H3ABionet Next Gen Accreditation Questions

The following are questions to keep in mind when running the NextGen Workflow during the H3ABioNet accreditation exercise. Use them to plan your work in a way that would allow gathering the necessary information for your final report. The report should not be limited to only providing brief answers to these questions; it is expected to be a well-rounded description of the process of running the workflow, and of the results. Please note that only phases I and II of the variant calling SOP (<http://h3abionet.org/node-assessments/human-variant-calling>) need to be performed.

* Questions related to the nature of the input dataset
  + Was the input dataset of sufficiently good quality to perform the analysis?
  + How did the reads’ quality and GC content affect the way analysis was run?
* Operational questions
  + At each step of the workflow, describe which software was used and why:
    - Was the choice affected by the nature and/or quality of the reads?
    - Was the choice made due to the time and cost of the analysis?
    - What are the accuracy and performance considerations for the chosen piece of software?
  + For each software, describe which input parameters were chosen, and why:
    - Was the choice affected by the nature and/or quality of the reads?
    - Did the available hardware play a role in the parameter choice?
    - How did the purpose of the study affect the parameter choice?
  + For each step of the workflow, how do you know that it completed successfully and that the results are usable for the next step?
* Runtime analysis: this is useful information for making predictions for the clients and collaborators
  + How much time and disk space did each step of the workflow take?
  + How did the underlying hardware perform? Was it possible to do other things, or run other analyses on the same computer at the same time?
* Analyzing the results
  + How many variants were called with sufficient confidence to be included in further analyses? Are the results good and trustworthy, and can you estimate the sensitivity and selectivity of the analysis? How do you know the workflow completed successfully and the results are worth analyzing further?
  + How many variants were located in intronic, exonic, or in non-genic regions? Put this in context of the nature of the input dataset as described in the README.
  + How many variants were found in dbSNP and how many were unique to your sample? What does it mean?
  + What is the fraction of simple variants (SNPs, small indels) versus complex variants (translocations, inversions, etc.)? How is this influenced by your choice of software and parameters?
  + What would be the next steps for your analysis, given this information?