Design and Implementation of an Automated SRA Submission Pipeline

Indraniel Das, Michael Kiwala, Todd Hepler, Craig Pohl, David Dooling, Elaine Mardis, Richard Wilson

The Genome Center at Washington University in St. Louis

`{idas,mkiwala,thepler,cpohl,ddooling,emardis,rwilson}@wustl.edu`

Introduction

The NCBI Short Read Archive (SRA) is a new resource for the archival of data produced by massively parallel next-generation sequencing technologies. The SRA was designed to scale against the large datasets produced by the massively parallel sequencing platforms, and has been gradually deployed since 2008.

The Genome Center at Washington University School of Medicine, as a large scale sequencing center, produces sequencing data from a diverse set of research studies, experiments, collaborators and technologies. The center currently has 14 Illumina/Solexa (GA II), 8 454 (FLX and Titanium), and 3 ABI SOLiD instruments. These instruments produce data for a wide range of areas, ranging from human and cancer studies to microbial and comparative genetics.

To effectively track the preparation, transfer, and receipt of data to the SRA from such a heterogeneous environment, an automated submissions pipeline is warranted.

Architecture

Shown below in Figure 1 is a general overview of the submissions process. The basic goal is to take sequence data produced by the instruments and have them published by the SRA. In general, a submission process to any type of data repository would involve the steps highlighted in the "submissions" box.

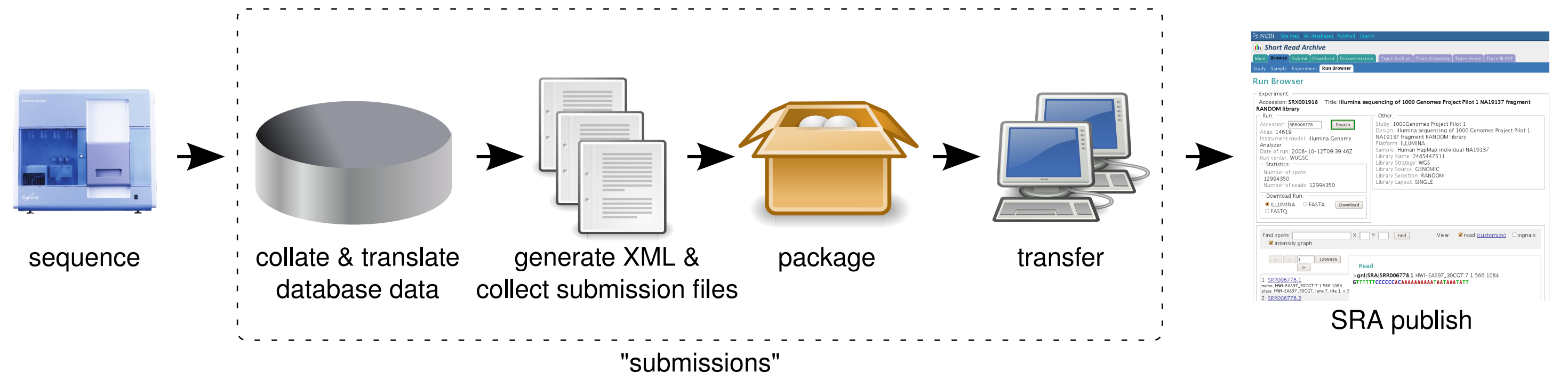
sequence
collate & translate database data
generate XML & collect submission files
package
transfer
"submissions"
SRA publish

Figure 1: General "submissions" process

An SRA submission is separated into two parts, *metadata* and *results*. The metadata provides the background information for the sequencing data. This is encompassed by the SRA Study, Sample and Experiment. The sequencing data results are contained within the SRA Run. Each of these SRA entities are independent of each other and can be submitted separately. The specification of these atomic SRA items are available at the SRA web site[1].

To generate these SRA items, the automated SRA submissions system interfaces at various points in already existing next-generation sequencing pipelines developed at The Genome Center. Each point chosen in the sequencing pipeline provides an entry point into the submission pipeline designed to generate a submission for a particular SRA Study, Sample, Experiment or Run. Most internal sequencing pipeline events are recorded within an OLTP database designed for processing efficiency. A new datawarehouse based submission database schema was created and designed around the SRA submission specifications. This transitional schema ensures that appropriate data integrity and constraints are met before transferring to the SRA.

