Opening apps and notebooks for cloud-scale cancer genomics

with Bioconductor
ITCR May 2019

Vince Carey (HMS), presenting results of efforts with Sean Davis of NCI CCR, John Readey (HDF Group), Herve Pages (FHCRC), BJ Stubbs, Shweta Gopaulakrishnan (HMS), Lori Shepherd (RPCI), Nathalie Pochet, Mohsen Nabian, Celine Everaert (Broad/HMS), Olivier Gevaert (Stanford), Marcel Ramos (CUNY), Nitesh Turaga, Martin Morgan (RPCI), Levi Waldron (CUNY)

A tension for cloud-scale genomic computing strategy

Integrated, mutually interdependent infrastructure (pre-cloud "ecosystem")

Isolated, high-speed, focused microservice, suited to cloud deployment (swagger, "serverless", ...)

Another tension

- Separation of concerns *vs.* silo-smashing
  - example from E. Mardis: seven fusion detectors employed -- separation of concerns manifested in multiple independent methods for a similar goal
  - Ravi Madduri -- variant calling APIs may change -- headache!
- Benefits of siloing
  - diversity of approaches to attacking very hard problems
  - evidential value of consensus
- Reduction of **costs** of siloing requires coordination/**governance**, test framework

Two concepts of governance for software ecosystems

- **Active:** users/developers get a seat on the board managing system components and vote on whether or not a pull request for an API change is merged
- **Passive:** tool developers commit in advance to grace periods during which
  - deprecated elements of APIs are noted as such
  - downstream tools have time to adapt
- In Bioconductor, this passive approach leads to "release" and "devel" branches -- costly to maintain, possibly confusing as we co-evolve with R, but likely **central to project durability and growth**
- Active and passive approaches are not mutually exclusive

Road map

- Example of an "app" that uses key cloud computing concepts implicitly: ivyGlimpse
- A project that is inevitably cloud-dependent: Sean Davis' BigRNA and OmicIDX systems, and the resulting ca43k app built from Bioconductor components
- Principles, tensions, and more examples

Part 1: IvyGAP resources: ISH

Data Overview

The Ivy Glioblastoma Atlas Project includes the following data sets

ISH: Image data at cellular resolution of *in situ* hybridization (ISH) tissue sections and adjacent hematoxylin and eosin (H&E)-stained sections annotated for anatomic structures

- **Anatomic Structures ISH Survey**: Primary screen of 8 tumors with probes for 343 genes enriched in glioblastoma.
- **Anatomic Structures ISH for Enriched Genes**: Subsequent screen of 29 tumors with probes for 37 genes enriched in glioblastoma structures identified in Anatomic Structures RNA-Seq Study (see below).
- **Cancer Stem Cells ISH Survey**: Primary screen of 16 tumors with probes for 55 genes enriched in putative cancer stem cells, resulting in a 20 probe reference set, which was then used in an extensive screen of 42 tumors.
- **Cancer Stem Cells ISH for Enriched Genes**: Subsequent screen of 37 tumors with probes for 76 genes enriched in clusters of putative cancer stem cells identified in the Cancer Stem Cells RNA-Seq Study (see below).

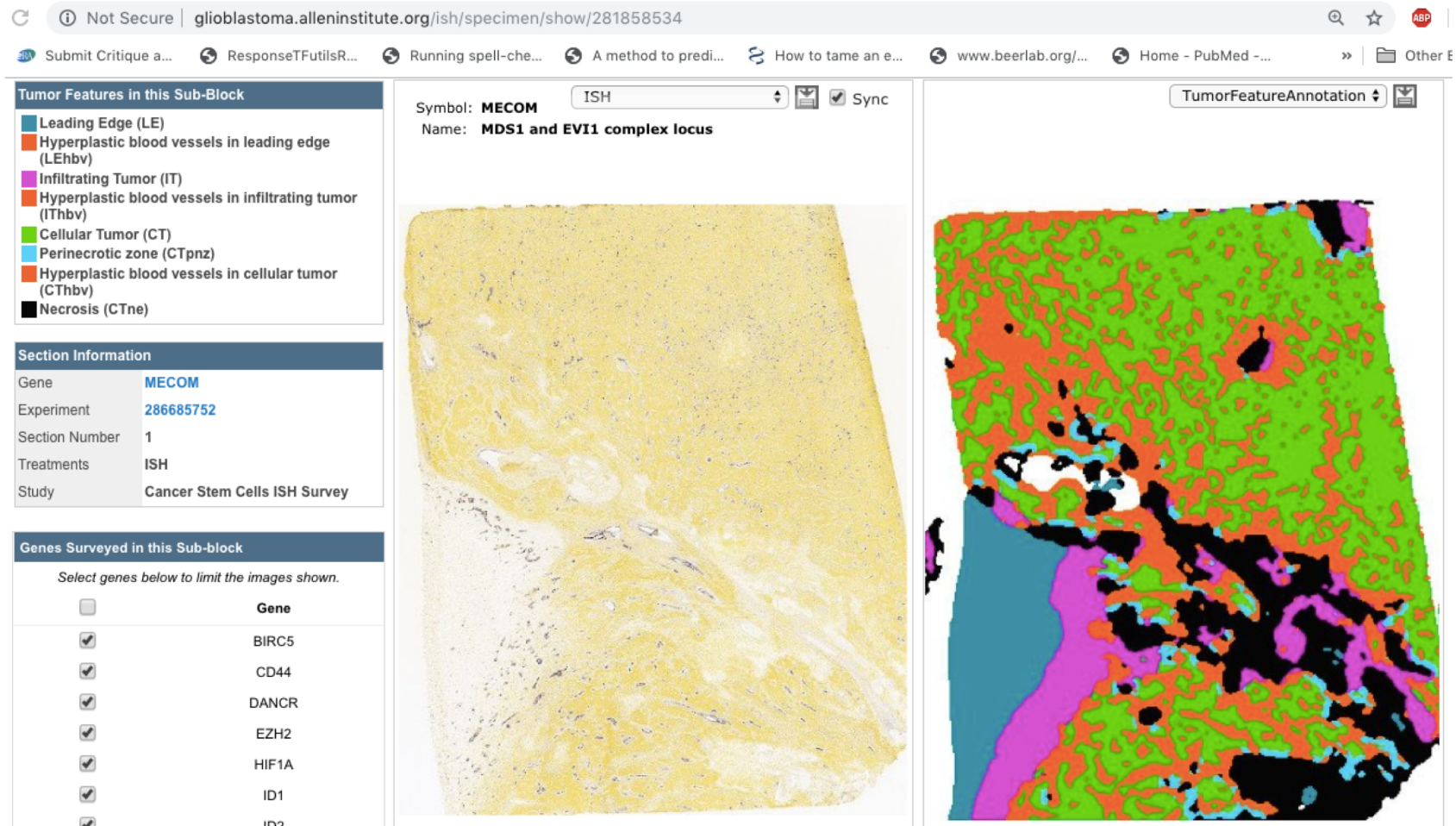
IvyGAP resources: RNA-seq -- complex design

