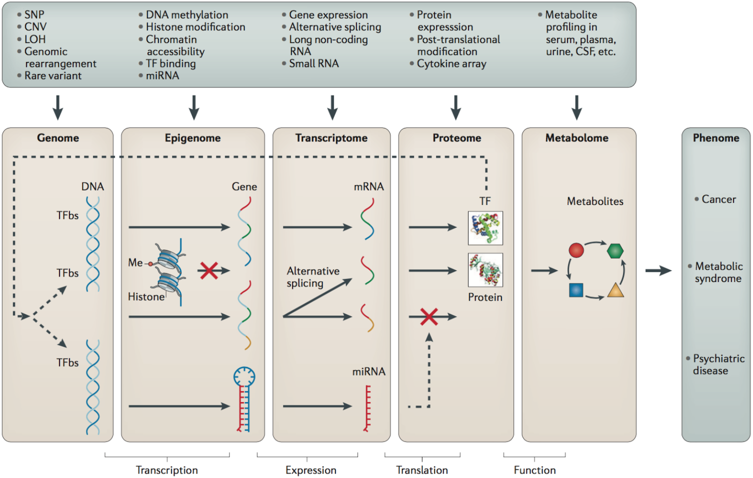
Integrative analysis intro

Mikhail Dozmorov

Spring 2018

Integrative analysis



- SNP
- CNV
- LOH
- Genomic rearrangement
- Rare variant

- DNA methylation
- Histone modification
- Chromatin accessibility
- TF binding
- miRNA

- Gene expression
- Alternative splicing
- Long non-coding RNA
- Small RNA

- Protein expresssion
- Post-translational modification
- Cytokine array

- Metabolite profiling in serum, plasma, urine, CSF, etc.

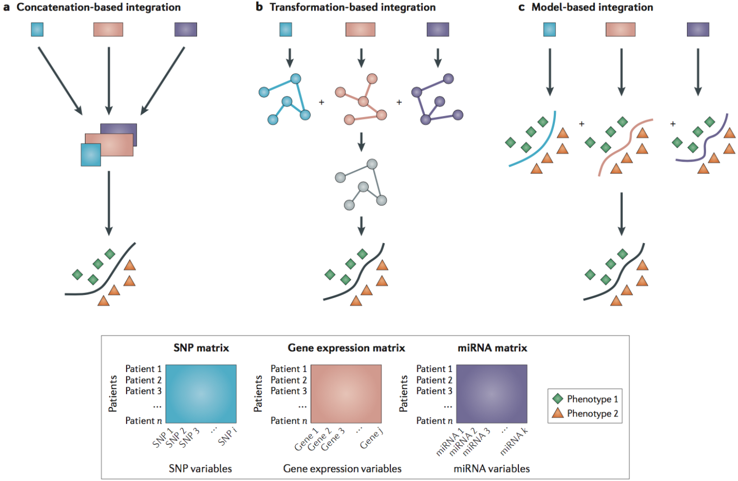
Ritchie, Marylyn D., Emily R. Holzinger, Ruowang Li, Sarah A. Pendergrass, and Dokyoon Kim. "Methods of Integrating Data to Uncover Genotype-Phenotype Interactions." Nature Reviews. Genetics 16, no. 2 (February 2015): 85–97. https://doi.org/10.1038/nrg3868.

Three main approaches

- **Concatenation-based integration** - combining data sets from different data types at the raw or processed data level before modelling and analysis
- **Transformation-based integration** - performing mapping or data transformation of the underlying data sets before analysis, and the modelling approach is applied at the level of transformed matrices
- **Model-based integration** - performing analysis on each data type independently, followed by integration of the resultant models to generate knowledge about the trait of interest

Ritchie, Marylyn D., Emily R. Holzinger, Ruowang Li, Sarah A. Pendergrass, and Dokyoon Kim. "Methods of Integrating Data to Uncover Genotype-Phenotype Interactions." Nature Reviews. Genetics 16, no. 2 (February 2015): 85–97. https://doi.org/10.1038/nrg3868.

