40_CONCORSO PUBBLICO, PER TITOLI ED ESAMI, PER LA COPERTURA A TEMPO DETERMINATO, DELLA DURATA DI CINQUE ANNI PER N. 1 POSTO DI RICERCATORE SANITARIO DA ASSEGNARE ALLA DIREZIONE SCIENTIFICA

PROVA I

1. Caratteristiche e differenze tra approcci di sequenziamento e analisi di short- e long-read sequencing

2. A cosa serve la funzione CERCA.VERT (VLOOKUP) in Excel?
   a) Calcola automaticamente la media dei valori contenuti in un intervallo
   b) Permette di cercare un valore in una colonna e restituire un valore corrispondente da un'altra colonna della stessa tabella
   c) Ordina i dati di una tabella in ordine crescente o decrescente

3. Leggere e tradurre il testo seguente

Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

Ricard Argelaguet[1,†], Britta Velten[2,†], Damien Arnol[1], Sascha Dietrich[3], Thorsten Zenz[3,4,5], John C Marioni[1,6,7], Florian Buettner[1,8,*], Wolfgang Huber[2,**] & Oliver Stegle[1,2,***]

Abstract

Multi-omics studies promise the improved characterization of biological processes across molecular layers. However, methods for the unsupervised integration of the resulting heterogeneous data sets are lacking. We present Multi-Omics Factor Analysis (MOFA), a computational method for discovering the principal sources of variation in multi-omics data sets. MOFA infers a set of (hidden) factors that capture biological and technical sources of variability. It disentangles axes of heterogeneity that are shared across multiple modalities and those specific to individual data modalities. The learnt factors enable a variety of downstream analyses, including identification of sample subgroups, data imputation and the detection of outlier samples. We applied MOFA to a cohort of 200 patient samples of chronic lymphocytic leukaemia, profiled for somatic mutations, RNA expression, DNA methylation and *ex vivo* drug responses. MOFA identified major dimensions of disease heterogeneity, including immunoglobulin heavy-chain variable region status, trisomy of chromosome 12 and previously underappreciated drivers, such as response to oxidative stress. In a second application, we used MOFA to analyse single-cell multi-omics data, identifying coordinated transcriptional and epigenetic changes along cell differentiation.

Keywords data integration; dimensionality reduction; multi-omics; personalized medicine; single-cell omics
Subject Categories Computational Biology; Genome-Scale & Integrative Biology; Methods & Resources
DOI 10.15252/msb.20178124 | Received 27 November 2017 | Revised 28 May 2018 | Accepted 29 May 2018
Mol Syst Biol. (2018) 14: e8124

Introduction

Technological advances increasingly enable multiple biological layers to be probed in parallel, ranging from genome, epigenome, transcriptome, proteome and metabolome to phenome profiling (Hasin *et al*, 2017). Integrative analyses that use information across these data modalities promise to deliver more comprehensive insights into the biological systems under study. Motivated by this, multi-omics profiling is increasingly applied across biological domains, including cancer biology (Gerstung *et al*, 2015; Iorio *et al*, 2016; Mertins *et al*, 2016; Cancer Genome Atlas Research Network, 2017), regulatory genomics (Chen *et al*, 2016), microbiology (Kim *et al*, 2016) or host-pathogen biology (Soderholm *et al*, 2016). Most recent technological advances have also enabled performing multi-omics analyses at the single-cell level (Macaulay *et al*, 2015; Angermueller *et al*, 2016; Guo *et al*, 2017; Clark *et al*, 2018; Colomé-Tatché & Theis, 2018). A common aim of such applications is to characterize heterogeneity between samples, as manifested in one or several of the data modalities (Ritchie *et al*, 2015). Multi-omics profiling is particularly appealing if the relevant axes of variation are not known *a priori*, and hence may be missed by studies that consider a single data modality or targeted approaches.

A basic strategy for the integration of omics data is testing for marginal associations between different data modalities. A prominent example is molecular quantitative trait locus mapping, where large numbers of association tests are performed between individual genetic variants and gene expression levels (GTEx Consortium, 2015) or epigenetic marks (Chen *et al*, 2016). While eminently useful for variant annotation, such association studies are inherently *local* and do not provide a coherent global map of the molecular differences between samples. A second strategy is the use of kernel- or graph-based methods to combine different

SM RP ES

MARCO BARRESI 10/12/1990 28/10/2025

Prova non estratta

Marco Barresi

40_CONCORSO PUBBLICO, PER TITOLI ED ESAMI, PER LA COPERTURA A TEMPO DETERMINATO, DELLA DURATA DI CINQUE ANNI PER N. 1 POSTO DI RICERCATORE SANITARIO DA ASSEGNARE ALLA DIREZIONE SCIENTIFICA

PROVA 2

1. Quali sono i principali vantaggi e svantaggi dell'uso di Whole Exome Sequencing rispetto a Whole Genome Sequencing in studi genomici, e in quali contesti ciascun approccio è preferibile?

