American Society of Plant Biologists | SEB SOCIETY FOR EXPERIMENTAL BIOLOGY | WILEY

CoGe LoadExp+: A web-based suite that integrates next-generation sequencing data analysis workflows and visualization

Jeffrey W. Grover[1] | Matthew Bomhoff[2,3] | Sean Davey[2,3] | Brian D. Gregory[4] | Rebecca A. Mosher[1,2,3] | Eric Lyons[2,3]

[1]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ, USA

[2]BIO5 Institute, University of Arizona, Tucson, AZ, USA

[3]School of Plant Sciences, University of Arizona, Tucson, AZ, USA

[4]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

Correspondence
Rebecca A. Mosher and Eric Lyons, BIO5 Institute, University of Arizona, Tucson, AZ, USA.
Emails: ericlyons@email.arizona.edu, rmosher@email.arizona.edu

Summary

To make genomic and epigenomic analyses more widely available to the biological research community, we have created LoadExp+, a suite of bioinformatics workflows integrated with the web-based comparative genomics platform, CoGe. LoadExp+ allows users to perform transcriptomic (RNA-seq), epigenomic (bisulfite-seq), chromatin-binding (ChIP-seq), variant identification (SNPs), and population genetics analyses against any genome in CoGe, including genomes integrated by users themselves. Through LoadExp+'s integration with CoGe's existing features, all analyses are available for visualization and additional downstream processing, and are available for export to CyVerse's data management and analysis platforms. LoadExp+ provides easy-to-use functionality to manage genomics and epigenomics data throughout its entire lifecycle using a publicly available web-based platform and facilitates greater accessibility of genomics analyses to researchers of all skill levels. LoadExp+ can be accessed at https://genomevolution.org.

KEYWORDS
bisulfite sequencing, ChIP-seq, DNA methylation, epigenomics, expression analysis, genomics, next-generation sequencing, population genetics, RNA-seq, SNP identification

1 | INTRODUCTION

As advanced next-generation sequencing (NGS) technologies become more powerful and affordable, a growing number of researchers find themselves in a position to answer genome-scale biological questions using both existing and newly generated data. However, despite rapid technological advances, the knowledge and computational resources needed to analyze and interpret multiple large genomic datasets create bottlenecks for scientific discovery. While initiatives exist to provide high-performance computing resources (e.g., CyVerse, XSEDE) (Merchant et al., 2016; Towns et al., 2014) to life

scientists, these resources require computational expertise to fully exploit. Researchers in the life sciences without this expertise would benefit from a platform focused on NGS data and genomics analyses that integrates data management, analysis, and visualization tools into a single user-friendly interface on a publicly available web-accessible server. Such a platform also benefits more advanced users by simplifying the process of analyzing, visualizing, managing, and distributing data with collaborators.

We have addressed these needs with the creation of LoadExp+, an addition to the comparative genomics platform, CoGe (https://genomevolution.org) (Lyons & Freeling, 2008). LoadExp+ is a web

portal through which numerous genomic and epigenomic analyses can be conducted. These analyses include RNA-seq, whole-genome bisulfite sequencing (BS-seq), ChIP-seq, SNP identification, and population genetics calculations. The collection of these tools in one location, integration with an advanced genome browser, and the use of CoGe's user-friendly interface present advantages over other web-based bioinformatics platforms for the life sciences. For advanced users, there is also an REST API available for programmatic access to data integrated through LoadExp+ (https://goo.gl/Pf4xjf).

2 | RESULTS

2.1 | Web interface

LoadExp+ is accessible from CoGe's main page from the menu bar under "Tools" or the User Data page by clicking "Load Experimental Data." It is intuitive to navigate, with mouse-over descriptions and conspicuous links to documentation for each tool. Users begin their experimental setup by selecting data to load from their local computer, specifying a web address (HTTP or FTP), inputting NCBI Short Read Archive (SRA) accession numbers, or selecting files from the user's CyVerse Data Store (Figure 1). Options for LoadExp+'s integrated programs, otherwise accessible as command line arguments, are then presented as checkboxes or entry fields. When setting up their desired analysis, users must provide minimal metadata by naming and describing their experiment, providing version information, and indicating the source of their data.

