Scientific Programming Practical 1 (QCB)

Introduction

Luca Bianco - Academic Year 2020-21
luca.bianco@fmach.it

Outline

- ❖ Personal introduction
- ❖ Introduction to the practical
- ❖ Hands-on practical

About me

Computer Science
 Ph.D. at the University of Verona, Italy, with thesis on Simulation of Biological Systems

Research Fellow at Cranfield University - UK
 Three years at Cranfield University working at proteomics projects (GAPP, MRMaid, X-Tracker...)
 Module manager and lecturer in several courses of the MSc in Bioinformatics

Bioinformatician at IASMA – FEM
 Currently bioinformatician in the Computational Biology Group at Istituto Agrario di San Michele all'Adige – Fondazione Edmund Mach, Trento, Italy

Collaborator uniTN - CiBio
I ran the Scientific Programming Lab for QCB for the last four years

Fondazione Edmund Mach

FEM – San Michele, Trento - Italy

Agricultural Institute

Research and Innovation Centre

Genomics, transcriptomics, metabolomics wet labs on fruits (apple, grape, small fruits,…)

Bioinformatics and computational biology

Bioinformatics @FEM (UBC)

❖ Genomics
  ➢ Assembly and annotation of complex genomes (plants, insects, etc.)
  ➢ Development of SNP Chips for genetic screening
  ➢ Resequencing of genomes / Variant discovery

❖ Metagenomics
  ➢ Targeted metagenomic data
  ➢ *Feature selection* algorithms
  ➢ Algorithms for strain-level identification from un-targeted metagenomics

❖ Transcriptomics
  ➢ RNA-seq data analysis, gene and pathway enrichment
  ➢ Data integration and compilation of expression atlases

❖ Metabolomics
  ➢ Data analysis pipelines for targeted and untargeted data
  ➢ Methods for MS imaging

❖ Statistical data analysis
  ➢ Integration of –omic data and analysis of correlation networks

Bix @FEM - Examples

Genome assembly
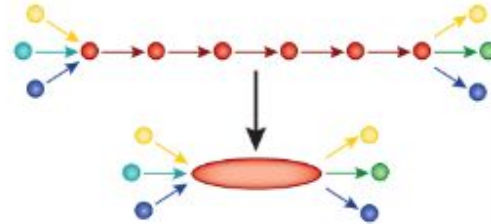

1. Fragment DNA and sequence

2. Find overlaps between reads

…AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT

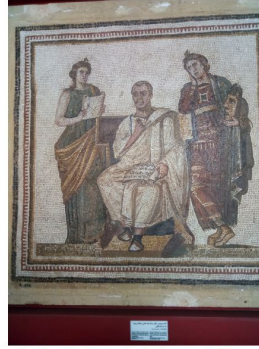3. Assemble overlaps into contigs

4. Assemble contigs into scaffolds
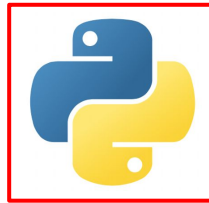
In a nutshell… (Tunis' version…)

Reads

Assembled genome

[Virgil and the Muses, Bardo Museum, Tunis]

[from M. Baker, Nature Methods, 2014]

Bix @FEM - Examples

Genome assembly of DH of Pear and Apple

Multiple sources of input data:

  Illumina: ~ 100x PE information (mate pairs - in the past)

  Pacific Biosciences > 50x

  Bionano optical maps: ~ 600x

  Hi-C: illumina sequencing of chromosome conformation capture libraries

  Genetic maps: genetic information coming from mapping populations

Output result (target):

  Chromosome scale assembly

  Ideally, we want to arrange all the sequences produced in N (= number of chromosomes) sequences

