2022 Computational Genomics Survey Comments and Response
Scientific Computing and Data & GGS
March 9, 2022

The Computational Genomics survey —distributed in January 2022—solicited feedback from 662 active Minerva users. Of these, 24 users responded (3.6% response rate).

Survey Questions:

1. **What pipelines do you use? Check all that apply:**

Pipeline	Yes	No
Alignment and variant (SNV/indel) calling from WGS or WES data	58%	42%
Copy number calling from WES data	37%	63%
Structural variant calling form WGS data	37%	63%
De novo mutation calling from trio WGS or WES data	21%	79%
Gene-level differential expression analysis from RNA-seq data	75%	25%
Detection of alternative splicing from RNA-seq data	27%	63%
Other	33%	67%

2. **If possible, please describe the specific type of analysis you (or your team) perform and the main software used for analysis?**
    - Comparing pipelines with high-C data, attackSEQ data, etc.
    - Single Cell Multiomics
    - epigenomic features calling (ChIP-Seq/ATAC-Seq peaks, HiC chromatin interactions)
    - association analysis using PLINK, fastQTL and REGENIE
    - Single cell RNA-seq data analysis
    - Hi-C analysis, ChIP-seq
    - I write my own

    Thanks for providing the info on your specific type of analysis. It helps us understand better your needs in software and resources.

3. **Would you utilize a centralized service for genomic pipelines?**
    - Yes: 38%

- No: 8%
- Maybe: 54%

Comments regarding other needs users reported related to computational genomics:

Storage and compute resources

- Storage and backing up data is important. Also, having high power computer cluster nodes for further analyses is a must; We develop several custom approaches for different tasks. Basically having sufficient storage and compute nodes is good for us, with support to give flexibility in what we do; Computational analysis of large images
    - Minerva has 32 PB of storage along with TSM tape as data archival and backup, about 24,000 Compute cores with different RAM up to 1.5 TB/node, 88 GPUs with up to 2 TB RAM and 80 GB in A100 (more details are on our HPC website). For using TSM system as archival and backup, please follow our instructions on our website.
      If you need help on this, please send us a ticket at to discuss your specific requirement at hpchelp@hpc.mssm.edu

Pipeline & Software

- How does Mount Sinai centralized pipeline differ from community-curated nf-core pipelines? https://nf-co.re/
    - Looks like this is a great resource. We appreciate the recommendation. The HPC group will look into this resource and provide support for Mount Sinai users who would like to utilize this pipeline to analyze their data; for example, we will create an institution config file for users who want to run nf-core pipelines at Minerva. At the same time, we will look into all the pipelines' possibilities once we make a decision on a centralized pipeline and the best practice to fit our specific data set and research requirements.

- It will be great if we can use IPA as well.
    - We will look into this to see if this is something that we can potentially add.

Service/Customized environment

- Ability to setup custom environment on Minerva and access using RStudio, Jupyter, Pluto.jl notebooks via web browser from personal computer.
    - We already have Rstudio and Jupyter access via web browser. Please follow the instruction on our slide (especially page 18-22)

- Inclusion of Hi-C data analysis software and deepTools
    - Please send us a ticket at hpchelp@hpc.mssm.edu to discuss the specific requirement for software installation. We may already have deepTools.

- Documentation in case of a centralized service for genomic pipelines together with some more information of the already available genomic datasets/reference genomes for analysis.
    - Currently, we do not have a centralized genomics data analysis pipeline. This is an item that the committee will discuss. We plan to create a space for commonly used reference genomes - human and mouse. The detailed information will be described on the Data Ark website.

- GATK pipelines (the "ngs" pipeline on Minerva) along with accompanied reference data in the "ngs" project directory used to be regularly maintained and updated on Minerva. Hopefully, it can be picked up and kept regularly maintained and updated on Minerva, as almost all WES/WGS related analyses may use these GATK pipelines and accompanied third-party software such as bwa, samtools, etc.
    - We researched this project, however, we are not going to host and maintain this project through Data Ark as the tools and the database are outdated. However, the data ark group will develop to meet the user's research requirements related to WES/WGS data analysis. Currently, the data ark contains the phase III data for the 1,000 Genome Project. In addition, We will continue to add tools and databases to help users with their WES/WGS study, for example, the Genome Aggregation Database–gnomeAD would be a good resource to work with WES/WGS data analysis and we are looking into it and may add it to our data ark commons soon. The software mentioned, GATK, bwa, and samtools, have already been installed on Minerva, and they are regularly maintained and updated. In addition, the group will describe the possibility of a centralized WES/WGS data analysis pipeline. We would love to have you contact us to discuss the importance and details of this project and your specific requirements so we could better assist you. The email address is data-ark-team@lists.mssm.edu.
    -
- A centralized workflow with troubleshooting help would be AMAZING!
    - Thank you for the comment. We will look into the possibility of creating and maintaining a centralized workflow.