1 **CoGe LoadExp+: A web-based suite that integrates next-gen sequencing data analysis**

2 **workflows and visualization**

3

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

5 Mosher[1, 2, 4*], and Eric Lyons[2, 4*]

6

7 [1]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721, USA

8 [2]BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA

9 [3]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

10 [4]School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

11 [*]To whom correspondence should be addressed.

12

13 **Contact:** [*]ericlyons@email.arizona.edu, [*]rmosher@email.arizona.edu,

14 groverj3@email.arizona.edu, mbomhoff@email.arizona.edu, sdavey@email.arizona.edu

15

16 **Running Title: LoadExp+ for NGS analysis**

17

18 **Keywords**

19 Genomics, Epigenomics, Next-generation sequencing, RNAseq, expression analysis, bisulfite

20 sequencing, DNA methylation, ChIP-seq, SNP identification, population genetics

21

22 **Abstract**

23 To make genomic and epigenomic analyses more widely available to the biological research

24 community, we have created LoadExp+, a suite of bioinformatics workflows integrated with the

25 web-based comparative genomics platform, CoGe. LoadExp+ allows users to perform

26 transcriptomic (RNA-seq), epigenomic (bisulfite-seq), chromatin-binding (ChIP-seq), variant

27 identification (SNPs), and population genetics analyses against any genome in CoGe, including

28 genomes integrated by users themselves. Through LoadExp+'s integration with CoGe's existing

29 features, all analyses are available for visualization and additional downstream processing, and

30 are available for export to CyVerse's data management and analysis platforms. LoadExp+

31 provides easy-to-use functionality to manage genomics and epigenomics data throughout its

32 entire lifecycle and facilitates greater accessibility of genomics analyses to researchers of all

33 skill levels. LoadExp+ can be accessed at https://genomevolution.org.

34

35 **Background**

36 As advanced next-generation sequencing (NGS) technologies become more powerful and

37 affordable, a growing number of researchers find themselves in a position to answer genome-

38 scale biological questions using both existing and newly generated data. However, despite rapid

39 technological advances, the knowledge and computational resources needed to analyze and

40 interpret multiple large genomic datasets create bottlenecks for scientific discovery. While

41 initiatives exist to provide high performance computing resources (e.g., CyVerse, XSEDE) [1, 2]

42 to life scientists, these resources require computational expertise to exploit. Researchers in the

43 life sciences without this expertise would benefit from a platform focused on NGS data and

44 genomics analyses that integrates data management, analysis, and visualization tools into a

45 single user-friendly interface. Such a platform also benefits more advanced users by simplifying

46 the process of analyzing, visualizing, managing, and distributing data with collaborators.

47

48   We have addressed these needs with the creation of LoadExp+, an addition to the comparative

49   genomics platform, CoGe (https://genomevolution.org) [3]. LoadExp+ is a web portal through

50   which numerous genomic and epigenomic analyses can be conducted. These analyses include

51   RNAseq, whole-genome bisulfite-sequencing (BS-seq), ChIPseq, SNP identification, and

52   population genetics calculations. The collection of these tools in one location, integration with an

53   advanced genome browser, and the use of CoGe's user-friendly interface present advantages

54   over other web-based bioinformatics platforms for the life sciences. For advanced users, there

55   is also an REST API available for programmatic access to data integrated through LoadExp+

56   (https://goo.gl/Pf4xjf).

57

58   **Results**

59   **Web Interface**

60   LoadExp+ is accessible from CoGe's main page from the menu bar under "Tools" or the User

61   Data page by clicking "Load Experimental Data." It is intuitive to navigate, with mouse-over

62   descriptions and conspicuous links to documentation for each tool. Users begin their

63   experimental set-up by selecting data to load from their local computer, specifying a web

64   address (HTTP or FTP), inputting NCBI Short Read Archive (SRA) accession numbers, or

65   selecting files from the user's CyVerse Data Store (**Figure 1**). Options for LoadExp+'s

66   integrated programs, otherwise accessible as command-line arguments, are then presented as

67   checkboxes or entry fields. When setting up their desired analysis users must provide minimal

68   metadata by naming and describing their experiment, providing version information, and

69   indicating the source of their data.

