

# Python Programming for Life Scientists

Hanna M. Schilbert, Katharina Sielemann, Bianca Frommer,  
Daniela Holtgräwe

Genetics & Genomics of Plants and Computational Biology, CeBiTec, Bielefeld University



# Introduction of participants

- Name
- Study program (semester)
- Bioinformatics experiences
- Research interests
- motivation?
- ...



## **Boas Pucker –**

one and likely most important former course developer

- Since 10.2021: Prof. Pucker: Plant Biotechnology and Bioinformatics, TU Braunschweig
- 2020-2021: Evolution and Diversity, Department of Plant Sciences, University of Cambridge
- 2019-2020: Molekulare Genetik und Physiologie der Pflanzen, Ruhr-Universität Bochum
- 2016-2019: Genetik und Genomik der Pflanzen, Center for Biotechnology (CeBiTec) & Fakultät für Biologie, Universität Bielefeld
- A lot of bioinformatics experiences by now
- Wants to teach python so that more biologists can benefit from using it for their own research
- Specialised plant metabolism

# Hanna M. Schilbert

## 2014-2016: **Molecular Biology**

Functional and comparative genomics, crops, flavonoid biosynthesis

## 2016-2018: **Genome-based Systems Biology**

Bioinformatics, phylogenetic analyses, nutrigenetics/genomics

## 2018: Visiting student in the Antonio Vidal-Puig lab, Clinical Biochemistry, Cambridge

Molecular nutrition and metabolism, differential gene expression analyses

## Since April 2019: **PhD student in Genetics and Genomics of Plants**

Dry lab: Applied bioinformatics in crops (rapeseed), genomics, transcriptomics, phylogeny

Wet lab: flavonoid biosynthesis, complementation experiments

## September 2021 - February 2022: Visiting PhD student in the Glover lab, Evolution and Development, Cambridge

Evolution of transcriptional flavonol biosynthesis regulators in plants



# Katharina Sielemann

## 2015-2017: **Molecular Biology**

- Bachelor thesis: “Bioinformatic identification and molecular biological validation of structural genome differences in *Arabidopsis thaliana* accessions”

## 2017-2019: **Genome-based Systems Biology**

- Master thesis: “Transcriptomics of human genetic cardiomyopathies”

## Since January 2020: **PhD student** in Genetics and Genomics of Plants

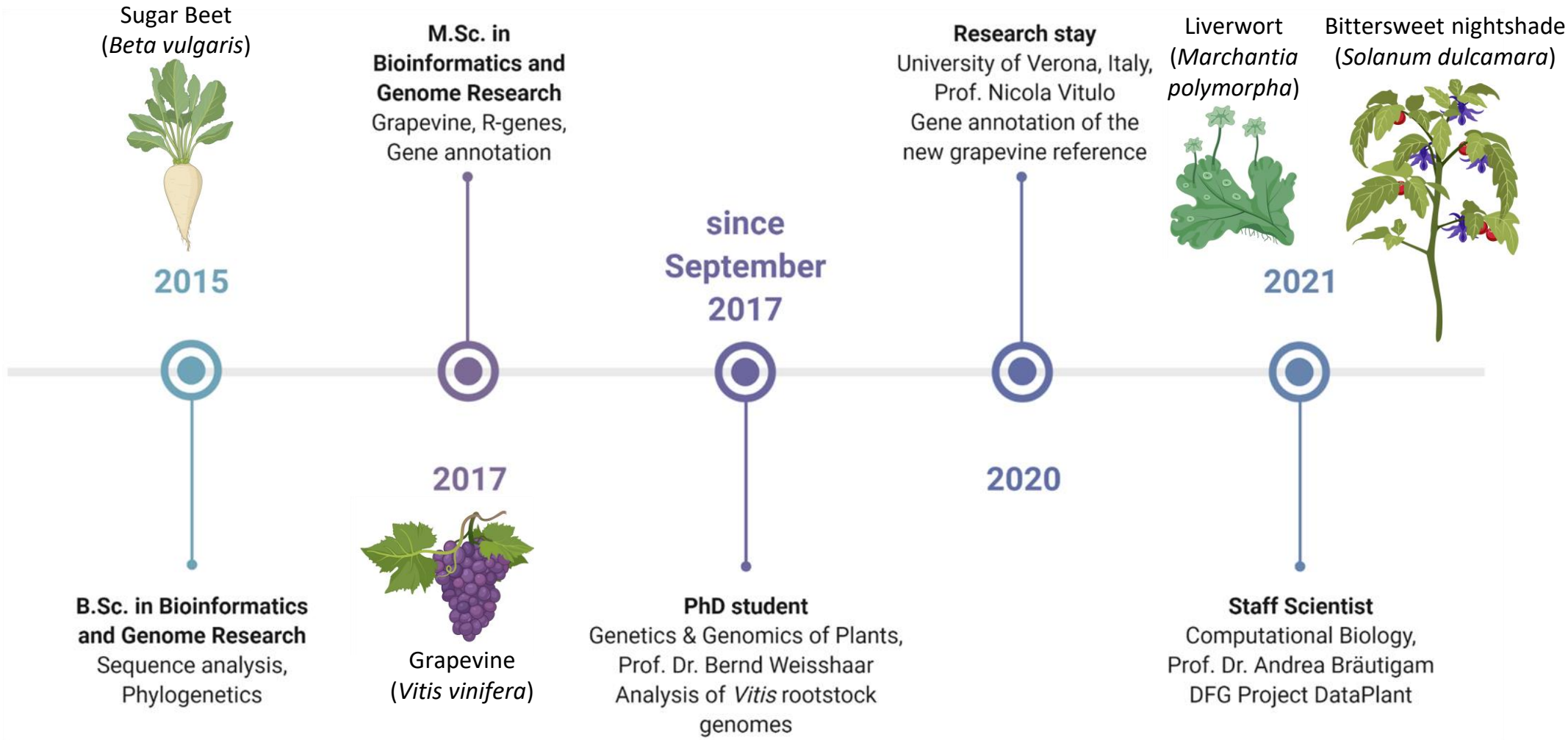
- Computational Pangenomics in Plants (sugar beet)
  - Genome assemblies, annotation, comparative genomics