RNA-Seq: RNA sequencing data for anatomic structures identified in the Anatomic Structures ISH Survey and putative cancer stem cell clusters isolated by laser microdissection

- **Anatomic Structures RNA-Seq**: Screen of 5 structures (Leading Edge, Infiltrating Tumor, Cellular Tumor, Microvascular Proliferation, and Pseudopalisading Cells Around Necrosis) identified by H&E staining. A total of 122 RNA samples were generated from 10 tumors.
- **Cancer Stem Cells RNA-Seq**: Screen of 35 clusters of putative cancer stem cells identified by ISH with a 17 reference probe subset (validated in the Cancer Stem Cells ISH Survey). A total of 148 RNA samples were generated from 34 tumors.

Specimen Metadata: De-identified clinical data for each patient and tumor.

App #1: Accessing annotation of an IvyGAP tumor tissue block: CThbv (hyperplastic blood vessels in cellular tumor) is unusually prevalent



https://vjcitn.shinyapps.io/ivyglimpse/

Apps | Submit Critique a... | ResponseTFutilsR... | Running spell-che... | A method to predi... | How to tame an e... | www.beerlab.org/... | Home - PubMed -.

IvyGAP explorer: expression, clinical, and image-based data for glioblastoma patients; see background panel for more details

subBlockDetail features for selectables scatterplot

x

normalized_area_cthbv

y

normalized_area_ct

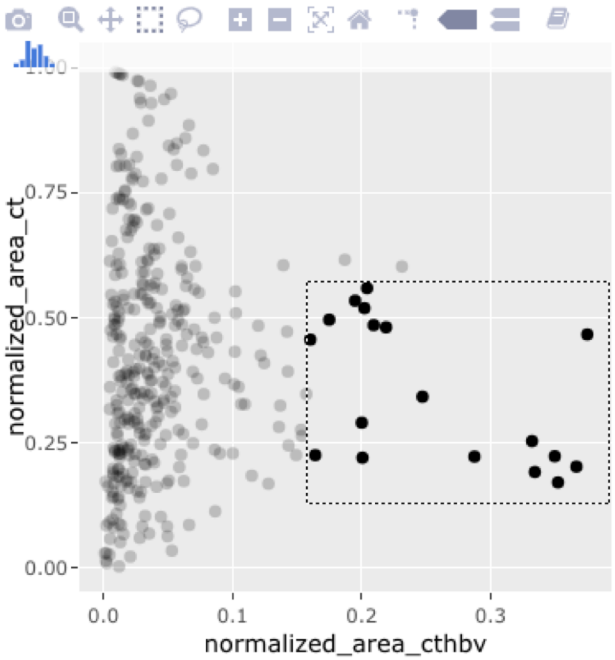
cbioP sets

General: Ras-Raf-MEK-Erk/JNK signaling (26 genes)

