BRIEF REPORT

REVISED Runcer-Necromancer: a method to rescue data from an interrupted run on MGISEQ-2000 [version 2; peer review: 1 approved, 1 approved with reservations]

Anna Pavlova, Vera Belova , Robert Afasizhev, Irina Bulusheva, Denis Rebrikov, Dmitriy Korostin

Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Pirogov Medical University, Moscow, Russian Federation

v2 First published: 14 Jan 2021, 10:22
<https://doi.org/10.12688/f1000research.27763.1>

Latest published: 14 Feb 2022, 10:22
<https://doi.org/10.12688/f1000research.27763.2>

Abstract

During the sequencing process, problems can occur with any device, including the MGISEQ-2000 (DNBSEQ-G400) platform. We encountered a power outage that resulted in a temporary shutdown of a sequencer in the middle of the run. Since barcode reading in MGISEQ-2000 takes place at the end of the run, it was impossible to use non-demultiplexed raw data. We decided to completely use up the same cartridge with reagents and flow cell loaded with DNB and started a new run in a shortened custom mode. We figured out how the MGISEQ-2000 converts preliminary data in .cal format into .fastq files and wrote a script named "Runcer-Necromancer" for merging .fastq files based on the analysis of their headers (available online: <https://github.com/genomecenter/runcer-necromancer>). Read merging proved to be possible because the MGISEQ-2000 flow cell has a patterned structure and each DNB has invariable coordinates on it, regardless of its position on the flow cell stage. We demonstrated the correctness of data merging by comparing sample analysis results with previously obtained .fastq files for them. Thus, we confirmed that it is possible to restart the device and save both parts of the interrupted run.

Keywords

MGISEQ-2000, DNBSEQ-G400, NGS, Paired-end sequencing, fastq merging

This article is included in the **Cell & Molecular Biology** gateway.

Open Peer Review

Approval Status ✓ ?

	1	2
version 2 (revision) 14 Feb 2022		? view
version 1 14 Jan 2021	✓ view	? view

1. **Simon Andrews** , Babraham Institute, Cambridge, UK
2. **Sergey Knyazev** , University of California, Los Angeles, Los Angeles, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Anna Pavlova (pavlova.a.s@gmail.com), Vera Belova (verusik.belova@gmail.com)

Author roles: **Pavlova A:** Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Original Draft Preparation; **Belova V:** Conceptualization, Methodology, Resources, Writing – Original Draft Preparation; **Afasizhev R:** Formal Analysis, Investigation, Methodology, Software; **Bulusheva I:** Investigation, Methodology, Software, Writing – Original Draft Preparation; **Rebrikov D:** Funding Acquisition, Project Administration; **Korostin D:** Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the Ministry of Science and Higher Education of the Russian Federation allocated to the Center for Precision Genome Editing and Genetic Technologies for Biomedicine [grant №075-15-2019-1789].

Copyright: © 2022 Pavlova A *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Pavlova A, Belova V, Afasizhev R *et al.* **Runcer-Necromancer: a method to rescue data from an interrupted run on MGISEQ-2000 [version 2; peer review: 1 approved, 1 approved with reservations]** F1000Research 2022, 10:22 <https://doi.org/10.12688/f1000research.27763.2>

First published: 14 Jan 2021, 10:22 <https://doi.org/10.12688/f1000research.27763.1>

REVISED Amendments from Version 1

No specific changes to the uploaded data, or affiliation or names. Some explanations are added in the conclusion section accordingly to the reviewer's comment. We pointed out that the proposed script provides merging of reads, but does not guarantee the good quality of the received data. In this case, the researcher must decide whether to restore the data or not, depending on the emergency situation, and be sure to check the data quality for in-lab reference samples.