2. Per URL si intende una sequenza di caratteri che:
   a) identifica univocamente l'indirizzo di una risorsa web
   b) un componente del sistema operativo
   c) un linguaggio di programmazione

3. Leggere e tradurre il testo seguente

Current best practices in single-cell RNA-seq analysis: a tutorial

Malte D Luecken[1] & Fabian J Theis[1,2,*]

Abstract

Single-cell RNA-seq has enabled gene expression to be studied at an unprecedented resolution. The promise of this technology is attracting a growing user base for single-cell analysis methods. As more analysis tools are becoming available, it is becoming increasingly difficult to navigate this landscape and produce an up-to-date workflow to analyse one's data. Here, we detail the steps of a typical single-cell RNA-seq analysis, including pre-processing (quality control, normalization, data correction, feature selection, and dimensionality reduction) and cell- and gene-level downstream analysis. We formulate current best-practice recommendations for these steps based on independent comparison studies. We have integrated these best-practice recommendations into a workflow, which we apply to a public dataset to further illustrate how these steps work in practice. Our documented case study can be found at https://www.github.com/theislab/single-cell-tutorial. This review will serve as a workflow tutorial for new entrants into the field, and help established users update their analysis pipelines.

Keywords analysis pipeline development, computational biology, data analysis tutorial, single-cell RNA-seq
DOI 10.15252/msb.20188746 | Received 16 November 2018 | Revised 15 March 2019 | Accepted 3 April 2019
Mol Syst Biol. (2019) 15: e8746

Introduction

In recent years, single-cell RNA sequencing (scRNA-seq) has significantly advanced our knowledge of biological systems. We have been able to both study the cellular heterogeneity of zebrafish, frogs and planaria (Briggs *et al*, 2018; Plass *et al*, 2018; Wagner *et al*, 2018) and discover previously obscured cellular populations (Montoro *et al*, 2018; Plasschaert *et al*, 2018). The great potential of this technology has motivated computational biologists to develop a range of analysis tools (Rostom *et al*, 2017). Despite considerable effort being undertaken by the field to ensure the usability of individual tools, a barrier of entry for novices in single-cell data analysis is the lack of standardization due to the relative immaturity of the field. In this paper, we present a tutorial for scRNA-seq analysis and outline current best practices to lay a foundation for future analysis standardization.

The challenges to standardization include the growing number of analysis methods (385 tools as of 7 March 2019) and exploding dataset sizes (Angerer *et al*, 2017; Zappia *et al*, 2018). We are continuously finding new ways to use the data at our disposal. For example, it has recently become possible to predict cell fates in differentiation (La Manno *et al*, 2018). While the continuous improvement of analysis tools is beneficial for generating new scientific insight, it complicates standardization.

Further challenges for standardization lie in technical aspects. Analysis tools for scRNA-seq data are written in a variety of programming languages—most prominently R and Python (Zappia *et al*, 2018). Although cross-environment support is growing (preprint: Scholz *et al*, 2018), the choice of programming language is often also a choice between analysis tools. Popular platforms such as Seurat (Butler *et al*, 2018), Scater (McCarthy *et al*, 2017), or Scanpy (Wolf *et al*, 2018) provide integrated environments to develop pipelines and contain large analysis toolboxes. However, out of necessity these platforms limit themselves to tools developed in their respective programming languages. By extension, language restrictions also hold true for currently available scRNA-seq analysis tutorials, many of which revolve around the above platforms (R and bioconductor tools: https://github.com/drisso/bioc2016singlecell and https://hemberg-lab.github.io/scRNA.seq.course/; Lun *et al*, 2016b; Seurat: https://satijalab.org/seurat/get_started.html; Scanpy: https://scanpy.readthedocs.io/en/stable/tutorials.html).

