October 20th, 2021

8th Week

Bioconductor

Advanced Bioinformatics 1

Kyulhee Han

Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea

Contents

Introduction

Papers

Published: 29 January 2015

Orchestrating high-throughput genomic analysis with Bioconductor

Wolfgang Huber, Vincent J Carey, [...] Martin Morgan

Nature Methods **12**, 115–121 (2015) | Cite this article

24k Accesses | **1498** Citations | **170** Altmetric | Metrics

Method | Open Access | Published: 15 September 2004

Bioconductor: open software development for computational biology and bioinformatics

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean YH Yang & Jianhua Zhang

Genome Biology **5**, Article number: R80 (2004) | Cite this article

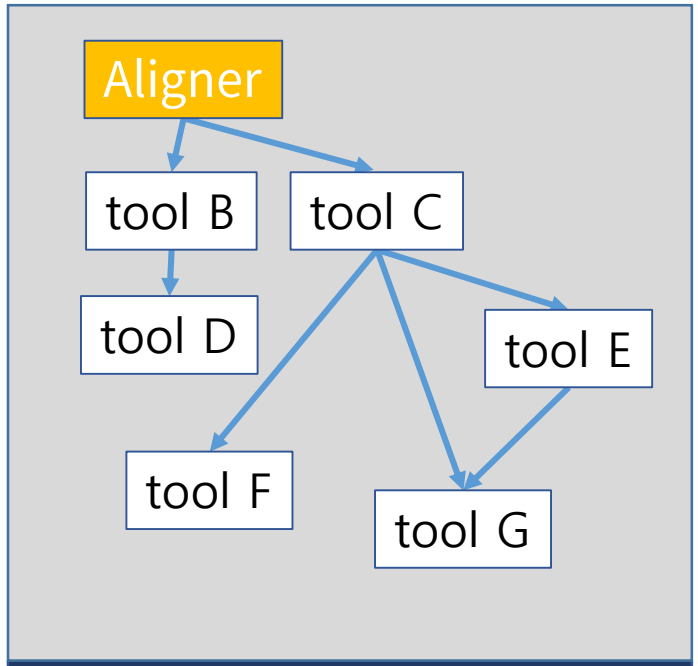
187k Accesses | **8757** Citations | **42** Altmetric | Metrics

Background

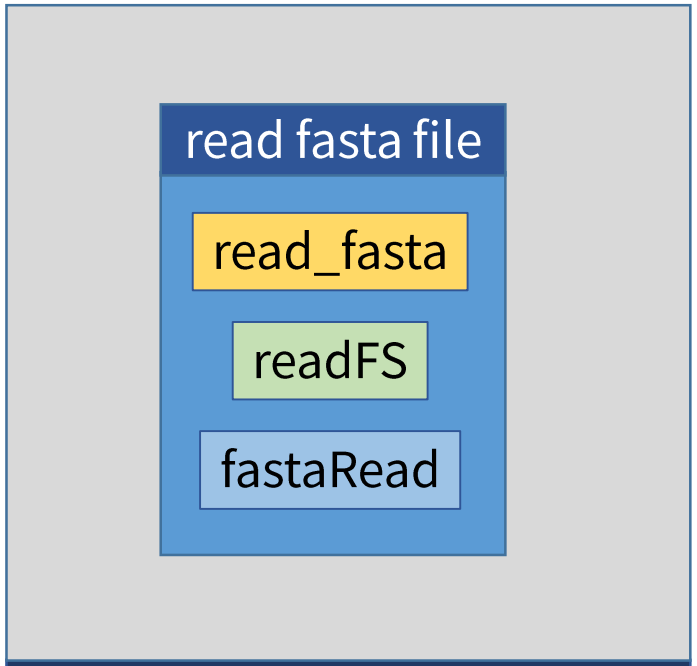
- A lot of analytical methods for bioinformatics
  - as biological data increases, bioinformatics analysis methodologies dealing with it have also been actively developed
  - using analysis tools separately have several problems

https://bioinformatics.stackexchange.com/questions/639/why-bioconductor

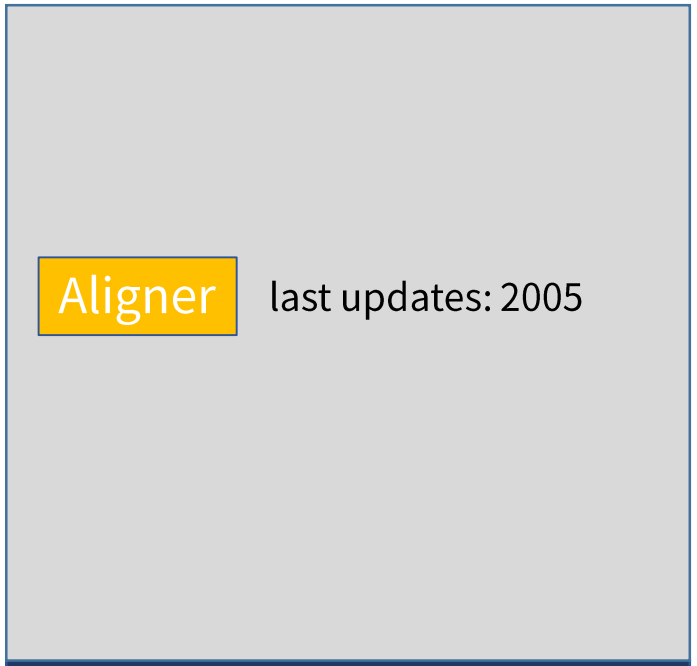
Problems



dependency & compatibility



duplicated tools



reliability

Solution: Bioconductor

- Open source & Open development project
  - install a package for genome analysis with proper dependencies
  - by using well-developed package, avoid developing same function
  - Bioconductor forces developers hard test, and remove not maintained packages

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Aims & Features

- Aims
  - Providing compelling user experiments for end user
    - full workflow, document of each tool, detailed example code
  - Activating new algorithm & software development community for bioinformatics
    - distributed development of each software components
- Features
  - Documentation and reproducible research
  - Statistical and graphical methods
  - Genome annotation
  - Open source

Feature: S4 class

- Specific class for structured analysis
  - Bioconductor packages defining several classes for specific data type
  - Have several benefits, such as less confusing of users and store extra information for experiment