Any further responses from the reviewers can be found at the end of the article

Introduction

At the end of 2017, Chinese company MGI Tech presented the MGISEQ-2000 sequencing platform¹, promoting it as a device for large and medium scale genome sequencing. MGISEQ is specific in harnessing cPAS sequencing technology and using nanoballs (DNB) generated from circular molecules of DNA library by rolling circle replication². MGISEQ is compatible with a wide range of reagents for sequencing in SE50, SE100, SE400, PE100, PE150, and PE200 modes. MGISEQ-2000 provides the quality of sequencing comparable with that of the Illumina platform³⁻⁶.

The first MGISEQ-2000 sequencer in Russia was installed in our lab (Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Pirogov Medical University) in February 2019, and we run it once a week in the paired-end 150 mode (PE150). According to our experience, one PE150 run usually takes 68 hours if one flow cell is used at a time. During one of these runs, about 23.00 on Saturday, there was a failure of the Moscow power grid leading to a 50-minute blackout of a whole district including Pirogov Medical University. UPS battery storage was sufficient only for 20 extra minutes, then the sequencer turned off until the power was restored. Therefore, the device with loaded reagents remained in sleep mode for 35 hours until Monday. Before the instrument was switched off, it had performed 138 full cycles of forward read sequencing (run 27). The specific feature of the MGISEQ-2000 sequencing program is that it reads a barcode at the end of a run after it completes sequencing of forward and reverse reads. So, to a first approximation, the data obtained could not be demultiplexed as information on the barcodes was absent.

According to the MGI Tech⁷ recommendations, after consulting with the MGI Tech service engineers, we were advised to dispose of the current tank with reagents as well as the flow cell and run the samples using new reagents. In the first place, it is linked to the high sensitivity of the MDA reagent to storage at +4°C as it loses its activity very quickly. We decided to continue the run using the reagents that had been loaded for the weekend and try to restore the data. Finally, we managed to rescue the data using the software from ZebraCall⁸ and our own script on C++, which is reported here <https://github.com/genomecenter/runcer-necromancer>.

Methods

Sequencing

We prepared 3 pools of circularized libraries following the standard MGI Tech protocol⁹. Then we synthesized DNB, loaded a flow cell using the MGIDL-200H manual loader, prepared a sequencing cartridge from the MGISEQ-2000RS High-throughput Sequencing Set with User manual version: A2, and started sequencing on A-side in PE150 mode. Run 27 was aborted at the 139th cycle of the read-1 sequencing phase. After 35 h, we restarted the run (run 27_2) using the same sequencing cartridge and flow cell in a custom mode with the following parameters: read 1 for 12 cycles, read 2 for 151 cycles, Start phase: Sequencing (Figure 1). For summary reports generated by MGISEQ-2000 for lane 1 of runs 27 and 27_2, see *Extended data*: File S1, S2.

fastq generation

The generation of .fastq files containing forward reads for the interrupted run was performed using ZebraCall v2⁸ framework (C:\ZebraCallV2\client.exe – the pathway to software on MGISEQ-2000), which transforms intermediate .cal files into fastq format and demultiplexes them using barcodes.

The appropriate work of ZebraCall requires a .txt file with barcode sequences used for demultiplexing. We created an empty file 'empty_barcode.txt' so that the last 10 nucleotides from 13 nucleotides that were read earlier would not be recognized as barcodes by ZebraCall.

We used the following command (we provide an example for lane 1):

```
client.exe D:\Result\workspace\run_name\L01
139 6 72 -B C:\ZebraCallV2\empty_barcode.txt
-N run_name -U 1 -F
```

It contains the options:

- the access to the folder with .cal files
- run_name — the name of a run
- 139 – the number of completed sequencing cycles
- 6 72 – the number of fields of view counted horizontally and vertically for a corresponding lane
- -B – a path to the file with barcodes
- -U – the number of a lane
- -F – fastq generation without generation of flow cell images

As a result, for each lane, we generated files 'run_name_L0N_read.fq.gz' where N is a lane number. Such file contained a read name and a sequence of 138 nucleotides long.

