GENOMIC BASIS OF EVOLUTIONARY INNOVATIONS
GEvol

JG|U
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Differential splicing analysis

Joe Colgan

JGU Mainz

tcolgan@uni-mainz.de

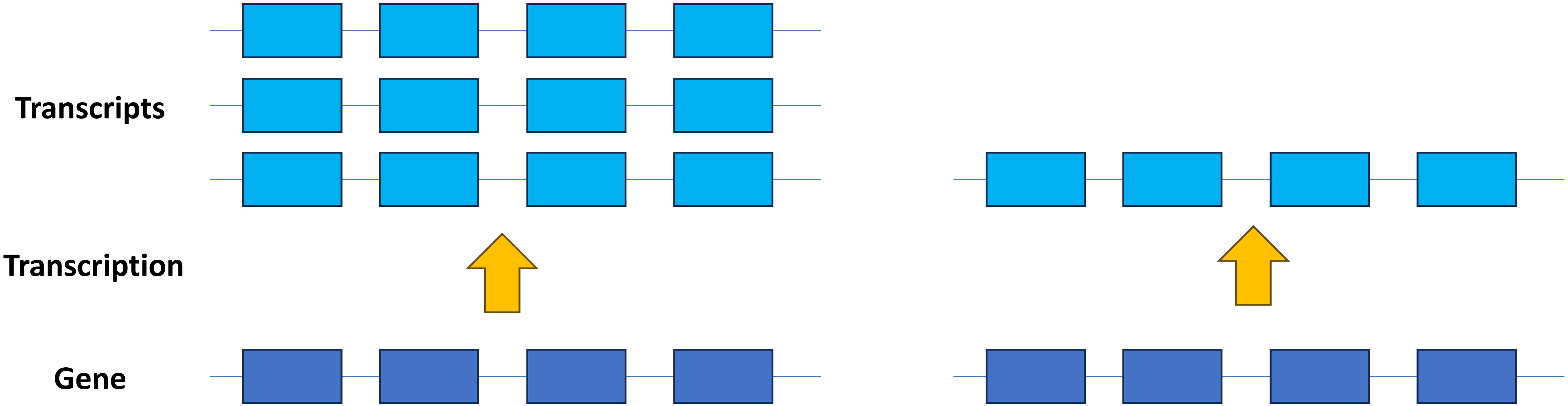
Workshop outline

Introduction into identifying and quantifying splicing events

Processing and visualisation of splicing events

Differential gene expression

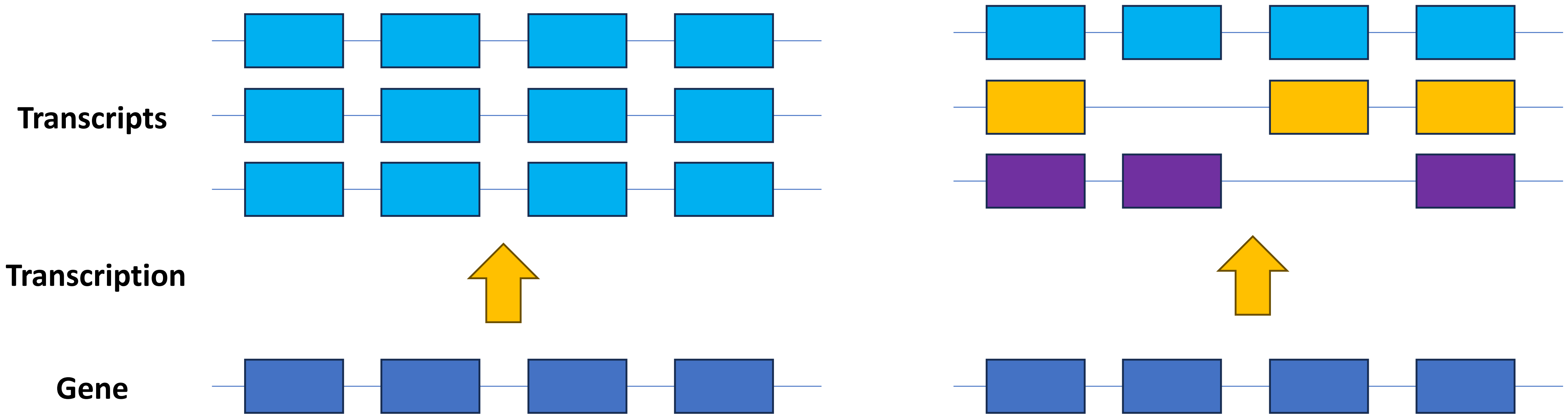
Differences in **amplitude**
(i.e., the amount a gene is transcribed)