Feature: S4 class

```
ALL                                   Large ExpressionSet (987.5 kB)
..@ experimentData :Formal class 'MIAME' [package "Biobase"] with 13 slots
.. .. ..@ name : chr "Chiaretti et al."
.. .. ..@ lab : chr "Department of Medical Oncology, Dana-Farber Cancer Institute, Department…
.. .. ..@ contact : chr ""
.. .. ..@ title : chr "Gene expression profile of adult T-cell acute lymphocytic leukemia ide…
.. .. ..@ abstract : chr "Gene expression profiles were examined in 33 adult patients with T-…
.. .. ..@ url : chr ""
.. .. ..@ pubMedIds : chr [1:2] "14684422" "16243790"
.. .. ..@ samples : list()
.. .. ..@ hybridizations : list()
.. .. ..@ normControls : list()
.. .. ..@ preprocessing : list()
.. .. ..@ other : list()
.. .. ..@ .__classVersion__:Formal class 'Versions' [package "Biobase"] with 1 slot
.. .. .. .. ..@ .Data:List of 1
.. .. .. .. .. ..$ : int [1:3] 1 0 0
..@ assayData :<environment: 0x0000029282bdb3c8>
..@ phenoData :Formal class 'AnnotatedDataFrame' [package "Biobase"] with 4 slots
.. .. ..@ varMetadata :'data.frame': 21 obs. of 1 variable:
.. .. .. ..$ labelDescription: chr [1:21] " Patient ID" " Date of diagnosis" " Gender of the …
.. .. ..@ data :'data.frame': 128 obs. of 21 variables:
.. .. .. ..$ cod : chr [1:128] "1005" "1010" "3002" "4006" ...
.. .. .. ..$ diagnosis : chr [1:128] "5/21/1997" "3/29/2000" "6/24/1998" "7/17/1997" ...
.. .. .. ..$ sex : Factor w/ 2 levels "F","M": 2 2 1 2 2 2 1 2 2 2 ...
.. .. .. ..$ age : int [1:128] 53 19 52 38 57 17 18 16 15 40 ...
.. .. .. ..$ BT : Factor w/ 10 levels "B","B1","B2",..: 3 3 5 2 3 2 2 2 3 3 ...
.. .. .. ..$ remission : Factor w/ 2 levels "CR","REF": 1 1 1 1 1 1 1 1 1 1 ...
.. .. .. ..$ CR : chr [1:128] "CR" "CR" "CR" "CR" ...
.. .. .. ..$ date.cr : chr [1:128] "8/6/1997" "6/27/2000" "8/17/1998" "9/8/1997" ...
.. .. .. ..$ t(4;11) : logi [1:128] FALSE FALSE NA TRUE FALSE FALSE ...
.. .. .. ..$ t(9;22) : logi [1:128] TRUE FALSE NA FALSE FALSE FALSE ...
```

```
> ALL@experimentData
Experiment data
  Experimenter name: Chiaretti et al.
  Laboratory: Department of Medical Oncology, Dana-Farber Cancer Institute,
en's Hospital, Harvard Medical School, Boston, MA 02115, USA.
  Contact information:
  Title: Gene expression profile of adult T-cell acute lymphocytic leukemia
 with different response to therapy and survival.
  URL:
  PMIDs: 14684422 16243790
```

```
> ALL@assayData$exprs
                              01005     01010     03002     04006
1000_at                    7.597323  7.479445  7.567593  7.384684
1001_at                    5.046194  4.932537  4.799294  4.922627
1002_f_at                  3.900466  4.208155  3.886169  4.206798
```

```
> ALL@phenoData@data
        cod  diagnosis  sex age BT
01005  1005  5/21/1997    M  53 B2
01010  1010  3/29/2000    M  19 B2
```

```
> ALL@annotation
[1] "hgu95av2"
```

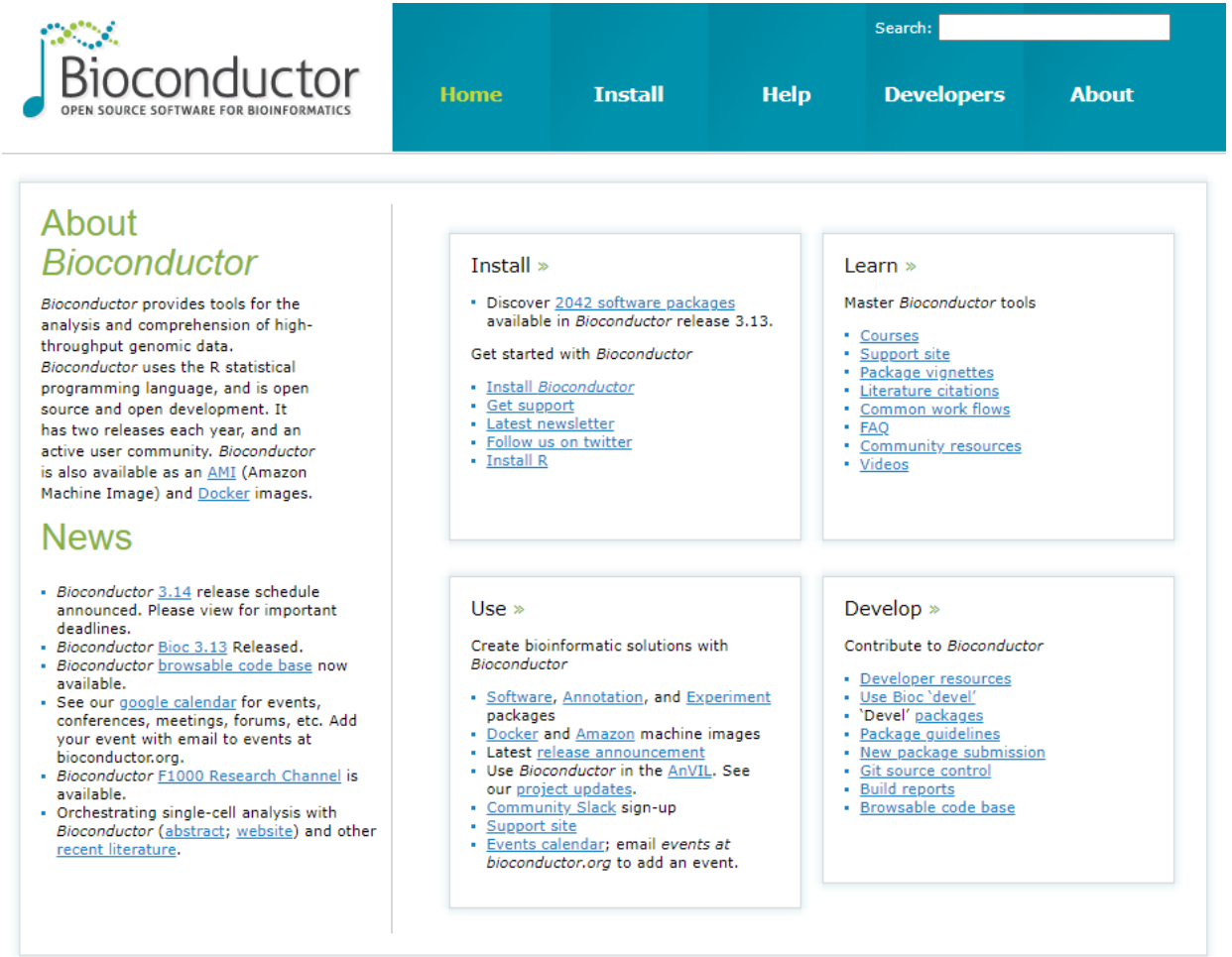
Why use R?

