Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

Bioinformatics

Jean-Claude Walser
jean-claude.walser [at] usys.ethz.ch

HS2024

Naming

Suggestions to name files:

- meaningful and unique file names
- keep it simple and logical → encoding?
- avoid special characters (e.g. ä)
- careful with numbers (e.g. 1 or 01)
- correct extensions (e.g. fasta, fastq, bam)

Sample-File-Names

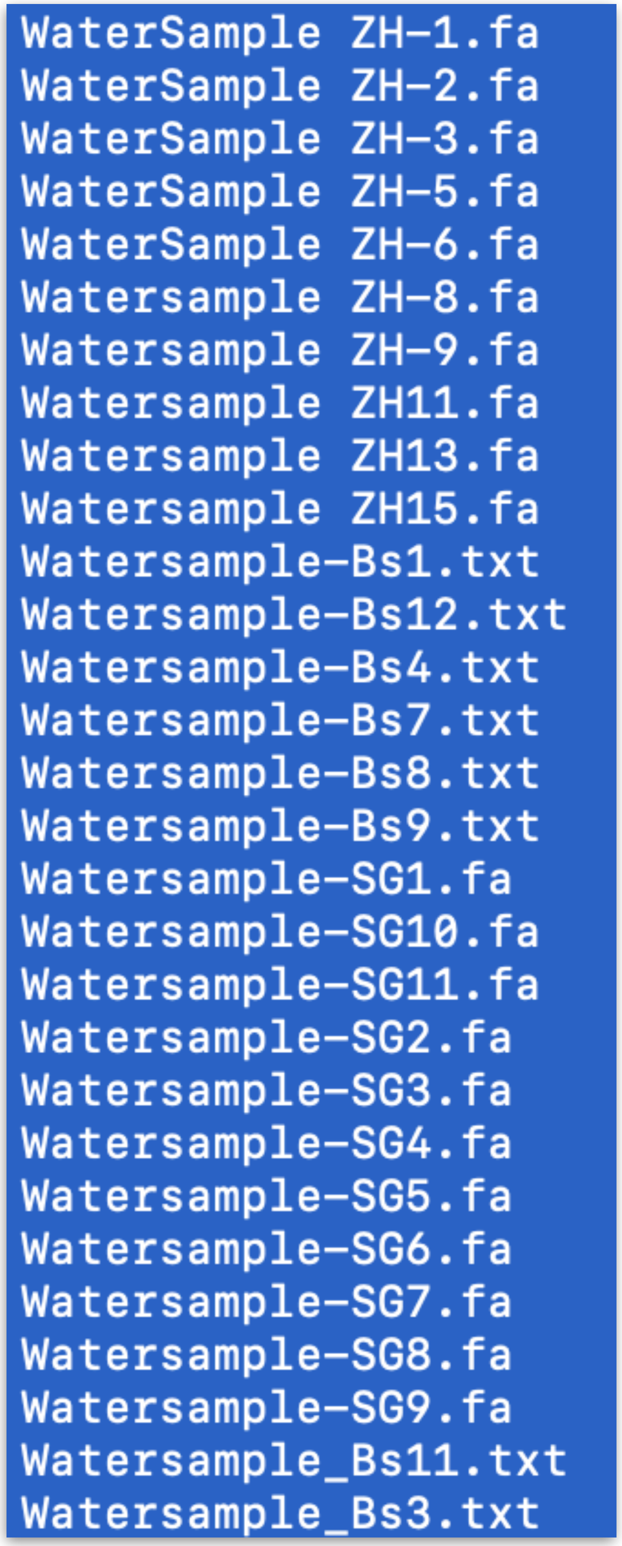
```
Watersample-SG1.txt
Watersample-SG2.txt
Watersample-SG3.txt
Watersample-SG4.txt
Watersample-SG5.txt
Watersample-SG6.txt
Watersample-SG7.txt
Watersample-SG8.txt
Watersample-SG9.txt
Watersample-SG10.txt
Watersample-SG11.txt
```

```
Watersample-Bs1.txt
Watersample_Bs3.txt
Watersample-Bs4.txt
Watersample-Bs7.txt
Watersample-Bs8.txt
Watersample-Bs9.txt
Watersample_Bs11.txt
Watersample-Bs12.txt
```