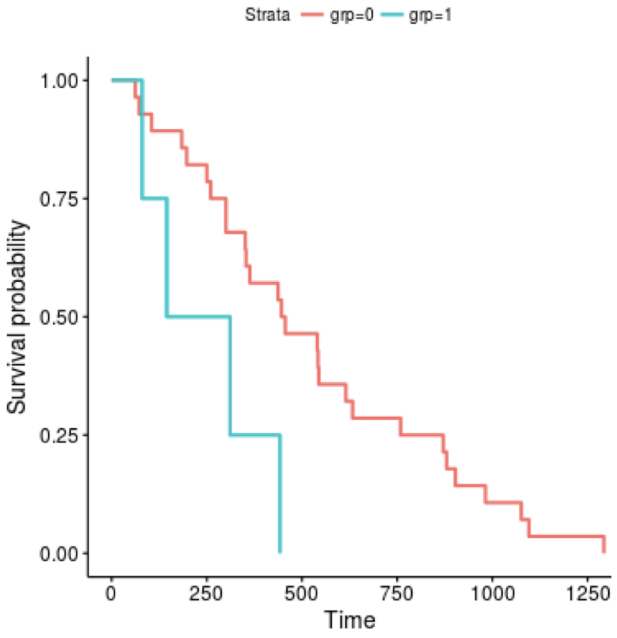
Supported by NCI ITCR U01 CA214846 and U24 CA180996

basic | vocabulary | background

hover over for tumor donor ID; partition data by dragging over points to select; click on a specific point to visit IvyGAP clinical specimen page for that sample



Kaplan-Meier, grp=1 for selected donors



Bioconductor/cloud in ivyGlimpse app

- Underlying Bioconductor package: ivygapSE (SummarizedExperiment unites sample-level and assay data)
- Expression data (not shown here) organized by cBioPortal pathway
- Rstudio shiny (and plotly) code in the package defines the app
- Rstudio shinyapps.io serves the app publicly ($39/month billed to the publishing author, various machine configurations available, performance diagnostics and usage logs)
- Developers and users take advantage of **containerization** and **elasticity** managed by Rstudio

Part 2: Data/metadata synthesis over NCBI SRA and other institutional archives: Sean Davis, NCI

OmicIDX 1.0 OAS3

/openapi.json

What is this?

This is the OmicIDX API for accessing and analyzing omics metadata.

Background

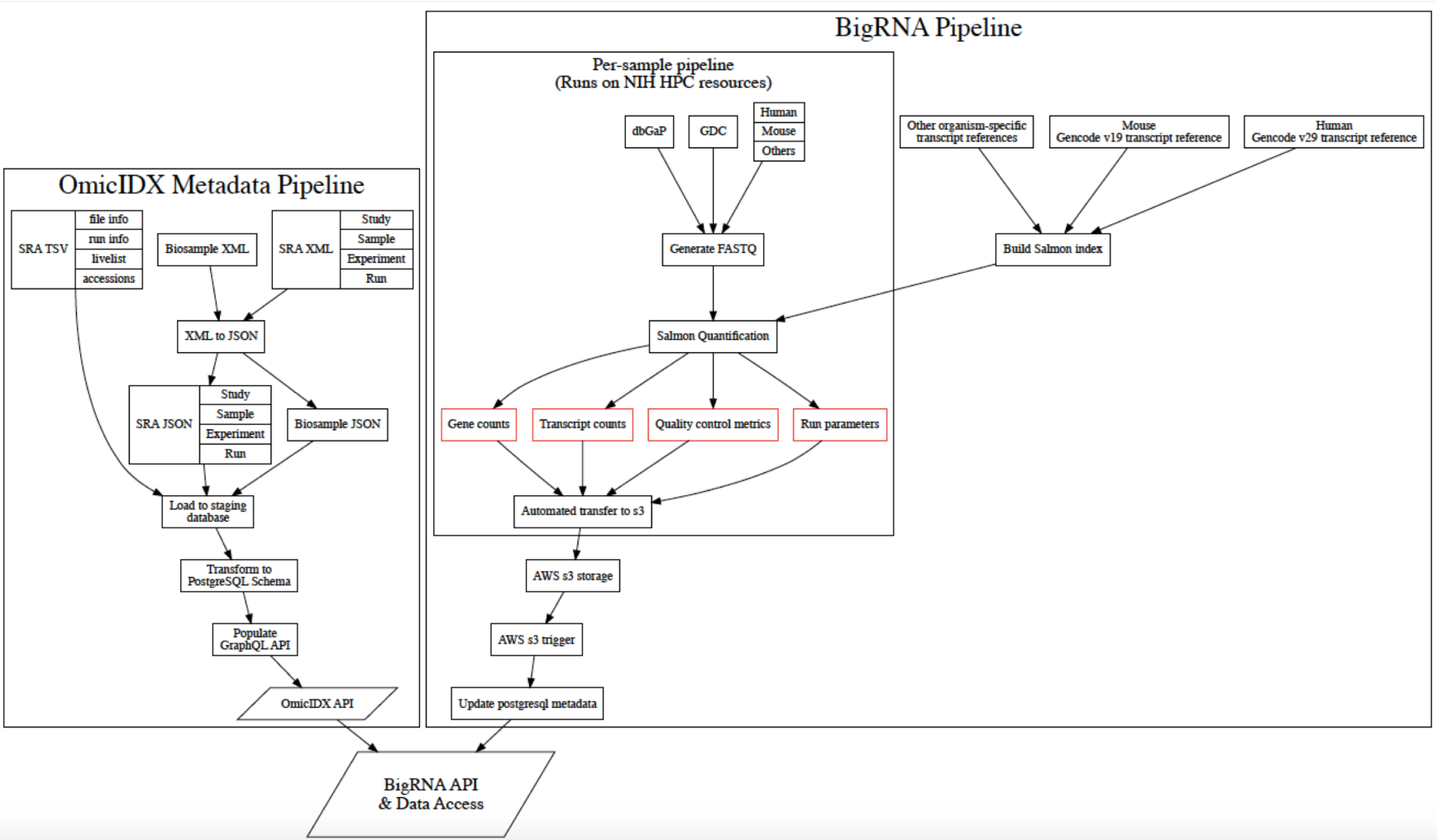
The practice of Data Science often starts with finding, extracting, and organizing the data into systems that are fit for purpose. With the growth of genomics data resources, there are opportunities for large scale data reuse. Furthermore, the corpus of so-called "metadata" that detail the biological materials, experimental variables, and protocols and methods is now a large and rich dataset itself.

