Supplementary Note 5: SCiAp job execution metrics

Pablo Moreno

1 Introduction

In this document we show execution metrics for different tool sets, as executed in real jobs since late 2019 in the EBI LSF Cluster Galaxy 19.05 instance (EBI LSF) and the Human Cell Atlas (@usegalaxy.eu) Galaxy instance (HCA) running at the DEN.bi cluster.

EBI LSF data is captured by the Platform LSF monitoring system; HCA metric data is captured through cgroups. Both clusters probably use different linux distributions, have different swap/real memory balancing strategies, different CPU clock speeds, different numbers of cores per processor, etc. These reasons, besides the fact that the loads where different as well, lead to different memory/CPU values shown in the coming plots. This re-affirms the point that the convenience of SCiAp is not about the performance, which will vary from setup to setup for so many reasons, but the reproducible access to tools in a non-programmatic manner and its ability to run them in different hardware settings, increasing in orders of magnitude the critical mass of people who can run these scRNA-Seq analysis beyond trained bioinformaticians only.

The metrics shown are:

- Memory usage: the maximum amount of memory in Gigabytes used by the job/process.
- CPU Time: the amount of time that one or more CPUs spent on the job/process. Note that for multi-process jobs, this time is the sum of all CPU time used, and can be hence longer than the time perceived by the user.
- Wall time: the clock time it takes the job to complete in the system where it is running. This is what is more closely perceived by the user, usually called Wall time or Wall clock time, as it is what the user sees in the clock mounted on the wall while running the job.

EBI LSF includes jobs from all users, which includes production pipelines, training and the group's exploratory analysis. HCA includes jobs from 2 users, one of them running training material preparation jobs and the other executing re-analysis of the datasets from the Sanger CoVid-19 Cell Atlas.

The amount of data varies by tool. Scanpy is heavily used by the Single Cell Atlas production pipeline since mid 2019, and so is the tool set that has the highest number of executions by far in the EBI LSF instance. For some tools like SC3 and Scater, which were part of early developments there is not much data because most of the runs happened before the establishment of current internal EBI LSF instance (late 2019) and the HCA instance (~August 2019), in a previous internal instance which has been decomissioned.

Given that the largest input is an important variable that will impact on the CPU and memory usage of each job, the section for each toolset starts by showing the distribution of the largest input sizes for all jobs of that tool, through an histogram. Then it explores the memory usage, CPU time and walltime of the executions of the tool, either through violin plots or jitter plots based on the number of observations (for too many observations, violin plots are used). These violin plots/jitter plots are faceted per sub functionality of each toolset, and the jitter plots will color code the input size in Gigabytes. For the case of violin plots (mostly Scanpy), that cannot easily show the input size variable, additional scatter plots are added right after to show the relation between memory/CPU time/walltime to the largest input size.

12 data points (CPU time and memory usage for 6 jobs) had to be imputed for the HCA instance for Seurat (imputed using similar input sizes jobs for the same tools), as they where clearly metric recollection errors (all of those recorded around the same dates).

1.1 Metrics per tool

The size of the main input will invariably play a role in execution and memory consumption, so we make it explicit here. The EBI LSF data part includes a total of 58992 jobs; the HCA data part includes a total of 3078

This is the distribution of input sizes for the datasets of all jobs.

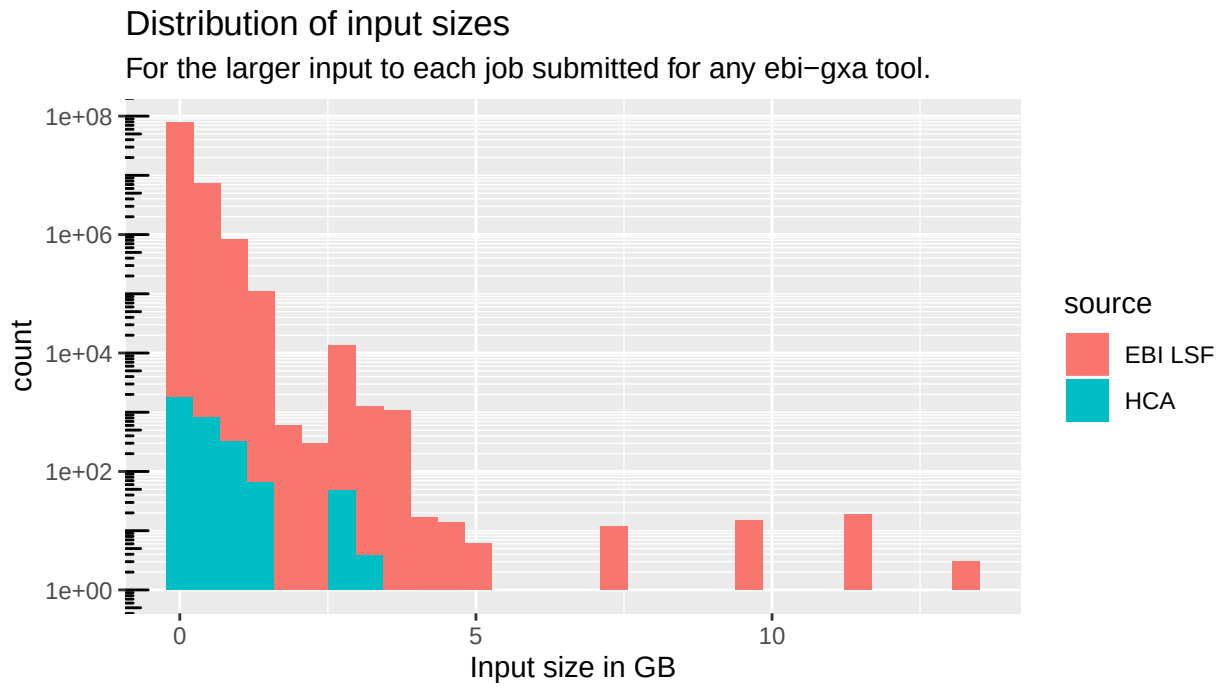

Figure 1: Largest input sizes distribution for all jobs associated to ebi-gxa tools in both the HCA and the EBI LSF galaxy instances. Given the sharp concentration of smaller datasets, log scale is used to be able to convey the counts of datasets at higher sizes as well. A dataset is counted here as each that it is used as input for a tool in a job, which means that each dataset can be counted more than once. The relevance is places on the job executions, which is the main matter of interest. The sizes here should be considered in the context of spare count matrices. For a reference, 75% of the sparse count matrices at Single Cell Expression Atlas (currently 175 datasets closing into 4 Million cells) are below the 450 MB, and the largest one being ~15 GB.

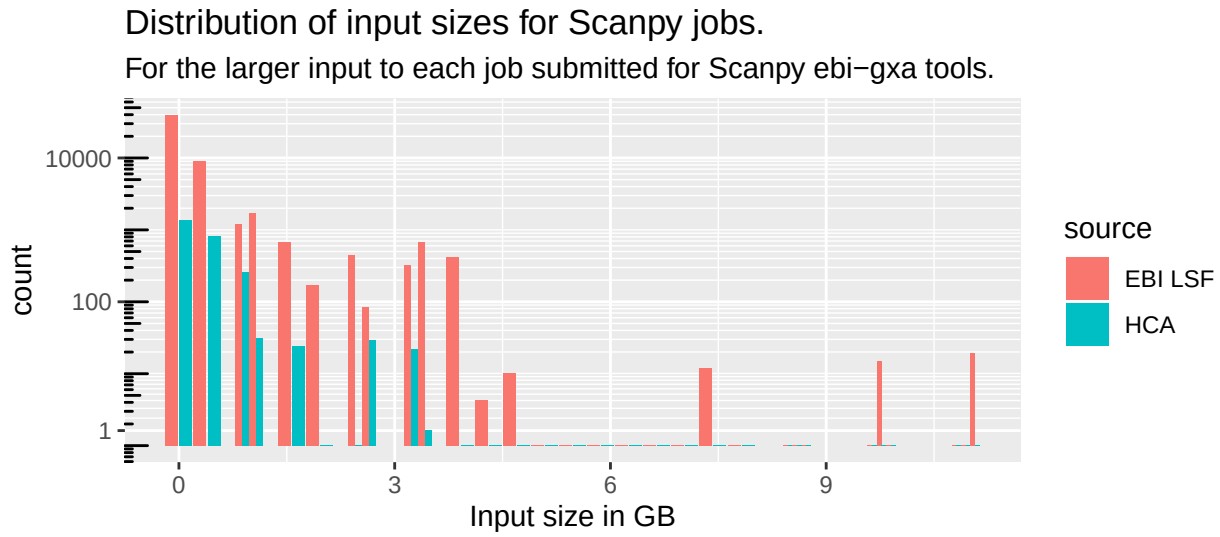

Figure 2: Largest input sizes distribution for Scanpy jobs associated to ebi-gxa tools in both the HCA and the EBI LSF galaxy instances.

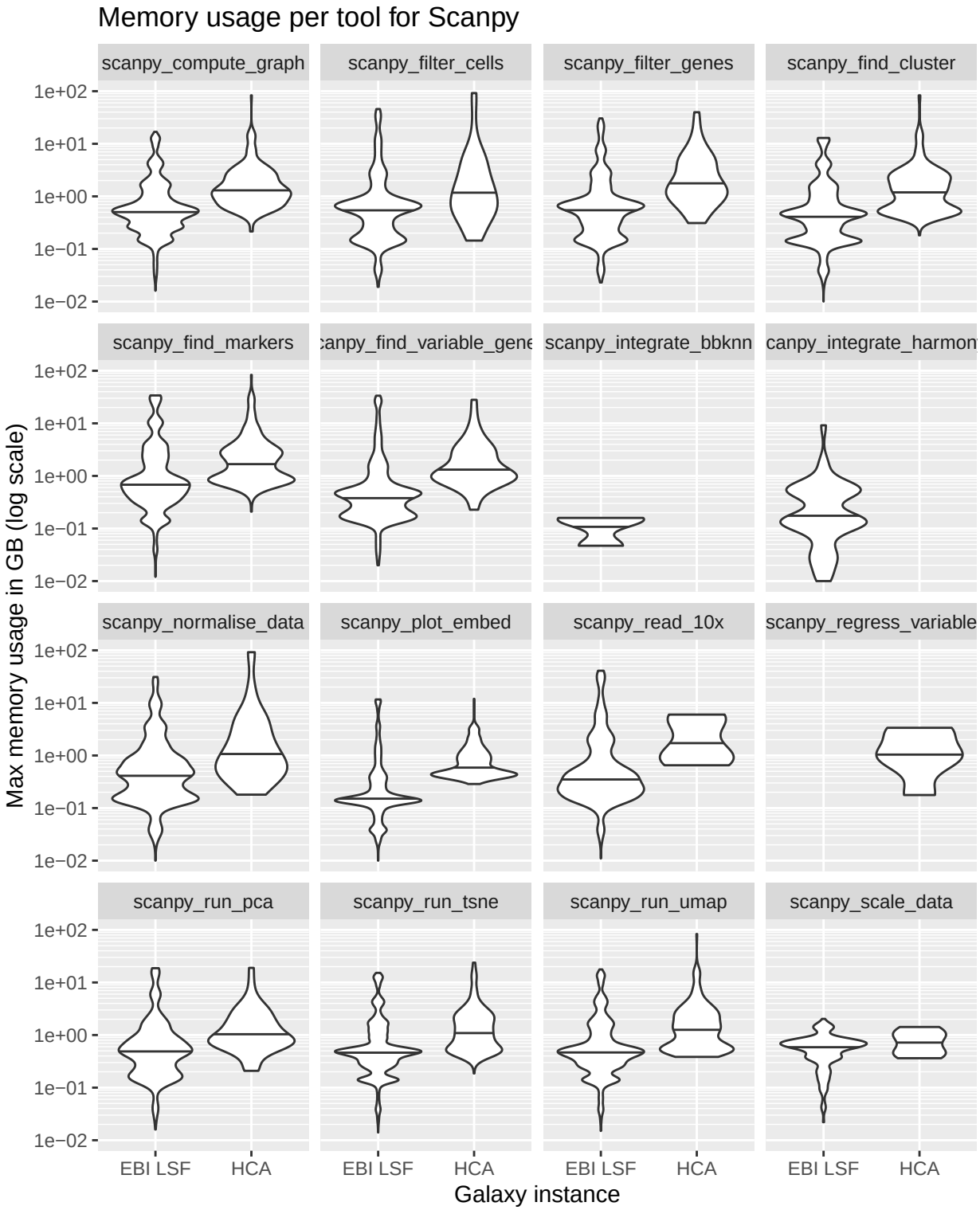

Figure 3: Memory consumption for Scanpy tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). The plot only considers real memory, not swap. Most of the runs for HCE are for the re-analysing the CoVid-19 Sanger Cell Atlas data (~30 different scRNA-Seq datasets); Most of the runs for EBI-LSF are analysis of datasets in EBI Single Cell Expression Atlas

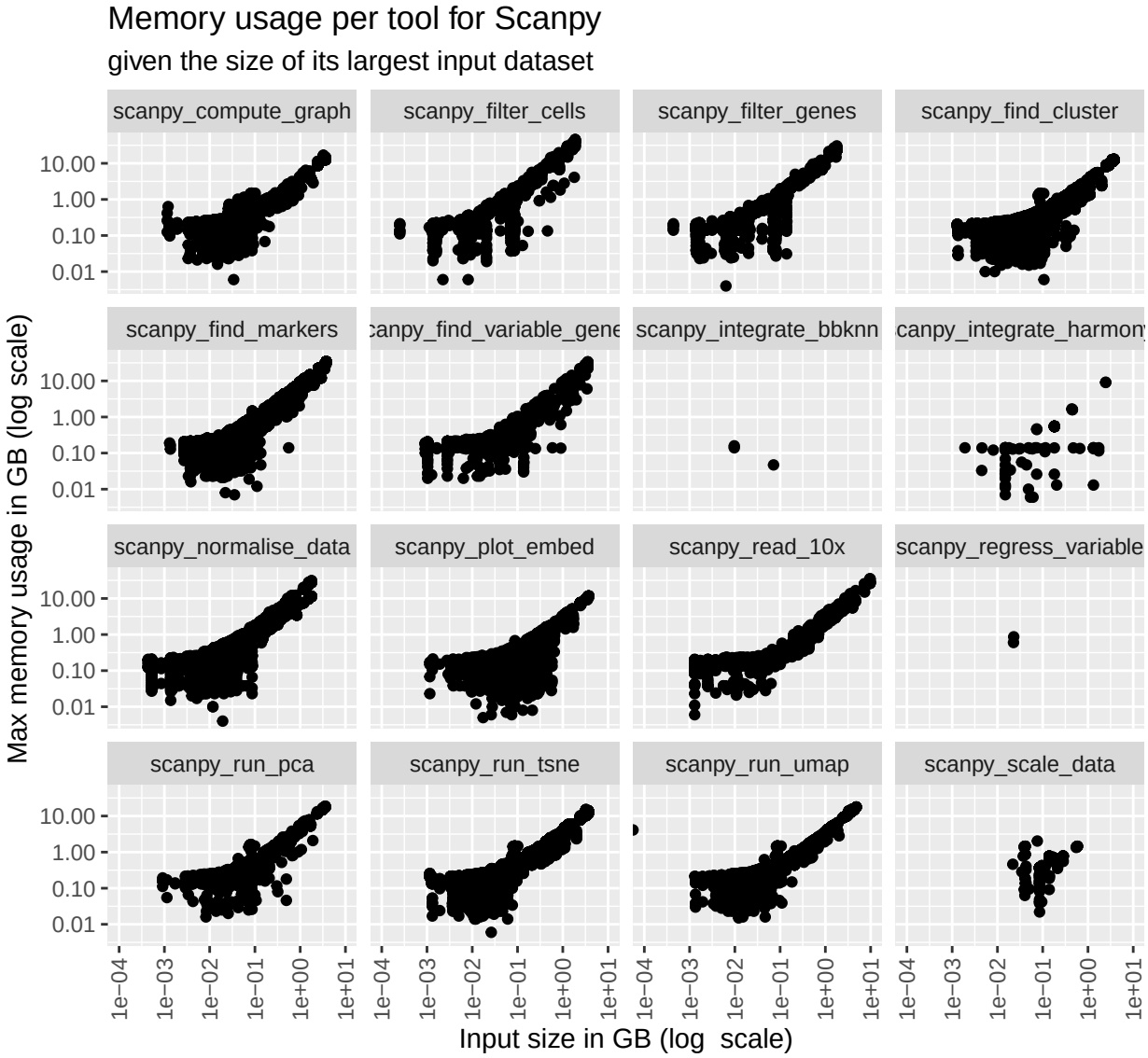

Figure 4: The EBI LSF dataset shows the relation between largest input size and max memory used by the Scanpy tool set jobs.

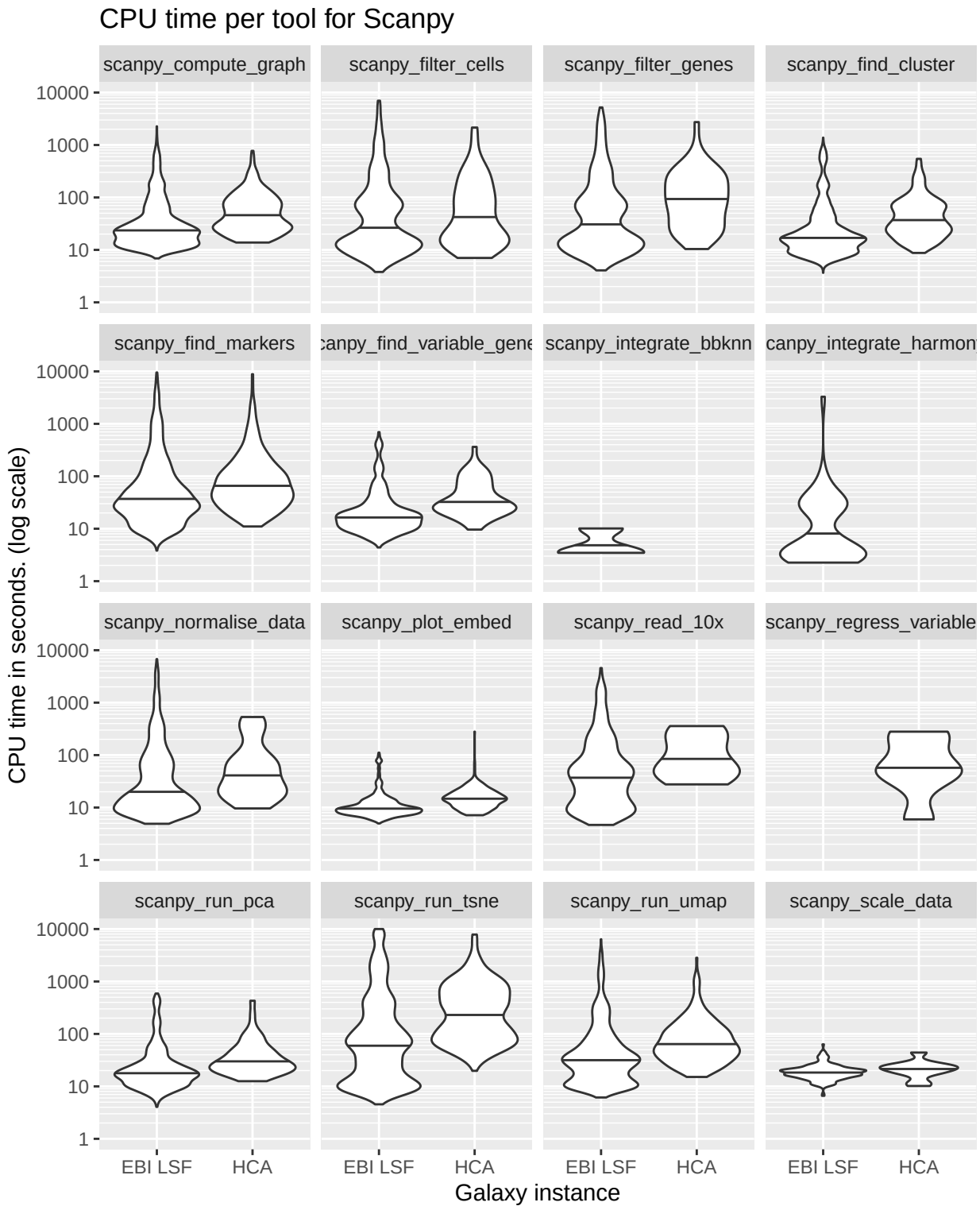

Figure 5: CPU time for Scanpy tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). Consider that a job that uses multiple CPUs at the same time will have the time used for each CPU added. This plot reflects the processing power needed by the tool, not the time that the user waits for.

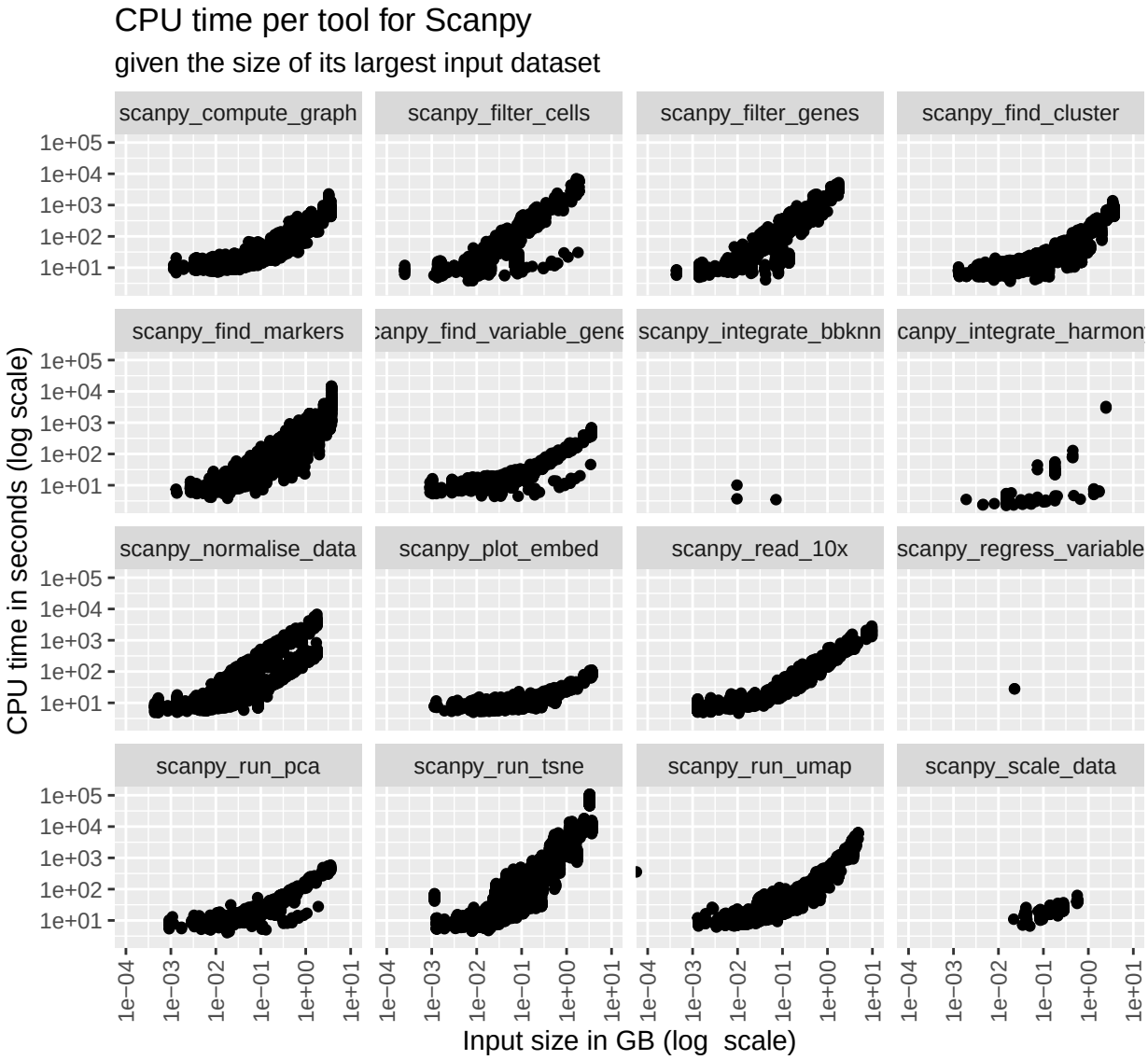

Figure 6: The EBI LSF dataset shows the relation between largest input size and CPU time used by the Scanpy tool set jobs.

