Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

MultiAssayExperiment

Software for the integration of multi-omics experiments in Bioconductor

U24CA180996 (PI: Martin Morgan)

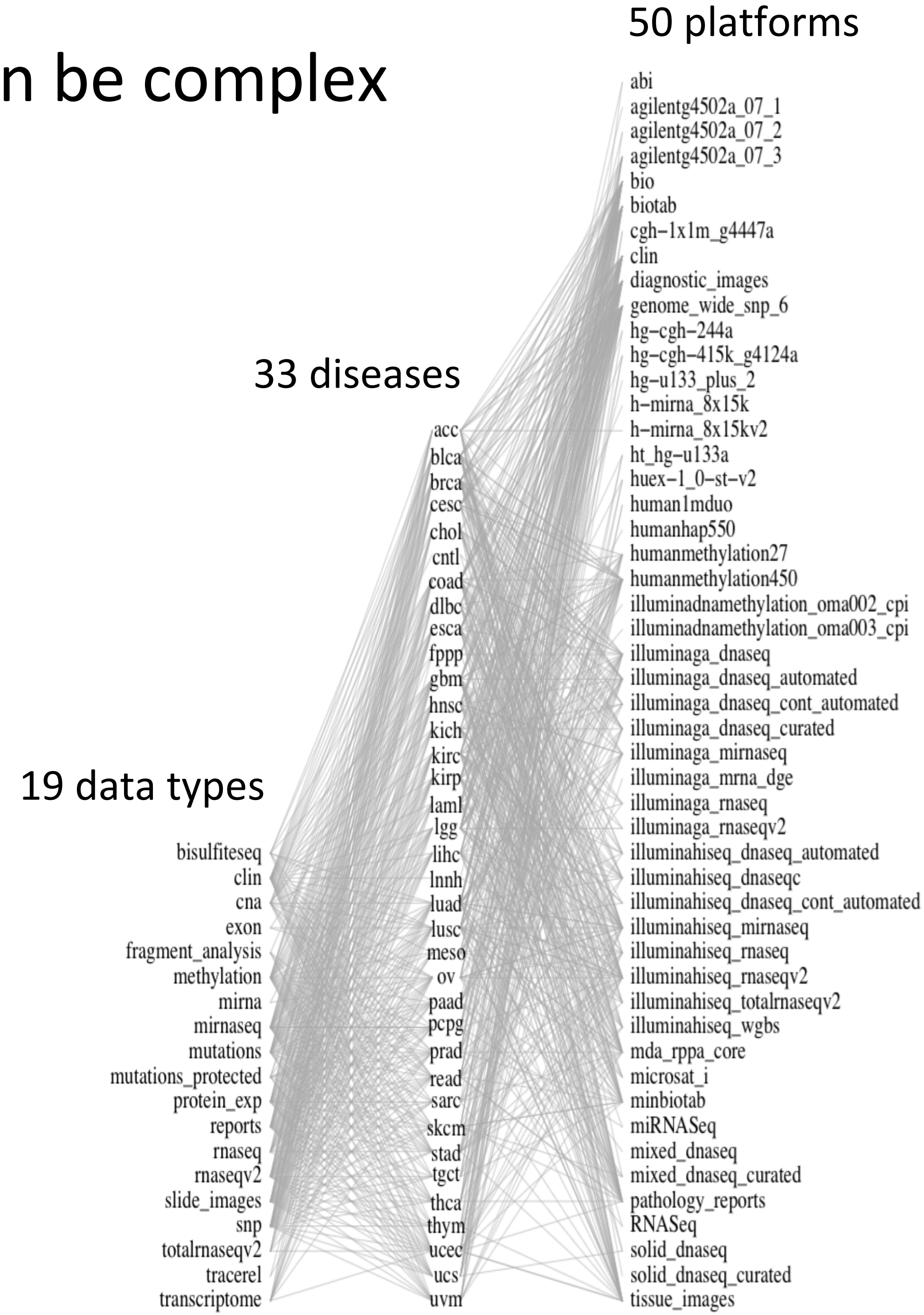
Co-investigators: Levi Waldron, Vincent Carey, Kasper Hansen

