**PROCESSING OF LONG TRANSCRIPTOM SEQUENING READS USING AMAZON WEB SERVICES CLOUD COMPUTING PLATFORM**

Shapovalova V.V..1, Radko S. P.2, Ptitsyn K. G. 2, Krasnov G. S. 2, Nakhod K.V.\* 2, Konash O.S.. 2, Vinogradina M. A. 2, E.A. Ponomarenko. 2, Druzhilovskiy D. S. 2, , Lisitsa A. V. 2

1 - Center for Strategic Planning and Management of Biomedical Health Risks, 119121, Moscow, st. Pogodinskaya, 10, building 1

2 - Research Institute of Biomedical Chemistry named after V. N. Orekhovich 119121, Russia, Moscow, st. Pogodinskaya, 10, building 8.

\*g-s2011@mail.ru

**APPENDIX # 1**

**Step-by-step instructions for processing the results of quantitative analysis of human transcriptome using a nanopore sequencer**

***STAGE 1*** - Launching and connecting to the virtual machine

***STAGE 2*** - Running Programs on the AWS Platform

***STAGE 3*** - Transferring Calculation Results and Shutting Down the Virtual Machine

***STAGE 1 -*** *Launching and connecting to the virtual machine*

Step 1. Launch a virtual machine on an AWS server (Amazon Web Services).

Using a web browser, navigate to the Sharing AWS Resources page. To do this, specify “Amazon AWS” in the search bar of your browser (Google, Yandex, etc.) and follow the corresponding link. Find the [Login to the Console] link button on the web page and enter the authorization information provided by the authors in the form of additional materials to the article upon request (Fig. 1-1). On the screen that appears after authorization, select the [EC2] links (Fig. 1-2) and then [Running Instances] (Fig. 1-3). Then click on the Mr. First (ONT) (fig. 1-4) and using the top menu option [Actions] launch the virtual machine by selecting [instance state] as shown in fig. 2-5.



**Figure 1**. Running a virtual machine on the Amazon Web Services platform.

**(1)** Entering authorization information to log into the user console. **(2)** AWS User Console, where you select the EC2 (Elastic Compute Cloud) hyperlink to navigate to cloud resources**. (3)** Select [Running Instance] in the resource list to go to the list of virtual machines to select Mr.First machine **(4).**

** Figure 2**. Starting, shutting down the virtual machine and changing its configuration Starting **(5)** and shutting down **(6)** the virtual machine in the [Instance state] menu. (6) Change the virtual machine **configuration (7)** in the [Instance setting] menu.

Please note that changing the configuration of a virtual machine can only be performed after it is turned off, then a restart is required. When shutting down, the connection to the PuTTY / WinSCP virtual computer is terminated and must be resumed (see below).

Step 2. Connecting to a virtual computer.

To connect to the selected virtual machine. on the Amazon platform, run the PuTTy terminal program on a Windows computer **(Figure 3-1)**. Connecting to a remote computer will require you to enter (a) the address of the remote machine (“host” address), which is copied from the cloud management console **(Figure 3-2, Figure 3-3)**. and (b) the path to the key file that protects access to the remote machine **(Figure 4-4, Figure 4-5).** The key file was generated when Mr. First (ONT) and named bober.ppk (available as part of supplementary materials for this article, it is recommended to copy it to the Windows desktop). When loading the terminal, enter the data at the login: "ubuntu" request. **(Figure 5-6).**

In the terminal that appears on a black background, a prompt for entering ("promt") in the form of a "$" sign will appear, indicating that the terminal is ready for operation **(Fig. 5-7).**

 **Figure 3.** Connecting to the file system of a remote machine using the PuTTy program (https://www.putty.org/). **(1)** Select the virtual machine in "Running" mode, go to the [Connect] menu item. **(2)** In the [Connect to instance] window, copy the line containing the temporary IP address of the virtual machine. **(3)** The copied address should be pasted into the “Host Name (or IP address)” field of PuTTy.

 **Figure 4.** Connecting to the file system of a remote machine using PuTTy (https://www.putty.org/). **(4)** In the hierarchical menu of the PuTTy program, select the Connections> SSH> Auth sub-item and, using the [Browse] button, specify in the "Public key file for authentication" field the name of the public key file .ppk **(5)**, which is included in the supplementary materials to the article (available upon request).