- High level language for data processing
  - easy to learn and quick start
  - deal with matrix type data well, and supports various statistical methods
- High quality visualization
  - comparing to other statistical methods, R provides high quality and customizable figures with free of charge
- Using R ecosystem
  - using well developed software derived from CRAN or Bioconductor, development could be more efficient

The user perspective

Quick Start

- Requirements
  - R
  - Rstudio (Recommended)
- Bioconductor
  - https://www.bioconductor.org/



Quick Start

- Bioconductor
  - Install BiocManager first

```
install.packages("BiocManager")
library(BiocManager)
```

- Bioconductor package
  - To install specific package in Bioconductor, use install function in BiocManger package

```
BiocManager::install("GenomicRanges")
library(GenomicRanges)
```

Quick Start

GenomicRanges

platforms all | rank 11 / 2041 | support 1 1 / 1 5 | in Bioc 11.5 years
build ok | updated before release | dependencies 16

DOI: 10.18129/B9.bioc.GenomicRanges

Representation and manipulation of genomic intervals

Bioconductor version: Release (3.13)

The ability to efficiently represent and manipulate genomic annotations and alignments is playing a central role when it comes to analyzing high-throughput sequencing data (a.k.a. NGS data). The GenomicRanges package defines general purpose containers for storing and manipulating genomic intervals and variables defined along a genome. More specialized containers for representing and manipulating short alignments against a reference genome, or a matrix-like summarization of an experiment, are defined in the GenomicAlignments and SummarizedExperiment packages, respectively. Both packages build on top of the GenomicRanges infrastructure.

Author: P. Aboyoun, H. Pagès, and M. Lawrence

Maintainer: Bioconductor Package Maintainer <maintainer at bioconductor.org>

Citation (from within R, enter `citation("GenomicRanges")`):

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M, Carey V (2013). "Software for Computing and Annotating Genomic Ranges." *PLoS Computational Biology*, **9**. doi: 10.1371/journal.pcbi.1003118, http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003118.

Installation

To install this package, start R (version "4.1") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("GenomicRanges")
```

For older versions of R, please refer to the appropriate Bioconductor release.

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("GenomicRanges")
```

HTML	R Script	1. An Introduction to the GenomicRanges Package
PDF	R Script	2. GenomicRanges HOWTOs
PDF	R Script	3. A quick introduction to GRanges and GRangesList objects (slides)
PDF	R Script	4. Ten Things You Didn't Know (slides from BioC 2016)
PDF	R Script	5. Extending GenomicRanges
PDF		Reference Manual
Text		NEWS

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("GenomicRanges")
```

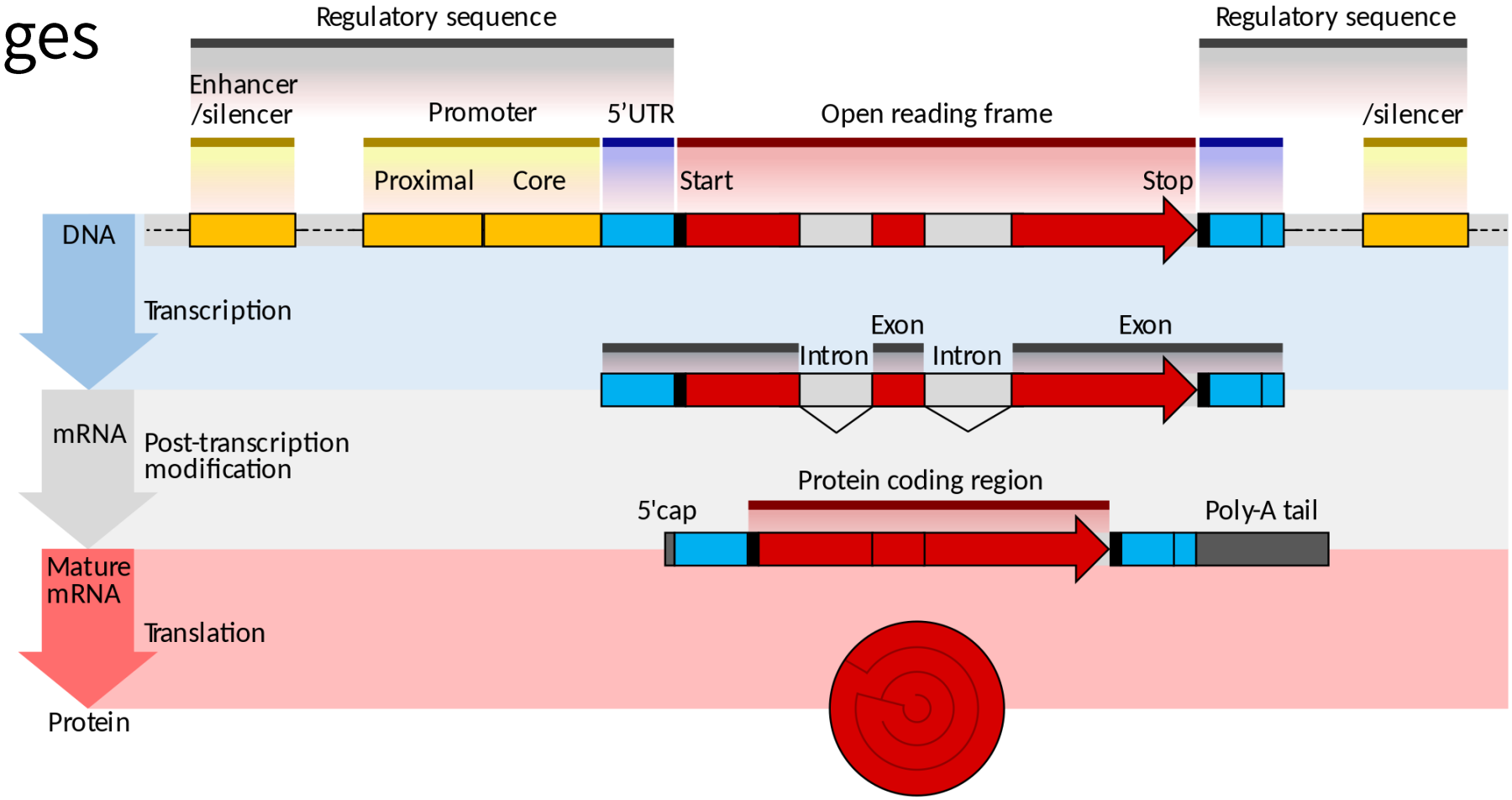
HTML	R Script	1. An Introduction to the GenomicRanges Package
PDF	R Script	2. GenomicRanges HOWTOs
PDF	R Script	3. A quick introduction to GRanges and GRangesList objects (slides)
PDF	R Script	4. Ten Things You Didn't Know (slides from BioC 2016)
PDF	R Script	5. Extending GenomicRanges
PDF		Reference Manual
Text		NEWS

User scenario

- Scientific need: Differential Expressed Gene analysis using RNA-seq data
- For particular problem
  - Range of interest: *GRanges*
  - Multiple samples: *SummarizedExperiment & ExperimentSet*
  - Differential Expression: *limma*
  - Visualization: *ggbio*
  - Annotation: *hgu95av2.db*
  - Gene Ontology analysis: *GOstats*

1) Range of interest

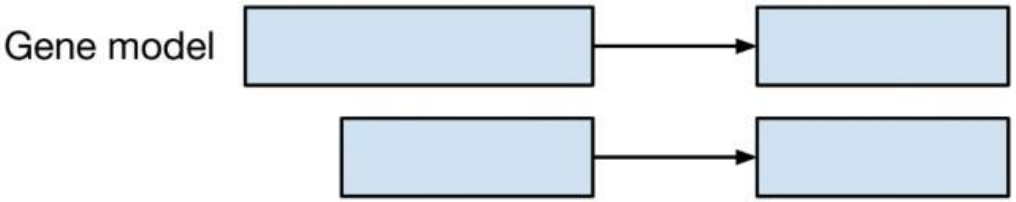
- Several ranges



https://en.wikipedia.org/wiki/Gene_structure

1) Range of interest

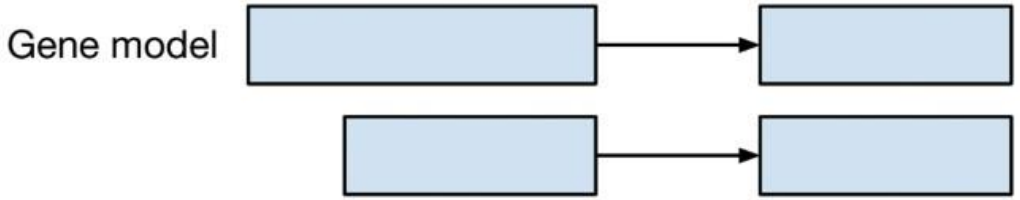
- ‘Ranges’
  - representation & analysis of genomic intervals
  - consist of several packages (*IRanges*, *GenomicRanges*⋯)



g (Grange object)

1) Range of interest

- *GenomicRanges* package



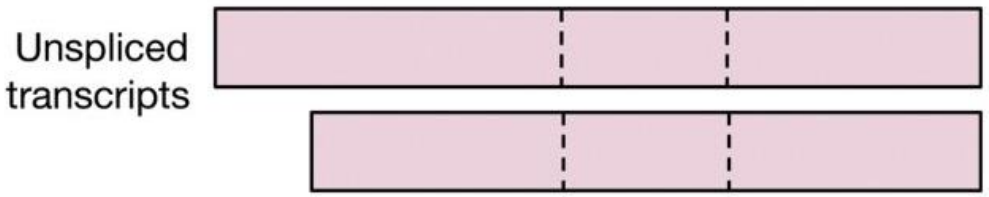

