Treehouse Storage Management

Proposal for the organization and lifecycle including backup and retention for Treehouse data.
Retention Key: (retention on local filesystem/retention in S3/retention in Glacier)

treehouse/archive/

- primary/
    - original/ (45/90/∞)
        - TH01_0001_S01/
            - Folder name is assigned Treehouse ID
            - Original partner names ie FOO.R1.fq.gz or FOO.bam or FOO.sra
    - derived/ (45 /90/180)
        - TH01_0001_S01/
            - Processed fastq names, IFF grooming/conversion was necessary
- downstream/
    - TH01_0001_S01 (∞/∞/∞)
        - secondary/
            - ucsc_cgl-rnaseq-cgl-pipeline-3.4.1-ab45e/
                - RSEM/
                - Kallisto/
                - QC/
                - sortedByCoord.md.bam (45 /90/180)
            - ucsctreehouse-fusion-0.0.1-a5gfe/
            - ucsctreehouse-mini-var-call-1.0.0-abcde/
        - tertiary/
            - treehouse-protocol-9.0.0-ab5f33/
                - compendium-v5/
                    - All the notebooks, json, automated output etc
                    - Includes automatically created Report templates
        - findings/
            - Manually edited analyst files including presentations
- compendium/
    - 4.0.0/
        - expression.h5
- references/
    - eg, tertiary external reference files
    - starIndex_hg38_no_alt.tar.gz.(md5 checksum)
    - Version via the name
- metadata/
    - Various site- or batch-specific metadata
    - Includes pdfs of emails conveying details of metadata

    - Organized into subdirectories, by TH Site ID (TH01, TH02, TH03_TH26_TH27, TH04, ... THR33)

Special Cases

Tertiary

See operations#166 : When a tertiary analysis is manually run with:

- a "standard" docker instance and compendium version
- but: some other nonstandard configuration
- then: the treehouse-protocol dir path will be suffixed with .1, .2 etc.
- and: a "README.md" file will be manually placed in the output dir explaining what happened.

Permissions

In general, archive/ is group read+writable, and other no read/no write. The automated process is responsible for making its output folders group-non-writable when appropriate.
primary, findings, metadata = always group-writable.

compendium/ and references/ must be other-readable (non-writable) for the tertiary protocol to access them. All subdirs to these must in addition be other-executable.

Backup

- All folders backed up to archive-treehouse-ucsc-edu weekly
- Backup via aws sync w/o delete to allow for pruning on the source
- Potentially entire downstream hierarchy (excluding bam's) checked into a local private git repo for tracking as presented...

Lifecycle

- IDs
    - We intend to retain the Treehouse ID in its current format, e.g., TH01_0681_S01
    - Most of the time we will have enough data to assign a TH ID to an incoming sample - we need to stress this with clinical sites!
    - If we mistakenly apply a new donor ID on what turns out to be an existing donor's later sample, we will mark the first run/TH ID as obsolete (in REDCap), retire that TH ID forever, and re-run the whole sample under its correct TH ID.
- Retention in S3 is only if we believe we would delete from the local file system and still need to access from EC2, otherwise immediate transition to Glacier probably better and no intermediate step.

- Versions: Use semantic version - short hash
- Transition and backup is automatic
- May need custom script to delete bam’s from downstream as S3 life cycle only supports paths