

BV-BRC

Bacterial and Viral (BV) -
Bioinformatics Resource Center (BRC)

Monthly System Performance Metrics Report

Performance Period: February 1, 2021 – February 28, 2021

Issued to:

National Institute of Allergy and Infectious Diseases
National Institute of Health

Contract No.: 75N93019C00076

Contract Title: Bioinformatics Resource Centers for Infectious Diseases

Submission Date:

March 10, 2022

Submitted by:

Rick Stevens (contact PI)
Associate Laboratory Director
Argonne National Laboratory
Professor, Computer Science
University of Chicago
5801 South Ellis Avenue
Chicago, IL 60637-5418
630.252.3378 (phone)
630.252.6333 (fax)

BV-BRC System Performance Metrics Report

Notes: As per the recent request from NIAID, we are working with the other BRC to provide jointly agreed plots showing accumulative usage and performance data over time. We will start including them in the monthly reports, starting with the next monthly report.

This monthly system performance metrics report provides a summary of the BV-BRC system performance for the current reporting period in accordance with the Joint-BRC Common System Performance Metrics Plan developed by the BRCs and subsequently approved by NIAID.

As per the plan, each BRC will report and aggregate performance metrics for their constituent parts, *i.e.*, PATRIC and IRD/ViPR for BV-BRC. These metrics will serve as a basis for collecting quantitative measures of performance of the BRC resources to identify trends, areas that are performing well, and areas for improvement. Once the system performance plan is approved by NIAID, each BRC will submit a system performance report for their resource monthly. Annual summaries will be included in the Annual Progress Reports.

It is important to note that metrics across the two BRCs are highly dependent on the relative sizes of the respective research communities, the associated quantities and types of available data, complexity of various analysis tools, and how each of the resources delivers the data and tools to the user. Thus, cross-BRC comparisons of individual metrics are not necessarily indicative of relative usage or performance.

Common system performance metrics covering both BRCs (note that this list is subject to modification, based on feasibility of collection, changes in availability technologies, BRC website development, suggestions from NIAID program and other stakeholders, *etc.*):

Website Performance

Every month, each BRC will report the performance of the key web pages from their website, starting with the pages listed in the table below and adding new pages as they are released on the website. For each page, the average page load time will be computed based on a predefined set of pages and compared against the target page load time set as a benchmark. This will help us ensure that the performance of the individual pages and the overall website is maintained as the amount of data and usage increase with time over the life of the project. If performance of any of the pages is below the set benchmark, we will address it by performing necessary hardware or software optimizations.

- **Target page load time**

- *Definition* - Target page load time measured in seconds, set as a benchmark. The target page load times may vary for various pages depending on their complexity and amount of data they present / visualize to the user.
- *Measurement mechanism* - Manual / custom performance measurement scripts.
- *Measure* - Page load time in seconds.

- **Average page load time**

- *Definition* - Average page load time measured in seconds, after N executions. The average page load times may vary for various pages depending on their complexity and amount of data they present / visualize to the user. Hence, average load time for a web page should be compared only to the benchmark set for that page.
- *Measurement mechanism* - Manual / custom performance measurement scripts.
- *Measure* - Average page load time measured seconds, after N executions.

- Note: For IRD/ViPR, the average page load times reported are representative of both resources, as they both are part of the same single application and database.

Table 1. BV-BRC Website Performance

Web Page	BRC Domain	Target Load Time (Seconds)	Avg. Load Time (Seconds)
PATRIC home page	PATRIC	5	1.56
Global search – genome	PATRIC	5	2.15
Global search – gene	PATRIC	5	1.30
Taxonomy overview	PATRIC	5	3.45
Genome overview	PATRIC	5	5.79
Genome list	PATRIC	5	0.232
Feature list	PATRIC	5	7.57
Feature overview	PATRIC	5	1.55
Genome browser	PATRIC	5	6.01
Circular viewer	PATRIC	5	3.36
Home Page	IRD/ViPR	5	1.42
Genome Search	IRD/ViPR	10	3.38
Protein Search	IRD/ViPR	10	4.08
Strain Search	IRD/ViPR	10	5.26
Epitope Search	IRD/ViPR	10	6.65
Surveillance Search	IRD/ViPR	10	4.3
SFVT Search	IRD/ViPR	10	1.23
Strain Details Page	IRD/ViPR	5	1.56
Protein Details Page	IRD/ViPR	5	2.86
Protein Structure Search	IRD/ViPR	5	4.35
SARS-CoV-2 Variants and Lineages of Concern (S-VLC) Overview	BV-BRC	5	1.22
S-VLC Lineages	BV-BRC	5	0.97
S-VLC Covariants	BV-BRC	5	1.13
S-VLC Variants	BV-BRC	5	1.23
S-VLC Genome Browser	BV-BRC	5	1.33

