Understanding Genomic Sequencing Workflows Through Data-Driven Analysis and Pilot Design

Claudia A. Rodríguez Rosario
Master in Manufacturing Competitiveness
Advisor: Miriam Pabón, Ph.D.
Polytechnic University of Puerto Rico
Graduate Project EXPO, May 2026

Abstract *— This study developed and designed a data-driven process improvement pilot for COVID-19 genomic sequencing at the Research Institute of the Puerto Rico Science, Technology & Research Trust, based on workflow mapping and analysis. The workflow includes key stages such as RNA extraction, library preparation, sequencing, and bioinformatics analysis, where variability can propagate across interconnected processes and impact overall performance. A data-driven case study approach was used, combining qualitative and quantitative analysis to evaluate operational performance over a three-month period. Key metrics analyzed included first-pass yield, rework rates, and turnaround time, providing visibility into workflow inefficiencies and performance gaps. Based on this analysis, the study develops a structured framework to identify sources of variability and proposes a data-driven pilot design to evaluate potential process improvements in a controlled environment. This work bridges the gap between process analysis and implementation planning in laboratory environments.*

Keywords *— Bioinformatics, Genomic Sequencing, Next-Generation Sequencing, Process Improvement.*

INTRODUCTION

Advancements and innovations within the genomic sequencing sector have positioned laboratories as strategic pillars for public health surveillance and research. As demand for sequencing increases, laboratories must operate with operational efficiencies. These protocols need to be followed to make research processes reliable and adaptable, especially when dealing with biological variability.

Similarly to most laboratory workflows, inefficiencies do not stem directly from a single step, rather than variability that is carried through interconnected stages. Through sequencing workflows, processes such as RNA extraction, library preparation, sequencing, and bioinformatics analysis are highly or completely-dependent on each other. Variability might be introduced in the process early on and carried out downstream in the process. This leads to rework and increased turnaround time, which will inevitably reduce the overall performance of the workflow.

The Research Institute (RI) of the Puerto Rico Science, Technology & Research Trust is a driver of health innovation in Puerto Rico. Through genome sequencing initiatives that allow disease surveillance and prevention, these activities rely on complex next-generation sequencing (NGS) workflows, which include sample preparation, sequencing, bioinformatic analysis, and quality reporting. All these stages of genomic sequencing processing rely on the interconnectedness and effectiveness of the team within the research production environment.

This project features a real-world genomic sequencing workflow process that includes 480-520 samples per month. Preliminary data received from the laboratory indicates that while all the individual stages demonstrate high first-pass success rates (above 90%), cumulative inefficiencies result in only 70-79% of samples completing the process without rework. Rework increases total turnaround time to over 13 days.

The purpose of this project was to identify sources of variability throughout the genomic sequencing workflow at the Research Institute, to develop a structured proposal of improvements that can be made throughout all phases, supported with a pilot implementation plan that will help staff

evaluate both the feasibility and impact of each recommendation.

LITERATURE REVIEW

One of the most widely recognized and effective approaches for process improvement is the use of structured, data-driven methodologies derived from Six Sigma principles. These approaches provide an organized framework for identifying the root causes of process variation and guiding improvement efforts. Although originally developed for the manufacturing sector, such methodologies have been successfully adapted across various industries, including healthcare systems and other labor-intensive environments.

Specifically, next-generation sequencing (NGS) technologies have caused genomic sequencing to become more effective and innovative. NGS workflows involve several interconnected stages that together allow for DNA and RNA sequencing, including sample preparation, library preparation, sequencing, and bioinformatics analysis [1]. All of these stages happen through extensive and detailed coordination across operational phases.

Laboratory environments, in particular, present unique operational challenges due to increasing demand and process complexity. Taking the Research Institute as an example, research laboratories must manage fluctuating workloads, varying sample sizes, and highly interdependent process stages. Previous studies have identified that many laboratories operate with significant process variation, leading to inefficiencies, delays, and rework [2]. Similarly, process variation in laboratory operations can negatively affect workflow performance and turnaround time [3]. This reinforces the value of structured process-improvement approaches in enabling targeted, effective interventions.

A key challenge that has been faced when implementing such methodologies in research environments is balancing standardization with the flexibility required for scientific procedures. Traditional manufacturing systems require little space for adaptability. Excessive standardization may limit this when demand and sampling change. Literature has consistently supported this perspective. Previous studies have identified that standardization requires a degree of flexibility to remain practical within scientific environments operating across different conditions and workflows [4]. This highlights the reality faced in research laboratories. Being able to adapt to workflows, equipment, sample conditions, and experimental conditions is key. Therefore, effective process improvement in these environments requires integrating structured analytical approaches with sufficient flexibility to support scientific work. In essence, little framework is not good, but too much structure is also not good for the given environment.

