

Introduction to Bioinformatics and Computational Biology

2021-03-07

Contents

1	Course information	11
1.1	Contributors	11
2	Introduction	13
2.1	Brief history of bioinformatics	14
2.2	Should I take this course?	14
2.3	Course information	14
2.4	Lab 1	14
3	High throughput sequencing	15
3.1	Three generations of sequencing technologies	15
3.2	FASTQ and FASTQC	15
3.3	Early sequence alignment (1 with 1)	15
3.4	Sequence search algorithms (1 with many)	16
3.5	Borrow-Wheeler Aligner (many with many)	16
3.6	Alignment output	16
4	RNA-seq Quantification	17
4.1	Introduction to RNA-seq experiment	17
4.2	RNA quality control and experimental design	17
4.3	Alignment	17
4.4	RNA-seq QC	17
4.5	RNA-seq expression index	17
4.6	RSEM and Salmon	17

4.7	RNA-seq read distribution	17
4.8	Lab 2	17
5	Differential expression, FDR, GO, and GSEA	19
5.1	DESeq2 library normalization	20
5.2	DESeq2 variance stabilization	20
5.3	Multiple hypotheses testing and False Discovery Rate	20
5.4	DESeq2 gene filtering	20
5.5	Gene Ontology (GO analysis)	20
5.6	Gene Set Enrichment Analysis (GSEA)	20
5.7	DESeq2 tutorial	20
6	Clustering	21
6.1	Heatmap and clustering quality	21
6.2	Hierarchical cluster	21
6.3	K means cluster	21
6.4	Pick K and consensus clustering	21
6.5	Batch effect removal	21
6.6	Lab3	21
7	Dimension Reduction	23
7.1	Principal Component Analysis: idea behind PCA.	23
7.2	Principal Component Analysis: PCA applications.	23
7.3	Multidimensional Scaling (MDS)	23
7.4	Linear discriminant Analysis (LDA)	24
8	Classification	25
8.1	Introduction	25
8.2	Supervised learning	26
8.3	Cross validation	26
8.4	Regression	26
8.5	Regularization	26
8.6	KNN	26

<i>CONTENTS</i>	5
8.7 Decision trees	26
8.8 Random forest	26
8.9 SVM	26
8.10 Lab 4	26
9 Module I Review	27
9.1 Module I review	27
9.2 Analysis Scenario 1	27
9.3 Analysis Scenario 2	27
10 Transcription Factor Motif Finding	29
10.1 Transcription regulation and motif representation	29
10.2 Motif finding using expectation maximization	29
10.3 Motif finding using Gibbs sampler	29
10.4 Gibbs sampler intuition and transcription factor motif databases	29
10.5 Motif finding general practices	29
11 Transcription Factor Motif Finding	31
11.1 Transcription regulation	31
11.2 Motif representation	31
11.3 EM	31
11.4 Gibbs sampler	31
11.5 Gibbs intuition	31
11.6 Motif finding in eukaryotes	31
11.7 Known motif database	31
12 ChIP-seq, Expression Integration	33
12.1 Motif finding in eukaryotes, and ChIP-seq	33
12.2 MACS and ChIP-seq QC	33
12.3 Identify TF interactions from ChIP-seq motifs	33
12.4 TF target genes and expression integration	33
12.5 Lab 5	33

13 Epigenetics, DNA Methylation	35
13.1 Intro to DNA Methylation	35
13.2 DNA Methylation Pattern and Function	35
13.3 DNA Methylation in Diseases	35
13.4 Techniques to Measure DNA Methylation	35
14 Histone Modifications , Chromatin Accessibility	37
14.1 Nucleosome Positioning	38
14.2 Introduction to Histone Modifications	38
14.3 Infer Transcription Factor Binding from Histone Mark Dynamics	38
14.4 Using Histone Marks to Infer Gene Functions	38
14.5 Introduction to DNase-seq and ATAC-seq	38
14.6 Infer TF from Differential Genes Using LISA	38
14.7 Caution on DNase/ATAC-seq footprint analysis	38
14.8 Summary of Epigenetics and Chromatin	38
14.9 Lab 6	38
15 Long Range Chromatin Interactions	39
15.1 Chromatin interactions	39
15.2 HiC	39
15.3 HiC contact map	39
15.4 HiC normalization	39
15.5 Fractal globule	39
15.6 Loops	39
15.7 Domains	39
15.8 Compartments	39
15.9 Phase separation	39
16 Hidden Markov Model	41
16.1 Intro to HMM	41
16.2 Pb1: Forward & backward procedure	41
16.3 Pb2: Viterbi algorithm	41
16.4 Pb3: Parameter estimation	41
16.5 HMM application	41