Categorization of meta-dimensional analysi



Ritchie, Marylyn D., Emily R. Holzinger, Ruowang Li, Sarah A. Pendergrass, and Dokyoon Kim. "Methods of Integrating Data to Uncover Genotype-Phenotype Interactions." Nature Reviews. Genetics 16, no. 2 (February 2015): 85–97. https://doi.org/10.1038/nrg3868.

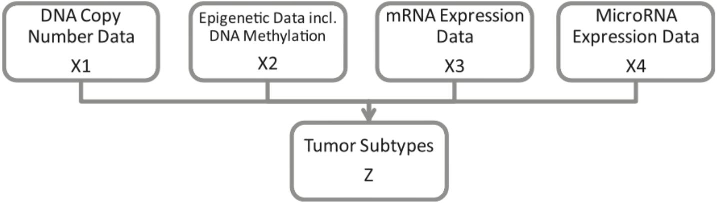
Clustering for data integration

- `iCluster` - uses a Gaussian latent variable model to infer clusters
  - It assumes that there is a common set of latent cluster membership variables across all datasets
  - Differences in structure between different datasets are accounted for only via individual noise terms, which correspond to within-dataset variances
  - It uses the k-means algorithm to extract the actual cluster assignments given latent variable values

Shen, Ronglai, Adam B. Olshen, and Marc Ladanyi. "Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis." Bioinformatics (Oxford, England) 25, no. 22 (November 15, 2009): 2906–12. https://doi.org/10.1093/bioinformatics/btp543.

https://www.mskcc.org/departments/epidemiology-biostatistics/biostatistics/icluster

iCluster



- The concept is to formulate the tumor subtypes as the joint latent variable Z that needs to be simultaneously estimated from multiple genomic data types measured on the same set of tumors

Shen, Ronglai, Adam B. Olshen, and Marc Ladanyi. "Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis." Bioinformatics (Oxford, England) 25, no. 22 (November 15, 2009): 2906–12. https://doi.org/10.1093/bioinformatics/btp543.

https://www.mskcc.org/departments/epidemiology-biostatistics/biostatistics/icluster

Regression for data integration

- remMap — REgularized Multivariate regression for identifying MAster Predictors for fitting multivariate response regression models under the high-dimension–low-sample-size setting.
- Dependence between two datasets, e.g., RNA levels and DNA copy numbers, is modeled through a multivariate response linear regression model
    - RNA levels are responses
    - DNA copy numbers are predictors

Regularized multivariate regression

Consider multivariate regression with Q response variables $y_1, ..., y_Q$ and P prediction variables $x_1, ..., x_P$

$$y_q = \sum_{p=1}^{P} X_p \beta_{pq} + \epsilon_q, \ \ q = 1, ..., Q$$

- The goal is to identify nonzero entries in the $P \times Q$ coefficient matrix $B = (\beta_{pq})$

`remMap` - Regularized Multivariate Regression for Identifying Master Predictors, https://cran.r-project.org/web/packages/remMap/index.html

Peng, Jie, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R. Pollack, and Pei Wang. "Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer." The Annals of Applied Statistics 4, no. 1 (March 2010): 53–77. https://doi.org/10.1214/09-AOAS271SUPP.

Proposed penalization

- L1 norm penalty to control the overall sparsity of the coefficient matrix B
- L2 norm penalty on regression coefficients for each predictor, i.e., the row vectors $C_p \cdot B_p$ - "group" sparse penalty inducing row sparsity of the product matrix $C \cdot B$ (some rows may be entirely zero)

$$L(B, \lambda_1, \lambda_2) = \frac{1}{2}\| Y - \sum_{p=1}^{P} X_p B_p \|_F^2 + \lambda_1 \sum_{p=1}^{P} \| C_p \cdot B_p \|_1 + \lambda_2 \sum_{p=1}^{P} \| C_p \cdot B_p \|_2$$

