МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

Кафедра генетики

ЗАЙКОВА
Ксения Михайловна

СРАВНИТЕЛЬНАЯ ХАРАКТЕРИСТИКА БИОИНФОРМАТИЧЕСКИХ ИНСТРУМЕНТОВ ДЛЯ КОЛИЧЕСТВЕННОГО АНАЛИЗА ЭКСПРЕССИИ ГЕНОВ

Аннотация
к дипломной работе

Научный руководитель: старший преподаватель И.Н. Ильюшёнак

Минск, 2021

РЕФЕРАТ

Дипломная работа 27 с., 7 рис., 6 табл., 47 источников.

СРАВНИТЕЛЬНАЯ ХАРАКТЕРИСТИКА БИОИНФОРМАТИЧЕСКИХ ИНСТРУМЕНТОВ ДЛЯ КОЛИЧЕСТВЕННОГО АНАЛИЗА ЭКСПРЕССИИ ГЕНОВ

Объект исследования: парноконцевые прочтения полного транскриптома клеток Kasumi-1, с номерами кода доступа SRR1145838, SRR1145839, SRR1145840.

Цель: сравнение эффективности различных биоинформатических инструментов анализа экспрессии генов.

Методы исследования: биоинформатрические, молекулярно-генетические.

Анализ крупных библиотек RNA-Seq является сложной задачей, и различные программные инструменты используют разные алгоритмы для её решения. Также особенности реализации алгоритмов в каждом конкретном случае приводят к тому, что программы для анализа RNA-Seq имеют различную эффективность – как при оценке экспрессии на уровне гена, так и отдельного транскрипта.

Поэтому на сегодняшний день не существует некоего общепринятого «золотого стандарта» в анализе данных RNA-Seq. Для проверки же эффективности их работы можно использовать «устоявшиеся», рутинные лабораторные методики – такие, как количественная ПЦР. Из-за всех вышеперечисленных факторов появляется необходимость проводить исследования, чтобы определить, какие биоинформатические алгоритмы работают лучше.

По результатам сравнения таких биоинформатических инструментов как Cufflinks, StringTie и Salmon было выяснено, что точность оценки экспрессии у Cufflinks и StringTie ниже, чем у Salmon. А сам Salmon имеет зависимость от сборщика транскриптов. При сравнении этих же результатов с оценкой, проведенной при помощи количественной ПЦР наиболее близкой, оказалась оценка полученная StringTie.

Исходя из этого был сделан следующий вывод, что наиболее точными из инструментов являются StringTie и Salmon.

РЭФРАТ

Дыпломная праца 27 с., 7 мал., 6 табл., 47 крыніц.

ПАРАЎНАЛЬНАЯ ХАРАКТЫРЫСТЫКА БІЯІНФАРМАТЫЧНЫХ ІНСТРУМЭНТАЎ ДЛЯ КОЛЬКАСНАГА АНАЛІЗУ ЭКСПРЭСІІ ГЕНАЎ

Аб'ект даследаванні: парнаканцавыя прачытання поўнага транскыптома клетак Kasumi-1, з нумарамі доступу SRR1145838, SRR1145839, SRR1145840.

Мэта: параўнанне эфектыўнасці розных біяінфарматычных інструментаў аналізу экспрэсіі генаў.

Метады даследавання: біяінфарматрычныя, малекулярна-генетычныя.

Аналіз буйных бібліятэк RNA-Seq з'яўляецца складанай задачай, і розныя праграмныя інструменты выкарыстоўваюць розныя алгарытмы для яе рашэння. Таксама асаблівасці рэалізацыі алгарытмаў ў кожным канкрэтным выпадку прыводзяць да таго, што праграмы для аналізу RNA-Seq маюць розную эфектыўнасць - як пры ацэнцы экспрэсіі на ўзроўні гена, так і асобнага транскрыптаў.

Таму на сённяшні дзень не існуе нейкага агульнапрынятага «залатога стандарту» у аналізе дадзеных RNA-Seq. Для праверкі жа эфектыўнасці іх работы можна выкарыстоўваць «ўстояныя», руцінныя лабараторныя методыкі - такія, як колькасная ПЦР. З-за ўсіх вышэйпералічаных фактараў з'яўляецца неабходнасць праводзіць даследаванні, каб вызначыць, якія біяінфарматычныя алгарытмы працуюць лепш.

Па вынікам параўнання такіх біяінфарматычных інструментаў як Cufflinks, StringTie і Salmon было высветлена, што дакладнасць ацэнкі экспрэсіі ў Cufflinks і StringTie ніжэй, чым у Salmon. А сам Salmon мае залежнасць ад зборшчыка транскрыптаў. Пры параўнанні гэтых жа вынікаў з ацэнкай, праведзенай пры дапамозе колькаснай ПЦР найбольш блізкай апынулася адзнака атрыманая StringTie.

Зыходзячы з гэтага был зроблен наступны выник, што найбольш дакладнымі з інструментаў з'яўляюцца StringTie і Salmon.

ABSTRACT

Graduate work 27 p., 7 pict., 6 tabl., 47 references.

COMPARATIVE CHARACTERISTICS OF BIOINFORMATIC TOOLS FOR QUANTITATIVE ANALYSIS OF GENE EXPRESSION.

Object of research: pair-terminal readings of the complete transcriptome of Kasumi-1 cells, with access code numbers SRR1145838, SRR1145839, SRR1145840.

Aim of work: to compare the effectiveness of various bioinformatics tools for gene expression analysis.

Methods: bioinformatric, molecular-genetic.

Analyzing large RNA-Seq libraries is challenging and different software tools use different algorithms to solve it. Also, the peculiarities of the implementation of the algorithms in each specific case lead to the fact that the programs for the analysis of RNA-Seq have different efficiency - both in assessing the expression at the gene level and in an individual transcript.

Therefore, today there is no generally accepted "gold standard" in the analysis of RNA-Seq data. To test the effectiveness of their work, you can use "established", routine laboratory techniques such as quantitative PCR. Due to all of the above factors, it becomes necessary to conduct research to determine which bioinformatics algorithms perform best.

Comparing bioinformatics tools such as Cufflinks, StringTie, and Salmon, it was found that the accuracy of expression estimation for Cufflinks and StringTie was lower than that for Salmon. And Salmon itself has a dependency on the transcript collector. When comparing the same results with the quantitative PCR estimate, the closest was the estimate obtained by StringTie.

Based on this, the following conclusion was made that the most accurate of the tools are StringTie and Salmon.