CUNY SPH GRADUATE SCHOOL OF PUBLIC HEALTH & HEALTH POLICY

ISPH CUNY Institute for Implementation Science in Population Health

Multi-assay experiments can be complex

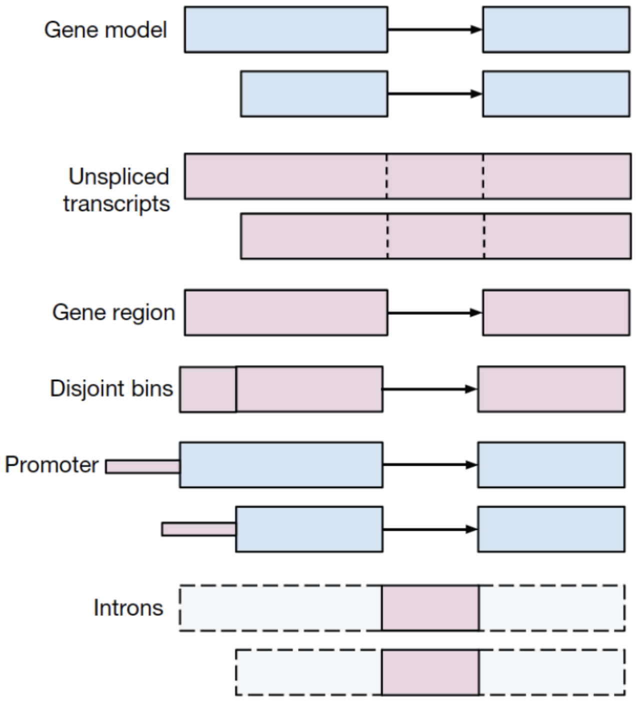
Diseases, platforms, and data types of The TCGA



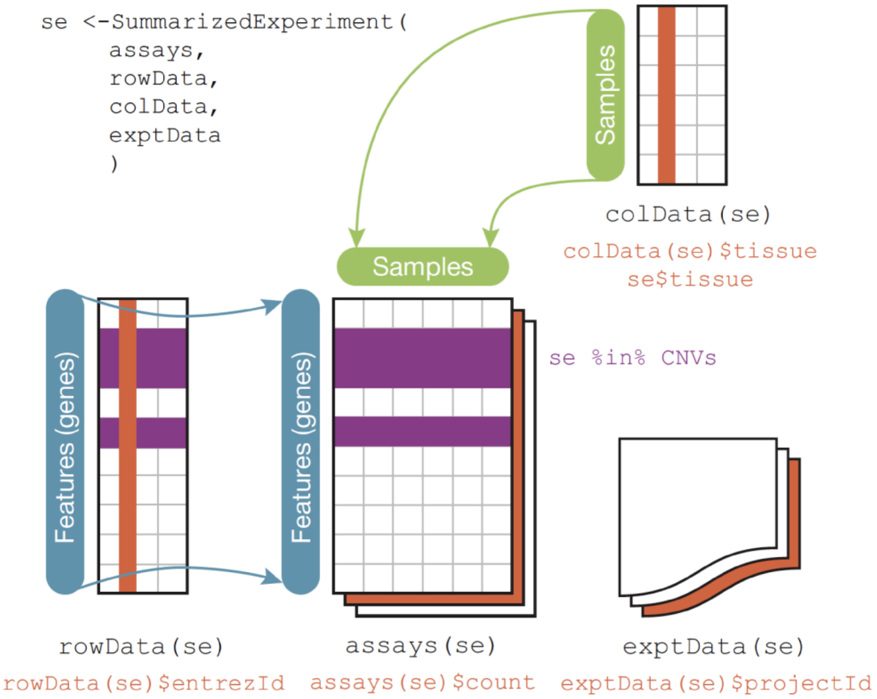
Credit: Marcel Ramos

Why Bioconductor?

