UNIVERSITÁ DEGLI STUDI DI SALERNO

DOTTORATO IN MANAGEMENT & INFORMATION TECHNOLOGY

CURRICULUM: INFORMATION SECURITY & INNOVATION SYSTEMS

COORDINATORE: Ch.mo. Prof. Antonelli Valerio Ciclo XVII N.S.

Novel tools for reproducible Next Generation Sequencing data analysis and integration

Relatori

Ch.mo. Prof. Tagliaferri Roberto

Dott. Angelini Claudia

ANNO ACCADEMICO 2017/2018

Candidato

Righelli Dario Matr. 8800800010

Abstract italiano:

Le massicce tecnologie di sequenziamento parallelo producono una grande quantità di dati sull'intero genoma di cellule, tessuti e organismi modello, utili per comprendere molti meccanismi biologici, come le interazioni proteina-cromatina (ad esempio ChIP-Seq), metilazione del DNA (Methyl-Seq o BS -Seq), accessibilità alla cromatina (ad es. Atac-Seq), attività globali di trascrizione e traslazione (es. RNA Sequencing (RNA-seq)) e organizzazione 3-D della cromatina (es. Hi-C), dando la possibilità di studiare lo stesso individuo o condizione sperimentale da molti punti di vista diversi (trascrittomica, epigenomica, ecc.) ad altissima risoluzione. Ogni tipo di questi dati omics spiega un diverso aspetto del comportamento cellulare. Al fine di estrapolare informazioni rilevanti da ogni omica, è necessario sviluppare metodologie statistiche specifiche per l'analisi di singoli dati e, al tempo stesso, metodologie computazionali per la gestione di enormi quantità di dati. Tuttavia, per dare una visione completa dei meccanismi di regolazione delle cellule, è necessario non solo analizzare singoli omics ma anche sviluppare nuovi modelli statistici e computazionali per l'integrazione di diversi tipi di omics all'interno di uno studio unificato.
Questa tesi è incentrata sullo sviluppo di tre strumenti computazionali principali (ticorser, DEScan2 e IntegrHO), che consentono l'analisi dei dati e l'integrazione di più esperimenti di sequenziamento di prossima generazione. Inoltre, contiene anche un quarto strumento (easyReporting) per una ricerca computazionale riproducibile. Ticorser (analizzatore di dati RNA-seq a tempo pieno) è un nuovo pacchetto R che mira ad analizzare i dati RNA-seq nel corso del tempo. Offre molteplici metodi per l'analisi dei dati di espressione differenziale e fornisce più trame utili per esplorare e visualizzare i risultati in ogni fase dell'analisi. Inoltre, fornisce anche metodi per l'integrazione funzionale annotando i geni in percorsi e termini GO. DEScan2 (Differential Enriched Scan 2) è un nuovo pacchetto R per analisi di dati ATAC-seq, una delle tecniche emergenti per lo studio dell'accessibilità alla cromatina. Consiste nella seguente procedura in tre fasi: 1) Identifica le regioni candidate all'interno di ciascun campione che implementa un picco chiamante; 2) Filtra gli artefatti potenziali allineando le regioni candidate tra i campioni e rimuovendo quelle regioni candidate che non erano riproducibili tra i campioni 3) Produce una matrice di conteggi di regioni e campioni, utile per l'arricchimento differenziale tra più condizioni e anche per integrazione di questo tipo di dati con altri dati omici, come RNA-seq.
IntegrHO (Integrazione di dati Omics High-Throughput) è un'interfaccia grafica utente (GUI), scritta in R e Shiny, finalizzata all'analisi e all'integrazione di tipi di dati multi-omici. Fornisce un'interfaccia amichevole agli strumenti sopra menzionati e incorpora anche un'ampia selezione di metodi e altri strumenti disponibili in letteratura. Questa piattaforma, attraverso un semplice approccio point-and-click, consente all'utente di analizzare ed esplorare singoli dati omici, come RNA-seq, ChIP-seq e ATAC-seq e, inoltre, offre la possibilità di integrarli in differenti livelli, come annotazione del picco genico e metodi di annotazione funzionale. Infine, poiché negli ultimi decenni la comunità scientifica ha vissuto una profonda crisi nota come "ricerca irreproduttiva", questa tesi presenta EasyReporting, un pacchetto R per una creazione automatica di report, sviluppato per supportare la riproducibilità della ricerca scientifica. Grazie al paradigma di classe R6 su cui si basa, è facile da usare e da estendere.
Nel complesso, questo lavoro propone e combina diversi strumenti computazionali per analizzare, visualizzare, confrontare, integrare e tracciare i diversi tipi di dati omici. I risultati sono illustrati con i dati scaricati dalla letteratura o dai progetti di collaborazione.