 **Figure 5.** Connecting to the file system of a remote machine using PuTTy (https://www.putty.org/). **(6)** When loading the terminal, enter the data at the login: "ubuntu" request**. (7)** The terminal is ready for operation.

Establish a [connect] connection using the WinSCP file manager (Figure 6-1). To do this, specify the name of the “host” (Fig. 6-2, 6-3). and the path to the key file (bober.ppk, see Fig. 6-4, Fig. 6-5), in the authorization window (6-6) using the [enter] button (6-7) a file manager window opens for Working with files on the AWS server (Figure 6-8).

 **Figure. 6.** Connecting to the file system of a remote machine using the WinSCP program (https://winscp.net). **(1)** Establishing a connection using the [connect] option **(2)** Copying the address of the remote machine ("host" address) from the cloud management console to **(3)** the WinSCP control system. **(4).**

 **Figure. 7.**Connecting to the file system of the remote machine using the WinSCP program (https://winscp.net). (4) In the hierarchical menu of the WinSCP program, select the Authentication sub-item and, using the [...] button, and specify the name of the public key file .ppk ( 5), which is part of the supplementary materials to the article (available upon request). Specifying the bober.ppk key file (5) in the authorization window (6) using the [login] button (7) opens access to the data disks on the local server and on the AWS virtual machine. (8).

When connecting using the PuTTy / WinSCP utilities, a warning appears about the absence of a key file in the cache, which must be ignored by selecting the "Yes" option.

Step 3. Checking the availability of free virtual disk space on the virtual machine.

The Linux operating system (OS) command “df” allows you to get information about the free disk space in the work \ directory, which is a virtual data storage (disk, or volume in Amazon Web Services terminology) connected to a virtual machine. The “df -h” command is launched in the PuTTy terminal and displays information about the availability of free space on the disks (the “-h” parameter specified after the “df” command means “human readable”, that is, the result of the command is displayed in an intuitive user view). If the working disk, in particular the work \ directory, is occupied by more than 60%, then it should be freed from the results of previous starts of the pipe-line (see the section "The final stage").

Step 3a. Copying data using the WinSCP file manager.

It is recommended to use the WinSCP file manager to copy data (see Figure 8). Before copying data to the AWS cloud platform, determine the size and location of the source folder with the sequencing results on the local computer (or on the storage system of the organization) and call it fast5. To copy, you need to move the folder with files from the file panel of the local computer to the work / directory of the file panel of the remote virtual machine in the WinSCP graphical interface. Copying 100 GB will take 8-10 hours on average, so it is advisable to set copying overnight, while using the cheapest configuration of a remote computer, for example, t2.micro (~ 1 cent per hour), plus there is a payment for data transfer. After you finish transferring files to the AWS VM, use the PuTTy utility to estimate the amount of virtual disk space in the work / directory using the “du - h” command and make sure that the size of the copied folder matches the original original.



**Figure 8.** WinSCP screen: transferring fast5 files from the local computer or organization's storage system to the work / fast5 directory on the Amazon Web Services cloud platform virtual machine.

***STAGE 2*** *- Running Programs on the AWS Platform*

WinSCP is used to upload fast5 data to the work \ directory on the AWS cloud machine virtual disk. The work \ fast5 \ subdirectory is specified as an input parameter to the guppy\_basecaller program, and the result of the base-calling execution is generated into the guppy\_results subdirectory. Then the data is analyzed by the sequencing quality control program, and then the fastQ format files that have passed the validation and located in the guppy\_results / pass / subdirectory are combined into one bigfile.fastQ file with the “cat” command. This file, in turn, is the input for the procedure for aligning reads on the transcript by the Minmap2 program. The mapping result, in the form of a file aln.sam, is transferred for statistical analysis of the alignment quality by the Samtools stats program, followed by the launch of the program for estimating the number of transcripts by the Salmon quant program.

At the output of the ONT data processing pipeline, a set of directories and files is generated, listed in the results description section (see Table 2 in the article).

Step 4. Base Calling..

Before performing the base-calling procedure, put the virtual computer in the p3.2xlarge configuration (Figure 2-7).