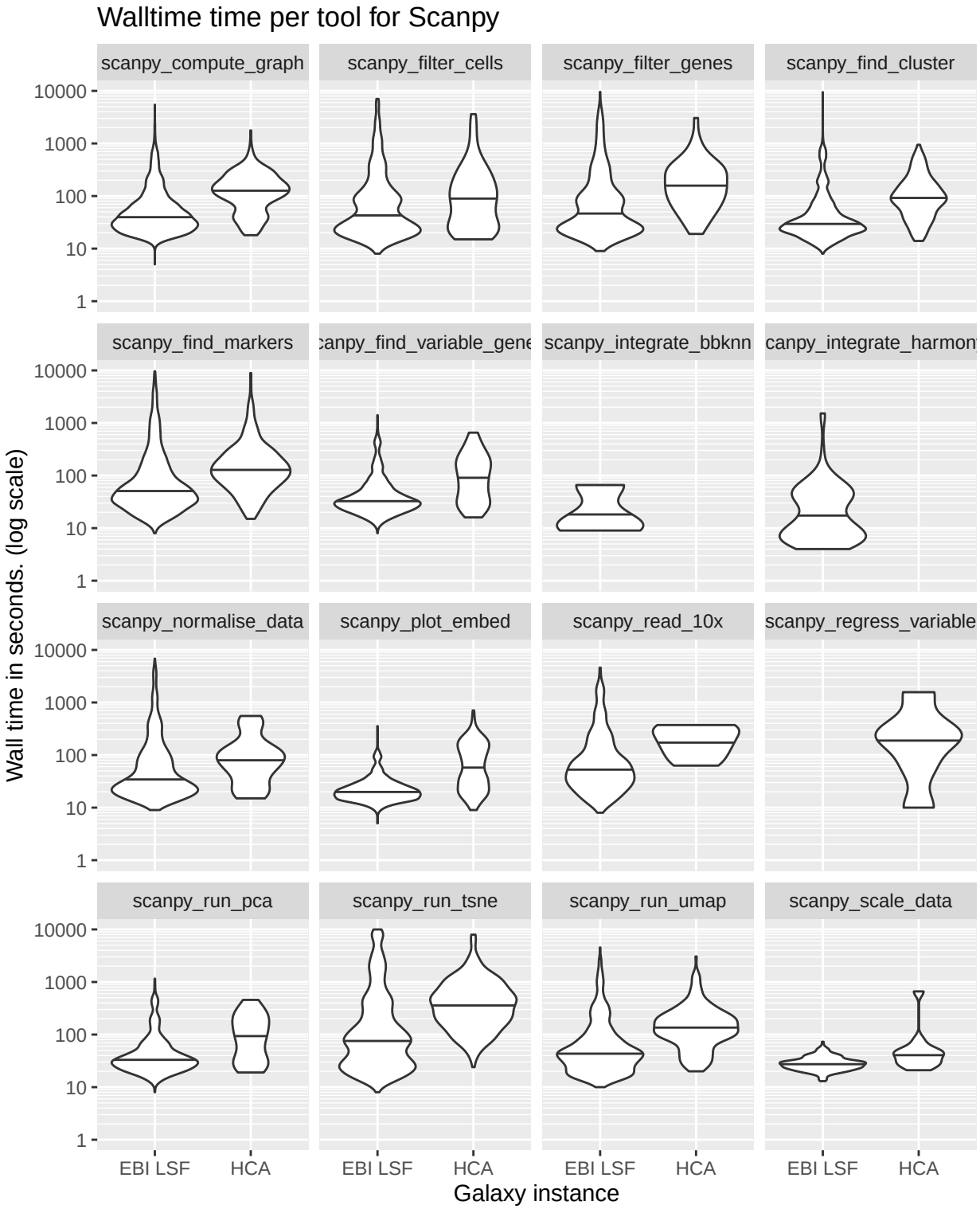

Figure 7: Wall time for Scanpy tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). This metric indicates the time that the user waits for the job, and includes time used in processing (serial CPU time), IO waits, file copies, metadata collection, etc.

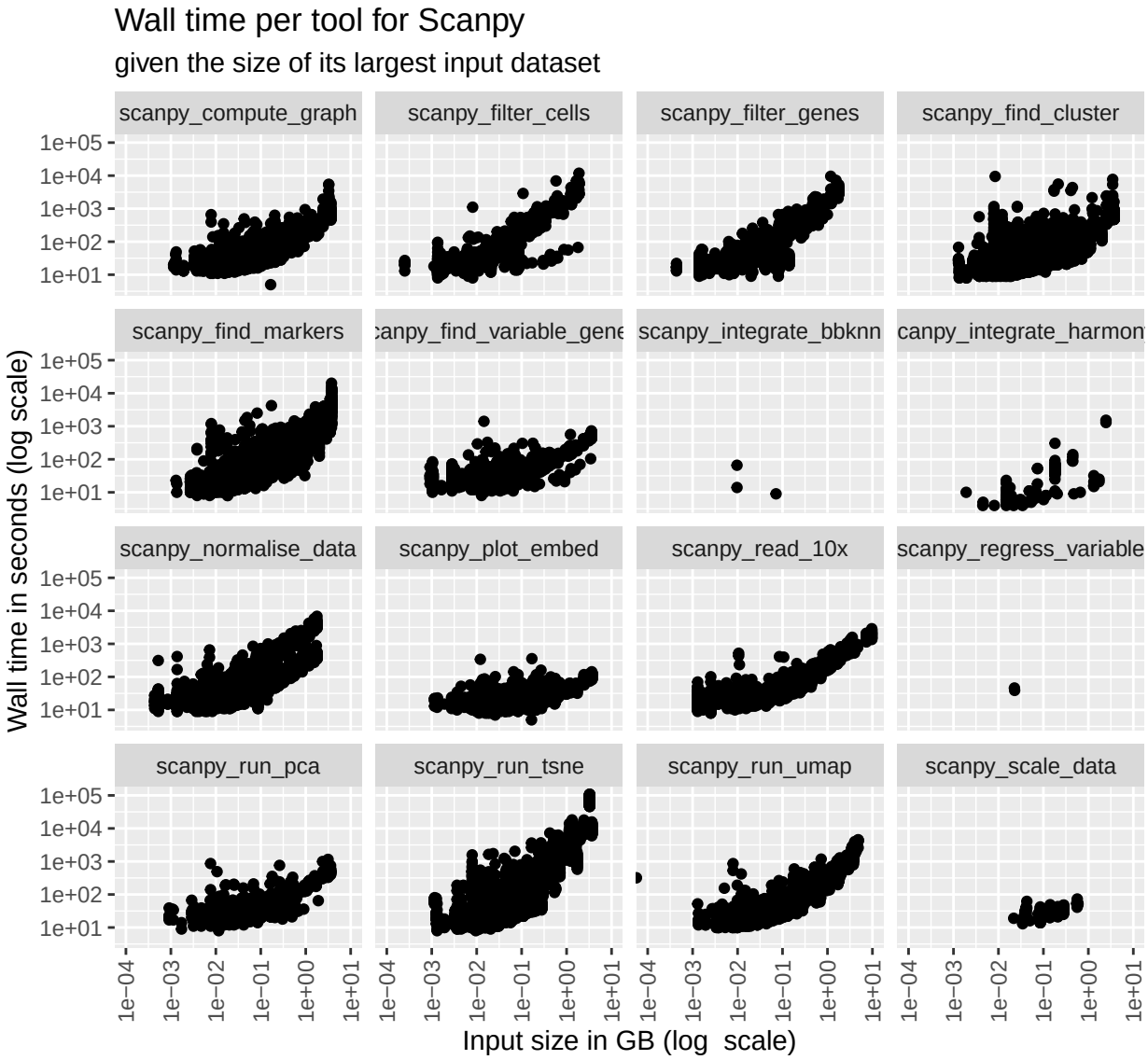

Figure 8: The EBI LSF dataset shows the relation between largest input size and Wall time used by the Scanpy tool set jobs.

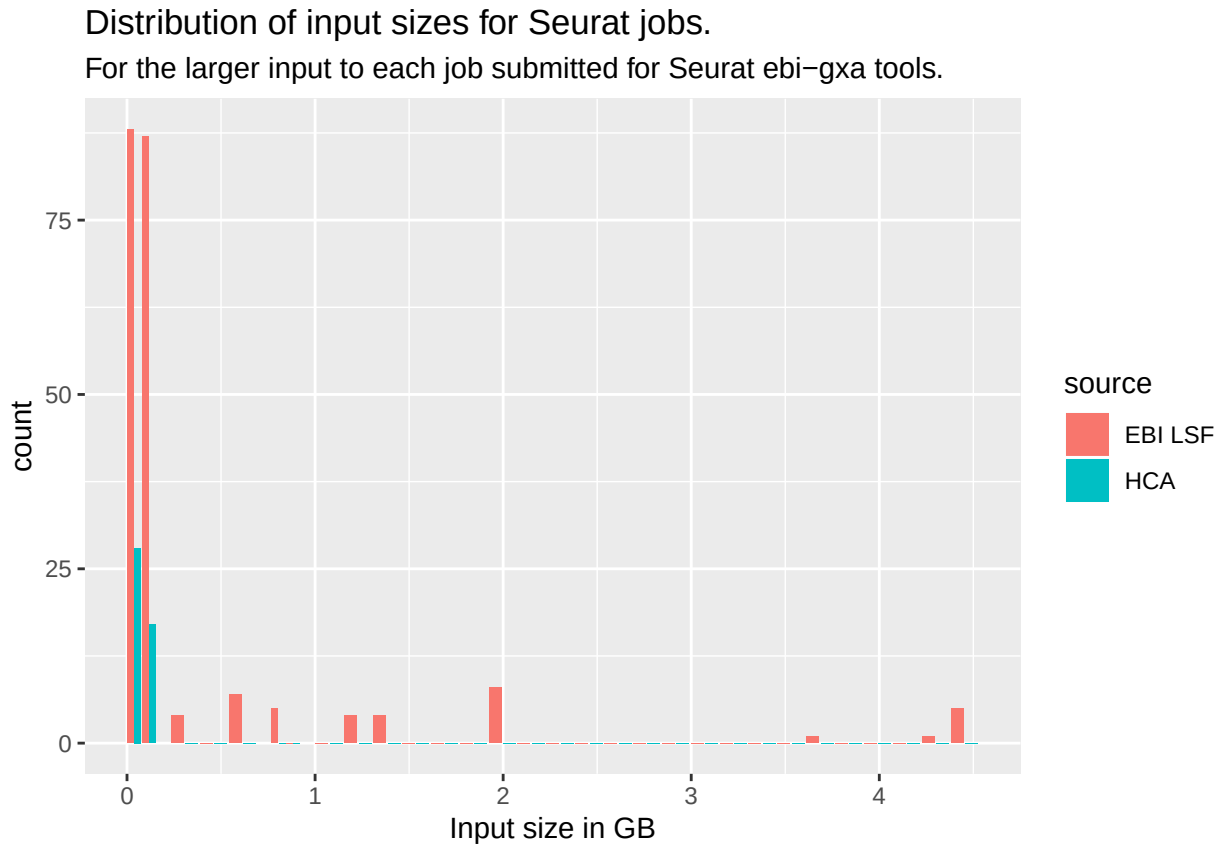

Figure 9: Largest input sizes distribution for Seurat jobs associated to ebi-gxa tools in both the HCA and the EBI LSF galaxy instances.

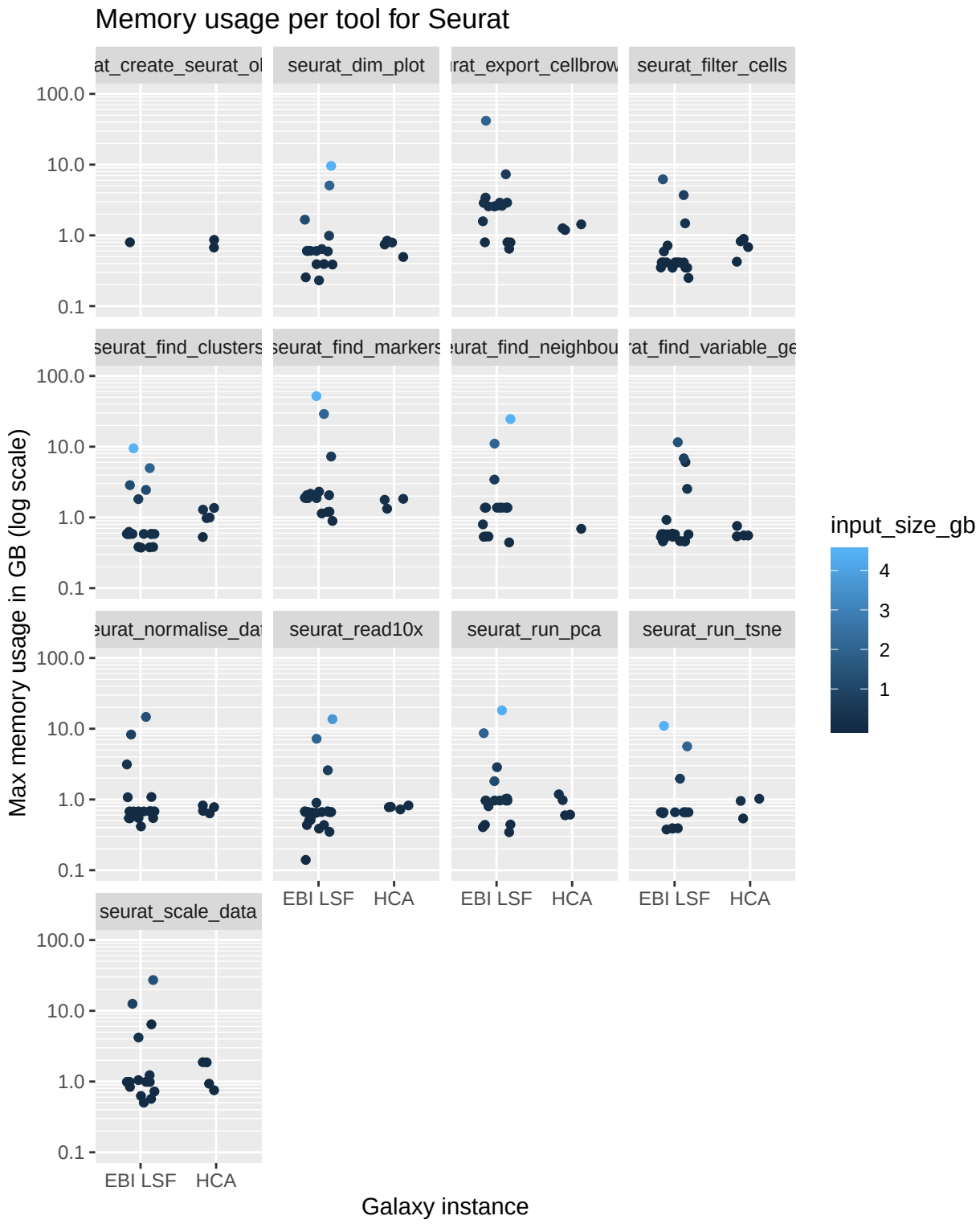

Figure 10: Memory consumption for Seurat tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). The plot only considers real memory, not swap.

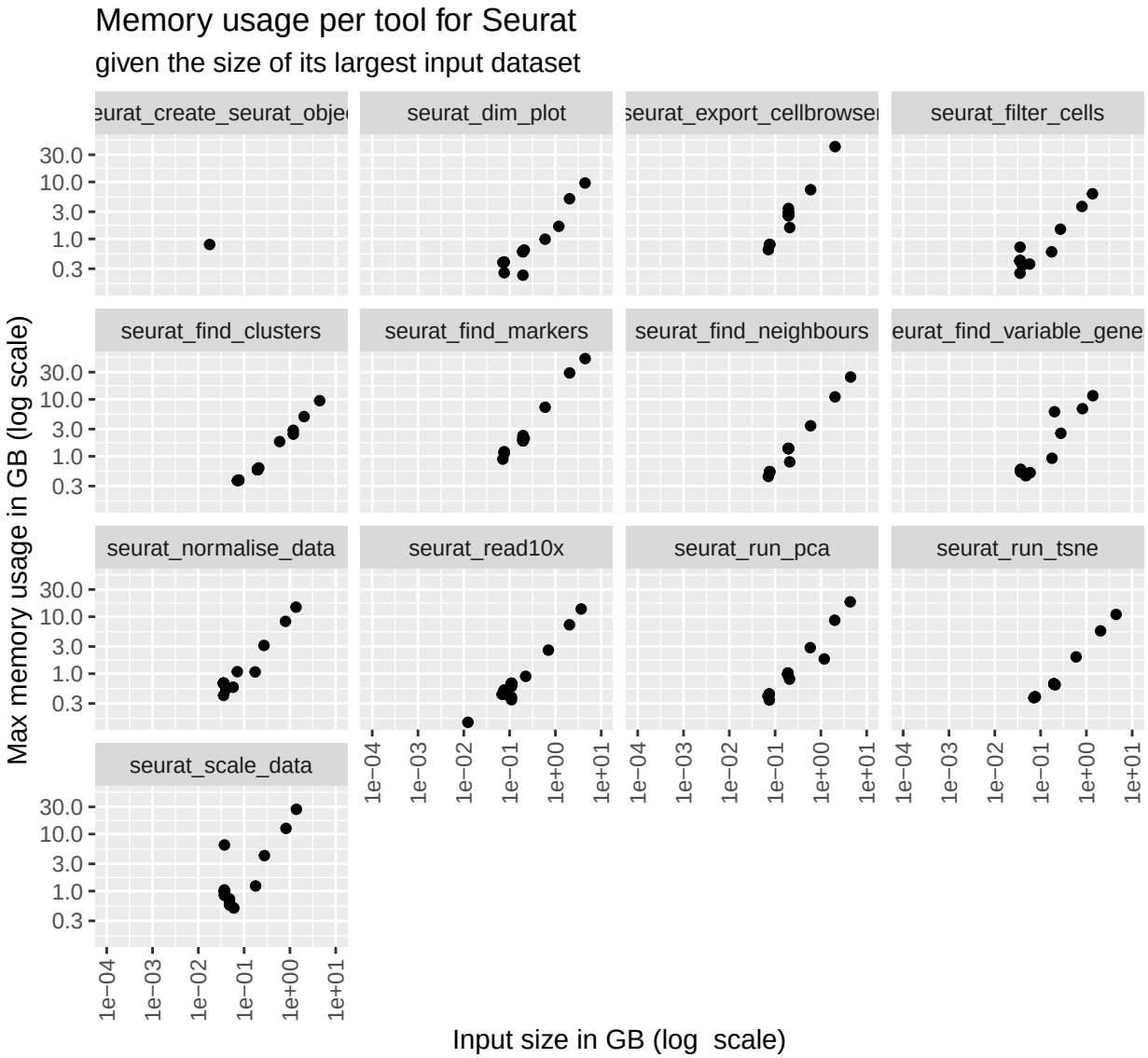

Figure 11: The EBI LSF dataset shows the relation between largest input size and max memory used by the Seurat tool set jobs.

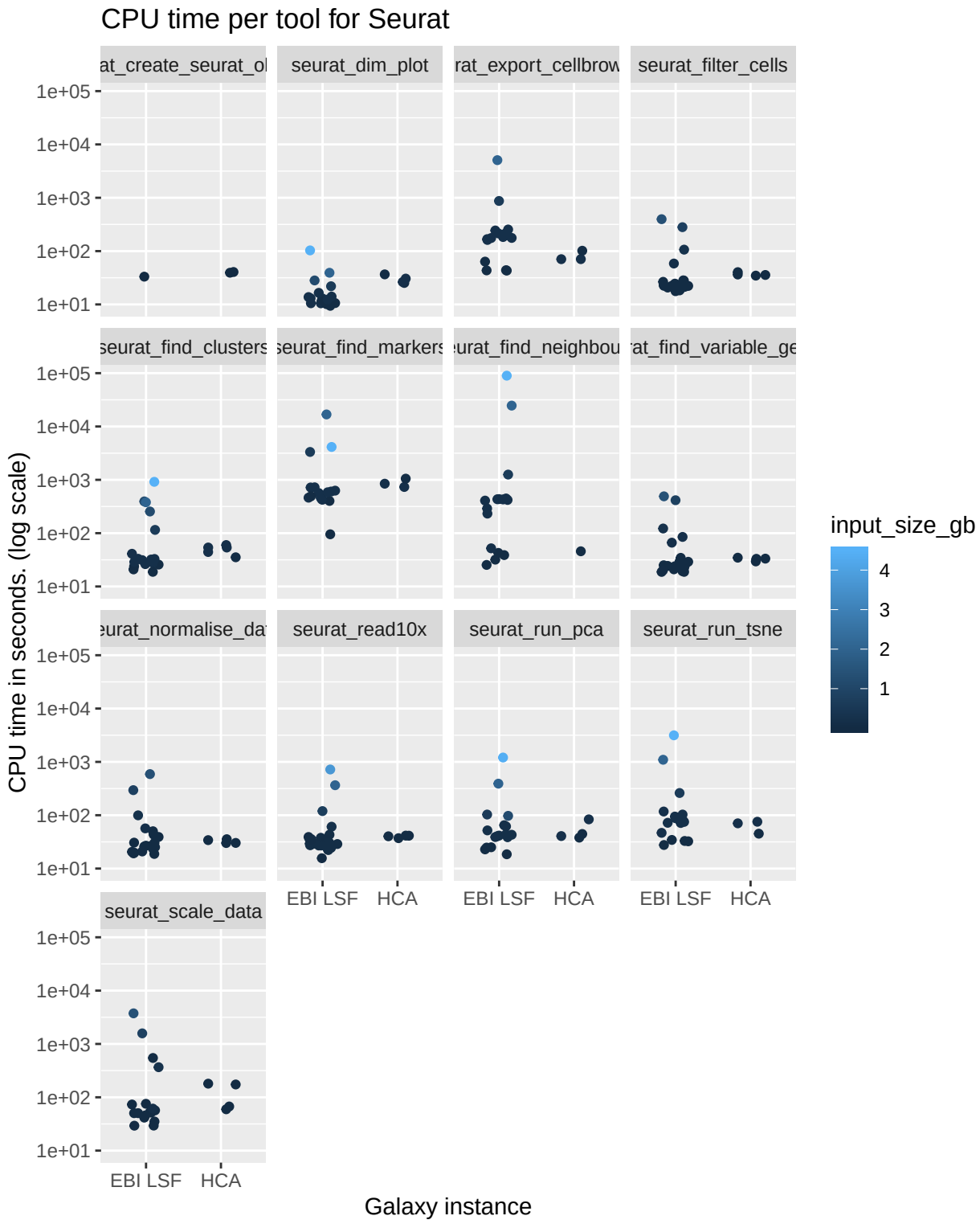

Figure 12: CPU time for Seurat tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). Consider that a job that uses multiple CPUs at the same time will have the time used for each CPU added. This plot reflects the processing power needed by the tool, not the time that the user waits for.