1,400 packages on a backbone of data structures



The Genomic Ranges algebra



```
se <-SummarizedExperiment(
      assays,
      rowData,
      colData,
      exptData
      )
```

The integrative data container *SummarizedExperiment*

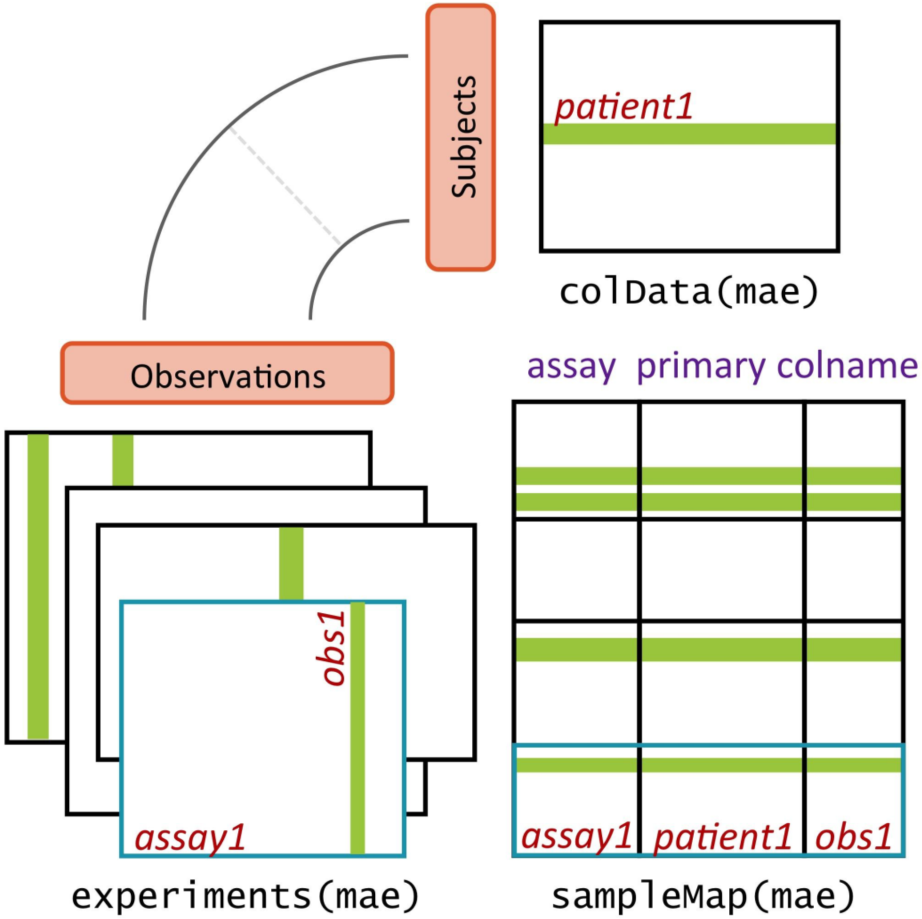
Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12,** 115–121 (2015).

The need for MultiAssayExperiment

Need a core data structure to:

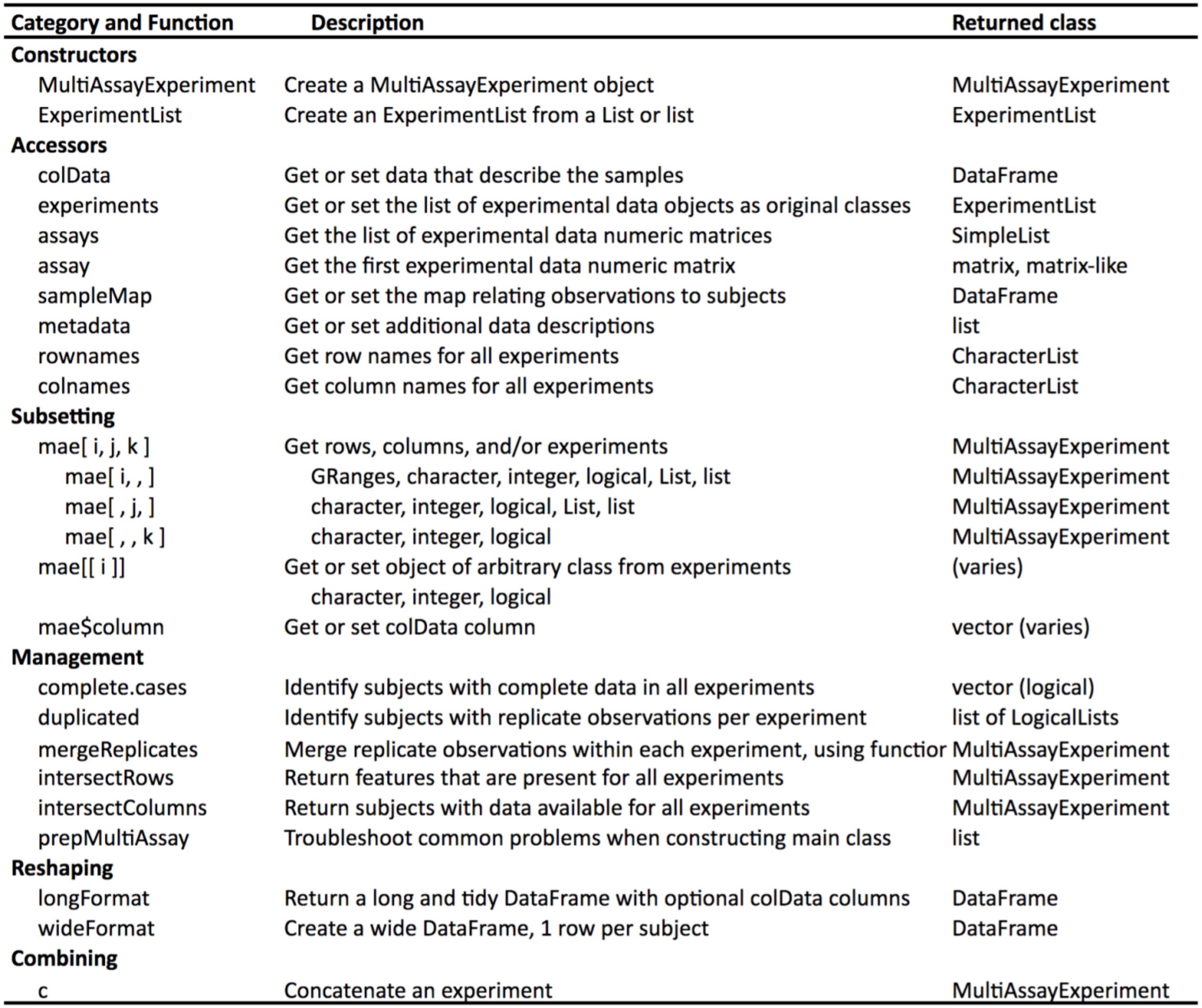
- harmonize single-assay data structures
- relate multiple assays & clinical data
- handle missing and replicate observations
- accommodate ID-based and range-based data
- support on-disk representations of big data

