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 данных высокопроизводительного секвенирования
по направлению: Биотехнология
профиль подготовки: Биомедицинские технологии
Физтех-школа Биологической и Медицинской Физики
кафедра биоинформатики и системной биологии
курс: 1
квалификация: магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:
лекции: 0 час.
семинары: 30 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 15 час.

Всего часов: 45, всего зач. ед.: 1

Программу составил: А.С. Касьянов, канд. физ.-мат. наук

Программа обсуждена на заседании кафедры биоинформатики и системной биологии 04.05.2020

Аннотация

Целью данной дисциплины является знакомство студентов с известными на данный момент способами обработки данных, получаемых в результате 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, вычислительные задачи, возникающие при обработке данных, полученных с использованием высокопроизводительного секвенирования.

1. Цели и задачи

Цель дисциплины

знакомство студентов с известными на данный момент способами обработки данных, получаемых в результате высокопроизводительного секвенирования.

Задачи дисциплины

- формирование базовых знаний об особенностях данных, получаемых с помощью платформ высокопроизводительного секвенирования;
- практическое освоение студентами методов для анализа биологических данных, полученных с помощью высокопроизводительного секвенирования;
- формирование у студентов основных навыков разработки методов для анализа данных и приобретение ими практического опыта, необходимого для проведения самостоятельных научных исследований в области вычислительной обработки биологических данных, полученных с помощью технологий высокопроизводительного секвенирования .

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
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-1.5 Способен создавать программные средства и базы данных, используемые в биоинженерии и биоинформатике
	ПК-1.3 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.2 Способен использовать специализированные знания фундаментальных разделов математики, физики, химии и биологии для постановки и решения научно-исследовательских задач в области биоинженерии и биоинформатики
	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.4 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-3 Способен профессионально работать с	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.5 Способен применять методы биоинженерии и биоинформатики для получения биологических объектов с целенаправленно измененными свойствами
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	ПК-3.4 Способен самостоятельно находить и осваивать новые информационные и программные ресурсы в области биоинженерии и биоинформатики

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.

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 большого объёма информации;
- культурой постановки и моделирования вычислительных задач обработки биологических данных, полученных с использованием технологий высокопроизводительного секвенирования.

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	Технологии высокопроизводительного секвенирования		1		
2	Основы работы с командной строкой Linux		1		1
3	Предобработка результатов секвенирования		4		2
4	de novo сборка геномов и транскриптомов		4		2
5	Аннотация геномных последовательностей		4		2
6	Ресеквенирование		4		2
7	RNA-seq		4		2
8	Метагеномика		4		2
9	ChIP-seq		4		2

Итого часов		30		15
Подготовка к экзамену	0 час.			
Общая трудоёмкость	45 час., 1 зач.ед.			

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

Семестр: 2 (Весенний)

1. Технологии высокопроизводительного секвенирования

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

2. 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.

3. Предобработка результатов секвенирования

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

4. 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.

5. Аннотация геномных последовательностей

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

6. 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.

7. RNA-seq

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

8. Метагеномика

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

9. ChIP-seq

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

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

Оборудование, необходимое для семинаров: аудитория, компьютер и мультимедийное оборудование (проектор, звуковая система),

6.Перечень рекомендуемой литературы

Основная литература

Предоставляется на кафедре:

1. Phillip Compeau, Pavel Pevzner, Bioinformatics Algorithms: An Active Learning Approach 2014 Book
2. Xinkun Wang Next-Generation Sequencing Data Analysis 2016 Book
3. Ion Mandoiu, Alexander Zelikovsky. Computational Methods for Next Generation Sequencing Data Analysis 2016 Book

Дополнительная литература

Предоставляется на кафедре:

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

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

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.

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

Доступ в Интернет, UNIX сервер с отдельным аккаунтом для каждого студента. 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 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 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.

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Биотехнология
профиль подготовки: Биомедицинские технологии
Физтех-школа Биологической и Медицинской Физики
кафедра биоинформатики и системной биологии
курс: 1
квалификация: магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

Разработчик: А.С. Касьянов, канд. физ.-мат. наук

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
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-1.5 Способен создавать программные средства и базы данных, используемые в биоинженерии и биоинформатике
	ПК-1.3 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.2 Способен использовать специализированные знания фундаментальных разделов математики, физики, химии и биологии для постановки и решения научно-исследовательских задач в области биоинженерии и биоинформатики
	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.4 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
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-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов
	ПК-3.5 Способен применять методы биоинженерии и биоинформатики для получения биологических объектов с целенаправленно измененными свойствами
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	ПК-3.4 Способен самостоятельно находить и осваивать новые информационные и программные ресурсы в области биоинженерии и биоинформатики

2. Показатели оценивания компетенций

В результате изучения дисциплины «Анализ данных высокопроизводительного секвенирования» 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.

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 большого объёма информации;
- культурой постановки и моделирования вычислительных задач обработки биологических данных, полученных с использованием технологий высокопроизводительного секвенирования.

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

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

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 выскопроизводительного 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, 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 выскопроизводительного секвенирования.
14. Дизайн RNA-seq эксперимента.
15. Основные способы нормализации экспрессионных данных.
16. Анализ диф. экспрессии.
17. Таргентное секвенирования 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, использующегося для обработки данных, полученных в результате ChIP-seq эксперимента.

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 и дифференцированном зачете.

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 платворм выскопроизводительного 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, 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 выскопроизводительного секвенирования.

14. Дизайн RNA-seq эксперимента.

15. Основные способы нормализации экспрессионных данных.

16. Анализ диф. экспрессии.

17. Таргентное секвенирования 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, использующегося для обработки данных, полученных в результате ChIP-seq эксперимента.

Примеры билетов:

Билет№ 1

Полнометагеномное секвенирование

Билет №2

Оценка качества аннотации.

Билет №3

SNP calling

Билет №4

Дизайн ChIP – seq эксперимента

Билет №5

Оценка качества геномных сборок

Билет №6

Характеристики основных платформ 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

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