Proposal: Building a large-scale single-cell multi-omics analysis platform for bench scientists

Xun Zhu

Introduction

With the astonishing advances in Single-cell omics techonologies, the integration of different omics analysis has been predicted to be the next big step. However, there has not been a dedicated platform that aims to enable this integration. In this proposal, we are aiming to develop a platform that integrates different components in single-cell genomics, epigenomics, transcriptomics, proteomics, phenomics, and metabolomics analysis. This platform, nick-named "Granatum", is going to bring the cutting-edge computational methods and algorithms to the large number of bench scientists who are trained in traditional biology background and thus are often frustrated by the time consuming process of letting a third-party bioinformatician to analyze the data.

In order to better track our progress, we break the seemingly massive project into three achievable aims.

Objectives and Approach

Aim 1: scRNA-Seq pipline

At this stage, we are going to develop and validate an RNA-Seq bioinformatics pipeline optimized for single-cell studies. This is set up as a pilot application which will showcase the value and potential of an accessible one-stop analysis platform. The pipline will include noval methods, as well as all essential steps commonly acknowledged in the single-cell research community. These steps should cover basic and some advance analysis procedures from removing the outlier, normalize the samples, and filtering the genes of the expression matries to biologically meaningful results such as clustering, differential expression, protein-protein interaction network, and pseudo-time construction. Each step will present the user with with a essential selection of the most widely used methods. Although at this initial stage the application will have an relative small size, it will consist of a fully functional graphical user interface and a operational

back-end supported by the Shiny framework in the R language. We will try to deploy it on the cloud for easy access. We will also provide the virtual machine installation instructions for the users.

To test the effectiveness of the pipline, we test it on several well-studied public datasets, demonstrating the pipline's ability to extract biologically relevant results from the data.

Aim 2: Methylation pipeline

Next, we are going to develop and validate a novel CpG methylome bioinformatics pipeline specifically tailored for single-cell studies. This pipeline will have a similar structure as that of the scRNA-Seq pipeline described in Aim 1. We will focus on developing a new computational method for detecting the methylation status of CpG sites by comparing sequencing patterns digested by methylation sensitive restriction enzymes between test data and control data. We will also develop noval computational methods for analyzing methylation data, as well as including state-of-the-art methods currently available.

The effectiveness of the pipeline will also be validated through testing on many publically available datasets.

Aim 3: Integrative platform

Finally, we will develop and validate an integrative bioinformatics toolset to analyze multi-dimensional single-cell NGS data, with single-cell RNA-Seq transcriptome, methylome data, and optionally other omics data such as the exome data. Based upon the previous experience of developing bioinformatics pipelines, we will develop an integrative platform that finds correlations at gene, pathway and network level among the single-cell NGS data from Aims 1, 2 and 3. The methodology will be generic and easily adaptable to other multi-dimensional high-throughput data types.

We will also aim to build and community eco-system for the platform. The single-cell computational method development has been extremely active in recent years, and new methods, tools, and algorithms are emerging every week. To make sure that our platform keep up with the latest development, we plan to build a unified and simple application program interface (API) to allow for simple integration for any method developer to easily integrate their tools with our platform.

Justification of the study

Single-cell technologies have demonstrated its massive potential in multiple domains such as stem cell, immunology, cancer research, and rare cell type

detection. However, the lack of unified, easy-to-use graphical analysis tools has been observed as a major barrier for many researchers. We believe the development of a standardized, intergrated, user-friendly pipline will prove to have of great value for the scientific community.