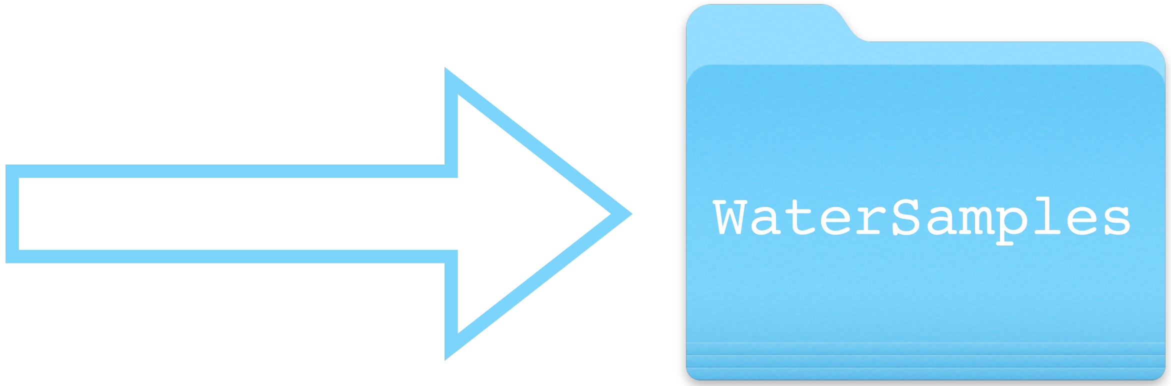
```
WaterSample ZH-1.fa
WaterSample ZH-2.fa
WaterSample ZH-3.fa
WaterSample ZH-5.fa
WaterSample ZH-6.fa
Watersample ZH-8.fa
Watersample ZH-9.fa
Watersample ZH11.fa
Watersample ZH13.fa
Watersample ZH15.fa
```

- Format: <Name>-<ID>.extention
- Extensions: txt and fa
- Naming: Watersample - WaterSample -> all are water samples
- Separator: -ID or _ID or white-space
- ID: lowercase (Bs) and uppercase (SG, ZH) mix
- Numbers: different length (SG1 and SG11) or format (Bs1 and ZH-1)

Bioinformatics ▷ Basics



```
WaterSample ZH-1.fa
WaterSample ZH-2.fa
WaterSample ZH-3.fa
WaterSample ZH-5.fa
WaterSample ZH-6.fa
Watersample ZH-8.fa
Watersample ZH-9.fa
Watersample ZH11.fa
Watersample ZH13.fa
Watersample ZH15.fa
Watersample-Bs1.txt
Watersample-Bs12.txt
Watersample-Bs4.txt
Watersample-Bs7.txt
Watersample-Bs8.txt
Watersample-Bs9.txt
Watersample-SG1.fa
Watersample-SG10.fa
Watersample-SG11.fa
Watersample-SG2.fa
Watersample-SG3.fa
Watersample-SG4.fa
Watersample-SG5.fa
Watersample-SG6.fa
Watersample-SG7.fa
Watersample-SG8.fa
Watersample-SG9.fa
Watersample_Bs11.txt
Watersample_Bs3.txt
```



```
BS01.fa
BS03.fa
BS04.fa
BS07.fa
BS08.fa
BS09.fa
BS11.fa
BS12.fa
SG01.fa
SG02.fa
SG03.fa
SG04.fa
SG05.fa
SG06.fa
SG07.fa
SG08.fa
SG09.fa
SG10.fa
SG11.fa
ZH01.fa
ZH02.fa
ZH03.fa
ZH05.fa
ZH06.fa
ZH08.fa
ZH09.fa
ZH11.fa
ZH13.fa
ZH15.fa
```

Directory (Folder) - Names

`Evolutionary Genetics Exercises` ✗

🐍 `Evolutionary_Genetics_Exercises` ✓

🐪 `EvolutionaryGeneticsExercises` ✓

`EvoGenEx` ✓

Script-Names

```
error_correction.ph
error_correction_PE.ph
error_correction_PE_1.ph
error_correction_PE_new.ph
error_correction_SR_newer.ph
error_correction_SR_v190423.ph
```

▼

```
ReadME.txt
error_correction_SR_v190116.ph.archive
error_correction_SR_v190423.ph.archive
error_correction_SR_v190502.ph
error_correction_PE_v200317.ph.archive
error_correction_PE_v200421.ph
```

Bioinformatics ▷ Basics

Careful with renaming original files!