70

71   Before submitting their desired analysis users can choose to have the resulting files

72   automatically organized into a new or existing notebook (**Supplemental Figure S1 and S2**),

73   and have an email sent to them when their analysis is complete. After submitting their analysis

74   users are automatically presented with a window in which to monitor its progress as well as a

75   persistent link that they can use to return to monitor the workflow's progress. Users can also

76   navigate to their "My Data" page to check the progress of submitted LoadExp+ workflows under

77   "Analyses." Results from the user's analysis are available for viewing and further analysis in the

78   EPIC-CoGe genome browser (see below). All files, including intermediate files, generated from

79   the analysis can also be downloaded from the Experiment View page (**Supplemental Figure**

80   **S3 and S4**) or exported to the CyVerse Data Store. These files are also available through a

81   RESTful API following completion of the run.

82

83   **Supported Data Formats and Metadata Attribution**

84   LoadExp+ supports the input of data in a variety of formats, including raw sequencing reads

85   (FASTQ, SRA), compressed data (.gz, .bz2, .zip), or preprocessed data in the form of

86   alignments (BAM, SAM), polymorphism data (VCF, GVCF), and quantitative data (CSV/TSV,

87   BED/BEDGraph, GFF/GTF, WIG, BigWig). These data can be uploaded directly from the user's

88   computer, retrieved from a remote server (FTP/HTTP), imported from their CyVerse Data Store

89   (recommended for files > 1GB), or retrieved from NCBI's Short Read Archive by accession or

90   project number. Transferring data between CoGe and the CyVerse Data Store uses iRODS [4]

91   for secure, high-performance, parallel file transfers.

92

93   Users are prompted to supply metadata which are captured for every experiment, including the

94   name, description, type, and source of the data. Additionally, the analysis program(s),

95   parameters, and the options chosen at run-time are captured automatically and are immutable.

96   After an experiment has been loaded, researchers can add additional metadata including text-

97   based descriptions, web links, and images (**Supplemental Figure S4**).

98

99   **Integrated Analysis Workflows**

100   LoadExp+ is a unified platform encompassing a variety of popular and powerful genomic and

101   epigenomic analyses. The following NGS workflows are currently available in LoadExp+: RNA-

102   seq, whole-genome BS-seq, ChIP-seq, SNP identification, and population genetics analysis.

103   More detailed information for each analysis workflow is collected in **Table 1**. A schematic view

104   of LoadExp+ workflows can be seen in **Figure 1** and a full list of all tools (with benchmarking

105   data) is provided in **Supplemental Table S1**. Many workflows share usage of the various

106   aligners integrated into LoadExp+ including GSNAP [12], Bowtie2 [13], TopHat2 [14], or HISAT2

107   [15]. Sequencing reads can be trimmed by either CutAdapt [16] or Trim Galore! [17].

108

109   Analyses are dispatched by a job scheduling engine, based on WorkQueue [18]. Users can

110   request a notification email upon completion and monitor the job under the Activity and Data

111   Loading tabs in the My Data page, or with a persistent link displayed on the progress window.

112   All data integrated by users are kept private by default and may be shared with collaborators

113   through the My Data page (**Supplemental Figure S5**). Each LoadExp+ workflow generates

114   output files that can be visualized in the EPIC-CoGe genome browser with a single click. Output

115   files can also be downloaded or exported to CyVerse for users to perform additional analyses.

116

117   **Flexible Data Visualization With EPIC-CoGe**

118   We have integrated a customized version of JBrowse [19], called EPIC-CoGe

119   (https://goo.gl/cWX0PB), within CoGe to visualize the data generated by LoadExp+ and other

120   appropriately formatted data. The browser will display any of the thousands of genomes

121   available in CoGe onto which experimental data tracks may be overlaid. Each file generated by

122   LoadExp+ is automatically loaded and made available as selectable tracks for viewing in EPIC-

123   CoGe. Critically, users may simultaneously display their own private data and publicly available

124   data already present in CoGe for the selected genome (**Figure 2**). Users can also mix and

125   match display of different visualization tracks, comparing the results of different NGS workflows

126   as well as data from different labs' experiments. While browsing a genome of interest and

127   exploring their experimental data users may export results from a genomic region of interest,

128   search quantitative tracks for minima, maxima, or a range of data, and search for SNPs

129   overlapping annotated genomic features. Users can also download complete experimental

130   results, export them directly to their CyVerse Data Store for additional downstream analyses, or

131   access them through CoGe's RESTful API (https://goo.gl/Pf4xjf).

132

133   **Benchmarking**

134   To verify that all workflows are functioning as intended we have re-analyzed publicly available

135   data using LoadExp+. In order to simulate the needs of users working in various organisms and

136   with a wide range of data sizes we performed analyses with data from *Arabidopsis thaliana*, a

137   common model organism, *Zea mays*, an agronomically important crop with a large genome, and

138   *Homo sapiens*. We also used both paired ended and single ended reads in our benchmarking.