Before submitting their desired analysis, users can choose to have the resulting files automatically organized into a new or existing notebook (Figure S1 and S2), and have an email sent to them when their analysis is complete. After submitting their analysis, users are automatically presented with a window in which to monitor its progress as well as a persistent link that they can use to return to monitor the workflow's progress. Users can also navigate to their "My Data" page

to check the progress of submitted LoadExp+ workflows under "Analyses." Results from the user's analysis are available for viewing and further analysis in the EPIC-CoGe genome browser (see below). All files, including intermediate files, generated from the analysis can also be downloaded from the Experiment View page (Figure S3 and S4) or exported to the CyVerse Data Store. These files are also available through a RESTful API following completion of the run.

2.2 | Supported data formats and metadata attribution

LoadExp+ supports the input of data in a variety of formats, including raw sequencing reads (FASTQ, SRA), compressed data (.gz, .bz2, .zip), or preprocessed data in the form of alignments (BAM, SAM), polymorphism data (VCF, GVCF), and quantitative data (CSV/TSV, BED/BEDGraph, GFF/GTF, WIG, BigWig). These data can be uploaded directly from the user's computer, retrieved from a remote server (FTP/HTTP), imported from their CyVerse Data Store (recommended for files > 1GB), or retrieved from NCBI's Short Read Archive by accession or project number. Transferring data between CoGe and the CyVerse Data Store uses iRODS (Hedges, Hasan, & Blanke, 2007) for secure, high-performance, parallel file transfers.

Users are prompted to supply metadata which are captured for every experiment, including the name, description, type, and source of the data. Additionally, the analysis program(s), parameters, and the options chosen at run-time are captured automatically and are immutable. After an experiment has been loaded, researchers can add additional metadata including text-based descriptions, web links, and images (Figure S4).

2.3 | Integrated analysis workflows

LoadExp+ is a unified platform encompassing a variety of popular and powerful genomic and epigenomic analyses. The following NGS
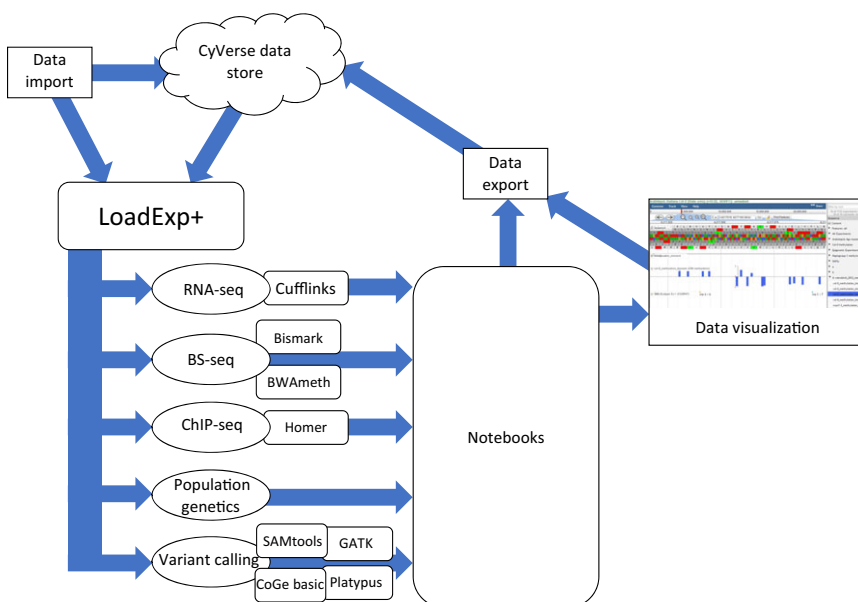