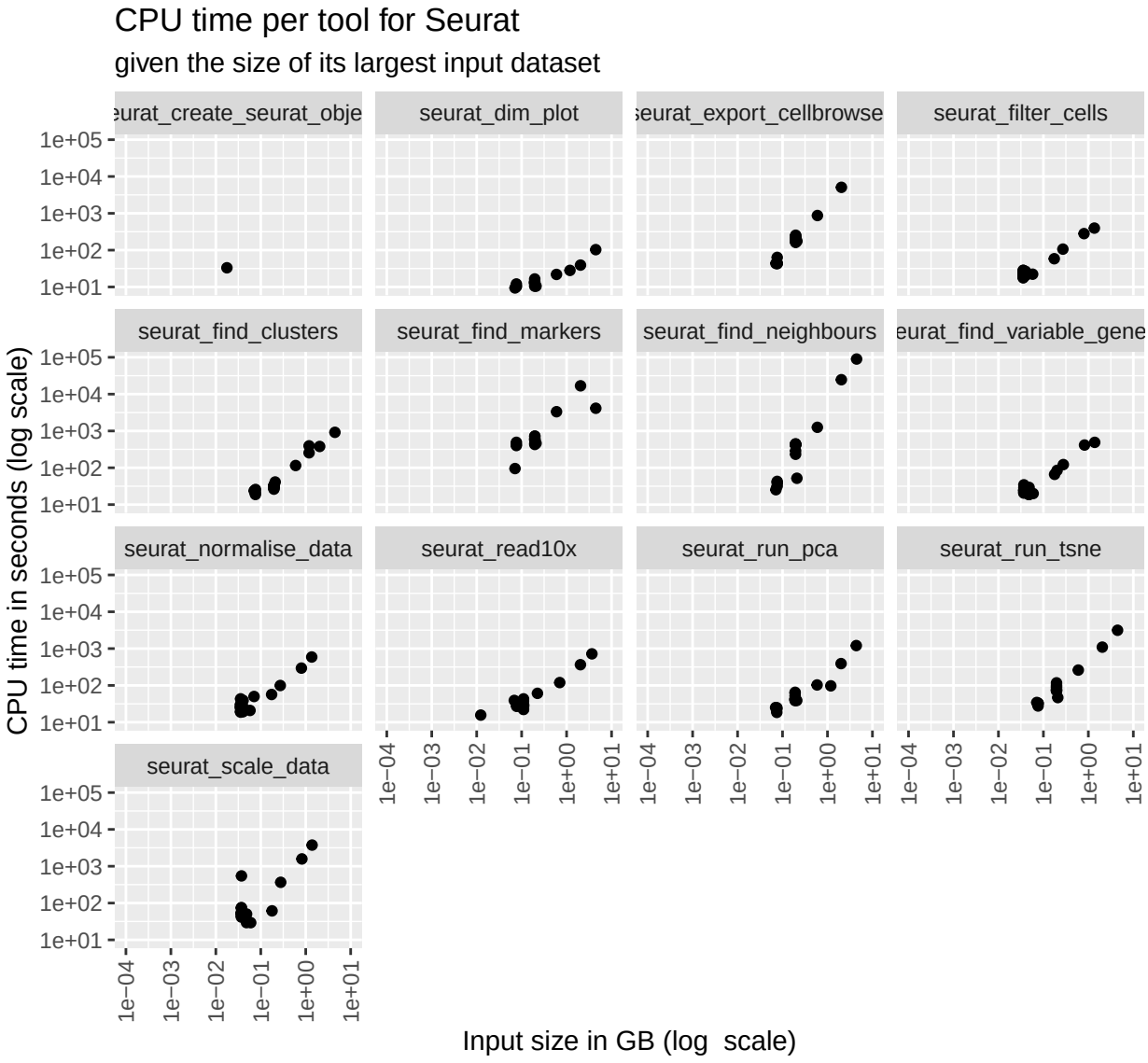

Figure 13: The EBI LSF dataset shows the relation between largest input size and CPU time used by the Seurat tool set jobs.

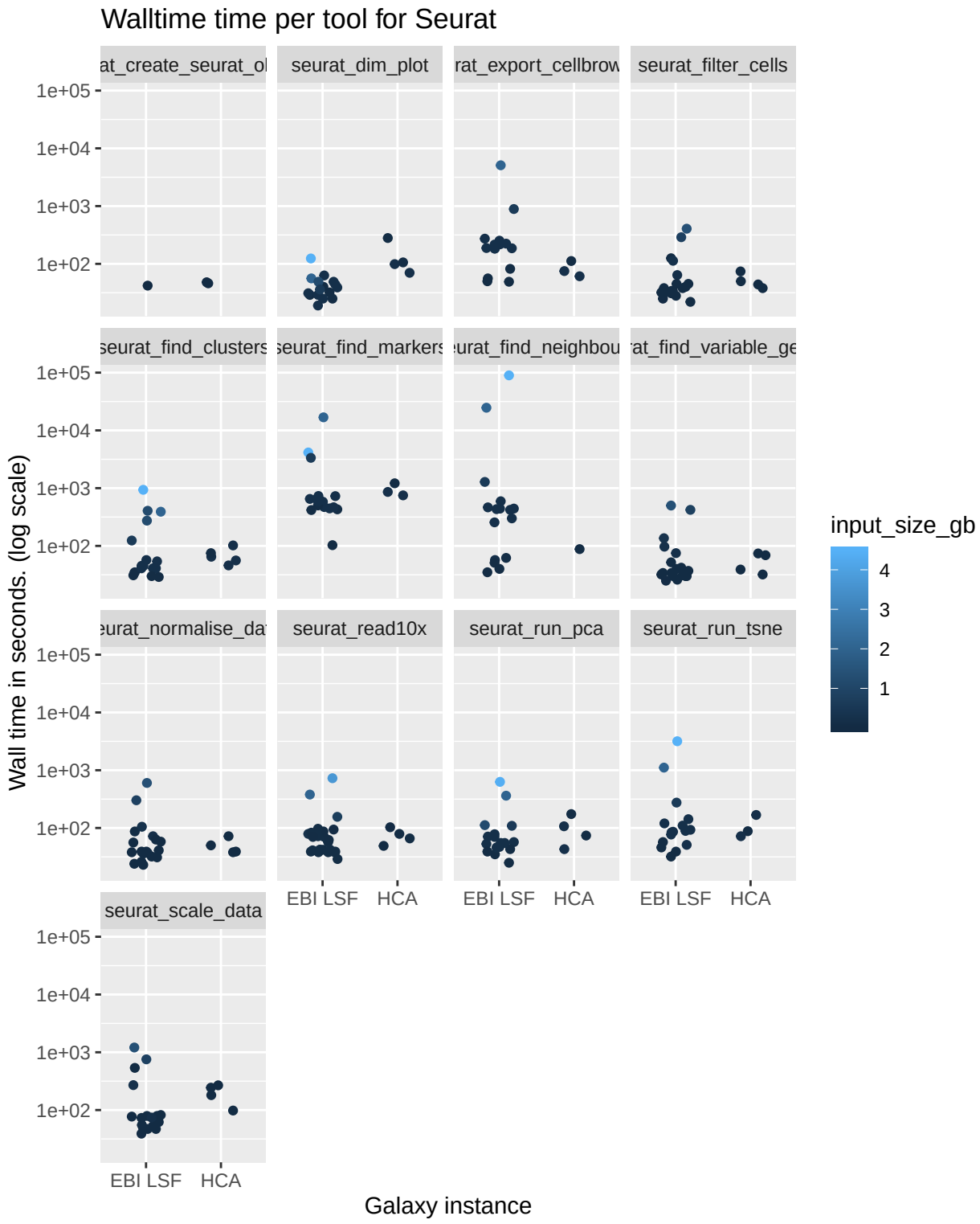

Figure 14: Wall time for Seurat tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). This metric indicates the time that the user waits for the job, and includes time used in processing (serial CPU time), IO waits, file copies, metadata collection, etc.

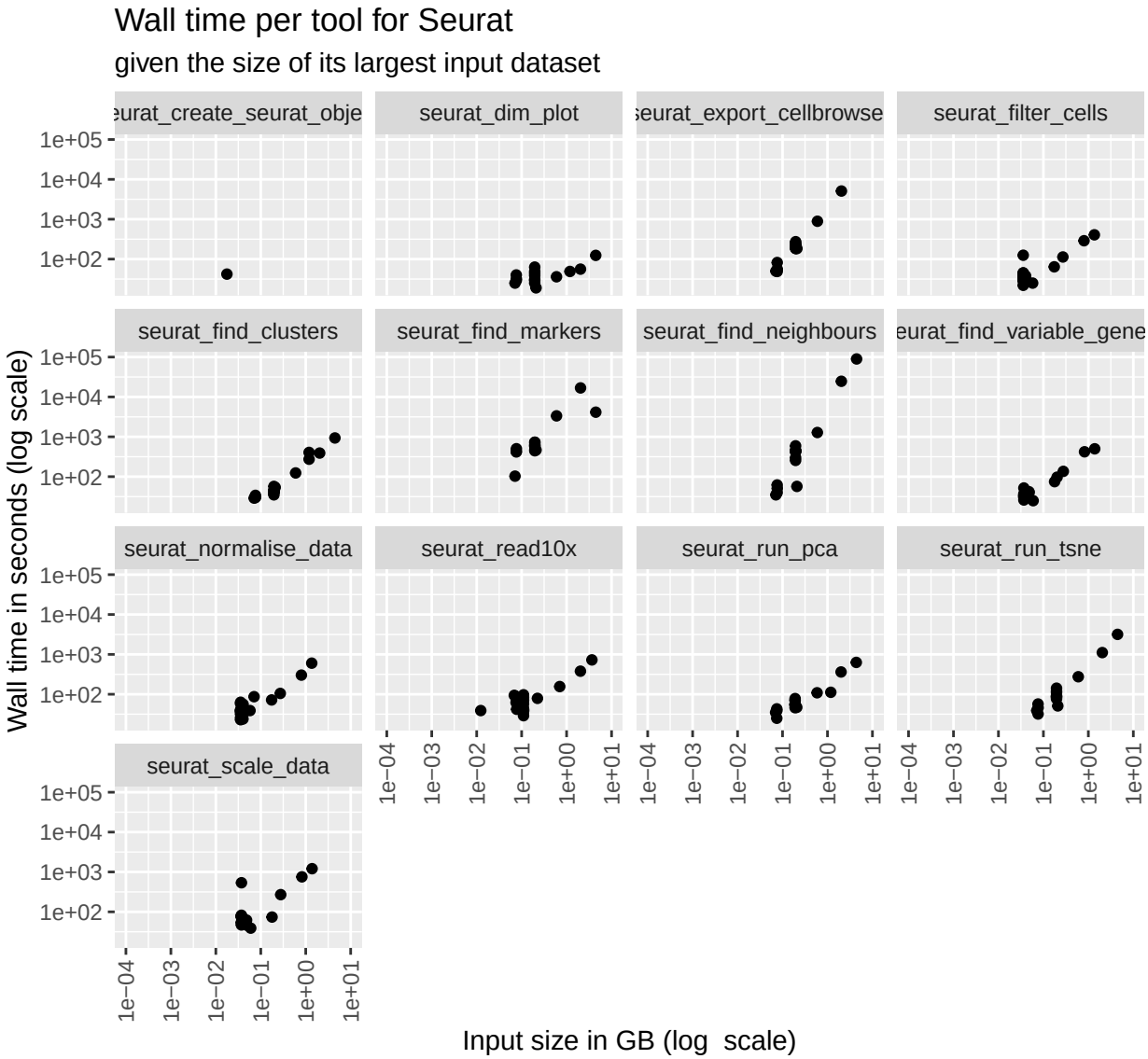

Figure 15: The EBI LSF dataset shows the relation between largest input size and Wall time used by the Seurat tool set jobs.

Figure 16: Largest input sizes distribution for Monocle jobs associated to ebi-gxa tools in both the HCA and the EBI LSF galaxy instances.

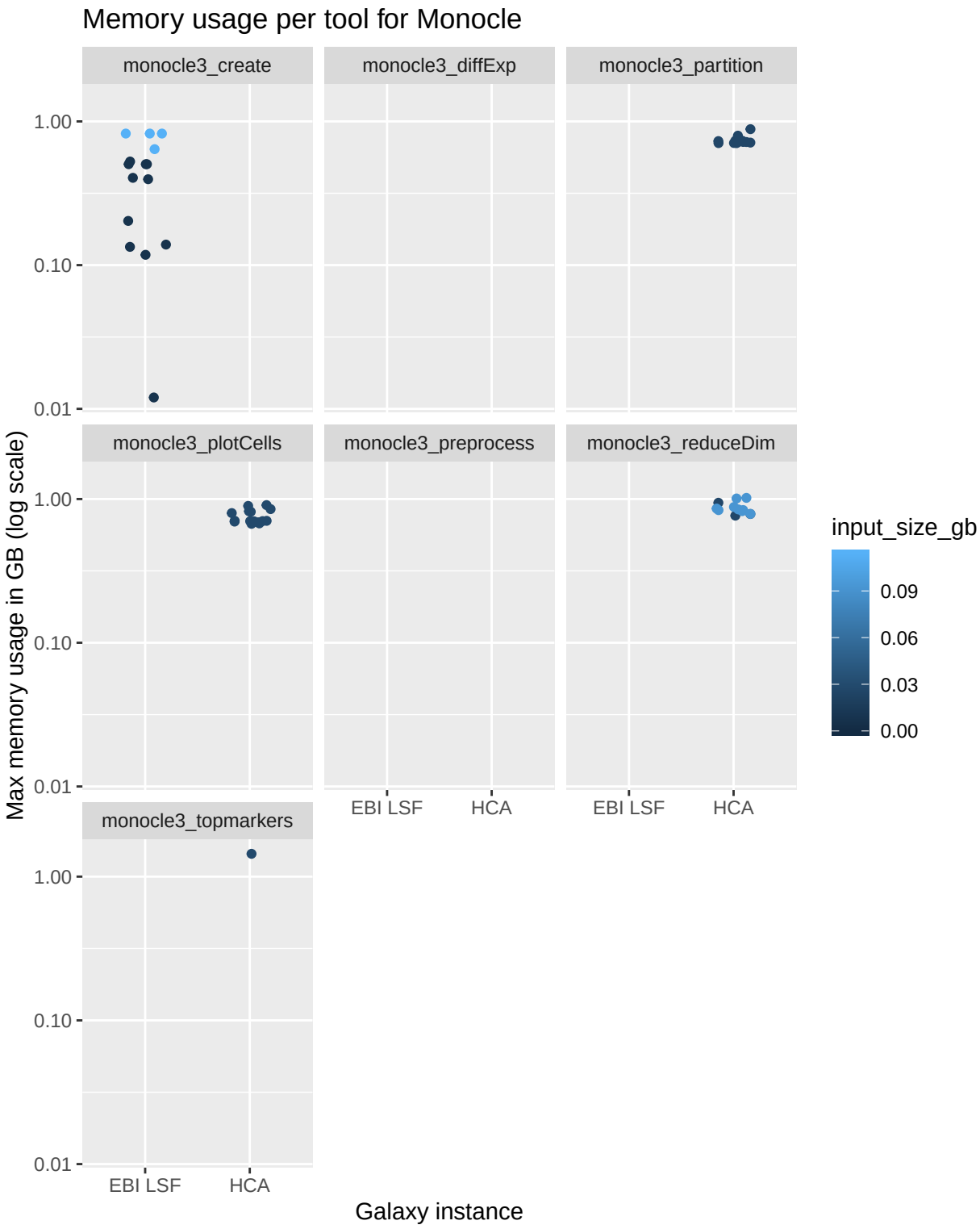

Figure 17: Memory consumption for Monocle tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). The plot only considers real memory, not swap.

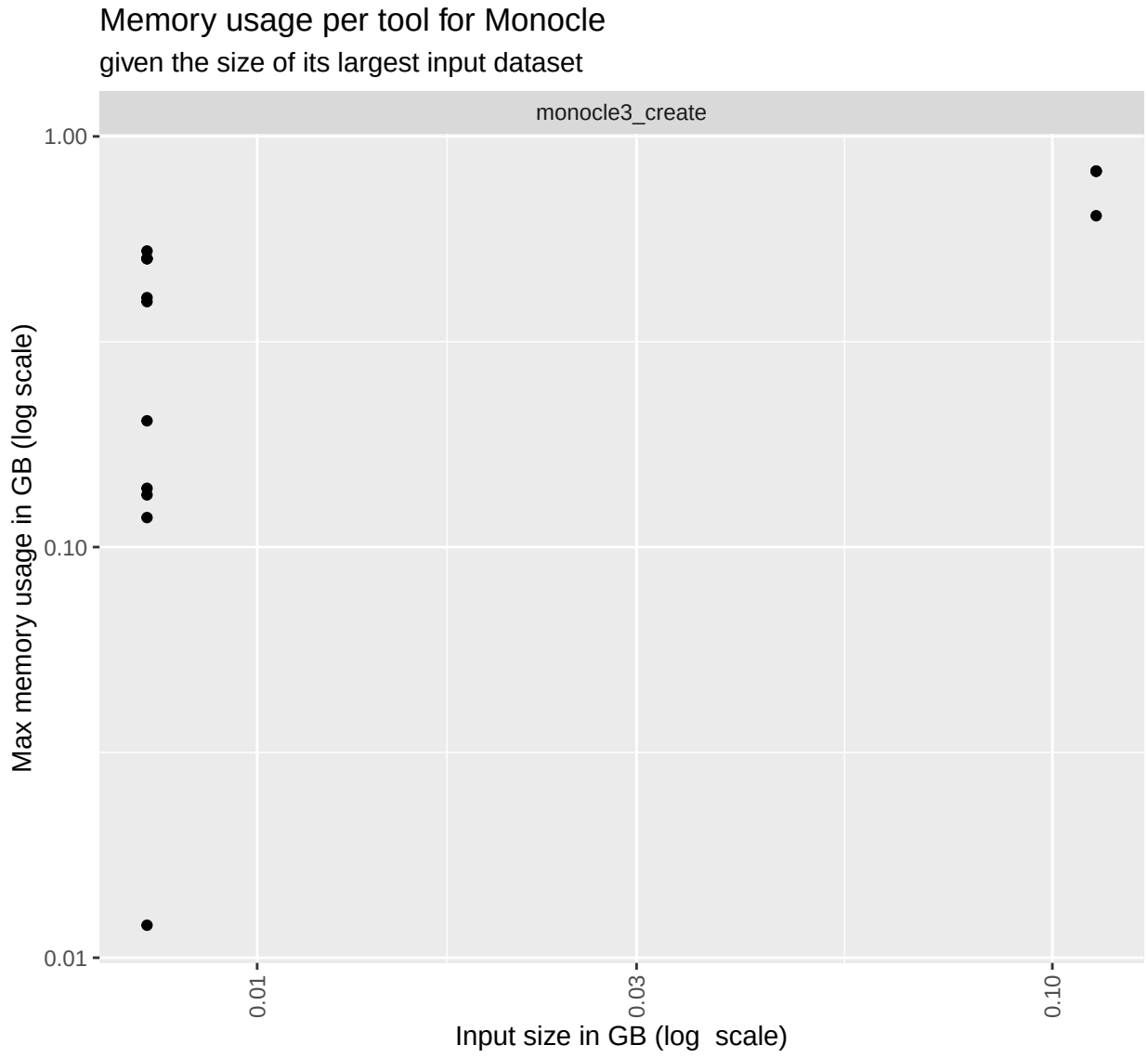

Figure 18: The EBI LSF dataset shows the relation between largest input size and max memory used by the Monocle tool set jobs.

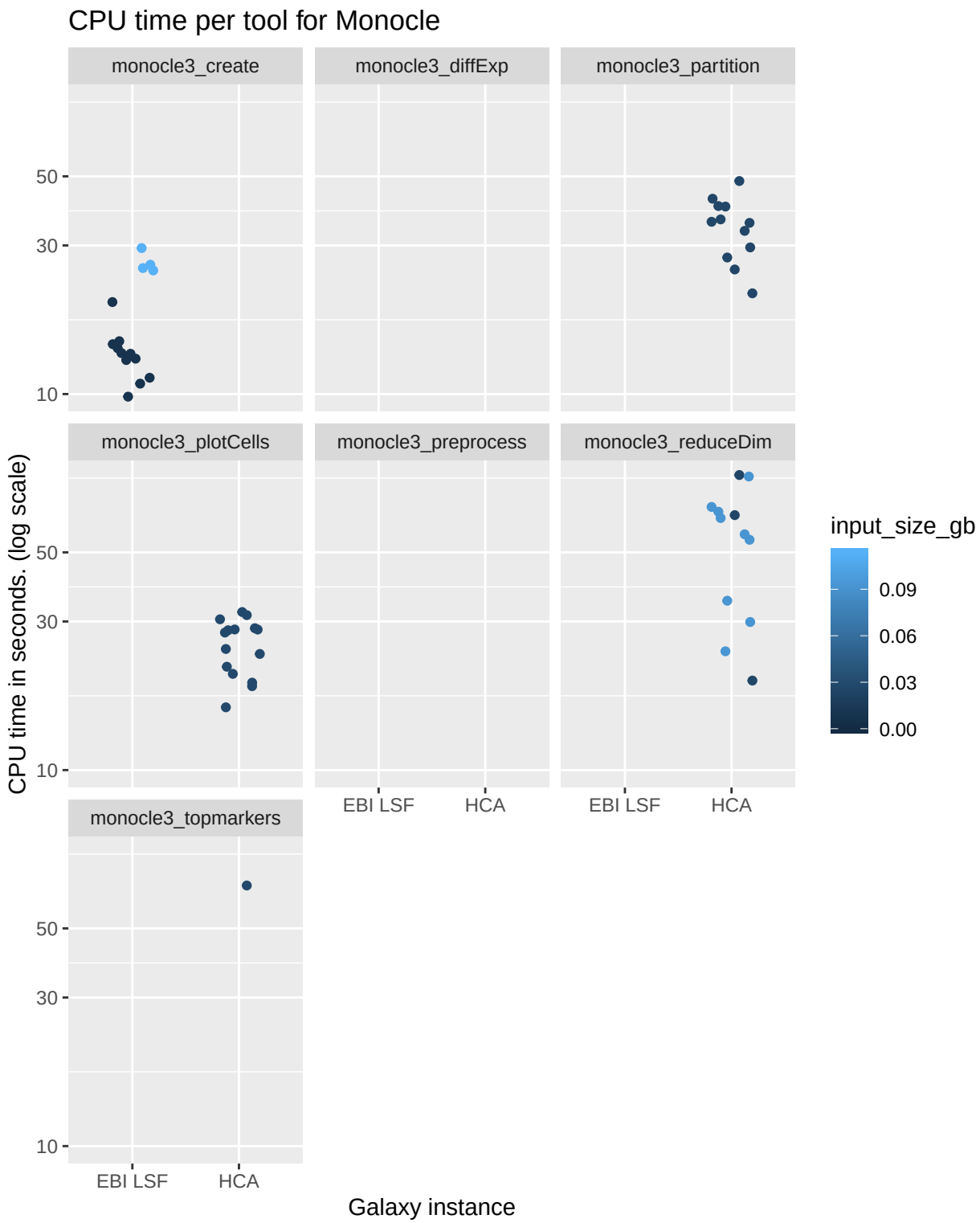

Figure 19: CPU time for Monocle tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). Consider that a job that uses multiple CPUs at the same time will have the time used for each CPU added. This plot reflects the processing power needed by the tool, not the time that the user waits for.