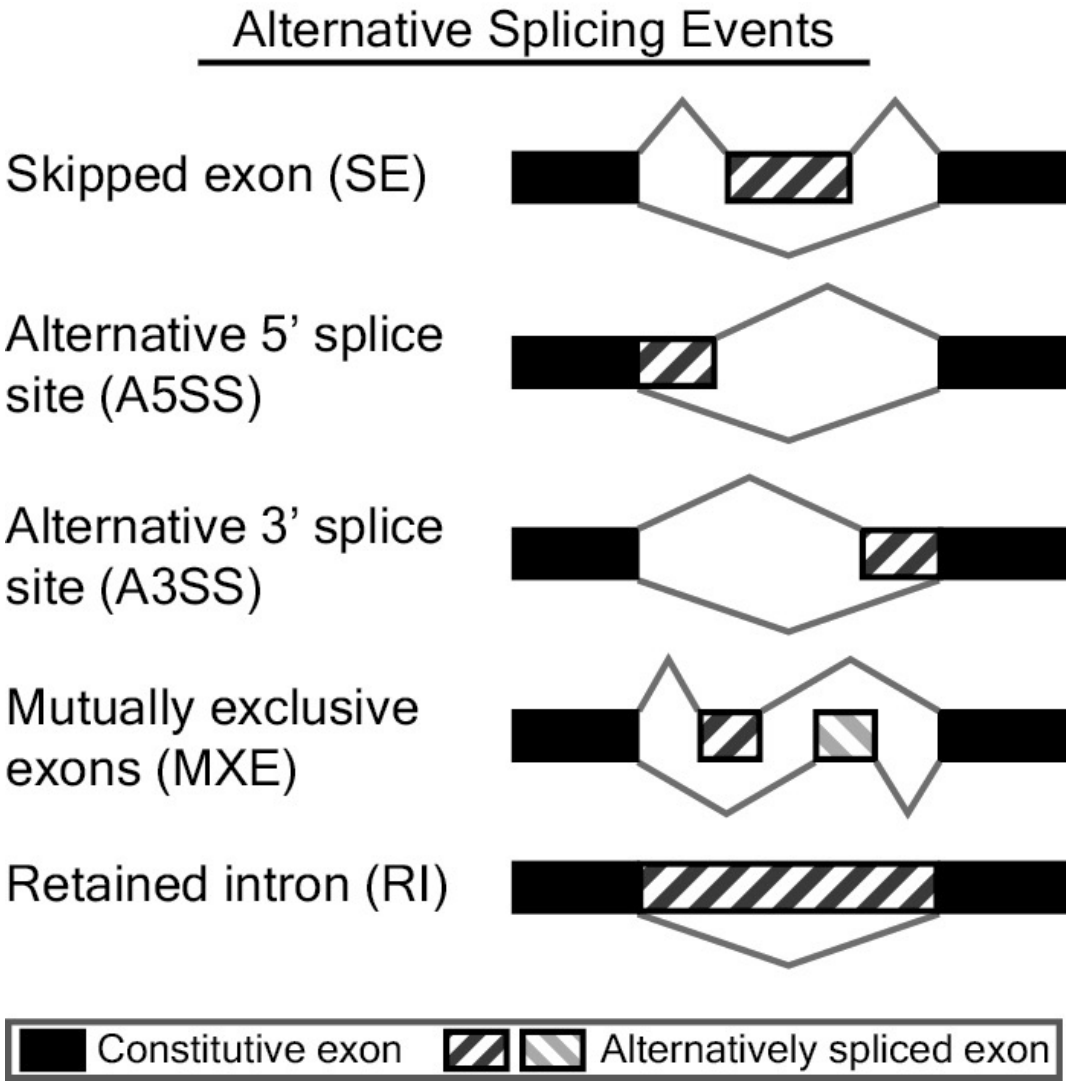
Differential gene expression

Changes in ***amplitude***
(i.e., the amount a gene is transcribed)

Changes in ***isoforms***
(i.e., what a gene is transcribing)



Common splicing events



Alternative Splicing Events

Skipped exon (SE)

Alternative 5’ splice site (A5SS)

Alternative 3’ splice site (A3SS)

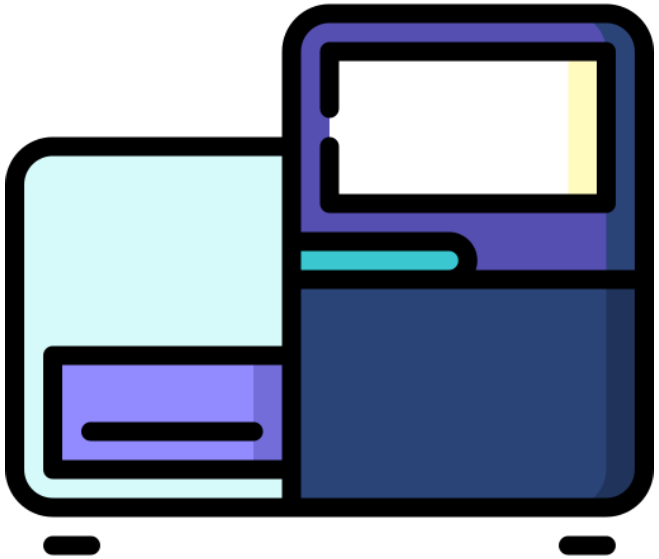
Mutually exclusive exons (MXE)

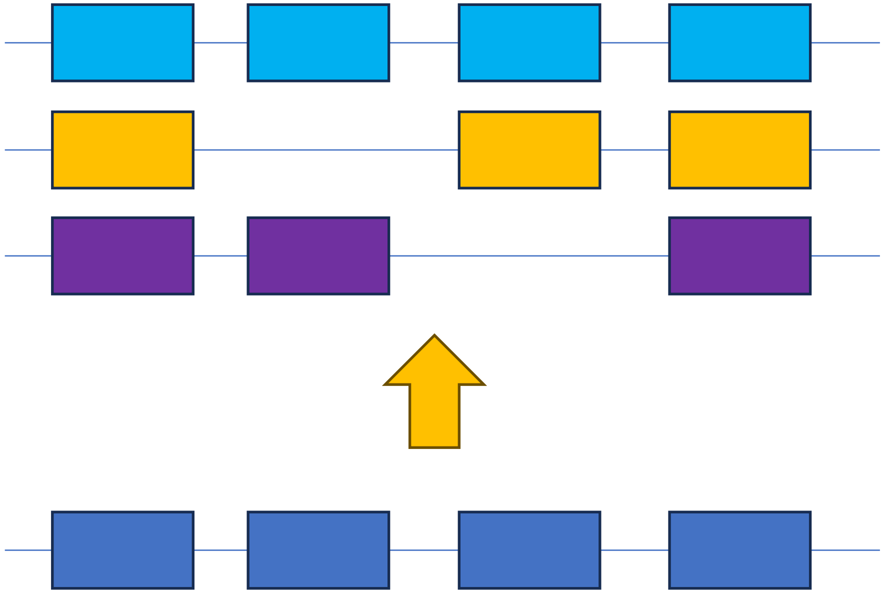
Retained intron (RI)