FIGURE 1 Schematic Representation of LoadExp+ Workflow. Data to be used in any workflow may be imported directly from a user's local machine, a remote address by HTTP/FTP, from SRA accession number, or from the CyVerse Data Store. After import, LoadExp+ allows users to choose which bioinformatics workflows to run. After the desired analyses have finished, experimental data can be organized into "Notebooks" and viewed in CoGe's integrated genome browser, EPIC-CoGe

workflows are currently available in LoadExp+: RNA-seq, whole-genome BS-seq, ChIP-seq, SNP identification, and population genetics analysis. More detailed information for each analysis workflow is collected in Table 1. A schematic view of LoadExp+ workflows can be seen in Figure 1 and a full list of all tools (with benchmarking data) is provided in Table S1. Many workflows share usage of the various aligners integrated into LoadExp+ including GSNAP (Wu & Nacu, 2010), Bowtie2 (Langmead, Trapnell, Pop, & Salzberg, 2009), BWA-MEM (Li, 2013), TopHat2 (Kim et al., 2013), or HISAT2 (Kim, Langmead, & Salzberg, 2015). Sequencing reads can be trimmed by CutAdapt (Martin, 2011), Trim Galore! (Krueger, http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), Trimmomatic (Bolger, Lohse, & Usadel, 2014), or BBDuk (Bushnell, http://jgi.doe.gov/data-and-tools/bbtools/).

Analyses are dispatched by a job scheduling engine, based on WorkQueue (Albrecht, Rajan, & Thain, 2013). Users can request a notification email upon completion and monitor the job under the Activity and Data Loading tabs in the My Data page, or with a persistent link displayed on the progress window. All data integrated by users are kept private by default and may be shared with collaborators through the My Data page (Figure S5). Each LoadExp+ workflow generates output files that can be visualized in the EPIC-CoGe genome browser with a single click. Output files can also be downloaded or exported to CyVerse for users to perform additional analyses.

2.4 | Flexible data visualization with EPIC-CoGe

We have integrated a customized version of JBrowse (Buels et al., 2016), called EPIC-CoGe (https://goo.gl/cWX0PB), within CoGe to visualize the data generated by LoadExp+ and other appropriately formatted data. The browser will display any of the thousands of genomes available in CoGe onto which experimental data tracks may be overlaid. Each file generated by LoadExp+ is automatically loaded and made available as selectable tracks for viewing in EPIC-CoGe. Critically, users may simultaneously display their own private data and publicly available data already present in CoGe for the selected genome (Figure 2). Users can also mix and match display of different visualization tracks, comparing the results of different NGS workflows as well as data from different laboratories' experiments. While browsing a genome of interest and exploring their experimental data, users may export results from a genomic region of interest, search quantitative tracks for minima, maxima, or a range of data, and search for SNPs overlapping annotated genomic features. Users can also download complete experimental results, export them directly to their CyVerse Data Store for additional downstream analyses, or access them through CoGe's RESTful API (https://goo.gl/Pf4xjf).

2.5 | Benchmarking

To verify that all workflows are functioning as intended, we have re-analyzed publicly available data using LoadExp+. In order to simulate the needs of users working in various organisms and with a wide range of data sizes, we performed analyses with data from *Arabidopsis thaliana*, a common model organism, *Zea mays*, an agronomically important crop with a large genome, and *Homo sapiens*. We also used both paired-ended and single-ended reads in our benchmarking. All analyses were completed successfully with no errors, and the options and the run-time are recorded in Table S1-S3. The slowest-running analysis is often Bisulfite-seq, which is expected because of the computational demands of bisulfite sequencing aligners due to alignment against wild-type and C->T converted genomic sequences. The slowest analysis was bisulfite sequencing in *Z. mays* with paired-end reads, requiring ~54 hours, while the fastest analysis was RNA-seq from *Arabidopsis thaliana*, needing only 36 minutes to complete. These benchmarks represent a general expectation of run times, but users should note that their precise run-time will also vary based on server load.

