Bioconductor:Cancer – Towards cloud-scale interactive cancer genomics, October 2018

S Davis, AC Culhane, M Ramos, H Pages, BJ Stubbs, S Gopaulakrishnan, S Pollack, B Halbe-Kains, L Waldron, MT Morgan, V Carey

2018-10-13

Introduction

Strategic thinking about computing for cancer genomics involves a panoply of concepts evolving in an unstable technological domain.[1] In this computable document, we lay out a series of concepts and examples around which our basic developmental strategy for cloud-oriented Bioconductor:Cancer can be understood. **All demonstrations are based on the "devel branch" of Bioconductor in R 3.5.**

Basic map of the situation

We will use the term "architecture" in conjunction with the following broad categories of relevance to computing for cancer genomics.

- **Conceptual architecture:** How are formal ontologies[2], APIs, and concrete data tables used together efficiently to identify key biological and therapeutic processes in cancer?[3]
- **Data architecture:** How are cohorts, trials, experiments, samples, quantifications, and annotations best represented for efficient solutions to problems arising in computational biology of cancer? Are cloud-scale approaches to managing and interrogating "big data" effectively usable by a community of researchers with diverse interests and computational skills?
- **Analysis architecture:** How can advances in conceptual and data architecture help to accelerate the development of compelling new interpretations of existing and new experiments in cancer genomics?

In what follows we aim to get very concrete about the ways in which the Bioconductor:Cancer project delivers advances in these areas.

[1] One snapshot of current approaches can be found at Broad's CGCA site `https://www.broadinstitute.org/cancer/cancer-genome-computational-analysis`

[2] See the OBO Foundry Principles document (`http://www.obofoundry.org/principles/fp-000-summary.html`).

[3] The Genomics API concept is helpfully reviewed in Swaminathan et al. [2016].

Conceptual architecture for FAIRness in cancer genomics

Ontology tools

Results of experiments and trials in human cancer produce information in many domains.

The ontoProc package simplifies usage of key ontologies in the Open Biological Ontologies Foundry. We consider the oncotree vocabulary produced at Sloan-Kettering.[4]

```
library(ontoProc)
otr = getOncotreeOnto()
grep("Breast", otr$name, value=TRUE)[1:4]
```

```
##                                 NCIT:C12971
##                                    "Breast"
##                                NCIT:C139532
##            "Breast Cancer by AJCC v8 Stage"
##                                NCIT:C139533
## "Breast Cancer by AJCC v8 Anatomic Stage"
##                                NCIT:C139534
##  "Anatomic Stage 0 Breast Cancer AJCC v8"
```

There are 43 terms involving `Breast` in the vocabulary [5]. Relationships among these terms are encoded in `otr`. The collection of ontologies usable in this way can be gleaned from

```
grep("Onto", ls("package:ontoProc"), value=TRUE)
```

```
## [1] "getCellLineOnto"    "getCellOnto"
## [3] "getCellosaurusOnto" "getChebiOnto"
## [5] "getDiseaseOnto"     "getEFOOnto"
## [7] "getGeneOnto"        "getHCAOnto"
## [9] "getOncotreeOnto"
```

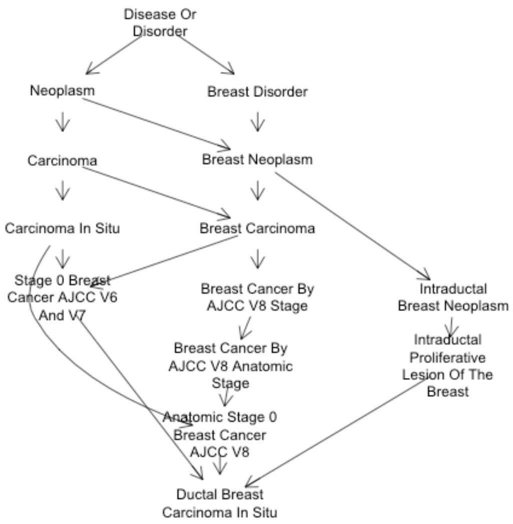
APIs and microservice concepts[6]

It may be fruitful to construe the solutions to many problems arising in cancer genome analysis as "microservices". Numerous projects have implemented REST APIs to expose metadata and data about resources relevant to cancer genomics.