- C_p is the pth row of $C = (c_{pq}) = (C_1^T : \ldots : C_P^T)^T$, which is a pre-specified $P \times Q$ 0-1 matrix indicating the coefficients on which penalization is imposed. Based on prior knowledge. Can be set to $c_{p,q} = 0$
- B_p is the pth row of B
- $\| \cdot \|_F$ - Frobenius norm of matrices
- $\| \cdot \|_1$ and $\| \cdot \|_2$ are the l_1 and l_2 norms for vectors
- "$\cdot$" - Hadamard product (entry-wise multiplication)

Estimating coefficients

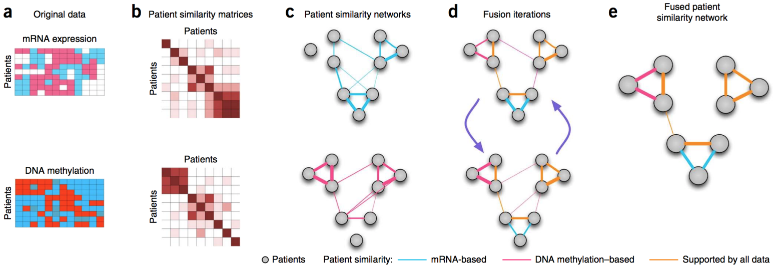
$$L(B,\lambda_1,\lambda_2) = \frac{1}{2}\|Y - \sum_{p=1}^{P} X_p B_p\|_F^2 + \lambda_1 \sum_{p=1}^{P} \|C_p \cdot B_p\|_1 + \lambda_2 \sum_{p=1}^{P} \|C_p \cdot B_p\|_2$$

- The C_p indicator matrix may be set to $c_{pq} = 1$
- An iterative algorithm for solving a convex optimization problem when the two tuning parameters are non-zero. Estimate of the coefficient matrix B is

$$\hat{B}(\lambda_1,\lambda_2) := argmin_B L(B;\lambda_1,\lambda_2)$$

Similarity Network Fusion for data integration

- SNF - Fusing correlation matrices for each data type into one network. Constructing sample similarity for each data type, then merging them into a single similarity network using a nonlinear combination method based on message passing theory



Wang, Bo, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. “Similarity Network Fusion for Aggregating Data Types on a Genomic Scale.” Nature Methods 11, no. 3 (March 2014): 333–37. https://doi.org/10.1038/nmeth.2810.

http://compbio.cs.toronto.edu/SNF/SNF/Software.html

MultiAssayExperiment

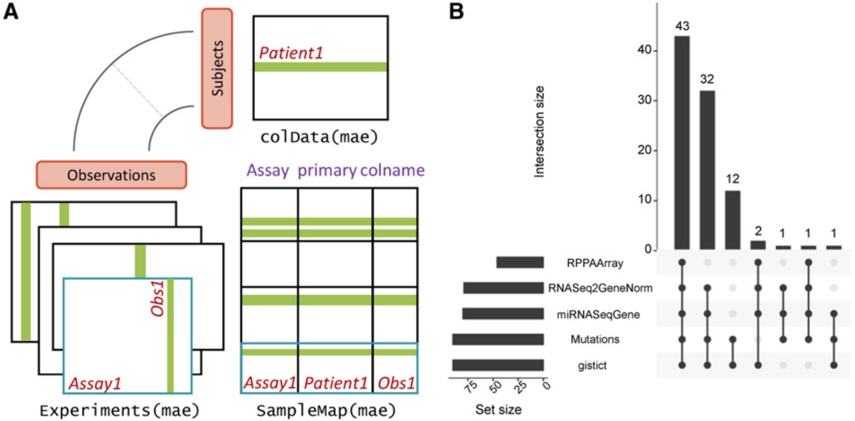
The need for MultiAssayExperiment

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

https://github.com/waldronlab/MultiAssayExperiment

MultiAssayExperiment R package



https://bioconductor.org/packages/release/bioc/html/MultiAssayExperiment.html

Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Cabrera CR, Chan T, Chapman P, Davis S, Gomez-Cabrero D, Culhane AC, Haibe-Kains B, Hansen K, Kodali H, Louis MS, Mer AS, Reister M, Morgan M, Carey V and Waldron L (2017). "Software For The Integration Of Multi-Omics Experiments In Bioconductor." Cancer Research. http://cancerres.aacrjournals.org/content/77/21/e39