2.6 | CoGe API

Data stored in CoGe, including genomes or the outputs and intermediate files generated by LoadExp+ analyses, are also accessible in a

TABLE 1 Workflow summary

| LoadExp+ Workflow	Analysis Tool(s)	Detailed Workflow Information (CoGe Wiki)	Visualization (Browser Track)	Analysis
RNA-seq	Cufflinks (Trapnell et al., 2010)	https://goo.gl/KCBtbH	Alignment, Read Depth (bar plot), FPKM (bar plot)	FPKM Differential expression
Bisulfite-seq	Bismark (Krueger & Andrews, 2011) BWAmeth (Pedersen, Eyring, De, Yang, & Schwartz, 2014)	https://goo.gl/2eegqK	Alignment, Percent Methylation (bar plot)	Whole-genome DNA methylation
ChIP-seq	HOMER (Heinz et al., 2010)	https://goo.gl/aFV2T5	Peaks (bar plot)	Protein–DNA Interactions
Variant calling	SAMtools (Li et al., 2009), Platypus (Rimmer et al., 2014), GATK (McKenna et al., 2010), CoGe Basic	https://goo.gl/GlQGu5	SNP Genome Browser Track	SNP Discovery
Population Genetics	Nucleotide diversity, Watterson's estimator, and Tajima's D	https://goo.gl/j4nNS1	Custom report	Identification of selection, selective sweeps
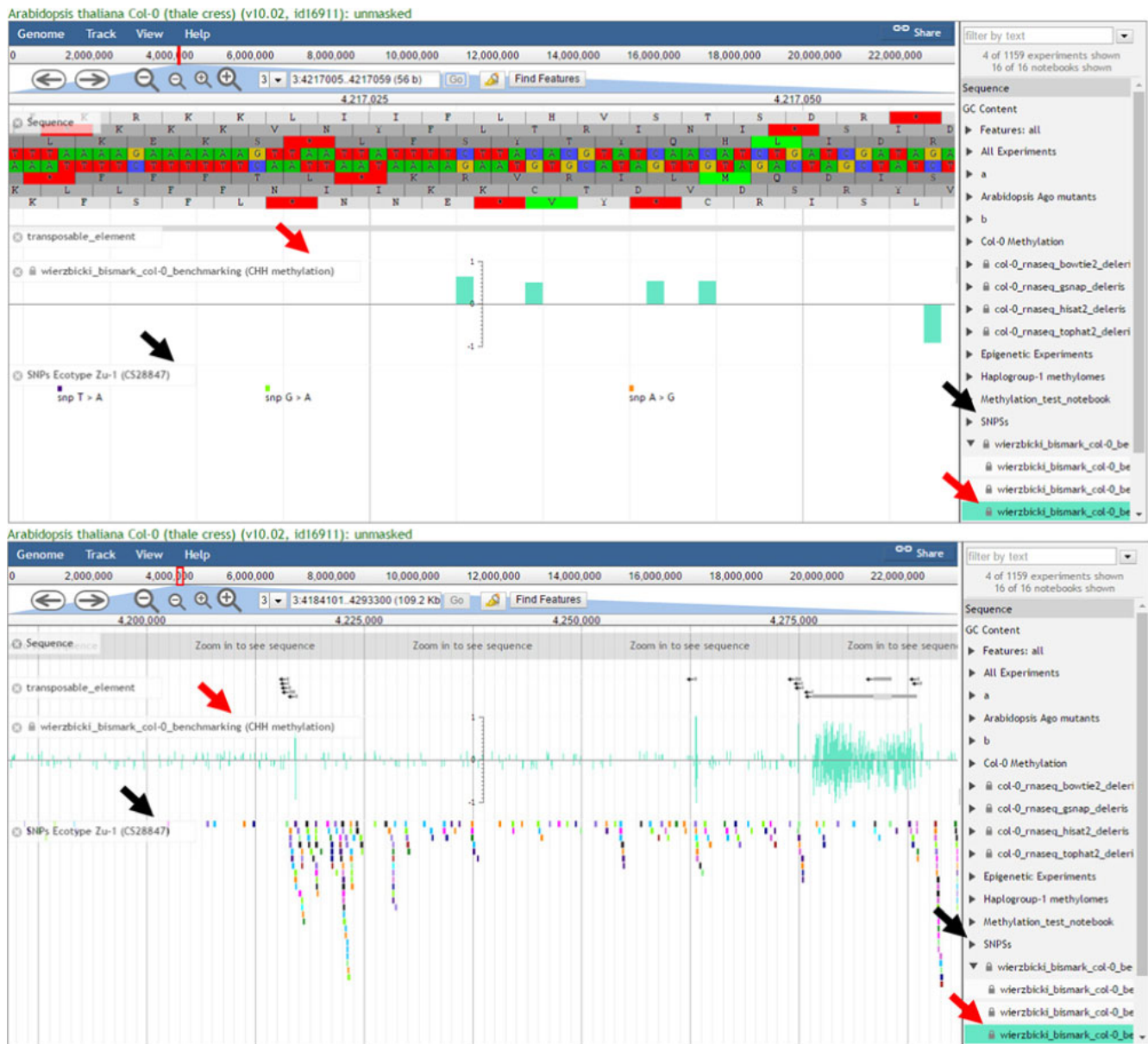
FIGURE 2 Data Visualization With EPIC-CoGe. At the completion of a LoadExp+ workflow, data are viewed in CoGe's genome browser, EPIC-CoGe (based on JBrowse). Experiments appear as selectable tracks within each notebook (right side). Shown here are private quantitative data (CHH methylation, red arrows) and public diversity data (SNPs, black arrows) displayed with annotated genomic features (transposable elements) and genomic DNA sequence. The upper and lower images display the same data at different scales to show both specific details (upper) and broad features (lower)