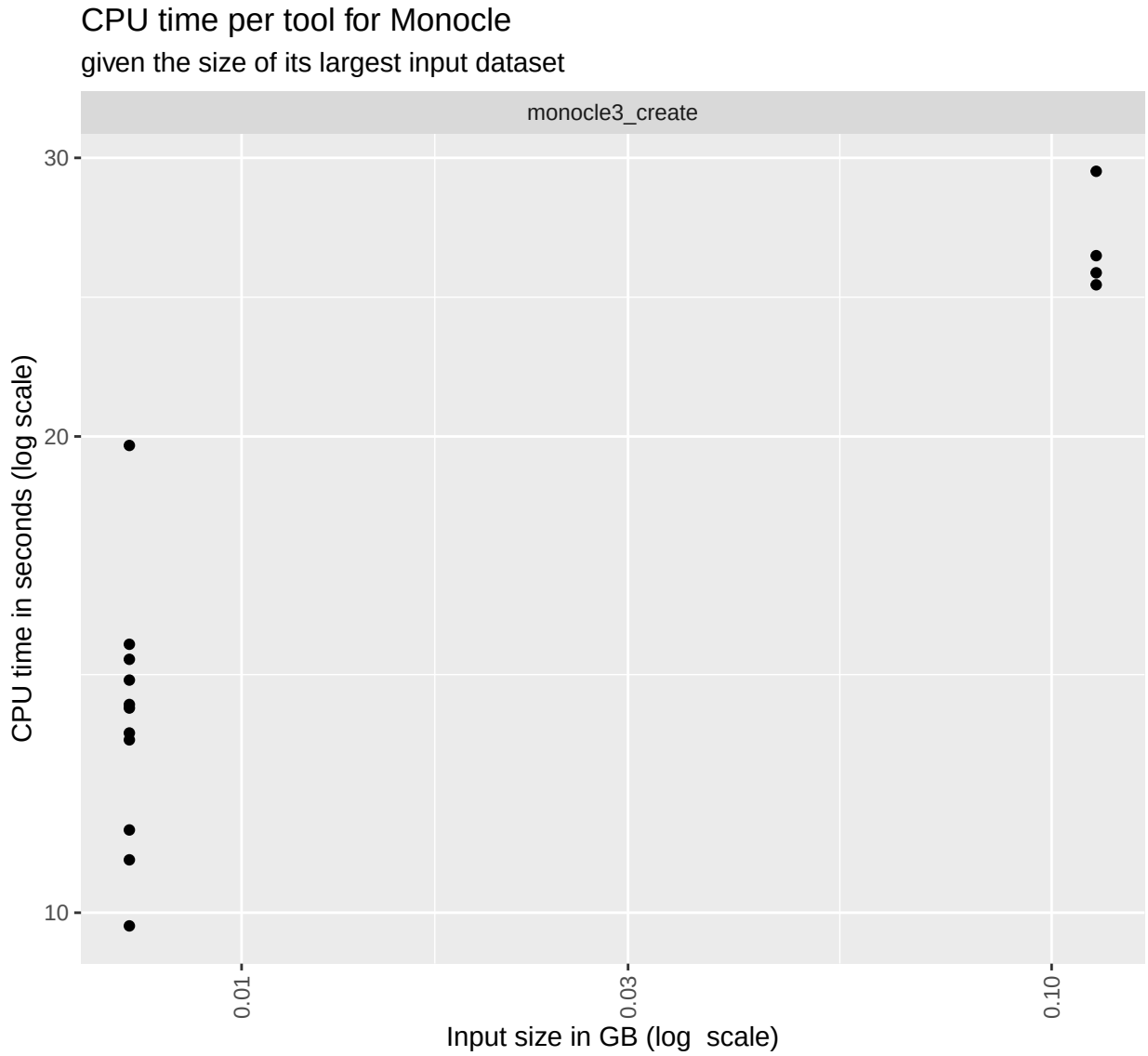

Figure 20: The EBI LSF dataset shows the relation between largest input size and CPU time used by the Monocle tool set jobs.

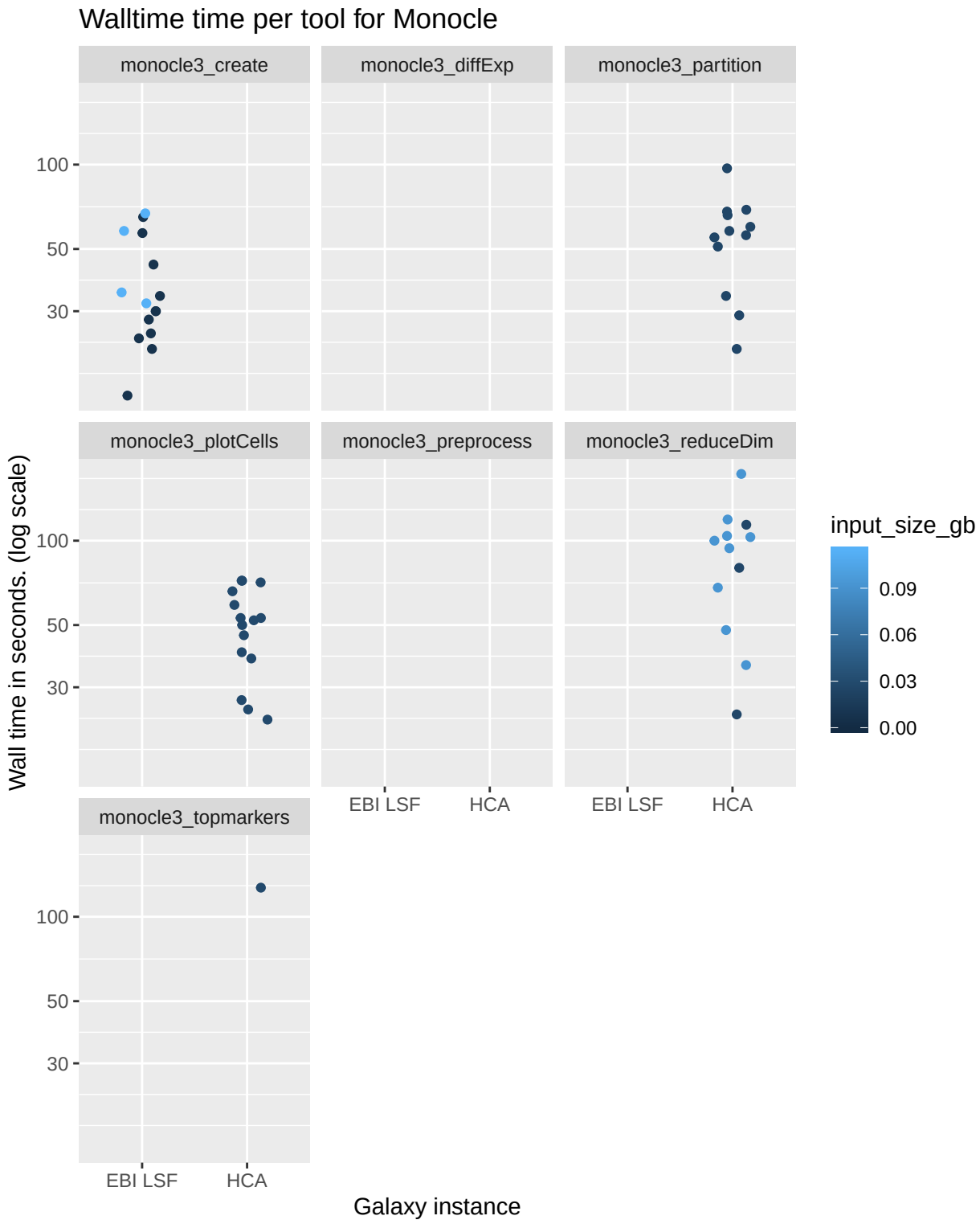

Figure 21: Wall time for Monocle tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). This metric indicates the time that the user waits for the job, and includes time used in processing (serial CPU time), IO waits, file copies, metadata collection, etc.

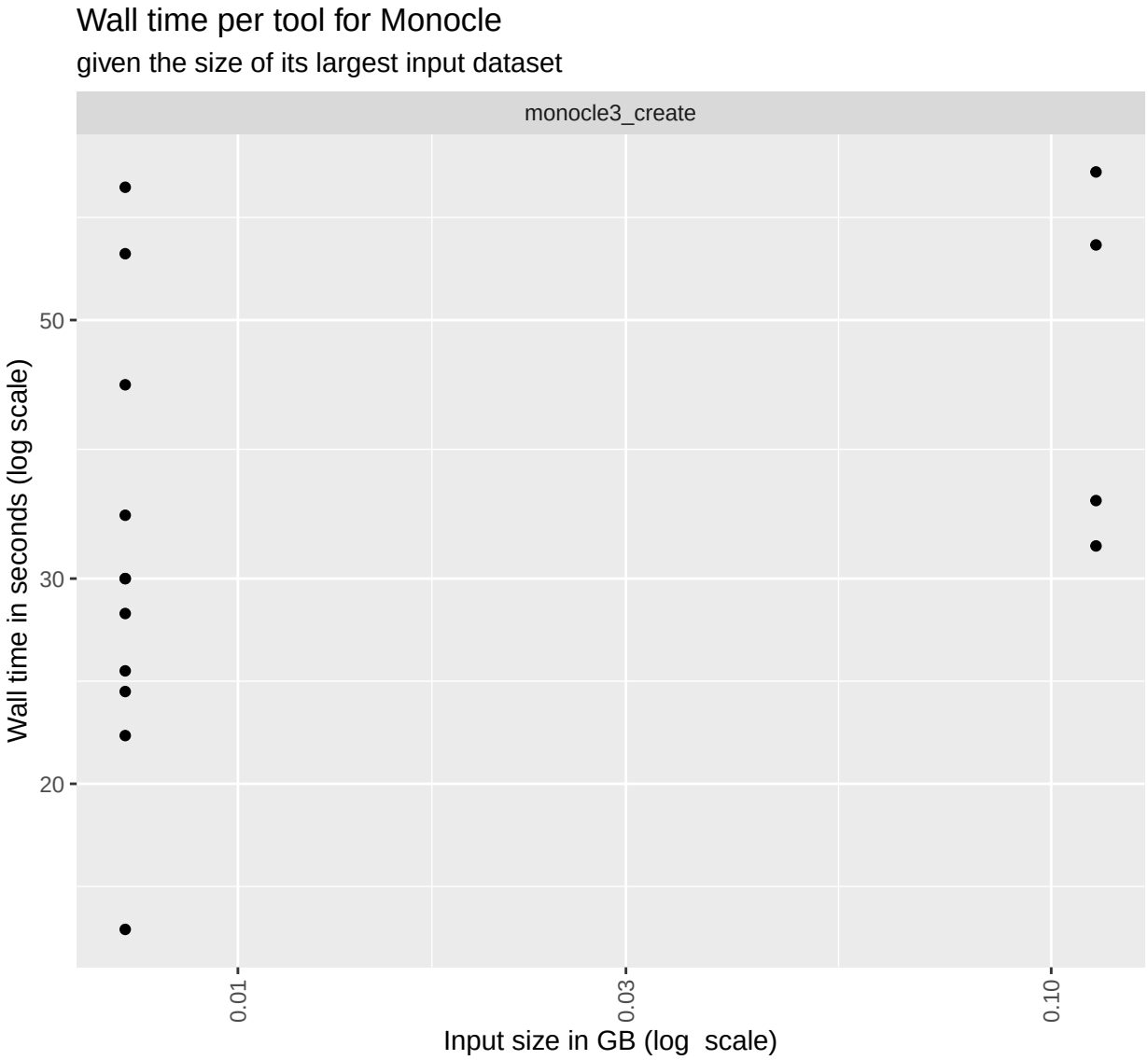

Figure 22: The EBI LSF dataset shows the relation between largest input size and Wall time used by the Monocle tool set jobs.

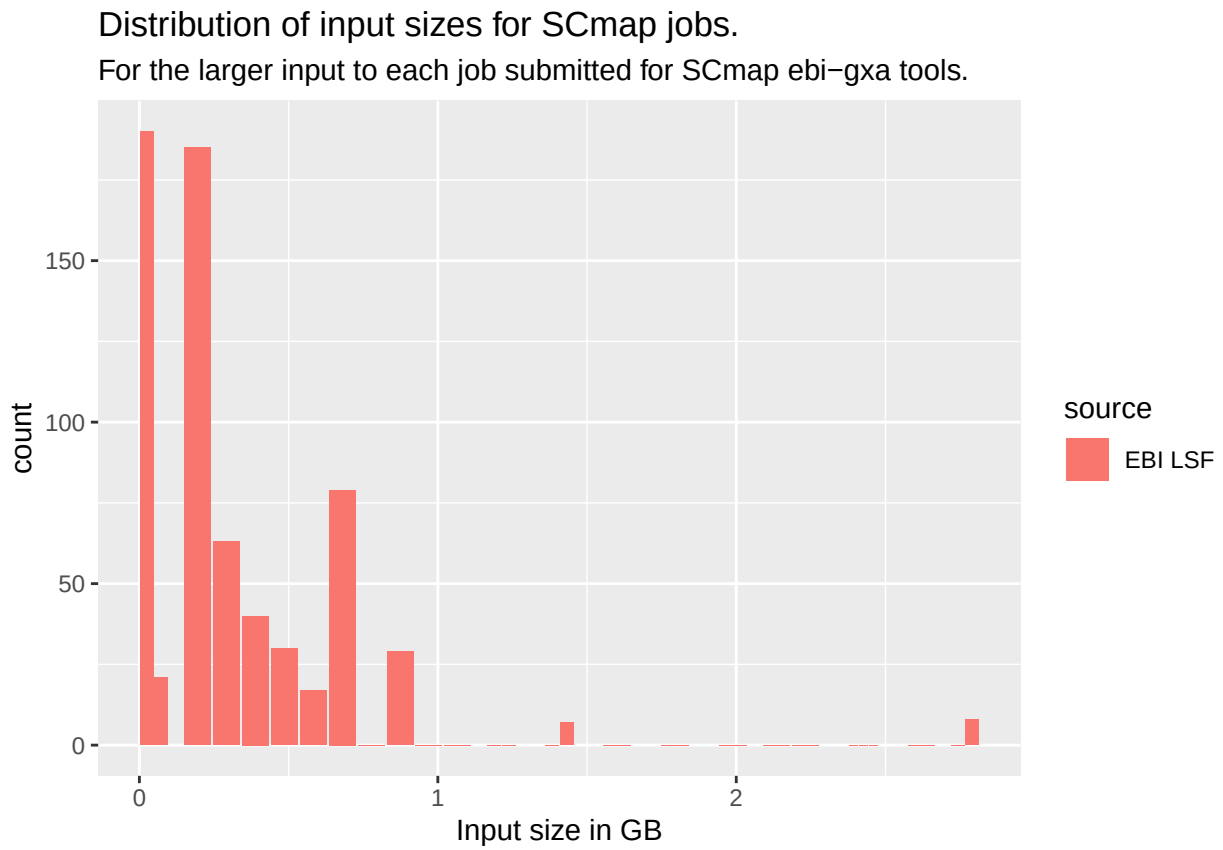

Figure 23: Largest input sizes distribution for SCmap jobs associated to ebi-gxa tools in both the HCA and the EBI LSF galaxy instances.

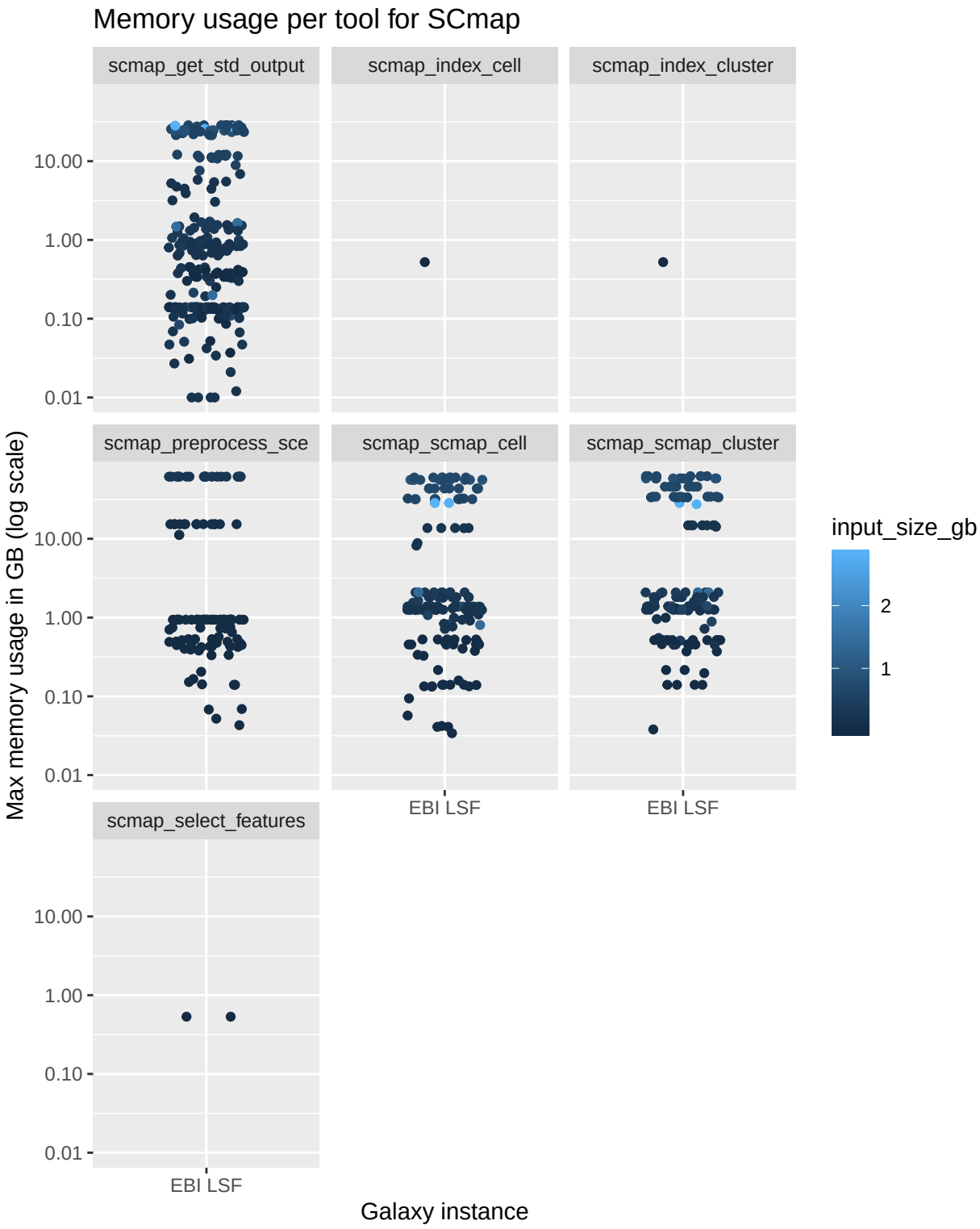

Figure 24: Memory consumption for SCmap tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). The plot only considers real memory, not swap.

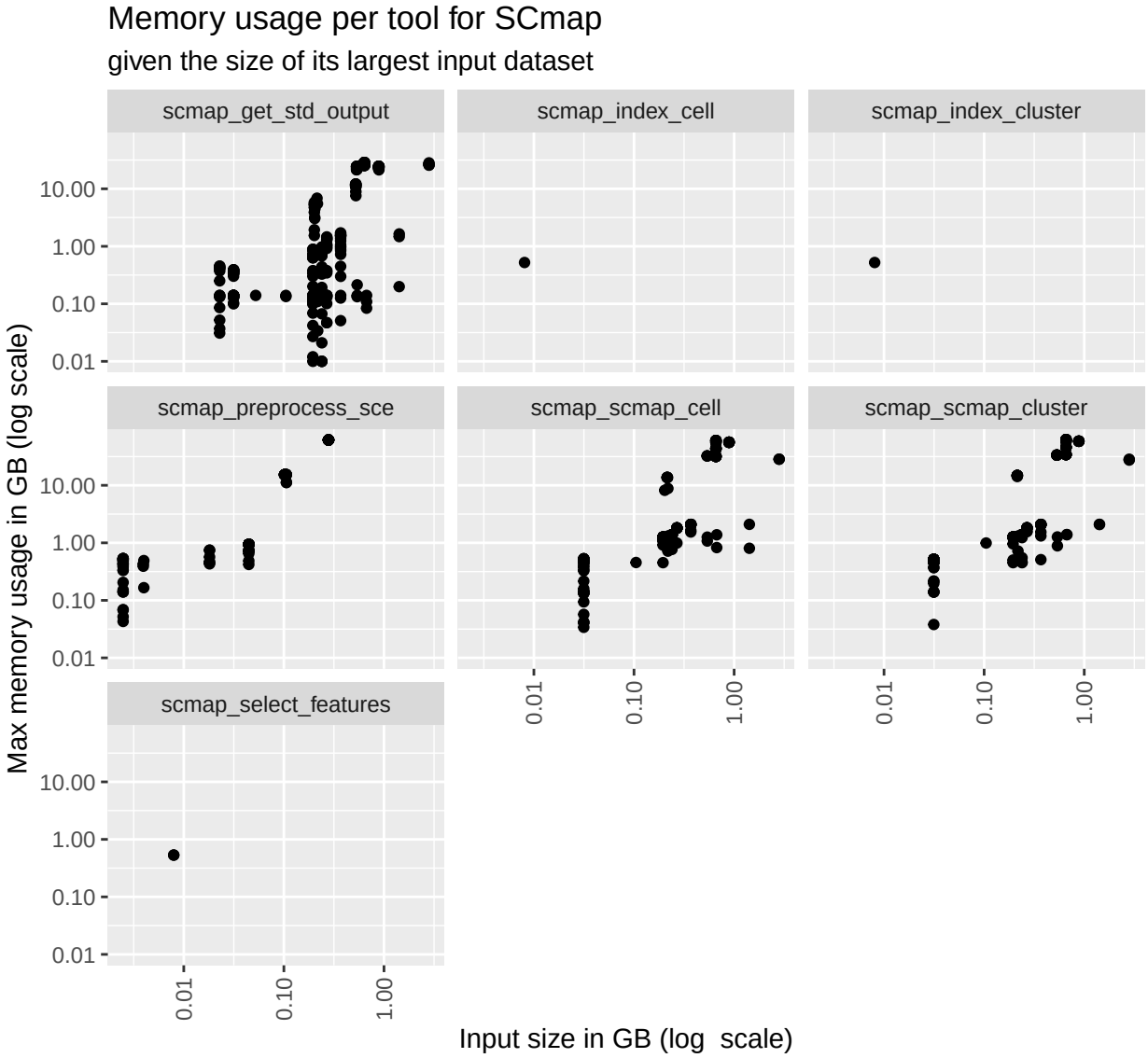

Figure 25: The EBI LSF dataset shows the relation between largest input size and max memory used by the SCmap tool set jobs.