■ Constitutive exon ▨ ▨ Alternatively spliced exon

How can we measure alternative splicing?

The basic principle in RNA-Seq analysis of alternative splicing is to use RNA-Seq reads mapped to different isoforms to estimate the isoform proportion





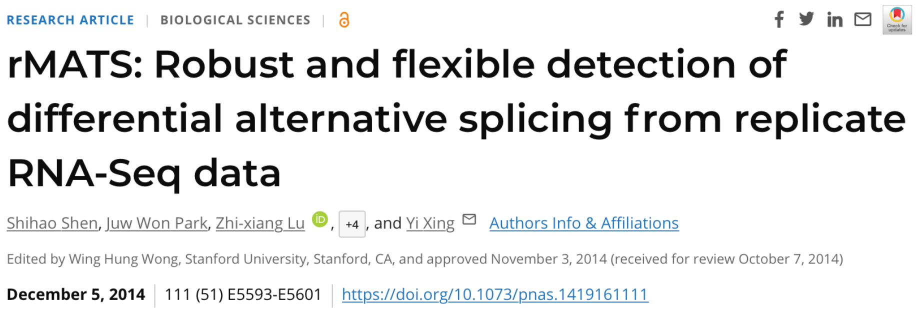
rMATs

Based on MATs (Multivariate analysis of transcript splicing)

Detect differential alternative splicing events from RNA-Seq data.

The statistical model of MATS calculates the P-value and false discovery rate that the difference in the isoform ratio of a gene between two conditions exceeds a given user-defined threshold.

From the RNA-Seq data, MATS can automatically detect and analyze alternative splicing events corresponding to all major types of alternative splicing patterns



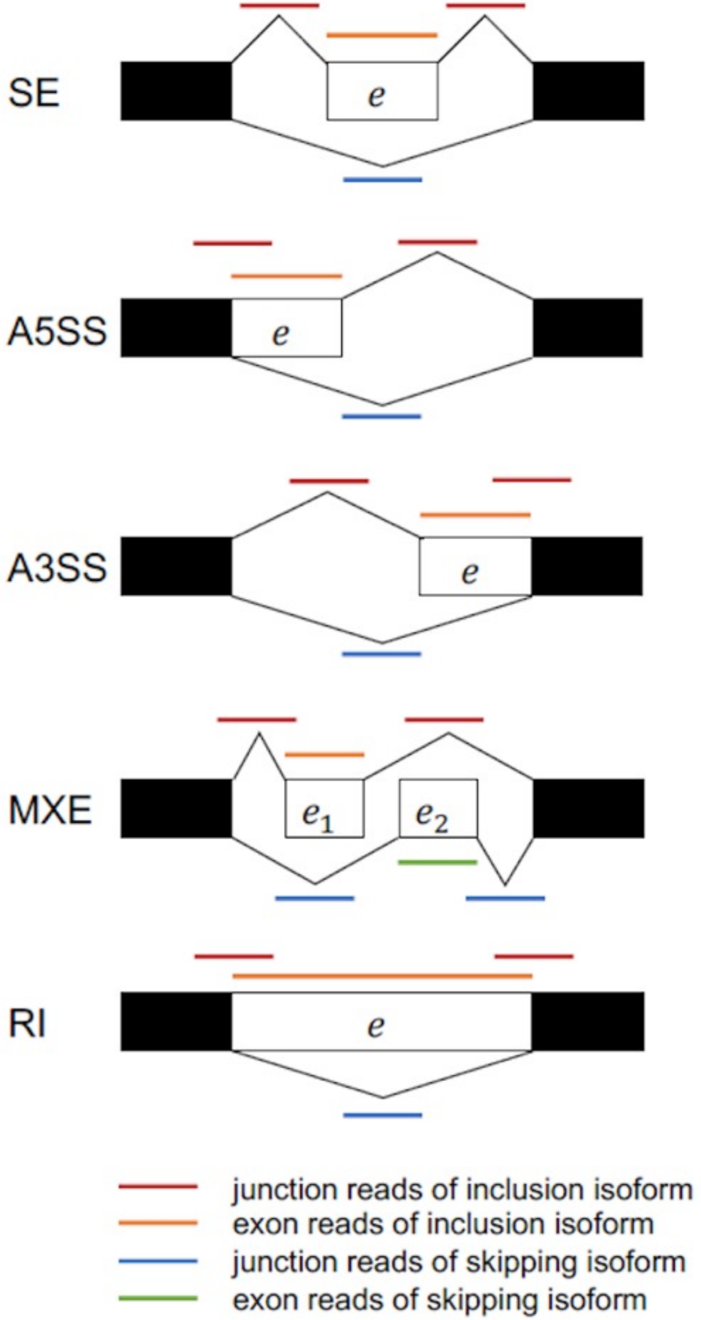
RESEARCH ARTICLE | BIOLOGICAL SCIENCES |

rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data

Shihao Shen, Juw Won Park, Zhi-xiang Lu, +4, and Yi Xing Authors Info & Affiliations

