UCGD Project and Pipeline Processing

Overview of changes for UCGD-Pipeline version 2.12.43+, to add project creation improvements, and automation.

Project setup is assumed to flow the following way:

- New project is created (`ucgd_db3 project new`)
- Make any needed updates to project info during or after setup (`ucgd_db3 project update`)
- Setup staging area for file download (`ucgd_db3 project stage`)
- Samples are added to the `IRB_Number/Staging/[A###_VAR_phenotype]` directory. There is also a softlink to the staging area directly under the project folder.
- Edit `source_files_ids.txt` in the staging area. Two column file (first column Sample_ID and second column is input fastq/bam/cram file). Files should be under `IRB_Number/Staging/[A###_VAR_phenotype]` as well.
- Project is updated to `queued` status (`ucgd_db3 project queue`). At which point.
    - *Note: Any errors/warnings from the above steps cannot be ignored. Processing will not start otherwise.*
- Project is processed to completion (automation), completed files are added to mosaic, db status is updated.

Database status

Project status	Sample status	Scripts	Stage	Type
new_project	pending	`ucgd_db3`	Newly created project	manual
staging	awaiting_data	`ucgd_db3`	Prepare download staging area and sample to file manifest	manual
queued*	ready*	`ucgd_db3`	Project updated: source_files_id & and original data in staging directory.	manual
built	ready	`UCGDPipeline`	Project setup validated and directory structure completed.	automated

processing	processing	`UCGDPipeline`	VAR/JGT pipeline currently processing	automated
processed	processed	`UCGDPipeline`	VAR/JGT pipeline processing complete	automated
var_complete	called	`UCGDPipeline`	VAR/JGT and Mosaic upload completed.	automated

*Last updated carried out by individuals. Expected that `source_files_ids.txt` and original sample data is in the correct location after this update.

Additional manifest requirements

New column	Table	Required
IRB_Number	Projects	Yes

New projects

When projects are created, `ucgd_db` will set the project status as `new_project` and the sample status as `awaiting_data` by default. Running `ucgd_db` new will also create the following:

- Database: entry in projects and samples table.
- New Mosaic project.
- Build `IRB` directory if it *doesn't* exist. Including `scratch` and `staging` directories under `IRB`.

New Projects Examples:

Example (`A1007-200401-VAR-Rynearson-Test`) of expected processing space created by ucgd_db

```
$> tree A1007-200401-VAR-Rynearson-Test
├── A1007-200401-VAR-Rynearson-Test
    └── Staging -> ../Staging/A1007-200401-VAR-Rynearson-Test
    └── Scratch -> ../Scratch/A1007-200401-VAR-Rynearson-Test
├── Scratch
    └── A1007-200401-VAR-Rynearson-Test
└── Staging
    └── A1007-200401-VAR-Rynearson-Test
```

Example of expected processing space created by project analysts:

```
$> A1007-200401-VAR-Rynearson-Test
.
├── A1007-200401-VAR-Rynearson-Test
├── Scratch
└── Staging
    └── A1007-200401-VAR-Rynearson-Test
                └──source_files_ids.txt
                └──Primary files(0-*)(fastq|bam|cram)
```

Processing

`UCGDPipeline` will:

- Check current status of project.
  - Begin processing if project status is: `ready`.
- Complete project directory build (under `IRB`). DB: `built`.
- Start and process both VAR and JGT files from `source_files_ids.txt.`
- Begin processing project.
  - Set DB status to `processing`.
- VAR or JGT projects complete.
- Updated DB project to `var_complete` and sample to `called` upon completion.

Project name syntax

All UCGD projects will be created using the following order and format (hyphen separated):

- [UCGD ID]
- [date in ‘`year(last two digits)monthday`’ format]
- [project type]
- [principal investigator|name]
- [phenotype term]

Example

```
A1007-200401-VAR-Rynearson-Test
```