Figure 1. The screenshot of MGISEQ software in a custom mode with the settings used for restarting the run.

Fastq merging

MGISEQ-2000 employs a patterned flow cell, so each DNB in a cell has unique coordinates at X and Y axes which do not depend on flow cell localization in a device and are not changed if the flow cell is displaced. When the power of the sequencer was off, the vacuum pump was switched off as well. The coordinates of each read were saved in a header of a .fastq file (Figure 2). This allowed us to integrate the data on forward reads obtained before and after the instrument was off.

As a read number being used for forward and reverse reading is unique, we managed to combine the 138-nucleotide sequences obtained during the first run with the nucleotide sequences obtained during the second run based on the information on F.O.V Column, F.O.V Row, and read numbers. To achieve this, we created a C++ script, which can be accessed at GitHub <https://github.com/genomecenter/runcer-necromancer>. The instruction for script running can be found below and in the file README.md in the repository.

Script manual

The script (<http://doi.org/10.5281/zenodo.4316802>¹⁰) is executable on Linux (was tested on Ubuntu 20.04) with GCC compiler with C++17 support and zlib (apt-get install zlib1g-dev). First step is a building: you need to run build.sh script inside the root folder. SaveReads program recovers sample files by placing fixed files into the fixed directory inside current directory. It is important to check that there are no identical filenames between samples files. SaveReads accepts N+1 argument, where first argument is _undecoded.fq.gz (pool of non-demultiplexed reads) file from interrupted run, and next N arguments are

standard samples files. Script SaveReads.py simplifies call to SaveReads. This file accepts pool of non-demultiplexed reads as its single argument. All files with _1.fq.gz ending from current folder will be taken as samples files.

Results

The most important parameter for sequencing quality is the ratio of the data with the quality level of no less than Q30. The Q30 value and other quality metrics did not decrease dramatically in spite of a 35-hour stand by (Figure 3, Table 1).

This implies that storing a loaded cartridge for 35 h leads to its decline, however, it can be still used for sequencing.

To check if merging reads from different runs was correct, we compared 7 samples of whole-exome sequencing from runs 27 and 27_2 with the data from the same samples obtained in run 22. We used the distribution of the size of an insert between left and right reads and the ratio of reads having the insert size exceeding 1000 nucleotides as control metrics. If read merging had been performed with errors, the portion of the reads mapped to various genome regions would have significantly increased. See Figure 4 for the distribution of an insert size for sample LWX777 from the control group.

The ratio of reads from sample LWX777 with the insert size exceeding 1000 nucleotides was 0.003% in case of the data combined from the different runs, while it was 0.005% in case of the previous sequencing without read integration. The obtained data imply that read merging was correct.

Figure 2. The structure of a read header in an MGISEQ-2000 .fastq file.

Figure 3. Q30 histograms for runs 27 (A) and 27_2 (B). The X-axis represents the number of sequencing cycles, the Y-axis represents the ratio of the data with the quality no less than Q30 (%). Blue arrows in histogram B indicate the cycles which reverse reading and barcode reading started.

Table 1. Metrics of 27 and 27_2 run for lane 1 from summary reports (see *Extended data: Files S1, S2*).

Run	27	27_2	Drop
Loaded DNBS (from the first base report)	1169312	1131610	3%
Chip productivity, %	76.82	68.18	11%
Q30, %	90	90.96	0%
Total reads per lane, M	467.3	397.67	15%

Conclusion

It is possible to use a sequence cartridge after 35-hour storage at +4°C, although the quality of the obtained data is reduced.