## January-May 2022: **Research stay** at SFU, Vancouver

- Prediction of isoforms in long read transcriptomics data



# Bianca Frommer



# Dr. Daniela Holtgräwe

- Since 2005 Bielefeld University, Faculty of Biology, **Teaching & Research**
- Chair of Genetics and Genomics of Plants / Prof. Dr. Bernd Weisshaar
- **Research:** leading subgroup „**Crop genomics**“
- **Main species:**

sugar beet (*Beta vulgaris*)



rapeseed (*Brassica napus*)



grapevine  
(*Vitis vinifera*)



....arabidopsis and crop wild relatives

- **Teaching:** e.g. Molekularbiologische Experimente für den Schulunterricht; Grundlagen der Genomforschung 2; Projekt- und Forschungsmodule, Betreuung von Bachelor, Master und Doktorarbeiten

# What we are doing in the subgroup “Crop Genomics”

## Functional genomics

- We study how genes and intergenic regions of genomes contribute to different processes
- We use **genomic data** to study **gene expression** and protein function on a genome-wide scale
- We aim to determine how components (genes/proteins/metabolites..) work together and produce a particular **phenotype**

## Topics (examples)

- Identification of disease resistant genes,
- Identification of genes (regions) crucial for metabolite and nutrition accumulation,
- Unrevealing genotype dependent differences in crop yield and fruit development

## Main overall goals of the projects

- Improving and speeding up crop breeding
- Gain general knowledge of processes in crop plants

# What we are doing in the subgroup “Computational Biology”

## Transcriptomics

- We analyse **transcriptional responses** and **transcriptomic data**
- We identify functions of **transcription factors** (TF) via **network analyses** and TF binding sites via **DAP-Seq**
- We study **photosynthesis**, especially **C4 photosynthesis** and **crassulacean acid metabolism (CAM)**

## Topics (examples)

- Studying CAM plants like *Kalanchoe*
- Network analysis of *A. thaliana* and C3 and C4 crop plants (tomato, barley and many other crop plants)
- Generation of *Marchantia polymorpha* expression atlas and genome sequence & gene annotation of ecotype BoGa