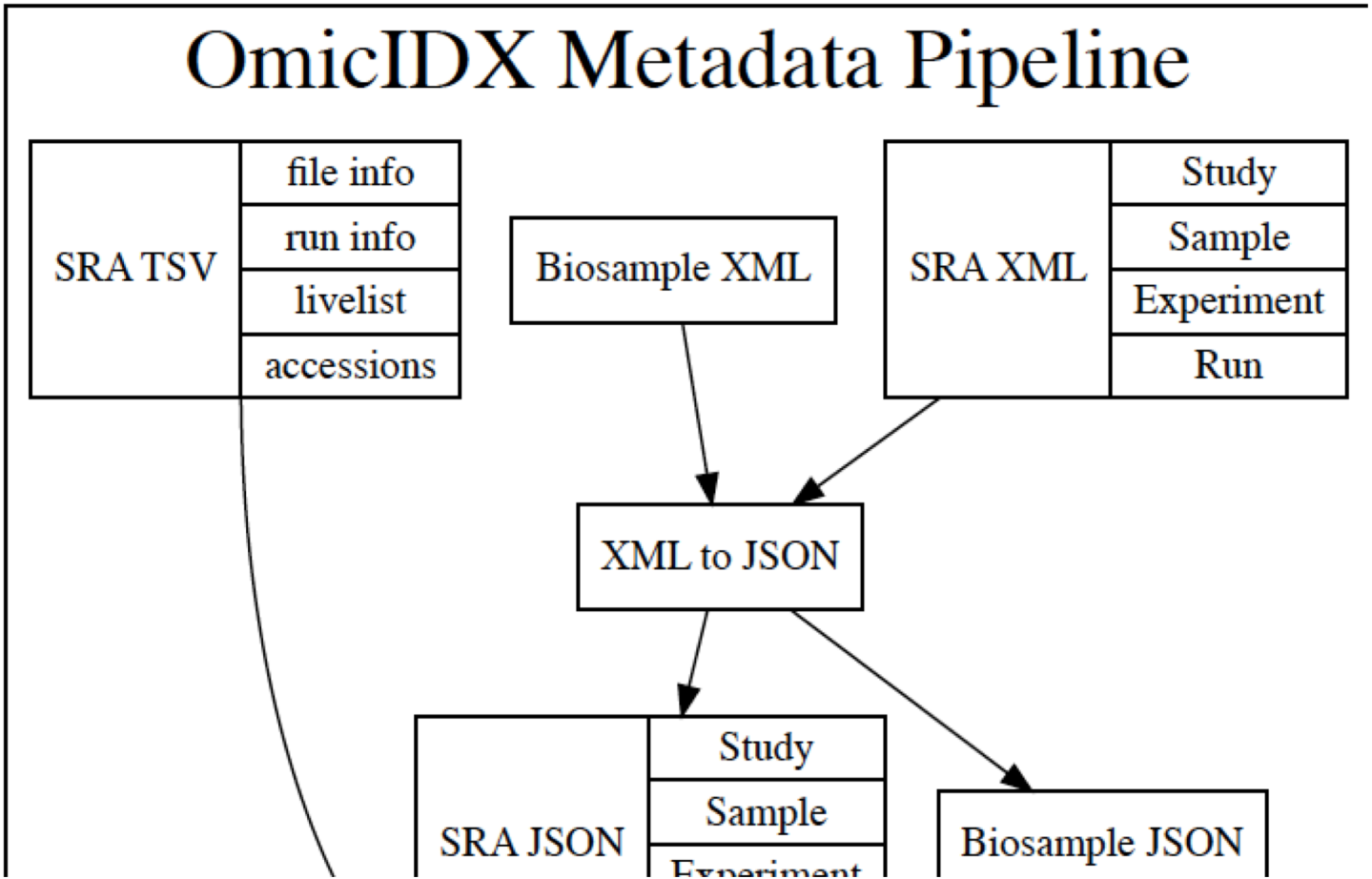
OmicIDX

The OmicIDX project collects disparate metadata from public genomics data repositories and transforms it into several forms that render it fit-for-purpose for large-scale and granular processing. Tasks such as indexing and searching, metadata enrichment with ontologies, and natural language processing all benefit from data resources that are available in bulk and computable formats.

