Methods 97 (2016) 3–10

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Analyzing HT-SELEX data with the Galaxy Project tools – A web based bioinformatics platform for biomedical research

William H. Thiel [a,b,*], Paloma H. Giangrande [a,b,c,d]

[a] Department of Internal Medicine, University of Iowa, Iowa City, IA 52242, USA
[b] The François M. Abboud Cardiovascular Research Center, University of Iowa, Iowa City, IA 52242, USA
[c] The Holden Comprehensive Cancer Center, University of Iowa, Iowa City, IA 52242, USA
[d] The Molecular and Cell Biology Program, University of Iowa, Iowa City, IA 52242, USA

ARTICLE INFO

Article history:
Received 30 July 2015
Received in revised form 12 October 2015
Accepted 15 October 2015
Available online 23 October 2015

Keywords:
Aptamer
SELEX
Bioinformatics
Galaxy
High-throughput sequencing (HTS)
Next-generation sequencing (NGS)

ABSTRACT

The development of DNA and RNA aptamers for research as well as diagnostic and therapeutic applications is a rapidly growing field. In the past decade, the process of identifying aptamers has been revolutionized with the advent of high-throughput sequencing (HTS). However, bioinformatics tools that enable the average molecular biologist to analyze these large datasets and expedite the identification of candidate aptamer sequences have been lagging behind the HTS revolution. The Galaxy Project was developed in order to efficiently analyze genome, exome, and transcriptome HTS data, and we have now applied these tools to aptamer HTS data. The Galaxy Project's public webserver is an open source collection of bioinformatics tools that are powerful, flexible, dynamic, and user friendly. The online nature of the Galaxy webserver and its graphical interface allow users to analyze HTS data without compiling code or installing multiple programs. Herein we describe how tools within the Galaxy webserver can be adapted to pre-process, compile, filter and analyze aptamer HTS data from multiple rounds of selection.

1. Introduction

Aptamers are small (20–100 nucleotide) structured DNA or RNA oligonucleotides that interact with a target molecule with a high degree of specificity and affinity. Aptamers can be generated with affinities similar to those of monoclonal antibodies and have been referred to as "chemical antibodies" or "nucleic acid antibodies" [1]. Several aptamers are currently undergoing various stages of clinical trials [2], and one aptamer, Macugen, has been FDA approved for the treatment of age-related macular degeneration [2].

Aptamers with affinity and selectivity for a target can be developed utilizing the SELEX process [3,4], Systematic Evolution of Ligands by EXponential enrichment. The SELEX process begins with a DNA or RNA oligonucleotide library that contains 5′ and 3′ constant regions flanking a variable region. The length of the variable region, which typically ranges from 20 to 60 nucleotides, dictates the complexity of the starting aptamer library, yielding 10^{12}–10^{36} different aptamer sequences, respectively. The SELEX process removes non-specific aptamers from the library and enriches for aptamers highly specific for the target molecule using a combination of negative and positive selection pressures. The SELEX process has been adapted to include a wide-range of conditions. For all SELEX processes, the sequence information of the enriched aptamers must be determined. Originally, the precise sequences of aptamers from the final round of selection was obtained using sub-cloning techniques followed by sequencing each clone one at a time. High-throughput sequencing (HTS) fundamentally changed the aptamer field by enabling the sequencing of hundreds of millions of reads from multiple rounds of a selection [5–8]. With aptamer HTS data came the need for more sophisticated aptamer bioinformatics methods [9]. Bioinformatics analysis of these HTS data is necessary to narrow down the hundreds of millions of sequences to a select few candidates that can be feasibly evaluated experimentally [5].

Aptamer bioinformatics begins with pre-processing the aptamer HTS data to remove adapter/barcode/constant region sequences and sequences with mismatches within the constant region. Pre-processing also consists of setting a variable region length cutoff and counting the number of identical reads [5,10,11]. The next step of aptamer bioinformatics is a first pass

Abbreviations: SELEX, Systematic Evolution of Ligands by EXponential enrichment; HTS, high-throughput sequencing.

* Corresponding author at: University of Iowa, 5241 MERF, 375 Newton Rd., Iowa City, IA 52242, USA.
E-mail addresses: william-thiel@uiowa.edu (W.H. Thiel), paloma-giangrande@uiowa.edu (P.H. Giangrande).