Afterwards, data from the submission database is appropriately transformed and packaged into the desired XML file formats specified by the SRA and then transferred electronically to the SRA. Upon acceptance to the SRA, the archival system issues an accession identifier for the submission of interest. This accession is noted in the submission database for future reference.

The implementation details of how these entities are created, stored and managed with respect to The Genome Center's internal sequencing pipeline and the SRA's public identifying accession numbers are described in the following event and data architecture diagrams.

Event Architecture

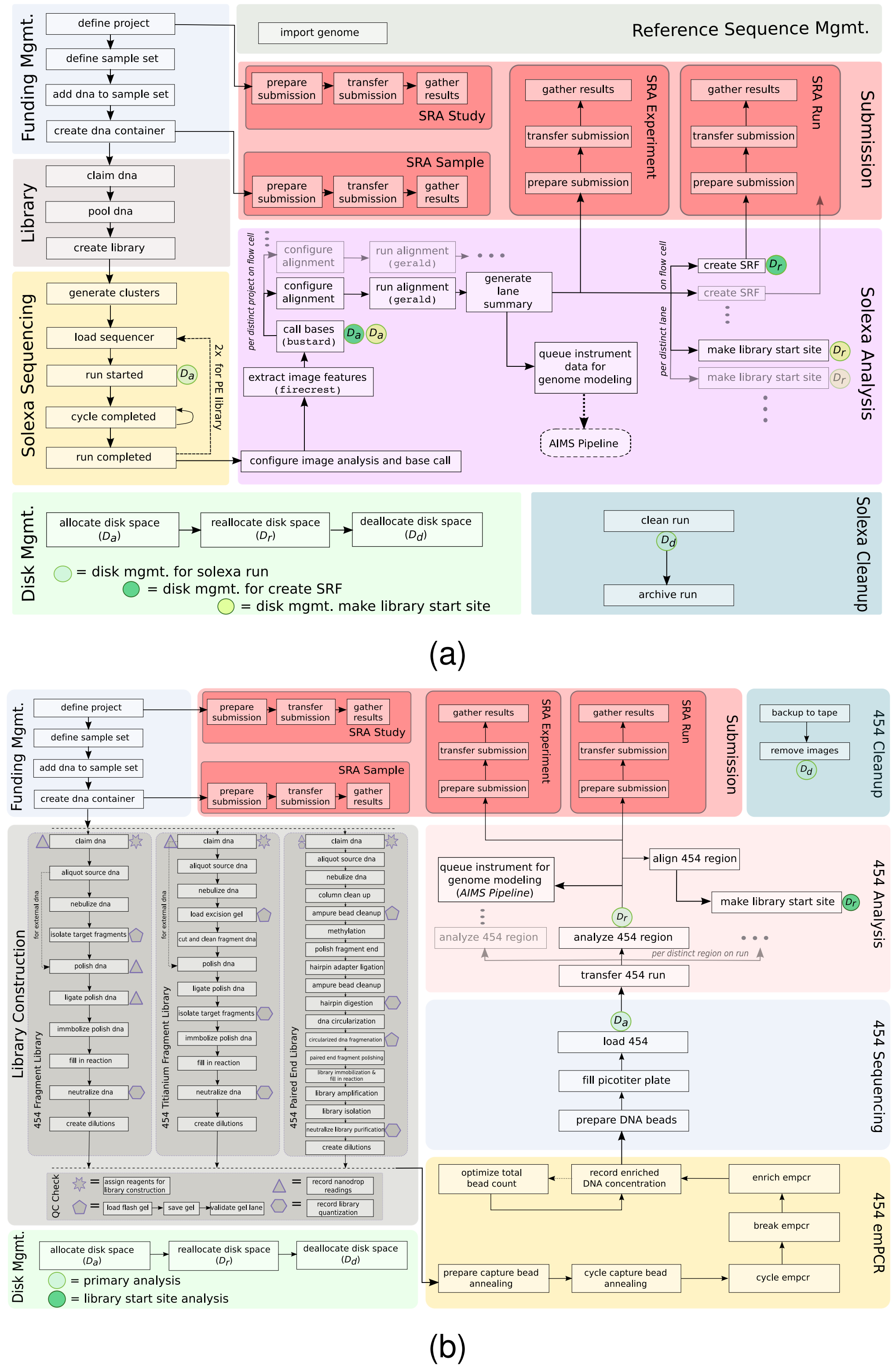
Funding Mgmt.
define project
define sample set
add dna to sample set
create dna container
Library
claim dna
pool dna
create library
Solexa Sequencing
generate clusters
load sequencer
run started
cycle completed
run completed
2x for PE library
Reference Sequence Mgmt.
import genome
prepare submission
transfer submission
gather results
SRA Study
SRA Sample
SRA Experiment
SRA Run
Submission
Solexa Analysis
per distinct project on flow cell
configure alignment
run alignment (gerald)
generate lane summary
call bases (bustard)
extract image features (firecrest)
configure image analysis and base call
queue instrument data for genome modeling
AIMS Pipeline
on flow cell
per distinct lane
create SRF
make library start site
Disk Mgmt.
allocate disk space (D_a)
reallocate disk space (D_r)
deallocate disk space (D_d)
= disk mgmt. for solexa run
= disk mgmt. for create SRF
= disk mgmt. make library start site
Solexa Cleanup
clean run
archive run
(a)
Library Construction
454 Fragment Library
454 Titanium Fragment Library
454 Paired End Library
for external dna
claim dna
aliquot source dna
nebulize dna
isolate target fragments
polish dna
ligate polish dna
immobilize polish dna
fill in reaction
neutralize dna
create dilutions
load excision gel
cut and clean fragment dna
column clean up
ampure bead cleanup
methylation
polish fragment end
hairpin adapter ligation
hairpin digestion
dna circularization
circularized dna fragmentation
paired end fragment polishing
library immobilization & fill in reaction
library amplification
library isolation
neutralize library purification
QC Check
= assign reagents for library construction
= record nanodrop readings
= load flash gel → save gel → validate gel lane
= record library quantization
= primary analysis
= library start site analysis
454 Cleanup
backup to tape
remove images
454 Analysis
queue instrument for genome modeling (AIMS Pipeline)
align 454 region
make library start site
analyze 454 region
per distinct region on run
transfer 454 run
454 Sequencing
load 454
fill picotiter plate
prepare DNA beads
454 emPCR
optimize total bead count
record enriched DNA concentration
enrich empcr
break empcr
prepare capture bead annealing
cycle capture bead annealing
cycle empcr
(b)