`original_file_name.txt` ▶ `original_file_name_`**`copy`**`.txt`

✗ ✓

`new_file_name.txt` `new_file_name.txt`

Basics

MyFile**.txt** / MyTable**.csv** / MySequence**.fa**

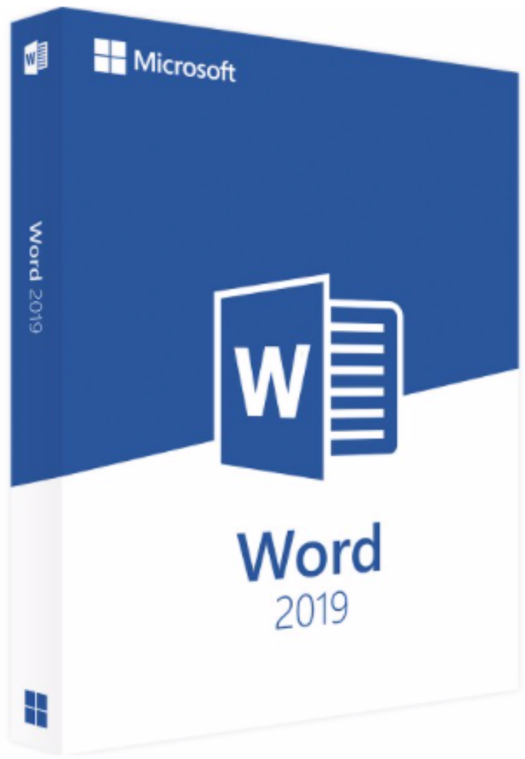
File extensions are important because they indicate the type of file and help both the operating system and users understand how to open and interact with the file. For example, .txt indicates a text file, .csv indicates a comma-separated values file, and .fa indicates an fastq formatted sequence file. Extensions allow software to correctly identify the file, open it, or run it with the appropriate application. They also play a role in security, as changing the extension of a file can sometimes hide its true nature, so understanding file extensions is crucial for both functionality and security in computing.