139   All analyses were completed successfully with no errors, and the options and the run-time are

140   recorded in **Supplemental Table S1-S3**. The slowest-running analysis is often Bisulfite-seq,

141   which is expected because of the computational demands of bisulfite sequencing aligners due

142   to alignment against wild type and C->T converted genomic sequences. The slowest analysis

143   was bisulfite sequencing in *Z. mays* with paired end reads, requiring ~54 hours, while the fastest

144   analysis was RNAseq from *Arabidopsis thaliana*, needing only 36 minutes to complete. These

145   benchmarks represent a general expectation of run times, but users should note that their

146   precise run-time will also vary based on server load.

147

148   **CoGe API**

149   Data stored in CoGe, including genomes or the outputs and intermediate files generated by

150   LoadExp+ analyses, are also accessible in a programmatic fashion. By using CoGe's RESTful

151   [20] application programming interface (API) (https://goo.gl/Pf4xjf) developers can integrate

152   CoGe's web services into their own websites or analysis pipelines. This extends CoGe and

153   LoadExp+ beyond use by individual researchers by making them available as web services to

154   be utilized by developers.

155

156   **Discussion**

157   LoadExp+ allows users to upload, analyze, and visualize a variety of public and private NGS

158   data using CoGe's web-accessible graphical user interfaces (GUIs) and application

159   programming interfaces (APIs). The streamlined user interface is designed to allow novice users

160   to quickly move from data analysis to visualization, and allows more experienced users to

161   customize their analyses and make them available to collaborators. Data generated, and

162   analyses performed, using LoadExp+ may be kept private, shared with collaborators, or made

163   fully public. Currently, LoadExp+ supports RNA-seq, ChIP-seq, whole-genome bisulfite

164   sequencing (BS-seq), variant analysis (SNP-calling), and population genetics analyses using

165   any genome within CoGe. Users are able to upload genomes for organisms not already present

166   in CoGe, new versions of genomes, and gene model annotations, extending the usefulness of

167   LoadExp+ to any organism with a sequenced genome. CoGe is powered by CyVerse

168   (www.cyverse.org) [1], providing access to high-performance, secure systems for computational

169   scalability and interoperability. LoadExp+ offers a unified approach to NGS data analysis for

170   both novice and experienced users.

171

172   LoadExp+ allows users to benefit from CoGe's CyVerse integration in various ways. The most

173   important of which is federated user identity management, allowing the seamless transfer of

174   large files from a user's CyVerse Data Store to CoGe in order to perform analyses using any of

175   LoadExp+'s workflows. Data transfers between the Data Store and LoadExp+ are multithreaded

176   and proceed in an automated fashion after submitting a job through the LoadExp+ GUI.

177   Additionally, files that are generated through LoadExp+ analyses can be easily exported to the

178   CyVerse Data Store as well, making them accessible to the bioinformatics applications available

179   through the rest of CyVerse, including the CyVerse Discovery Environment for managing and

180   running additional analyses, and Atmosphere, for applications that require on-demand cloud

181   computing.

182

183   LoadExp+ distinguishes itself from similar web-based tools due to its integration with CoGe and

184   easily accessible workflows for numerous genomic and epigenomic analyses. While tools such

185   as Galaxy [21] also combine a user-user friendly interface with high performance computing

186   resources, at the time of publication it is not possible to perform the same analyses available

187   through LoadExp+ on any of the publicly accessible Galaxy servers. Users that wish to perform

188   bisulfite sequencing, in particular, are unserved by public Galaxy installations. In order for users

189   to perform these or other analyses not currently part of a public Galaxy installation they will

190   need to set up their own and install their desired workflows, negating the user-friendly nature of

191   the platform. Many epigenomics analysis tools are also available through the CyVerse

192   Discovery Environment [1]. However, because Discovery Environment applications are

193   integrated by various users, usually for a specific purpose, not all options for the applications

194   may be available. In LoadExp+ we have tried to make visible the most relevant options for each

195   analysis, but users can contact our active development team in the event that an option they

196   desire is not available.

197

198   Additionally, LoadExp+ leverages CoGe's automatic integration of data into our advanced

199   genome browser, EPIC-CoGe. Once data loading or workflows through LoadExp+ are

200   complete, output datasets are automatically available for visualization and additional analyses

201   as tracks in EPIC-CoGe, CoGe's integrated genome browser which is based on JBrowse [19].

202   EPIC-CoGe allows users to manipulate visualizations of their data in various ways including

203   changing colors, normalizing or rescaling data, or searching for data that overlaps genomic

204   features of interest. Users can also export raw or processed data, or data from genomic regions

205   of interest either to their local machine or to the CyVerse Data Store.

