VIDO BIOINFO Bioinformatics and Systems Biology Lab

Canadian Bioinformatics Hub

Canadian Bioinformatics Workshops

# Introduction to R Programming for Bioinformatics

## Day 2- Module 3B: Exploring Key Bioconductor Packages and Datasets

**Mohamed Helmy, PhD**
Principal Scientist and Adjunct Professor
Bioinformatics and Systems Biology Lab
VIDO, University of Saskatchewan

6-7 October 2025, VIDO, Saskatoon

# Learning Objectives

By the end of this module, we should have knowledge on:

- Key Bioconductor packages
- Exploring assays, metadata, and annotations
- Subsetting treated vs untreated samples
- Using ExperimentHub and AnnotationHub
- Annotating genes

# Key Bioconductor Packages to Know

- **SummarizedExperiment**
  - Standard container for omics data (counts + metadata).
- **GenomicRanges**
  - Working with genomic coordinates (chromosome intervals).
- **AnnotationHub**
  - Access to genomes, gene models, regulatory features.
- **ExperimentHub**
  - Curated public datasets ready for analysis.
- **Biostrings**
  - Efficient manipulation of DNA/RNA/protein sequences.
- **org.Hs.eg.db (and similar org.* packages)**
  - Organism-level gene annotations.
- **limma / DESeq2 / edgeR**
  - Differential expression workflows.


**This is the “starter toolbox” for most bioinformatics projects.**

# The airway Dataset

- **RNA-seq from human airway smooth muscle cells.**
- **8 samples: treated (dexamethasone) vs untreated.**
- **Stored as a RangedSummarizedExperiment.**

```r
# Install airway package
BiocManager::install("airway")

# load package and data
library("airway")
data("airway")   # loads the dataset into your environment
airway
```

The type of object (the standard Bioconductor container for RNA-seq and other omics data)

The dataset has 63,677 rows (genes) and 8 columns (samples)

Extra dataset information (here, minimal)

The assay is the actual numeric data matrix. Here we have 1 assay, called counts (raw RNA-seq read counts)

The dataset has 63,677 rows (genes) and 8 columns (samples)

The rows are genes, identified by ENSEMBL gene IDs.

Extra metadata for each gene (row), includes gene name, chromosome, etc

The columns are RNA-seq samples

Metadata for the samples, includes info like cell line, treatment (dex), experiment ID, etc

```
class: RangedSummarizedExperiment
dim: 63677 8
metadata(1): ''
assays(1): counts
rownames(63677): ENSG00000000003 ENSG00000000005 ... ENSG00000273492 ENSG00000273493
rowData names(10): gene_id gene_name ... seq_coord_system symbol
colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
colData names(9): SampleName cell ... Sample BioSample
```

# Navigating Dataset Contents

- **Any Bioconductor dataset has:**
  - assays()
    - expression/counts data
    - actual data matrix you'll analyze
  - colData()
    - information about samples
    - sample metadata (treatment, sex, batch, etc.)
  - rowData()
    - information about genes
    - feature metadata (genes, probes)

| | SRR1039508 | SRR1039509 | SRR1039512 | SRR1039513 | SRR1039516 |
|---|---|---|---|---|---|
| ENSG00000000003 | 679 | 448 | 873 | 408 | 1138 |
| ENSG00000000005 | 0 | 0 | 0 | 0 | 0 |
| ENSG00000000419 | 467 | 515 | 621 | 365 | 587 |
| ENSG00000000457 | 260 | 211 | 263 | 164 | 245 |
| ENSG00000000460 | 60 | 55 | 40 | 35 | 78 |

DataFrame with 5 rows and 9 columns

| | SampleName | cell | dex | albut | Run | avgLength | Experiment | Sample | BioSample |
|---|---|---|---|---|---|---|---|---|---|
| | <factor> | <factor> | <factor> | <factor> | <factor> | <integer> | <factor> | <factor> | <factor> |
| SRR1039508 | GSM1275862 | N61311 | untrt | untrt | SRR1039508 | 126 | SRX384345 | SRS508568 | SAMN02422669 |
| SRR1039509 | GSM1275863 | N61311 | trt | untrt | SRR1039509 | 126 | SRX384346 | SRS508567 | SAMN02422675 |
| SRR1039512 | GSM1275866 | N052611 | untrt | untrt | SRR1039512 | 126 | SRX384349 | SRS508571 | SAMN02422678 |
| SRR1039513 | GSM1275867 | N052611 | trt | untrt | SRR1039513 | 87 | SRX384350 | SRS508572 | SAMN02422670 |
| SRR1039516 | GSM1275870 | N080611 | untrt | untrt | SRR1039516 | 120 | SRX384353 | SRS508575 | SAMN02422682 |