MultiAssayExperiment design



Credit: Marcel Ramos

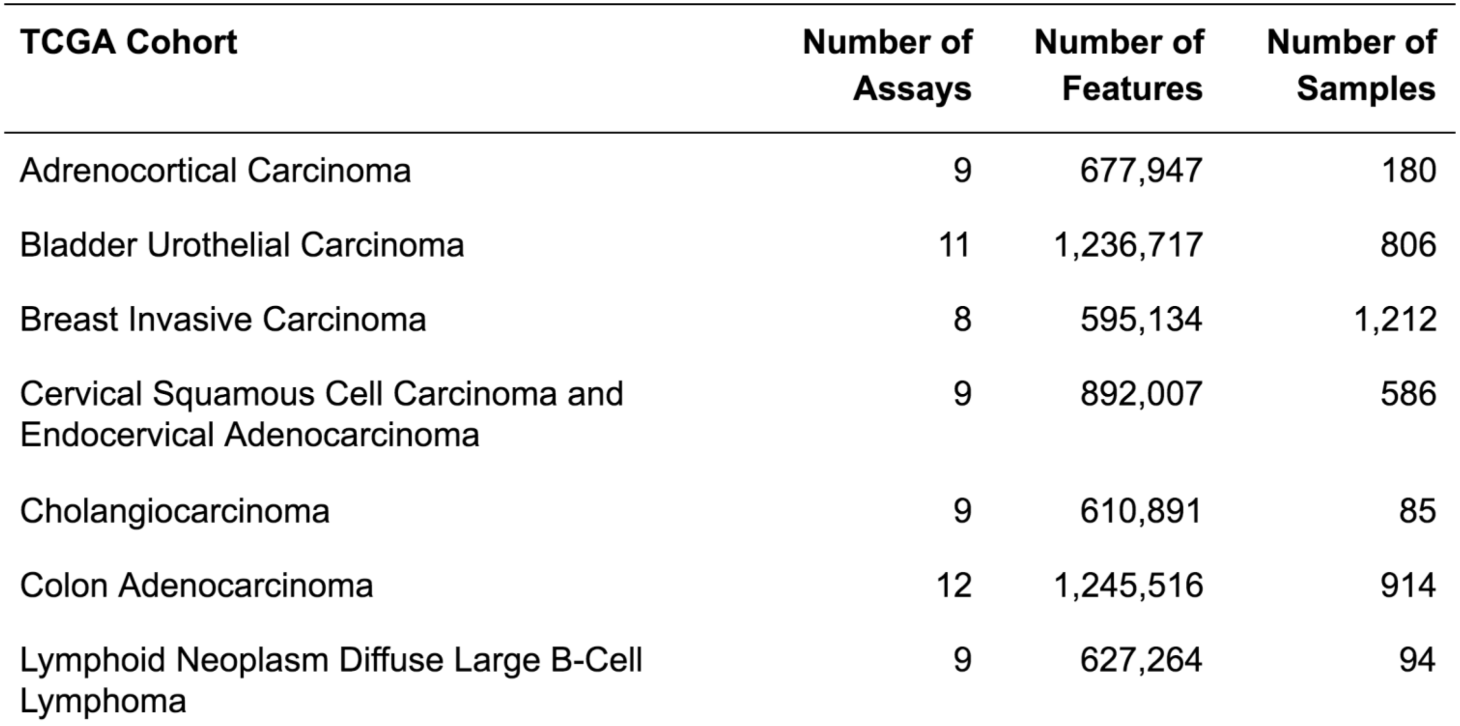
The MultiAssayExperiment API



Category and Function	Description	Returned class
Constructors		
MultiAssayExperiment	Create a MultiAssayExperiment object	MultiAssayExperiment
ExperimentList	Create an ExperimentList from a List or list	ExperimentList
Accessors		
colData	Get or set data that describe the samples	DataFrame
experiments	Get or set the list of experimental data objects as original classes	ExperimentList
assays	Get the list of experimental data numeric matrices	SimpleList
assay	Get the first experimental data numeric matrix	matrix, matrix-like
sampleMap	Get or set the map relating observations to subjects	DataFrame
metadata	Get or set additional data descriptions	list
rownames	Get row names for all experiments	CharacterList
colnames	Get column names for all experiments	CharacterList
Subsetting		
mae[i, j, k]	Get rows, columns, and/or experiments	MultiAssayExperiment
mae[i, ,]	GRanges, character, integer, logical, List, list	MultiAssayExperiment
mae[, j,]	character, integer, logical, List, list	MultiAssayExperiment
mae[, , k]	character, integer, logical	MultiAssayExperiment
mae[[i]]	Get or set object of arbitrary class from experiments character, integer, logical	(varies)
mae$column	Get or set colData column	vector (varies)
Management		
complete.cases	Identify subjects with complete data in all experiments	vector (logical)
duplicated	Identify subjects with replicate observations per experiment	list of LogicalLists
mergeReplicates	Merge replicate observations within each experiment, using functior	MultiAssayExperiment
intersectRows	Return features that are present for all experiments	MultiAssayExperiment
intersectColumns	Return subjects with data available for all experiments	MultiAssayExperiment
prepMultiAssay	Troubleshoot common problems when constructing main class	list
Reshaping		
longFormat	Return a long and tidy DataFrame with optional colData columns	DataFrame
wideFormat	Create a wide DataFrame, 1 row per subject	DataFrame
Combining		
c	Concatenate an experiment	MultiAssayExperiment