206

207   Finally, using LoadExp+ and CoGe simplifies the management of experimental data. Data

208   generated or loaded through LoadExp+ can be made public or kept private, and can be shared

209   with other researchers easily. Data can also be organized into "Notebooks" that can, similarly,

210   kept private, made public, or shared with collaborators. LoadExp+ and CoGe offer a unified

211   approach to genomics and epigenomics data analysis and visualization for novice and

212   experienced users, as well as facilitating collaboration between scientists.

213

214   **Conclusions**

215   To fully exploit NGS technologies in biological research we must increase access to NGS

216   analysis tools. We have tackled this problem through the creation of LoadExp+, an integrated

217   suite of NGS workflows for analysis of genomic and epigenomic data within the CoGe platform.

218   These workflows enable users to easily perform a variety of analyses, share their data with

219   collaborators, and visualize their results on a single, web-based platform with an intuitive GUI.

220   While many web-based platforms exist to assist researchers analyzing NGS data, LoadExp+

221   provides additional features for managing public and private data, support for many types of

222   NGS data, and seamless integration with the EPIC-CoGe genome browser. We have verified

223   that all LoadExp+ workflows can be run successfully and benchmarked performance with

224   publicly available data (**Supplemental Tables S1-S3**).

225

226   Our primary motivation is to lower the barrier of entry into genomics research, but the

227   convenience of LoadExp+ genomics workflows makes their use an attractive option to

228   researchers of all levels. In the time that LoadExp+ has been available, over 250 experiments

229   are loaded each month by researchers (~8 per day), demonstrating the usefulness of this

230   platform.

231

232   **Methods**

233   **Benchmarking**

234   Benchmarking was performed using standard options for each workflow and publicly available

235   data (**Supplemental Tables S1-S3**). Analyses were performed one at a time to avoid multiple

236   concurrent analyses affecting run-time, however, run-time was still dependent on day-to-day

237   server load. The results from the benchmarking analyses have also been made public in a

238   CoGe notebook (https://goo.gl/dQoEGJ) and are accessible as experiment files tied to their

239   respective genomes. The experiment files generated through benchmarking are also visible as

240   selectable tracks in the EPIC-CoGe browser.

241

242   **Availability of Data and Materials**

243   LoadExp+ is freely available on the web at https://genomevolution.org under the MIT open

244   source license (source code on GitHub https://github.com/LyonsLab/coge). LoadExp+ and

245   EPIC-CoGe user interfaces are written in JavaScript and are compatible with modern web

246   browsers. Server-side components are written in PERL and Python.

247

248   **Competing interests**

249   The authors declare that they have no competing interests

250

251   **Funding**

254

255 **Authors' contributions**

256 JWG and MB integrated, tested, and benchmarked, the workflows. MB and EL contributed

257 code. SD led development of EPIC-CoGe. EL, RAM, and BDG conceived the project and EL

258 and RAM supervised the development of LoadExp+. JWG, EL, and RAM wrote the manuscript.

259 All authors were involved in editing and approving the final version of this manuscript.

260

261 **Acknowledgements**

266

267 **References**

268 1. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, et al. The iPlant Collaborative:

269 Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLOS Biol.

270 2016;14:e1002342. doi:10.1371/journal.pbio.1002342.

271 2. Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, et al. XSEDE: Accelerating

272 Scientific Discovery. Comput Sci Eng. 2014;16:62–74. doi:10.1109/MCSE.2014.80.

273 3. Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as

274 DNA sequences. Plant J. 2008;53:661–73.

275 4. Hedges M, Hasan A, Blanke T. Management and preservation of research data with iRODS.

276 In: Proceedings of the ACM first workshop on CyberInfrastructure: information management in

277 eScience - CIMS '07. New York, New York, USA: ACM Press; 2007. p. 17.

278 doi:10.1145/1317353.1317358.

279 5. Trapnell C, Williams B a, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript

280 assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching

281   during cell differentiation. Nat Biotechnol. 2010;28:511–5. doi:10.1038/nbt.1621.

282   6. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq

283   applications. Bioinformatics. 2011;27:1571–2. doi:10.1093/bioinformatics/btr167.

284   7. Pedersen BS, Eyring K, De S, Yang I V, Schwartz D a. Fast and accurate alignment of long

285   bisulfite-seq reads. arXiv Prepr. 2014;:1–2. http://arxiv.org/abs/1401.1129v2.

