



UNIVERSITÀ DEGLI STUDI DI PALERMO

DIPARTIMENTO	Scienze Economiche, Aziendali e Statistiche
ANNO ACCADEMICO OFFERTA	2024/2025
ANNO ACCADEMICO EROGAZIONE	2024/2025
CORSO DILAUREA MAGISTRALE	STATISTICA E DATA SCIENCE
INSEGNAMENTO	STATISTICA GENETICA E APPLICAZIONI BIOINFORMATICHE
TIPO DI ATTIVITA'	C
AMBITO	84545-Attività formative affini o integrative
CODICE INSEGNAMENTO	22375
SETTORI SCIENTIFICO-DISCIPLINARI	SECS-S/01
DOCENTE RESPONSABILE	CHIODI MARCELLO Professore Ordinario Univ. di PALERMO
ALTRI DOCENTI	
CFU	6
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	108
NUMERO DI ORE RISERVATE ALLA DIDATTICA ASSISTITA	42
PROPEDEUTICITA'	
MUTUAZIONI	
ANNO DI CORSO	1
PERIODO DELLE LEZIONI	2° semestre
MODALITA' DI FREQUENZA	Facoltativa
TIPO DI VALUTAZIONE	Voto in trentesimi
ORARIO DI RICEVIMENTO DEGLI STUDENTI	CHIODI MARCELLO Martedì 15:00 17:00 stanza del docente (edificio 13); eccezionalmente su teams Venerdì 12:00 13:00 stanza del docente (edificio 13); eccezionalmente su teams

DOCENTE: Prof. MARCELLO CHIODI

PREREQUISITI	Buona dimestichezza con la programmazione informatica.
RISULTATI DI APPRENDIMENTO ATTESI	Il corso si propone di introdurre lo studente ai principali metodi statistici per l'analisi di espressività genica e di trasmettere le competenze informatiche di base per l'analisi e la gestione di grosse moli di dati. La complessità della materia non permetterà allo studente di essere completamente autonomo nello svolgimento di analisi su dati genetici. Ciononostante si ritiene che lo studente acquisirà contezza delle principali metodologie e avrà una buona comprensione delle applicazioni. Inoltre lo studente sarà in grado di risolvere i principali problemi tecnici legati alla gestione e all'analisi di grosse moli di dati.
VALUTAZIONE DELL'APPRENDIMENTO	L'esame consiste in una prova orale nella quale lo studente dovrà mostrare la conoscenza degli argomenti trattati durante il corso. Ciò avverrà grazie ad alcune domande sugli argomenti del corso e alla presentazione da parte dello studente di un programma R da lui creato. Prima dell'esame lo studente riceverà le indicazioni per la creazione di un programma R parallelizzato ed eseguibile dal terminale. Il suddetto programma dovrà rispondere e specifiche esigenze di analisi concordate con il docente. Durante la prova orale lo studente dovrà mostrare l'utilizzo del programma da lui creato evidenziandone l'utilità. La valutazione viene espressa in trentesimi. In particolare, l'esito della valutazione può essere descritto come di seguito. 30 e lode: ottima conoscenza degli argomenti, ottima proprietà di linguaggio, ottima capacità di applicare autonomamente le conoscenze acquisite 30: ottima conoscenza degli argomenti, ottima proprietà di linguaggio, buona capacità di applicare autonomamente le conoscenze acquisite 26-29: Buona padronanza degli argomenti, buona proprietà di linguaggio, discreta capacità di applicare autonomamente le conoscenze acquisite. 24-25: Discreta conoscenza dei principali argomenti, discreta proprietà di linguaggio, limitata capacità di applicare autonomamente le conoscenze acquisite. 21-23: conoscenza dei principali argomenti dell'insegnamento, ma senza una approfondita comprensione. Limitata proprietà di linguaggio, scarsa capacità di applicare autonomamente le conoscenze acquisite. 18-20: minima conoscenza degli argomenti principali dell'insegnamento e del linguaggio tecnico, scarsissima capacità di applicare autonomamente le conoscenze acquisite. insufficiente: lo studente non possiede una conoscenza accettabile degli argomenti trattati durante il corso e non ha alcuna capacità applicativa.
OBIETTIVI FORMATIVI	Il corso si propone di introdurre lo studente ai principali metodi statistici per l'analisi di espressività genica e di trasmettere le competenze informatiche di base per l'analisi e la gestione di grosse moli di dati.
ORGANIZZAZIONE DELLA DIDATTICA	La prima parte del corso verte sull'analisi statistica dell'espressività genica. La trattazione della metodologia statistica verrà contestualizzata nell'ambito della biologia molecolare e della genetica, sebbene tali metodi siano utilizzati anche in altri contesti applicativi. Lo studente sarà quindi coinvolto nell'analisi di dati di espressività genica tramite l'utilizzo del software R. La seconda parte del corso si concentra su aspetti tecnici legati alla programmazione in parallelo e all'uso del terminale. Tali aspetti sono di fondamentale importanza per la gestione e l'analisi di grosse moli di dati (indipendentemente dalla natura del dato in esame). Questi argomenti verranno accompagnati da applicazioni bioinformatiche come l'allineamento dei genomi, la predizione di legami molecolari e la stima di frazioni cellulari su tessuti. Lo studente avrà contezza delle principali metodologie statistiche per l'analisi di espressività genica. Inoltre, sarà in grado di connettersi autonomamente a un server e di analizzare grosse moli di dati tramite la programmazione da remoto e il calcolo in parallelo. Lo studente sarà in grado di creare un programma R eseguibile dal terminale.
TESTI CONSIGLIATI	materiale fornito in aula

PROGRAMMA

ORE	Lezioni
2	Introduzione alla biologia computazionale, alla statistica genetica e alla trascrittomica. Cenni di biologia molecolare; dogma centrale della biologia e sintesi proteica
5	Inferenza di larga scala e cenni sull'approccio bayesiano. Introduzione ai dati di espressività genica e tecniche di normalizzazione
6	Differential expression analysis: Significant analysis of microarray, moderated t-test, Mann-Whitney test, Kruskal-Wallis one-way ANOVA. Identificazione di firme geniche. Gene set enrichment analysis. Introduzione alla Gene Ontology, Reactome pathway analysis, utilizzo del software GSEA.
6	Armonizzazione di dati biologici. Analisi del batch effect tramite le proiezioni PCA, t-SNE e UMAP e rimozione di tale effetto tramite pacchetti R specifici. Analisi di clustering su dati di espressività genica.
5	Analisi di dati di sequenziamento su singole cellule. Utilizzo del software Seurat. Approccio bootstrap per la verifica di ipotesi. Applicazione del bootstrap t-test all'analisi di espressività genica.

ORE	Laboratori
8	Strumenti informatici avanzati per l'analisi di vaste moli di dati. Introduzione al terminale. Introduzione alla programmazione da remoto. Esecuzione di software tramite il terminale. Creazione di programmi R eseguibili dal terminale.
6	Calcolo in parallelo in R. Monitoraggio dei processi e della RAM tramite il terminale. Introduzione alle applicazioni bioinformatiche. Introduzione ai file FASTA e FASTQ. Allineamento dei genomi tramite STAR.
4	Introduzione alla metagenomica Introduzione ai meccanismi di regolazione genica. Predizione di legami tra molecole di RNA messaggero e microRNA. Stima in silico delle frazioni di cellule del sistema immunitario in tessuti tumorali. Utilizzo di CIBERSORTX