Supplementary Material

A cloud-based workflow to quantify transcript-expression levels in public cancer compendia

PJ Tatlow[1], Stephen R. Piccolo[1,2,*]

1 - Department of Biology, Brigham Young University, Provo, Utah, USA

2 - Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA

* - Correspondence should be addressed to S. R. P. (801-422-7116; stephen_piccolo@byu.edu).

Supplementary Figures

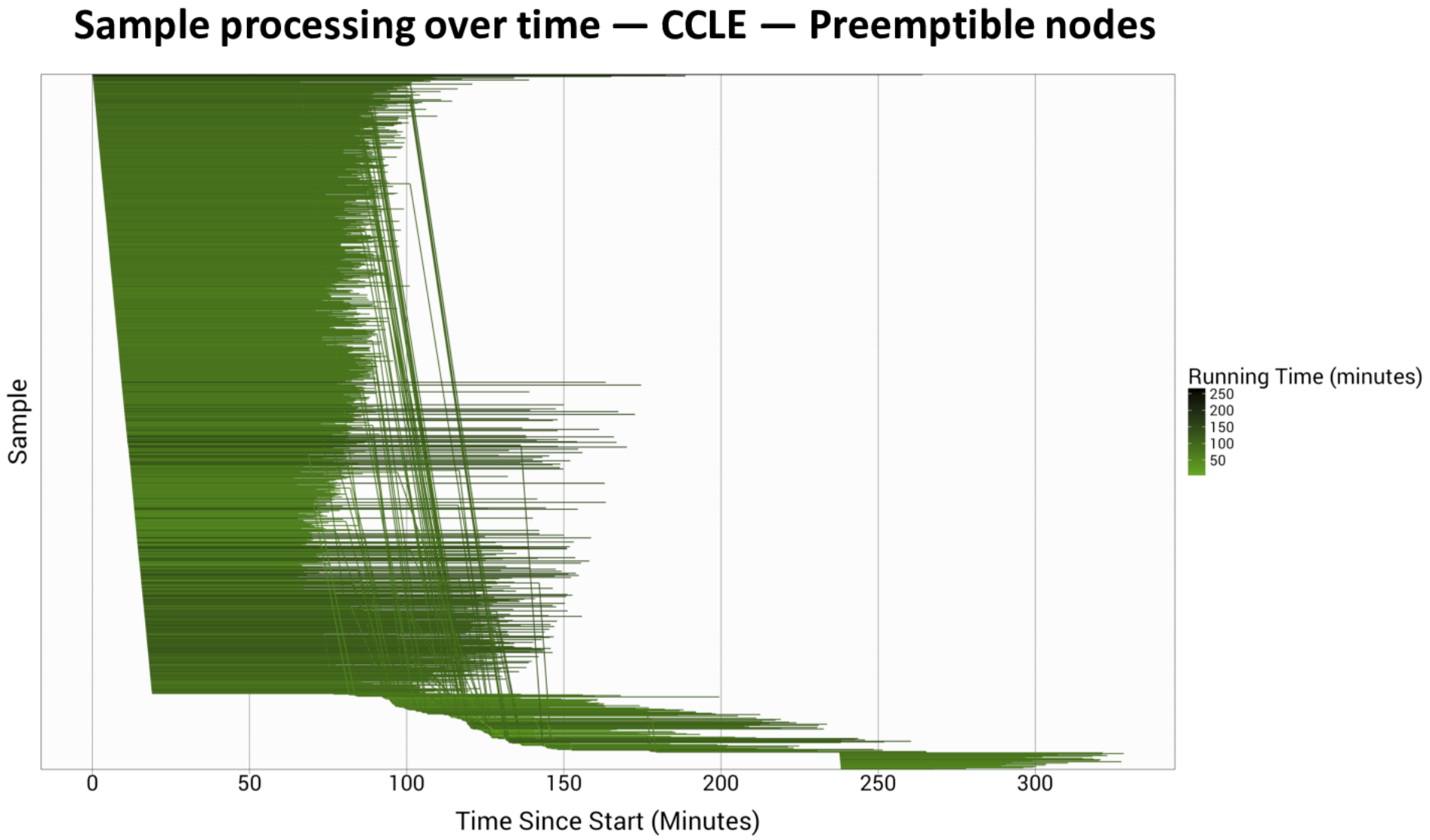


Figure S1. Processing time per CCLE sample using the preemptible-node configuration. The 934 CCLE samples were processed using a variable number of preemptible virtual machines. The horizontal lines represent the relative start and stop times at which each sample was processed. Darker lines identify samples that took longer to process. Vertical lines indicate times at which samples were preempted and then resubmitted for processing. In total, 78 preemptions occurred.