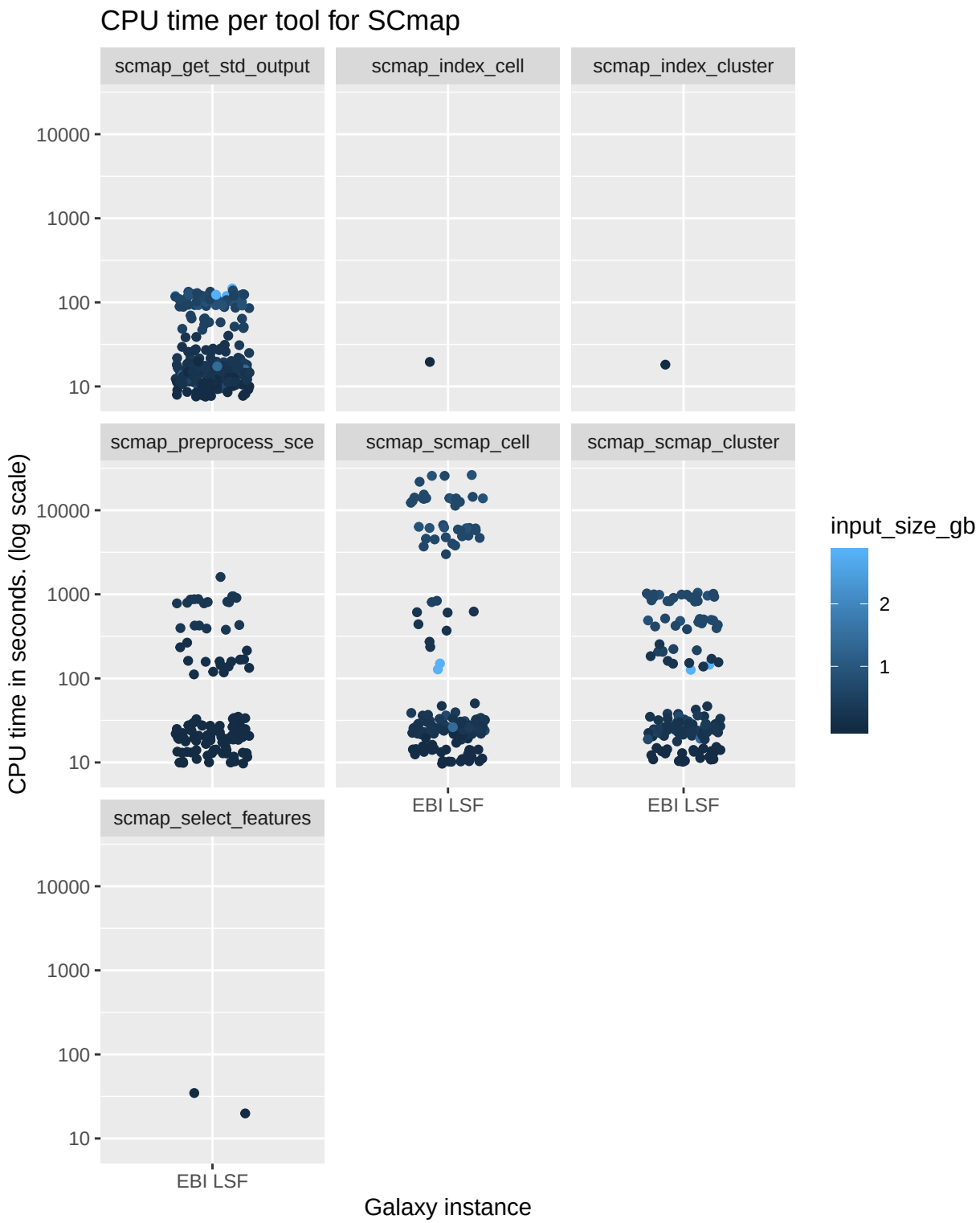

Figure 26: CPU time for SCmap tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). Consider that a job that uses multiple CPUs at the same time will have the time used for each CPU added. This plot reflects the processing power needed by the tool, not the time that the user waits for.

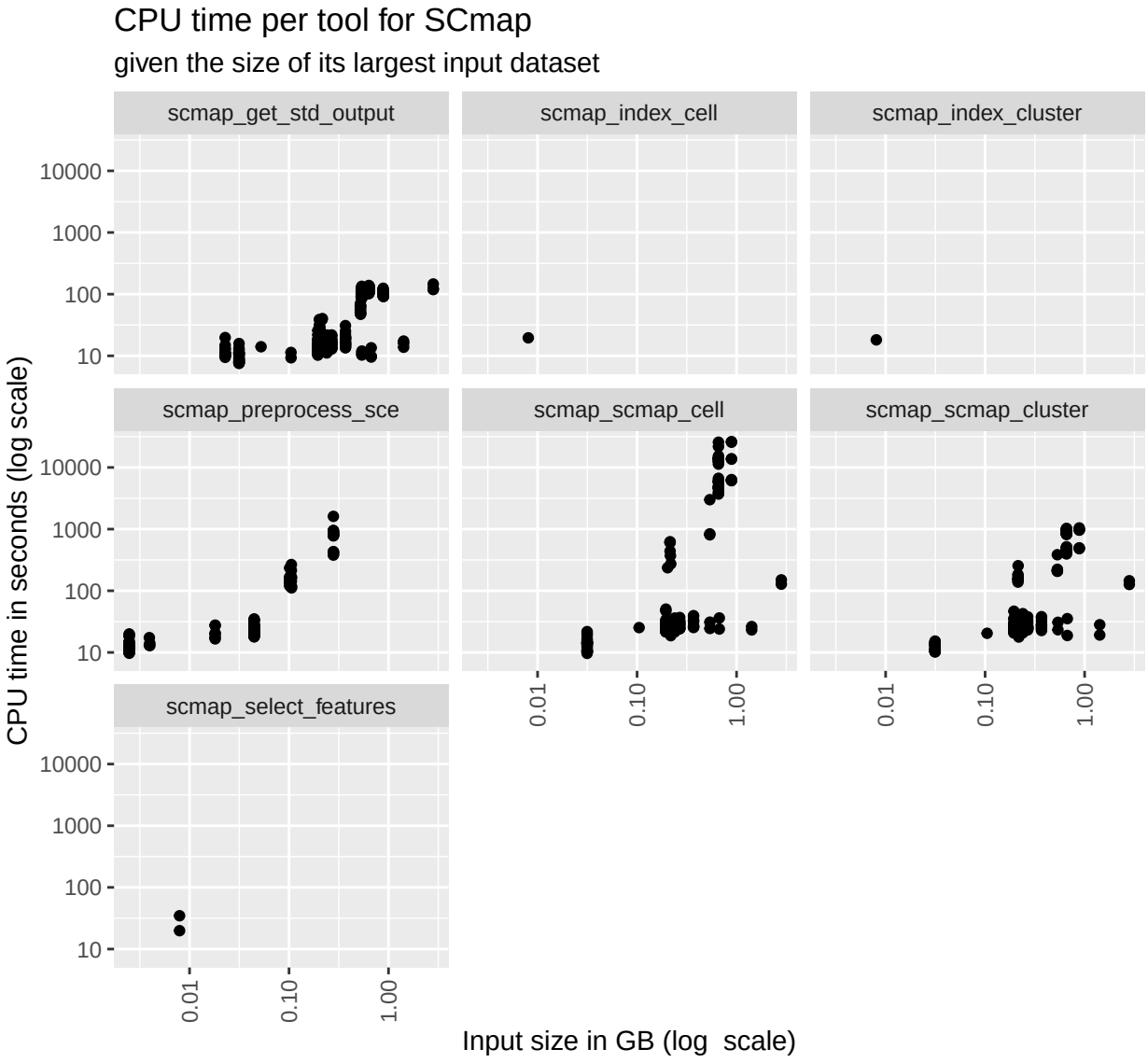

Figure 27: The EBI LSF dataset shows the relation between largest input size and CPU time used by the SCmap tool set jobs.

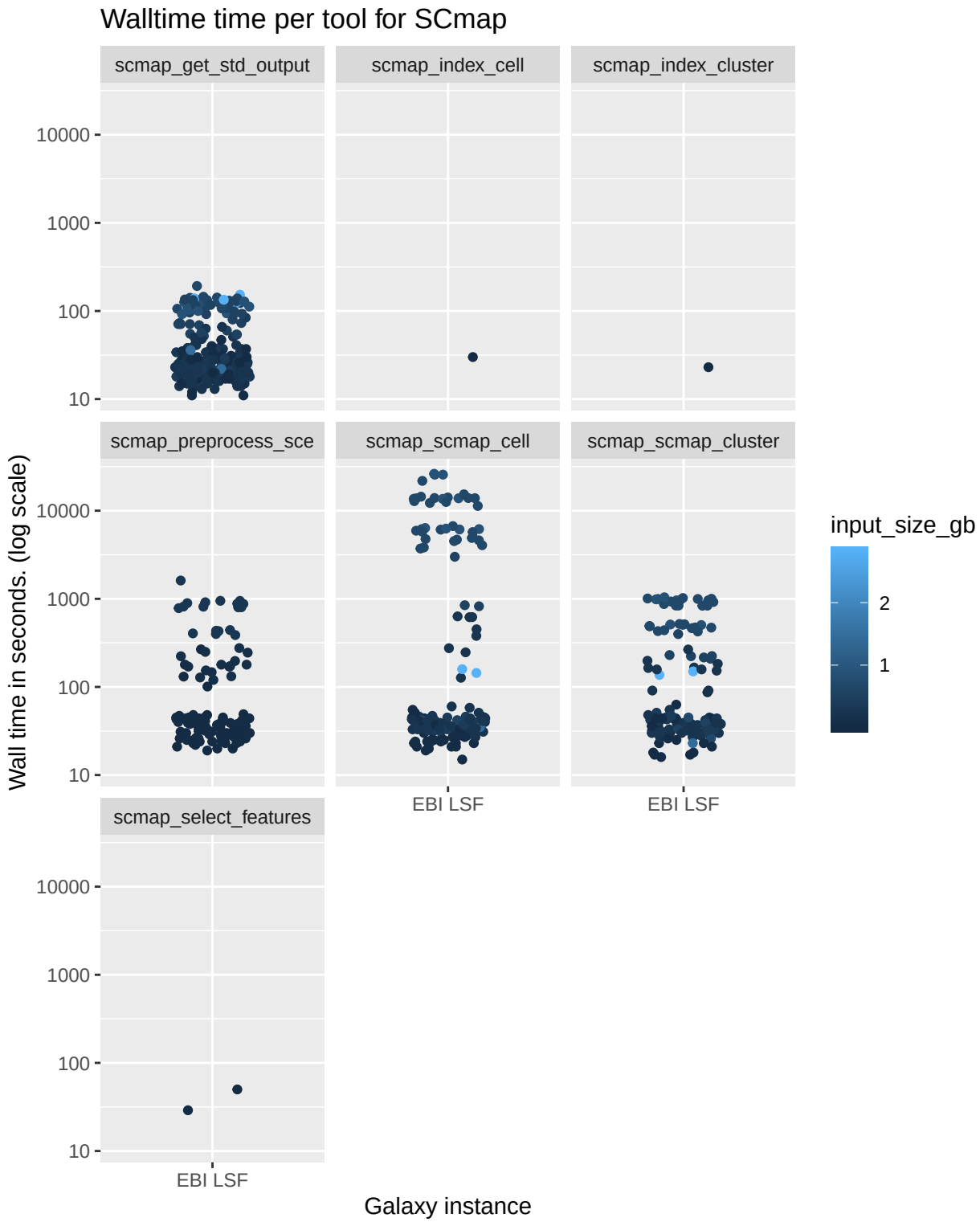

Figure 28: Wall time for SCmap tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). This metric indicates the time that the user waits for the job, and includes time used in processing (serial CPU time), IO waits, file copies, metadata collection, etc.

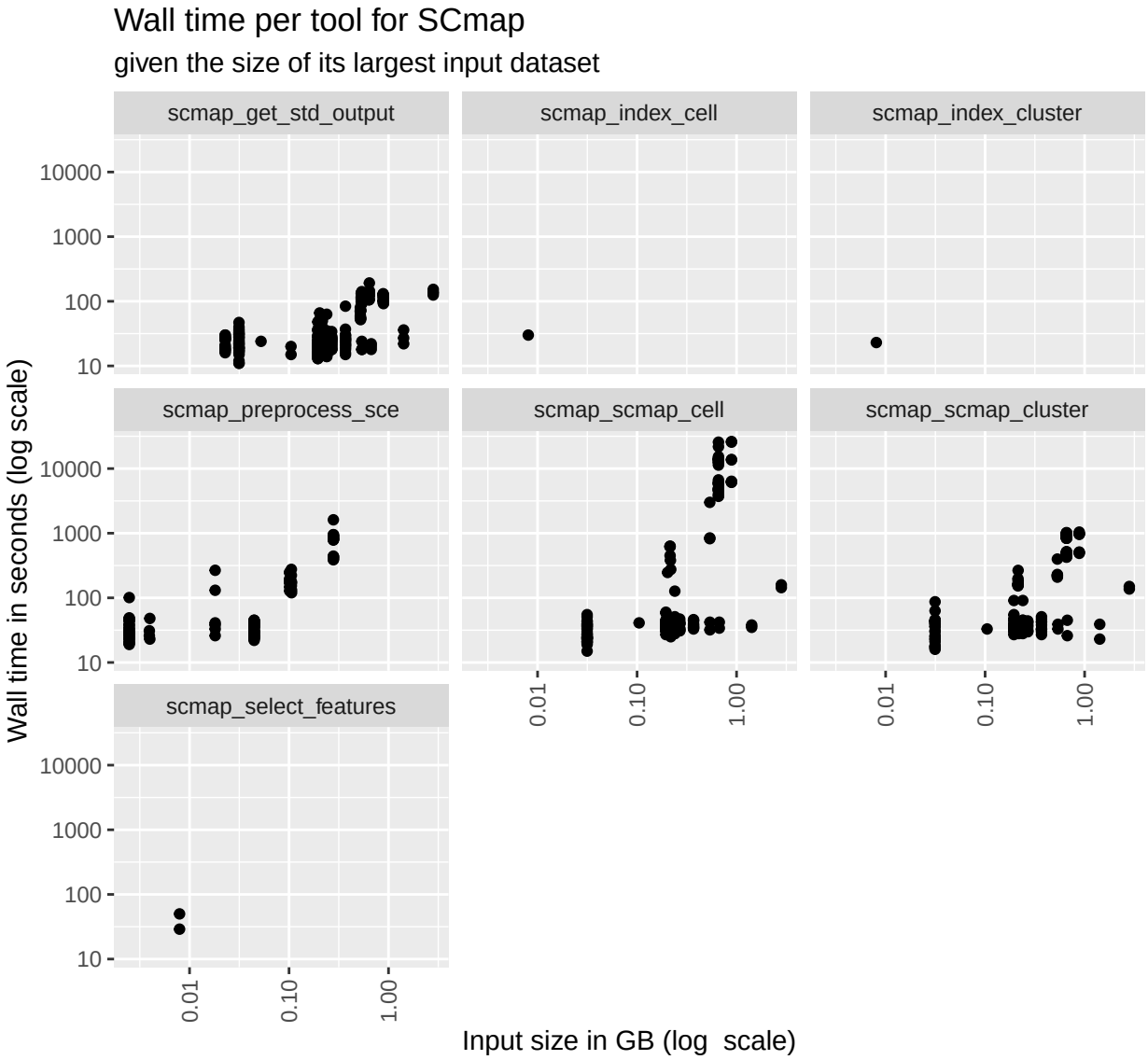

Figure 29: The EBI LSF dataset shows the relation between largest input size and Wall time used by the SCmap tool set jobs.

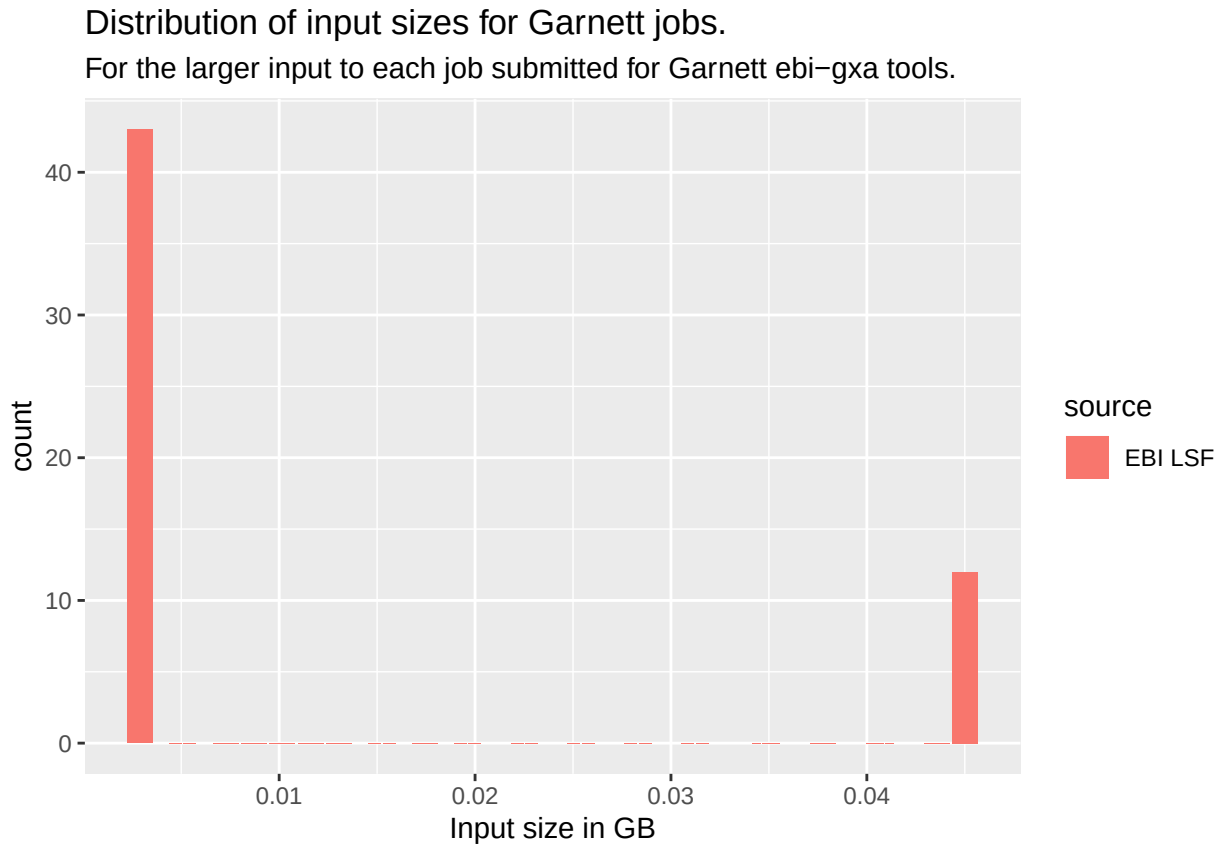

Figure 30: Largest input sizes distribution for Garnett jobs associated to ebi-gxa tools in both the HCA and the EBI LSF galaxy instances.

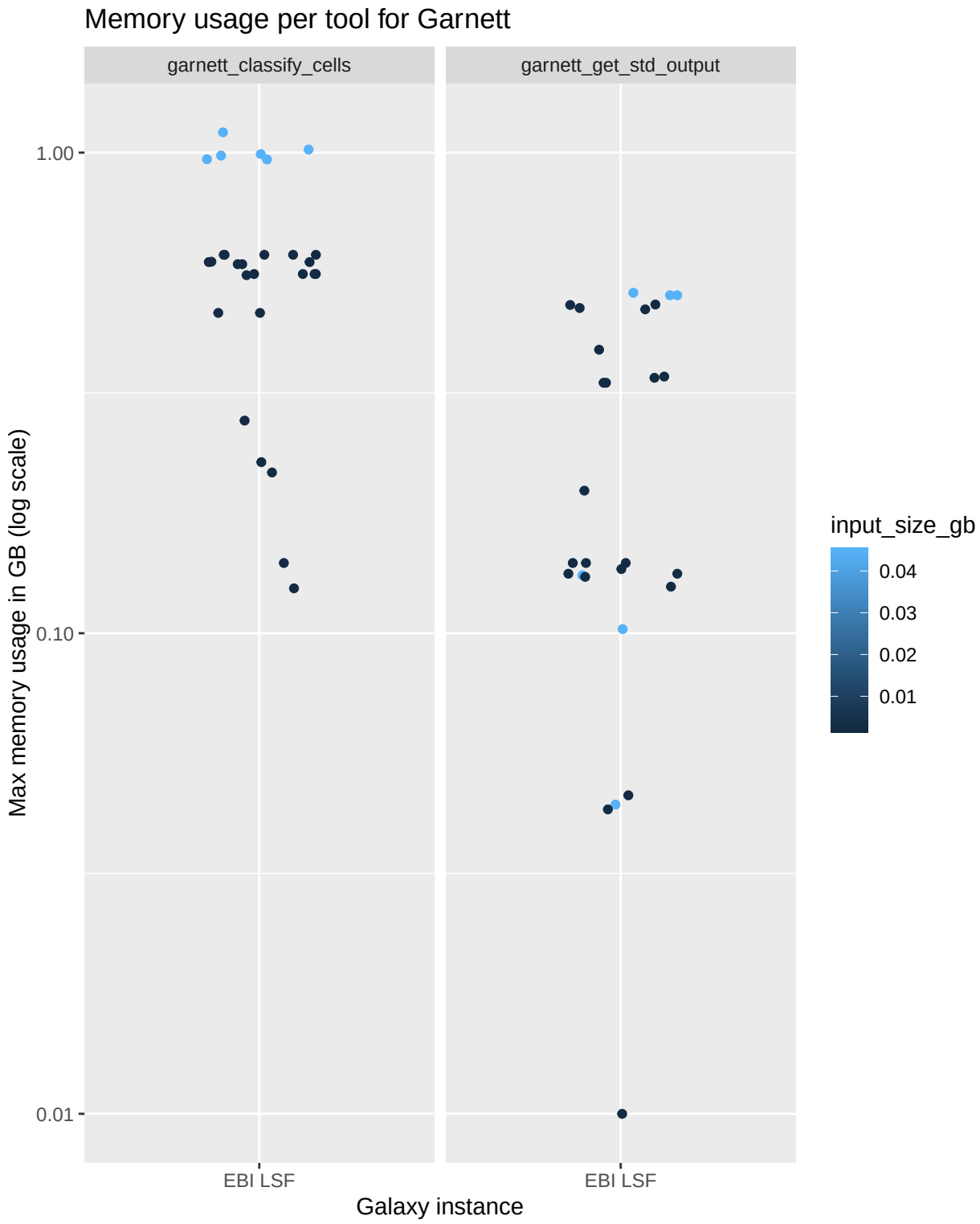

Figure 31: Memory consumption for Garnett tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). The plot only considers real memory, not swap.

