Bioconductor:Cancer -- Genome-scale data science for precision oncology

Sean Davis[1], Aedin Culhane[2], Marcel Ramos[3], Herve Pages[4], BJ Stubbs[5], Shweta Gopaulakrishnan[5], Samuela Pollack[2], Benjamin Haibe-Kains[6], Levi Waldron[3], Martin Morgan[7], Vincent Carey[5]*,

[1]National Cancer Institute, [2]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, [3]City University of New York, [4]Fred Hutchinson Cancer Research Center, [5]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, [6]Princess Margaret Cancer Centre, [7]Roswell Park Cancer Institute, *stvjc@channing.harvard.edu

Bioconductor:Cancer (B:C): a software ecosystem that modernizes the data models of R/Bioconductor to allow general statistical analysis and visualization of cloud-scale cancer data. B:C is a collection of rigorously tested software modules addressing reference and variant genome annotation, assay preprocessing and summarization, and integrative statistical learning in multiomic contexts, with a focus on applications in cancer research.

Aim: *Contribute to design and deployment of a Data Science Ecosystem for discovery in cancer genomics*

Needed: Many things!

Conceptual architecture

- *Ontologies of all relevant factors supporting discovery and use*
- *APIs for data and analysis production and consumption*

Data architecture

- *FAIR principles*
- *scalable/reliable/evolvable*

Analysis architecture

- *Verifiable, interoperable components*
- *Environment-agnostic, scalable*

Use cases:

*1) Programmatically survey/query **all** available metadata in NCBI SRA*

*2) Provide scalable access to uniformly preprocessed quantifications for **all** human RNA-seq samples in SRA*

*3) **Interactively** combine **novel** molecular assay results with PanCancer Atlas expression data for new cross-tumor inferences*

*4) **Interactively** assess **signature variation** in immuno-panning for GBM scRNA-seq*

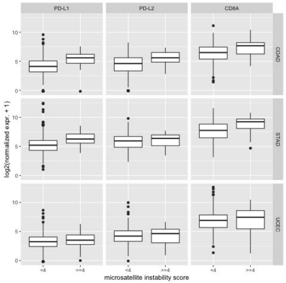
Solutions:

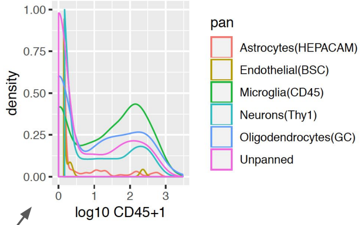
1) `SRAdbv2 (github.com/seandavi)` supports lucene queries over 5 million samples

2) `htxcomp (github.com/vjcitn)` uses HDF Scalable Data Service (HSDS) + Bioconductor `restfulSE` for elementwise access to 181000 human RNA-seq studies

3) `BiocOncoTK` simplifies use of BigQuery pancancer-atlas through `restfulSE`, and binds MSIsensor results

4) `BiocOncoTK::darmGBMcls` is a richly annotated HSDS-backed SummarizedExperiment based on Darmanis et al. Cell Rep. (2017)





B:C directly implements the *FAIR:* (findable, accessible, interoperable, reusable) principles, an initial step in the development of a comprehensive statistical learning environment for the pan-cancer/multi-omic context. Take a flyer or see github.com/vjcitn/aacrPamphlet to pick up the PDF!