<i>CONTENTS</i>	7
17 Module II Review	43
17.1 Module II Review	43
17.2 Practive Questions	43
18 SNP and GWAS	45
18.1 SNP and LD	45
18.2 Family-based vs case-control association studies	45
18.3 GWAS studies and catalog	45
18.4 GTEx and eQTL	45
19 GWAS and Epigenomics	47
19.1 Find tissue / cell type	47
19.2 Identify causal SNPs and genes	47
19.3 Predict phenotypes	47
20 Single-cell RNA-seq (1)	49
20.1 Intro to scRNA-seq	49
20.2 Smart, Droplet, microwell, SCI-based	49
20.3 QC	49
20.4 Normalization	49
20.5 Imputation	49
20.6 Dimension reduction	49
20.7 Clustering	49
20.8 t-SNE and UMAP	49
21 Single-cell RNA-seq (2)	51
21.1 Annotate scRNA-seq clusters	51
21.2 Differential expression	51
21.3 Batch effect removal	51
21.4 Pseudotime	51
21.5 Overload 10X	51
21.6 Other applications (CITE-seq, multi-seq, spatial transcriptomics)	51

22 scATAC-seq	53
22.1 Intro to scATAC-seq	53
22.2 Sample and cell QC	53
22.3 Dimension reduction, clustering & visualization	53
22.4 Differential peaks and annotations	53
22.5 Integration with scRNA-seq	53
23 Module III Review	55
23.1 Module III Review	55
24 Cancer Genome Sequencing , Mutation analyses	57
24.1 Intro to TCGA	57
24.2 Cancer mutation characterization	57
24.3 Cancer mutation patterns	57
24.4 Tumor purity and clonality	57
24.5 Interpret tumor mutations	57
24.6 Find cancer genes	57
24.7 Summary and future	57
25 Cancer Subtyping, Survival Analyses	59
25.1 TCGA expression	59
25.2 Tumor subtypes	59
25.3 Survival analysis	59
25.4 GoF Oncogenes and LoF TS	59
25.5 Chromatin regulator mutations in cancer	59
25.6 DNA methylation and CIMP	59
26 Targeted Therapy, Drug Resistance, Compound and Genetic Screens	61
26.1 Hallmarks of cancer	62
26.2 Chemo vs targeted therapy	62
26.3 Drug resistance	62
26.4 Synthetic lethality	62

<i>CONTENTS</i>	9
26.5 Precision medicine	62
26.6 Tumor (bulk vs scRNA-seq), mice, cell lines	62
26.7 Compound screens	62
26.8 Genetic screens	62
26.9 Tumor heterogeneity	62
27 Cancer Immunotherapy (1)	63
27.1 Systemic immunotherapy	63
27.2 Personalized immunotherapy	63
27.3 HLA and neoantigens	63
27.4 Tumor immune deconvolution	63
27.5 T cell signaling (PD1/PDL1, etc)	63
27.6 Other immune-cells (scRNA-seq)	63
28 Cancer Immunotherapy (2)	65
28.1 TCR analysis	65
28.2 BCR analysis	65
28.3 Microbiome	65
28.4 Immunotherapy response biomarkers	65
28.5 Targeted therapy as immune-modulators	65
28.6 Epigenetic therapy as immune-modulators	65
29 CRISPR Screens	67
29.1 CRISPR and KO	67
29.2 CRISPRa and CRISPRi	67
29.3 CRISPR design and outcome	67
29.4 CRISPR screens & DepMap	67
29.5 CRISPR screen analysis	67
29.6 CRISPR screens in drug response	67
29.7 CRISPR screens in immunology	67
29.8 Enhancer CRISPR screen	67
29.9 CRISPR screens + scRNA-seq	67

30 Module IV Review and Course Review	69
30.1 Module IV Review	69
30.2 Course Review	69

Chapter 1

Course information

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

All the YouTube videos in this course are organized under the 2021 STAT115 playlist.

