Seqs-Extractor: Automated sequences extraction to reduce tedious manual corrections of large datasets.

Patrick D. C. Pereira[1], Cleyssian Dias[2], Mauro A. D. Melo[1], Nara G. M. Magalhães[3], Cristovam Guerreiro-Diniz[1], Cristovam W. Picanço-Diniz[3]

[1] Laboratório de Biologia Molecular e Neuroecologia, Instituto Federal de Educação, Ciência e Tecnologia do Pará, Bragança, Pará, Brasil.

[2] Grupo de Genética e Conservação, Universidade Federal do Pará, Bragança, Pará, Brasil.

[3] Laboratório de Investigações em Neurodegeneração e Infecção no Hospital Universitário João de Barros Barreto, Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Pará, Brasil.

Corresponding Author:

Patrick Pereira[1]

Rua da Escola Agrícola, S/N, Vila Sinhá, Bragança, Pará, Brasil, 68600-000.

Email address: patrick@ufpa.br

Abstract

The analysis of large numbers of sequences requires the reduction of ambiguities during the analytical work to ensure that the effort will focus only on confirmed sequences. Performing this work automatically may help to minimize potential errors associated with tedious manual correction, allowing more effective results. Basic local alignment search tool (BLAST) seems to be the most widely used sequence analysis program. It is free, but commercial parties enhanced BLAST applications and charge a fee for their uses. There are some tools of public domain that can perform the search of microsatellites in the next generation sequencing (NGS) data, as the microsatellite identification tool (MISA), which has some features to discover microsatellites in large datasets. Here, we developed a basic shell script (BASH script) to be ran under Linux environment that can be used to extract from a sequence dataset only confirmed (BLASTed) sequences from both nucleotide

(BLASTN) and protein (BLASTX) databases and extract sequences that contains microsatellites using MISA tool, using a friendly interface and no fees charged. Seqs-Extractor is a helpful tool that may enhance the analysis of large datasets in BLAST+ and MISA by minimizing the time of management, reducing potential errors caused by manipulating data and no fees charged. Seqs-Extractor is available at https://github.com/patrick-douglas/Seqs-Extractor/wiki.

Subjects: Bioinformatics, Computational Biology, Genomics.

Keywords: Sequences analysis, Next-generation sequencing, Databases, Bash.

Introduction

Evaluation of sequences homology is the most common way to help understand the biology of specific organisms (Donkor et al. 2014). Indeed, from sequence homology studies we may recognize for example, the homology between protein or DNA sequences, that may shed light on shared ancestry in the evolutionary history of life. The main program used to achieve this goal is the popular Basic Local Alignment Search Tool (BLAST), that uses a heuristic algorithm which performs comparisons between pairs of sequences, searching for regions with some similarity (Altschul et al. 1990).

BLAST+ is the standalone BLAST suite used in an offline environment, when is not possible to work in online-BLAST, normally due to the large size of datasets to be processed (Camacho et al. 2009). To perform a BLAST search of large datasets using standalone tools, a minimum command line knowledge is required, because the standalone application relies on that. Indeed, the command line interface introduces

technical difficulties ranging from installation to generation of results (Camacho et al. 2009). Free easier to use software, with graphic interface may help users to circumscribe operational difficulties of command line interface, but this free facility is limited to smaller datasets (Excoffier & Lischer 2010). Alternatively, the users may pay for software with friendly user interface, like Blast2GO (Conesa et al. 2005) that perform BLAST search and many others analyses of larger datasets, however the cost to use paid programs can be high.

In the next generation sequencing (NGS) the BLAST search is a good way to validate the assembled sequences by comparing it with a valid database. Moghadam et al. (2013) used this approach to validate their transcripts of *Charadrius vociferous* (obtained in *De novo* RNA-Seq assembly) BLASTing it against some datasets of genome sequences from Chicken, Turkey and Zebra finch, using only the top hit sequences (sequences that matched 80% to 100% with a valid database). This procedure may avoid considering contaminant derived sequences during analysis. However, get only the top hit sequences are challenger because a BLAST search generates output files containing only details about alignment scores of entered sequences and the subject database, but does not generated a file containing sequences that match with a specific percentage value.

Simple sequence repeats (SSR) or microsatellite markers are known as a good tool to identification of genetic distances among organisms, and is commonly used in the population genetic studies (Fernandez-Silva et al. 2013; Koohi-Dehkordi et al. 2006). In the NGS studies the use of SSR is very common the production of a large amount of sequences containing microsatellites (Castoe et al. 2012; Donkor et al. 2014; Fernandez-Silva et al. 2013; Guichoux et al. 2011).

There are some tools of public domain that can perform the search of microsatellites

in the NGS data, as the microsatellite identification tool (MISA), which has some features to discover microsatellites in large datasets (Thiel et al. 2003) and has already been used in several other studies (Khlestkina et al. 2004; Varshney et al. 2005; Varshney et al. 2002; Yu et al. 2004).