In a terminal window, change to the working directory using the "cd" (change directory) command:$ cd /home/ubuntu/work/

In the work / directory, where a folder with files in the fast5 format was previously prepared, run the guppy\_basecaller program (command C1), or in test mode use the command (C1a), in which the fast5.tutor folder is specified instead of the fast5 folder:

|  |  |  |
| --- | --- | --- |
| **(C1)\*** | :**~/work**$ guppy\_basecaller –i fast5 –s guppy\_results  -qscore\_filtering –flowcell FLOW-MIN106 -kit SQKRNA002  -device auto | \  \ |
| **(C1a)\*** | :**~/work**$ guppy\_basecaller -i fast5.tutor -s guppy\_results  -qscore\_filtering --flowcell FLOW-MIN106 -kit SQKRNA002  -device auto | \  \ |
| \* to copy the text of the command being executed, use Appendix # 2. | | |

Notice the “backslash” (“\”) icons in the command. This designation indicates that the command is a single line, but spread over several lines for ease of presentation in the article. To execute the protocol, copy the command text from application # 1 to the terminal window, then go to the terminal window of the PuTTy program and right-click the "mouse", which is analogous to the Paste function.

Using the command (C1) specifying the -i (input) parameter, set the fast5 subdirectory as the input subdirectory, which contains the previously copied output files received from the nanopore sequencer, and as the subdirectory into which the processed data must be "saved" (the -s parameter, save ), specify the guppy\_results subdirectory (see C1).

ATTENTION! In the standard character system, ASCII (American Standard Code for Information Interchange), there are two similar characters - “minus” (short horizontal bar) and “dash” (long horizontal bar). In the command line of the Unix operating system, only minus signs are used, and when performing the base-calling procedure, the Guppy\_basecaller program displays a “progress indicator” reflecting the percentage of processed data (Fig. 9). After execution, a subdirectory guppy\_results / is created, in the pass / subdirectory of which files of the fastQ format are located, which are necessary for performing further steps (Fig. 10). (For reference: to view the list of files in the guppy\_results / pass subdirectory, you can use the "ls guppy\_results / pass" command).

**Figure 9.** Progress indicator of the base-calling procedure.

After guppy\_basecaller is finished, take the virtual machine out of p3.2xlarge mode and switch to t2.2large (or p3.micro) saving mode, see (Fig. 2-6).



**Figure 10.** After the base-call procedure, the guppy\_results subdirectory is formed.

Step 5 (optional). Assessment of the quality of the results of the base-calling procedure.

Quality assessment (QC, quality control) is performed using the MiniIONQC.R script developed in the R programming language (see command C2). This language was preinstalled during the preparation of the virtual machine.

|  |  |
| --- | --- |
| **(C2)\*** | :**~/work**$ MinIONQC.R \  -i guppy\_results\sequencing\_summary.txt –o qc\_results |

\* to copy the text of the command being executed, use Appendix # 2.

The input to the command (C2) is the sequencing\_summary.txt file, obtained as a result of the guppy\_basecaller program, as the input directory where the quality control results, qc\_results, will be located (Fig. 11).



**Figure 11**. The qc\_results directory where the quality control results are located.

Step 6. Mapping (Minimap2) and Mapping Statistics (samtools stats)

To continue working, you need to combine the set of files resulting from the guppy\_basecaller program into one fastQ file. For this, the standard “cat” command of the Linux / Ubuntu operating system is used (see C3). As a result of the execution of the C3 command, the files located in the guppy\_results / pass / folder will be combined into one file, which we named bigfile.fastQ (Fig. 12). Use the “ls –hl” command to verify that the file has been successfully created and is of the same size as the human genome (several gigabytes, at least). For reference: the human genome occupies about 3 gigabytes.

|  |  |
| --- | --- |
| **(C3)\*** | :**~/work**$ cat guppy\_results/pass/\*.fastq > bigfile.fastQ |

**Figure 12.** Bigfile.fastQ as a result of the “cat” command.

|  |  |
| --- | --- |
| **(C3)\*** | :**~/work**$ cat guppy\_results/pass/\*.fastq > bigfile.fastQ |

\* to copy the text of the command being executed, use Appendix # 2.

While in the work / directory, run the Minimap2 program to map the sequencing results to the reference transcript, as indicated in the command (C3). As the initial parameters for the Minimap2 program, the name of the input file with "reads" in the fastq format and the reference transcript - gencode.v32.transcripts.fa are indicated. The reference transcript is located in the ref folder.

|  |  |
| --- | --- |
| **(C4)\*** | :**~/work**$ minimap2 -ax map-ont -N 100 \ ref/gencode.v32.transcripts.fa bigfile.fastQ > aln.sam |

** Figure 13.** The aln.sam file obtained as a result of mapping the sequencing results to the reference transcript.

|  |  |
| --- | --- |
| **(C4)\*** | :**~/work**$ minimap2 -ax map-ont -N 100 \ ref/gencode.v32.transcripts.fa bigfile.fastQ > aln.sam |

\* to copy the text of the command being executed, use Appendix # 2.