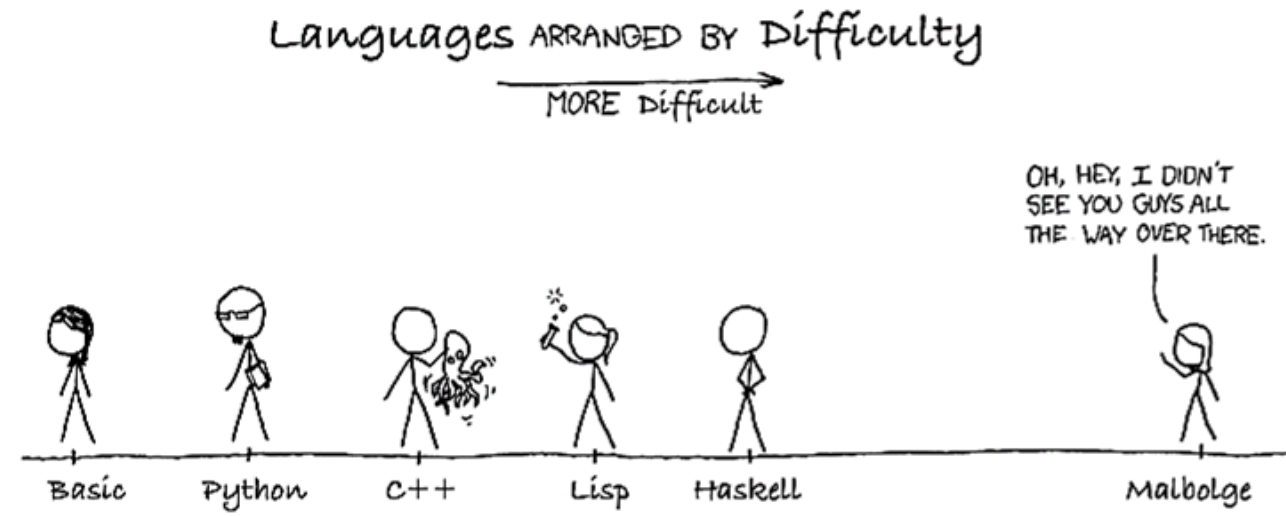
## Main overall goals of the projects

- Gain knowledge of specific interaction and photosynthesis pathways
- Identification the function of transcriptional regulators
- Application of machine learning algorithms to plant transcriptomics





- Open source Programming language
  - Well readable, concise programming style
  - Simple syntax (e.g. with similarities to English)
- Designed by Guido van Rossum and published in 1991
- Latest version: python 3.10.2 (February 2022)
- Platform independent
- Pre installed on many PCs
- Can be written in a text editor (.py file), but can be used in an Integrated Development Environment (Pycharm, Netbeans, Eclipse, Visual Studio)
  - Only run 'python my\_file.py' in a terminal to run a script



## 205062 Python Programming for Life Scientists (S) (WiSe 2021/2022)

### Inhalt, Kommentar

Scientists in molecular biology and genome research are confronted with a very broad range of numerous bioinformatic challenges. Presentations of example scripts and general concepts in this seminar will enable participants to develop Python scripts to address their own problems. Practical exercises are performed in 205063.

### Lehrende

Frau Dr. Daniela Holtgräwe  
Frau Prof. Dr. Andrea Bräutigam  
Herr Prof. Dr. Bernd Weisshaar  
*Bianca Frommer, Hanna Schilbert, Katharina Sielemann*

### Termine ( Kalendersicht )

Rhythmus	Tag	Uhrzeit	Format / Ort	Zeitraum	
<b>Block</b>		9-17	VOR ORT & ONLINE G1-106	21.03.2022-01.04.2022	

### Klausuren

- Keine gefunden


### Fachzuordnungen

Modul	Veranstaltung	 Leistungen	
20-EB_10 <i>Ergänzungsmodul Biologie</i>	<u>2 std. Seminar 1</u>	Studienleistung	Studieninformation
20-EB_5 <i>Ergänzungsmodul Biologie</i>	<u>2 std. Seminar 1</u>	Studienleistung	Studieninformation
20-GBSB-EM <i>Ergänzungsmodul</i>	<u>Ergänzungsveranstaltungen</u>	<u>unbenotete Prüfungsleistung</u>	Studieninformation
20-GBSB-EM_a <i>Ergänzungsmodul</i>	<u>Ergänzungsmodul</u>	<u>unbenotete Prüfungsleistung</u>	Studieninformation



## 205063 Python Programming for Life Scientists (Ü) (WiSe 2021/2022)

### Fachzuordnungen

Modul	Veranstaltung	 Leistungen	
20-EB_10 <i>Ergänzungsmodul Biologie</i>	<u>3 std. Übung 1</u>	Studienleistung	Studieninformation
20-EB_5 <i>Ergänzungsmodul Biologie</i>	<u>3 std. Übung 1</u>	Studienleistung	Studieninformation
20-GBSB-EM <i>Ergänzungsmodul</i>	<u>Ergänzungsveranstaltungen</u>	<u>unbenotete Prüfungsleistung</u>	Studieninformation
20-GBSB-EM_a <i>Ergänzungsmodul</i>	<u>Ergänzungsmodul</u>	<u>unbenotete Prüfungsleistung</u>	Studieninformation