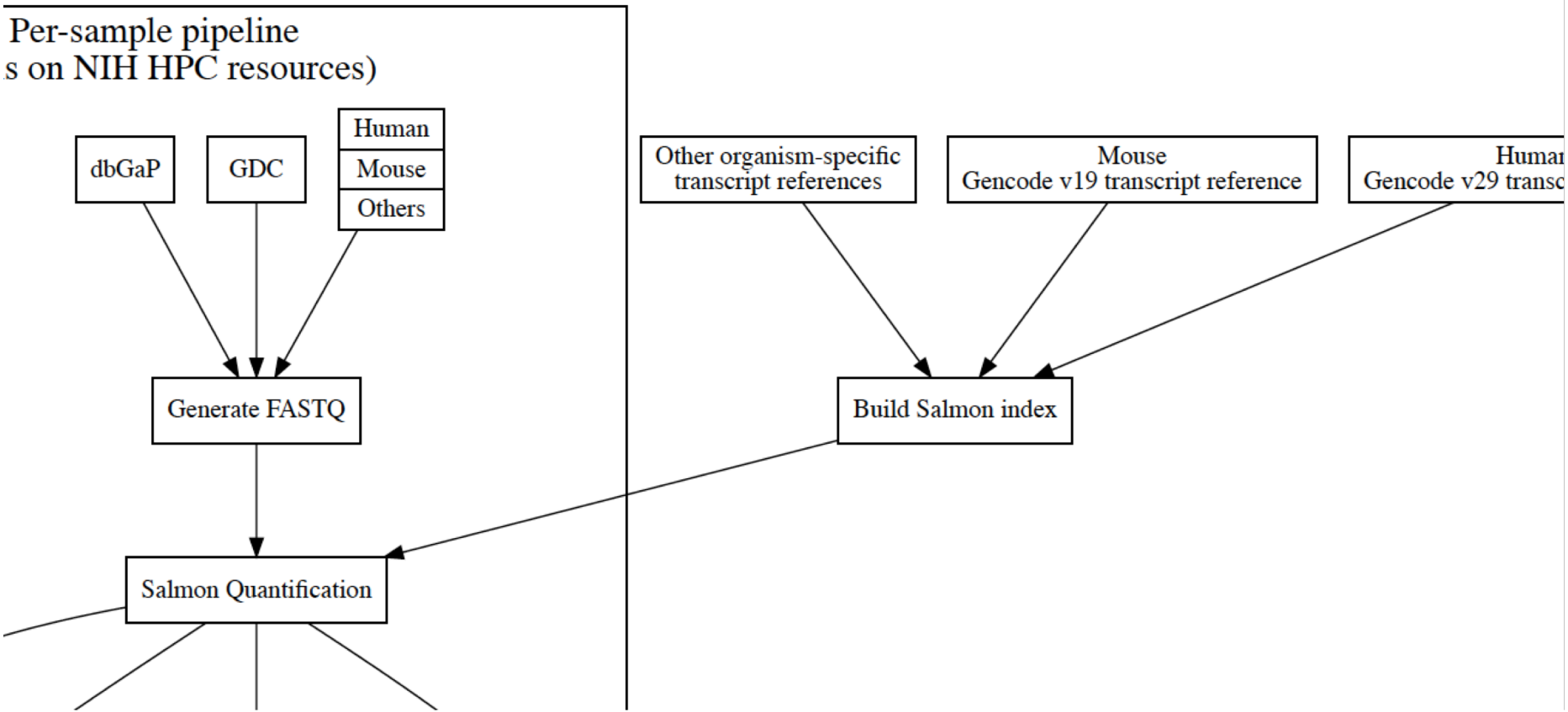
What is an API

A web-based Application Programming Interface (API) uses the same technology as your browser. However, rather than you directing your browser to access information, an API is typically accessed by another piece of sofware. This software sends a request to the API (just a webserver running somewhere) in a format that the server will understand. The server then processes the request and returns a result, typically not in the form that is meant to be viewed on the screen but instead in a format that computers (and often humans) can read.


BigRNA Pipeline
Per-sample pipeline
(Runs on NIH HPC resources)
dbGaP
GDC
Human
Mouse
Others
Other organism-specific transcript references
Mouse Gencode v19 transcript reference
Human Gencode v29 transcript reference
Generate FASTQ
Build Salmon index
Salmon Quantification
Gene counts
Transcript counts
Quality control metrics
Run parameters
Automated transfer to s3
AWS s3 storage
AWS s3 trigger
Update postgresql metadata
OmicIDX Metadata Pipeline
SRA TSV
file info
run info
livelist
accessions
Biosample XML
SRA XML
Study
Sample
Experiment
Run
XML to JSON
SRA JSON
Study
Sample
Experiment
Run
Biosample JSON
Load to staging database
Transform to PostgreSQL Schema
Populate GraphQL API
OmicIDX API
BigRNA API
& Data Access


OmicIDX Metadata Pipeline
SRA TSV
file info
run info
livelist
accessions
Biosample XML
SRA XML
Study
Sample
Experiment
Run
XML to JSON
SRA JSON
Study
Sample
Biosample JSON

BigRNA Pipeline



App #2: ca43k -- 43000 cancer transcriptomes **not in TCGA**

Objectives

- Unified representation of uniformly preprocessed RNA-seq from SRA -- Sean Davis' BigRNA project (now up to 700,000 transcriptomes quantified using salmon)
- Unified multiplexed access to quantifications via HDF Scalable Data Service (thanks to John Readey of HDF Group)
- Searchable comprehensive sample-level metadata
- Familiar programming patterns to filter and analyze
- "Fill your cart" with the transcriptomes you want

Bioconductor:Cancer43K

Full text search over genomic metadata on 43000 cancer transcriptomes exclusive of TCGA, retrieved from NCBI SRA.