http://dx.doi.org/10.1016/j.ymeth.2015.10.008
1046-2023/

course filtering to separate non-selected sequences from selected sequences. This first pass filtering can be accomplished by a variety of methods: (1) calculating fold enrichment of an aptamer sequence between selection rounds [5,10–12]; (2) comparing the round representation of each sequence to a non-selected round (e.g. round 0) [5]; (3) comparing the number of identical reads of a sequence (read count) to a non-selected round (e.g. round 0) [5]; and (4) isolating aptamer sequences with shared homology [13]. Frequently, a subsequent second pass in-depth analysis is performed to identify sequence families [5,7,12,13], structure families [5], motifs [14,15] and potential beneficial mutations [16]. Several bioinformatics tools have been generated to analyze aptamer HTS data [9,10,13,16]. Unfortunately, many of these tools are either not easily accessible, have extensive requirements prior to use or require significant computational/coding expertise. These requirements exclude most molecular biologists with expertise in aptamer selections from conducting their own data analysis. To meet this critical need we have adapted the web-based Galaxy Project [17–19] tools for analyzing HTS genome, exome and transcriptome datasets for the analysis of HTS aptamer data. The Galaxy Project's public webserver, termed the "Main instance," is a freely available collection of bioinformatics tools that are powerful, flexible, dynamic, and most importantly easy to use with minimal computational expertise/knowledge. All that is needed to start using Galaxy is a web browser, a freely available account with Galaxy and aptamer HTS data. Herein, we describe methods where we have adapted the tools within Galaxy to pre-process and conduct a first-pass course analysis of aptamer HTS data.

2. Equipment

2.1. Computer

A computer is necessary to store and upload data files. These files will include HTS data (e.g. FASTQ) and text files with output data. Any computer capable of running a current web browser along with an internet connection will be sufficient.

2.2. Web browser

All major web browsers (e.g. Chrome, Firefox, Internet Explorer, Safari, Opera) are supported by Galaxy.

2.3. Galaxy Project main webserver account

This article is based on a registered account of the web-based Main instance of Galaxy (https://usegalaxy.org/). The registration process (https://usegalaxy.org/user/create?use_panels=True) requires a valid email address and an approval with the Galaxy Web Portal Service Agreement (https://usegalaxy.org/static/terms.html). Each user is allowed one registered account, which includes 250 GB server space for data and permits running a maximum of six concurrent jobs. The Galaxy Wiki includes additional information about Galaxy Main user accounts (https://wiki.galaxyproject.org/Learn/UserAccounts) and a video walk-through (https://vimeo.com/75925027) of the Galaxy Main registration is available online.

3. Methods

3.1. Getting started

3.1.1. Galaxy webserver workspace

The Main instance of the Galaxy webserver workspace includes three panels. The left-most panel contains all of the tools used for data analysis and is referred to as "Tools". These tools are reference throughout this article (e.g. NGS: QC and manipulation/Collapse). The right-most panel keeps track of data files and work history of the data analysis and will be referred to as "History". Selecting a data file within the History will expand the data files options. These options include view data, edit attributes, deleting and download. The central panel is a dynamic space and will contain the menu options for selected tools from Tools or a sample of data from files selected within the History.

3.1.2. Getting help

Galaxy includes extensive documentation through the Galaxy Wiki (https://wiki.galaxyproject.org/). When a tool under Tools is selected, the middle panel of Galaxy will include instructions and examples of data input and output.

3.1.3. Moving files

The analysis of HTS aptamer data using Galaxy will require uploading data files to the Galaxy web server. These files may include compressed HTS raw data files and text files for using the Barcode Splitter tool. The Galaxy webserver has multiple routes for uploading data files. Small files, such as the barcode text file, can be uploaded using The Get Data/Upload File > Choose Local File option. However, large files, such as the raw HTS data files, will need to be uploaded using FTP and may take several hours (10+) depending on network connection speeds and file sizes. More information on uploading data files by FTP is available on the Galaxy Wiki (https://wiki.galaxyproject.org/FTPUpload) and a video walk-through describing all of the methods for uploading data to the Galaxy web server can be found at (https://vimeo.com/75938324). Downloading data files from the Galaxy web server can be accomplished by selecting the download option of a file within the work history.

3.1.4. Running multiple jobs

Each tool running is referred to as a "job" and Galaxy allows registered users to run up to six jobs concurrently. Each "job" may take several minutes to several hours to complete depending on server load and the complexity of the "job". Tools within Galaxy can be stacked sequentially in the History and several jobs may be executed prior to the previous tool completing a task. The History panel color codes each "job" to reflect status; green = completed, yellow = currently working; gray = queued; red = failed.

3.1.5. Edit attributes

Data files in the History are numbered sequentially and are referenced with the tool from the previous job. This information can be edited by expanding the data file in the History and selecting the "Edit attributes" pencil icon.

3.1.6. HTS aptamer data format

The characteristics of the HTS aptamer data (e.g. FASTQ) should be understood prior to bioinformatics analysis. To date, Illumina sequencers are the most common HTS platform for acquired HTS aptamer data. Three criteria should be noted from DNA amplicon molecule (Fig. 1) used to attain HTS aptamer data.

3.1.6.1. Read length. Read length refers to the number of nucleotides sequenced and is a fixed number ranging from 50 to 150 nucleotides. The read length is important with aptamer HTS data in determining how far the sequence data extends through the variable region into the 3′ constant region. Longer read lengths will likely extend through the entirety of the 3′ constant region into the Illumina priming/adapter sequence.

3.1.6.2. Single-end read or paired-end read. Single-end read describes HTS data attained from one end of the amplicon DNA