DataFrame with 5 rows and 10 columns

| | gene_id | gene_name | entrezid | gene_biotype | gene_seq_start | gene_seq_end | seq_name | seq_strand | seq_coord_system | symbol |
|---|---|---|---|---|---|---|---|---|---|---|
| | <character> | <character> | <integer> | <character> | <integer> | <integer> | <character> | <integer> | <integer> | <character> |
| ENSG00000000003 | ENSG00000000003 | TSPAN6 | NA | protein_coding | 99883667 | 99894988 | X | -1 | NA | TSPAN6 |
| ENSG00000000005 | ENSG00000000005 | TNMD | NA | protein_coding | 99839799 | 99854882 | X | 1 | NA | TNMD |
| ENSG00000000419 | ENSG00000000419 | DPM1 | NA | protein_coding | 49551404 | 49575092 | 20 | -1 | NA | DPM1 |
| ENSG00000000457 | ENSG00000000457 | SCYL3 | NA | protein_coding | 169818772 | 169863408 | 1 | -1 | NA | SCYL3 |
| ENSG00000000460 | ENSG00000000460 | C1orf112 | NA | protein_coding | 169631245 | 169823221 | 1 | 1 | NA | C1orf112 |

```
# Explor airway package
ex <- assay(airway)[1:5, 1:5]   # expression counts
cols <- colData(airway)[1:5, ]    # sample metadata
rows <- rowData(airway)[1:5, ]    # gene metadata
```

# Hands-on: Exploring and Subsetting

**For the airway dataset**

- **Get number of genes**
  - Use *nrow(),* it's a data frame
- **Subsetting treated vs untreated**
  - Use indexing with []
- **Count treated vs untreated**
  - Use *table()* on the dex column
- **Extract samples from a specific cell line**
  - Use the *cell* column
  - Extract data for the cell line named "N061011"

```
# Hands on
# Subsetting treated vs untreated
treated <- airway[, airway$dex == "trt"]
untreated <- airway[, airway$dex == "untrt"]

dim(treated)
dim(untreated)

# Count treated vs untreated
table(airway$dex)

# Extract samples from a specific cell line
subset_cell <- airway[, airway$cell == "N061011"]

# Get number of genes
nrow(airway)
```

# ExperimentHub-Accessing Curated Datasets

**What is that?**

- **A Bioconductor service that gives direct access to curated experimental datasets.**
- **Datasets are stored in the cloud and downloaded on demand.**
- **Once downloaded, they're cached locally for fast reuse.**
- **Great for exploring published data without manually downloading from GEO/ArrayExpress.**

**Why is it useful?**

- **Easy way to find datasets by keyword (e.g. "RNA-seq", "single-cell").**
- **Ensures you're working with standardized, curated data.**
- **Supports reproducibility in teaching and research.**

# Demo: ExperimentHub

- The query() function lets you filter datasets by organism, data type, or keywords.
- Each dataset has a unique ID (EHxxx) that you can load directly.

```r
# Load ExperimentHub
library(ExperimentHub)

# Create a hub object
eh <- ExperimentHub()

# Search for RNA-seq datasets
query(eh, "RNA-seq")

# Access a specific dataset by ID (example)
eh[["EH1234"]]   # Loads dataset into R
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 1988 features, 5 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: SRS014465 SRS014466 ... SRS062752 (5 total)
  varLabels: subjectID body_site ... NCBI_accession (18 total)
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
  pubMedIds: 22699609
Annotation:
```

# AnnotationHub - Accessing Annotation Resources

**What is that?**

- **A Bioconductor service providing genomic annotation resources.**
- **Gives access to reference genomes, gene models, regulatory features, and functional annotations.**
- **Data is stored in the cloud and cached locally for reuse.**

**Why is it useful?**

- **Easy to find reference annotations for your organism (e.g. human GRCh38).**
- **Ensures consistent, curated annotation data for reproducible analysis.**
- **No need to manually download GTF/GFF files or genome sequences from Ensembl/NCBI.**