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved November 3, 2014 (received for review October 7, 2014)

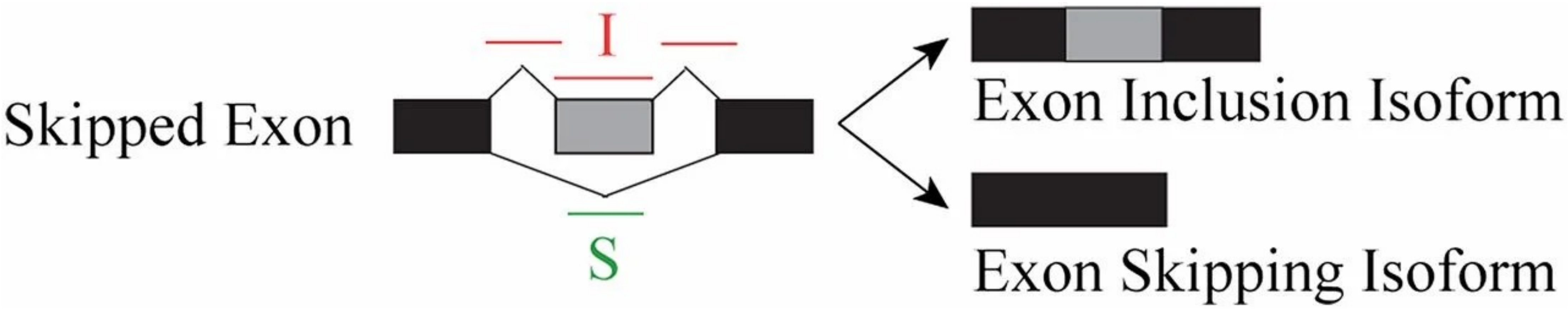
December 5, 2014 | 111 (51) E5593-E5601 | https://doi.org/10.1073/pnas.1419161111


SE
e
A5SS
e
A3SS
e
MXE
e_1
e_2
RI
e
junction reads of inclusion isoform
exon reads of inclusion isoform
junction reads of skipping isoform
exon reads of skipping isoform

How can we measure alternative splicing?

Percent Spliced in (PSI): Ratio between reads including (I) or excluding exons (S)