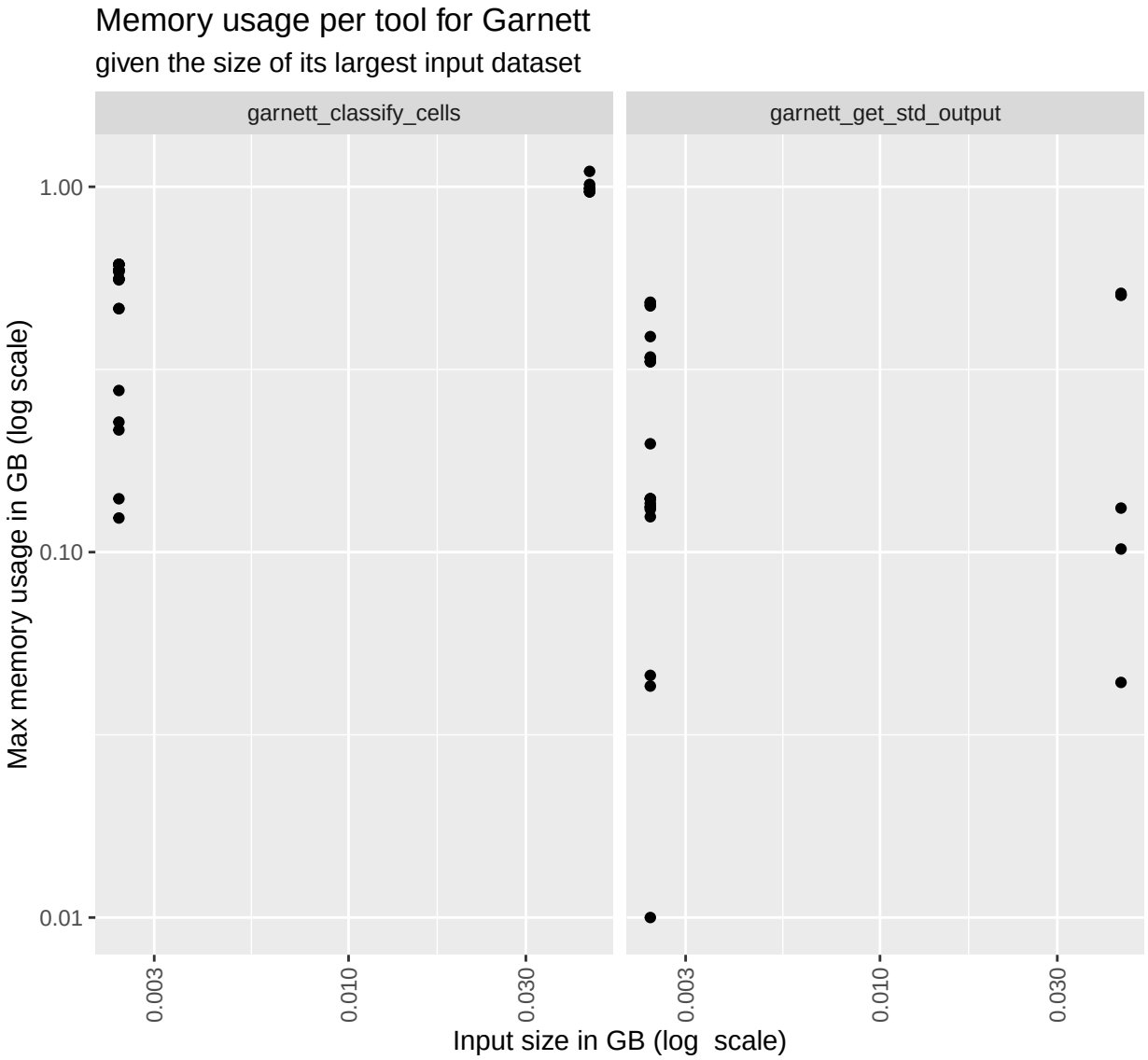

Figure 32: The EBI LSF dataset shows the relation between largest input size and max memory used by the Garnett tool set jobs.

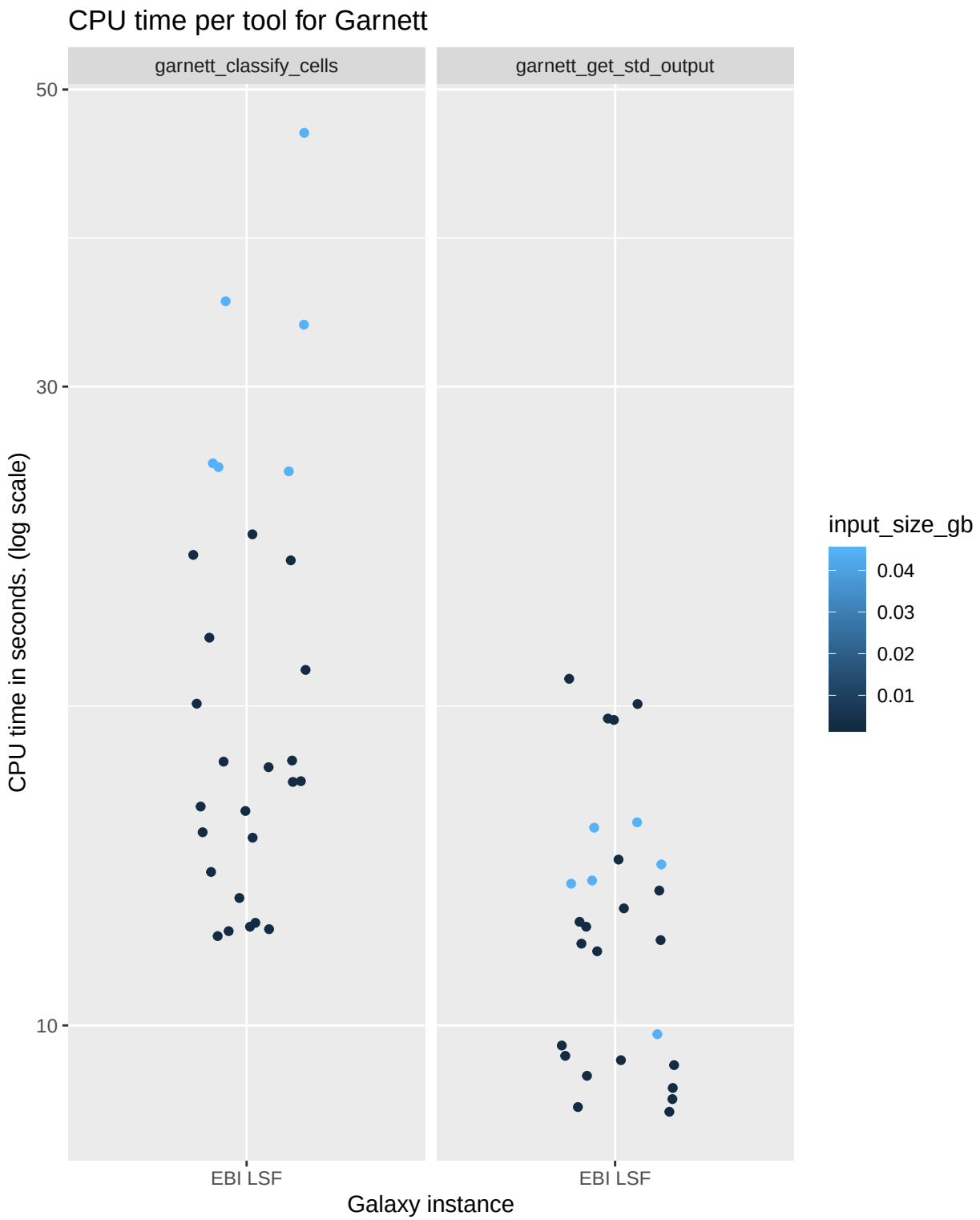

Figure 33: CPU time for Garnett tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). Consider that a job that uses multiple CPUs at the same time will have the time used for each CPU added. This plot reflects the processing power needed by the tool, not the time that the user waits for.

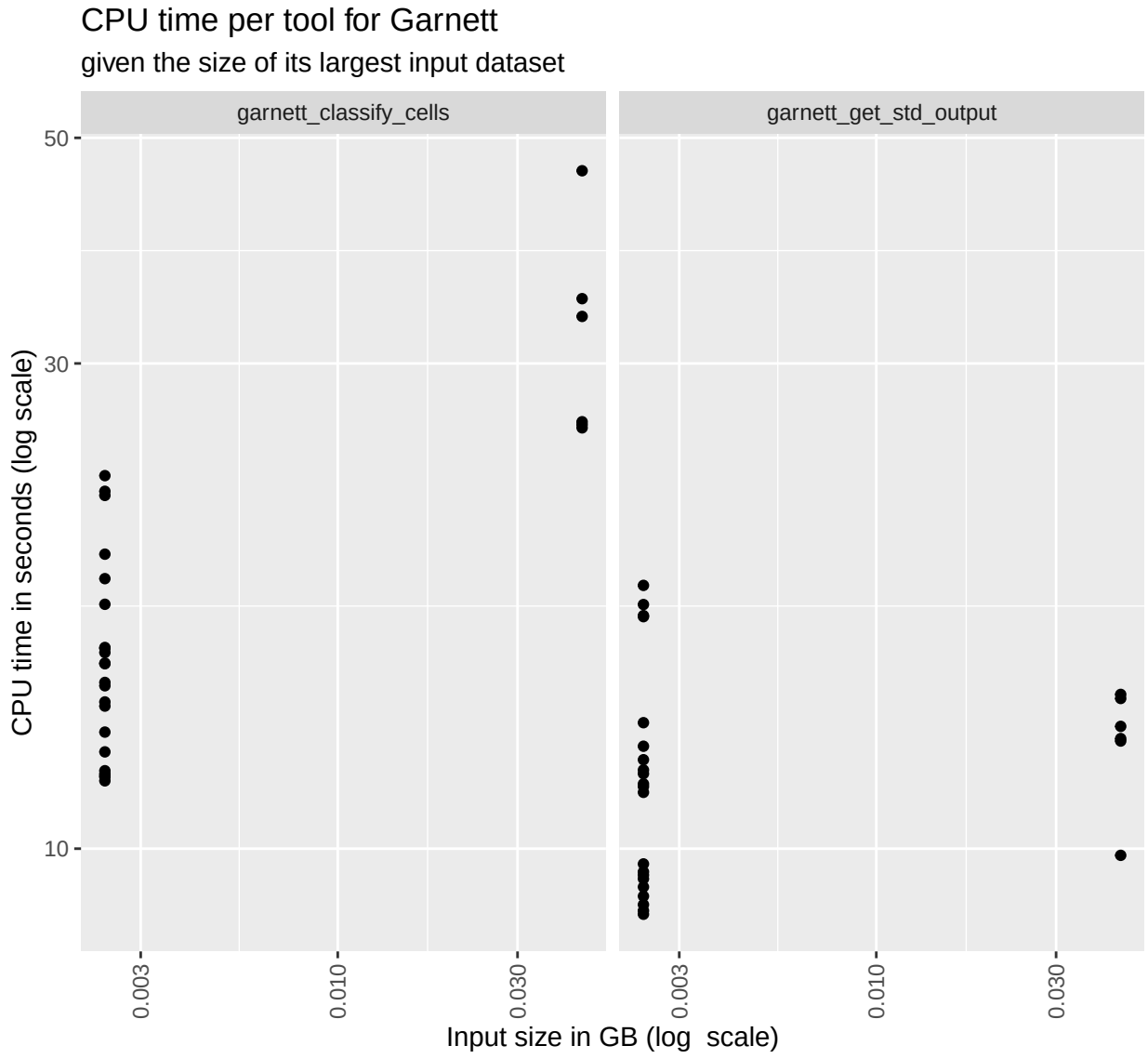

Figure 34: The EBI LSF dataset shows the relation between largest input size and CPU time used by the Garnett tool set jobs.

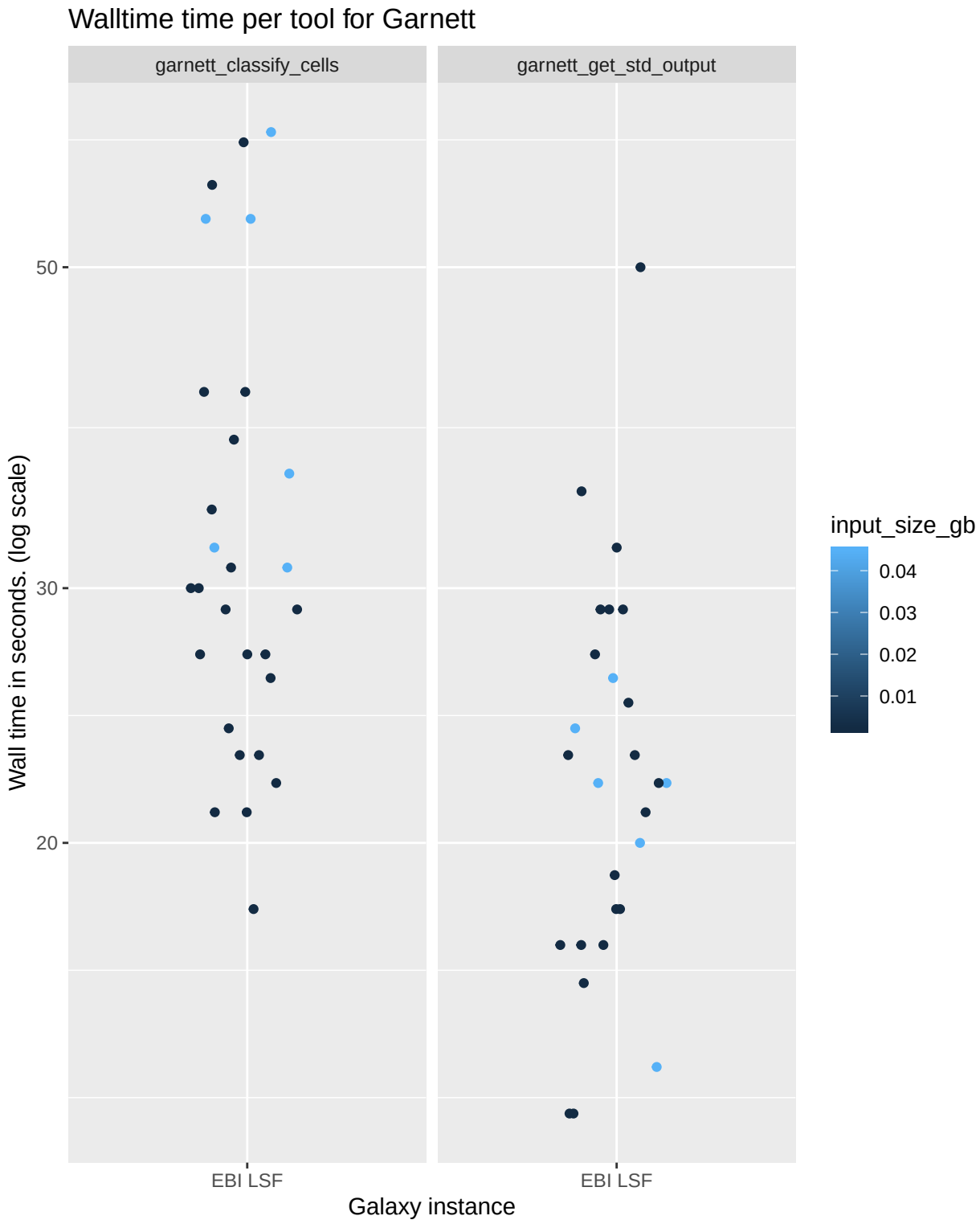

Figure 35: Wall time for Garnett tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). This metric indicates the time that the user waits for the job, and includes time used in processing (serial CPU time), IO waits, file copies, metadata collection, etc.

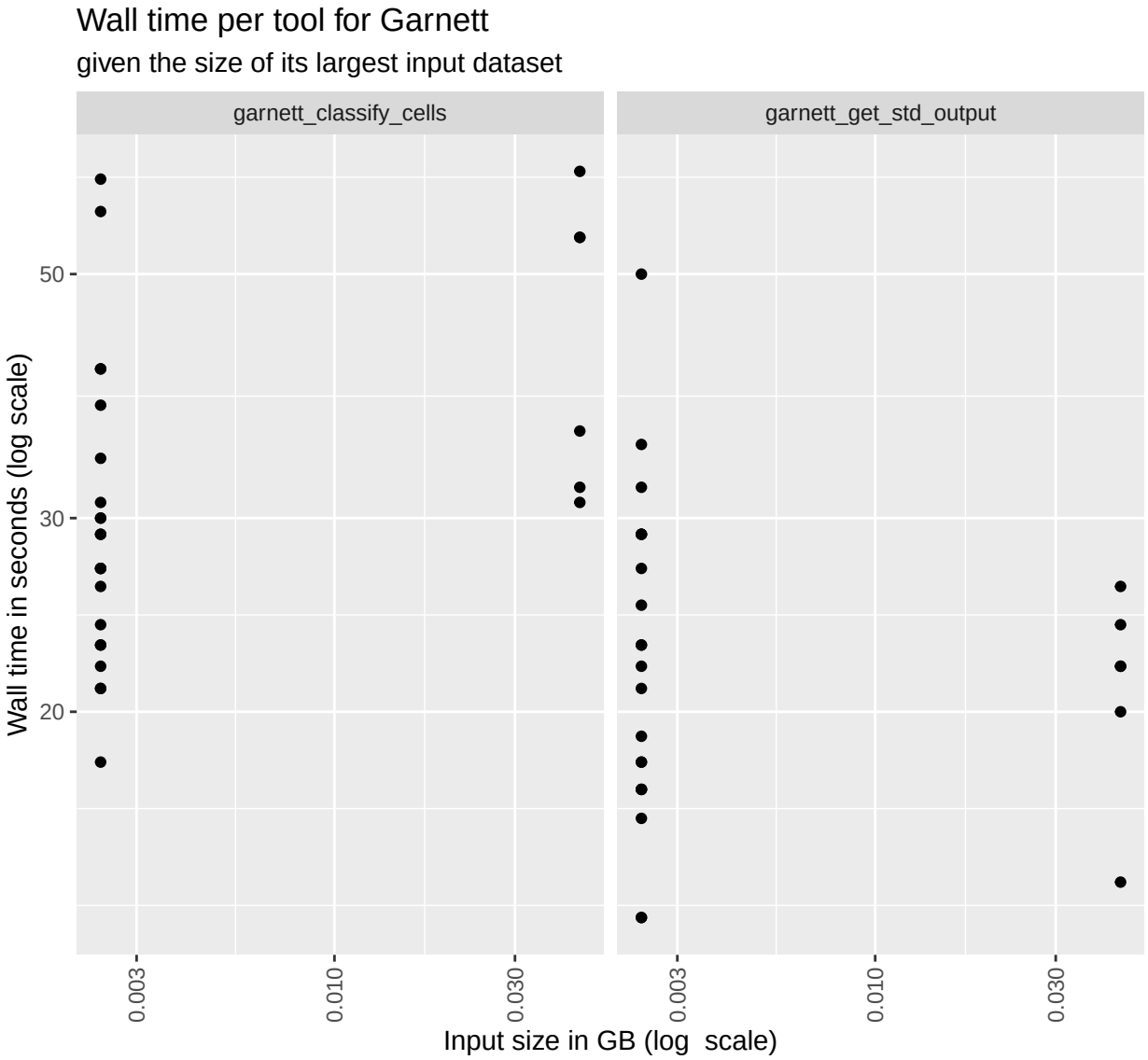

Figure 36: The EBI LSF dataset shows the relation between largest input size and Wall time used by the Garnett tool set jobs.

Figure 37: Largest input sizes distribution for SCpred jobs associated to ebi-gxa tools in both the HCA and the EBI LSF galaxy instances.

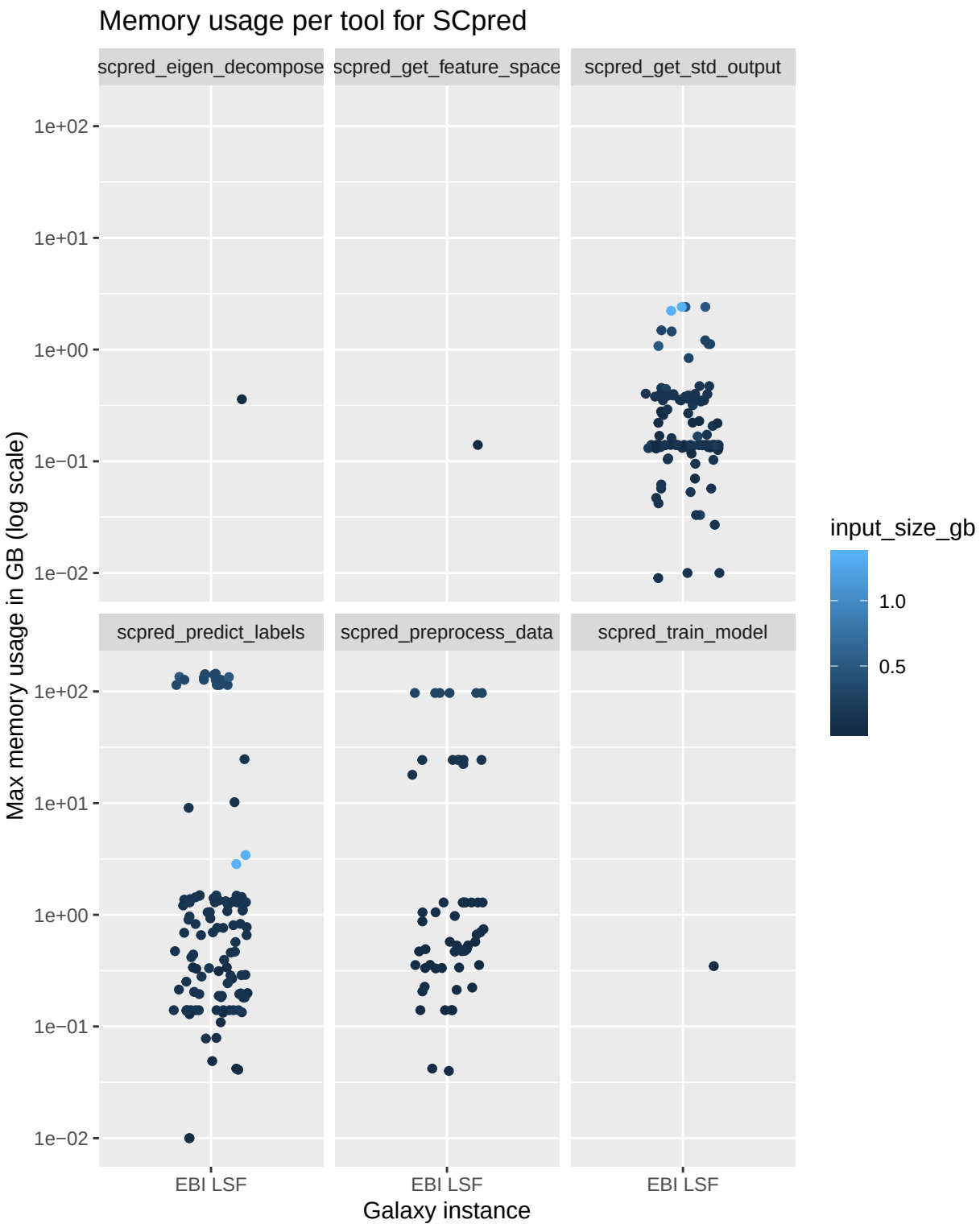

Figure 38: Memory consumption for SCpred tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). The plot only considers real memory, not swap.