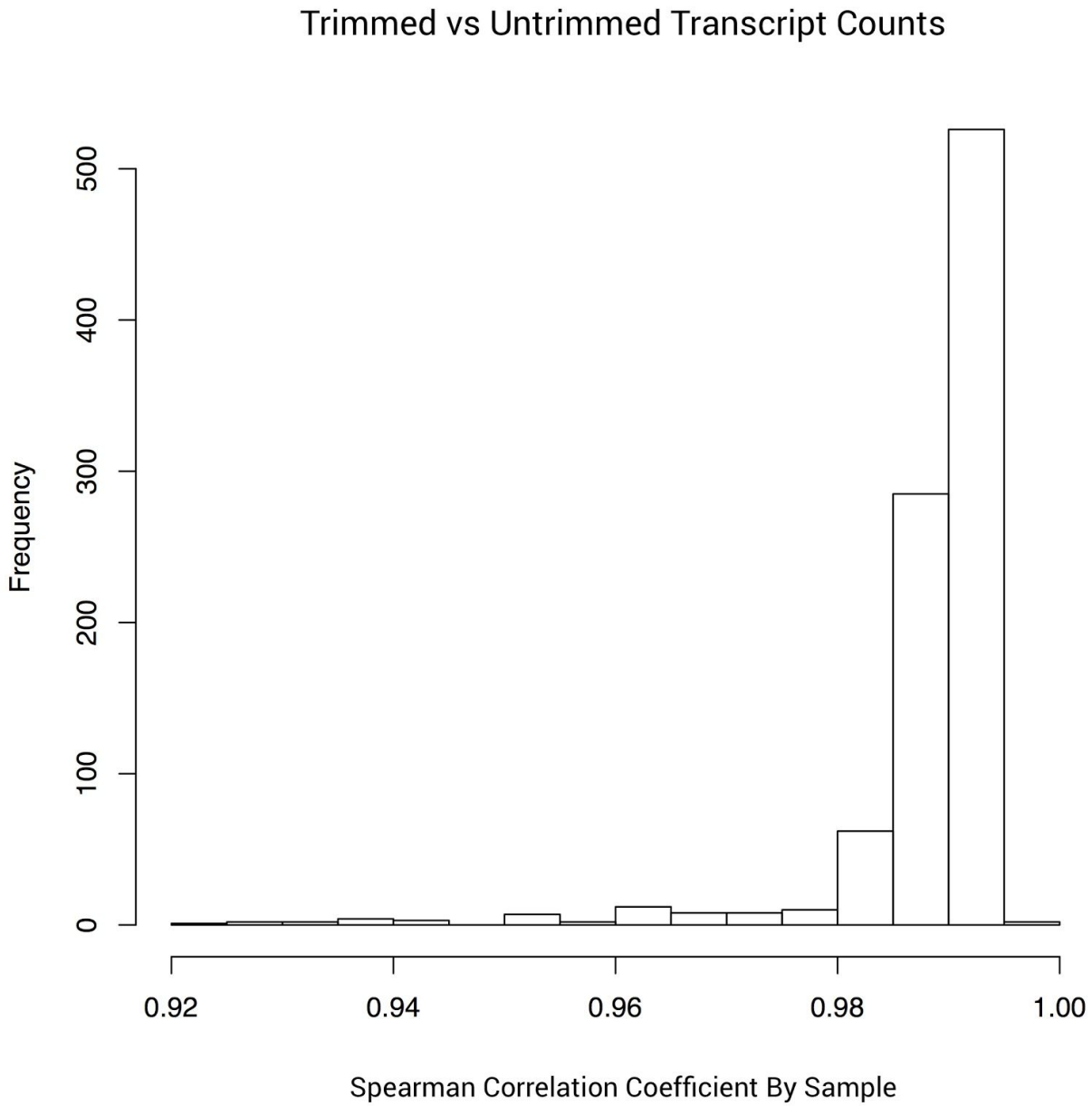


Figure S2. Histogram of Spearman correlation coefficients comparing read counts for trimmed vs. untrimmed CCLE data.