Figure 2: *Event processing architecture: (a) Illumina/Solexa sequencing pipeline; (b) 454 sequencing pipeline*

As DNA preparations traverse through the sequencing pipeline, each SRA entity is created at a point where there is enough information collected for its generation. The SRA pipeline (denoted in the red sections in Figure 2) is implemented upon a Perl based event processing and data framework that was developed earlier at The Genome Center [2] and already in use by the established sequencing pipelines.

The generation of each SRA entity is composed of three ordered events:

1. **Prepare Submission** *Translate the database data* represented in the internal sequencing pipeline into the specifications desired by the SRA in the transitional SRA schema. (Described in the following "Data Architecture" subsection.)
2. **Transfer Submission** Properly *generate*, *package* and electronically *transfer* the submission entity file(s) to the SRA based on the data created in The Genome Center's SRA transitional schema.
3. **Gather Results** Record the acknowledgement of the submission by the SRA via its *published* accession number.

Data Architecture

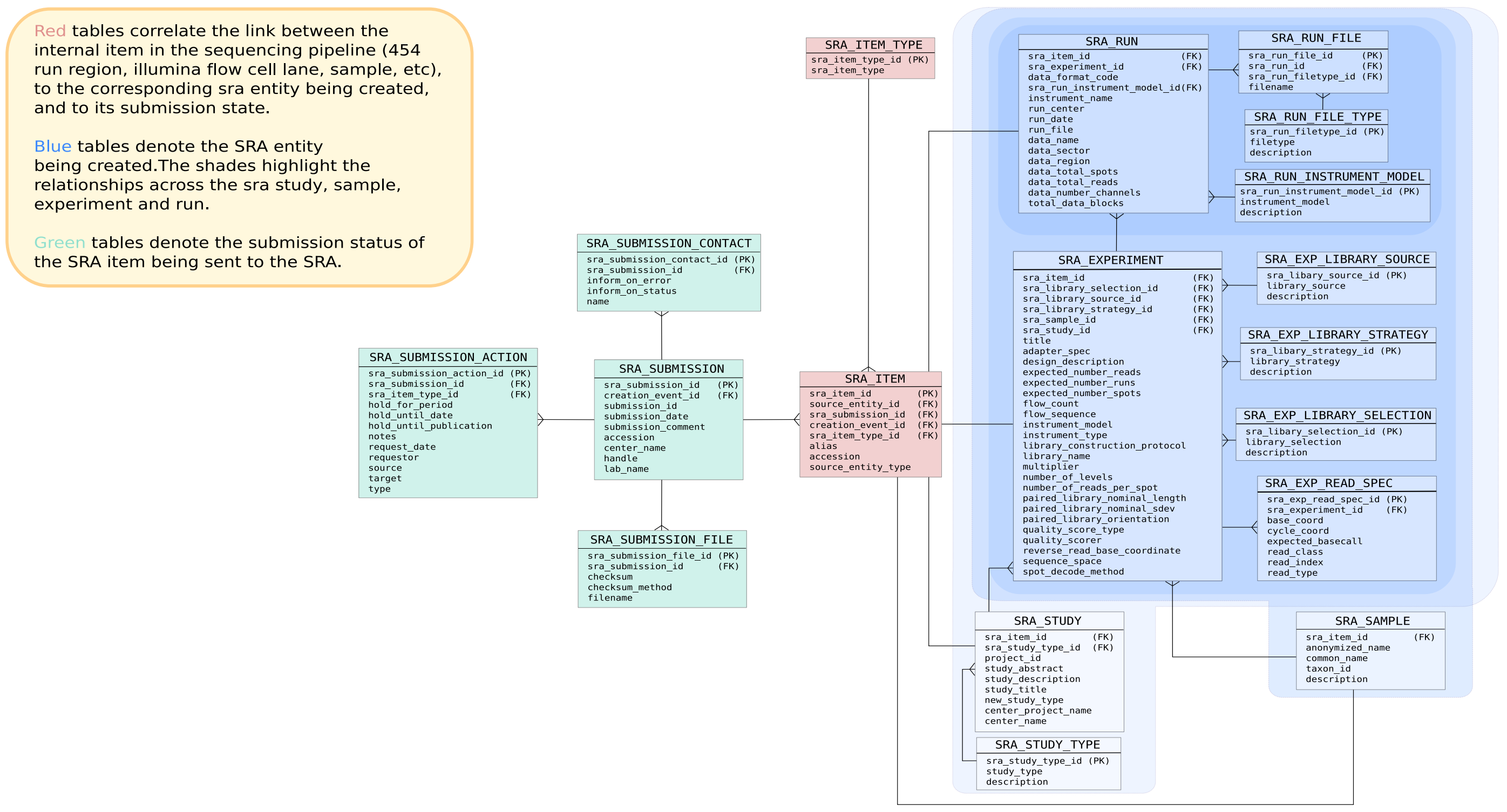
Red tables correlate the link between the internal item in the sequencing pipeline (454 run region, illumina flow cell lane, sample, etc), to the corresponding sra entity being created, and to its submission state.
Blue tables denote the SRA entity being created.The shades highlight the relationships across the sra study, sample, experiment and run.
Green tables denote the submission status of the SRA item being sent to the SRA.
SRA_ITEM_TYPE
sra_item_type_id (PK)
sra_item_type
SRA_SUBMISSION_CONTACT
sra_submission_contact_id (PK)
sra_submission_id (FK)
inform_on_error
inform_on_status
name
SRA_SUBMISSION_ACTION
sra_submission_action_id (PK)
sra_submission_id (FK)
sra_item_type_id (FK)
hold_for_period
hold_until_date
hold_until_publication
notes
request_date
requestor
source
target
type
SRA_SUBMISSION
sra_submission_id (PK)
creation_event_id (FK)
submission_id