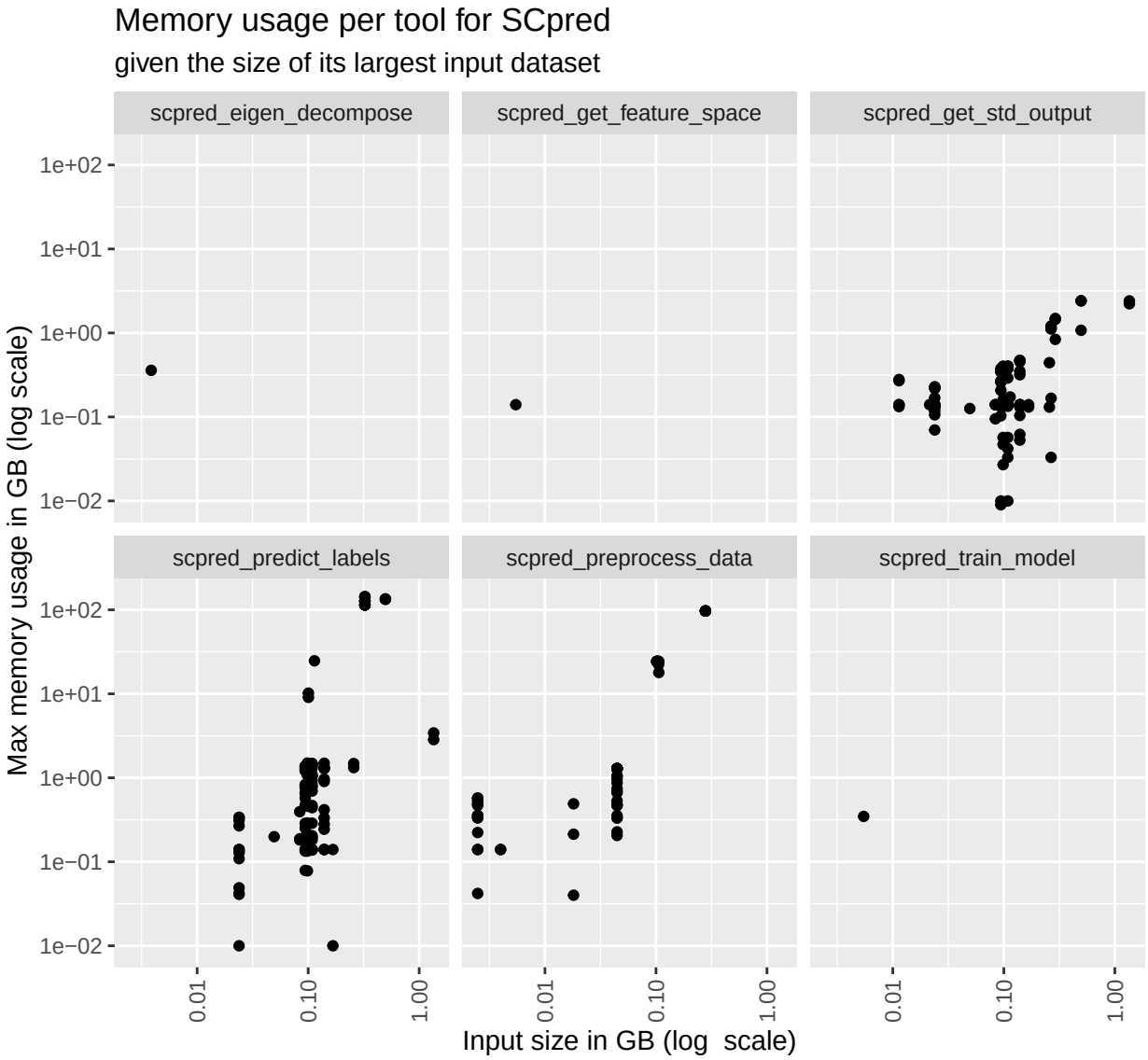

Figure 39: The EBI LSF dataset shows the relation between largest input size and max memory used by the SCpred tool set jobs.

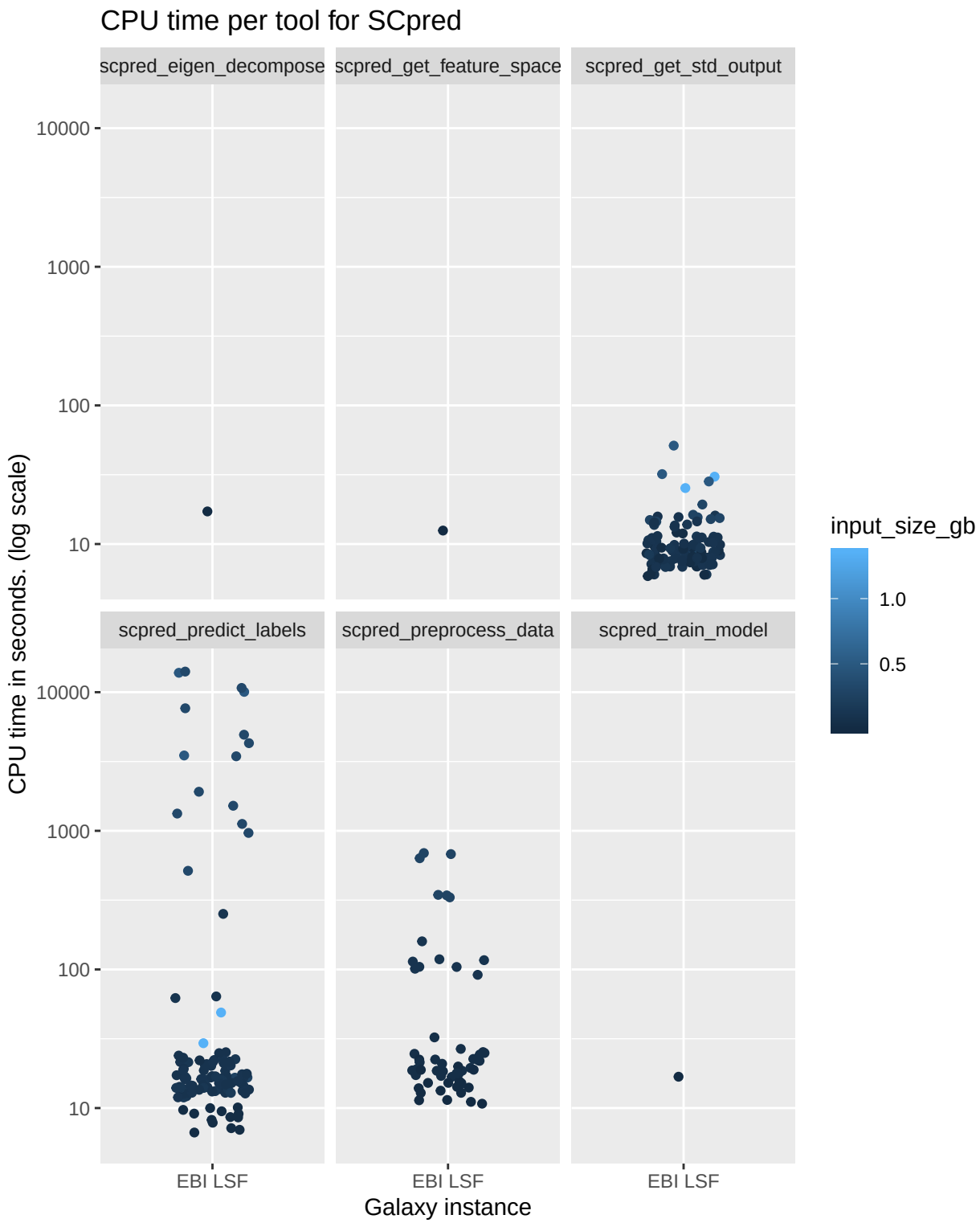

Figure 40: CPU time for SCpred tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). Consider that a job that uses multiple CPUs at the same time will have the time used for each CPU added. This plot reflects the processing power needed by the tool, not the time that the user waits for.

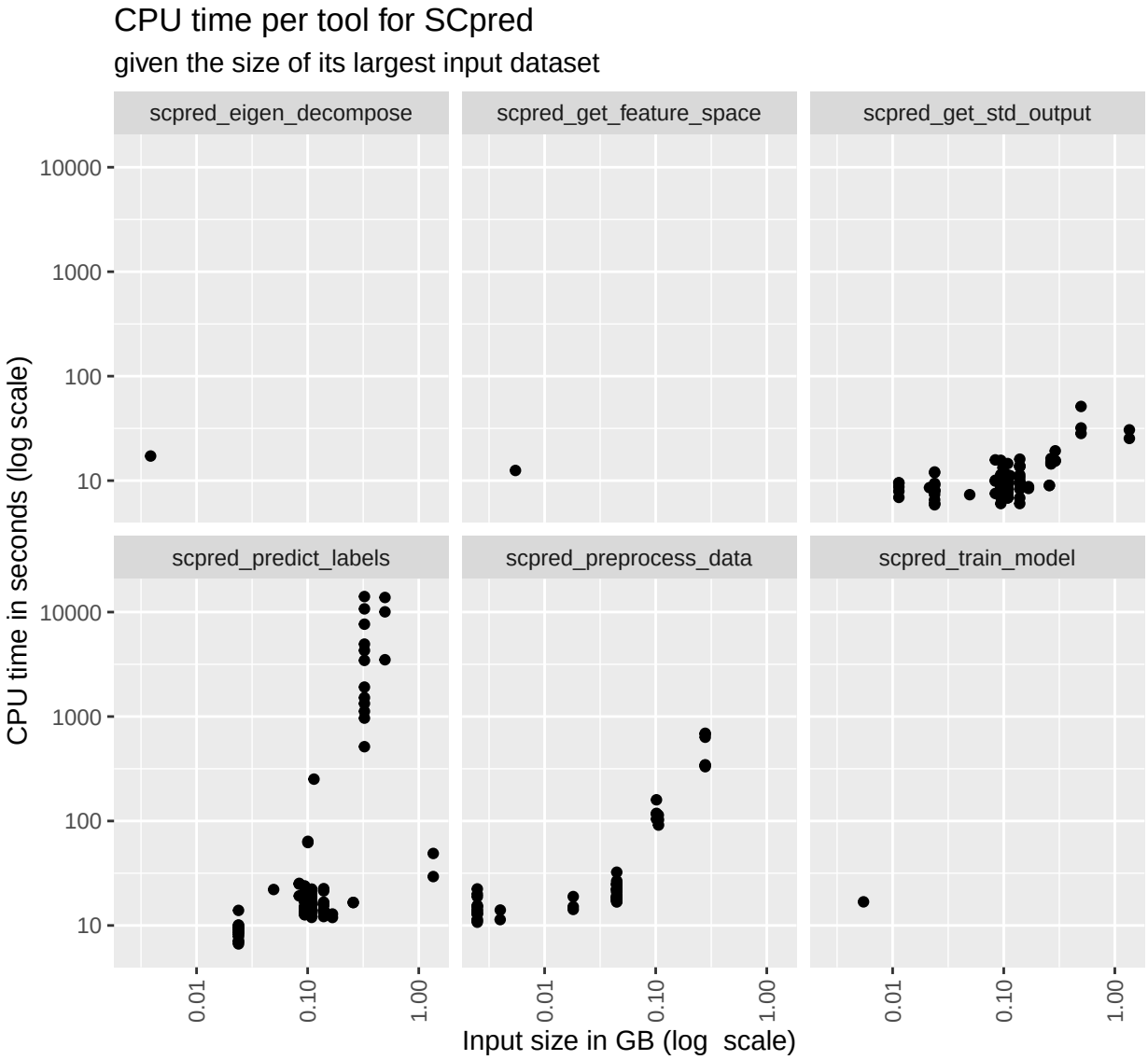

Figure 41: The EBI LSF dataset shows the relation between largest input size and CPU time used by the SCpred tool set jobs.

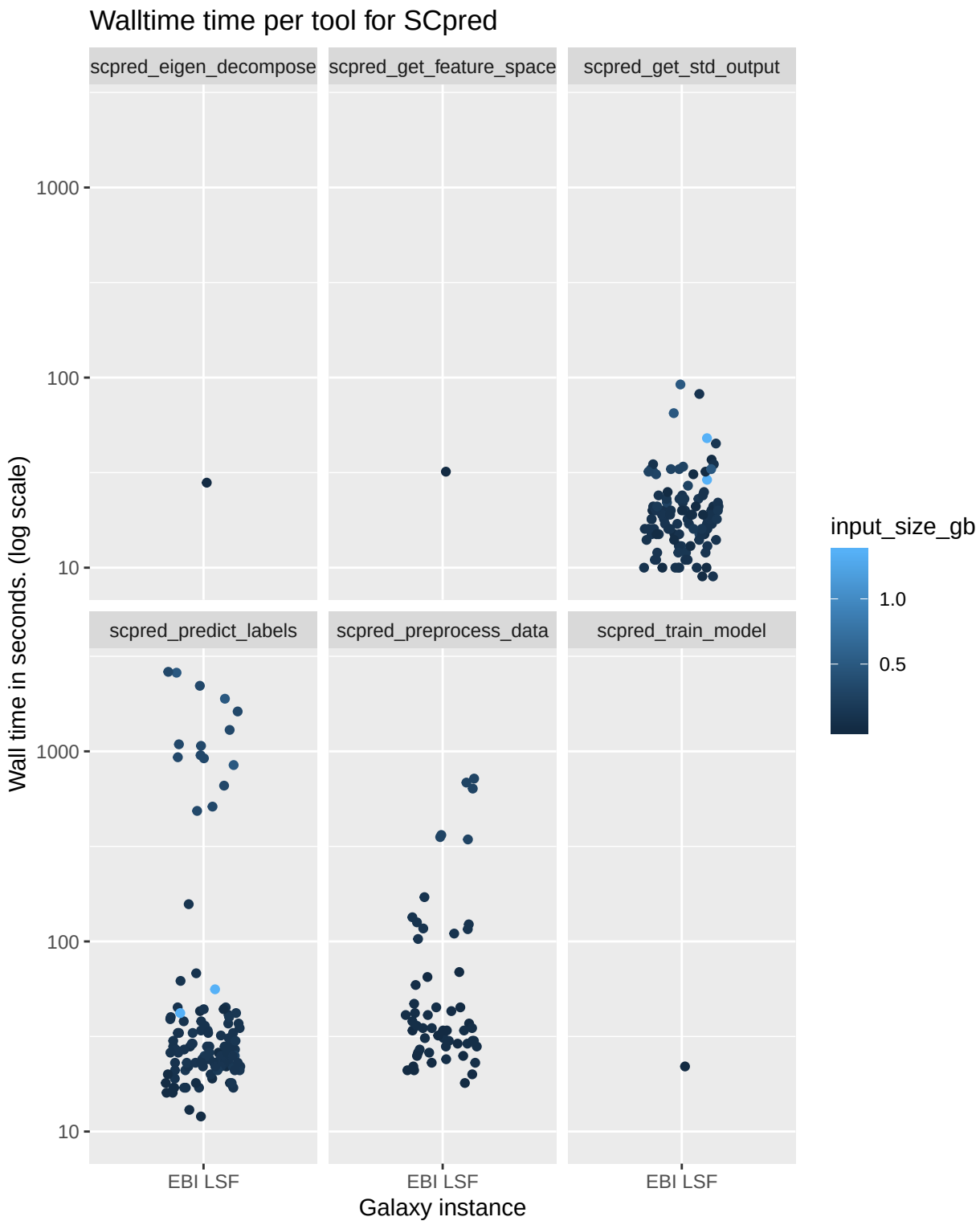

Figure 42: Wall time for SCpred tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). This metric indicates the time that the user waits for the job, and includes time used in processing (serial CPU time), IO waits, file copies, metadata collection, etc.

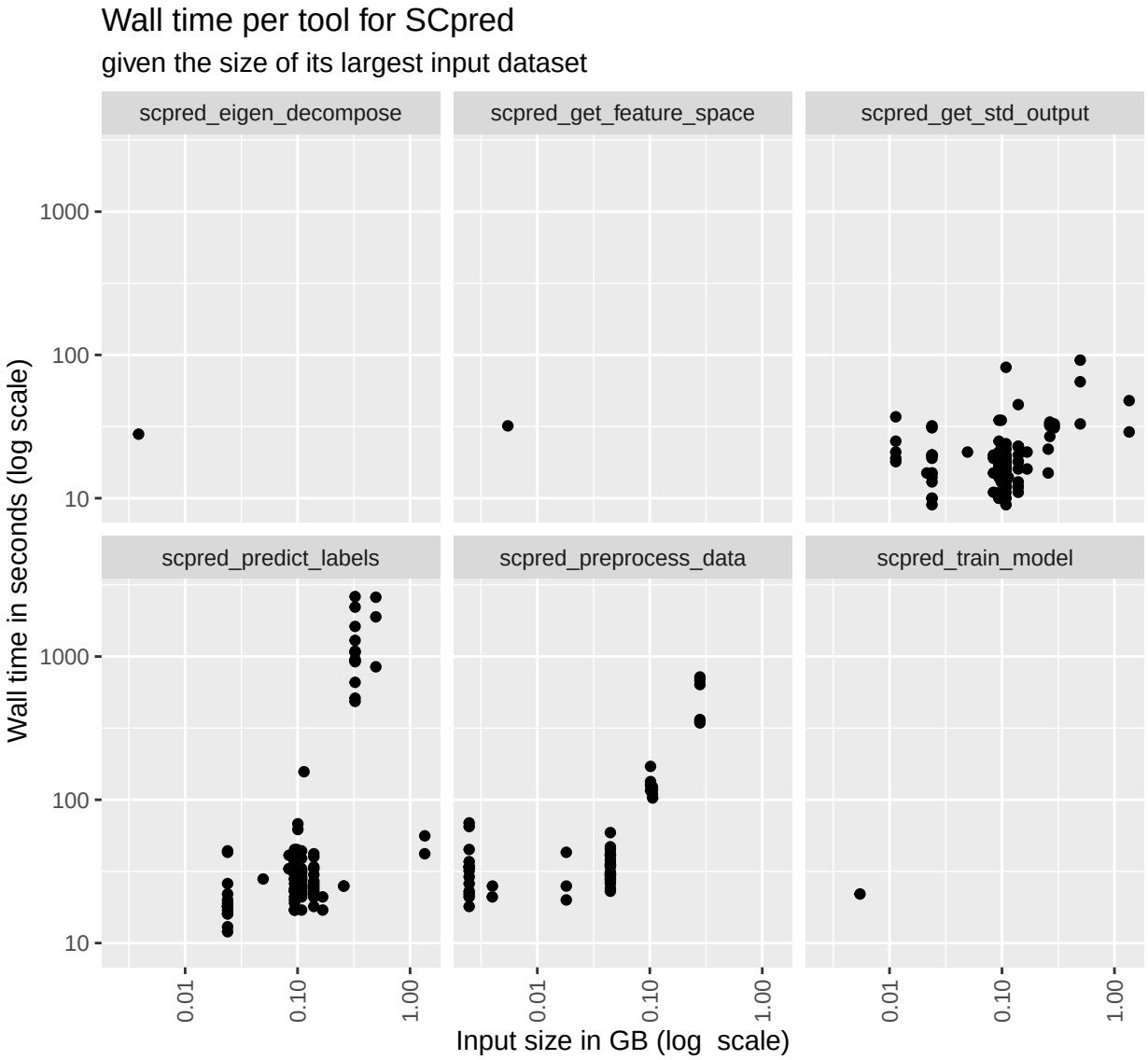

Figure 43: The EBI LSF dataset shows the relation between largest input size and Wall time used by the SCpred tool set jobs.

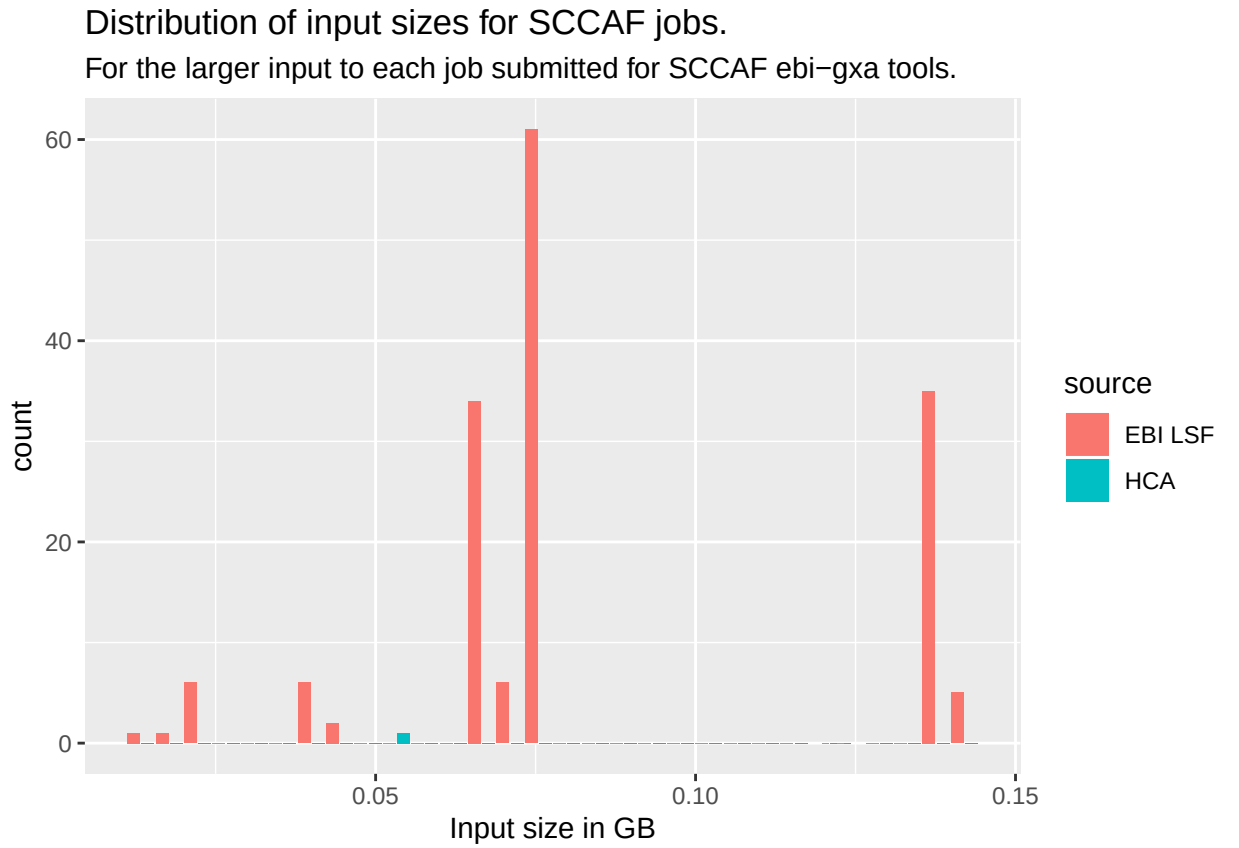

Figure 44: Largest input sizes distribution for SCCAF jobs associated to ebi-gxa tools in both the HCA and the EBI LSF galaxy instances.