```r
BiocManager::install("GenomicRanges")
library(GenomicRanges)
```

gg

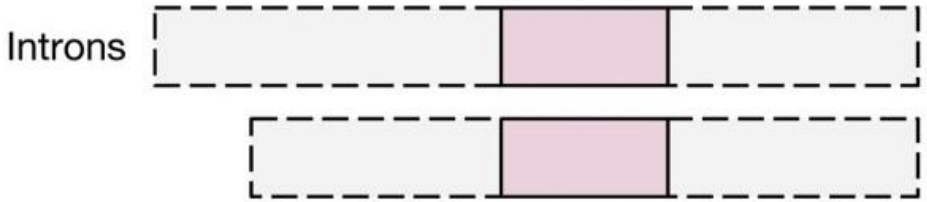
```
> gene_model
GRangesList object of length 2:
$txn1
GRanges object with 2 ranges and 0 metadata columns:
      seqnames    ranges strand
         <Rle> <IRanges>  <Rle>
  [1]        A   100-500      *
  [2]        A   700-900      *
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths

$txn2
GRanges object with 2 ranges and 0 metadata columns:
      seqnames    ranges strand
         <Rle> <IRanges>  <Rle>
  [1]        A   200-500      *
  [2]        A   700-900      *
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

1) Range of interest

- GenomicRanges package



range(g)

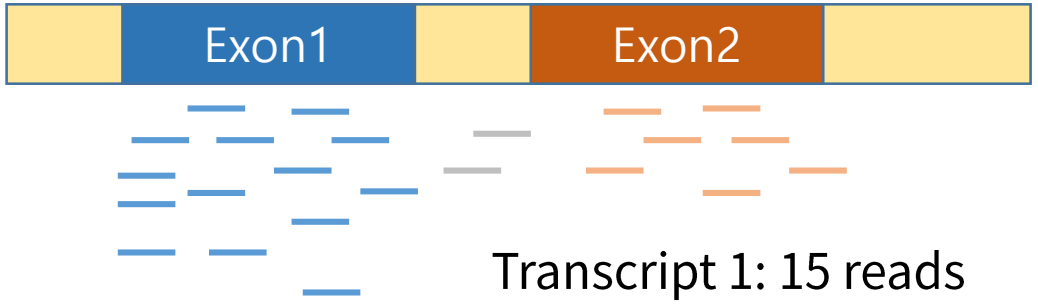


setdiff(range(g), g)

```
> range(gene_model)
GRangesList object of length 2:
$txn1
GRanges object with 1 range and 0 metadata columns:
      seqnames    ranges strand
         <Rle> <IRanges>  <Rle>
  [1]        A   100-900      *
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths

$txn2
GRanges object with 1 range and 0 metadata columns:
      seqnames    ranges strand
         <Rle> <IRanges>  <Rle>
  [1]        A   200-900      *
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths

> setdiff(range(gene_model), gene_model)
GRangesList object of length 2:
$txn1
GRanges object with 1 range and 0 metadata columns:
      seqnames    ranges strand
         <Rle> <IRanges>  <Rle>
  [1]        A   501-699      *
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths

$txn2
GRanges object with 1 range and 0 metadata columns:
      seqnames    ranges strand
         <Rle> <IRanges>  <Rle>
  [1]        A   501-699      *
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

1) Range of interest

- Summarize
  - After conducting several range of interest, users can summarize the data in their own purpose



Transcript 1: 15 reads
Transcript 2: 7 reads
…

RNA-seq
counting #cDNA fragments for each transcript

call: … A/A … A/T

DNA-seq
calling DNA sequence variants

2) Multiple Samples