1.1 Contributors

Xiaole Shirley Liu Harvard University and Dana-Farber Cancer Institute

Joshua Starmer StatQuest

Martin Hemberg Wellcome Sanger Institute

Ting Wang Washington University

Feng Yue Northwestern University

Gad Getz Harvard University and Broad Institute

Ming Tang

Yang Liu

Bo Yuan

Jack Kang

Scarlett Qian

Jiazhen Rong

Phillip Nicol

Maartin De Vries

We thank many colleagues in the community, who helped Dr. Liu in prepare the STAT115/215 BIO/BST282 course over the years. Some of the lecture slides acknowledged their contributions, but these contributors are not individually acknowledged here.

Chapter 2

Introduction

2.1 Brief history of bioinformatics

2.1.1 Protein structure wave

2.1.2 Gene expression wave

2.1.3 Genome sequencing wave

2.1.4 Big data challenge from sequencing

2.2 Should I take this course?

2.2.1 Bioinformatics vs computational biology

2.2.2 Is this class for me?

2.3 Course information

2.3.1 Logistics

2.3.2 X Shirley Liu lab introduction

2.4 Lab 1

2.4.1 Introduction

2.4.2 Introduction to R

2.4.3 Introduction to Bash

2.4.4 Getting started with Cannon

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 Needleman-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 algorithms (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 Borrow-Wheeler transformation and LF mapping

The basic idea of Borrow-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 simplest, although there is information loss.

Chapter 4

RNA-seq Quantification

- 4.1 Introduction to RNA-seq experiment
- 4.2 RNA quality control and experimental design
- 4.3 Alignment
- 4.4 RNA-seq QC
- 4.5 RNA-seq expression index
- 4.6 RSEM and Salmon
- 4.7 RNA-seq read distribution
- 4.8 Lab 2
 - 4.8.1 STAR tutorial
 - 4.8.2 RSeQC tutorial
 - 4.8.3 RSEM/Salmon Tutorial

Chapter 5

Differential expression, FDR, GO, and GSEA

DESeq2 is a popular and accurate computational algorithm to detect differential gene expression from RNA-seq data. It includes many elegant quantitative considerations, such as:

- Normalize the gene read counts by library size and composition
- Model gene read counts with negative binomial distribution
- Use hierarchical modeling to stabilize the gene variance
- Use Benjamini-Hochberg to calculate control for false discovery rate of calling differentially expressed genes
- Filter lowly expressed genes to reduce the number of hypotheses to be tested

5.1 DESeq2 library normalization

5.2 DESeq2 variance stabilization

5.3 Multiple hypotheses testing and False Discovery Rate

5.4 DESeq2 gene filtering

5.5 Gene Ontology (GO analysis)

5.6 Gene Set Enrichment Analysis (GSEA)

5.7 DESeq2 tutorial

Chapter 6

Clustering

6.1 Heatmap and clustering quality

6.2 Hierarchical cluster

6.3 K means cluster

6.4 Pick K and consensus clustering

6.5 Batch effect removal

6.6 Lab3

6.6.1 PCA tutorial

6.6.2 Clustering tutorial

6.6.3 Combat tutorial

6.6.4 DESeq2 Tutorial

6.6.5 DAVID/GSEA Tutorial

Chapter 7

Dimension Reduction

RNA-seq samples have tens of thousands of genes, although many genes might not vary much between samples and many others have correlated gene expression. Dimension reduction techniques aim to reduce the dimension of representing each sample with tens of thousands of genes to much fewer dimensions, e.g. 2 to 100.

7.1 Principal Component Analysis: idea behind PCA.

PCA / SVD automatically outputs PC1, PC2, PC3, etc, with earlier PCs capturing the highest level of variability in the original data. Each PC is a linear combination of raw gene expression, and is orthogonal to all other PCs.