S-VLC Protein Structure	BV-BRC	5	0.96
S-VLC Phylogenetic Tree	BV-BRC	5	8.76
S-VLC Resources	BV-BRC	5	1.02

Service/Tool Performance

Both BRC analysis services and tools allow users to analyze data pulled from the respective BRC databases and their own private data, compare to other datasets, and save the results in their private workspaces. Both the BRCs will monitor and report the performance of all analysis services/tools available in their resource monthly. The performance reports will be generated based on the actual usage of these services/tools by BRC users in each month. For each analysis service, we will compute the total number of jobs submitted by users, number of jobs completed successfully, failed, average wait time for the jobs queued in the system, and average run time. Monitoring the fraction of jobs that fail and/or reported by users will allow us to identify recurring problems and address them in a timely manner to make the services more robust and reliable. The job wait time depends on the variation in the usage patterns and system load, while the run time depends heavily on the size of the input data and the parameters selected. Monitoring these metrics will allow us to identify factors affecting the overall performance of the application services and tools and address them by performing necessary software and/or hardware scaling or optimization to meet the user expectations.

- **Analysis tasks submitted**
 - *Definition* - A breakdown of total number of analysis tasks submitted by users, summarized by service/tool, during the specified date range.
 - *Measurement mechanism* - Captured via website and server logs, which are used to tally the number.
 - *Measure* - Jobs per month, tallied by service/tool.
- **Analysis tasks completed**
 - *Definition* - A breakdown of total number of analysis tasks submitted by users and completed successfully, summarized by service/tool, during the specified date range.
 - *Measurement mechanism* - Captured via website and server logs, which are used to tally the number.
 - *Measure* - Jobs per month, tallied by service/tool.
- **Analysis tasks failed**
 - *Definition* - A breakdown of total number of analysis tasks submitted by users and failed, summarized by service/tool, during the specified date range.
 - *Measurement mechanism* - Captured via website and server logs, which are used to tally the number.
 - *Measure* - Jobs per month, tallied by service/tool.
- **Average run time by service/tool**
 - *Definition* - A breakdown of average run time for all analysis tasks submitted by users, summarized by service/tool, during the specified date range.
 - *Measurement mechanism* - Captured via website and server logs, which are used to tally the number.
 - *Measure* - Average run time measured in seconds, tallied by service/tool.
- **Input limits**
 - *Definition* - Maximum size of the input supported by a service/tool, beyond which it may degrade the performance or fail to produce results.

- *Measurement mechanism* - Defined by requirements, design and/or testing of a service/tool.
- *Measure* - Input size defined as number or size of the input parameters. The units can vary depending on tool/service.

Table 2. BV-BRC Tools/Services Performance Metrics

Tool/Service	BRC Domain	Jobs Submitted	Jobs Completed¹	Jobs Failed	Avg Run Time	Input Limits²
Codon Tree	PATRIC	533	504	29	14,312	100 genomes
Comprehensive Genome Analysis	PATRIC	2,333	2,186	147	5,863	No limit
Differential Expression	PATRIC	3	1	2	31	No limit
FastqUtils	PATRIC	84	72	12	4,989	No limit
Genome Alignment	PATRIC	147	137	10	464	20 genomes
Genome Annotation	PATRIC	5,660	5,525	135	618	No limit
Genome Assembly	PATRIC	2,576	2,375	201	4,797	No limit
Genome Comparison	PATRIC	275	257	18	582	10 genomes
Metagenome Binning	PATRIC	289	264	25	6,357	No limit
Metagenomic Read Mapping	PATRIC	56	56	0	549	No limit
RNASeq Analysis	PATRIC	46	25	21	14,674	No limit
Taxonomic Classification	PATRIC	368	365	3	1,198	No limit
TnSeq Analysis	PATRIC	17	2	15	2,463	No limit
Variation Analysis	PATRIC	147	133	14	2,087	No limit
Alignment Viewer	IRD/ViPR	22	22	0	35	No limit
Antiviral-Resistance-Risk	IRD/ViPR	6	6	0	3	No limit
BLAST	IRD/ViPR	357	349	8	610	No limit
Enrichment	IRD/ViPR	1	1	0	444	No limit
Genotype-Recombination	IRD/ViPR	9	7	2	8,106	No limit
H1-Clade Classifier	IRD/ViPR	89	89	0	32	No limit
H1N1-classifier	IRD/ViPR	8	8	0	12	No limit
H5N1-classifier	IRD/ViPR	58	56	2	1,062	No limit
Ha Numbering	IRD/ViPR	115	115	0	658	No limit