MultiAssayExperiment R package

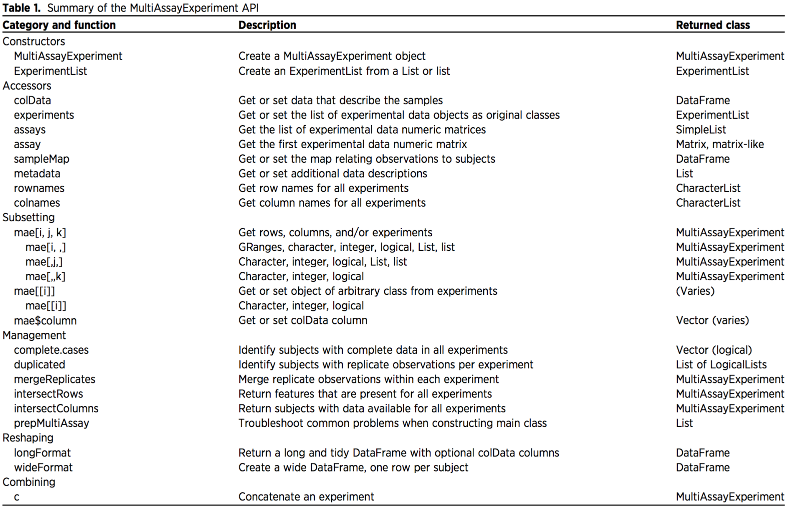


Table 1. Summary of the MultiAssayExperiment API

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	(Varies)
mae[[i]]	Character, integer, logical	
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	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, one row per subject	DataFrame
Combining		
c	Concatenate an experiment	MultiAssayExperiment

http://cancerres.aacrjournals.org/content/77/21/e39

MultiAssayExperiment object

```
> acc
A MultiAssayExperiment object of 9 listed
experiments with user-defined names and respective classes.
Containing an ExperimentList class object of length 9:
[1] RNASeq2GeneNorm: ExpressionSet with 20501 rows and 79 columns
[2] miRNASeqGene: ExpressionSet with 1046 rows and 80 columns
[3] CNASNP: RaggedExperiment with 79861 rows and 180 columns
[4] CNVSNP: RaggedExperiment with 21052 rows and 180 columns
[5] Methylation: SummarizedExperiment with 485577 rows and 80 columns
[6] RPPAArray: ExpressionSet with 192 rows and 46 columns
[7] Mutations: RaggedExperiment with 20166 rows and 90 columns
[8] gistica: SummarizedExperiment with 24776 rows and 90 columns
[9] gistict: SummarizedExperiment with 24776 rows and 90 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
```

MultiAssayExperiment

- MultiAssayExperiment, Bioconductor package for management of multi-assay data
- TCGA data
- How to get the data,

https://github.com/waldronlab/MultiAssayExperiment

https://docs.google.com/spreadsheets/d/1Ih64DDS5mqDlYFzDyCY9HAUnxvI1b6hapKP_akFuNPY/edit#gid=0

https://github.com/waldronlab/curatedTCGAData

Ramos, Marcel, Lucas Schiffer, Angela Re, Rimsha Azhar, Azfar Basunia, Carmen Rodriguez Cabrera, Tiffany Chan, et al. "Software For The Integration Of Multi-Omics Experiments In Bioconductor," June 1, 2017. https://doi.org/10.1101/144774.