In this case, improving internal workflows has broader implications. For organizations like the Research Institute of Puerto Rico Science, Technology & Research Trust, enhancing operational improvement is both an internal objective and a strategic priority. Following this context, applying structured improved methodologies is an opportunity to align internal operational practices to the institutional goal of being recognized as a global center for innovation. For this to happen, efficient and reliable processes must be present to strengthen their capacity to support research needs.

There is still an existing gap in the application of pilot strategies to research laboratory environments. Several available studies focus on identifying inefficiencies or proposing general recommendations, but few translate these findings into actionable and real improvement plans. This study seeks to address this gap by not just analyzing process variability but also proposing targeted improvements through all phases to allow staff to evaluate their feasibility.

METHODOLOGY

This study utilized a case study approach focused on the genomic sequencing workflow at the Research Institute of the Puerto Rico Science, Technology & Research Trust. This is an applied process management project that combines qualitative and quantitative analysis to identify variability in every phase of the

workflow. The specific phases analyzed are RNA extraction, library preparation, sequencing, and bioinformatics analysis- which are the phases identified by the research staff where time elapsed can be varied. The objective was to understand the operational factors that influenced directly its performance and where opportunities for improvement exist. The analysis was conducted using Illumina best practice guidelines and standard operating procedures provided by the laboratory staff, ensuring that performance assessments aligned with established industry benchmarks and manufacturer specifications defined in the Illumina COVIDSeq RUO Kits Reference Guide [5].

Data for this study were collected from the Research Institute genomic sequencing laboratory over a three-month period from October to December 2025, during COVID-19 genomic surveillance operations. The dataset captures key operational metrics such as total samples processed, first-pass yield (FPY) at each stage, rework rates, and processing times. Additional data include the number of samples requiring re-analysis and the average time associated with bioinformatics processing. These comprehensive metrics provided visibility into process performance and enabled quantification of inefficiencies across the workflow, hile also supported competitive benchmarking against industry standards. Performance evaluation referenced Illumina's recommended metrics and quality thresholds to ensure assessments reflect both internal capabilities and adherence to industry-leading practices.

The first step included defining the sequencing workflow was mapped by establishing benchmarks utilizing process boundaries, stakeholders, and performance boundaries. Gaining an understanding about how samples arrive to the Research Institute was key to better understand the unique time frame and scope of the process. This was a foundational step that helped to determine the competitiveness of the institute when compared to other facilities.

As the process was evaluated, the sequencing workflow was clearly mapped, as shown in Figure 1. Data collected over the three-month included total samples processed, first-pass samples at each stage, rework rates, and processing times. Key performance metrics identified include turnaround times, quality metrics, and operational efficiency metrics.

A root cause analysis was conducted using advanced statistical evaluation to identify sources of variability and performance gaps. Factors that were taken into consideration include methods, materials, equipment, staffing, and workflow design. Lean principles were performed to assess performance, identify possible bottlenecks, and highlight phases where rework occurs to gain a clearer understanding of where inefficiencies take place in the workflow.