Explanatory video

File an issue

Search studies for

squamous cell carcinoma

Click on rows of 'titles' table to add studies to cart.

SRP002326 SRP006575

A restfulSE is returned; RNA-seq quantifications by salmon.

download SE | Clear tabs. | Clear titles. | Stop app.

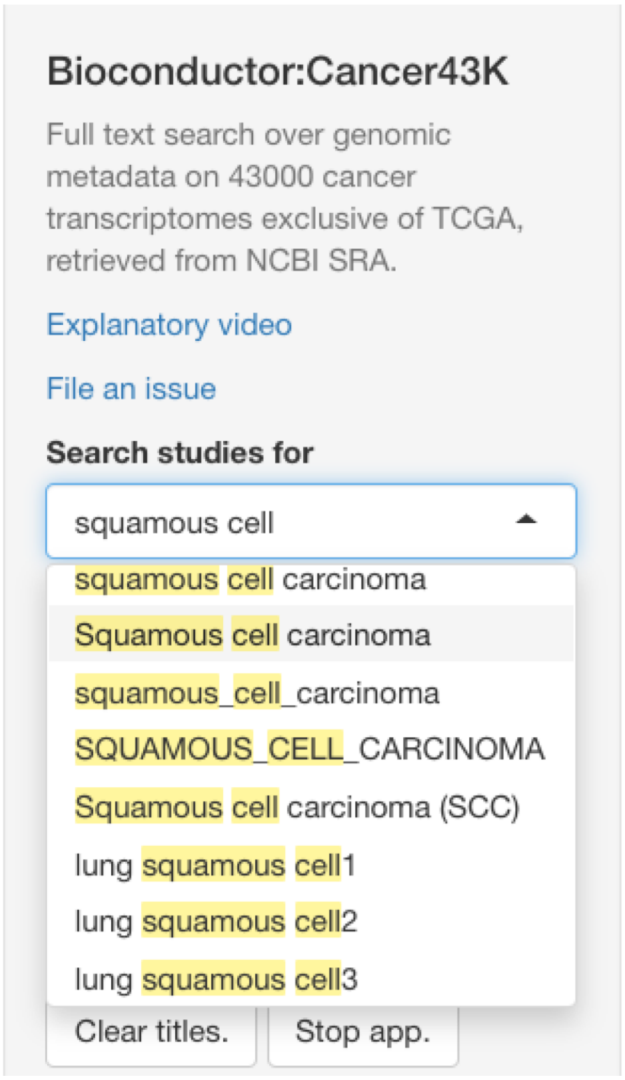
Tabs will appear for studies using selected terms in metadata

Click on tab to see sample.attributes for all experiments in the study, derived with SRAdbV2

titles SRP002326 SRP114925 SRP070156 SRP057020 SRP068450 SRP075613 SRP010129 SRP006575 SRP090586 SRP060234 SRP076496 SRP002009 ERP022968 SRP078505 SRP126169 SRP020473 SRP028180 SRP034698 SRP031478 SRP041647 SRP072163 about

Show 10 entries Search:

	pmid	hits	study	title
1	NA	squamous cell carcinoma	ERP022968	RNAseq experiment of non-exposed skin, Actinic Keratosis, Intraepdimeral Carcinoma and Squamous Cell Carcinoma
2	20174472	oral squamous cell carcinoma	SRP002009	RNA-Seq of oral squamous cell carcinomas and matched normal tissues
3	20459774	adenosquamous cell carcinoma	SRP002326	Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumors and matched controls
4	22929540	head and neck squamous cell carcinoma	SRP006575	Transcriptional profiling of lncRNAs and novel transcribed regions across a diverse panel of archived human cancers



Autocompleting search widget

- selectize.js very performant over vectors of options
- What should the options be?
- Tokenize study title, [abstract], and all field names and field values
- Lots of tokens, a 'facilitated' lexicographic search