```
BiocManager::install("SummarizedExperiment")
library(SummarizedExperiment)
```

- *SummarizedExperiment* class
  - Provided by *SummarizedExperiment* package
  - summarized matrix
    - e.g. reads for each transcript for multiple samples
  - sample & feature information
    - e.g. transcript information, age and gender of each sample
  - other metadata
    - e.g. experiment condition, researcher

2) Multiple Samples

assays

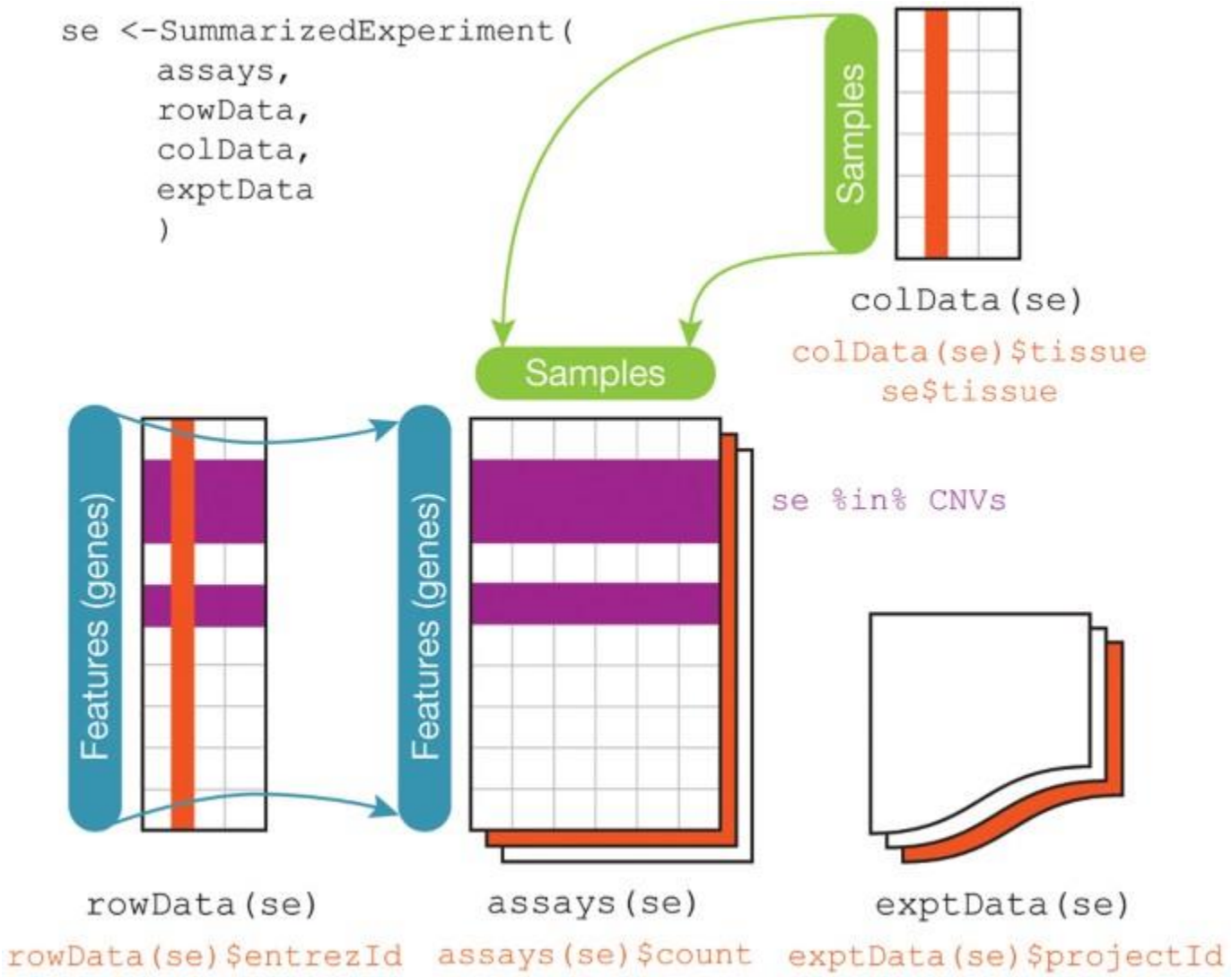
⇒ rows for features, cols for samples. multiple dataframe could be stored

rowData & colData

⇒ feature & sample information. one dataframe

exptData

⇒ experiment date, project II etc. no need to have same dimension with assay's col or row



2) Multiple Samples

```
library(SummarizedExperiment)

nrows ← 200
ncols ← 6
counts ← matrix(runif(nrows * ncols, 1, 1e4), nrows)
rowRanges ← GRanges(rep(c("chr1", "chr2"), c(50, 150)),
                    IRanges(floor(runif(200, 1e5, 1e6)), width=100),
                    strand=sample(c("+", "-"), 200, TRUE),
                    feature_id=sprintf("ID%03d", 1:200))
colData ← DataFrame(Treatment=rep(c("ChIP", "Input"), 3),
                    row.names=LETTERS[1:6])

exp ← SummarizedExperiment(assays=list(counts=counts),
                    rowRanges=rowRanges, colData=colData)
```

2) Multiple Samples

- *ExperimentSet* class
  - Provided by *biobase* package, which is base component of Bioconductor (no need to be installed separately)
  - *ExpressionSet* is generally used for array-based experiments, where the rows are features, and the SummarizedExperiment is generally used for sequencing-based experiments

http://www.sthda.com/english/wiki/expressionset-and-summarizedexperiment

3) Differential Expression

- Differential Expression
  - Various methods are available to find the DEGs
  - This scenario consider the significance test based on simple linear model

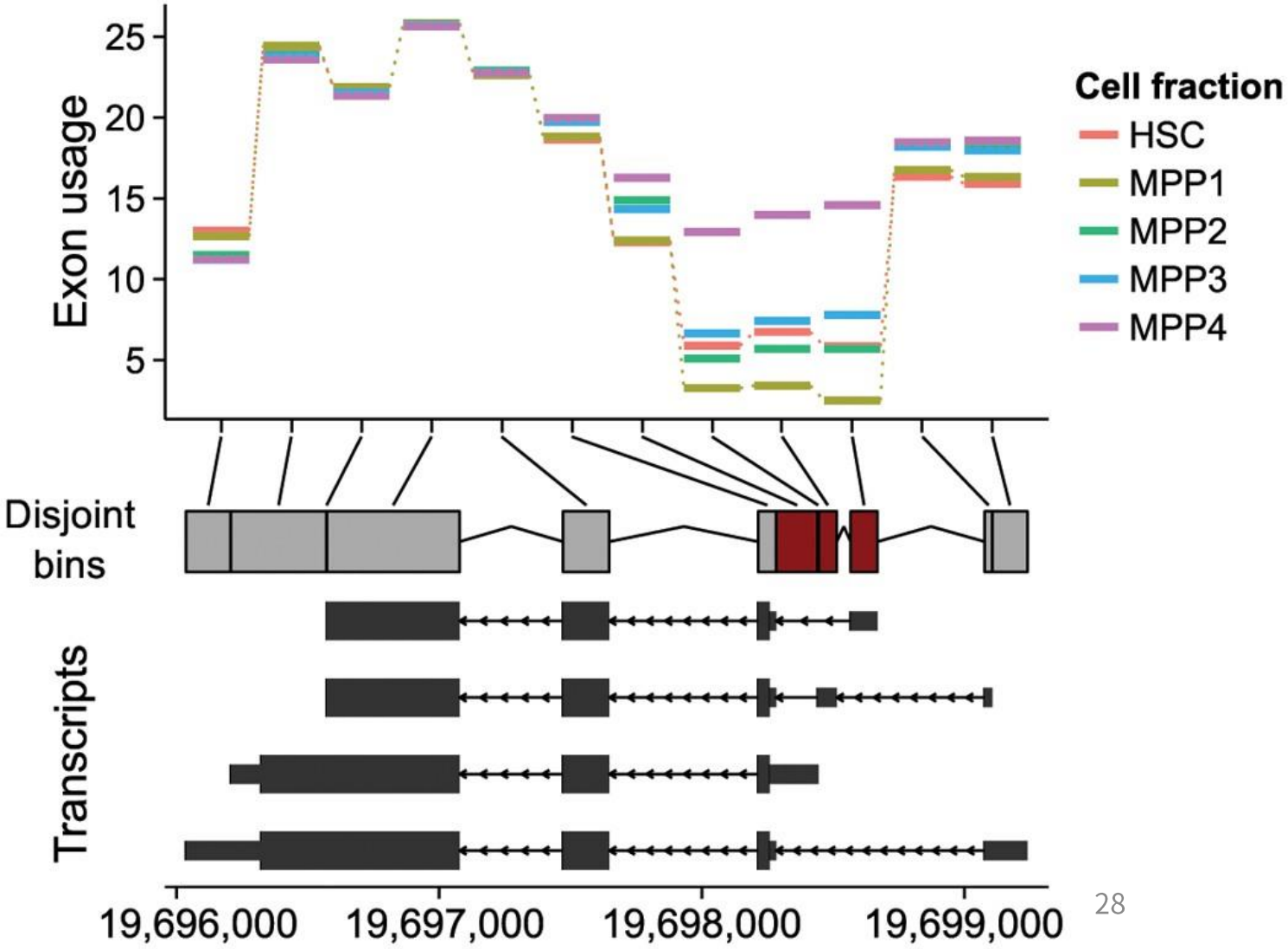
3) Differential Expression