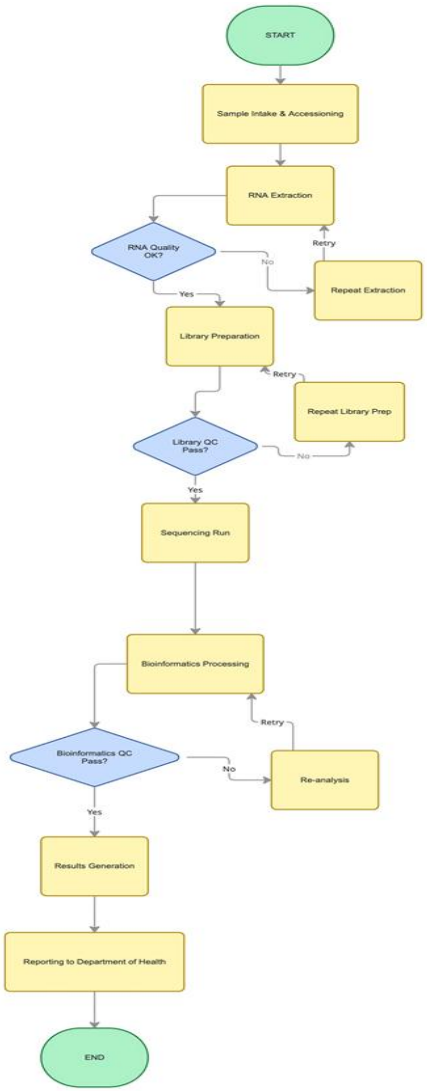

Figure 1
Genomic Sequencing Workflow Flow Chart

Based on the data collected, the study developed a structured set of improvement that targets identified sources of variability. Included recommendations is the optimization of quality control checkpoints, improved coordination between stages with the assistance of laboratory staff, and the standardization of critical steps. Rather than implementing these recommendations and presenting its results at full scale, the study presents a designed data-driven pilot framework to evaluate the selected improvements in a controlled environment. The pilot project included defined variables like duration, targeted workflow stages, and volume. Performance metrics that were taken into account are turnaround time, rework rates, and process consistency rates.

The study advances from analysis to design and planning of targeted interventions based on the identified sources of variability. This approach enables the development of a structured pilot framework to evaluate proposed changes in a controlled environment, without disrupting ongoing laboratory operations. The study also establishes a monitoring structure to support future implementation. This includes defining key performance indicators (KPIs), assigning process ownership roles, and establishing documentation practices to guide performance tracking and ensure sustained evaluation over time. These specific interventions are based on the observed areas of inefficiency throughout the workflow. This is not only useful for laboratory staff to better understand their workflow, but it also allows for external reporting to external players, stakeholders, and public health entities. By improving the consistency and traceability of the workflow, the proposed framework can assist in having more available data to align and boost the institute to surpass industry expectations - which is exactly what the Puerto Rico Science, Technology & Research Trust is vouching for. Overall, this approach provides a scalable, implementation-ready framework that offers a practical foundation for integrating process improvements within complex laboratory environments.

Following the workflow illustrated above in Figure 1, the next-generation genome sequencing for Covid-19 begins with RNA extraction, where samples are prepared for analysis. Once the RNA is extracted, library preparation takes place, a key stage that involves quality control checks to ensure that the samples are apt for genomic sequencing. Quality control procedures during library preparation are essential to maintain sequencing consistency and reduce downstream sequencing variability [6]. Sequencing is performed to generate the genomic data that will be analyzed and reported on- this happens during the bioinformatics phase. Samples that do not pass the quality control checks during the bioinformatics phase, need to be re analyzed- which can introduce potential rework into the system. These quality control check points, are seen in the library preparation and bioinformatics phases. If failure is identified in either, it can represent a direct impact on processing time. To better understand the performance of this workflow, we identified key metrics like processing time, rework rates, and first-pass yields- which helped us assess variability. This analysis was conducted with data collected from the last quarter of 2025 (October - December).

RESULTS

This section presents workflow performance by phase, including end-to-end, processing time, sample flow and yield, and rework distribution across the evaluated stages. Table 1 summarizes in depth the operational metrics collected across the major phases of the genomic sequencing workflow.

Table 1
Detailed Workflow Performance Metrics
October - December 2025

Key Metric	October	November	December	
Samples processed (monthly)	520	500	480	Input (total samples per month
				It can happen because genomic DNA contamination, low RNA yield, degraded RNA, or missing target RNAs
RNA extraction rework (%)	0.08	0.07	0.06	
% passing RNA extraction (1st pass)	0.92	0.93	0.94	The proportion of samples that successfully complete RNA extraction without rework.
RNA extraction rework (samples) (2 days)	42	35	29	The total number of samples that must be re-extracted.
RNA passing 1st pass (samples)	478	465	451	The number of samples that successfully pass RNA extraction on the first attempt.
				The percentage of RNA-passing samples that fail library preparation quality control on the first pass.
Library preparation rework (%)	0.09	0.11	0.08	
% libraries passing QC (1st pass)	0.91	0.89	0.92	The proportion of samples that pass library QC without rework.
Library preparation rework (samples) (2 days)	43	51	36	The number of samples that must be repeated at the library preparation stage.
				The number of libraries that successfully complete wet-lab processing and move to bioinformatics.
Libraries passing QC 1st pass (samples)	435	414	415	
Bioinformatics re-analysis (%)	0.16	0.11	0.09	The percentage of datasets that fail bioinformatics QC and require re-analysis.
% datasets passing bioinformatics QC (1st pass)	0.84	0.89	0.91	The proportion of datasets that pass bioinformatics without needing re-analysis.
Bioinformatics re-analysis (samples)	70	46	37	The number of datasets that require bioinformatics reprocessing.
Datasets passing bioinformatics QC 1st pass (samples)	365	368	378	The number of samples that successfully pass bioinformatics analysis on the first attempt.
Avg bioinformatics processing time (days)	4.8	4.6	4.4	
Avg end-to-end lead time (days)	8.5	8.5	8.5	This represents intake → RNA → library → sequencing when everything passes first time.
	4.8	4.6	4.4	Phase 1 - no rework
	0.16	0.14	0.12	*Rework 1 - RNA extraction*
	0.18	0.22	0.16	Rework 2 - Library preparation
	0.16	0.22	0.09	Rework 3 - Bioinfomatics re-analysis
	13.8	13.68	13.27	End-to-end performance
	520	500	480	Starting samples
	365	368	378	Samples that will be reported on to the Department of Health
	155	132	102	Total samples that did not pass the sequencing process
Total number of samples that pass without rework	70%	74%	79%	