- **TWO** week crash course for **BEGINNERS**
- **„Studienleistungen“ – study requirements:**
  1. Active participation
  2. Documentation
  3. Project work and presentation
- Additional information in the syllabus = organizational framework
- All important documents can be found in the corresponding **“Lernraum”**

### Quicklinks

In Stundenplan setzen  
Fragen oder Korrekturen?  
Termine  
Klausuren  
Fachzuordnungen  
Teilnehmer\*innen: 16  
**Teilnahmebegrenzung: 12**  
E-Mailverteiler

# ‘Studienleistungen’ – study requirements

## 1. Active participation

- Participate! Only those who participate and try to solve the exercises can learn something!

## 2. Documentation of the learning progress

- Solutions of the exercises + your own notes
- A simple text file is possible

## 3. Own programming project + presentation (10-15 min)

- Allocation of a own programming project at the end of the course; work on the project at home; own ideas for a project can be suggested

📅 **submission until 07.04.2022 (Thursday) 1:00 pm**

- **Presentation of the project: 08.04.2022 (Friday) 10:00 am, online via Zoom**
- Feel free to ask any questions via email: [hschilbe@cebitec.uni-bielefeld.de](mailto:hschilbe@cebitec.uni-bielefeld.de), [ksielemann@cebitec.uni-bielefeld.de](mailto:ksielemann@cebitec.uni-bielefeld.de), [frommer@cebitec.uni-bielefeld.de](mailto:frommer@cebitec.uni-bielefeld.de), [dholtgra@cebitec.uni-bielefeld.de](mailto:dholtgra@cebitec.uni-bielefeld.de)

# Learning goals

1. Be able to understand what code is
2. Get to know the basics of Python as one of THE programming languages in the life sciences
3. Learn to use Jupyter Notebooks
4. Be able to find solutions for error messages during programming
5. Write and understand your own scripts
6. Gain experience as to whether programming could play a bigger role in the further course of one's own (scientific) career



# Contents

- Introduction: Software background (Jupyter notebooks, Python)
- Basics of the Python syntax
  - Data types, if/else operators, loops, functions



```
def print_hello_world():  
    """function prints 'hello world!'"""  
    print("hello world!")  
print_hello_world() #calling function
```

```
hello world!
```

# Contents

- Parsing files, filtering, exporting, converting

Genbank (.gb)

```
ncRNA              11 1
Frameshifted Genes  31 35
##Genome-Annotation-Data-END##
COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source            1..5879002
                        /organism="Xanthomonas campestris pv. campestris"
                        /mol_type="genomic DNA"
                        /strain="B100"
                        /db_xref="taxon:340"
                        /pathovar="campestris"
     gene              1..1329
                        /locus_tag="XCCB100_RS00005"
                        /old_locus_tag="xcc-b100_0001"
                        /old_locus_tag="xccb100_0001"
     CDS               1..1329
                        /locus_tag="XCCB100_RS00005"
                        /old_locus_tag="xcc-b100_0001"
                        /old_locus_tag="xccb100_0001"
                        /inference="EXISTENCE: similar to AA
                        sequence:SwissProt:Q8PRG2.1"
                        /note="Derived by automated computational analysis using
                        gene prediction method: Protein Homology."
                        /codon_start=1
                        /transl_table=11
                        /product="chromosomal replication initiator protein DnaK"
                        /protein_id="WP_011035259.1"
                        /db_xref="GI:499345728"
                        /translation="MDAMPRLERLEAEFPPEVHTLKPQLAEDRGDSIVLYAPNAF
                        IVEQVREYLPRIREL LAYFAGNGEVALAVGSRPAPPEPLPAPQAVASAPAAAPIVPF
                        AGNLDSHYTFANPVEGRSNQLGLAAAIQAQKPCDRAHNPILLVYSTGLGKTHLPFAA
                        GNALRQANPAKAWYLRSEQFFSARIRALQKAMQKQKQFQIIDLIDDIQFFAGK
                        DRTQEEFFHTNALPDGRQIILTCDRYPREVEGLEPRLKSLRAAGLSVAIDPPDET
                        RAAITLAKARERGAETPDQVAFI AKKMSNVRLLEGALNTLVARANFTGRSITVEFA
                        QETLRDLRAQQQAI GIPNIQKTVADYVGLQMDLLSKRRTSLARPRQVAMALAKEL
                        TEHSLPEIGDAFAGRHITVLHACQIRITLREADCKLRDEKLLRKLS*"
     gene              1605..2705
                        /locus_tag="XCCB100_RS00010"
                        /old_locus_tag="xcc-b100_0002"
                        /old_locus_tag="xccb100_0002"
     CDS               1605..2705
                        /locus_tag="XCCB100_RS00010"
                        /old_locus_tag="xcc-b100_0002"
                        /old_locus_tag="xccb100_0002"
                        /EC_number="2.7.7.7"
                        /inference="EXISTENCE: similar to AA
                        sequence:RefSeq:WP_006451791.1"
                        /note="binds the polymerase to DNA and acts as a sliding
                        clamp; Derived by automated computational analysis using
                        gene prediction method: Protein Homology."
                        /codon_start=1
                        /transl_table=11
```