As an illustration of the basic concept, S. Davis of NCI has created the SRAdbV2 package[7], which provides capacity for R/Bioconductor users to survey all sequencing experiments housed in NCBI SRA. Metadata that are continuously developed at SRA are transformed for "serverless" interrogation using Elasticsearch. Thus

[4] `http://oncotree.mskcc.org/#/home?tab=news` shows that this is an evolving resource, as one might expect. What we work with is a slight variation created using NCIT terms that coincide with oncotree terms, as distributed at `http://purl.obolibrary.org/obo/ncit/ncit-oncotree.obo`

[5] A view of relationships among 10 oncoTree terms related to "Breast".


Disease Or Disorder
Neoplasm
Breast Disorder
Carcinoma
Breast Neoplasm
Carcinoma In Situ
Breast Carcinoma
Stage 0 Breast Cancer AJCC V6 And V7
Breast Cancer By AJCC V8 Stage
Intraductal Breast Neoplasm
Breast Cancer By AJCC V8 Anatomic Stage
Intraductal Proliferative Lesion Of The Breast
Anatomic Stage 0 Breast Cancer AJCC V8
Ductal Breast Carcinoma In Situ

[6] "Microservice architecture has taken hold in application design for a number of reasons ..., but ultimately, the fundamental driver has been the opportunity for significant increase in productivity. (Williams et al. [2016])

[7] `https://github.com/seandavi/SRAdbV2`

```
library(SRAdbV2)
oidx = Omicidx$new()
allhrec = oidx$search(q="sample_taxon_id: 9606")
allhrec$count()
```

```
## [1] 1291933
```

gives the number of records involving human samples. A selection of fields that can be interrogated to acquire accession numbers for experiments of interest is

```
lk = allhrec$scroll()
sample(mm <- names(lk$yield()), size=5)
```

```
## [1] "experiment_design"
## [2] "study_LastMetaUpdate"
## [3] "study_Insdc"
## [4] "experiment_broker_name"
## [5] "sample_description"
```

There are 86 fields whose semantics must be known to enable effective searching[8]. Some of these fields include subfields whose structure varies from experiment to experiment.

It is useful to contrast this approach to the predecessor package, SRAdb, which was based on a 35GB SQLite database, updated semiannually, that was installed on each user's system. SRAdbV2 is based on a chain of transformations from NCBI's XML metadata compendium, harvested every other week, leading to a service that responds to detailed queries in the Lucene query idiom. The stability and local nature of the SRAdb approach is traded for a potentially more volatile resource that is very light weight and universally accessible.

The REST API underlying the SRAdbV2 service is documented at `https://api-omicidx.cancerdatasci.org/sra/1.0/ui/`.

Data architecture: Exploring BigQuery and HDF Cloud as back ends for cancer genomics

We consider two cloud-scale solutions for multiomic data in cancer: Google BigQuery, as employed in the ISB CGC project, extended from TCGA to the Pancancer Atlas, and the HDF Scalable Data Service. Our approach emphasizes hybridized data interfaces: certain data such as focused annotation and sample attributes can be resident in memory, while voluminous quantifications are managed remotely and queried only when needed.

[8] These fields constitute a small fraction of the overall data model of the NCI Genomic Data Commons; see `https://gdc.cancer.gov/developers/gdc-data-model/gdc-data-model-components`.

Lazy MultiAssayExperiments with BigQuery back end

The SummarizedExperiment class unifies genomic quantifications with metadata about samples, assay features, and experimental protocol. This has been generalized for multi-assay experiments (where tissue from a given sample is characterized along various molecular dimensions). We used the `buildPancanSE` function of the BiocOncoTK package, and the MultiAssayExperiment construction API to develop `blcaMAE`, the Pancancer Atlas data on RNA-seq, miRNA-seq, and Illumina Infinium methylation results for both tumor and normal tissues derived from the TCGA Bladder Cancer cohort.

```
blcaMAE
```

```
## A MultiAssayExperiment object of 6 listed
##  experiments with user-defined names and respective classes.
##  Containing an ExperimentList class object of length 6:
##  [1] rnaseq: SummarizedExperiment with 20531 rows and 408 columns
##  [2] rnaseq_n: SummarizedExperiment with 20531 rows and 19 columns
##  [3] meth: RangedSummarizedExperiment with 396065 rows and 409 columns
##  [4] meth_n: RangedSummarizedExperiment with 396065 rows and 21 columns
##  [5] mirna: SummarizedExperiment with 743 rows and 409 columns
##  [6] mirna_n: SummarizedExperiment with 743 rows and 19 columns
## Features:
##  experiments() - obtain the ExperimentList instance
##  colData() - the primary/phenotype DataFrame
##  sampleMap() - the sample availability DataFrame
##  `$`, `[`, `[[` - extract colData columns, subset, or experiment
##  *Format() - convert into a long or wide DataFrame
##  assays() - convert ExperimentList to a SimpleList of matrices
```