We would like to point out that the method we propose can be applied in exceptional cases at your own risk and always followed by a quality check of the obtained data. As the procedure described above is not a complete experiment with

Figure 4. The distribution of inserts for sample LWX777 from runs 22 (A) and 27+27_2 (B). The X-axis represents the number of reads, the Y-axis represents an insert size. The diagrams were obtained using Picard CollectInsertSizeMetrics v2.22.4.

replication and control samples, we cannot guarantee the quality of the data after merging reads of sequencing run under different problematic conditions. Therefore, we recommend that researchers in such situations adequately assess the conditions: how long a sequencing run was interrupted (hours or days), how the temperature and humidity in the laboratory room and inside the device changed, at which stage of sequencing (reading forward or reverse read, MDA reaction, reading a barcode) the power was turned off, etc. It is also desirable to use an in-lab reference sample in each run of the instrument to assess data quality and batch-effect. The described aborted run had a DNA library sample that we had previously sequenced under normal conditions, so we were able to assess the quality of the data obtained after merging reads in terms of GC-content, Ti/Tv and het/hom ratio, coverage statistics, variant calling.

Merging sequencing data can be successfully performed if the information about the localisation in flow cells is saved in a read header. The researcher must then compare the data quality on their own reference samples to decide whether to use the data from the aborted run.

Data availability

Underlying data

Raw data for the sample LWX777 from runs 22 and 27_1+27_2 available at Sequence Read Archive (SRA), BioProject ID PRJNA683755: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA683755/>

LWX777_run27_2_united (SRA: [SRS7871577](https://www.ncbi.nlm.nih.gov/sra/SRS7871577)) is an example of reconstructed fastq file from the 2 parts of interrupted run 27. First 139 nucleotides were received from run_name_L04_read.fq.gz fastq file with non-demultiplexed left reads for lane 4. LWX777_run22 (SRA: [SRS7871575](https://www.ncbi.nlm.nih.gov/sra/SRS7871575)) is fastq files from previous run 22.

Extended data

Zenodo: genomecenter/runcer-necromancer: Runcer Necromancer updated release (December 2020), [http://doi.org/10.5281/zenodo.4340350](https://doi.org/10.5281/zenodo.4340350)¹⁰.

This project contains the following extended data:

- File S1. Summary report for run 27 lane 1 from MGISEQ-2000
- File S2. Summary report for run 27_2 lane 1 from MGISEQ-2000

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0).

Software availability

Script available from: <https://github.com/genomecenter/runcer-necromancer>

Archived script as at time of publication: [http://doi.org/10.5281/zenodo.4340350](https://doi.org/10.5281/zenodo.4340350)¹⁰.

License: MIT

References

1. MGI Tech official site. (Accessed on 14 September 2020).
[Reference Source](#)
2. Fehlmann T, Reinheimer S, Geng C, *et al.*: **cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs.** *Clin Epigenetics*. 2016; **8**(1): 123.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Jeon SA, Park JL, Kim JH, *et al.*: **Comparison of the MGISEQ-2000 and Illumina HiSeq 4000 sequencing platforms for RNA sequencing.** *Genomics Inform*. 2019; **17**(3): e32.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Senabouth A, Andersen S, Shi Q, *et al.*: **Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing.** *NAR Genom Bioinform*. 2020; **2**(2): lqaa034.
[Publisher Full Text](#)
5. Korostin D, Kulemin N, Naumov V, *et al.*: **Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing.** *PLoS One*. 2020; **15**(3): e0230301.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Chen J, Li X, Zhong H, *et al.*: **Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers.** *Sci Rep*. 2019; **9**(1): 9345.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. MGI Tech official site. (Accessed on 14 September 2020).
[Reference Source](#)
8. Huang J, Liang X, Xuan Y, *et al.*: **A reference human genome dataset of the BGISEQ-500 sequencer.** *GigaScience*. 2017; **6**(5): 1–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. MGI Tech official site. (Accessed on 14 September 2020).
[Reference Source](#)
10. Pavlova A: **genomecenter/runcer-necromancer: Runcer Necromancer updated release (December 2020) (Version v1.0.1).** *Zenodo*. 2020.
<http://www.doi.org/10.5281/zenodo.4340350>