Compare between mean PSI between two groups using likelihood ratio test

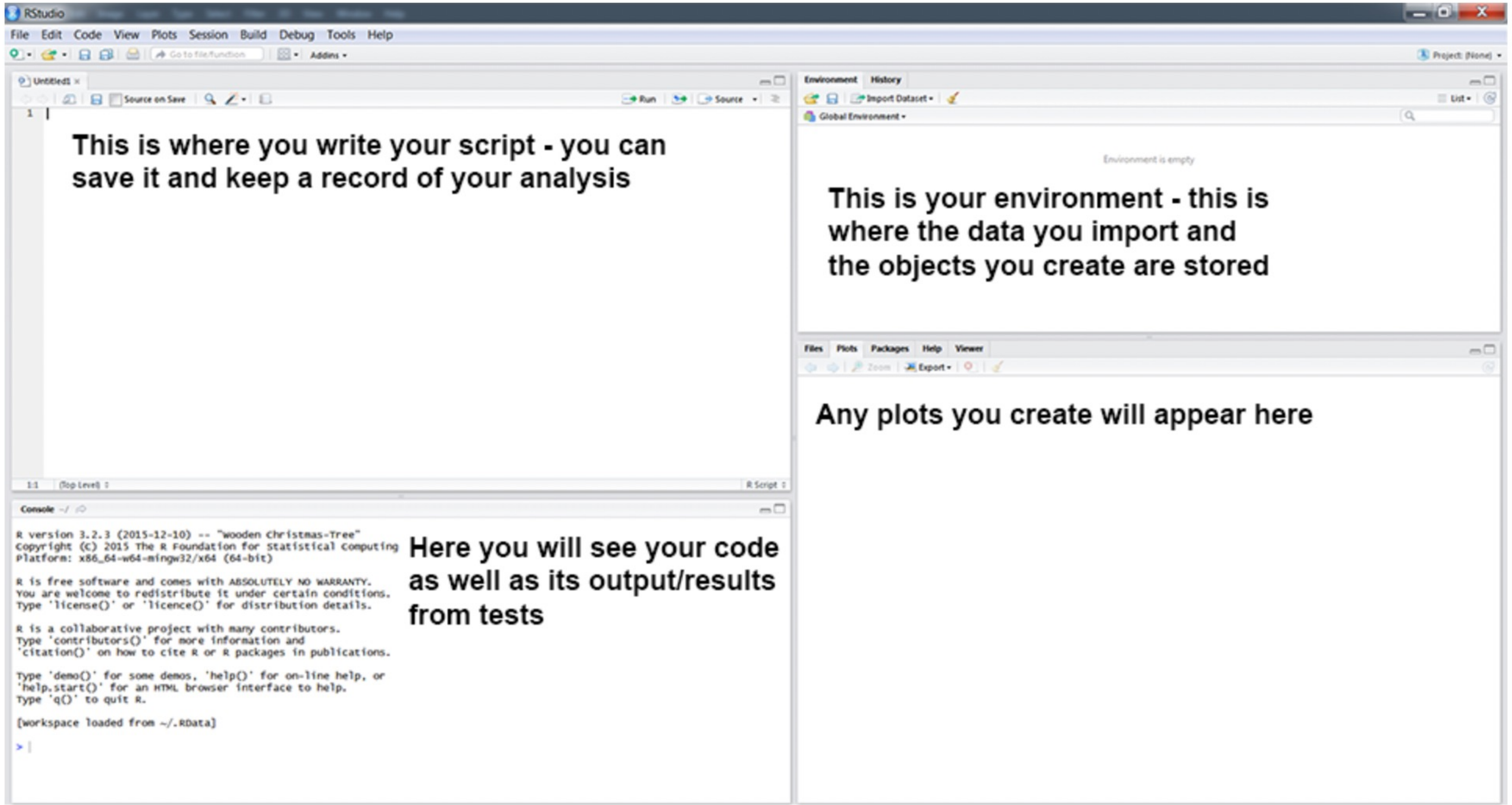


Website: https://rnaseq-mats.sourceforge.io

Introduction to R

Today, we will do our analysis in RStudio:

- A standard RStudio consists of four panels



This is where you write your script - you can save it and keep a record of your analysis

This is your environment - this is where the data you import and the objects you create are stored

Any plots you create will appear here

```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[workspace loaded from ~/.RData]

> |
```

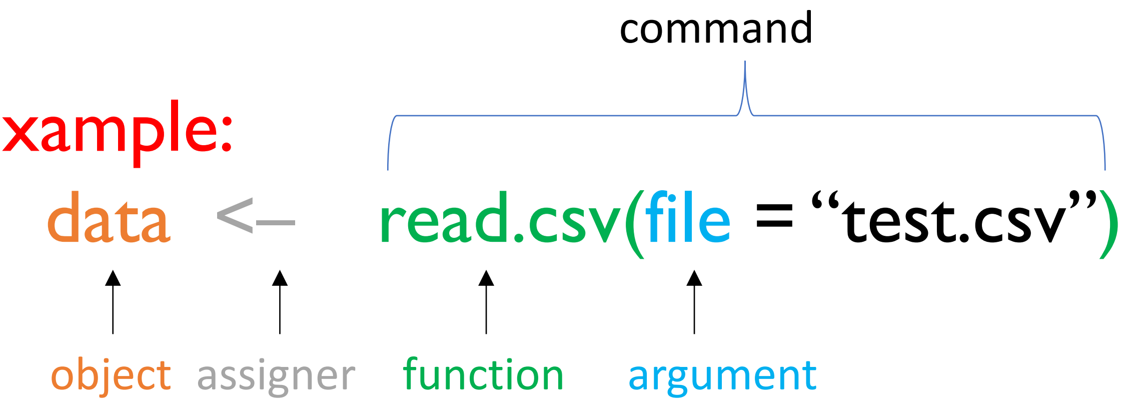
Here you will see your code as well as its output/results from tests

Introduction to R

R is an object-based language

- This means the data you import or values you create during your R session are stored in objects

For example:



command

```
data  <–   read.csv(file = “test.csv”)
```

object assigner function argument

In this command:

- the ‘read.csv()’ takes the file ’test.csv’ as input and assigns the information to the object ‘data’ where the information is stored.
- objects can contain different data such as vectors, data frames, matrices, and lists.

Introduction to R

How data are stored:

-There are four main classes of data:

- Integer: only whole numbers (1, 2, 3, 4, 5, 6, 7, 8, 9, 100, 1000, 260000)
- Numerical: whole and floating values (0.1, 2.3, 3.5556, 5, 6, 7.8, 8.9, 0.0007)
- Character: (‘a’, ‘b’, ’dog’, ‘bee’, ‘1’, ‘7.8’, ‘4000’)
- Factor:
  - Objects where you can add additional information (categories) to your data structure

For example: Categorise cities by country

Cities	**Country**
Berlin	Germany
Paris	France
Frankfurt	Germany
Mainz	Germany
London	UK

These categories are then known as **levels**

You can find out the class of an object by using the ‘str()’ command, which stands for structure

Introduction to R

Functions:

There are three main classes of functions:

- Base R functions: these are functions that come already come with Rstudio (e.g., nrow(), dim(), str())