What comes back for a Bioconductor user

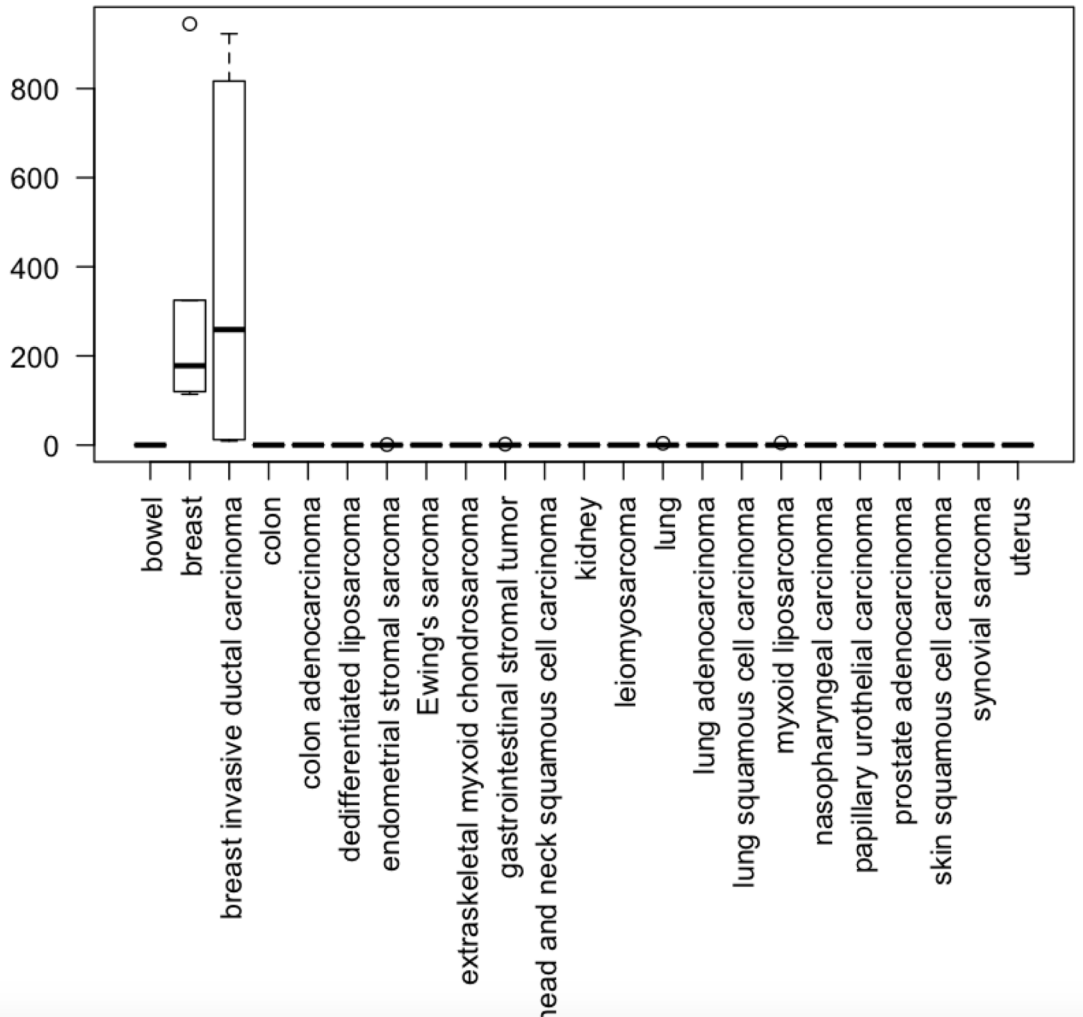
```
> x = readRDS("SE-2019-05-23.rds")
> x
class: RangedSummarizedExperiment
dim: 58288 133
metadata(3): rangeSource SRP002326 SRP006575
assays(1): counts
rownames(58288): ENSG00000000003.14 ENSG00000000005.5 ...
  ENSG00000284747.1 ENSG00000284748.1
rowData names(4): gene_type gene_id gene_name havana_gene
colnames(133): SRX019280 SRX019281 ... SRX176119 SRX176120
colData names(4): experiment_accession experiment_platform
  study_accession study_title
> table(x$study_title)

                                          Transcriptional profiling of lncRNAs
and novel transcribed regions across a diverse panel of archived human cancers

                                                                             96
Ultra-high throughput sequencing-based small RNA discovery and discrete statisti
cal biomarker analysis in a collection of cervical tumors and matched controls

                                                                             37
```

Cart checked out with two studies selected

salmon quantifications are in the cloud, ready

```
> assay(x)
<58288 x 133> DelayedMatrix object of type "double":
                    SRX019280 SRX019281 SRX019282 ...  SRX176119  SRX176120
ENSG00000000003.14          0         0         0   .  109.64368  242.28540
 ENSG00000000005.5          0         0         0   .    0.00000    0.00000
ENSG00000000419.12          0         0         0   . 1168.00027  257.99989
ENSG00000000457.13          0         0         0   .  218.83636  572.78447
ENSG00000000460.16          0         0         0   .   85.75431  244.28797
                ...         .         .         .   .          .          .
 ENSG00000284744.1          0         0         0   .   2.703193   0.000000
 ENSG00000284745.1          0         0         0   .   0.000000   0.000000
 ENSG00000284746.1          0         0         0   .   0.000000   0.000000
 ENSG00000284747.1          0         0         0   .  30.687233  21.091629
 ENSG00000284748.1          0         0         0   .   2.036036   0.000000
```

The authors of SRP006575 identify a lincRNA on chr10 that seemed specific -- a breast cancer biomarker? We can swiftly confirm



Components employed in the evolution of OmicIDX/BigRNA + (current) - (not in use at present)

- **Apache Spark** (google and AWS)
+ **Apache Beam** - Implement batch and streaming data processing jobs that run on any execution engine/
+**GCP Cloud Dataflow** Simplified stream and batch data processing, with equal reliability and expressiveness
- **AWS Lambda** functions
- **AWS DynamoDB** - key-value and document database that delivers single-digit millisecond performance at any scale
+ **AWS RDS** - relational database service
+ **Google BigQuery**
+ **Kubernetes**
+ **Docker**
- **Github actions** - "automate your workflow from idea to production."
+ **Github**
+ **AWS container registry**
- **Google container registry**
- **AWS Athena** - serverless interactive query service
+ **AWS S3**
+ **Google Cloud Storage**
+ **Elasticsearch** (hosted service)
- **Google Genomics Pipeline API**

