической и медицинской физики 08.06.2023

Аннотация

Дисциплина направлена на освоение базовых технологий высокопроизводительного секвенирования. Студент после освоения курса будет понимать 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, методы мета-анализа, методы сокращения числа переменных при анализе больших массивов данных, основы байесовского анализа данных.

1. Цели и задачи

Цель дисциплины

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

Задачи дисциплины

- формирование базовых знаний в области анализа данных NGS;
- обучение студентов принципам секвенирования, их сильные стороны и лимитирующие факторы, основным данным результатов секвенирования NGS и инструментами для анализа их качества и оценки успешности проведенного эксперимента.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-4 Способен применять современные коммуникативные технологии, в том числе на иностранном(ых) языке(ах), для академического и профессионального взаимодействия	УК-4.1 Способен вести обмен деловой информацией в устной и письменной формах на государственном языке Российской Федерации и не менее чем на одном иностранном языке
	УК-4.2 Владеет навыками, необходимыми для написания, письменного перевода и редактирования различных академических текстов (рефератов, эссе, обзоров, статей и т.д.)
	УК-4.3 Способен представлять результаты академической и профессиональной деятельности на различных научных мероприятиях, включая международные
	УК-4.4 Способен использовать современные средства информационно-коммуникационных технологий для академического и профессионального взаимодействия
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	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	1	1		3
2	Анализ сопряженности признаков	1	2		2
3	Многомерные методы статистического анализа	1	1		2
4	Байесовская статистика	1	1		2
5	Технологии высокопроизводительного секвенирования	1	2		3
6	Основы работы с командной строкой Linux	2	1		2
7	Предобработка результатов секвенирования	1	1		2
8	de novo сборка геномов и транскриптомов	2	1		3
9	Аннотация геномных последовательностей	1	1		3
10	Ресеквенирование	1	1		2
11	RNA-seq	1	1		2

12	Метагеномика	1	1		2
13	ChIP-seq	1	1		2
Итого часов		15	15		30
Подготовка к экзамену		30 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

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

Семестр: 1 (Осенний)

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)

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

6.Перечень рекомендуемой литературы

Основная литература

Предоставляется на кафедре:

Biswas, A., Datta, S., Fine, J. P., Segal, M. R. (eds.) Statistical Advances in the Biomedical Sciences Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics. - Germany: WILEY-VCH Verlag GmbH & Co, 2008

Statistical Human Genetics / edited by Robert C. Elston. - Springer Science+Business Media, LLC, 2012

Phillip Compeau, Pavel Pevzner, Bioinformatics Algorithms: An Active Learning Approach 2014 Book
Xinkun Wang Next-Generation Sequencing Data Analysis 2016 Book

Ion Mandoiu, Alexander Zelikovsky. Computational Methods for Next Generation Sequencing Data Analysis 2016 Book

Дополнительная литература

Предоставляется на кафедре:

Eija Korpelainen, Jarno Tuimala, Panu Somervuo, Mikael Huss, Garry Wong RNA-seq Data Analysis: A Practical Approach. 2014 Book.

Topics in Biostatistics. Edited by Walter T. Ambrosius 2007 Humana Press Inc. 999 Riverview Drive, Suite 208, Totowa, New Jersey 07512

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)

Для части занятий потребуется Zoom. Google Drive для доступа к материалам курса.

Приветствуется наличие во время занятий смартфонов/ноутбуков для участия в интерактивных упражнениях.

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, понятия.

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- чтение и конспектирование рекомендованной литературы,
- проработку учебного материала (учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- подготовку к экзамену.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

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

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Прикладные математика и физика
профиль подготовки:	Алгоритмическая биология Физтех-школа Биологической и Медицинской Физики центр образовательных программ Физтех-школы биологической и медицинской физики
курс:	<u>1</u>
квалификация:	магистр
Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен	
Разработчик:	Ф.Е. Френкель, канд. биол. наук

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-4 Способен применять современные коммуникативные технологии, в том числе на иностранном(ых) языке(ах), для академического и профессионального взаимодействия	УК-4.1 Способен вести обмен деловой информацией в устной и письменной формах на государственном языке Российской Федерации и не менее чем на одном иностранном языке
	УК-4.2 Владеет навыками, необходимыми для написания, письменного перевода и редактирования различных академических текстов (рефератов, эссе, обзоров, статей и т.д.)
	УК-4.3 Способен представлять результаты академической и профессиональной деятельности на различных научных мероприятиях, включая международные
	УК-4.4 Способен использовать современные средства информационно-коммуникационных технологий для академического и профессионального взаимодействия
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

2. Показатели оценивания компетенций

В результате изучения дисциплины «Анализ NGS данных человека» 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

Во время текущего контроля студент должен уметь ответить на следующие вопросы:

1. Дизайн ChIP – seq эксперимента.
2. Основные элементы вычислительного конвейера, используемого для обработки данных, полученных в результате ChIP-seq эксперимента.
3. Отношение шансов и относительный риск
4. Множественная регрессия и парадокс Фридмана
5. Методы оценки публикационного сдвига.
6. Графики-воронки.
7. FDR-метод учета множественности сравнений.
8. Байесовские оценки частот редких событий.
9. Полнометагеномное секвенирование.
10. Таксономический анализ и анализ биоразнообразия.
11. De novo сборка и аннотация данных, полученных в результате полнометагеномного секвенирования.

Во время занятий могут проходить интерактивные обсуждения в чатах курса, что будет являться домашним заданием. Возможно выполнение патентного поиска в качестве самостоятельной задачи. Успешное выполнение всех заданий по курсу и выполнение контрольных срезов знаний дает преимущество на экзамене.

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 геномных сборок.
7. Особенности сборки транскриптомов de novo.
8. Оценка качества транскриптомной сборки.
9. Основные методы, используемые 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.

Пример экзаменационного билета:

Билет №1.

1. Особенности геномных последовательностей, приводящие к трудностям при сборке de novo.
2. Оценка качества геномных сборок.

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

При проведении экзамена обучающемуся предоставляется 40 минут на подготовку.

Во время проведения экзамена обучающиеся могут пользоваться программой дисциплины, конспектами лекций и любой другой литературой.