Open Peer Review

Current Peer Review Status:

Version 2

Reviewer Report 15 February 2022

<https://doi.org/10.5256/f1000research.120655.r123530>

© 2022 Knyazev S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sergey Knyazev

University of California, Los Angeles, Los Angeles, CA, USA

The conclusion states that the researchers were able to assess the quality of recovery by measuring batch-effect including accessing GC-content, Ti/Tv, het/hom ratio, coverage statistics, variant calling. However, I was not able to find support for this statement in the results. In my opinion, adding the methodology of measuring batch-effect coupled with bootstrapping statistics to the results would significantly add to the quality of the research. If it is possible, I would also like to see in the discussion some elaboration on the questions like:
Was batch-effect observed at all after an outage? Does read quality score correlate with batch-effect?

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 22 November 2021

<https://doi.org/10.5256/f1000research.30701.r98664>

© 2021 Knyazev S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sergey Knyazev ¹ University of California, Los Angeles, Los Angeles, CA, USA² University of California, Los Angeles, Los Angeles, CA, USA

The article describes the successful attempt of recovery of the genomic sequencing data after a power outage in the middle of a sequencing experiment. As a result of the outage, the sequencing machine stopped working prematurely that hindered the sequencing experiment. The manufacturer recommended throwing out the results of the entire experiment and starting another run. The article describes the methodology on how to overcome these recommendations and reconstruct results from both parts of the experiment before and after the outage. To reconstruct the experiment before the outage, the article proposes to use already generated files in a raw format that can be extracted from the machine. To reconstruct the experiment after the outage, the article proposes to run another experiment on the same cartridge with the same chemistry that was in the machine during the outage.

While the idea looks interesting, the results and the conclusion are based on misleading assumptions. To prove that the restoration of the experiment was successful, the article tries to use quality control metrics that are designed for normal sequencing runs. The sequencing protocol that the article describes significantly diverges from the standard one because there is an outage in the middle of the experiment and the chemistry may be expired because of the delayed rerun. All that should be described in the article and all the assumptions should be explicitly discussed. Otherwise, the experiment requires rigorous case/control studies with running the sequencing machine in different outage settings with the comparing the results of the experiments with the outage with the experiments without the outage. In this case, at least two experiments are required, for example, one without protocol violation, and one after an outage on the same sequencing library.

Having said that, I recommend the article just simply describe all the limitations of its methodology very explicitly just to keep the audience from potentially misleading conclusions.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

No

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 24 Jan 2022

Vera Belova

We would like to thank Sergey Knyazev for the careful reading of our short report. We appreciate the time and effort that you have dedicated to providing your suggestions. We agree with Sergey's general comment. For sure our short report contains a description of the method of merging reads for the MGI platform that we applied in a force majeure situation and it does not fit into the standard experiment. More explanation has been added in the conclusion section, we encourage readers to evaluate data quality on reference samples in this kind of situation. We hope that you find our response satisfactory.

Competing Interests: No competing interests were disclosed.

Reviewer Report 08 April 2021

<https://doi.org/10.5256/f1000research.30701.r82441>

© 2021 Andrews S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Simon Andrews

¹ Bioinformatics Group, Babraham Institute, Cambridge, UK

² Bioinformatics Group, Babraham Institute, Cambridge, UK

This is a nice example of being able to rescue data from a technically failed run by understanding the processing pipeline well enough to be able to adapt it to your needs. This information will be useful for others who end up in similar situations and may be able to adapt the methods or code presented here to their own situation. I'm somewhat surprised that the indexing of the sequences can be preserved even if the flowcell is moved, and it wasn't entirely clear whether that had actually happened in this run. The results presented make a convincing case that the data was indeed correctly re-paired. I would also have been interested to see the stats on the separation of the barcodes at the end of the run to see whether there was a higher proportion of miscalled bases within those sequences.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, ngs

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research