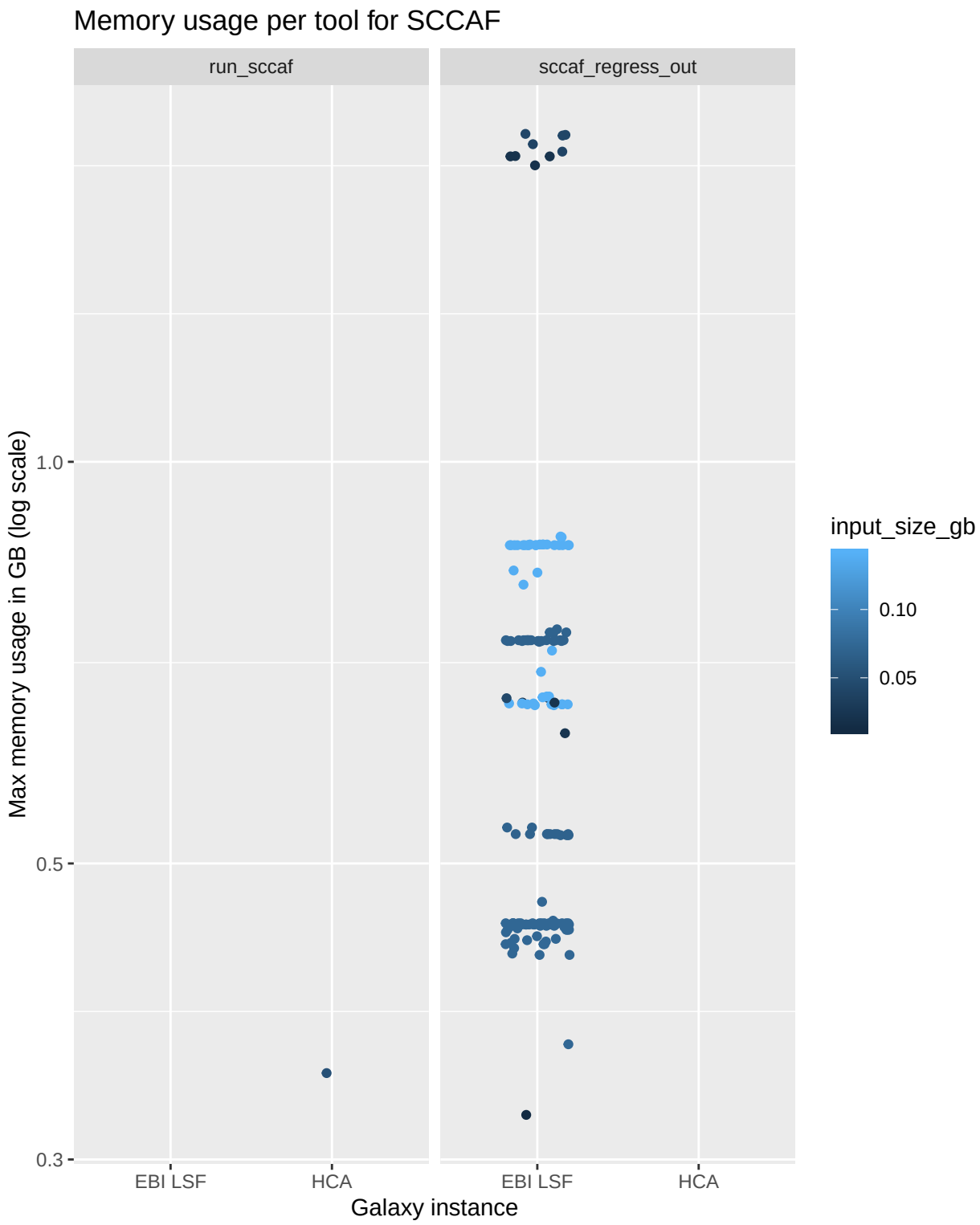

Figure 45: Memory consumption for SCCAF tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). The plot only considers real memory, not swap.

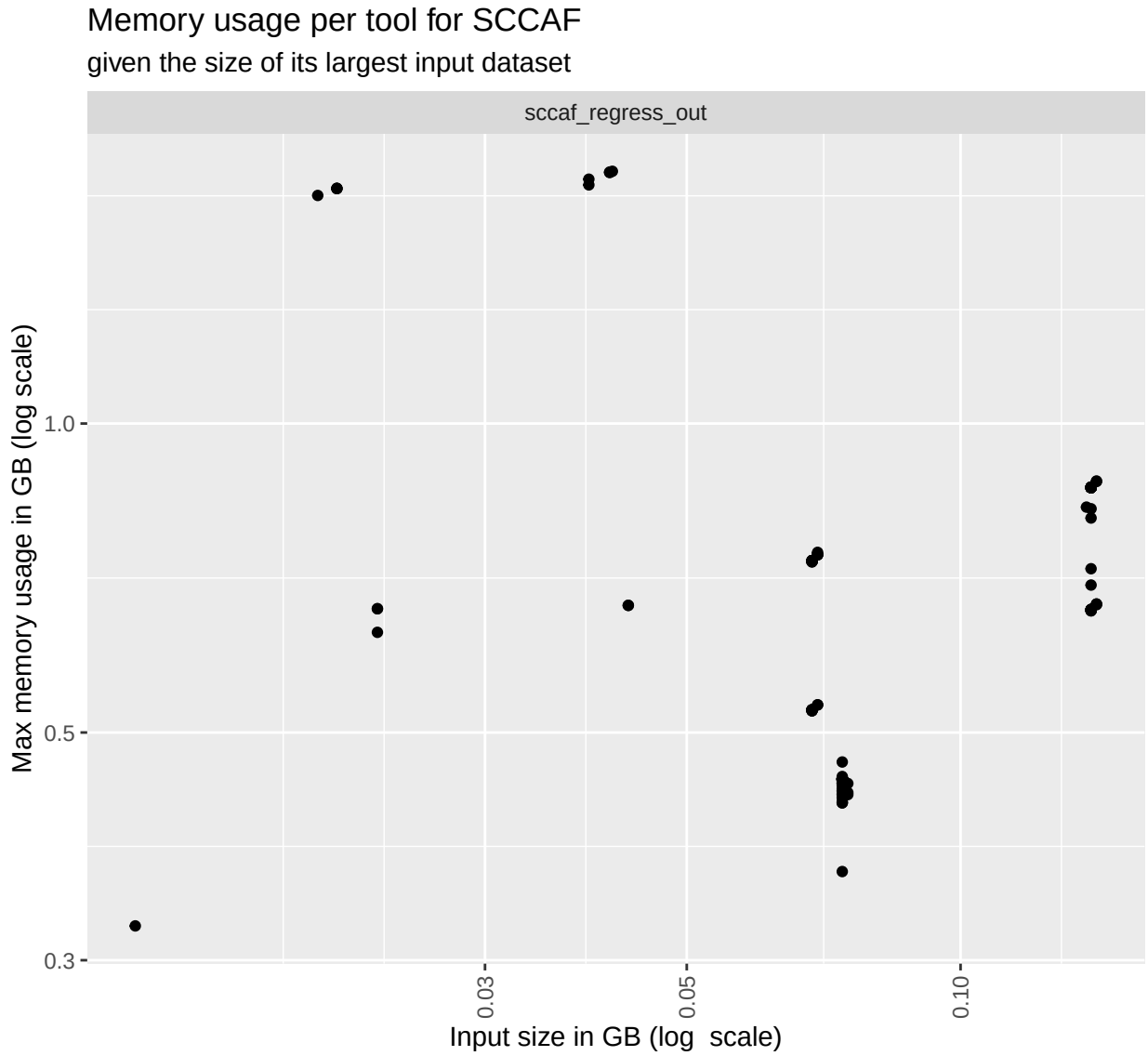

Figure 46: The EBI LSF dataset shows the relation between largest input size and max memory used by the SCCAF tool set jobs.

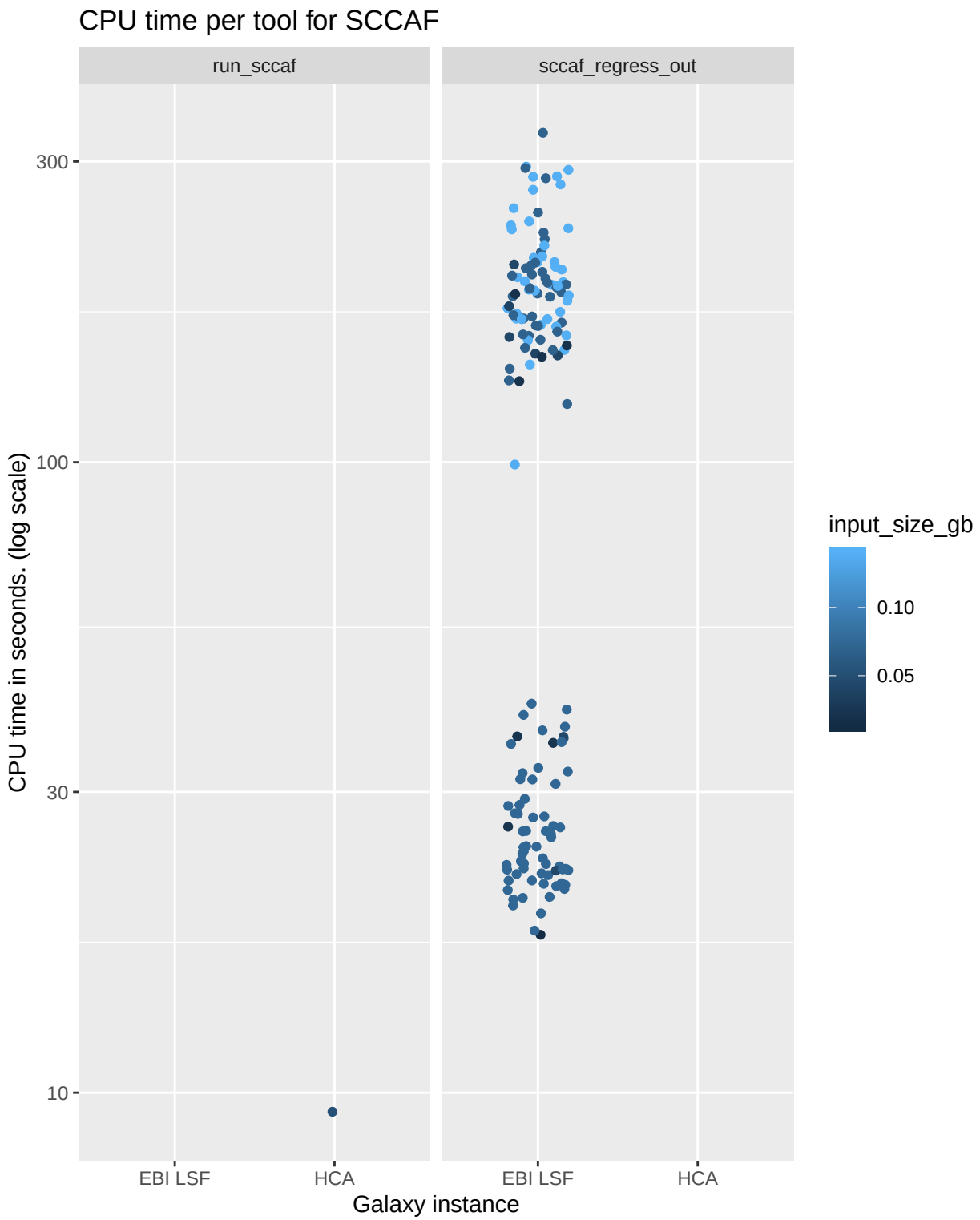

Figure 47: CPU time for SCCAF tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). Consider that a job that uses multiple CPUs at the same time will have the time used for each CPU added. This plot reflects the processing power needed by the tool, not the time that the user waits for.

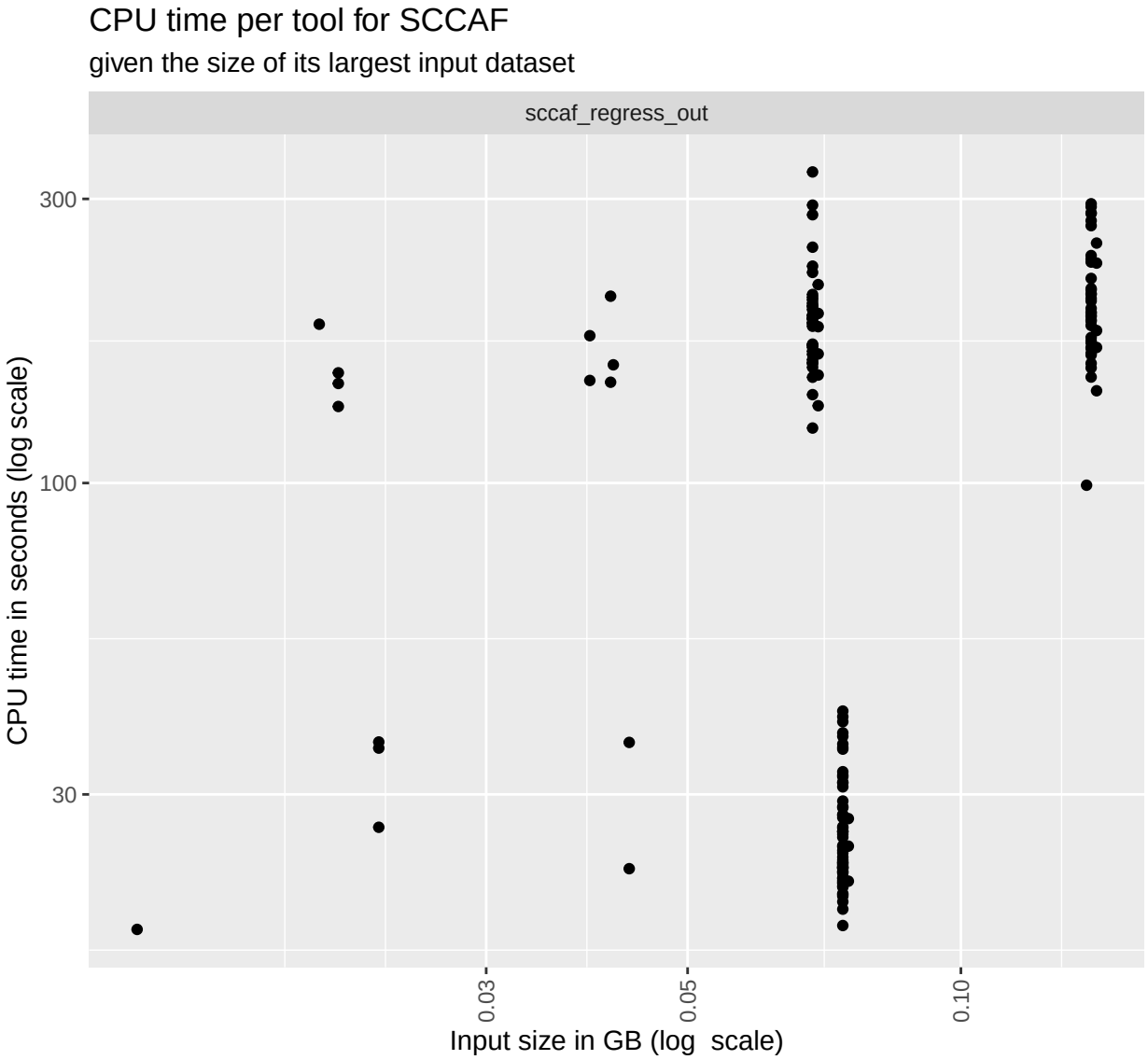

Figure 48: The EBI LSF dataset shows the relation between largest input size and CPU time used by the SCCAF tool set jobs.

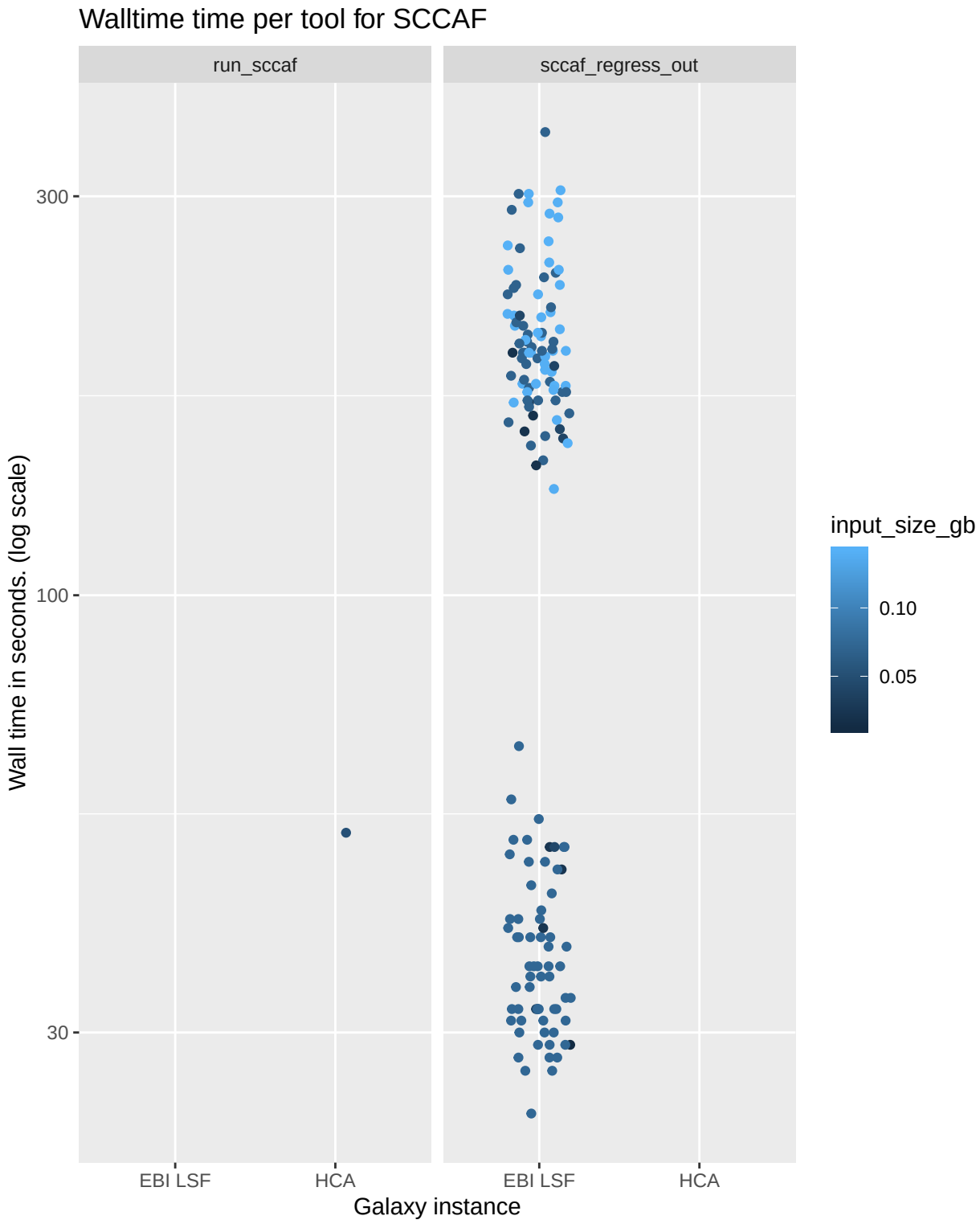

Figure 49: Wall time for SCCAF tool-set tool jobs in the Human Cell Atlas (HCA) instance and in the internal EBI LSF Galaxy instance. EBI LSF has more jobs because they comprise executions from all users (including production pipelines and two training sessions). This metric indicates the time that the user waits for the job, and includes time used in processing (serial CPU time), IO waits, file copies, metadata collection, etc.

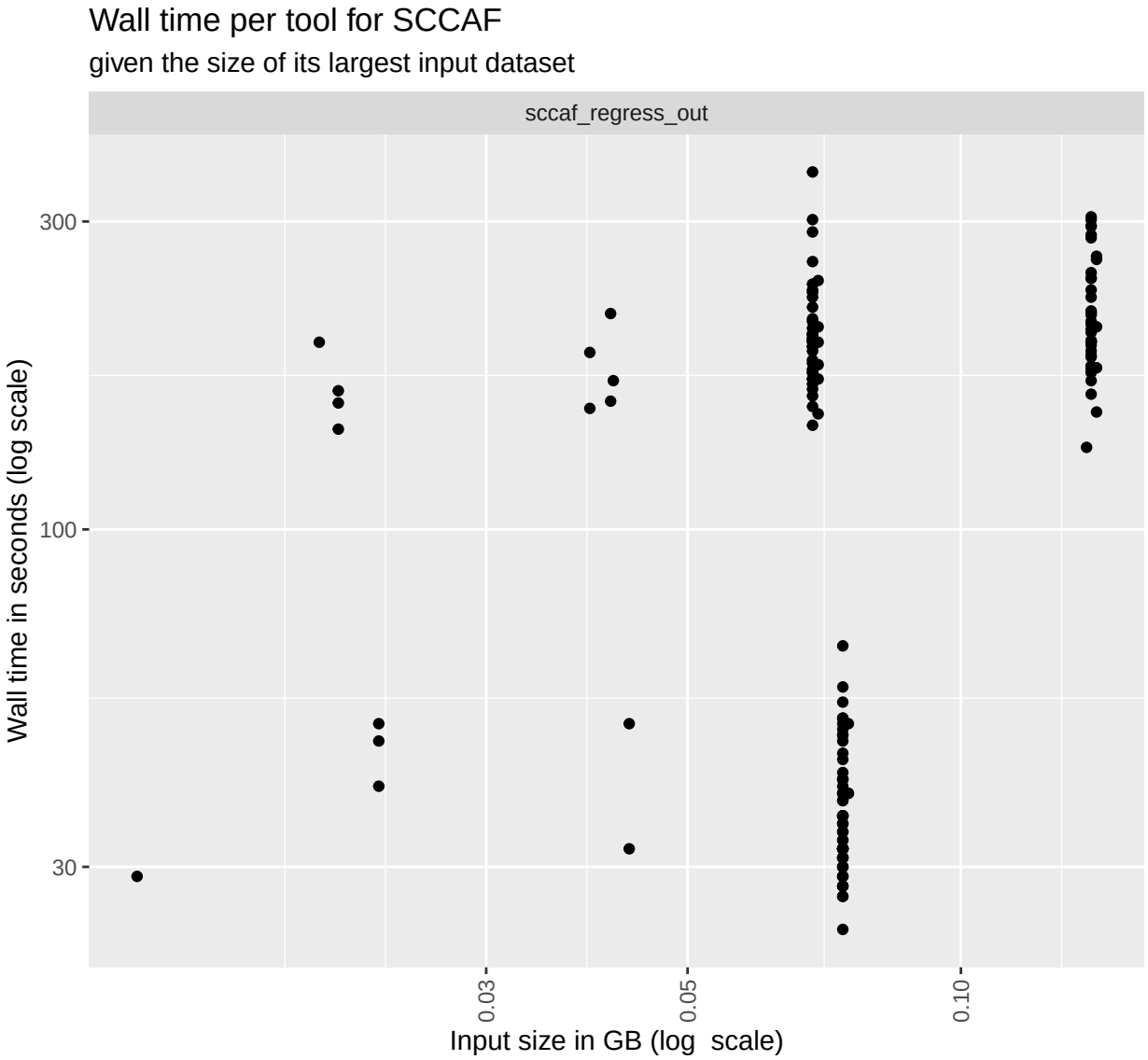

Figure 50: The EBI LSF dataset shows the relation between largest input size and Wall time used by the SCCAF tool set jobs.