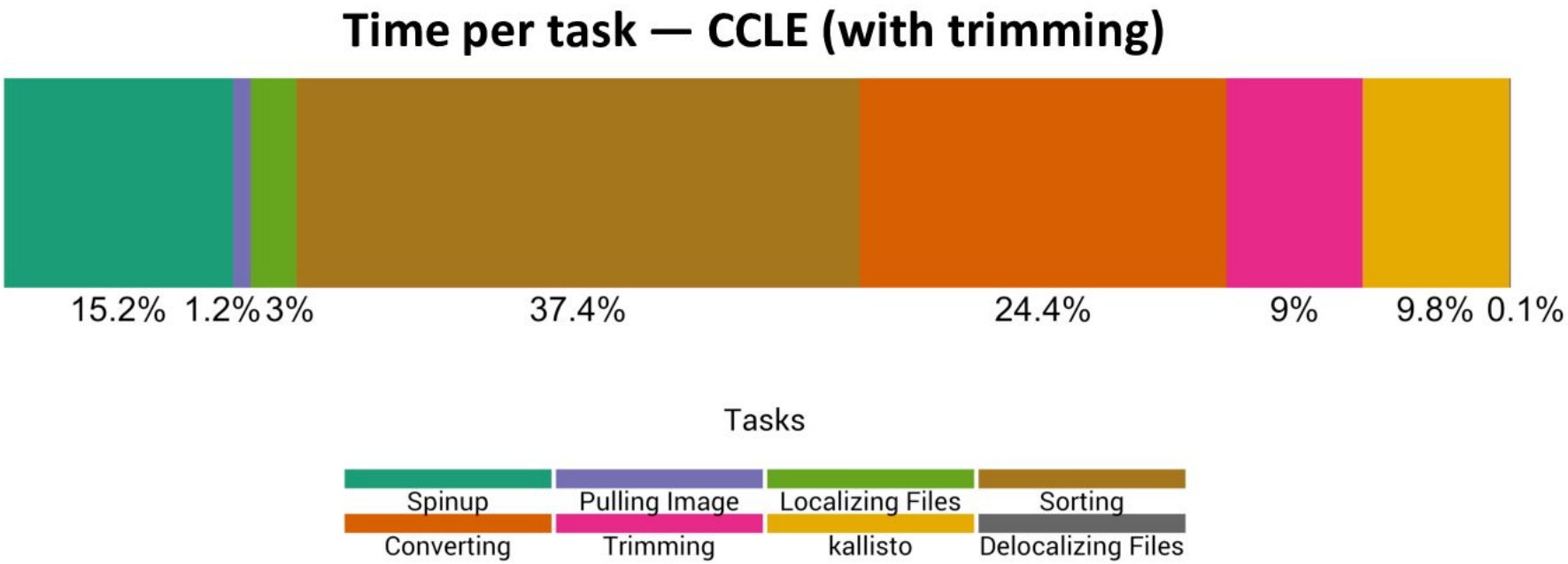
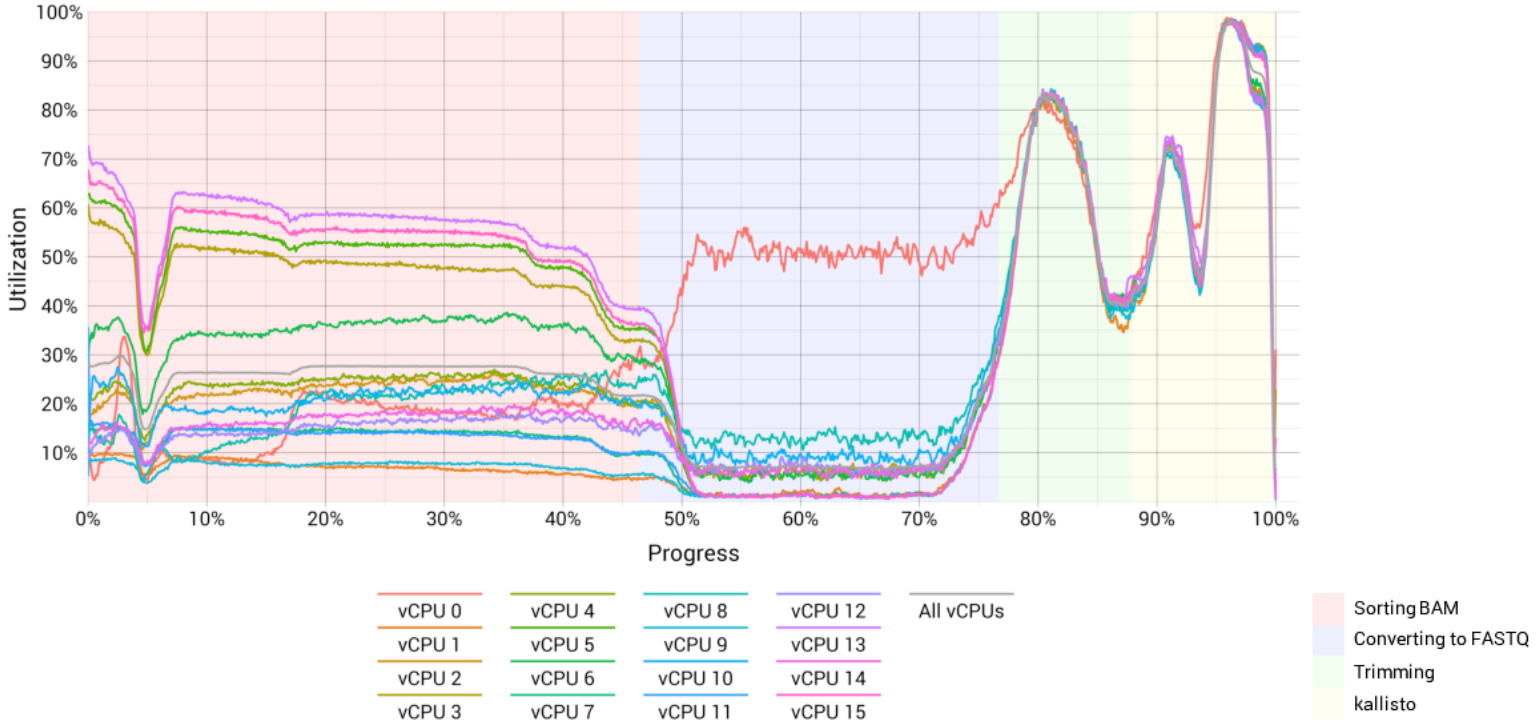