The command (C4) denotes a call to the mapping program (minimap2) and determines the execution parameters for this program (in particular, the –ax parameter indicates that it is necessary to map “reads” obtained using the Oxford nanopore technology, map-ont), then it points to the transcriptome file in FASTA-format (gencode.v32.transcripts.fa), which is located in the work / ref subdirectory. Then the name of the file with sequencing readings bigfile.fastQ, which was obtained as a result of the command (C3), is indicated. Let us remind once again that the \ character in the command syntax means that the line must be uniform, as shown in Appendix # 2.

Step 7 (optional). Mapping results statistics.

The results of genome mapping (the previous step, the minimap2 command) are sent to the aln.sam file (Fig. 13). Check with the command “ls –l” that such a file is created and it is not zero size. Next, run the command (C5), which allows you to get a statistical reference on the results of genome mapping. To do this, use the Samtools program with the “stats” parameter. Aln.sam is specified as an input file, (generated as a result of mapping, see (C4), and the result is placed in the samtools\_stats.txt file (Fig. 14).

|  |  |
| --- | --- |
| **(C5)\*** | :**~/work**$ samtools stats aln.sam > samtools\_stats.txt |

**\* to copy the text of the command being executed, use Appendix # 2.**

 **Figure 14**. The samtools\_stats.txt file containing statistical information about the results of mapping nanopor readings on a human transcript.

Step 8. Quantification of transcripts.

Execute the command (C6) intended to start the salmon program with the basic parameter “quant”. The result of the program execution will be placed in the work / working directory in the salmon\_quant.out folder (Fig. 15). Wait for the command to complete and make sure that the resulting folder was created using the “ls –hl” command.

|  |  |
| --- | --- |
| **(C6)\*** | :**~/work**$ salmon quant -t ref/gencode.v32.transcripts.fa -l U \  -a aln.sam -o salmon\_qnt.out |

**Figure 15.** Result of executing the salmon program in the salmon\_quant.out folder.

Let's consider the command execution parameters (C6), taking into account that the "\" symbol means that the command is executed without line breaks (see Appendix # 1). The –t parameter points to the file containing the human transcript; in this case: ref / gencode.v32.transcripts.fa, where ref / is a subdirectory in the work / directory, where this file is located in FASTA format. The next parameter "-а" indicates a file of SAM (Sequence Alignment Map) format, which was created as a result of the minimap2 program at the stage of command execution (C4). Finally, the "-o" (output) parameter is required to specify the subdirectory where the Salmon output should be placed.

Step 9. Selection for protein-coding genes.

The key object in the salmon\_quant\_out subdirectory is the quant.sf file (Fig. 13), which contains in a tabular form the name of the gene, transcript and the corresponding number of reads, expressed either in absolute or normalized (TPM - transcripts per million) units. The file is more than 2 GB in size, since it contains information about different types of transcripts, including long noncoding RNAs, ribosomal and transport RNAs, etc. Therefore, before uploading the file, it is advisable to reduce its size by selecting transcripts corresponding to protein-coding genes. For this we use the “grep” command (from the work / directory). Executing the “grep” command allows you to select from the resulting file work / salmon\_qnt.out / quant.sf only those lines that relate to protein-coding genes (Fig. 16).

~/work$ grep “protein\_coding” salmon.qnt.out/quant.sf > quant\_protcoding.sf

 **Figure 16.** Executing grep "protein\_coding" command and output file size information quant\_protcoding.sf.

***STAGE 3*** *- Transferring Calculation Results and Shutting Down the Virtual Machine*

Step 10. Copying files.

Use the WinSCP utility to copy files to a local computer or to an organization's storage drive (Figure 5). Not all of them need to be downloaded, for further work it is enough to download the file quant\_proteincoding.sf, the volume of which, when analyzing a human transcriptome using a nano-pore, is only 12 megabytes. (fig. 16).

Step 11. Completion of work.

Using the AWS website, select the MrFirst ONT virtual machine and then use the menu to stop using the “instate state” menu item (Figure 2-6).