7.2 Principal Component Analysis: PCA applications.

PCA is a widely used method to project samples with high dimensions (e.g. with gene expression data) onto two dimensions for better visualization. It is an intuitive way to identify sample clusters, and identify batch effect.

7.3 Multidimensional Scaling (MDS)

MDS can use differential ways to calculate pair-wise distance, then use lower dimensions to satisfy the pair-wise distance. PCA is a special case of MDS.

7.4 Linear discriminant Analysis (LDA)

LDA is not only a dimension reduction method, but also a supervised machine learning method.

Chapter 8

Classification

8.1 Introduction

Imagine you have RNA-seq of a collection of labeled normal lung and lung cancer tissues. Given a new sample of RNA-seq from the lung with unknown diagnosis, will you be able to predict based on the existing labeled samples and the expression data whether the new sample is normal or tumor? This is a sample classification problem, and it could be solved using **unsupervised** and **supervised** learning approaches.

Unsupervised learning is basically clustering or dimension reduction. You can use hierarchical clustering, MDS, or PCA. After clustering and projection the data to lower dimensions, you examine the labels of the known samples (hopefully they cluster into separate groups by the label). Then you can assign label to the unknown sample based on its distance to the known samples.

Supervised learning considers the labels with known samples and tries to identify features that can separate the samples by the label. Cross validation is conducted to evaluate the performance of different approaches and avoid over fitting.

StatQuest has done an amazing job with machine learning with a full playlist of well organized videos. While the full playlist is worth a full course, for the purpose of the course, we will just highlight a number of widely used approaches. They include logistic regression (this is considered statistical machine learning), K nearest neighbors, random forest, and support vector machine (these are considered computer science machine learning).

8.2 Supervised learning

8.3 Cross validation

8.4 Regression

8.5 Regularization

8.5.1 Ridge regression

8.5.2 LASSO regression

8.5.2.1 LASSO tutorial in R

8.6 KNN

8.7 Decision trees

8.8 Random forest

8.9 SVM

8.10 Lab 4

8.10.1 K-Nearest Neighbors tutorial

8.10.2 Regression/Ridge/LASSO Tutorial

8.10.3 Logistic Regression Tutorial

8.10.4 Support Vector Machine Tutorial

8.10.5 Random Forest Tutorial

Chapter 9

Module I Review

9.1 Module I review

9.2 Analysis Scenario 1

9.3 Analysis Scenario 2

Chapter 10

Transcription Factor Motif Finding

- 10.1 Transcription regulation and motif representation
- 10.2 Motif finding using expectation maximization
- 10.3 Motif finding using Gibbs sampler
- 10.4 Gibbs sampler intuition and transcription factor motif databases
- 10.5 Motif finding general practices