Similar to BLAST+, the MISA tool provides the microsatellites results in a table containing the information about length, nucleotides repeats, motifs but does not allow the access of sequences that contains microsatellites. A common run of MISA in a NGS dataset can produce hundreds of microsatellites results, and get manually only the sequences that contains these results is usually unfeasible due to the tedious and the long time spent on it.

Here we provided a bash script, named as Seqs-Extractor that can run natively under Linux, which can extract in isolation from .FASTA dataset, the sequences that match in a BLASTX or a BLASTN search (with a match percentage defined by the user), as well as positive sequences of a MISA search.

Our script can also perform an independent BLAST+ search, using a friendly interface and no fees charged, considering cases when the user did not run BLAST+ yet. In addition, Seqs-Extractor can extract sequences from a .FASTA dataset using only a simple text file containing only the sequence IDs.

Methods

Seqs-Extractor is implemented in BASH command language and run natively in Linux systems, which are based in the Debian version. This script was tested under Linux Mint 17.3 Xfce and Ubuntu 16.4LTS. At time *Seqs-Extractor* does not work in CentOS/Red Hat systems. The software is freely available to be run locally in personal computer.

Seqs-Extractor uses the following free and open source third-party software: BLAST+ (Altschul et al. 1990), MISA (Thiel et al. 2003) and SAMTOOLS (Li et al. 2009).

Results

To easily perform a BLAST search in the command line version, *Seqs-Extractor* provides a friendly interface where all users should inform some minimum required parameters.

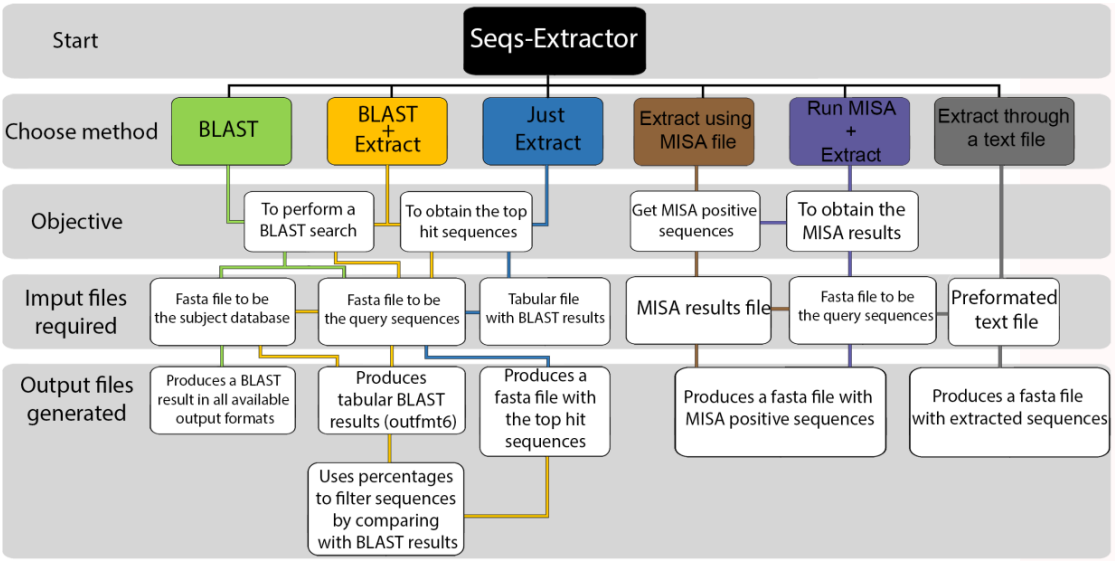
To test the script operation we used a .FASTA file that contains 107,185 sequences of the entire mouse genome (*Mus musculus*) obtained in NCBI database (NCBI Annotation Release 106) and run a BLASTX search against 168,031 sequences of revised and manually annotated mouse protein database obtained in (The universal protein resource (UniProt)). After BLASTX ran the resulting file contained 123,371 lines with the search results (considering expected value of 1^{-3}), where each line represents a result of alignment. To get only the sequences that show 100% homology with the subject we used *Seqs-Extractor* generates a new .FASTA file containing only 47,184 sequences with 100% confirmed results aligned with the subject database.

Thus, *Seqs-Extractor* uses the query ID provided in the tabular file, to search inside the .FASTA file and extract only sequences that match in a specific percentage level defined by user. *Seqs-Extractor* will then generate two files: the extracted sequences and the tabular results of BLAST search. All methods use the BLASTN or BLASTX.

We also used MISA (Thiel et al. 2003) through *Seqs-Extractor* to perform a search (using default parameters) and extraction of microsatellites inside the .FASTA file that contains entire mouse genome. This run generated three output files: the MISA results file, the default MISA statistics file and the .FASTA file containing the positive MISA sequences.