Writing

Scientific Writing

- Reports
- Publications

Bioinformatics ▷ Basics



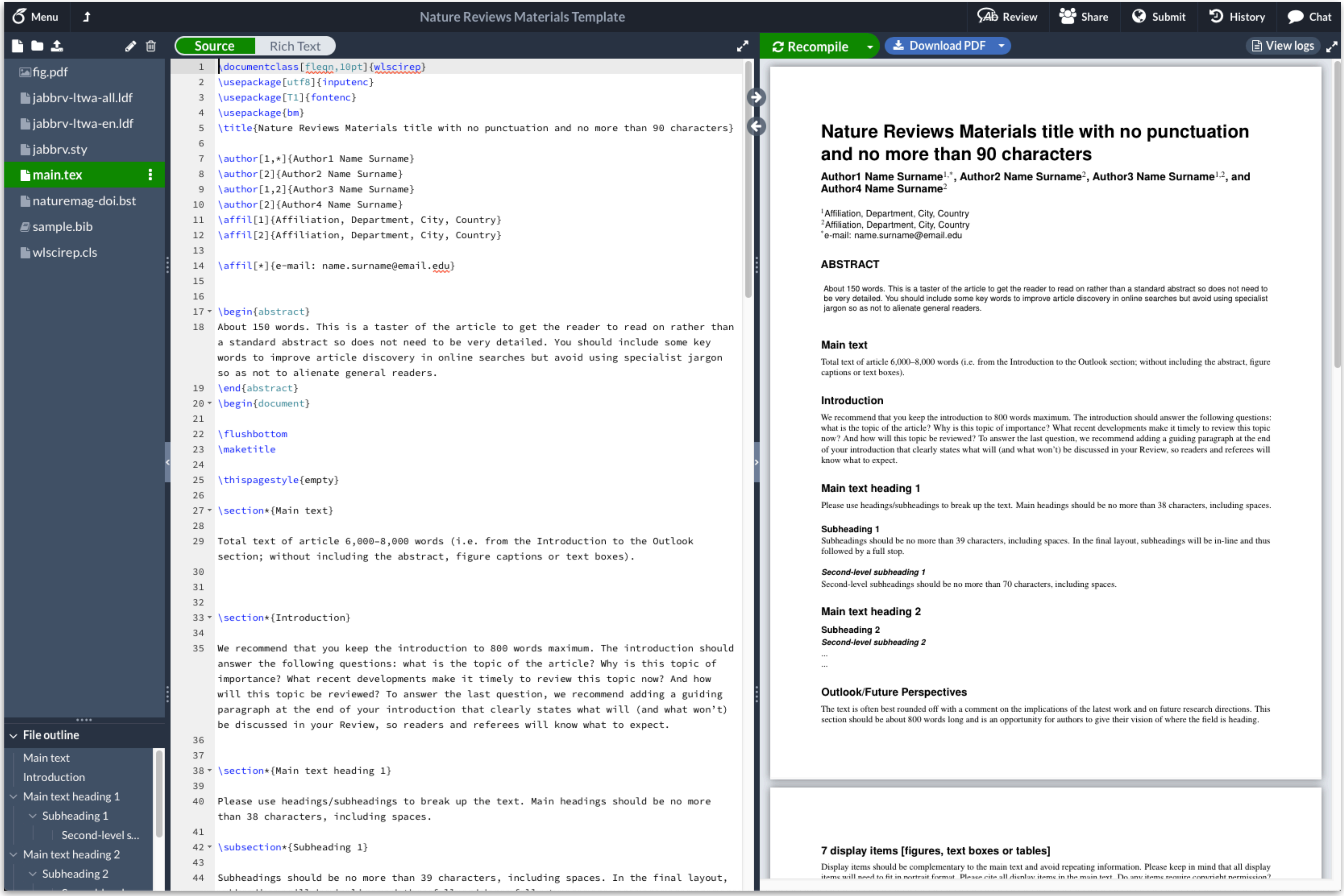
BestPaperEver_v1.1_jcw.doc →
← BestPaperEver_v.1.1_jcw_ar.doc
← BestPaperEver_v.1.2_jcw_mf.doc
BestPaperEver_v2_jcw.docx →
← BestPaperEver_v.2.1_md.docx
BestPaperEver_v3.1_jcw.docx →
← BestPaperEver_v.3.2_jcw_ar.docx
← BestPaperEver_v.3.3_jcw_mf.docx
BestPaperEver_v4_ready.docx

Bioinformatics ▷ Basics

Overleaf
by writeLaTeX

Overleaf is a collaborative cloud-based LaTeX editor used for writing, editing and publishing scientific documents. It partners with a wide range of scientific publishers to provide official journal LaTeX templates, and direct submission links.

Bioinformatics ▷ Basics



```
\documentclass[fleqn,10pt]{wlscirep}
\usepackage[utf8]{inputenc}
\usepackage[T1]{fontenc}
\usepackage{bm}
\title{Nature Reviews Materials title with no punctuation and no more than 90 characters}

\author[1,*]{Author1 Name Surname}
\author[2]{Author2 Name Surname}
\author[1,2]{Author3 Name Surname}
\author[2]{Author4 Name Surname}
\affil[1]{Affiliation, Department, City, Country}
\affil[2]{Affiliation, Department, City, Country}

\affil[*]{e-mail: name.surname@email.edu}


\begin{abstract}
About 150 words. This is a taster of the article to get the reader to read on rather than a standard abstract so does not need to be very detailed. You should include some key words to improve article discovery in online searches but avoid using specialist jargon so as not to alienate general readers.
\end{abstract}
\begin{document}

\flushbottom
\maketitle

\thispagestyle{empty}

\section*{Main text}

Total text of article 6,000-8,000 words (i.e. from the Introduction to the Outlook section; without including the abstract, figure captions or text boxes).



\section*{Introduction}

We recommend that you keep the introduction to 800 words maximum. The introduction should answer the following questions: what is the topic of the article? Why is this topic of importance? What recent developments make it timely to review this topic now? And how will this topic be reviewed? To answer the last question, we recommend adding a guiding paragraph at the end of your introduction that clearly states what will (and what won't) be discussed in your Review, so readers and referees will know what to expect.


\section*{Main text heading 1}

Please use headings/subheadings to break up the text. Main headings should be no more than 38 characters, including spaces.

\subsection*{Subheading 1}

Subheadings should be no more than 39 characters, including spaces. In the final layout,
```