Chapter 11

Transcription Factor Motif Finding

11.1 Transcription regulation

11.2 Motif representation

11.3 EM

11.4 Gibbs sampler

11.5 Gibbs intuition

11.6 Motif finding in eukaryotes

11.7 Known motif database

Chapter 12

ChIP-seq, Expression Integration

12.1 Motif finding in eukaryotes, and ChIP-seq

12.2 MACS and ChIP-seq QC

12.3 Identify TF interactions from ChIP-seq motifs

12.4 TF target genes and expression integration

12.5 Lab 5

12.5.1 MACS Tutorial

12.5.2 ChIP-seq QC Tutorial

12.5.3 TF Motif Finding Tutorial

12.5.4 TF Collaborator Tutorial

Chapter 13

Epigenetics, DNA Methylation

13.1 Intro to DNA Methylation

13.2 DNA Methylation Pattern and Function

13.3 DNA Methylation in Diseases

13.4 Techniques to Measure DNA Methylation

Chapter 14

Histone Modifications , Chromatin Accessibility

14.1 Nucleosome Positioning

14.2 Introduction to Histone Modifications

14.3 Infer Transcription Factor Binding from
Histone Mark Dynamics

14.4 Using Histone Marks to Infer Gene Func-
tions

14.5 Introduction to DNase-seq and ATAC-seq

14.6 Infer TF from Differential Genes Using
LISA

14.7 Caution on DNase/ATAC-seq footprint
analysis

14.8 Summary of Epigenetics and Chromatin

14.9 Lab 6

14.9.1 ChIP-seq Expression Integration

14.9.2 Cistrome-GO Tutorial

14.9.3 ATAC-seq Analysis and LISA Tutorial

Chapter 15

Long Range Chromatin Interactions

15.1 Chromatin interactions

15.2 HiC

15.3 HiC contact map

15.4 HiC normalization

15.5 Fractal globule

15.6 Loops

15.7 Domains

15.8 Compartments

15.9 Phase separation

Chapter 16

Hidden Markov Model

16.1 Intro to HMM

16.2 Pb1: Forward & backward procedure

16.3 Pb2: Viterbi algorithm

16.4 Pb3: Parameter estimation

16.5 HMM application

Chapter 17

Module II Review

17.1 Module II Review

17.2 Practive Questions

Chapter 18

SNP and GWAS

18.1 SNP and LD

18.2 Family-based vs case-control association studies

18.3 GWAS studies and catalog

18.4 GTEx and eQTL

Chapter 19

GWAS and Epigenomics

19.1 Find tissue / cell type

19.2 Identify causal SNPs and genes

19.3 Predict phenotypes

Chapter 20

Single-cell RNA-seq (1)

20.1 Intro to scRNA-seq

20.2 Smart, Droplet, microwell, SCI-based

20.3 QC

20.4 Normalization

20.5 Imputation

20.6 Dimension reduction

20.7 Clustering

20.8 t-SNE and UMAP

Chapter 21

Single-cell RNA-seq (2)

21.1 Annotate scRNA-seq clusters

21.2 Differential expression

21.3 Batch effect removal

21.4 Pseudotime

21.5 Overload 10X

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

Chapter 22

scATAC-seq

22.1 Intro to scATAC-seq

22.2 Sample and cell QC

22.3 Dimension reduction, clustering & visualization

22.4 Differential peaks and annotations

22.5 Integration with scRNA-seq

Chapter 23

Module III Review

23.1 Module III Review

Chapter 24

Cancer Genome Sequencing , Mutation analyses

24.1 Intro to TCGA

24.2 Cancer mutation characterization

24.3 Cancer mutation patterns

24.4 Tumor purity and clonality

24.5 Interpret tumor mutations

24.6 Find cancer genes

24.7 Summary and future

Chapter 25

Cancer Subtyping, Survival Analyses

25.1 TCGA expression

25.2 Tumor subtypes

25.3 Survival analysis

25.4 GoF Oncogenes and LoF TS

25.5 Chromatin regulator mutations in cancer

25.6 DNA methylation and CIMP

Chapter 26

Targeted Therapy, Drug Resistance, Compound and Genetic Screens

26.1 Hallmarks of cancer

26.2 Chemo vs targeted therapy

26.3 Drug resistance

26.4 Synthetic lethality

26.5 Precision medicine

26.6 Tumor (bulk vs scRNA-seq), mice, cell lines

26.7 Compound screens

26.8 Genetic screens

26.9 Tumor heterogeneity

Chapter 27

Cancer Immunotherapy (1)

27.1 Systemic immunotherapy

27.2 Personalized immunotherapy

27.3 HLA and neoantigens

27.4 Tumor immune deconvolution

27.5 T cell signaling (PD1/PDL1, etc)

27.6 Other immune-cells (scRNA-seq)

Chapter 28

Cancer Immunotherapy (2)

28.1 TCR analysis

28.2 BCR analysis

28.3 Microbiome

28.4 Immunotherapy response biomarkers

28.5 Targeted therapy as immune-modulators

28.6 Epigenetic therapy as immune-modulators

Chapter 29

CRISPR Screens

29.1 CRISPR and KO

29.2 CRISPRa and CRISPRi

29.3 CRISPR design and outcome

29.4 CRISPR screens & DepMap

29.5 CRISPR screen analysis

29.6 CRISPR screens in drug response

29.7 CRISPR screens in immunology

29.8 Enhancer CRISPR screen

29.9 CRISPR screens + scRNA-seq

Chapter 30

Module IV Review and Course Review

30.1 Module IV Review

30.2 Course Review