In the extraction process *Seqs-Extractor* uses SAMTOOLS application (Li et al. 2009). This script works by comparing the data in the tabular BLAST format and automatically search for sequences in the .FASTA file that match a percentage of alignment specified by the user, creating a new .FASTA file containing only the sequences filtered by percentage criteria. Similarly, SAMTOOLS extract sequences with microsatellites by comparing data in the .FASTA file with IDs of these sequences in the MISA file results. All tests performed and example files can be found in the Seqs-Extractor webpage on Git-hub website (https://github.com/patrick-douglas/Seqs-Extractor/wiki). The flowchart showing the *Seqs-Extractor* workflow is in Figure 1.

Figure 1 – Flowchart. Working steps of the six possible methods of Seqs-Extractor. The steps of each method can be followed according to the color of the fluxes.



Conclusions

Seqs-Extractor is an automated tool that enhances BLAST analysis using a friendly interface, besides extracting from a sequence dataset only 100% confirmed (BLASTed) sequences. It also allows an extraction of positive microsatellites sequences from a MISA

search. It is a free of charge program that provides automatically search analysis in larger datasets that may help to minimize potential errors associated with tedious manual correction, thus allowing more effective results.

Acknowledgements

We acknowledge Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Instituto Federal do Pará (IFPA) and Universidade Federal do Pará (UFPA).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Patrick D. C. Pereira: Development, source code analysis, manuscript conception writing

Cleyssian Dias: Source code analysis, manuscript conception and writing, testing

Mauro A. D. Melo: Manuscript conception and writing, organization of data

Nara G. M. Magalhães: Manuscript conception and writing, organization of data, data examples building.

Cristovam Guerreiro-Diniz: Source code analysis, manuscript conception and writing, testing

Cristovam W. Picanço-Diniz: Manuscript conception, writing and final revision, testing, organization of data, bug fixes.

Data Availability

The following information was supplied regarding data availability:

Seqs-Extractor is available at:

https://github.com/patrick-douglas/Seqs-Extractor/wiki

References

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215:403-410.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. 2009. BLAST+: architecture and applications. *BMC bioinformatics* 10:1.

Castoe TA, Poole AW, de Koning AJ, Jones KL, Tomback DF, Oyler-McCance SJ, Fike JA, Lance SL, Streicher JW, and Smith EN. 2012. Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS one* 7:e30953.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, and Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674-3676.

Consortium U. 2008. The universal protein resource (UniProt. *Nucleic acids research* 36:D190-D195.

Donkor ES, Dayie NT, and Adiku TK. 2014. Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). *Journal of Bioinformatics and Sequence Analysis* 6:1-6.

Excoffier L, and Lischer HE. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources* 10:564-567.

Fernandez-Silva I, Whitney J, Wainwright B, Andrews KR, Ylitalo-Ward H, Bowen BW, Toonen RJ, Goetze E, and Karl SA. 2013. Microsatellites for next-generation ecologists: a post-sequencing bioinformatics pipeline. *PLoS one* 8:e55990.

185 Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C,
186 Malausa T, Revardel E, and Salin F. 2011. Current trends in microsatellite
187 genotyping. *Molecular ecology resources* 11:591-611.

188 Khlestkina EK, Than MHM, Pestsova EG, Röder MS, Malyshev SV, Korzun V, and
189 Börner A. 2004. Mapping of 99 new microsatellite-derived loci in rye (Secale
190 cereale L.) including 39 expressed sequence tags. *Theoretical and Applied*
191 *Genetics* 109:725-732.

192 Koohi-Dehkordi M, Sayed-Tabatabaei B, Yamchi A, and Danesh-Shahraki A. 2006.
193 Microsatellite markers in pomegranate. XXVII International Horticultural
194 Congress-IHC2006: II International Symposium on Plant Genetic Resources of
195 Horticultural 760. p 179-184.

196 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and
197 Durbin R. 2009. The sequence alignment/map format and SAMtools.
198 *Bioinformatics* 25:2078-2079.

199 Moghadam HK, Harrison PW, Zachar G, Székely T, and Mank JE. 2013. The plover
200 neurotranscriptome assembly: transcriptomic analysis in an ecological model
201 species without a reference genome. *Molecular ecology resources* 13:696-705.

202 Thiel T, Michalek W, and Varshney R. 2003. Exploiting EST databases for the
203 development of cDNA derived microsatellite in barley (Hordeum vulgare L,).
204 *Theoretical and Applied Genetics* 106:411-422.

205 Varshney RK, Graner A, and Sorrells ME. 2005. Genic microsatellite markers in plants:
206 features and applications. *TRENDS in Biotechnology* 23:48-55.

207 Varshney RK, Thiel T, Stein N, Langridge P, and Graner A. 2002. In silico analysis on
208 frequency and distribution of microsatellites in ESTs of some cereal species.
209 *Cellular and Molecular Biology Letters* 7:537-546.

210 Yu J-K, Dake TM, Singh S, Benscher D, Li W, Gill B, and Sorrells ME. 2004.
211 Development and mapping of EST-derived simple sequence repeat markers for
212 hexaploid wheat. *Genome* 47:805-818.

213

214