Nature Reviews Materials title with no punctuation and no more than 90 characters

Author1 Name Surname[1,*], Author2 Name Surname[2], Author3 Name Surname[1,2], and Author4 Name Surname[2]

[1]Affiliation, Department, City, Country
[2]Affiliation, Department, City, Country
[*]e-mail: name.surname@email.edu

ABSTRACT

About 150 words. This is a taster of the article to get the reader to read on rather than a standard abstract so does not need to be very detailed. You should include some key words to improve article discovery in online searches but avoid using specialist jargon so as not to alienate general readers.

Main text

Total text of article 6,000–8,000 words (i.e. from the Introduction to the Outlook section; without including the abstract, figure captions or text boxes).

Introduction

We recommend that you keep the introduction to 800 words maximum. The introduction should answer the following questions: what is the topic of the article? Why is this topic of importance? What recent developments make it timely to review this topic now? And how will this topic be reviewed? To answer the last question, we recommend adding a guiding paragraph at the end of your introduction that clearly states what will (and what won't) be discussed in your Review, so readers and referees will know what to expect.

Main text heading 1

Please use headings/subheadings to break up the text. Main headings should be no more than 38 characters, including spaces.

Subheading 1

Subheadings should be no more than 39 characters, including spaces. In the final layout, subheadings will be in-line and thus followed by a full stop.

Second-level subheading 1

Second-level subheadings should be no more than 70 characters, including spaces.

Main text heading 2

Subheading 2

Second-level subheading 2

...

...

Outlook/Future Perspectives

The text is often best rounded off with a comment on the implications of the latest work and on future research directions. This section should be about 800 words long and is an opportunity for authors to give their vision of where the field is heading.

7 display items [figures, text boxes or tables]

Display items should be complementary to the main text and avoid repeating information. Please keep in mind that all display items will need to fit in portrait format. Please cite all display items in the main text. Do any items require copyright permission?

