🏠

New Featured workspace showcasing the GenoMetric Query Language

Posted by Tiffany_at_Broad on 1 Jun 2018   💬 (0)

We are excited to introduce a new Featured workspace that demonstrates the GenoMetric Query Language (GMQL) created by a team from Politecnico di Milano in Italy. For some context on Featured workspaces, please read our previous blog post.

GMQL is a high-level, declarative language supporting queries over thousands of heterogeneous datasets and samples; as such, it enables genomic "big data" analysis. Based on Hadoop framework and the Apache Spark platform, GMQL is designed to be highly scalable, flexible, and simple to use. You can try the system here through its several interfaces, with documentation and biological query examples on ENCODE, TCGA and other public datasets or clone the Featured workspace and launch an example analysis.

The **GMQL 101 workspace** features three methods, each with increasing levels of complexity to give you a taste of how the query language works. One method shows how to join two datasets, and then extracts a third dataset based on a specific condition: pairs of regions that are less than 1000 bases apart. The second method takes a VCF and performs an epigenomic analysis using gene annotation and Chip-Seq results. It shows how you can select high confidence regions, use RefSeq annotations to find regions that overlap a gene, and count the mutations falling within the high confidence regions. Finally, the third method is a combination of GATK4's Mutect 2 pipeline and the second method, showing an epigenomic analysis from start (calling somatic variants) to finish (annotating variants). For any GMQL-specific questions or problems you can visit the GMQL GitHub page.

Many thanks to Luca Nanni, Arif Canakoglu, Pietro Pinoli, and Stefano Ceri for putting together this workspace. It takes a lot of thought and effort to create a valuable learning resource like this, and we are still figuring out the most successful way to do this. Please share your thoughts in the Comments section below on the effectiveness of this workspace and any other Featured workspaces you try out. If you are interested in featuring examples of your methods in this way, please tell us here, and we can talk to you about the process.

💬 Comment on this article

🏠

🐦 - FOLLOW US ON TWITTER

FireCloud
@BROADFIRECLOUD

The blog is hot this week 🔥 . Read about the latest GMQL Featured workspace, how to get your work featured, and our...
https://t.co/yndshEhYVk
6 JUN 18

RT @drzoomie: 5k samples processed in 2 days using @Trinity_CTAT + @BroadFireCloud $2 per sample! #transcriptomics #RNAseq #informatic...
23 MAY 18

RT @keesvanbochove: Geraldine van der Auwera @broadinstitute presenting @BroadFireCloud which interestingly (but not surprisingly) has almo...
17 MAY 18

RT @broadinstitute: "Making Data More FAIR on the Cloud." Geraldine Van der Auwera on uses of @BroadFireCloud at #BioIT18. TODAY, 11:40am (...
17 MAY 18

Want to hear the latest on WDL, Cromwell, FireCloud, and GATK #BioIT18 ? See this blog for tomorrow's schedule of t...
https://t.co/S7kf58tECu
16 MAY 18

♥ - OUR FAVORITE TWEETS FROM OTHERS

5k samples processed in 2 days using @Trinity_CTAT + @BroadFireCloud $2 per sample! #transcriptomics #RNAseq...
https://t.co/fnqRdKC0xN
23 MAY 18

Geraldine van der Auwera @broadinstitute presenting @BroadFireCloud which interestingly (but not surprisingly) has... https://t.co/DOkBfNkUbb
17 MAY 18

"Making Data More FAIR on the Cloud." Geraldine Van der Auwera on uses of @BroadFireCloud at #BioIT18. TODAY, 11:40...
https://t.co/csJCfc5RII
17 MAY 18

Geraldine from @broadinstitute @gatk_dev is still at the @googlecloud #BioIT18 booth (#410). She'll be talking abou...
https://t.co/QRT0BLe7AX
15 MAY 18

on the @BroadFireCloud #bioinformatics
https://t.co/m3pzDTkk6X
11 MAY 18

See more of our favorite tweets...