```
BiocManager::install("limma")
library(limma)
```

- *limma* package
  - package for the analysis of gene expression microarray data
  - Provide linear model to distinguish DEG

4) Visualization

- Genomic-specific
  - plots along genomic coordination
  - *Gviz* and *ggbio*
  - Several package helps interaction of R and Genome Browser



5) Annotation

- Why annotation?

Microarray

Probes	P-value	…
Probe A	0.03	…
Probe B	0.4	…
Probe C	0.01	…

DB

Gene IDs	Gene Symbol
Probe A	*EGFR*
Probe B	*ERBB4*
Probe C	*PIK3CA*

Annotation

Gene Symbol	Pathway	…
1956	4-1BB Pathway	…
2066	Actin Nucleation	…
5290	PI3K pathway	…

5) Annotation

- Annotation data repository
  - 894 standardized annotation packages
  - present data using standardized interface

Name	Contents	Source
BSgenome	Whole genome sequence	UCSC / BioMart
TxDb	Transcript	
org	Identifier cross-reference	US National Center for Biotechnology Information / NCBI

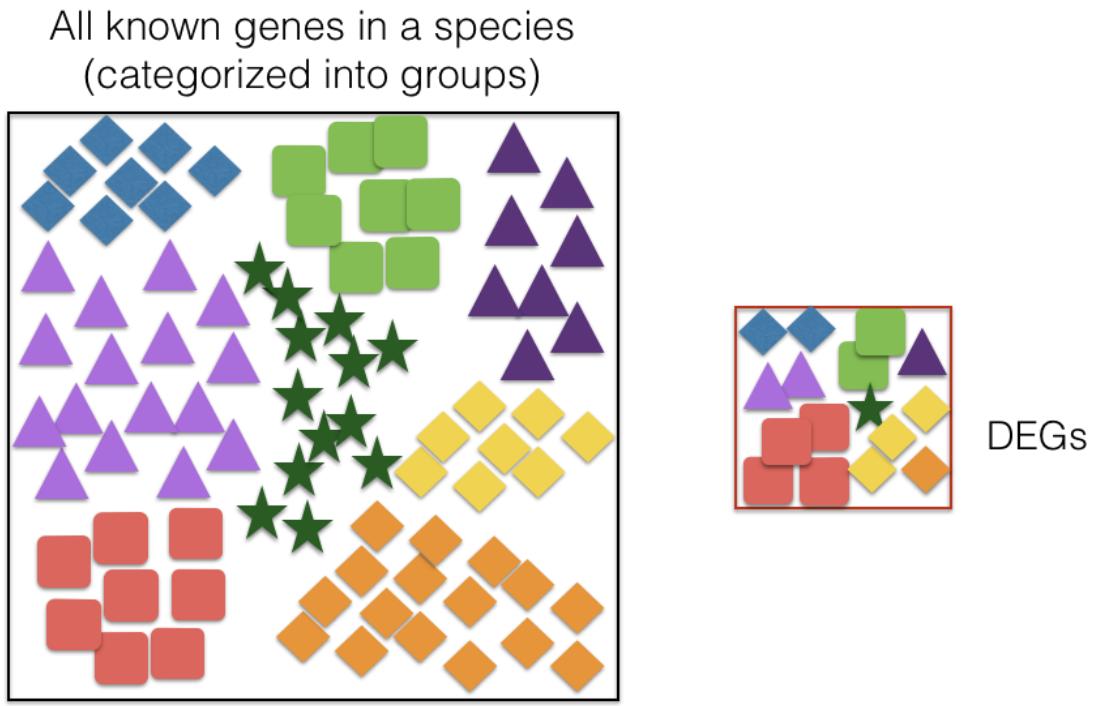
5) Annotation

```
BiocManager::install("hgu95av2.db")
library(hgu95av2.db)
```

- hgu95av2.db package
  - Reference for human genome Affymetrix Affymetrix HG_U95Av2 Array
  - Mapping probe of microarray and it’s target gene

6) Gene Ontology analysis

- Gene Ontology analysis
  - Test which category is ‘over-represented’ in DEGs
  - Interpretation of findings in gene set level



https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/10_FA_over-representation_analysis.html

6) Gene Ontology analysis