programmatic fashion. Using CoGe's RESTful (Masse, 2011) application programming interface (API) (https://goo.gl/Pf4xjf), developers can integrate CoGe's web services into their own websites or analysis pipelines. This extends CoGe and LoadExp+ beyond use by individual researchers by making them available as web services to be utilized by developers.

3 | DISCUSSION

LoadExp+ allows users to upload, analyze, and visualize a variety of public and private NGS data using CoGe's web-accessible graphical user interfaces (GUIs) and application programming interfaces (APIs). The streamlined user interface is designed to allow novice users to quickly move from data analysis to visualization, and allows more experienced users to customize their analyses and make them available to collaborators. Data generated, and analyses performed, using LoadExp+ may be kept private, shared with collaborators, or made fully public. Currently, LoadExp+ supports RNA-seq, ChIP-seq, whole-genome bisulfite sequencing (BS-seq), variant analysis (SNP calling), and population genetics analyses using any genome within CoGe. Users are able to upload genomes for organisms not already present in CoGe, new versions of genomes, and gene model annotations, extending the usefulness of LoadExp+ to any organism with a

sequenced genome. CoGe is powered by CyVerse (www.cyverse.org) (Merchant et al., 2016), providing access to high-performance, secure systems for computational scalability and interoperability. LoadExp+ offers a unified approach to NGS data analysis for both novice and experienced users.

LoadExp+ allows users to benefit from CoGe's CyVerse integration in various ways. The most important of which is federated user identity management, allowing the seamless transfer of large files from a user's CyVerse Data Store to CoGe in order to perform analyses using any of LoadExp+'s workflows. Data transfers between the Data Store and LoadExp+ are multithreaded and proceed in an automated fashion after submitting a job through the LoadExp+ GUI. Additionally, files that are generated through LoadExp+ analyses can be easily exported to the CyVerse Data Store as well, making them accessible to the bioinformatics applications available through the rest of CyVerse, including the CyVerse Discovery Environment for managing and running additional analyses, and Atmosphere, for applications that require on-demand cloud computing.

