Introduction to Bioinformatics and Computational Biology

2021-01-18

Contents

Chapter 1

Course information

This is the course material for STAT115/215 BIO/BST282 at Harvard University.

1.1 Contributors

Xiaole Shirley Liu (lead instructor)
Joshua Starmer
Martin Hemberg
Ting Wang
Feng Yue
Ming Tang
etc.

1.2 Supporting staff

Jack Kang
Scarlett Ge
etc.

Chapter 2

Introduction

2.1 Welcome

2.2 Brief history of bioinformatics

2.2.1 Protein structure wave

2.2.2 Gene expression wave

2.2.3 Genome sequencing wave

2.2.4 High throughput sequencing

2.2.5 Big data challenge from sequencing

2.3 Should I take this course?

2.3.1 Bioinformatics vs computational biology

2.3.2 Levels of bioinformatics

2.3.3 Is this class for me?

2.3.4 All biology is computational biology

2.4 Course information

2.4.1 Course organization and material

2.4.2 Course instructor and TAs

2.4.3 Homework and grading

2.4.4 Lab and Odyssey sign up

Chapter 3

High throughput sequencing

3.1 Three generations of sequencing technologies

First generation sequencing is Sanger sequencing. It is the technology that was used to obtain the first human genome sequence.

Second generation sequencing is also called next generation sequencing (NGS) and is the start of high throughput sequencing. It is what scientists use most often nowadays, and Illumina is the market leader. Most of the rest of this course will cover data analysis using second generation sequencing.

Third generation sequencing is single-molecule sequencing. There are many new technologies still under active development, although none has reached market penetration.

3.2 FASTQ and FASTQC

NGS generates FASTQ files. FASTQC is an computational approach to evaluate the quality of your NGS data.

3.3 Early sequence alignment (1 with 1)

In the early days (1970s), scientists were not worried about having to align too many sequences. They wanted to find the best alignment between two

sequences. Many bioinformatics courses start with learning these, although it is not the main focus of our course. We included two videos in case you are interested.

The Needlemen-Wunsch algorithm is the earliest algorithm to find the alignment between two sequences and score their similarity.

When two sequences are long, and only a portion of them can align well with each other, the Smith-Waterman algorithm can find the best local sequence alignment. It is still considered the best alignment approach, although it is slow.

3.4 Sequence search algorihtms (1 with many)

With more and more sequences available in the public in the 1980s, scientists were interested in finding whether their newly sequenced string has been sequenced before in the public database. Therefore, the fast search algorithm BLAST was developed, using one sequence as the query to find similar sequences from a database.

3.5 Borrow-Wheeler Aligner (many with many)

With NGS, scientists need much faster search (aka mapping) algorithms in order to align the millions of sequences to the reference genome. The current best algorithm is called Borrow-Wheeler Aligner or BWA.

In order to understand BWA, we first need to introduce Borrows-Wheeler transformation and LF mapping

The basic idea of Borrows-Wheeler alignment

3.6 Alignment output

NGS raw data is in FASTQ. Alignment gives you SAM (alignment) or BAM (binary version of SAM) files which contain the sequence information in FASTQ and the mapping locations. BED file is the simpliest, although there is information loss.

Chapter 4

RNA-seq quantification

4.1 RNA-seq & RNA QC (DV200)

4.2 RNA-seq replicates and study design

4.3 STAR and PE BAM file

4.4 RNA-seq QC

4.5 FPKM vs TPM

4.6 RSEM & Salmon

Chapter 5

Differential expression, False discovery rate, Gene ontology

5.1 RNA-seq NB distribution

5.2 DESeq2 and variance stabilization

5.3 Multiple hypotheses testing and FWER

5.4 FDR

5.5 GO

Chapter 6

GSEA, Clustering

6.1 GSEA

6.2 Heatmap and clustering quality

6.3 H-cluster

6.4 K-means

6.5 Pick K and consensus clustering

6.6 Batch effect removal

Chapter 7

Dimension Reduction

7.1 MDS

7.2 LDA

7.3 PCA

Chapter 8

Classification

8.1 Intro to machine learning

8.2 Cross validation

8.3 Regression

8.4 Regularization

8.5 KNN

8.6 Decision trees

8.7 Random forest

8.8 SVM

Chapter 9

Module I Review

9.1 Gene Expression Module Summary

9.2 Gene Expression Analysis Scenarios

Chapter 10

Transcription Factor Motif Finding

Chapter 11

ChIP-seq, Expression Integration

11.1 ChIP-seq

11.2 BWA and MACS

11.3 ChIP-seq QC

11.4 TF interactions (motif)

11.5 TF target genes (expression integration)

Chapter 12

Epigenetics, DNA Methylation

12.1 Epigenetics

12.2 DNA methylation

12.3 Promoter function

12.4 Gene body function

12.5 Enhancer function

12.6 Repetitive region function

12.7 Early cancer detection

Chapter 13

Histone Modifications , Chromatin Accessibility

Chapter 14

Long Range Chromatin Interactions

Chapter 15

Hidden Markov Model

15.1 Intro to HMM

15.2 Pb1: Forward & backward procedure

15.3 Pb2: Viterbi algorithm

15.4 Pb3: Parameter estimation

15.5 HMM application

Chapter 16

Module II Review

16.1 Module II Review

16.2 Practive Questions

Chapter 17

SNP and GWAS

17.1 SNP and LD

17.2 Family-based vs case-control association studies

17.3 GWAS studies and catalog

17.4 GTEx and eQTL

Chapter 18

GWAS and Epigenomics

18.1 Find tissue / cell type

18.2 Identify causal SNPs and genes

18.3 Predict phenotypes

Chapter 19

Single-cell RNA-seq (1)

Chapter 20

Single-cell RNA-seq (2)

20.1 Annotate scRNA-seq clusters

20.2 Differential expression

20.3 Batch effect removal

20.4 Pseudotime

20.5 Overload 10X

20.6 Other applications (CITE-seq, multi-seq, spatial transcriptomics)

Chapter 21

scATAC-seq

Chapter 22

Module III Review

22.1 Module III Review

Chapter 23

Cancer Genome Sequencing , Mutation analyses

Chapter 24

Cancer Subtyping, Survival Analyses

Chapter 25

Targeted Therapy, Drug Resistance, Compound and Genetic Screens

Chapter 26

Cancer Immunotherapy (1)

26.1 Systemic immunotherapy

26.2 Personalized immunotherapy

26.3 HLA and neoantigens

26.4 Tumor immune deconvolution

26.5 T cell signaling (PD1/PDL1, etc)

26.6 Other immune-cells (scRNA-seq)

Chapter 27

Cancer Immunotherapy (2)

Chapter 28

CRISPR Screens

Chapter 29

Module IV Review and Course Review

29.1 Module IV Review

29.2 Course Review