NICU Processing overview

Version 1.0.0

Introduction

I just briefly wanted to describe at a high level, how projects will enter the UCGD NeoSeq pipeline, where data will live, and how it can be accessed.

After Steven Boyden coordinates with the clinical & ARUP teams, he will review and complete the required manifest, then place it in a pre-defined ubox location.

Next, automation will acquire this completed manifest and initially do the following:

- Create a project name using the following format:
    - [Analysis id]-[kindred id]-RPN-[First phenotype term used in manifest]
    - Example: *A1000-104414-RPN-Coarctation*
- Update the UCGDDB
- Create a ubox project directory here
- Create a Fabric Genomics project.
- Create a Mosaic project.

Once sequencing is completed, ARUP will drop WGS Fastq files in a predetermined staging location, and inform us via a UCGD API we've created.

Processing

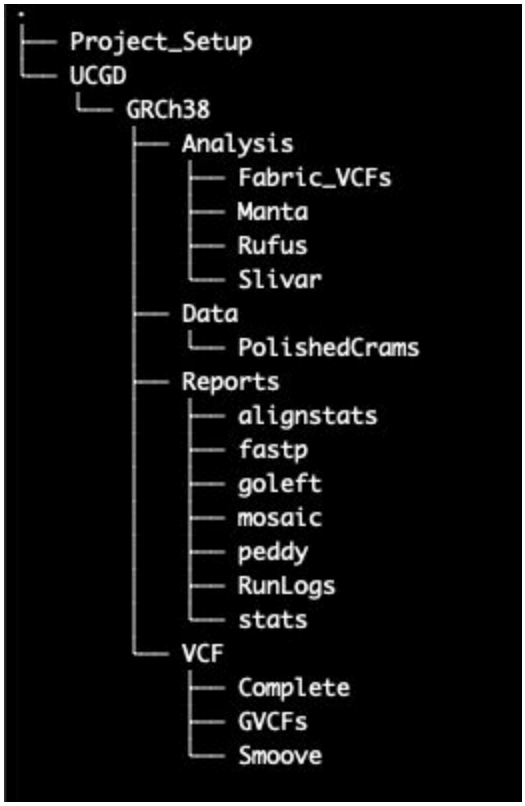
After receiving data for *all* individuals, and primary data validation, processing will commence, with the following step performed (not shown in order):

- Alignment.
- GVCF generation.
- Final VCF (w\VEP annotations, --everything, GO, LoFtool, SpliceRegion), VQSR (sensitivity 99.9) generation.
- Joint-genotyped with the new 30X WGS 1000 backgrounds (CEU, GBR, FIN).
- peddy, and ped file generation.
- QC (fastp, alignstats, lossless validation, bcftoots stats (vcf), goleft indexcov, multiqc).
- Lossless CRAM creation for review, and long term storage. *No BAM files will be kept*.
- Smoove SV calling following 'population calling protocol'.
- Final version of completed VCF will be kept, but the following step will additionally be ran for upload to Fabric (following this SOP)
    - bcftools normalize

    - All cohort individuals split out into individual VCF files.
    - Individual VCF files will be pushed to Fabric Genomes via their API.
    - A normalized Cohort VCF will be kept.
- MosiacCLI statistic computed and uploaded with files to mosaic.

Directory structure

The following data structure will be used:



An important point to mention here is that Analysis directories are designed to store *completed* result after each individual SOP is finalized.

A couple of important points:

- All data including original fastqs, will be processed and stored under the NICU IRB path, so in order to access (r+w) any data there, you ***MUST*** be added to the IRB. UCGD can not modify permission.
- The above shown project id is expected to be the following format used for all project once the transition to mosaic is complete, which is why A1000 will be the first id used for

NICU projects. This is expected to give us a grace period while the transition occurs. However when this format is used in conjunction with our move to mosaic, the following will change:
    - [Analysis id]-[kindred id]-RPN-[First phenotype term used in manifest]
    - [Analysis id]-[date]-[project type]-[First phenotype term used in manifest]
- If any additional directories are needed, or if anyone has additional questions or comments, please let us know.