MGC	IRD/ViPR	20	17	3	5,179	No limit
MSA	IRD/ViPR	595	490	105	6,600	No limit
Mutation-analysis	IRD/ViPR	58	58	0	64	No limit
Primer3	IRD/ViPR	49	49	0	1	No limit
Read-seq	IRD/ViPR	24	23	1	1,060	No limit
Rva Genotyper	IRD/ViPR	1,203	1,197	6	466	No limit
Short-seqsearch	IRD/ViPR	28	21	7	9,694	No limit
SNP-analysis	IRD/ViPR	621	429	192	9,711	No limit
Surveillance-data-mapping	IRD/ViPR	3	3	0	5	No limit
Tbl-formatter	IRD/ViPR	5	1	4	19,183	No limit
Tree	IRD/ViPR	482	446	36	3,964	No limit
VIGOR Annotator	IRD/ViPR	92	91	1	1,616	No limit
SARS-2 Genome Assembly and Annotation	BV-BRC	39	34	5	849	No limit

1. Jobs completed will not equal submitted minus failed if the user canceled the job before completion.
2. These services do not have an imposed hard limit on the input. Job completion time is however proportional to the input size, and it may fail or hang if it is too large in some cases.

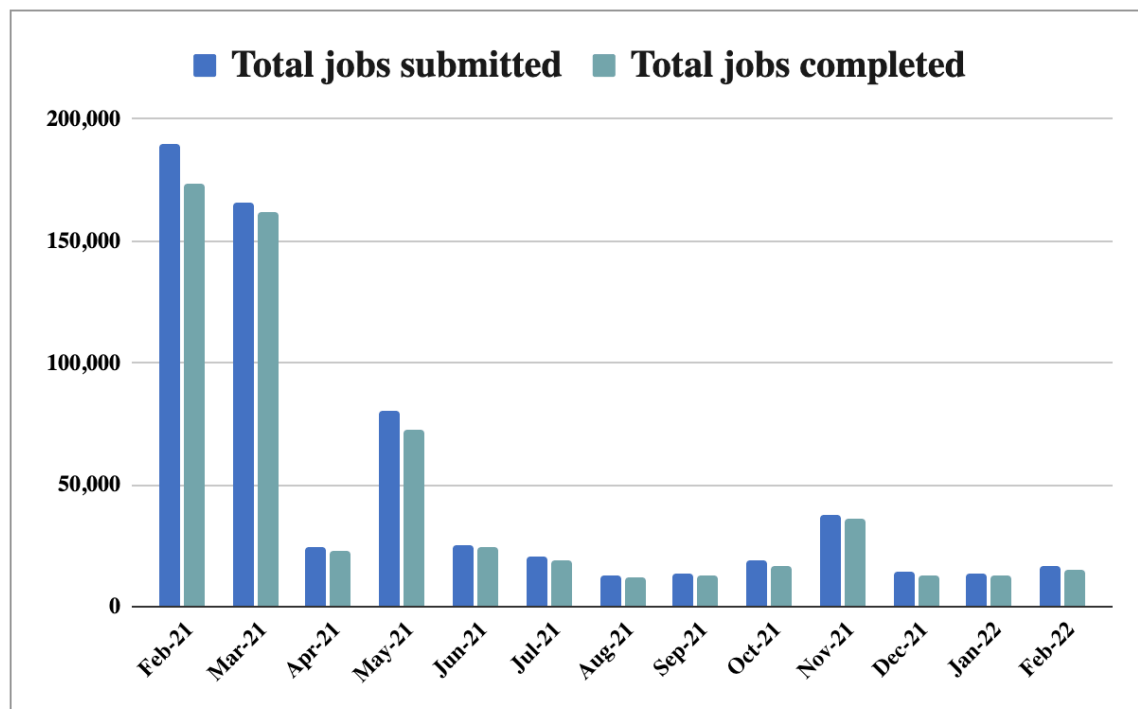


Figure 1. Analysis service jobs submitted vs. completed. February, March, and May had significantly higher job submissions stemming from collaborative SARS-CoV-2 sequence analysis.