submission_date
submission_comment
accession
center_name
handle
lab_name
SRA_SUBMISSION_FILE
sra_submission_file_id (PK)
sra_submission_id (FK)
checksum
checksum_method
filename
SRA_ITEM
sra_item_id (PK)
source_entity_id (FK)
sra_submission_id (FK)
creation_event_id (FK)
sra_item_type_id (FK)
alias
accession
source_entity_type
SRA_RUN
sra_item_id (FK)
sra_experiment_id (FK)
data_format_code
sra_run_instrument_model_id(FK)
instrument_name
run_center
run_date
run_file
data_name
data_sector
data_region
data_total_spots
data_total_reads
data_number_channels
total_data_blocks
SRA_RUN_FILE
sra_run_file_id (PK)
sra_run_id (FK)
sra_run_filetype_id (FK)
filename
SRA_RUN_FILE_TYPE
sra_run_filetype_id (PK)
filetype
description
SRA_RUN_INSTRUMENT_MODEL
sra_run_instrument_model_id (PK)
instrument_model
description
SRA_EXPERIMENT
sra_item_id (FK)
sra_library_selection_id (FK)
sra_library_source_id (FK)
sra_library_strategy_id (FK)
sra_sample_id (FK)
sra_study_id (FK)
title
adapter_spec
design_description
expected_number_reads
expected_number_runs
expected_number_spots
flow_count
flow_sequence
instrument_model
instrument_type
library_construction_protocol
library_name
multiplier
number_of_levels
number_of_reads_per_spot
paired_library_nominal_length
paired_library_nominal_sdev
paired_library_orientation
quality_score_type
quality_scorer
reverse_read_base_coordinate
sequence_space
spot_decode_method
SRA_EXP_LIBRARY_SOURCE
sra_libary_source_id (PK)
library_source
description
SRA_EXP_LIBRARY_STRATEGY
sra_libary_strategy_id (PK)
library_strategy
description
SRA_EXP_LIBRARY_SELECTION
sra_libary_selection_id (PK)
library_selection
description
SRA_EXP_READ_SPEC
sra_exp_read_spec_id (PK)
sra_experiment_id (FK)
base_coord
cycle_coord
expected_basecall
read_class
read_index
read_type
SRA_STUDY
sra_item_id (FK)
sra_study_type_id (FK)
project_id
study_abstract
study_description
study_title
new_study_type
center_project_name
center_name
SRA_STUDY_TYPE
sra_study_type_id (PK)
study_type
description
SRA_SAMPLE
sra_item_id (FK)
anonymized_name
common_name
taxon_id
description

