课程大纲
COURSE SYLLABUS

1.	课程代码/名称 **Course Code/Title**	基因组和组学数据分析/Genomics Data Analysis
2.	课程性质 **Compulsory/Elective**	专业−选修/Elective
3.	课程学分/学时 **Course Credit/Hours**	3
4.	授课语言 **Teaching Language**	英文为主，必要时辅以少量中文解释；教材、课件、考试为英文 English, with a few Chinese. Textbooks, ppts and examinations are in English
5.	授课教师 **Instructor(s)**	靳文菲,生物系 Dr. Wenei JIN, Department of Biology, SUSTech jinwf@sustech.eud.cn
6.	先修要求 **Pre-requisites**	Prerequisites include a college level mathematics, statistics and molecular biology

7. 教学目标
Course Objectives

Course Objectives

Genomics is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes. This subject will help students to understand life in a whole picture --- functional genomics, comparative genomics, evolutionary genomics, transcriptomics, 3D genomics, their interrelations and influence on the organism. Furthermore, this course emphasize on computational analyses of the genomics. Various existing methods will be critically described and the strengths and limitations of each will be discussed, with practical assignments utilizing the tools. It is to train students’ vigorous Scientific Spirit and inspire their scientific curiosity.

Learning Outcomes

With the completion of this course, The student could

1) Be familiar with the major genomic database and database searching

2) Be familiar with Linux and master at least one programming language

3) Conduct various genomic analysis

4) Analyze next generation sequencing data including DNA-seq, RNA-seq, ChIP-seq, single cell sequencing data.

8. 教学方法
Teaching Methods

PPT presentation, class discussion, written assignments, computational practice and quizzes

9. 教学内容
Course Contents

Section 1	I Introduction of Genomics and Basic computational skills (Linux/shell+ python/R)

	Hours: 10 1. Past, Present and Future of Genomics and Course Introduction 1.1 What is genomics? 1.2 The origin and development of genomics 1.3 Present Genomics 1.4 Challenges and future of Genomics 1.5 Course Introduction: Goals, outline, evaluation/examination and learning guidelines 2. Linux and Linux commands 2.1 Server and operating systems 2.2 Linux operating system and Open Source Software 2.3 Terminal and basic Linux commands 2.4 File system and server management 2.5 Personal setting 3. Programing language and shell 3.1 Principles of programming languages 3.2 Script languages and bash shell 3.3 Basic shell functions 3.4 I/O Redirection and file descriptors 3.5 Pattern matching in shell 3.6 Biological data analysis: Modularization and pipeline 4. Programming language Python 4.1 The features of Python 4.2 Data types and variable 4.3 Control structures 4.4 Functions and procedures 4.5 Classes & instances 4.6 Modules & packages 5. R Language Statistics and Drawing 5.1 Quick start R 5.2 Basic principles and concepts 5.3 Data operation in R (Vectors, matrices, arrays, data frames) 5.4 Plot figures 5.5 Statistical Analysis of R 5.6 Function definition and programing 5.7 packages
Section 2	II Basic sequence analysis Hours: 6 6 Pairwise sequence alignments 6.1 Sequence change over time 6.2 Pairwise sequence comparisons 6.3 Dynamic programming alignment 6.3.1 Global alignment (Needleman-Wunsch) 6.3.2 Local alignment (Smith-Waterman) 6.4 Sequence Similarity Searching

		6.4.1 FASTA Algorithm 6.4.2 BLAST Algorithm 7. Multiple Sequence Alignment and Phylogenetics 7.1 Significance of multiple sequence alignment 7.2 Progressive Alignment (ClustalW) 7.3 Basics of phylogeny: Characters, traits, nodes, branches, lineages 7.4 Molecular clock and modeling sequence evolution 7.5 Distances and clustering algorithm: UPGMA and Neighbor Joining (NJ) 7.6 From sequence alignments to trees: Parsimony methods 7.7 Probability based approach: Maximum likelihood methods
	Section 3	III Next Generation Sequencing (NGS) and cancer genomics Hours: 8 1. NGS and Short reads mapping Introduction to Genomic Technologies From Sanger sequencing to NGS Principles of NGS: Massive parallel sequencing Features of NGS data: Short reads Uses Trie structure (Trie and Suffix Array) to search a reference genome Burrows－Wheeler transform（BWT) 2. Variant calling and output Genetic variants: structure variants, SNV, CNV SAM format for mapped reads Approaches for variants calling VCF format for saving called variants 3. Cancer genomics and single cell cancer genomics Calling variants in cancer genomics Single cell cancer genomes Tumor microevolution
	Section 4	IV Transcriptomic and epigenomic analysis Hours: 10 1.Gene expression profiling and RNA-seq What’ s the advantage of RNA-seq compared with microarray? What factors should we consider for RNA-seq data normalization? What’ s the advantage of single cell sequencing over bulk cells? 2. Single cell RNA-seq Cellular heterogeneity Single cell RNA-seq technologies Distinct cell populations Pseudo-time inference 3. Epigenome and data anlysis Definition of epigenetics? How to detect genome-wide DNA methylation? How to detect genome-wide nucleosome positioning and chromatin accessibility? How to identify genome-wide TF binging sites? How to do the peak calling?

