Automated NICU Processing Outline

Version 1.0.0

List of manifest and database fields used for automated launch of NICU Pipeline:

Manifest (Clinical team) to UCGD mapping.

Manifest Field	Projects Table (UCGD)	Samples Table (UCGD)	type
sample_id		sample_id	int
arup_accession		ss_uuid	int
project	project	project	str
pi_first_name	pi_first_name	pi_first_name	str
pi_last_name	pi_last_name	pi_last_name	str
irb_institution		irb_institution	str
irb_number		irb_number	str
kindred_id		kindred_id	int
paternal_id		kindred_id	int
maternal_id		maternal_id	int
sex		sex	str
affection_status		affection_status	str
phenotype_description	phenotype	phenotype_description	str
hpo_terms		hpo_terms	str
birth_year		birth_year	int
assessment_year		birth_year	int
ethnic_group		race	str

ancestry		ancestry	str
consanguinity		consanguinity	str
tissue_type		tissue_type	str
tissue_condition		tissue_condition	str
molecule_type		molecule_type	str
sequence_center		sequence_center	str
seq_design		seq_design	str
status	status		str values: * *new_project* * ready * processing * hold * var_complete * analysis_complete * complete
sample_staus		sample_staus	str values: * *awaiting_data* * queue * ready * processing * called
scope_work	scope_work		* RPN
capture		seq_design	str
library_kit		library_kit	str
library_pcr		library_pcr	str
instrument		instrument	str
target_depth		target_depth	int
reference	assembly		str
notes	notes		str

ARUP (via API)

All items necessary for launch

DB Fields	Database	Table	SQL Type	Example
datatransfer_id**	webportal	portal_api_datatransfer	VARCHAR	ARUP-53000000003
src**	webportal	portal_api_datatransfer	VARCHAR	ARUP
src_id**	webportal	portal_api_datatransfer	INT	53000000003
filename**	webportal	portal_api_datatransferfile	VARCHAR	53000000003/53000000003_S2_R1_001.fastq.gz
md5sum**	webportal	portal_api_datatransferfile	VARCHAR	346321d02ab18b51ae77c02b1b2003fe

Required Manifest Fields

Manifest Required fields
sample_id
arup_accession
pi_last_name
kindred_id
paternal_id (affected only)
maternal_id (affected only)
sex
affection_status
phenotype_description
hpo_terms

** Values **not** captured from manifest: *Date_of_Birth, First_Name, Last_Name*

New columns and tables added

Table	Column	type
projects	mosaic_id	int
samples	mosaic_sample_id	int
project	ucgd_id	str
project_iterator**	iterator**	int

**New Table and column added.

Automated tasks.

Data should flow from the Clinic to ARUP to UCGD. We should never find primary data that hasn't been uploaded to UCGDDB from a manifest.

Manifest to UCGDDB sync

- Collect new manifest files from ubox
  - Validate required manifest fields.
  - Validate pedigree.
  - Upload to UCGDDB.
- Creates project name.
- Build project in:
  - $IRBS processing space.
  - Ubox.
  - Fabric.
  - Mosaic (Project, QC and Reports)
- Move manifest to built project in Ubox.
- Delete local copy of manifest.

Data found in webportal not in UCGDDB

Accession numbers found in webportal but not discovered in UCGDDB

- Collect all accession numbers from webportal API.

- ❖ Do samples and accessions match:
    - ■ Yes:
        - ● Does project exist in UCGDDB:
            - ◆ No:
                issue ‘project_issue’ SNS
            - ◆ Yes: (status: awaiting_data)
                - ➢ Validate MD5
                    - ■ Exit on failure and send SNS
                - ➢ Move data into projects Project_Setup directory.
                - ➢ Update sample_status from ‘awaiting_data’ to ‘queue’
                - ➢ Add mosiac_sample_id to UCGDDB.
    - ■ No:
        - ● Collect samples as ‘rouge_data’, issue alert SNS.
        - ● Possible pattern match check could be run?