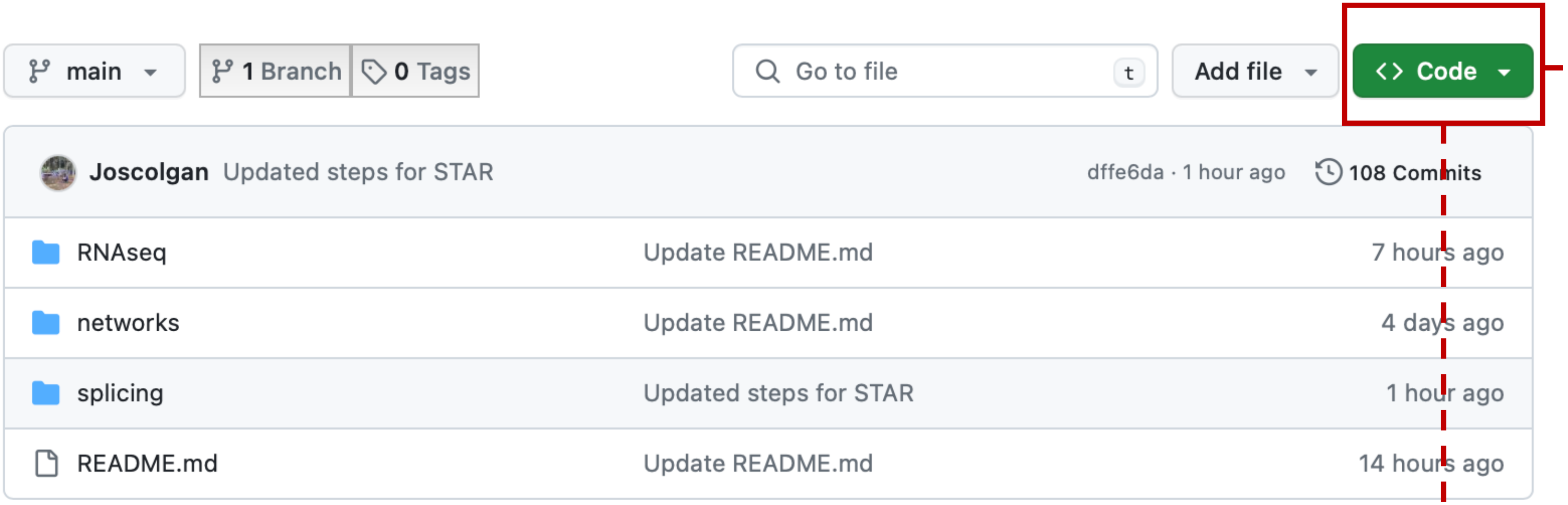
- Package functions: these are functions that come with an R package and the package must be loaded before the function can be used (e.g., ggplot(), dplyr(), euler()).
  - NB: Sometimes functions from different packages may be called the same name (e.g., plot()) – to ensure you are using the right function, we can specify the function from a particular function by using the syntax: package::function() (e.g., dplyr::select())

- Custom functions: functions that you can write yourself

What you have for today

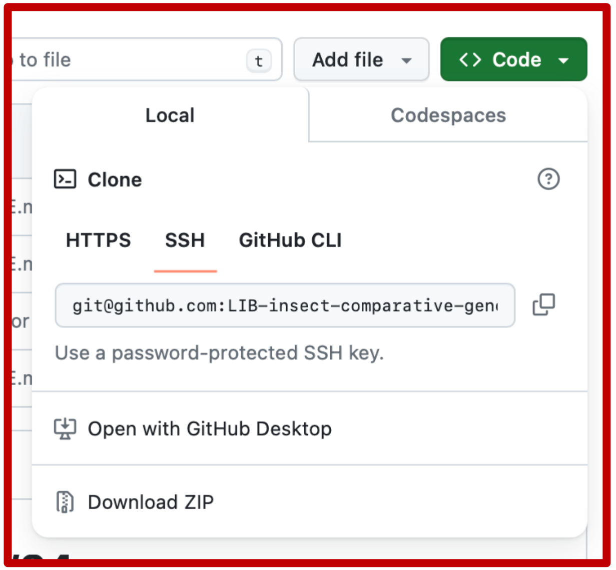
We have a directory on Github that you can download/clone:

https://github.com/LIB-insect-comparative-genomics/RNAseq-workshop2023



Joscolgan Updated steps for STAR · dffe6da · 1 hour ago · 108 Commits

RNAseq	Update README.md	7 hours ago
networks	Update README.md	4 days ago
splicing	Updated steps for STAR	1 hour ago
README.md	Update README.md	14 hours ago