```
BiocManager::install("GOstats")
library(GOstats)
```

- *GOstats* package
  - Package for GO analysis
  - Provide which GO term is possibly related to DEGs

A data.frame: 5 × 7

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
	<chr>	<dbl>	<dbl>	<dbl>	<int>	<int>	<chr>
1	GO:1903037	1.548343e-06	4.770861	3.903268	16	247	regulation of leukocyte cell-cell adhesion
2	GO:1903039	1.605129e-06	5.423247	3.002514	14	190	positive regulation of leukocyte cell-cell adhesion
3	GO:0019221	2.503935e-06	3.207579	9.639650	26	610	cytokine-mediated signaling pathway
4	GO:0007159	6.546285e-06	4.223342	4.361547	16	276	leukocyte cell-cell adhesion
5	GO:0071345	9.527182e-06	2.747318	13.037232	30	825	cellular response to cytokine stimulus

The developer perspective

What Should Developer Consider

- The package ecosystem
  - developer should distribute ‘package’
  - continuous maintenance and upgrade are needed
- Interoperability
  - encapsulation of shared structure using S4 class is needed
- Shared infrastructure
  - recommend using basic references (e.g. reference genome) and well developed library

Package Development

- Package Structure

Package		
DESCRIPTION	SETUP	describes dependency, copyright, etc.
R/	WRITE CODE	codes included in package
tests/	TEST	store several test results
man/	DOCUMENT	documentation
vignettes/	TEACH	more user-friendly document to teach users
data/	ADD DATA	data for the package
NAMESPACE	ORGANIZE	avoid interference of packages

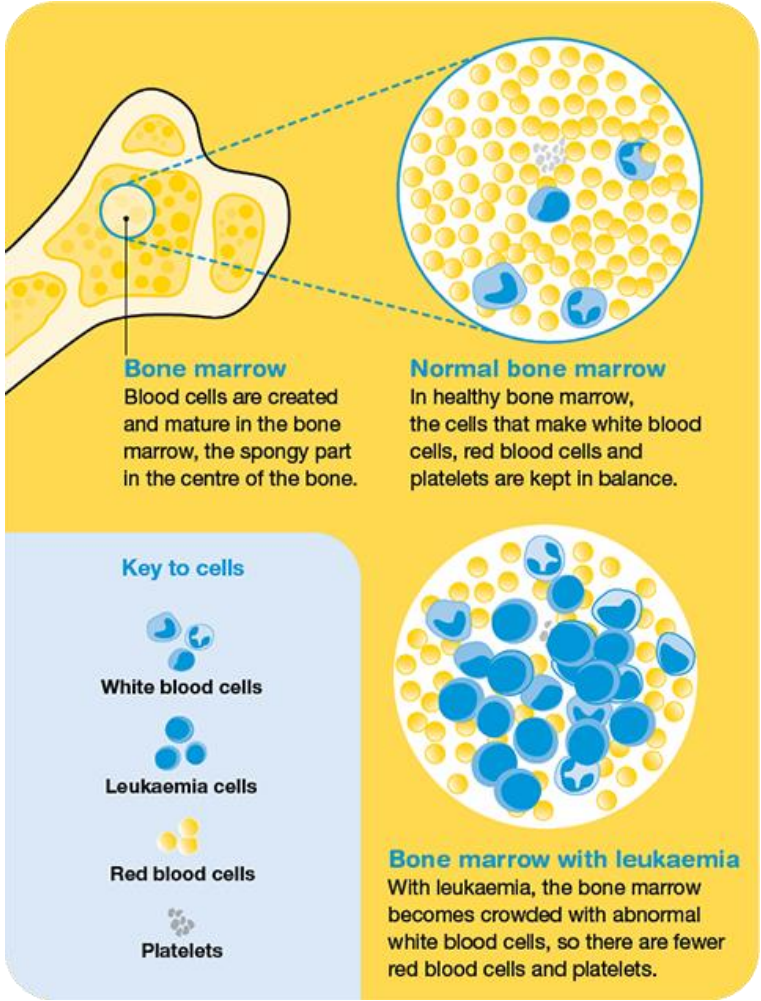
https://rawgit.com/rstudio/cheatsheets/master/package-development.pdf

Bioconductor Package Developer

- Submission
  - Opening a new issue in the Bioconductor Contributions repository following the guidelines of the README.md file
  - Add the link to developer's repository to the issue that is opened (default branch only)
- Guidelines
  - For each contents of packages, specific restrictions exist
