VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

VEuPathDB BRC contract HHSN75N93019C00077

Performance Metrics Report

Reporting Period: July 1-31, 2021
Submission Date: August 10, 2021

Notes & Change Log

Date	Version/release	Description & Notes
8/10/2021	1	VEuPathDB Performance Metrics for July 2021 ● Response to COR feedback - Updated Service/Tool Performance -> **Analysis tasks failed** -> *Measurement mechanism* to specify how job failures are monitored (page 4) and added a note to Table 2. ● In response to NIAID's request, we are working with our sister BRC to provide jointly agreed plots showing accumulative metrics over time. These will be included in the next reporting period.

Joint-BRC Common System Performance Metrics Plan

This report will be made available from all VEuPathDB sites, *e.g.,* https://veupathdb.org/ , from the About menu.

This monthly systems performance metrics report provides a summary of the VEuPathDB BRC 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.,* FungiDB, PlasmoDB, OrthoMCL-DB, VectorBase, *etc.* for VEuPathDB. 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 on a monthly basis. 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 target 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 run on all project sites (VEuPathDB.org + all component sites) except for Gene Record Pages which can only be run on the component sites.
  - *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 run on all project sites (VEuPathDB.org + all component sites) except for Gene Record Pages which can only be run on the component sites.
- *Measure* - Average page load time measured seconds, after N executions.

Table 1 VEuPathDB Website Performance (July 1-31, 2021)

Web Page	BRC Domain	Target Load Time (Seconds)	Avg. Load Time (Seconds)
Home page	VEuPathDB	5	4.81
Gene search form with filterParam	VEuPathDB	5	4.22
Gene search result (default organism)	VEuPathDB	5	4.77
Gene record page	VEuPathDB	5	2.97
Site search result	VEuPathDB	5	3.57
Organism table (search result on strategy panel)	VEuPathDB	5	2.86
Data Sets table (search result on answerController)	VEuPathDB	5	4.24
Fasta SRT result (click submit)	VEuPathDB	5	1.71

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 on a monthly basis. The performance reports will be generated based on the actual usage of these services/tools by BRC users in a given 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 across all project sites.
- *Measure* - Jobs per month, tallied by service/tool.

- **Analysis tasks completed**
  - *Definition* - A breakdown of the 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 across all project sites.
  - *Measure* - Jobs per month, tallied by service/tool.

- **Analysis Tasks Deleted**
  - *Definition* - A breakdown of total number of analysis tasks submitted by users and deleted, summarized by service/tool, during the specified date range.
  - *Measurement mechanism* - Captured via website and server logs, which are used to tally the number across all project sites.
  - *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 across all project sites. We monitor for significant change compared to previous reporting periods. We also rely on real-time user feedback to alert us to issues. We expect some number of failures because of user input error, and this may vary each month depending on usage. For Galaxy jobs we receive monthly error reports from Globus and review these to understand reasons for job failures. If the error logs indicate issues with software, Globus is asked to address the problem. For user input errors our Outreach team is informed so that training materials can be updated if needed.
  - *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 across all project sites.
  - *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.*
  - *N/A* - We are not aware of any limits. If there are limits, they will be imposed as part of the standard Galaxy implementation outside our control.

Table 2. VEuPathDB Tools/Services Performance Metrics July 1-31, 2021)

**Note - Inspection of Globus error logs indicates two sources of job failures in this reporting period.*

1. *There is an issue with the 'deepTools BAM coverage' executable causing all jobs to fail (as long as they were not deleted by the user before failure). This has been reported to Globus to be addressed and there is an alternative tool, 'Bam to BigWig' that can provide the same service.*
2. *All other job failures are due to user input error.*

Tool/Service	BRC Domain	Jobs Submitted	Jobs Completed	Jobs Deleted	Jobs Failed	Avg Run Time (sec)	Input limits
BLAST	VEuPathDB	11573	11493	N/A	80	5.3	31kb
Galaxy Jobs - Details below:	VEuPathDB						
FastQC	VEuPathDB	513	499	13	1	1620	N/A
Data Upload	VEuPathDB	414	395	11	8	17	N/A
FASTQ Groomer	VEuPathDB	498	459	18	21	9384	N/A
Trimmomatic	VEuPathDB	318	240	48	30	2529	N/A
HTSeqCountToTPM	VEuPathDB	372	195	89	88	6009	N/A
BAM to BigWig	VEuPathDB	223	178	36	9	2120	N/A
HISAT2	VEuPathDB	319	238	79	2	2180	N/A
OrthoMCL Blast Parser	VEuPathDB	125	77	34	14	12	N/A
OrthoMCL Preprocess Fasta File	VEuPathDB	76	67	2	7	11	N/A
OrthoMCL Map Proteome To Groups	VEuPathDB	67	38	22	7	11	N/A
MCL Clustering	VEuPathDB	53	25	28	0	2	N/A
Tophat2	VEuPathDB	0	0	0	0	0	N/A
Deeptools BAM Coverage	VEuPathDB	46	0	4	42	2416	N/A

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 run on https://plasmodb.org/plasmo/app as a reliable indicator of performance on all project websites.
- *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 run on https://plasmodb.org/plasmo/app as a reliable indicator of performance on all websites.
  - *Measure* - Average response time measured seconds, after N executions.

Table 3 VEuPathDB Database / Data API Performance (July 1-31, 2021)

Database Query	BRC Domain	Target Response Time (milliseconds)	Avg Response Time (milliseconds)
Data analysis searches (breakdown below):	VEuPathDB	NA	NA
Epigenomics	VEuPathDB	1000	55
Function prediction	VEuPathDB	1000	45
Gene models	VEuPathDB	1000	3525
Genetic variation	VEuPathDB	1000	104
Genomic Location	VEuPathDB	1000	35
Immunology	VEuPathDB	1000	46
Orthology and synteny	VEuPathDB	1000	141
Pathways and interactions	VEuPathDB	1000	86
Phenotype	VEuPathDB	1000	355
Protein features and properties	VEuPathDB	1000	88
Protein targeting and localization	VEuPathDB	1000	144
Proteomics	VEuPathDB	1000	188
Sequence analysis	VEuPathDB	1000	247
Structure analysis	VEuPathDB	1000	60
Taxonomy	VEuPathDB	1000	48
Text	VEuPathDB	1000	48
Transcriptomics	VEuPathDB	1000	740
Popset Isolate Sequences	VEuPathDB	1000	66

Genomic Sequences	VEuPathDB	1000	31
Genomic Segments	VEuPathDB	1000	28
SNPs	VEuPathDB	1000	72
ESTs	VEuPathDB	1000	31
Metabolic Pathways	VEuPathDB	1000	29
Compounds	VEuPathDB	1000	44
Sequence retrieval tool	VEuPathDB	1000	561
Site Search	VEuPathDB	1000	472
User Comments	VEuPathDB	1000	56
Multiple sequence alignment (isolates)	VEuPathDB	10000	5364