Monitoring and resolving job failures:

The BV-BRC team monitors jobs queue daily for the number of jobs submitted, queued, and failed and addresses any issues as they occur. The primary causes of failed jobs and actions taken to address them are as follows:

1. **System errors:** Typically, if there are large number of failed jobs over a short period of time, it is due to a system-level problem, such as one or more nodes in the scheduler are temporarily unavailable, out of disk space or hung. We resolve such issues as quickly as possible, automatically resubmit failed user jobs, and notify the users who submitted the jobs.
2. **User errors:** Another common reason for failed jobs is incorrect or mal-formatted input data or parameters provided by the user. We address such errors by informing users about the correct input formats by pointing them to the relevant documentation and/or examples or working with them to resolve the issue. We identify any recurring patterns and update user documentation and/or job submission forms as needed to avoid such errors in future.
3. **Bug in the application:** Sometimes job failure can result from an unknown bug in a specific application, triggered by the combination of the format or size of the user input data and/or parameters provided by the user. We address this by identifying the cause of failure and then fixing it as soon as possible.

In future, we will establish limits on failure rate that will trigger further investigation of the root causes of the problem and potential solutions to address the and include them in the monthly report.

Database / Data API Performance

Both the BRCs will monitor database performance using predefined search and retrieval queries for various data types, measure average response time in seconds, and report it on a monthly basis. These database queries will capture the most common data queries used by various web pages and tools on the BRC websites as well as user queries used to download large amounts of data in batch mode using the data API, web services, or Command Line Interface (CLI). For each query, the average response time will be compared to the set benchmark. This will help us ensure that the performance of individual data queries as well as the overall database meets the performance benchmarks as well as user expectations. If the performance of any query does not meet the benchmark, we will address it by performing necessary database, query, or hardware optimizations.

- **Target response time**

- *Definition* - Target response time measured in seconds, set as a benchmark. The target response times may vary for various queries depending on the complexity of the query and amount of data retrieved.
- *Measurement mechanism* - Manual / custom performance measurement scripts.
- *Measure* - Page load time in seconds.

- **Average response time**

- *Definition* - Average response time measured in seconds, after N executions. The average response times may vary for various pages depending on the complexity of the query and amount of data retrieved. Hence, average load time for a web page should be compared only to the benchmark set for that page.
- *Measurement mechanism* - Manual / custom performance measurement scripts.
- *Measure* - Average response time measured seconds, after N executions.
- *Note: For IRD/ViPR, the average response times reported are representative of both resources, as they both are part of the same single application and database.*

Table 3. BV-BRC Database / Data API Performance

Database Query	BRC Domain	Target Response Time (Seconds)	Avg Response Time (Seconds)
Genome	PATRIC	5	0.05
Genome sequence	PATRIC	5	1.45
Genome feature	PATRIC	5	0.59
Genome AMR	PATRIC	5	0.04
Feature sequence	PATRIC	5	0.16
Specialty gene	PATRIC	5	0.03
Pathway	PATRIC	5	1.18
Subsystem	PATRIC	5	1.17
Taxonomy	PATRIC	5	0.03
Transcriptomics Experiment	PATRIC	5	0.03
Transcriptomics Sample	PATRIC	5	0.09
Transcriptomics Genes	PATRIC	5	1.51
PPI	PATRIC	5	0.67
Sequence & Strain	IRD/ViPR	5	0.05
Animal Surveillance	IRD/ViPR	5	0.001
3D Protein Structures	IRD/ViPR	5	0.01
Human Surveillance	IRD/ViPR	5	0.01
Serology Experiments	IRD/ViPR	5	0.01
Antiviral Drugs	IRD/ViPR	5	0.02
Host Factor Data	IRD/ViPR	5	0.02
Immune Epitopes	IRD/ViPR	5	0.02
Phenotypes	IRD/ViPR	5	0.02
PCR Primer Probe Data	IRD/ViPR	5	0.001
Sequence Feature Variant Types	IRD/ViPR	5	0.01
Human Clinical Studies and Lab	IRD/ViPR	5	0.001
Experiments	IRD/ViPR	5	0.01
ORFeome Plasmid Data	IRD/ViPR	5	0.01

SARS-CoV-2 Lineages	BV-BRC	5	0.12
SARS-CoV-2 Variants	BV-BRC	5	0.14