# Demo: AnnotationHub

- query() lets you filter by species, genome build, or type of annotation.
- Each resource has a unique ID (AHxxxx) that you can load directly.

```r
# AnnotationHub Demo
# Load AnnotationHub
library(AnnotationHub)
library("rtracklayer")
# Create a hub object
ah <- AnnotationHub()

# Search for human genome resources
query(ah, "Homo sapiens")

# Access an annotation dataset by ID (example)
ah[["AH83281"]]   # Loads GRCh38 GTF annotation into R
```

```
AnnotationHub with 26727 records
# snapshotDate(): 2024-10-28
# $dataprovider: BroadInstitute, UCSC, Ensembl, GENCODE, Google DeepMind, UWashington, Stanford, Gencode, ENCODE, BioMart
# $species: Homo sapiens, homo sapiens
# $rdataclass: GRanges, BigWigFile, Rle, ChainFile, TwoBitFile, TxDb, list, data.frame, EnsDb, SQLiteFile
# additional mcols(): taxonomyid, genome, description, coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags, rdatapath, sourceurl,
#   sourcetype
# retrieve records with, e.g., 'object[["AH5012"]]'

            title
  AH5012   | Chromosome Band
  AH5013   | STS Markers
  AH5014   | FISH Clones
  AH5015   | Recomb Rate
  AH5016   | ENCODE Pilot
  ...        ...
  AH117067 | org.Hs.eg.db.sqlite
  AH117076 | TxDb.Hsapiens.UCSC.hg38.knownGene.sqlite
  AH117134 | MeSHDb for Homo sapiens (Human, v008)
  AH117228 | LRBaseDb for Homo sapiens (Human, v008)
  AH119325 | Ensembl 113 EnsDb for Homo sapiens
```

```
EnsDb for Ensembl:
|Backend: SQLite
|Db type: EnsDb
|Type of Gene ID: Ensembl Gene ID
|Supporting package: ensembldb
|Db created by: ensembldb package from Bioconductor
|script_version: 0.3.5
|Creation time: Sat Aug 22 04:51:11 2020
|ensembl_version: 101
|ensembl_host: localhost
|Organism: Oryzias sinensis
|taxonomy_id: 183150
|genome_build: ASM858656v1
|DBSCHEMAVERSION: 2.1
| No. of genes: 24022.
| No. of transcripts: 54551.
|Protein data available.
```

# The Bioconductor's family of org. packages

**What are they?**

- **A family of species-specific annotation databases.**

- **Provide mappings between:**
  - Gene IDs (ENSEMBL, Entrez, UniProt, RefSeq, etc.)
  - Gene symbols
  - Full gene names
  - Chromosome location and GO terms

- **Work the same way across all organisms using *mapIds()* or *select()* (from AnnotationDbi)**

- **Examples**
  - **Human** → org.Hs.eg.db (*Homo sapiens*)
  - **Mouse** → org.Mm.eg.db (*Mus musculus*)
  - **Rat** → org.Rn.eg.db (*Rattus norvegicus*)
  - **Fruit fly** → org.Dm.eg.db (*Drosophila melanogaster*)
  - **Worm** → org.Ce.eg.db (*Caenorhabditis elegans*)
  - **Arabidopsis** → org.At.tair.db (*Arabidopsis thaliana*)

# The org.Hs.eg.db package

- **What is it?**
  - org.Hs.eg.db is an annotation package for human genes.
  - It's part of Bioconductor's family of org.*.eg.db packages (one for each model organism).
  - Provides mappings between different types of gene identifiers and biological information.

- **Why use it?**
  - Datasets often use different gene IDs (ENSEMBL, Entrez, Affymetrix probes, etc.).
  - org.Hs.eg.db allows you to translate IDs into human-readable symbols and gene names.
  - Essential for downstream analysis (differential expression, pathway enrichment, reporting results).

- **What it contains:**
  - Gene identifiers: ENSEMBL, Entrez, UniProt, RefSeq, etc.
  - Gene symbols and full gene names.
  - Chromosomal locations.
  - GO terms and pathway annotations.