286   8. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of

287   Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for

288   Macrophage and B Cell Identities. Mol Cell. 2010;38:576–89. doi:10.1016/j.molcel.2010.05.004.

289   9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence

290   Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

291   10. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, et al. Integrating

292   mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing

293   applications. Nat Genet. 2014;46:912–8. doi:10.1038/ng.3036.

294   11. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The

295   Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA

296   sequencing data. Genome Res. 2010;20:1297–303. doi:10.1101/gr.107524.110.

297   12. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short

298   reads. Bioinformatics. 2010;26:873–81.

299   13. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of

300   short DNA sequences to the human genome. Genome Biol. 2009;10:R25.

301   14. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate

302   alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome

303   Biol. 2013;14:R36. doi:10.1186/gb-2013-14-4-r36.

304   15. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory

305   requirements. Nat Methods. 2015;12:357–60. doi:10.1038/nmeth.3317.

306   16. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.

307   EMBnet.journal. 2011;17:10. doi:10.14806/ej.17.1.200.

308   17. Felix Krueger. Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

309   18. Albrecht M, Rajan D, Thain D. Making work queue cluster-friendly for data intensive

310   scientific applications. Proc - IEEE Int Conf Clust Comput ICCC. 2013.

311   19. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic

312   web platform for genome visualization and analysis. Genome Biol. 2016;17:1–12.

313   doi:10.1186/s13059-016-0924-1.

314   20. Mark Masse. REST API Design Rulebook. O'Reilly Media; 2011.

315   21. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy

316   platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.

317   Nucleic Acids Res. 2016;44 May:gkw343. doi:10.1093/nar/gkw343.

318

319   **Table 1.** Workflow Summary

LoadExp+ Workflow	Analysis Tool(s)	Detailed Workflow Information (CoGe Wiki)	Visualization (Browser Track)	Analysis
RNAseq	Cufflinks [5]	https://goo.gl/KCBtbH	Alignment, Read Depth (bar plot), FPKM (bar plot)	FPKM Differential expression
Bisulfite-seq	Bismark [6] BWAmeth [7]	https://goo.gl/2eegqK	Alignment, Percent Methylation (bar plot)	Whole-genome DNA methylation
ChIPseq	HOMER [8]	https://goo.gl/aFV2T5	Peaks (bar plot)	Protein-DNA Interactions
Variant calling	SAMtools [9], Platypus [10], GATK [11], 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

320

321

322   **Figure Titles and Legends**

323   **Figure 1. Schematic Representation of LoadExp+ Workflow.** Data to be used in any

324   workflow may be imported directly from a user's local machine, a remote address by HTTP/FTP,

325   from SRA accession number, or from the CyVerse Data Store. After import LoadExp+ allows

326   users to choose which bioinformatics workflows to run. After the desired analyses have finished,

327   experimental data can be organized into "Notebooks" and viewed in CoGe's integrated genome

328   browser, EPIC-CoGe.

329

330   **Figure 2. Data Visualization With EPIC-CoGe.** At the completion of a LoadExp+ workflow,

331   data are viewed in CoGe's genome browser, EPIC-CoGe (based on JBrowse). Experiments

332   appear as selectable tracks within each notebook (right side). Shown here are private

333   quantitative data (CHH methylation, red arrows) and public diversity data (SNPs, black arrows)

334   displayed with annotated genomic features (transposable elements) and genomic DNA

335   sequence. The upper and lower images display the same data at different scales to show both

336   specific details (upper) and broad features (lower).

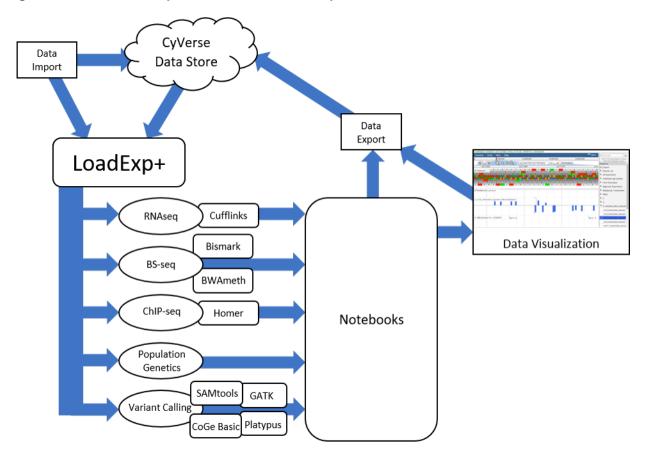337 **Figure 1. Schematic Representation of LoadExp+ Workflow.**

338

339

340 **Figure 2. Data Visualization With EPIC-CoGe.**

341

342