		What is Hi-C? How to identify the significant interaction？ 4. Single cell epigenomics challenges scDNAse-seq scMNase-seq scATAC-seq multipe-omics 5. Gene Ontology and enrichment analysis Gene ontology (GO) program Structure of GO Gene annotation in GO GO/pathway enrichment analysis Gene set enrichment analysis (GSEA)
	Section 5	V Population Genomics and association study Hours: 12 1. Haplotype and linkage disequilibrium What is Haplotype? What is linkage disequilibrium? Calculation of linkage disequilibrium Complete LD and perfect LD Recombination rate and LD block 2. Population genomics Effective population size (Ne) The major forces shaping population Population substructure Measure population structure (F-statistics) Approaches for analysis of population structure Analysis of molecular variance (AMOVA) Dimensionality reduction Model based approaches 3. Approaches for natural selection detection Divergence rate and phylogenetic shadowing Changed function-altering mutation, e.g., dN/dS or KN/KS Polymorphism deviating from interspecies divergence e.g. Hudson-Kreitman-Aguade (HKA) test and McDonald-Kreitman (MK) test Changed allele frequency spectrum e.g., Tajima’s D Increased derived allele frequencies Extended haplotype homozygosity (EHH), e.g., iHS Locus-specific population differentiation, e.g., FST Biased ancestry contribution in admixed population. Composite strategies. e.g. combine multiple factors and Likelihood-ratio test 4. Genomics and evolution theory Evolution is a unifying theme in biology History of “evolutionary thought” Darwin’s Four Postulates 1) Individuals within species vary.

		2) Some variations are heritable. 3) More offspring are produced than can survive 4) Survival and reproduction are nonrandom Modern evolutionary theory 5. Genomics and human evolution Classic approaches for study human evolution Human origin models Mitochondrial and Y-chromosome detailed “Out of African” theory Genomic approach revolutionized our understanding of human evolution Human origin model based on genomic data Human migration and natural selection 6.Gene mapping for identifying disease associated variants Linkage analysis for rare disease/Mendelian diseases Genetic model for complex disease: 1) Common disease common variant (CDCV), 2) Common disease rare variant (CDRV) Methods for identifying disease associated variants 1) Family based association study 2) Case control based association study 3) Association study based on next generation sequencing (NGS) Challenge in identifying disease associated variants
	Section 6	VI Genomics and Big data Hours: 2 Accumulation of genomic data Clustering algorithms 1) Hierarchical agglomerative clustering 2) Partitioning methods Two approaches for dimensionality reduction (Feature Selection and feature extraction) Linear reduction 1) Principal component analysis (PCA) 2) Singular Value Decomposition (SVD) 3) Multi-Dimensional Scaling (MDS) Non-linear reduction 1) t-distributed stochastic neighbor embedding （t-SNE） 2) Uniform Manifold Approximation and Projection (UMAP)
	Section 7	
	Section 8	
	Section 9	
	Section 10	
	…………	
10.	课程考核 **Course Assessment**	

	请再此注明：①考查/考试；②分数构成。 Total score 100 Attendance 10 Class Performance 20 Assignments 20 Mid-term Exam 20 Final Presentation/Exam 30 I encourage you to ask questions during the class, and you will get credit of Class Performance.
11.	**教材及其它参考资料** **Textbook and Supplementary Readings**
	Reference books: Introduction to Genomics. Arthur M. Lesk. Oxford University Press; 3 edition. ISBN-10: 0198754833. Bioinformatics and Functional Genomics. Jonathan Pevsner. Wiley-Blackwell; 3 edition. ISBN-10: 1118581784 Medical Genetics And Genomics, Csaba Szalai, ISBN: 9789632791876 Will have some docs at the beginning of the course