Quality Control of Whole Genome Sequencing

HEBRARD Maxime1, HOFMANN Oliver2, BERTIN Nicolas1, SMITH Lindsay3, on behalf of GHIF QC for WGS Workgroup

1. Precision Health Research Singapore (PRECISE), Genome Institute of Singapore (GIS), Singapore
2. Australian Genomics, Melbourne, Australia
3. Global Alliance for Genomics and Health (GA4GH), Ontario Institute for Cancer Research (OICR), Canada

When accessing genomics data from 3rd parties, researchers will often seek to only select datasets that meet certain quality control (QC) criteria requirements. For example, only retrieving datasets for which a nominal minimum sequencing depth or Q30 coverage has been reached. If this information is not available upon data discovery phase, is not defined with a common language or is not computed with a common process, data requestors will need to spend time and resources retrieving the data and computing QC metrics before eventually realizing it is unfit for purpose. Eventually if the burden of this preliminary exploration is too high, researchers might ignore the dataset in favor of more accessible ones. Therefore, (pre)computing and making available key sequencing and variant-calling QC metrics will increase the reach of data-generating initiatives, but also help them monitor their process internally and demonstrate good reliability.

Many tools are available to calculate a wide range of QC metrics and it is not uncommon for multiple tools to use similar names despite disparities in concept and method being employed resulting in different values reported. In this context, it is challenging both for data generators and requestors to choose which metrics to carry forward and to refer to. While existing ontologies and terminologies provide standard data concept definitions for capturing clinical information, standardized definitions and unambiguous reference implementations for NGS derived QC metrics have yet to be addressed.

Here, we propose to engage with GHIF/GA4GH participants and relevant tool developers in order to provide (i) standardized definition for key QC metrics, (ii) a reference implementation of selected tools for calculating them, and (iii) a set of benchmarking resources to assist further development of reference implementations. We envision the availability of standardized QC metrics will complement GA4GH’s Genomic Data Discovery ecosystem (DUO, DRS, phenopacket, etc).

[ 300/300 words ]