Interactive Shared Solutions

Google Docs

EduPad

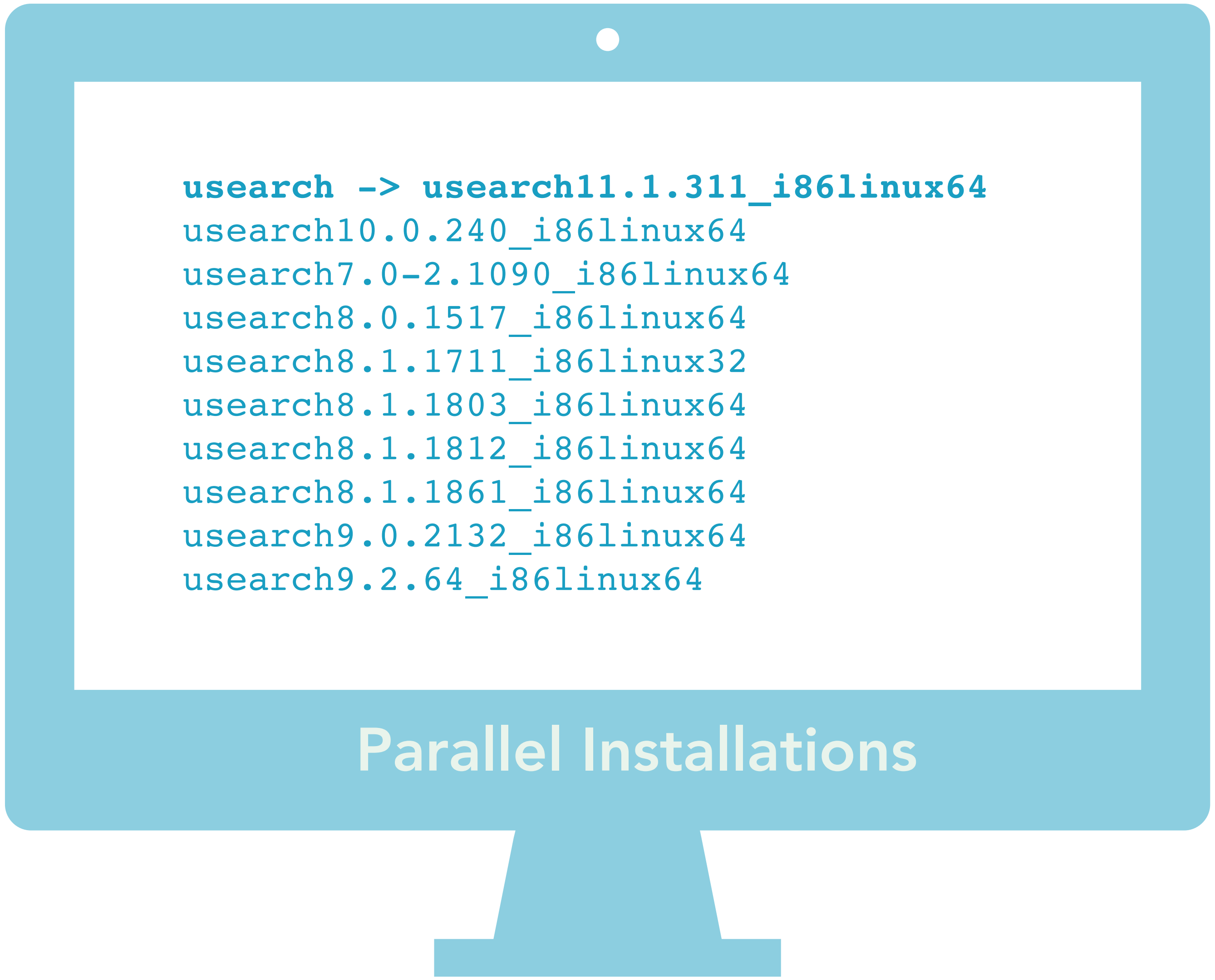
Etherpad

Overleaf

Manubot

Software Management

Bioinformatics ▷ Basics



```
usearch -> usearch11.1.311_i86linux64
usearch10.0.240_i86linux64
usearch7.0-2.1090_i86linux64
usearch8.0.1517_i86linux64
usearch8.1.1711_i86linux32
usearch8.1.1803_i86linux64
usearch8.1.1812_i86linux64
usearch8.1.1861_i86linux64
usearch9.0.2132_i86linux64
usearch9.2.64_i86linux64
```

Parallel Installations

Version Control

Application **A**
Requirements: Python 2.7

Application **B**
Requirements: Python >3.0

Bioinformatics ▷ Basics

CONDA

Package, ***dependency and environment management*** *for any language—Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN. Conda as a package manager helps you find and install packages. If you need a package that requires a different version of Python, you do not need to switch to a different environment manager, because conda is also an environment manager. With just a few commands, you can set up a totally separate environment to run that different version of Python, while continuing to run your usual version of Python in your normal environment.*