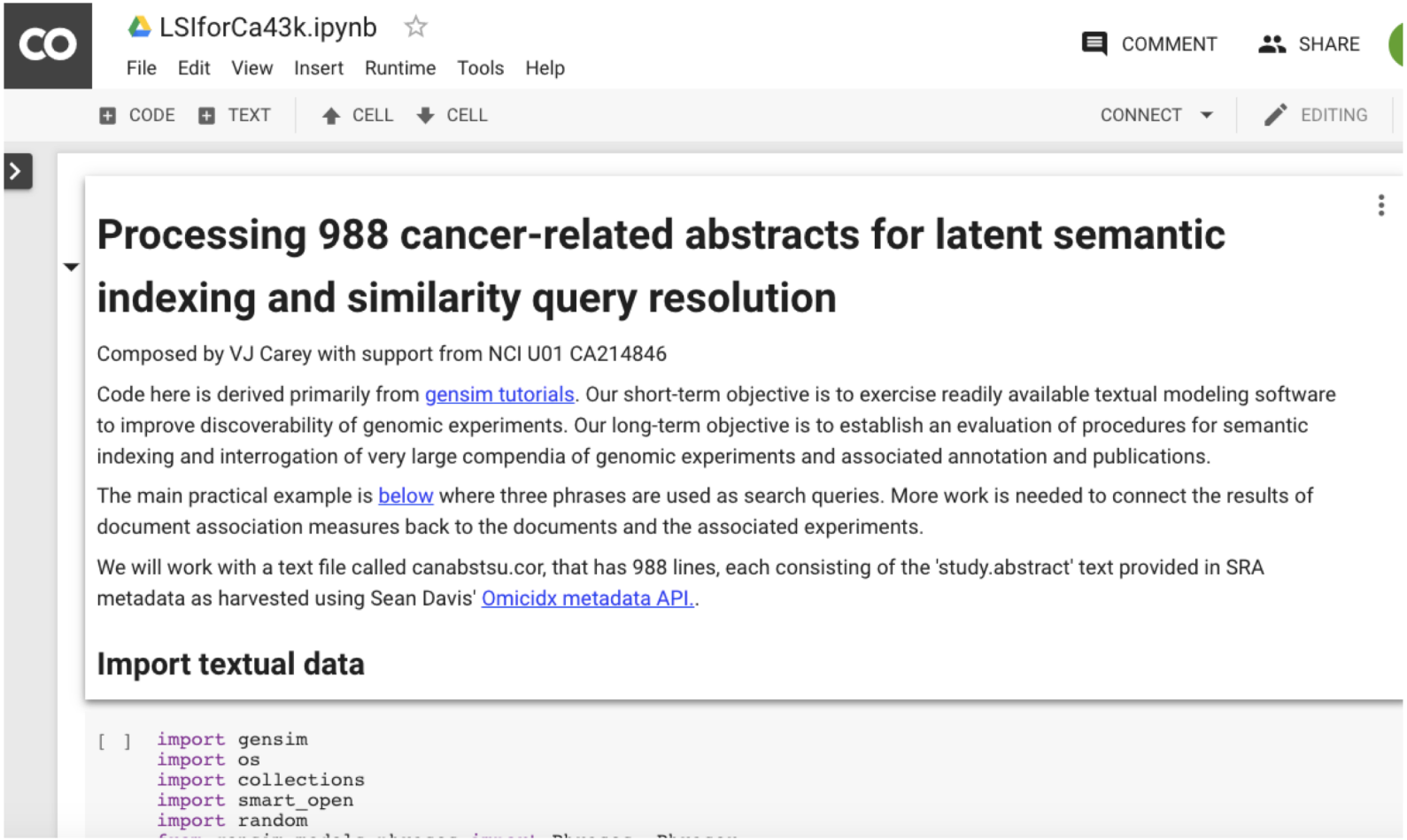
Concluding comments (1)

- Tension: integrated software ecosystem *vs.* distributed independent microservices glued together
  - Bioconductor gains a lot of momentum and reliability by using a **uniform packaging protocol** and **comprehensive continuous integration** for analytic software, genome annotation, and exemplary experiments
  - DelayedArray protocol provides one approach to working with microservice-based utilities without abandoning familiar programming patterns
  - Reliability assurance for distributed resource coordination is complicated and significant attention to fault-tolerance cannot be avoided
    - exception handling cannot be an afterthought

Concluding comments (2)

- Hypothesis: We can get **elasticity** of task-focused environments and **scalability** of tools using Bioconductor's methods
  - Martin Morgan: BiocParallel over Kubernetes clusters -- developers use familiar code patterns but back end is determined at run time
  - Levi Waldron: bioconductor_devel github repository addresses endowment of docker containers, and use of singularity
  - dockstore.org: Bind workflow programs (CWL, WDL, …) to docker containers -- launch in hosted environments with a button

Parting shots -- silos I'd like to smash. 1) Statistical semantics of genomic metadata archives



2) Integrative proteogenomics -- many potential points of contact between CPTAC and Bioconductor

pgconsis: assess consistency of proteogeomics data in breast cancer

gene

BCL2

main about