With the recent increased accessibility of sequencing-based analyses, several NGS workflow management tools have become available to bridge the gap between bioinformatics expertise and researchers who want to analyze data quickly and efficiently. LoadExp+ distinguishes itself from similar tools due to its integration with CoGe's thousands of available genomes, a modern and feature-rich genome browser (EPIC-CoGe), and easily accessible workflows for numerous genomic and epigenomic analyses on a public server. Tools such as Galaxy (Afgan et al., 2016) also combine a user-friendly interface with high-performance computing resources, yet at the time of publication it is not possible to perform the same analyses available through LoadExp+ on any of the publicly accessible Galaxy servers. In order for users to perform these analyses with Galaxy, they must set up a private Galaxy server and install their desired workflows, negating the user-friendly nature of the platform. Other NGS workflow management platforms such as QuickNGS (Wagle, Nikolić, & Frommolt, 2015) or Chipster (Kallio et al., 2011) lack public servers accessible to all users, requiring laboratories or institutes to maintain an installation themselves. Available web-based genomics resources and workflow management tools, in particular, underserve users that wish to analyze bisulfite sequencing. Web-based tools for bisulfite sequencing, such as WBSA (Liang et al., 2014), exist and are publicly available, but lack the integration with other NGS workflows that is present in LoadExp+. Many epigenomics analysis tools are also available through the CyVerse Discovery Environment (Merchant et al., 2016). However, because Discovery Environment applications are integrated by various users, usually for a specific purpose, not all options for the applications may be available. In LoadExp+, we have tried to make visible the most relevant options for each analysis, but users can contact our active development team in the event that an option they require is not available.

Additionally, LoadExp+ leverages CoGe's automatic integration of data into our advanced genome browser, EPIC-CoGe. Once data loading or workflows through LoadExp+ are complete, output

datasets are automatically available for visualization and additional analyses as tracks in EPIC-CoGe, CoGe's integrated genome browser which is based on JBrowse (Buels et al., 2016). EPIC-CoGe allows users to manipulate visualizations of their data in various ways including changing colors, normalizing or rescaling data, or searching for data that overlaps genomic features of interest. Users can also export raw or processed data, or data from genomic regions of interest either to their local machine or to the CyVerse Data Store. The combination of LoadExp+'s workflows with EPIC-CoGe's visualizations and data manipulation functions provide LoadExp+ with unique capabilities compared to other NGS workflow applications.

Finally, using LoadExp+ and CoGe simplifies the management of experimental data. Data generated or loaded through LoadExp+ can be made public or kept private and can be shared with other researchers easily. Data can also be organized into "Notebooks" that can, similarly, be kept private, made public, or shared with collaborators. LoadExp+ and CoGe offer a unified approach to genomics and epigenomics data analysis and visualization for novice and experienced users, as well as facilitating collaboration between scientists.

4 | CONCLUSIONS

To fully exploit NGS technologies in biological research, we must increase access to NGS analysis tools. We have tackled this problem through the creation of LoadExp+, an integrated suite of NGS workflows for analysis of genomic and epigenomic data within the CoGe platform. These workflows enable users to easily perform a variety of analyses, share their data with collaborators, and visualize their results on a single, web-based platform with an intuitive GUI. While many web-based platforms exist to assist researchers analyzing NGS data, LoadExp+ provides additional features for managing public and private data, support for many types of NGS data, and seamless integration with the EPIC-CoGe genome browser. We have verified that all LoadExp+ workflows can be run successfully and benchmarked performance with publicly available data (Tables S1-S3).

Our primary motivation is to lower the barrier of entry into genomics research, but the convenience of LoadExp+ genomics workflows makes their use an attractive option to researchers of all levels. In the time that LoadExp+ has been available, over 250 experiments are loaded each month by researchers (~8 per day), demonstrating the usefulness of this platform.

5 | METHODS

5.1 | Benchmarking

Benchmarking was performed using standard options for each workflow and publicly available data (Tables S1-S3). Analyses were performed one at a time to avoid multiple concurrent analyses

affecting run-time; however, run-time was still dependent on day-to-day server load. The results from the benchmarking analyses have also been made public in a CoGe notebook (https://goo.gl/dQoEGJ) and are accessible as experiment files tied to their respective genomes. The experiment files generated through benchmarking are also visible as selectable tracks in the EPIC-CoGe browser.