Credit:
Marcel Ramos

TCGA as MultiAssayExperiments

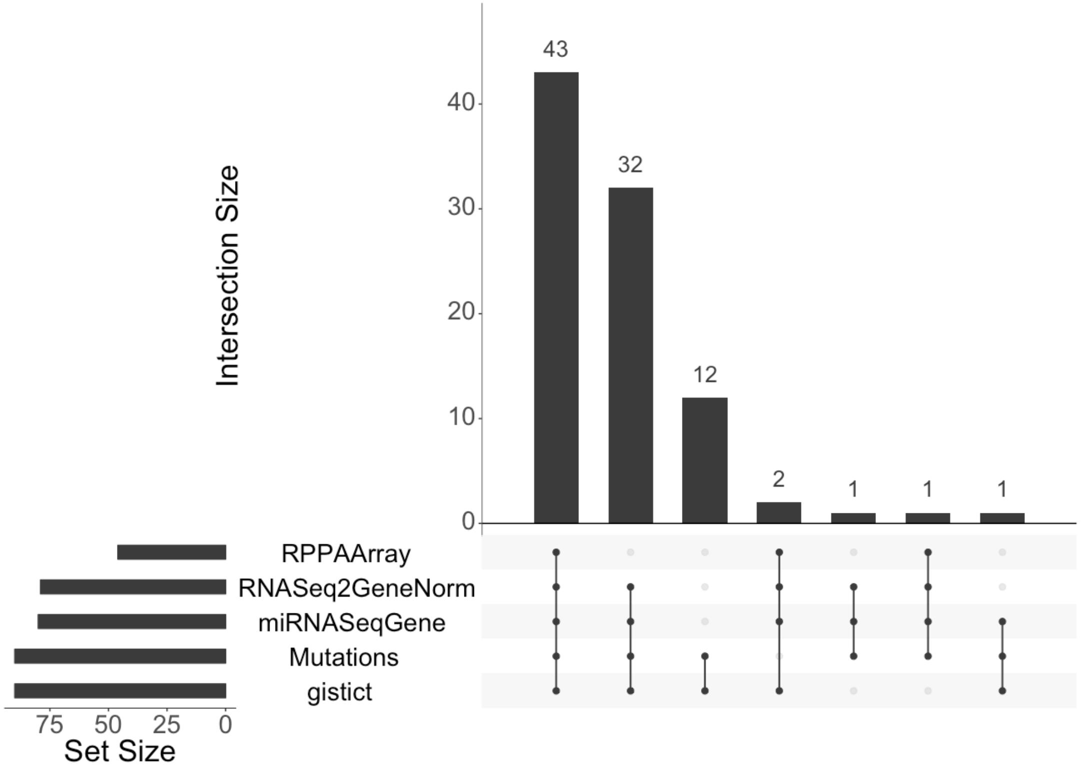


TCGA Cohort	Number of Assays	Number of Features	Number of Samples
Adrenocortical Carcinoma	9	677,947	180
Bladder Urothelial Carcinoma	11	1,236,717	806
Breast Invasive Carcinoma	8	595,134	1,212
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	9	892,007	586
Cholangiocarcinoma	9	610,891	85
Colon Adenocarcinoma	12	1,245,516	914
Lymphoid Neoplasm Diffuse Large B-Cell Lymphoma	9	627,264	94