FASTA (.fa/.fas/.fasta)

```
>xccb100_0001 chromosomal replication initiation protein
MDAMPRLERLEAEFPPEVHTLKPQLAEDRGDSIVLYAPNAFIVEQVREYLPRIREL
LAYFAGNGEVALAVGSRPAPPEPLPAPQAVASAPAAAPIVPFAGNLDSHYTFANP
VEGRSNQLGLAAAIQAQKPCDRAHNPILLVYSTGLGKTHLPFAAGNALRQANPAK
AWYLRSEQFFSARIRALQKAMQKQKQFQIIDLIDDIQFFAGKDRTEQEEFFHTNA
LPDGRQIILTCDRYPREVEGLEPRLKSLRAAGLSVAIDPPDETAAITLAKARERGA
ETPDQVAFIAKKMSNVRLLEGALNTLVARANFTGRSITVEFAQETLRDLRAQQQA
IGIPNIQKTVADYVGLQMDLLSKRRTSLARPRQVAMALAKELTEHSLPEIGDAFA
GRHITVLHACQIRITLREADCKLRDEKLLRKLS*

>xccb100_0002 DNA polymerase III subunit beta
NRFTLQREAFKPLAQVWVVERRQTLVLANLLVQVNNQSLSTGTDLEVENISRTWVE
DAQDGETTIPARKLPDILRALPGDSRYTVSQTGKVTVQAGRSHPTLATLPANFP
SVDEVEATERVAVPEAGLKELMERTAFAMAQDQVRYVNGLLPDLRDGLRCVATDQ
HRLALCETELEKSGSAKRQIIVPRKQVTELLRLLEAARDVLEELGRSHIRVKGQD
VFTSKLIDGRFPDVEAVIPIGADREVKVDREARSLQRAAILSHKRYGVRVVEVSP
GLKISAHNPQEEEAQEEIEADTKVDDLAIFGWNNYLLDALSAIRDHVVTQLRDANS
SALVREASSEKSRHVWNPRL*

>xccb100_0003 recombination protein F
NSTADHVCAPSAQGLQGQADRSMMHVARLSIHLRRFEEVFPASTLLNLTGZNGAG
KTSVLIALHVMVYGRSFRGVRVDGLTRQGGQDLTFVEMIRAGDSTERTRAGLRHS
GGQETGRLDGEDIWAQGLCSLCAALAVVTFFEPQSHWLSGGGEPRRRFLDNLG
FHVPEPFLALHRRYARALQDRNALKQGAQPMQLDANDHMLAESGETLSRRLQVLR
LQERLVPWATAIAPS LGLSALTFAPIQWRHVEVSLADALLARERDRQNGVTSQ
GPHRADWAPLFQALPGKDALSRGQAKLTALACLAAQAEFPAHERGEMPLMALDGL
SELDHMHQAVVIQRLASAPAGVLITATELPPGLADAGKTLHRFVHEHQLVPQ
PLPTDPRIIA*

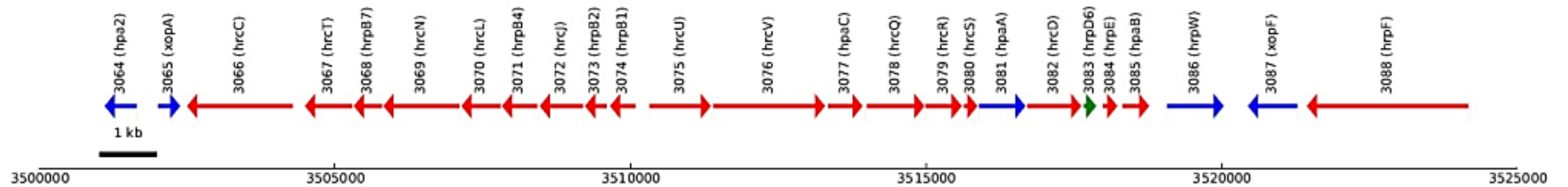
>xccb100_0004 DNA gyrase subunit B
MTDQGTTPPTPNCITYDSKCTVLRLGLAVRKRPCHYIGDVMGDTGLHMHVPEV
VDNSIDEALAGADDIVAKIHKQSVHAYSONGRCQPYDIDKEGVSAEVELTVLH
ACGKFDONSYSVSGGLHGVQSVVNALSEHLNLDINRGGHHYQGEVALGEPQYPL
KQLEASTKRGTLPFKPAVEIFSDVEFHYDILARRRELSPFNSGVKIALIDERGE
CRDRDHFVEGGTSFVEHLAQLKTPLHMWISVTGEHNGIVVQVALQTDAYQETNYC
FTNNIPQKGGTHLAGFRGALT RVLSMYTCQNGTAKQAKITLTGQDNREGNA