# Demo: org.Hs.eg.db

- **Human gene annotation package.**

- **Maps between IDs, symbols, Entrez, descriptions.**

```
# load packages
library(org.Hs.eg.db)
library(AnnotationDbi)

ids <- rownames(airway)[1:5]
mapIds(org.Hs.eg.db,
       keys = ids,
       keytype = "ENSEMBL",
       column = "SYMBOL")
```

```
ENSG00000000003 ENSG00000000005 ENSG00000000419 ENSG00000000457 ENSG00000000460
      "TSPAN6"          "TNMD"          "DPM1"         "SCYL3"         "FIRRM"
```

# Hands-on: Annotating Genes

**Student Tasks:**

- **Task 1:**
  - Take the first 20 genes from airway.
  - Map ENSEMBL IDs → gene symbols.
  - Retrieve gene descriptions.

- **Task 2**
  - Subset airway to treated samples only.
  - Select the first 5 genes.
  - Annotate them with symbols + full names using org.Hs.eg.db.

# Hands-on: Annotating Genes

```r
# Task 1: Take the first 20 genes from airway. Map ENSEMBL IDs → gene symbols.
# Retrieve gene descriptions.
library(airway)
data("airway")

library(org.Hs.eg.db)
library(AnnotationDbi)

# Get first 20 ENSEMBL IDs from airway
ids20 <- rownames(airway)[1:20]

# Map ENSEMBL → Gene Symbol
symbols <- mapIds(org.Hs.eg.db,
                  keys = ids20,
                  keytype = "ENSEMBL",
                  column = "SYMBOL")

# Map ENSEMBL → Full Gene Name
descriptions <- mapIds(org.Hs.eg.db,
                       keys = ids20,
                       keytype = "ENSEMBL",
                       column = "GENENAME")

# Combine into a data frame
annotated20 <- data.frame(ENSEMBL_ID = ids20,
                          Symbol = symbols,
                          Description = descriptions)

head(annotated20)
```

| | ENSEMBL_ID | Symbol | Description |
|---|---|---|---|
| ENSG00000000003 | ENSG00000000003 | TSPAN6 | tetraspanin 6 |
| ENSG00000000005 | ENSG00000000005 | TNMD | tenomodulin |
| ENSG00000000419 | ENSG00000000419 | DPM1 | dolichyl-phosphate mannosyltransferase subunit 1, catalytic |
| ENSG00000000457 | ENSG00000000457 | SCYL3 | SCY1 like pseudokinase 3 |
| ENSG00000000460 | ENSG00000000460 | FIRRM | FIGNL1 interacting regulator of recombination and mitosis |
| ENSG00000000938 | ENSG00000000938 | FGR | FGR proto-oncogene, Src family tyrosine kinase |

# Hands-on: Annotating Genes

```r
# Task 2: Subset airway to treated samples only. Select the first 5 genes.
# Annotate them with symbols + full names.
# Subset treated samples
treated <- airway[, airway$dex == "trt"]

# Get first 5 ENSEMBL IDs from treated dataset
ids5 <- rownames(treated)[1:5]

# Map ENSEMBL → Symbol
symbols5 <- mapIds(org.Hs.eg.db,
                   keys = ids5,
                   keytype = "ENSEMBL",
                   column = "SYMBOL")

# Map ENSEMBL → Gene Name
names5 <- mapIds(org.Hs.eg.db,
                 keys = ids5,
                 keytype = "ENSEMBL",
                 column = "GENENAME")

# Combine results
annotated5 <- data.frame(ENSEMBL_ID = ids5,
                         Symbol = symbols5,
                         Full_Name = names5)

annotated5
```

| | ENSEMBL_ID | Symbol | Full_Name |
|---|---|---|---|
| ENSG00000000003 | ENSG00000000003 | TSPAN6 | tetraspanin 6 |
| ENSG00000000005 | ENSG00000000005 | TNMD | tenomodulin |
| ENSG00000000419 | ENSG00000000419 | DPM1 | dolichyl-phosphate mannosyltransferase subunit 1, catalytic |
| ENSG00000000457 | ENSG00000000457 | SCYL3 | SCY1 like pseudokinase 3 |
| ENSG00000000460 | ENSG00000000460 | FIRRM | FIGNL1 interacting regulator of recombination and mitosis |

# THANK YOU

VIDO

VACCINE AND INFECTIOUS DISEASE ORGANIZATION

VIDO.ORG

UNIVERSITY OF SASKATCHEWAN