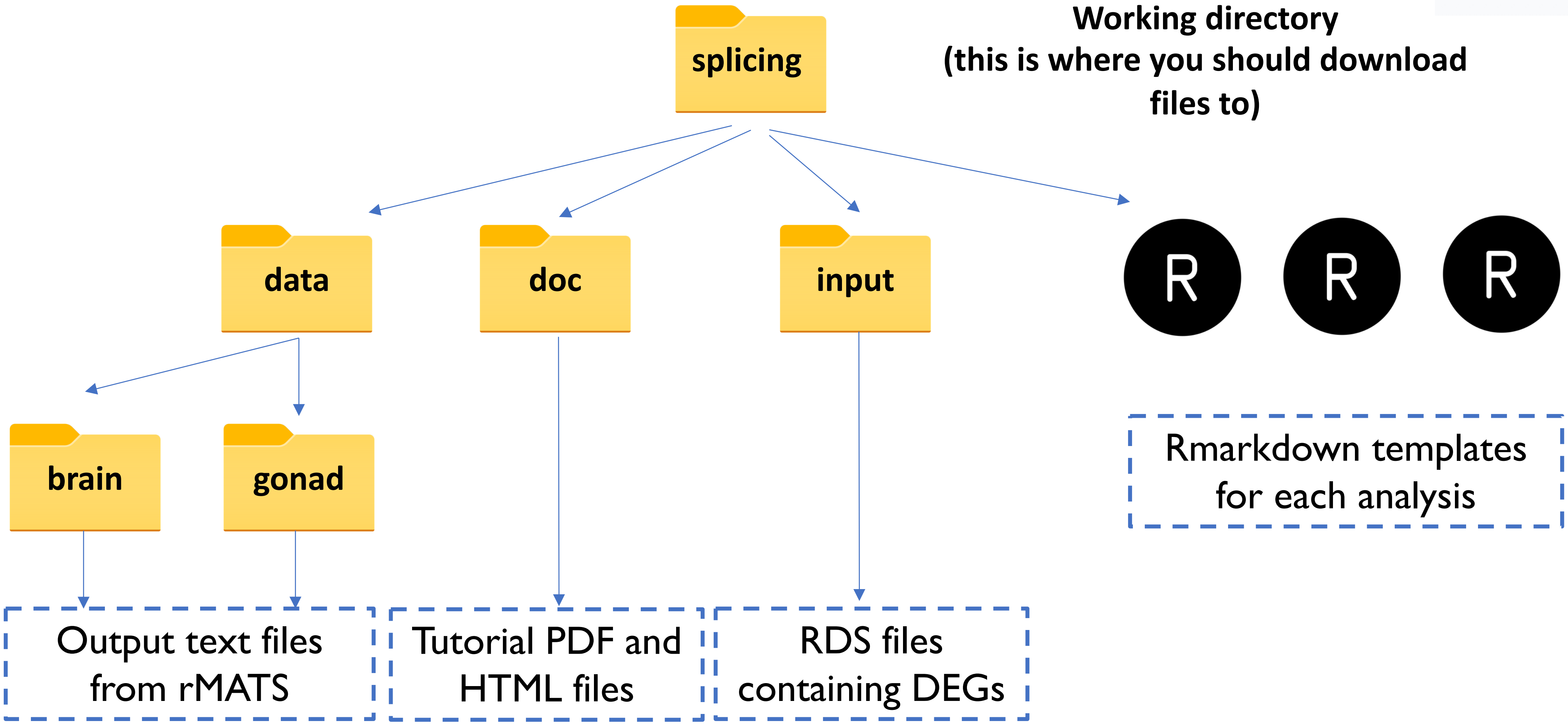


git@github.com:LIB-insect-comparative-gen

Use a password-protected SSH key.

```
git clone git@github.com:LIB-insect-comparative-genomics/RNAseq-workshop2023.git
```

What you have for today



splicing

Working directory
(this is where you should download files to)

data

doc

input

R R R

brain

gonad

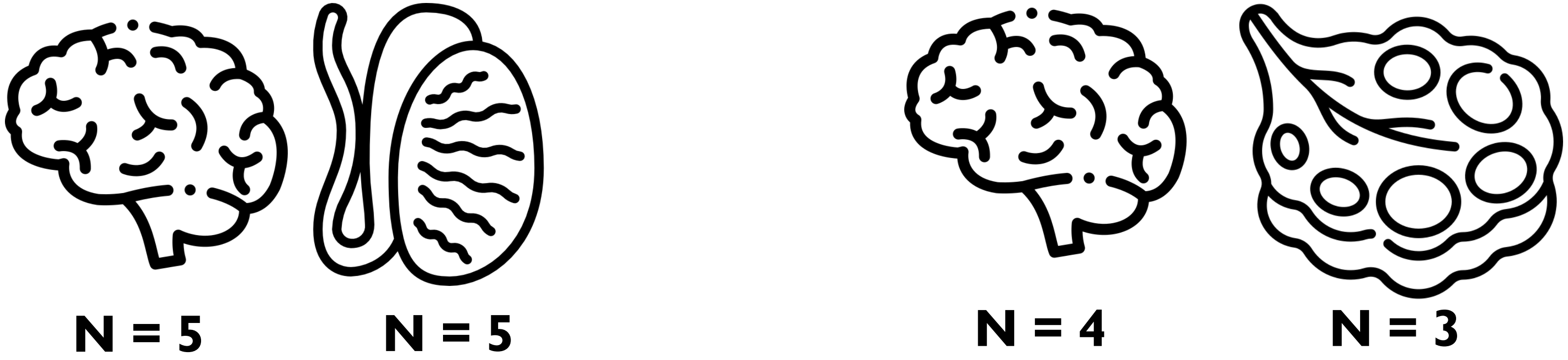
Output text files from rMATS

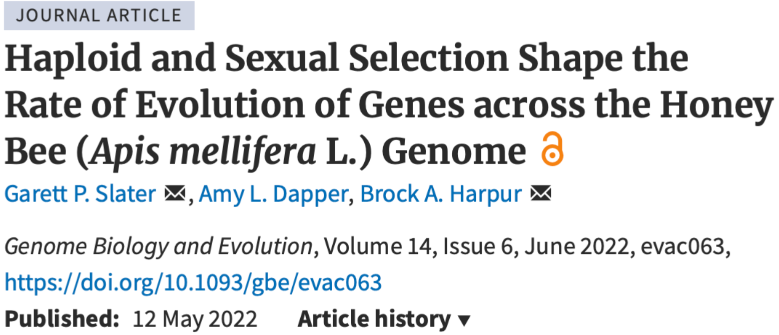
Tutorial PDF and HTML files

RDS files containing DEGs

Rmarkdown templates for each analysis

Dataset for today: Honeybees





JOURNAL ARTICLE

Haploid and Sexual Selection Shape the Rate of Evolution of Genes across the Honey Bee (*Apis mellifera* L.) Genome