Table 2 reflects the performance of each phase in the genomic sequencing workflow that was identified by the laboratory staff as highly variable. RNA Extraction exhibits relative stable performance throughout the three months presented with an average rework rate of 7%. However, we see a high average of rework in both library preparation and bioinformatics. Rework here is high, with a score of up to 16%.

Table 2
Summary of Workflow Performance Phase
October - December 2025

Phase	Average Rework (%)	First Pass Yield (%)	Interpretation
RNA Extraction	7%	93%	Relative stable performance
Library Preparation	9-11%	89-92%	Quality control failures, primary bottleneck
Bioninformatics	9-16%	84-91(%)	Highest Variability

In Table 3, we have a visual representation of end-to-end time, which measures the total time a workflow takes from start to end. Based on the Illumina guide, the average time the workflow should take it 8.5 days. In practice, the workflow during the fourth quarter of 2025 fell between the 13.3-13.8 days average. This represents an operational delay of over 5 days, which reflects the effect that rework has on overall turnaround time.

Table 3
Summary End-to-End Processing Time
October - December 2025

Metric	Time (days)
Baseline processing time (no rework)	8.5
Additional time due to rework	4.8-5.3
Actual average processing time	13.27-13.80

In Table 4 below, we see the full flow of samples through the workflow. Most importantly, we see an increase in final samples reported to the Department of Health of Puerto Rico from October to December 2025. Additionally, we reported an increase in first-pass yield, which suggests a slow and gradual improvement in performance internally.

Table 4
Summary of Sample Flow and Yield Across Workflow
October - December 2025

Month	Starting Samples	Final Samples (Reported)	Sample Loss	First-Pass Yield (%)
October	520	365	155	70%
November	500	368	132	74%
December	480	378	102	79%

Lastly, Table 5 depicts the specific phases where rework was reported in the workflow. Consistently, Bioinformatics accounts for the highest rework number, with Library preparation following closely. This highlights the concentration of rework being carried downstream through the workflow.

Table 5
Summary of Rework Volume Per Phase
October - December 2025

Phase	October (samples)	November (samples)	December (samples)
RNA Extraction	42	35	29
Library Preparation	43	51	36
Bioinformatics	70	46	37

DISCUSSION

Through these results and table analysis, rework has consistently been highlighted as the primary driver of inefficiency and variability within the next-generation genomic sequencing workflow. More than that, it has shown that rework is a collective of various steps- not just an individual issue. All phases are part of a collective to ensure that the sequencing results are ultimately given to the Puerto Rico Health Department, but quality control outcomes can inhibit this from happening in an efficient manner. Samples are lost, for the most part, during the end stages of the workflow. This ultimately creates an impact on

time and resource utilization. It is key to see this workflow as an interconnected system, where performance is designed by the coordination between each step. Also observed was the handoff between stages, for example, the amount of time each phase of rework takes, can contribute to an uneven or slow performance. Having a clear rework timeframe for each phase is necessary. To address this, there is a need for a more coordinated and standardized operation management. Following the data, improving efficiency is directly correlated on reducing variability. With this data, a targeted improvement framework can be developed for the Research Institute.