Bioinformatics ▷ Basics

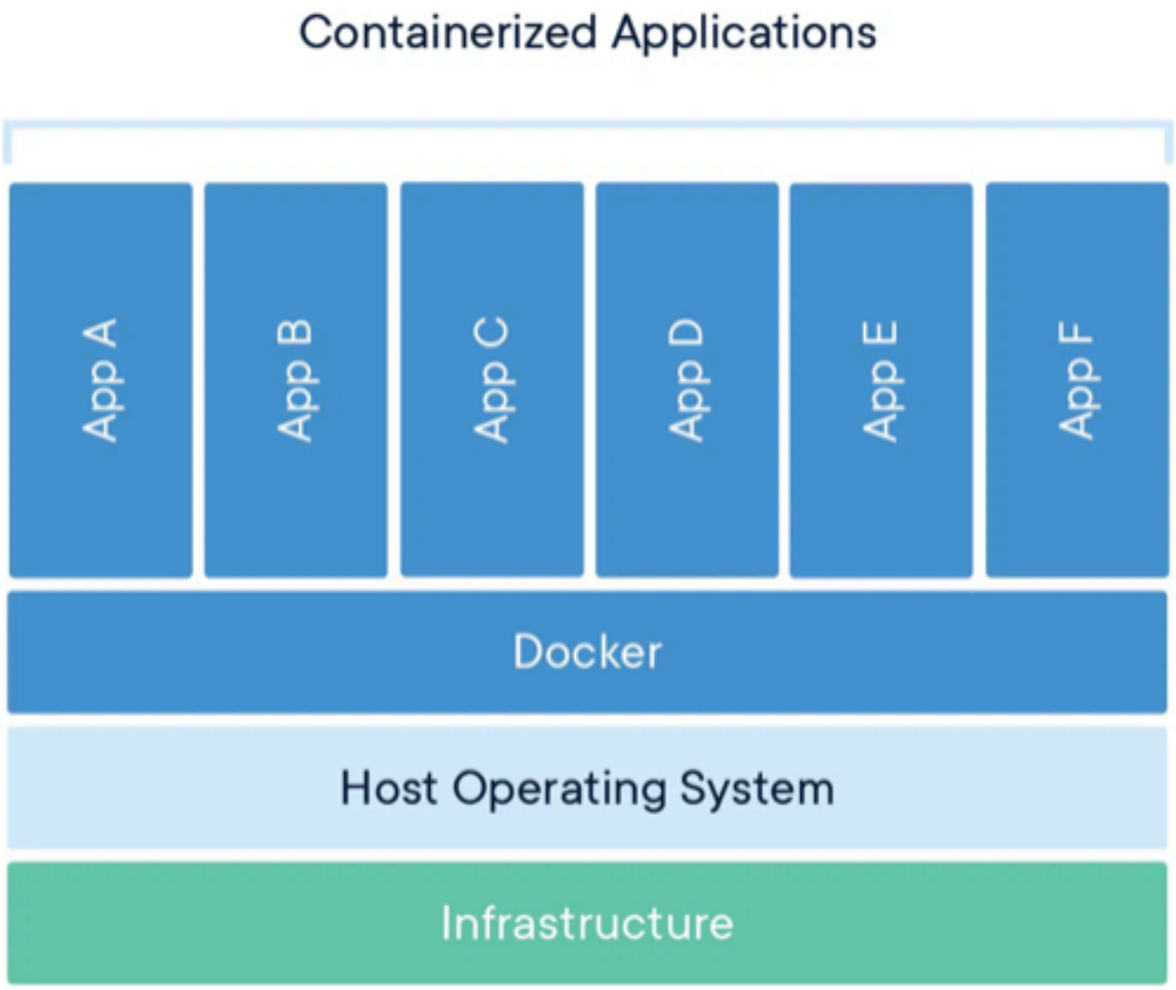
BIOCONDA®

*Bioconda is a **channel** for the conda package manager specialising in bioinformatics software. The conda package manager makes installing software a vastly more streamlined process. Conda is a combination of other package managers you may have encountered, such as pip, CPAN, CRAN, Bioconductor, apt-get, and home-brew. Conda is both language- and OS-agnostic, and can be used to install C/C++, Fortran, Go, R, Python, Java etc programs on Linux, Mac OSX, and Windows.*

Bioinformatics ▷ Basics

```
python --version
# Python 2.7.15
bwa
# -bash: bwa: command not found
blast -help
# -bash: blast: command not found

conda info --envs
source activate aligners
conda info --envs
python --version
# Python 3.6.7
bwa
blastn -help
```



Singularity (Apptainer) is a free and open source computer program that performs operating system level virtualisation, also known as containerisation. One of the main uses of Singularity is to bring containers and reproducibility to the world of scientific computing and high performance computing.

Singularity (Apptainer) is a free and open source computer program that performs operating system level virtualisation, also known as containerisation. One of the main uses of Singularity is to bring containers and reproducibility to the world of scientific computing and high performance computing.