Quantifications are accessed as needed through the DelayedArray protocol:

```
suppressMessages(assay(experiments(blcaMAE)$meth["rs939290",]))
```

```
## <1 x 409> DelayedMatrix object of type "double":
```

```
## Auto-refreshing stale OAuth token.
```

```
## Downloading 2 rows in 1 pages.
```

```
##          TCGA-FD-A5BV ... TCGA-UY-A78L
## rs939290      0.97691   .     0.966996
```

Lazy SummarizedExperiments over HDF Scalable Data Service

HDF5 is widely used for array-structured quantification sets in genomics. The HDF Scalable Data Service (HSDS)[9] leverages the HDF

[9] https://www.hdfgroup.org/solutions/hdf-kita/

data model for efficient design of hierarchical organizations of numerical arrays and their metadata, deployed in an S3 object store, exposed through a REST API. Query resolution is multiplexed.

In this example, we acquire a lazy but richly annotated HSDS-resident image of the BigRNA compendium assembled by Sean Davis of NCI (`http://bigrna.cancerdatasci.org/`). This is a collection of all RNA-seq studies of human-derived samples, uniformly preprocessed and quantified[10] to gene models of GENCODE version 27.

We present it to the user as a SummarizedExperiment, and construct a query to filter the content of the compendium based on transcript location.

```
library(htxcomp) # github.com/vjcitn/htxcomp
htxg = loadHtxcomp()
dim(htxg)
```

```
## [1]  58288 181134
```

```
subsetByOverlaps(htxg, GRanges("chr1", IRanges(1e6,2e6)))
```

```
## class: RangedSummarizedExperiment
## dim: 76 181134
## metadata(1): rangeSource
## assays(1): counts
## rownames(76): ENSG00000008128.22
##   ENSG00000008130.15 ...
##   ENSG00000284372.1 ENSG00000284740.1
## rowData names(0):
## colnames(181134): DRX001125 DRX001126
##   ... SRX999990 SRX999991
## colData names(4): experiment_accession
##   experiment_platform study_accession
##   study_title
```

Notice the number of samples, 181134. Selection of experiments of interest is accomplished using SRAdbV2 to obtain accession codes, say `KP`, and then working with `htxg[,kp]`.

Thanks to John Readey of the HDF Group, HSDS images of the RNA-seq archives from GTEx, Tabula Muris, HapMap, and two single-cell RNA-seq experiments in glioblastoma are available for public use via the HDF Kita Lab referenced above. The Bioconductor restfulSE package can also be used to work with these datasets.

Analysis architecture

We now demonstrate some virtues of the hybrid data architecture consisting of numerical data in a remote bigdata back end, and anno-

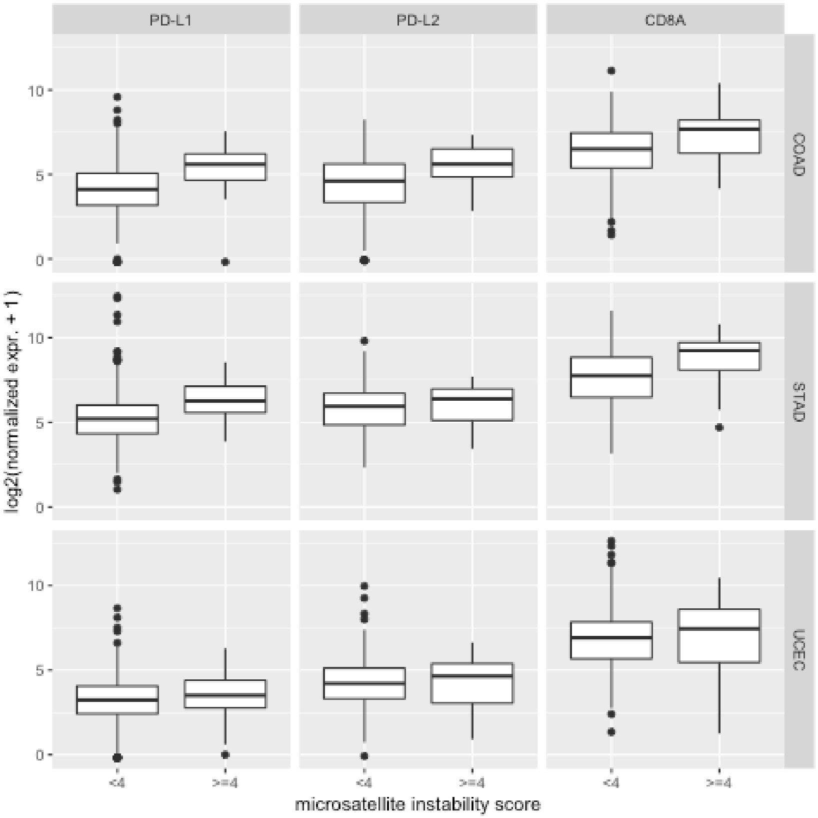
[10] `https://combine-lab.github.io/salmon/getting_started/`

