BRM5012 R practical: Bioconductor

## Flavours of R

### Base R

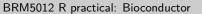
```
# Example
install.packages("glmnet")
```

#### Tidyverse

```
# Install entire Tidyverse
install.packages("tidyverse")
```

#### Bioconductor

```
# Setup
source("https://bioconductor.org/biocLite.R")
biocLite()
# Install a specific package
biocLite("limma")
```



2018-08-21

Flavours of R



R has 4 decades of history.

"Base R" is standard packages and CRAN, mostly designed for use by statisticians.

Tidyverse is a set of modern packages, also on CRAN.

Tidyverse is easy to pick up, has made R much more widely accessible. Key concepts are tidy data and pipelines.

Bioconductor is a separate package repository.

Bioconductor packages are for dealing with biological data efficiently. Bioconductor is not tidy. Expect to have to work a little harder to make use of these.

### What Bioconductor is for

#### Differential expression analysis from microarray or RNA-seq

- ▶ limma, edgeR, DESeq2, ...
- ► Further packages to normalize, impute, batch correct, check quality, . . .

#### Visualization

► ComplexHeatmap, Gviz, ggbio, ...

### Peaks, regions, motifs in genome (eg ChIP-seq)

▶ GenomicRanges, GenomicFeatures, ...

### Single-cell sequencing

Statistical methods for  $p \gg n$  data

. . .



Too many uses to count, ask a different person and get a different list.

50% of bioinformatics is read this file format, wrangle, write other file format.

Often used in combination with command-line software (eg read aligner).

# More types

Up to now we have been using vectors and data frames (specifically Tidyverse "tibbles").

To use Bioconductor, several further types are important.

- ▶ matrix
- ▶ DNAString, DNAStringSet, AAString
- GRanges
- ▶ Others: SeqInfo, TxDb, EnsDb, OrgDb, SummarizedExperiment, ...

matrix is another tabular data type from base R. Conceptually, each cell is an observation, rather than each row as in a tidy data frame.

More types

The rest are Bioconductor types. Note the use of CamelCase, this is typical in Bioconductor. Tidyverse separates words with underscores, and base R uses dots!

Bioconductor has evolved over time to cope with more complex data.

## Getting reference data

#### Analysis usually occurs in the context of:

- ▶ a genome assembly: DNA sequences of chromosomes
- gene annotations: ranges of positions in chromosome and strand

More on this later in the week!

#### In Bioconductor:

- ▶ AnnotationHub package gives access to reference data from various sources
  - access specific versions, good for reproducibility
- rtracklayer package to load files

#### Others

- ▶ biomaRt to access BioMart servers
- Data packages

BRM5012 R practical: Bioconductor

2018-08-2

-Getting reference data

Getting reference data

Analysis usually occurs in the context of a genome assembly: DNA sequences of chromosomes · pene annotations: ranges of positions in chromosome and strand

· AppotationBub package gives access to reference data from vario

. access specific versions, good for reproducibility rtracklayer package to load files

► biomalit to access BioMart sensors

► Data nackansa

## Documentation

Most packages have "vignettes" explaining normal usage.

Information about types available with, eg:

?"GRanges-class"



Aim for this session is to cover fundamentals of Bioconductor, so you can look up the package you need for your particular task and understand how to use it.