…… 33 cancer types

Access from www.github.com/waldronlab/MultiAssayExperiment

For building visualizations

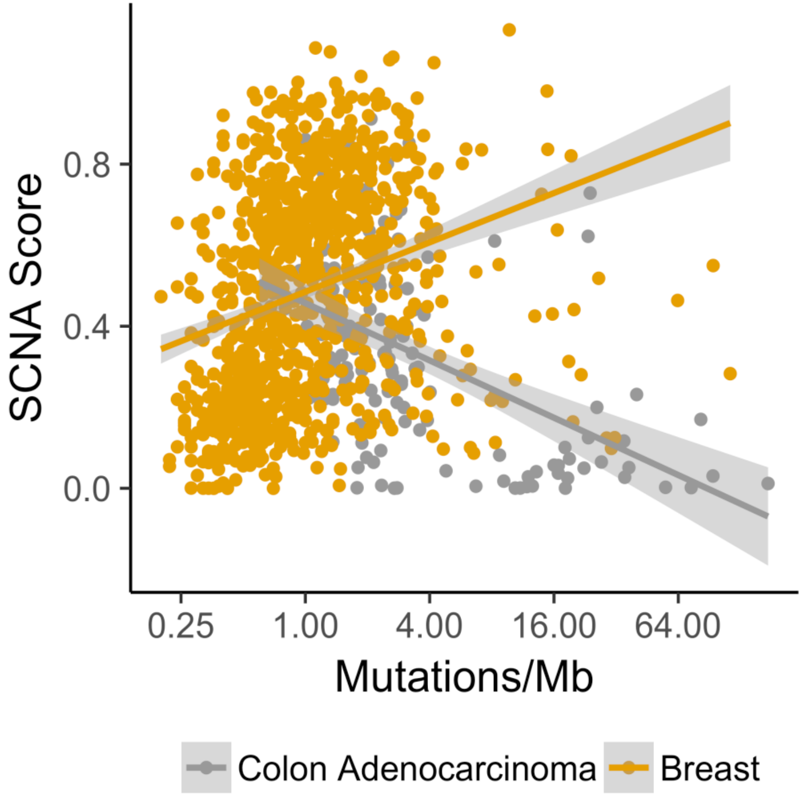


Upset Venn diagram for adrenocortical carcinoma TCGA

```
> data(miniACC)
> upsetSamples(miniACC)
```

Ramos *et al.*, Software for the integration of multi-omics experiments in Bioconductor (submitted).

For multi-omics analysis

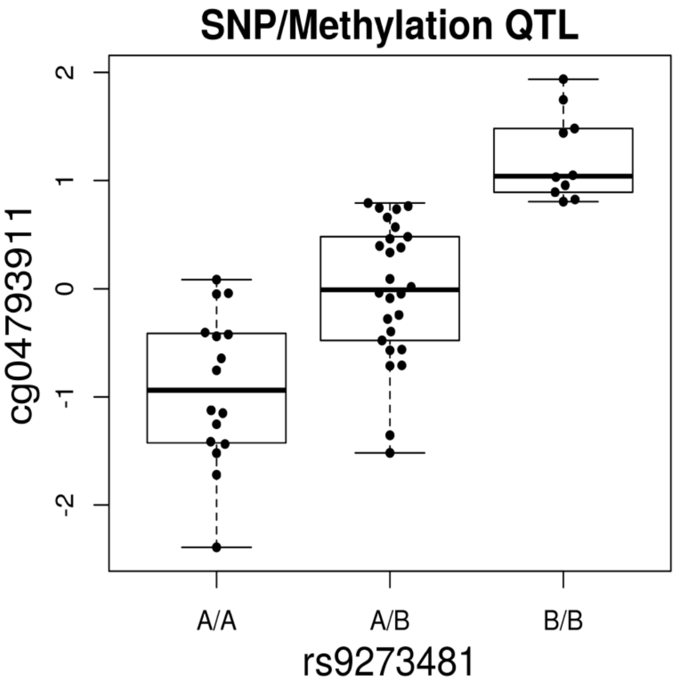


Davoli *et al*. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science 355, (2017).

```
> mae <- mae[, , c("Mutations", "gistict")]
> mae <- intersectColumns(mae)
> mae$cnload <-
colMeans(abs(assay(mae[["gistict"]])))
```

Ramos *et al.*, Software for the integration of multi-omics experiments in Bioconductor (submitted).

For integrating remotely stored data



Using tabix-indexed SNP VCFs from 1000 genomes on Amazon S3

credit: Vince Carey

```
> st <- ldblock::stack1kg() #Create a URL referencing 1000 genomes content in AWS S3
> multiban <- MultiAssayExperiment(
                    list(meth = banovichSE, snp = st),
                    colData = colData(banovichSE))
> multibanfocus <- multiban[rowRanges(banovichSE)[“cg04793911”], ]
> assoc <- cisAssoc(multibanfocus[[“meth”]],
                                TabixFile(files(multibanfocus[[“snp”]])))
```

Ramos *et al.*, Software for the integration of multi-omics experiments in Bioconductor (submitted).

Demo video

https://youtu.be/XziAMLf_AYI

Future work

- Distribute TCGA, cBioPortal through *ExperimentHub*
  - integrating clinical data and supplemental data in MAE
- Recognize relationships between:
  - genomic ranges – gene IDs – microRNAs – proteins
  - regulatory elements

THANK YOU

- **Lab** (www.waldronlab.org)
  - Marcel Ramos, Lucas Schiffer, Andy Samedy, Abzal Bacchus, Carmen Rodriguez, Audrey Renson, Ludwig Geistlinger
- **U24 CA180996 Collaborators**
  - Martin Morgan, Vincent Carey, Kasper Hansen
- **CUNY high-performance computing center**