Garett P. Slater, Amy L. Dapper, Brock A. Harpur

Genome Biology and Evolution, Volume 14, Issue 6, June 2022, evac063,

https://doi.org/10.1093/gbe/evac063

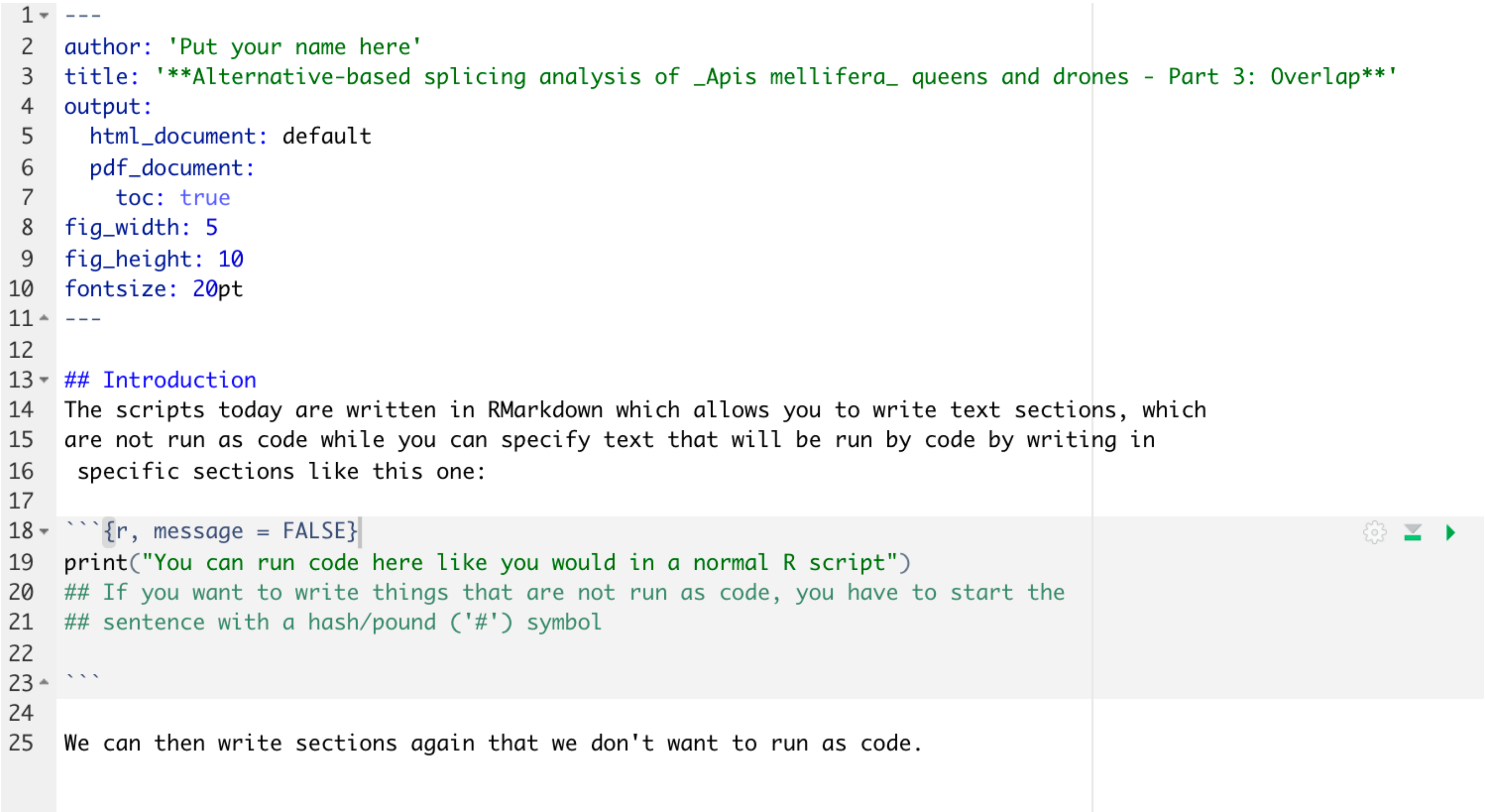
Published: 12 May 2022 **Article history** ▾

- Reads downloaded from NCBI SRA database [sratool]
- Aligned against latest reference genome [STAR]
- Identification of splicing events [rMATs]

https://academic.oup.com/gbe/article/14/6/evac063/6584681

What is an R Markdown script?

It looks like a fancier version of a regular R script but has some great features:
- You can clearly separate sections with and without code (easier to read and add notes/annotations)
- Using the knitR function, you can convert your script into a html, PDF, or word document



~~~
---
author: 'Put your name here'
title: '**Alternative-based splicing analysis of _Apis mellifera_ queens and drones - Part 3: Overlap**'
output:
  html_document: default
  pdf_document:
    toc: true
fig_width: 5
fig_height: 10
fontsize: 20pt
---

## Introduction
The scripts today are written in RMarkdown which allows you to write text sections, which
are not run as code while you can specify text that will be run by code by writing in
 specific sections like this one:

```{r, message = FALSE}
print("You can run code here like you would in a normal R script")
## If you want to write things that are not run as code, you have to start the
## sentence with a hash/pound ('#') symbol

```

We can then write sections again that we don't want to run as code.
~~~

What you have for today

We have three analyses today – three scripts that we want you to work through that will cover:

- **Processing** and **visualising** results of splicing events between queens and males for two tissues (brains and gonads)

- **Comparison** between alternatively spliced genes and those that differ in amplitude

Helpful resources for self-learning

- Structuring folders for computational analyses
- Improving in R and statistics (https://ourcodingclub.github.io)
- Introduction to Unix shell (https://swcarpentry.github.io/shell-novice/)
- Statistics & Machine learning (https://www.youtube.com/@statquest)
- Scripts for non-model organisms (mainly bees and fish): https://github.com/Joscolgan

Questions or feedback: tcolgan@uni-mainz.de