  - ex) for DESCRIPTION file, Authors@R field should be filled with name, active mail address
  - Full description: http://contributions.bioconductor.org/

Use Case (ALL)

ALL (Acute lymphocytic leukemia)

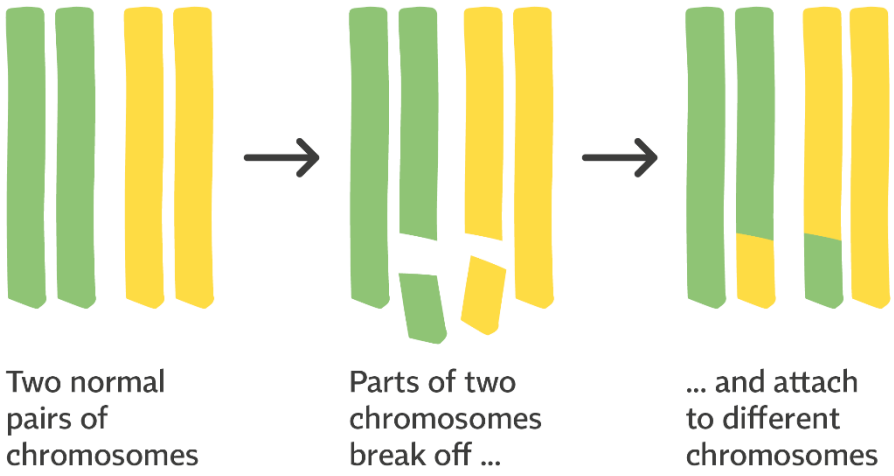


- Type of Leukemia characterized by undeveloped cells, occurs suddenly and grows quickly
- Develop from lymphocytes, including B-cells or T-cells

https://www.cancercouncil.com.au/acute-lymphoblastic-leukaemia/

ALL data

- Gene expression data of ALL patients
  - microarray from Ritz laboratory at the Dana Farber Cancer Institute (2004)
  - 128 patients, 12625 features
- Includes several subtypes
  - characterized by translocation of the specific regions



https://www.undiagnosed.org.uk/support-information/what-are-chromosome-translocations/

ALL data analysis

- Finding Differential Expressed Gene between subtypes
  - comparing two subtypes of ALL (BCR/ABL vs. ALL1/AF4)
- Heatmap analysis
  - clustering with genetic signature
- GO analysis
  - Annotate to EntrezGene IDs
  - Overrepresentation test for GO terms

https://www.undiagnosed.org.uk/support-information/what-are-chromosome-translocations/

Exercise with CoLab

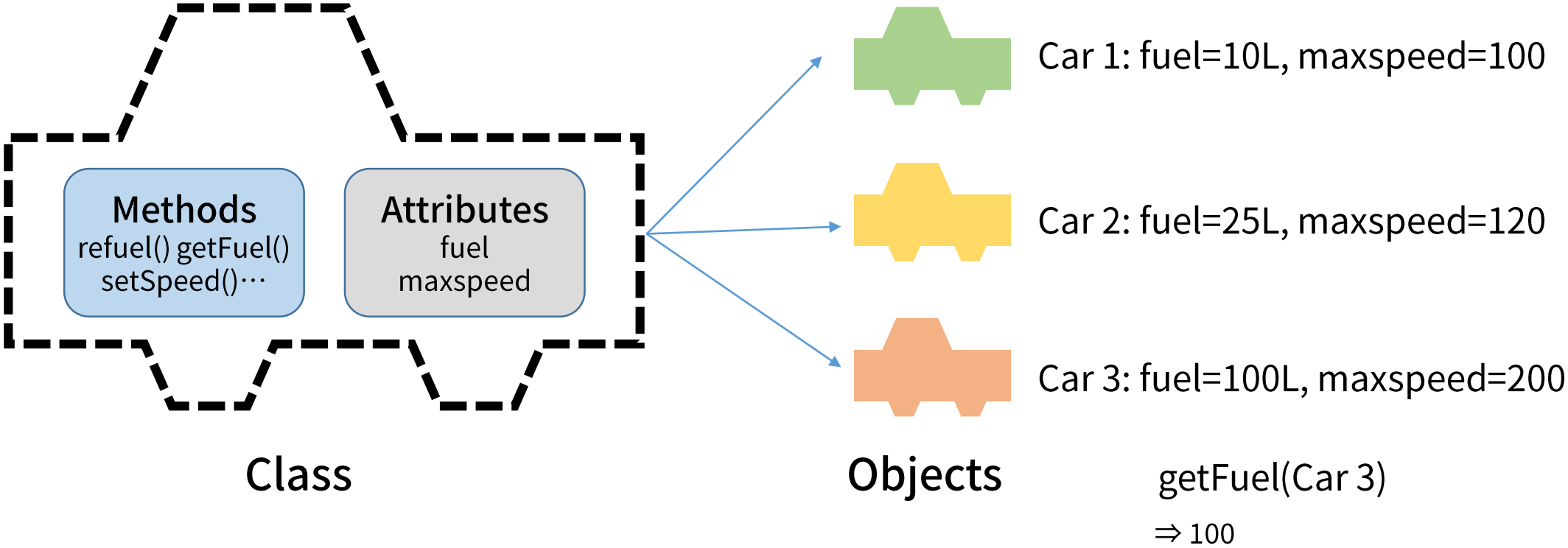
- https://colab.research.google.com/drive/1OrJMOipqL_XkaxtRCRLtmD8weiBIUUT_#scrollTo=X7R-fMAaiYL8

Q&A

Supplementary

Class & Object

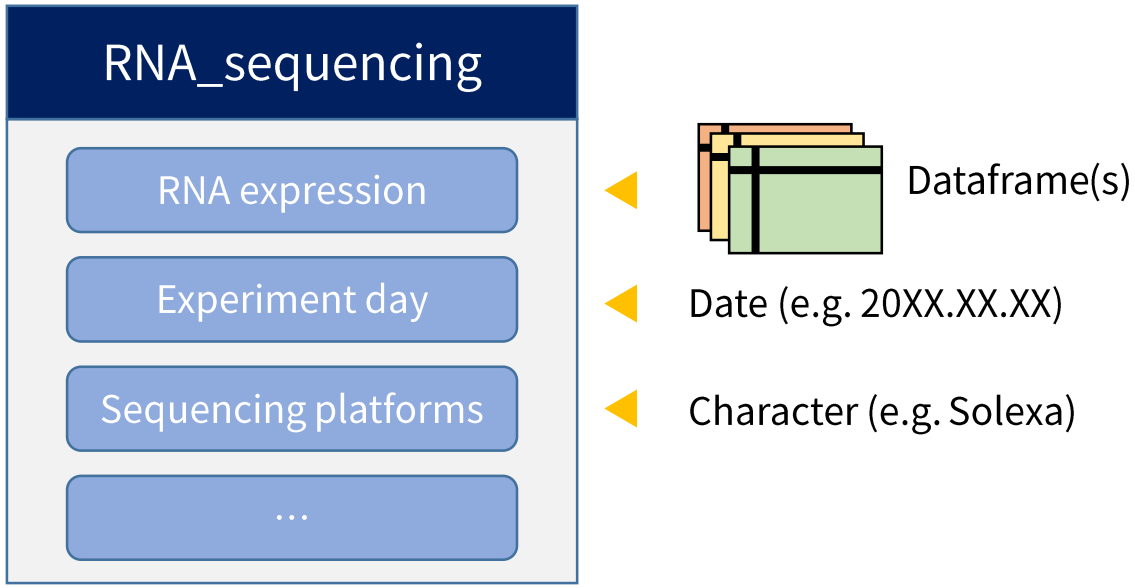
- Class: similar to data type, but it also have ‘methods’



- original image: https://commons.wikimedia.org/wiki/File:CPT-OOP-objects_and_classes_-_attmeth.svg

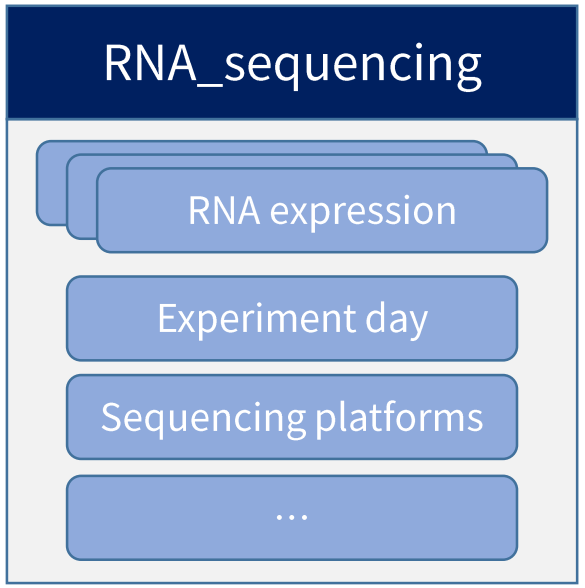
S4 Object System

- One of the Object System in R
  - S4 class includes slots (attributes) for storing specific data
  - S4 class could be defined by user for several purpose

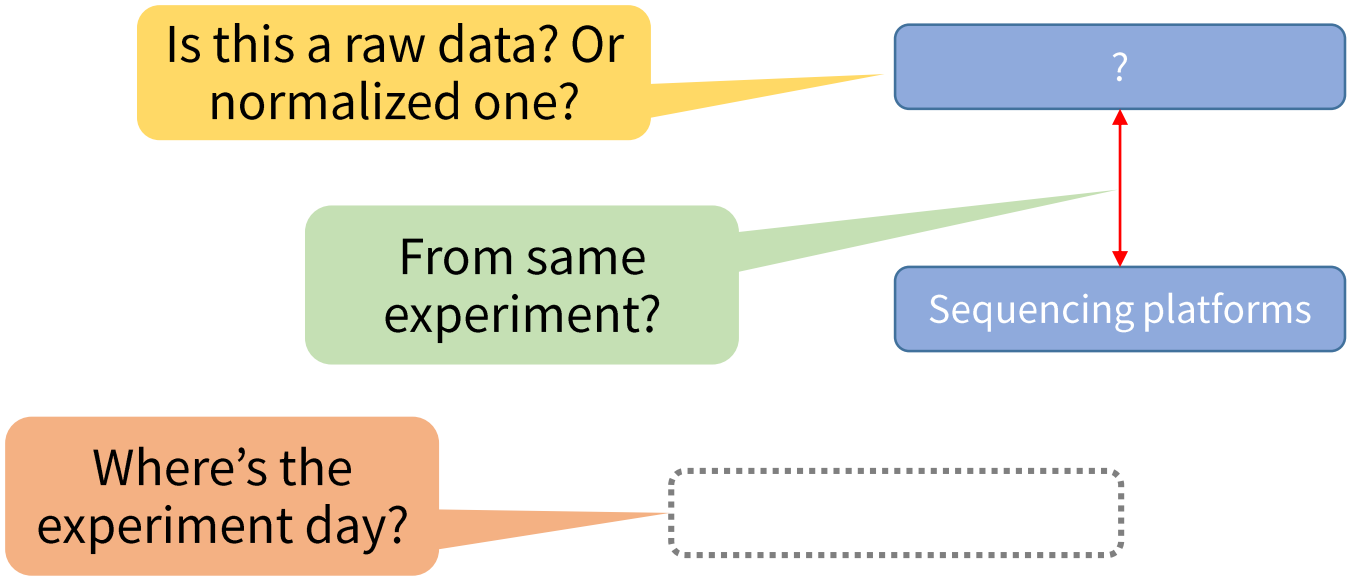


Why use S4 Class?

- Structured analysis
  - It helps end users to avoid confusing or mistakes



S4 type class



Using data separately