Figure 3: *SRA submission database schema*

The SRA specifications for the SRA Study, Sample, Experiment and Run were translated into the database schema shown in Figure 3. These entities are also linked to their respective internal sequencing pipeline equivalent and their submission transaction status via the "`sra_item`" table.

After these SRA entities are populated in the above schema, an XLST based process is used to translate the entries into the appropriate XML file format. This allows for an additional level of data integrity checking.

Results

The automated system was initially applied on sequence data for the 1000 Genomes Pilot Project [3] (Pilots 1 & 2). The following highlights were observed:

- 46 distinct SRA Samples
- 176 distinct SRA Experiments
- 2079 distinct SRA Runs were generated
  - 1767 Illumina SRA Runs (each SRA Run entity corresponds to a single SRF for a single flow cell lane)
  - 312 454 SRA Runs (each SRA Run entity corresponds to a single SFF for a single 454 run region)
- The SRA Runs corresponded to approximately 17.82 TB of data. Transferring each of the submission entities via the Aspera [4] `ascp` software took approximately 19 days with an average transfer rate of approximately 11MB/sec.

Future Directions

Both The Genome Center's pipelines and the SRA specifications are a snapshot in time. They are continously evolving to accomodate the latest trends in sequencing technology and research. Currently the following issues are being actively considered on the automated submissions pipeline:

- *Interaction with the SRA Receipt Infrastructure*. Currently the submission pipeline is web scrapping off the SRA's website for the acknowledgement of the submission by the SRA. The SRA is currently developing a process to acknowlege receipt of the submission for the submitor. It is intended for the automated pipeline to interact with this once it is deployed by the SRA.
- *Appropriate flagging and handling of public and private/protected data.*
- *SRA Experiment Design/Specification updates.* Account for capture library handling in the SRA schema. The Genome Center is currently awaiting updated SRA specifications.

References

[1] NCBI Short Read Archive. `http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi`

[2] M. C. Wendl, S. Smith, C. S. Pohl, D. J. Dooling, A. T. Chinwalla, K. Crouse, T. Hepler, S. Leong, L. Carmichael, M. Nhan, B. Oberkfell, E. R. Mardis, L. W. Hillier and R. K. Wilson. Design and implementation of a generalized laboratory data model. *BMC Bioinformatics*, 8:362, 2007.

[3] 1000 Genomes Project. `http://www.1000Genomes.org`

[4] Aspera Inc. `http://www.asperasoft.com`

AGBT 2009 10th Annual Advances in Genome Biology and Technology Februrary 4-7. 2009, Marco Island, Florida