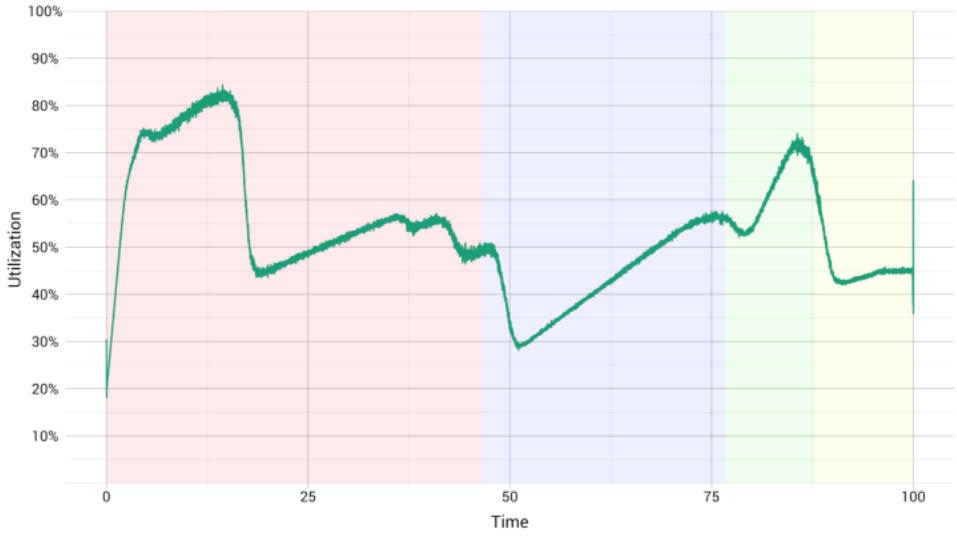