5.2 | Availability of data and materials

LoadExp+ is freely available on the web at https://genomevolution.org under the MIT open source license (source code on GitHub https://github.com/LyonsLab/coge). LoadExp+ and EPIC-CoGe user interfaces are written in JavaScript and are compatible with modern web browsers. Server-side components are written in PERL and Python.

AUTHORS' CONTRIBUTIONS

J.W.G. and M.B. integrated, tested, and benchmarked the workflows. M.B. and E.L. contributed code. SD led development of EPIC-CoGe. E.L., R.A.M., and B.D.G. conceived the project and E.L. and R.A.M. supervised the development of LoadExp+. J.W.G., E.L., and R.A.M. wrote the manuscript. All authors were involved in editing and approving the final version of this manuscript.

REFERENCES

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., ... Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses 2016 update. *Nucleic Acids Research*, *44*, gkw343. https://doi.org/10.1093/nar/gkw343

Albrecht, M., Rajan, D., & Thain, D. (2013). Making work queue cluster-friendly for data intensive scientific applications. Proc - IEEE Int Conf Clust Comput ICCC.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., ... Holmes, I. H. (2016). JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biology*, *17*, 1–12. https://doi.org/10.1186/s13059-016-0924-1.

Hedges, M., Hasan, A., & Blanke, T. (2007). Management and preservation of research data with iRODS. In *Proceedings of the ACM first workshop on CyberInfrastructure: Information management in eScience.* (pp. 17–22). New York, NY: ACM. https://doi.org/10.1145/1317353.1317358

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., ... Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

Kallio, M. A., Tuimala, J. T., Hupponen, T., Klemelä, P., Gentile, M., Scheinin, I., ... Korpelainen, E. I. (2011). Chipster: User-friendly analysis software for microarray and other high-throughput data. *BMC Genomics.*, *12*, 507. https://doi.org/10.1186/1471-2164-12-507.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*, 357–360. https://doi.org/10.1038/nmeth.3317.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*, R36. https://doi.org/10.1186/gb-2013-14-4-r36.

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, *27*, 1571–1572. https://doi.org/10.1093/bioinformatics/btr167.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*, R25.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997. https://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Liang, F., Tang, B., Wang, Y., Wang, J., Yu, C., Chen, X., ... Li, R. (2014). WBSA: Web service for bisulfite sequencing data analysis. *PLoS ONE*, *9*, 1–9.

Lyons, E., & Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, *53*, 661–673.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, *17*, 10. https://doi.org/10.14806/ej.17.1.200

Masse, M. (2011). *REST API design rulebook: Designing Consistent RESTful Web Service Interfaces*. O'Reilly Media, Inc.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., & Antin, P. (2016). The iPlant collaborative: Cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biology*, *14*, e1002342. https://doi.org/10.1371/journal.pbio.1002342.

Pedersen, B. S., Eyring, K., De, S., Yang, I. V., & Schwartz, D. A. (2014). Fast and accurate alignment of long bisulfite-seq reads. *ArXiv Preprint ArXiv*: 1401.1129. http://arxiv.org/abs/1401.1129v2

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., WGS500 Consortium, ... Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, *46*, 912–918. https://doi.org/10.1038/ng.3036.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., ... Wilkins-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. *Computing in Science & Engineering*, *16*, 62–74. https://doi.org/10.1109/MCSE.2014.80.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., ... Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*, 511–515. https://doi.org/10.1038/nbt.1621

Wagle, P., Nikolić, M., & Frommolt, P. (2015). QuickNGS elevates Next-Generation Sequencing data analysis to a new level of automation. *BMC Genomics.*, *16*, 487. https://doi.org/10.1186/s12864-015-1695-x.

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, *26*, 873–881.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.