tation and query support provided by R/Bioconductor.

Multitumor, multigene survey of association of gene expression with microsatellite instability: BigQuery back end

Figure[11] 5C of Bailey et al. [2018] indicates that microsatellite instability (MSI) is associated with different expression signatures of immune cell infiltration for adenocarcinomas of colon (COAD) and stomach (STAD), and uterine corpus endometrial carcinoma (UCEC).

The MSI scores developed using MSIsensor are found in Table S5 of Ding et al. [2018]. To reproduce aspects of this finding using the BigQuery Pancancer-atlas back end, we a) authenticate to the BigQuery platform, b) select tumor types and assay for `SummarizedExperiment` construction, c) bind Ding et al.'s MSI values d) acquire and transform the expression values of interest, and e) form the stratified boxplots. The basic findings of Bailey et al. are replicated.



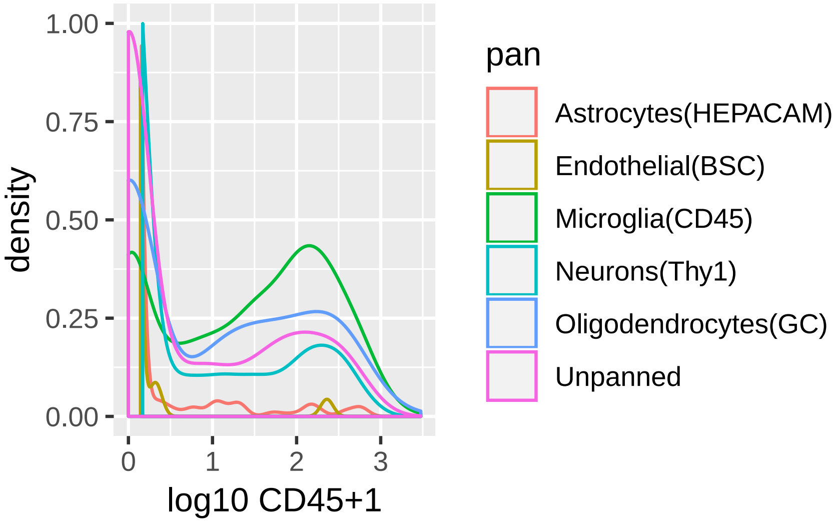
In the code snippet above, `exprByMSI` binds information not available in Pancancer atlas to the RESTful SummarizedExperiment, and, using `assay`, generates a series of SQL queries to BigQuery.

[11] Bioconductor code segment for boxplot panel.

```
library(BiocOncoTK)
# set up gene x tumor table
infilGenes = c("CD274",
  "PDCD1LG2", "CD8A")
names(infilGenes) = c("PD-L1", "PD-L2",
   "CD8A")
tumcodes = c("COAD", "STAD", "UCEC")
combs = expand.grid(tumcode=tumcodes,
  ali=names(infilGenes),
    stringsAsFactors=FALSE)
combs$sym = infilGenes[combs$ali]

# establish BigQueryConnection
bq = pancan_BQ() #billing=[project id])

# build a data.frame for ggplot, using
# bindMSI to add microsatellite instability
# scores  and replaceRownames to convert
# ENTREZ ids (used
# natively in ISB BigQuery
# pancan-atlas) to HGNC symbols
exprByMSI = function(bq, tumcode, genesym,
    alias) {
  if (missing(alias)) alias=genesym
  ex = bindMSI(buildPancanSE(bq, tumcode,
    assay="RNASeqv2"))
  ex = replaceRownames(ex)
  data.frame(
   patient_barcode=colnames(ex),
   acronym=tumcode,
   symbol = genesym,
   alias = alias,
   log2ex=log2(
     as.numeric(
      SummarizedExperiment::assay(
       ex[genesym,]))+1),
   msicode = ifelse(ex$msiTest >= 4,
      ">=4", "<4"))
 }

# apply exprByMSI to each tumor
updateL = lapply(1:nrow(combs), function(x)
   exprByMSI(bq, combs$tumcode[x],
     combs$sym[x], combs$ali[x]))
updated = do.call(rbind, updateL)
# visualize
library(ggplot2)
ggplot(updated,
    aes(msicode, log2ex)) + geom_boxplot() +
    facet_grid(acronym~alias) +
    ylab("log2(normalized expr. + 1)") +
    xlab("microsatellite instability score")
```

