DCC dicer Rev 3

Gordon Saksena

Rev 1

- Normalized everything*

- Misformatting killed whole sdrf
- SLOW

Rev2

- Most failures just killed single annotation
- Faster, C-based dicing, single threaded to avoid LSF farm
- Organized to separate filetypes from sdrf formatting
- Overrides more systematic

- Caching complicated reruns, provenance
- Error reporting too distributed

Rev 3 - Phases

- SDRF discovery
- Normalize/dice per tumor/sdrf/annotation-type.
- Autovalidate
- Manual accept
- Generate loadfile

SDRF discovery

- Fetch magetab archives
- Fetch non-conforming files, generate sdrf
- Report dates of sdrf's, number of samples in each, annotations present, delta in samples from previous

Normalization/Dicing

- Fetch/cache tarfiles
- Hack file per tumor/sdrf/annotation-type
- Use LSF farm for speed
- Generate SDRF file for dicing process.

Auto validate

- Eg merge all, generate IGV sif
- Eg run old version of an algorithm

Manual Accept

- Email would report that action was needed
- Store diced in tumor/annotation/ magetab_rev/, use hard links to save disk space
- Old directory structure was tumor/annotation/ archive_rev

Generate Loadfile

- Rewire to new output directory structure