Considering the above-mentioned challenges, instead of targeting a standardized analysis pipeline, we outline current best practices and common tools independent of programming language. We guide the reader through the various steps of a scRNA-seq analysis pipeline (Fig 1), present current best practices, and discuss analysis pitfalls and open questions. Where best practices cannot be determined due to novelty of the tools and lack of comparisons, we list popular available tools. The outlined steps start from read or count matrices and lead to potential analysis endpoints. Earlier pre-processing steps are covered in Lun et al (2016b). A detailed case study that integrates the established current best practices is available on our github at: https://github.com/theislab/single-cell-tutorial/. Here, we have applied the current best practices in a practical example workflow to analyse a public dataset. The analysis workflow

MARCO BARROSI 10/12/1930 28/10/2015

Prove non esatte

Marco Barrosi

40_CONCORSO PUBBLICO, PER TITOLI ED ESAMI, PER LA COPERTURA A TEMPO DETERMINATO, DELLA DURATA DI CINQUE ANNI PER N. 1 POSTO DI RICERCATORE SANITARIO DA ASSEGNARE ALLA DIREZIONE SCIENTIFICA

PROVA 3

1. Qual è la differenza tra RNA-seq bulk e single-cell RNA-seq, anche dal punto di vista analitico?

2. Con il termine "database" si intende:
   a) un linguaggio di programmazione
   b) una collezione di dati, relativi ad una specifica attività, opportunamente strutturati e accessibili tramite un software di gestione
   c) un insieme di dati distribuiti sulla rete e accessibili solo tramite un browser

3. Leggere e tradurre il testo seguente

Recommendations for bioinformatics in clinical practice

Ksenia Lavrichenko[1], Emilie Sofie Engdal[2], Rasmus L. Marvig[2], Anders Jemt[3,4], Jone Marius Vignes[5], Henrikki Almusa[6], Kristine Bilgrav Saether[4], Eirikur Briem[7], Eva Caceres[1,4], Edda María Elvarsdóttir[7], Magnús Halldór Gíslason[2], Maria K. Haanpää[8], Viktor Henmyr[9], Ronja Hotakainen[10,11], Eevi Kaasinen[10,11], Roan Kanninga[1,2], Sofia Khan[10,11], Mary Gertrude Lie-Nielsen[5], Majbritt Busk Madsen[2], Niklas Mähler[12], Khurram Maqbool[2], Ramprasad Neethiraj[13], Karl Nyrén[4], Minna Paavola[8], Peter Pruisscher[3,4], Ying Sheng[1], Ashish Kumar Singh[13], Aashish Srivastava[5], Thomas K. Stautland[5], Daniel T. Andreasen[14], Esmee ten Berk de Boer[3,4], Søren Vang[14], Valtteri Wirta[3,4] and Frederik Otzen Bagger[2*]

Abstract

Background Next-generation sequencing (NGS) is well established in clinical diagnostics, and whole-genome sequencing (WGS) is increasingly becoming the method of choice, as a result of lower prices and robust comprehensive data. While guidelines exist for variant interpretation and laboratory quality considerations, there remains a need for standardised bioinformatics practices to ensure clinical consensus, accuracy, reproducibility and comparability.

Methods This article presents consensus recommendations developed by 13 clinical bioinformatics units participating in the Nordic Alliance for Clinical Genomics (NACG) by expert bioinformaticians working in clinical production. The recommendations are based on clinical practice and focus on analysis types, test and validation, standardisation and accreditation, as well as core competencies and technical management required for clinical bioinformatics operations.

Results Key recommendations include adopting the hg38 genome build as reference, and a standard set of recommended analyses, including the use of multiple tools for structural variant (SV) calling and in-house data sets for filtering recurrent calls. Clinical bioinformatics in production should operate at standards similar to ISO 15189, utilising off-grid clinical-grade high-performance computing systems, standardised file formats and strict version control. Reproducibility should be ensured through containerised software environments. Pipelines must be documented and tested for accuracy and reproducibility, minimally covering unit, integration and end-to-end testing. Standard truth sets such as GIAB and SEQC2 for germline and somatic variant calling, respectively, should be supplemented by recall testing of real human samples that have been previously tested using a validated method. Data integrity must be verified using file hashing, while sample identity must be confirmed through fingerprinting and genetically inferred identification markers such as sex and relatedness. Finally, clinical bioinformatics should encompass diverse skills, including software development, data management, quality assurance and domain expertise in human genetics.

RP

ES

TARCO BARESI 10/12/1930 29/10/2015

Fiore estratto

Tarco Baresi