Figure S3. Relative time spent on computational tasks for CCLE samples using the preemptible-node configuration. We logged the durations of individual processing tasks for the CCLE samples, averaged these values, and calculated the percentage of overall processing time for each task. The “spinup,” image pulling, and file localization steps enabled the virtual machines to begin executing. The “spinup,” image pulling, and file localization steps enabled the virtual machines to begin executing. For sample preprocessing, the BAM files were sorted, converted to FASTQ format, and trimmed for quality; these steps took 61.8% of the overall processing time. The *kallisto* alignment and quantification steps took only 9.8% of the overall processing time.

Virtual Central Processing Units (a)



Random Access Memory (b)



Disk Storage (c)

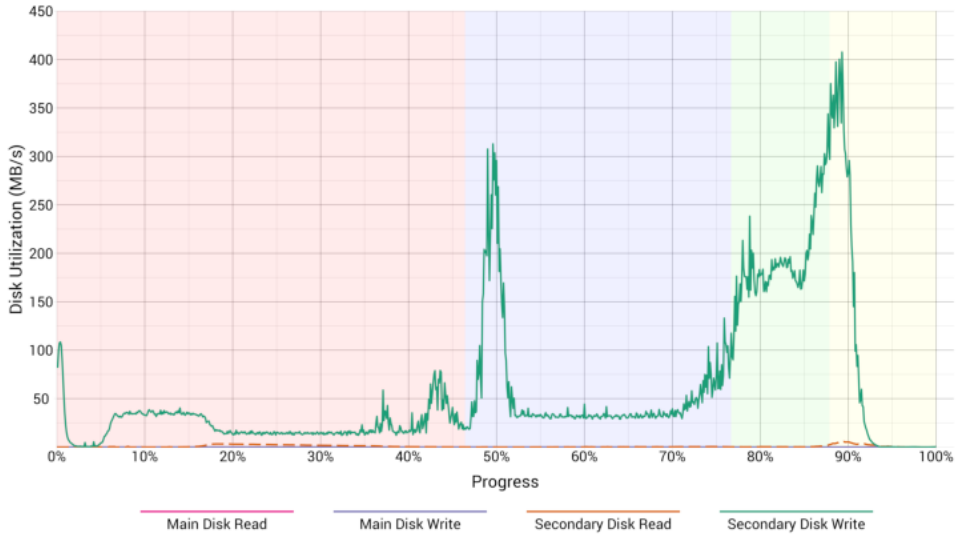


Figure S4. Computational resource utilization while CCLE samples were processed using a preemptible-node configuration. These graphs show the (a) percentage of user and system vCPU utilization, (b) percentage of memory usage, and (c) disk activity. The “main” disks had only 10 gigabytes of storage space and stored operating-system files. The “secondary” disks, which stored all data files, had 350 gigabytes of space. The background colors represent the computational tasks shown in Figure S3 and correspond with expected resource utilization for these tasks. (We were unable to collect performance metrics for preliminary tasks, such as file localization, because these tasks were not performed within the software container.) Each graph summarizes data from all 934 CCLE samples.