Processing status check

It has been decided that no projects will run unless all samples are present (UCGD Team meeting).

- ❖ Collect all NICU projects (projects table).
- ❖ Collect corresponding data from samples table.
    - ➢ All sample ‘queue’:
        - ■ Update sample and project status to: ‘ready’
    - ➢ Sample still ‘awaiting_data’:
        - ■ Issue SNS message “data waiting greater than >72 hours”

Pipeline run

Overview of how the pipeline will run

- ❖ Check UCGDDB if project and samples are in ‘ready’ state.
    - ➢ True:
        - ■ Update sample and project status to ‘processing’
        - ■ Launch pipeline [ALL|VAR|SV|POST|CLEAN]
            - ● On error:
                - ◆ UCGDDB project status updated to ‘hold’
                - ◆ SNS ‘project_issue’

Workflow overview.

The pipe operate in three steps:

1. Standard variant calling. [VAR]
2. Structural variant calling. [SV]
3. Filters, normalize and split individuals to upload to fabric [POST]
4. Clean up of processing directory and files. [CLEAN].
5. Or above steps are ran in order VAR, SV, POST, CLEAN. [ALL]

VAR calling steps:

- run_data_prep
- check_bgzf
- fastp
- fastq2bam
- bam2gvcf
- losslessValidate
- samtoolsCRAMer
- alignstats
- gvcfTyper
- mergeGVCFs
- varCalSnp
- applyVarCalSNP
- varCalIndel
- applyVarCalINDEL
- generateSampleFile
- finalStats
- updateBED
- runVEP
- mergeVEP
- finalVCF
- makePedFile
- peddy
- goleftIndexCov
- multiqc

SV calling steps:

- smoove_call
- smoove_merge
- smoove_genotype
- smoove_paste
- Smoove_annotate
- Clean_up

POST calling steps:

- process reheader
- process normalize
- process split_vcf

Clean calling steps:

- process clean_up
- process mosaic_post_samples
- process update_project_status

SNS messages

Topics and subscribers can be added, removed or modified as needed. Message can be predefined or sent ad-hoc as needed. SNS or emails can be used.

*** All SNS messages will be sent once daily.***

Current set

Name	Message	AWS Topic	Members
project_issue	Project issue discovered: {}	Project_Issue	Shawn
manifest_issue	Manifest Issue	Manifest_Issue	Shawn

	discovered: {}		Steve
new_project_discovered	Project {} discovered. Processing will begin when data arrives ~48 hours	New_Project	UCGD members
rouge_data	ARUP sequence data accession: {} discovered without proper clinical associated manifest.	Rouge_Data	Shawn Steve Mary-Ann ARUP
late_sample	Awaiting (> 72 hours) accession data {}.	Late_Sample	Shawn Steve Mary-Ann ARUP seq team
checksum_issue	Fastq MD5 validation failed for samples {}	Check_Sum_Issue	Shawn Steve ARUP seq team
missing_checksum	Missing MD5 checksum file for accession sample set: {}	Check_Sum_Issue	Shawn ARUP seq team
processing_started	Processing for project {} has started	Processing_Started	UCGD members
processing_completed	Processing for project {} has started. ? fabric url ? mosaic url	Processing_Completed	UCGD members ARUP analysis team

Crontab runtimes

Runtime (every hour)	Step
0:00:00	NICUWatch -mc
0:15:00	NICUWatch -dc
0:30:00	NICUWatch -pc
0:58:00	NICUWatch -rp -ps ALL

Automation Requirements

- All manifest files MUST be in .xlsx format and meet input requirements.
- Pedigree MUST meet input requirements.
- Pedagree sex MUST be filled out as ‘male’ or ‘female’.
- No directories WILL exist in ARUP Drop outside of ARUP generated accession directories. And all dropped data directories will adhere to agreed upon structure.
- All accessions MUST be unique.
- All files from ARUP MUST have *.fastq.gz file extension and MD5 files.