Comparing cell-type-specific expression distributions after immunopanning in GBM samples: HSDS back end

As a prelude to single-cell RNA-sequencing of glioblastoma (GBM) tumors from four patients, Darmanis et al. [2017] used immunopanning to increase the proportion of non-neoplastic cells that constitute the "migrating front" of progression of glioblastoma. Antibody to CD45 was used to capture microglial cells. Using the code to the right[12], we compare the distribution of CD45 expression among the classes of cells as labeled in the metadata of GSE84465, the NCBI GEO archive from which the quantifications were derived.



Discussion

We are probing towards a coherent environment of data and annotation services, query facilities, and statistical learning utilities sufficient for advancing knowledge in cancer genomics as rapidly and reliably as possible. We must admit that definition and adoption of best approaches to conceptual architecture (specifically use of ontologies and standard APIs) still needs substantial work throughout the development community. Options for data architecture proliferate rapidly and investments in comparative benchmarking are imperative. Progress in analysis architecture must address adaptability of data ingestion components, propagation of information on data and tool provenance, and mechanisms for contributions of new methods. Developers for Bioconductor:Cancer are striving for progress in all these areas.

[12] Code segment to create density traces.

```
library(rhdf5client)
library(SummarizedExperiment)
library(BiocOncoTK)
library(ggplot2)
cdar = BiocOncoTK::darmGBMcls
ind = match("PTPRC", rowData(cdar)$symbol)
var = gsub("selection: ", "",
        cdar$characteristics_ch1.8)
vals = log10(assay(cdar[ind,])+1)
ddd = data.frame(log10norm=vals, pan=var)
ggplot(ddd, aes(x=log10norm, colour=pan)) +
  geom_density() + ylim(0,1) +
  xlab("log10 CD45+1")
```

Funding support

This work was supported by NCI U01 CA214846, V. Carey, PI, NCI U24 CA180996, M. Morgan PI. Work of Sean Davis was supported by the NCI Center for Cancer Research.

References

Matthew H. Bailey, Collin Tokheim, Eduard Porta-Pardo, Rachel Karchin, Li Ding, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 173(2):371–385.e18, 2018. ISSN 10974172. DOI: 10.1016/j.cell.2018.02.060.

Spyros Darmanis, Steven A. Sloan, Derek Croote, Marco Mignardi, Sophia Chernikova, Peyman Samghababi, Ye Zhang, Norma Neff, Mark Kowarsky, Christine Caneda, Gordon Li, Steven D. Chang, Ian David Connolly, Yingmei Li, Ben A. Barres, Melanie Hayden Gephart, and Stephen R. Quake. Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Reports*, 21(5):1399–1410, 2017. ISSN 22111247. DOI: 10.1016/j.celrep.2017.10.030. URL `https://doi.org/10.1016/j.celrep.2017.10.030`.

Li Ding, Matthew H. Bailey, Eduard Porta-Pardo, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*, 173(2):305–320.e10, 2018. ISSN 10974172. DOI: 10.1016/j.cell.2018.03.033.

Rajeswari Swaminathan, Yungui Huang, Soheil Moosavinasab, Ronald Buckley, Christopher W. Bartlett, and Simon M. Lin. A Review on Genomics APIs. *Computational and Structural Biotechnology Journal*, 14:8–15, 2016. ISSN 20010370. DOI: 10.1016/j.csbj.2015.10.004. URL `http://dx.doi.org/10.1016/j.csbj.2015.10.004`.

Christopher L Williams, Jeffrey C Sica, Robert T Killen, and Ulysses G. J. Balis. The growing need for microservices in bioinformatics. *Journal of Pathology Informatics*, 7(1):45, 2016. ISSN 2153-3539. DOI: 10.4103/2153-3539.194835. URL `http://www.jpathinformatics.org/text.asp?2016/7/1/45/194835`.