



Oct 19, 2022

Nanopore sequencing data analysis using Microsoft Azure cloud computing service

 [PLOS One](#)

 Peer-reviewed method

DOI

<https://dx.doi.org/10.17504/protocols.io.x54v9dj7pg3e/v1>

Linh Truong¹

¹PathWest Laboratory Medicine WA

PLOS ONE Lab Protocols
Tech. support email: plosone@plos.org



Linh Truong

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.x54v9dj7pg3e/v1>

External link: <https://doi.org/10.1371/journal.pone.0278609>

Protocol Citation: Linh Truong 2022. Nanopore sequencing data analysis using Microsoft Azure cloud computing service.
protocols.io <https://dx.doi.org/10.17504/protocols.io.x54v9dj7pg3e/v1>

Manuscript citation:

Truong L, Ayora F, D'Orsogna L, Martinez P, Santis DD (2022) Nanopore sequencing data analysis using Microsoft Azure cloud computing service. PLoS ONE 17(12): e0278609. doi: [10.1371/journal.pone.0278609](https://doi.org/10.1371/journal.pone.0278609)

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: October 10, 2022

Last Modified: October 19, 2022

Protocol Integer ID: 71070

Keywords: raw data from oxford nanopore sequencing, oxford nanopore sequencing, sequencing data analysis, using microsoft azure cloud computing service, microsoft azure cloud server, microsoft azure cloud computing service this protocol, entire data flow from the hospital premise, nanopore, sequencing output, fastq format, hospital site at fiona stanley hospital, raw data in fast5 format, raw data, fiona stanley hospital, resources available in microsoft azure cloud server, pipeline, cloud, entire data flow, fast5 format, data, hospital site, analytic pipeline

Funders Acknowledgements:

Microsoft Australia

Grant ID: Microsoft Partner of the year

Abstract

This protocol provides instruction to set up the analytic pipeline to process raw data from Oxford Nanopore Sequencing. This pipeline leverages the computing resources available in Microsoft Azure cloud server as well as hospital site at Fiona Stanley Hospital. The raw data in FAST5 format would be converted to FASTQ format, demultiplexed, renamed to appropriate sample ID and filtered based on pre-determined quality threshold. The QC plots would also be generated for ongoing monitoring purposes of sequencing output and quality. The entire data flow from the hospital premise to the cloud and vice versa is completely automated and secured.



Section 1: Generation of data on-site

- 1 Load the multiplexed HLA library pool consisting of 48 individuals onto a MinION flow cell. The data is acquired using MinKNOW software for 16 hours using default settings.

16h

Equipment

MinION	NAME
Sequencer	TYPE
Oxford Nanopore Technologies	BRAND
MinION 1B / MinION 1C	SKU

 16:00:00

- 2 The raw FAST5 files are stored in a local folder on the MinION-connected PC.

Equipment

MinION-connected PC	NAME
Computer	TYPE
Dell	BRAND
N/A	SKU
Intel® Core™ i&-7700K CPU @ 4.20Ghz, 32 GB RAM, 64-bit operating system and GPU driver NVIDIA GTX 1080 Ti	SPECIFICATIONS

- 3 An automation agent for Loom Integrate runs on the MinION-connected PC and checks for new FAST5 files every 30 minutes.



Software

Loome Integrate

NAME

BizData Pty Ltd

DEVELOPER

<https://www.loomesoftware.com>

SOURCE LINK

Section 2: Data migration to Microsoft Azure

- 4 The input files are automatically uploaded by the Loome Integrate agent into a container in an Azure blob storage account, deployed within the PathWest Azure subscription. The files are uploaded using Transport Layer Security (TLS), and are encrypted at rest using 256-bit AES encryption.

Command

Command to upload data to Azure using the AzCopy command-line tool (<https://learn.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-v10>).

AzCopy upload

```
azcopy copy <local_folder> <remote_container> --recursive
```

Section 3: Orchestration of analysis pipeline in Microsoft Azure

- 5 The Loome Integrate agent detects that the sequencing job has been completed when it finds a file named "final_summary_<GUID>.txt", and then triggers a new job to deploy the necessary resources and to start the processing steps using the Azure Batch service.



Software

Loome Integrate

NAME

BizData Pty Ltd

DEVELOPER

<https://www.loomesoftware.com>

SOURCE LINK

- 6 Loome communicates with the Azure Batch service and tells it to run the analysis using a Docker container that is automatically pulled by Azure Batch from a private Azure Container Registry in PathWest's Azure subscription.

Software

Azure Batch service

NAME

Microsoft

DEVELOPER

<https://azure.microsoft.com/products/batch>

SOURCE LINK

Section 4: Workflow in the cloud server

7

Azure Batch automatically deploys a GPU-enabled Virtual Machine (VM) for basecalling, de-multiplexing, quality trimming and QC overview using the following commands.

1h 16m 25s

**Command****Guppy basecaller**

```
guppy_basecaller --input_path XX --save_path XX --flowcell FL0-MIN111  
--kit SQK-109 --device cuda:0
```

 01:07:10 (representative runtime)**Command****Guppy barcoder**

```
guppy_barcode --input_path XX --save_path XX --config  
configuration.cfg --device cuda:0 --records_per_fastq 0 --  
trim_barcodes
```

 00:03:06 (representative runtime)**Command****Concatenate & rename file**

```
cd /each_barcode_folder  
cat *.fastq > barcodeXX.fastq
```

 00:00:30 (representative runtime)

Command**NanoFilt**

```
cat barcodeXX.fastq | NanoFilt -q 7 -l 500 > barcodeXX_sampleID.fastq
```

 00:05:39 (representative runtime)

- 8 When each of the VMs was running, the input data is copied into their local disk for faster processing, run the analyses, and then copied the results back into blob storage so that the VMs could be deleted when processing had been completed. Loomer Integrate, in coordination with Azure Batch, orchestrates these steps.

Command

Command to download data from Azure using the AzCopy command-line tool (<https://learn.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-v10>).

AzCopy download

```
azcopy copy <remote_container> <local_folder> --recursive
```

**Command****AzCopy upload**

```
azcopy copy <local_folder> <remote_container> --recursive
```

- 9 Loomer Integrate detects the completion of all tasks in the Azure Batch job and sends an email to notify that the analysis has been successfully completed or to report an error.

Software**Loomer Integrate**

NAME

BizData Pty Ltd

DEVELOPER

<https://www.loomesoftware.com>

SOURCE LINK

Section 5: Data migration from Microsoft Azure server

- 10 If the analysis has been successfully completed, the Loomer Integrate agent downloads the results in FASTQ format into the MinION-connected PC.

Software**Loomer Integrate**

NAME

BizData Pty Ltd

DEVELOPER

<https://www.loomesoftware.com>

SOURCE LINK

Command

AzCopy download

```
azcopy copy <remote_container> <local_folder> --recursive
```

Section 6: Final analysis of results

- 11 The demultiplexed FASTQ file is analysed by a commercial HLA allele assignment software, GenDX NGSengine.

Software

NGSengine

NAME

GenDX

DEVELOPER

https://www.gendx.com/product_line/ngsengine/

SOURCE LINK

The HLA alleles are curated by laboratory staff for accuracy and suitability to reporting.