VLSVKKVDPSPSSQTKKLVSSQVRPAVENAFGARLQEFLEQENPEAKALTCK
TVDAARAREAAKAROLDTRKRGALDIAGLPCKLADQEKDPALSELPIVEG
DSAGSGAKGGRNRKQAVLPARGKILNVERARPDRPLASDQVCTLTITALTGTC
GRDEYNPQKLRYHRIITLNTDAGVDSHIRTLLTFFYRQMPLEIERGVYIYG
LPPYKLKQKQSELVLKDDAALNAYLASAVEGAALIPASDEPPICTGEALEKLL
LLPAGAKEAJARNHRYDPALITLIDLPLOVQLQAEQDVHPTLDALQAVLNRTG
LTARYHLRFDPATDSAAASLVSVKHNKEEFTQVLPNGAFESGELIRLREVALA
HGLVREGAQILRGNKSHPTISFAQAQWILLEAKRCRQVQRKGLGEMNAEQHWETT
VNPOTRRLQVRIEDAAVAADQIFSTLNGOVVEPRDFIEDNALKVSNDI*

>xccb100_0005 putative membrane protease
NSAIVPPSPAPVSVPPPSLSAVALGFCIDLLIATGLLLLSVAGFVAVGFLRSM
GEVQAVIAGGSPSPAAIMAAIQGPQVWQLLIALVSTATPAVLVFWRRHATPAE
QATSHALRRSTWGTATAAVAVGFMNLGNLSVLASALGICKPVPTNLPLMEET
AKQNPLALVWVAIAIPAYEELLFRVLFRLLAAGRPMGLGVLLSSLTALVHEV
PGTSGNVAIAQLMLVYGMGAAPFAMLYRITGLWAPILAHGNNATLAALYFFLG*

>xccb100_0006 putative exported peptidase/protease
MKVRLILIVAVLALTACATTISPTGRRQVVGQVTDQGLDKLGAEFATKAKEK
VSTGKQNAVYQCVNALVQQLPQWRERHETALFYDREANAFALPGKQVGVNTG
TFVAKTQQLAAVLGHEIGHVISHHHEERITRLGAQTGLGIIGALGAAYVGG
ASAVNQVSGMTAQTVPLPCSRTESEADVVQRLHAGAGDPDPAQVSLQNMMA
SGNRQPMLSITHPDANRIRELQDVNALQPVYQQAQGRVPRIC*
```

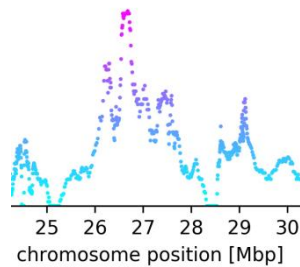
# Contents

- Create graphics
  - Learn to use modules - matplotlib



# Contents

- Statistics for biological data
- Bioinformatic solutions to real experimental data



- Choose and prepare your own Python programming project
- Feel free to 'bring' your own programming challenges!



# Participants need a Google ('colab') account!

- Use your existing account or generate a new one!
- Save your password!
- We will need access to Google Drive and Google Colab
- Feel free to 'bring' your own programming challenges!

## What is Colaboratory?

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier.