PROPOSED PROCESS IMPROVEMENT AND PILOT DESIGN

Following the analysis of workflow variability and rework, this study addresses the identified gap between process analysis and actionable implementation in laboratory environments. While prior studies often identify inefficiencies or propose general recommendations, limited work translates these findings into structured and pilot-ready improvement strategies. In response, this study developed a set of targeted interventions and frames them within a pilot design to enable controlled evaluation of their potential impact.

The pilot project focuses on the workflow stages with the highest variability, which were identified as:

- Library preparation
- Bioinformatics processing

These stages exhibit the highest levels of rework, significantly impacting overall workflow performance and turnaround time.

The proposed framework includes the following key interventions:

1. Defining Standard Operating Procedures (SOPs) to ensure consistency across operators to ease the level of rework on these steps.
2. Standardization of critical process steps to reduce variability propagation.
3. Reinforcement of quality control checkpoints throughout the workflow. This is key, since we know that variability tends to propagate throughout the workflow. If quality is ensured early on, we can minimize delays throughout the end-to-end process.
4. Definition of clear pass/fail thresholds to support consistent quality control decision-making.
5. Process ownership that focuses on assigning a designated staff member (e.g., lab technician or supervisor) responsible for monitoring performance metrics, tracking rework occurrences, and ensuring adherence to defined procedures.
6. A structured pilot implementation approach is proposed to evaluate these interventions in a controlled environment before full-scale deployment.

LIMITATIONS

This study presents a number of limitations that should be considered when interpreting its results and findings. The access to extensive and detailed operational data was constricted due to data privacy considerations. This limited the extensive availability for more in depth analysis. The Research Institute manages various initiatives and projects that cannot be halted, which make the working environment dynamic and ever-evolving. Hence , this why it is important for these recommendations to be put into practice slowly and not at once. Moreover, personal input and observations from laboratory staff were very helpful to provide valuable context for the study. Even with ongoing laboratory activities that can sometimes compete with the time available to dedicate to Covid-19 genomic sequencing, the information collected was sufficient to create a structured and data-backed assessment. These realities reflect and show the real-world context in which the study was conducted.

CONCLUSION

Through this study, Next-Generation Sequencing (NGS) for Covid-19 at the Research Institute of the Puerto Rico Science, Technology & Research Trust was examined. Through this

investigation, key sources of variability and inefficiency were identified through a data-driven approach using the Illumina best practices guidelines as our baseline. The results demonstrate that rework is the primary cause of extend production and processing time. This increases the total baseline goal of 8.5 days to 13.5 days. Variability was mostly found in the steps of library preparation and bioinformatics. The lack of clear pass or fail quality control structures early on in the process, allow for quality to be compromised later on in the workflow.

Through the analysis, the importance of having process consistency in laboratory environments is key. Having a clear coordination between workflow stages can ensure that productivity and efficiency is maintained. Based on the findings stated through this study, a targeted pilot plan was proposed to be implemented in phases to better evaluate their potential impact under a controlled environment. The framework serves as a foundation for future optimization efforts that are being sought by the laboratory staff. Overall, this work shows how applying a structured problem-solving approach can enhance efficiency, reduce variability, and create a stronger and scalable laboratory operation moving forward.

REFERENCES

[1] Illumina. (n. d.). *Next-Generation Sequencing (NGS) | Explore the Technology* [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing.html.

[2] C. Holmes, F. McDonald, M. Jones, V. Ozdemir & J. E. Graham, "Standardization and Omics science: Technical and social dimensions are inseparable and demand symmetrical study," in *OMICS a Journal of Integrative Biology*, vol. 14, no. 3, 2010, pp. 327–332. Available: https://doi.org/10.1089/omi.2010.0022.

[3] A. Harte. (2018, Sept. 18). *Improving Lab Performance with Six Sigma | BSM* [Online]. Available: https://bsmlean.com/blog/andrew-harte/improving-lab-performance-six-sigma.

[4] K. Athanasopoulou, V. Michalopoulou, A. Scorilas & P. G. Adamopoulos, "Integrating Artificial Intelligence in Next-Generation Sequencing: Advances, challenges, and future directions," in *Current Issues in Molecular Biology*, vol. 47, no. 6, 2025, pp. 470. Available: https://doi.org/10.3390/cimb47060470.

[5] Illumina, Inc., *Illumina COVIDSeq RUO kits reference guide (Document No. 1000000126053 v08)*, 2022.

[6] CD Genomics. (n. d.). *Quality control in NGS Library Preparation workflow* [Online]. Available: https://www.cd-genomics.com/resource-ngs-library-preparation-quality-control.html.