Abstract inglese:

Massive parallel sequencing technologies are producing a vast amount of genome- wide data about cells, tissues and model organisms, useful to understand many of biological mechanisms, like protein-chromatin interactions (e.g. ChIP-Seq), DNA methylation (Methyl-Seq or BS-Seq), chromatin accessibility (e.g. Atac- Seq), global transcriptional and translational activities (e.g. RNA Sequencing (RNA-seq)) and 3-D organisation of chromatin (e.g. Hi-C), giving the possibility to study same individual or experimental condition from many different points of view (transcriptomics, epigenomics, etc.) with a very high resolution. Each type of these omics data explains a different aspect of cellular behaviour. In order to extrapolate relevant information from each omics, it is required to develop specific statistical methodologies for single data analysis and, at the same time, computational methodologies for handling huge amount of data. However, to give a comprehensive view of the cell regulatory mechanisms, it is necessary not only to analize a single omics but also to develop novel statistical and computational models for integrating different omics types within a unified study.

This thesis is focused on the development of three main computational tools (ticorser, DEScan2 and IntegrHO), allowing data analysis and integration of multiple next-generation sequencing experiments. Additionally, it also contains a fourth tool (easyReporting) for reproducible computational research. Ticorser (time course RNA-seq data analyser) is a novel R package aimed to analyse time-course RNA-seq data. It offers multiple methods for differential ex- pression data analysis and provides multiple plots useful to explore and visualize the results at each step of the analysis. Furthermore, it also provides methods for functional integration by annotating genes in pathways and GO-terms. DEScan2 (Differential Enriched Scan 2) is a novel R package for ATAC-seq data analysis, one of the emerging techniques for investigating chromatin accessibil- ity. It consists in the following three-step procedure: 1) It identifies candidate regions inside each sample implementing a peak caller; 2) It filters out potential artefacts by aligning the candidate regions between the samples and removing those candidate regions that were not reproducible between samples 3) It pro- duces a count matrix of regions and samples, useful for differential enrichment between multiple conditions and also for integrating this data type with other omics data, such as RNA-seq.

IntegrHO (Integration of High-Throughput Omics data) is a Graphical User Interface (GUI), written in R and Shiny, aimed to analyze and integrate multi- omics data types. It provides a friendly interface to the above-mentioned tools and also incorporates a wide selection of methods and other tools available in the literature. This platform, through an easy point-and-click approach, enables the user to analyze and explore single omics data, such as RNA-seq, ChIP-seq and ATAC-seq and, moreover, it offers the possibility to integrate them at different levels, such as gene-peak annotation and functional annotation methods. Finally, since in last decades the scientific community has experienced a deep crisis known as ”irreproducible research”, this thesis presents EasyReporting, an R package for an automatic report creation, developed to support the repro- ducibility of scientific research. Thanks to the R6 class paradigm on which is based on, it is easy to use and to extend.

Overall, this work proposes and combines several computational tools for properly analyzing, visualizing, comparing, integrating and tracing different omics data types. The

results are illustrated with downloaded data from the literature or from collaboration projects.