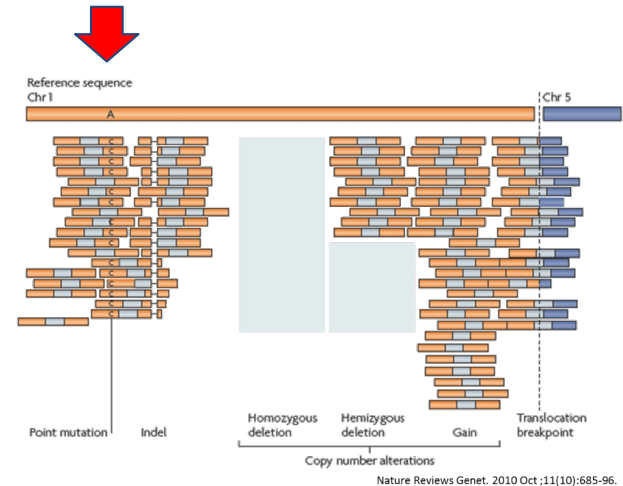[Daccord et al, Nature Genetics, 49, 2017; Linsmith et al., GigaScience, 2019; Marrano et al., GigaScience, 2020]
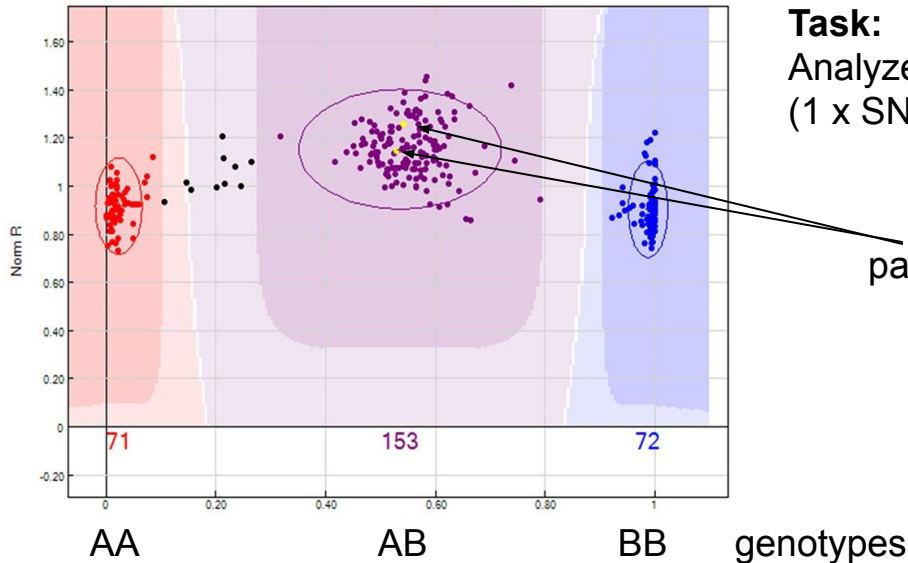
Bix @FEM - Examples

SNP-Chips development for GWAS

20K SNP Illumina Infinium II Array (reseq of 16 Apple cultivars, Illumina 30x)

487K SNP Affymetrix Axiom Array (reseq of 63 Apple cultivars, Illumina 20-30x)

600K SNP Affymetrix Axiom Array Walnut (reseq. 18 cultivars, Illumina 80x)

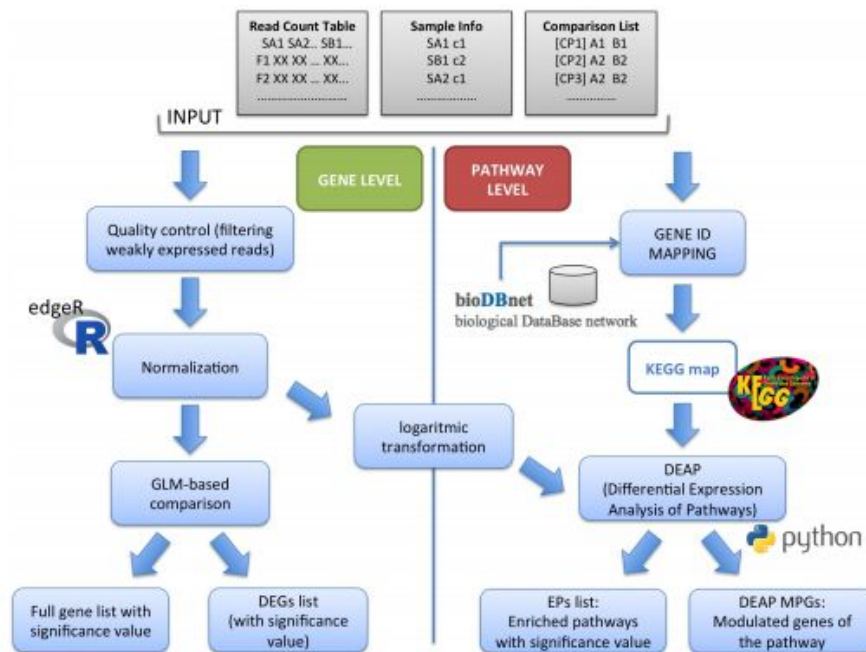70K SNP Affymetrix Axiom Array Pear (reseq. 55 cultivars, Illumina ~5x)

1. Reads alignment and filtering

2. SNP calling

3. Identification of most reliable SNPs

4. Selection of (20K) 487K target SNPs

Nature Reviews Genet. 2010 Oct ;11(10):685-96.

Several Terabytes of data produced!!!!

[Bianco et al., PloS One, 2014; Bianco et al., the Plant Journal 2016; Marrano et al., the Plant Journal 2018; Montanari et. al, BMC Genomics, 2019]

Bix @FEM - Examples

SNP-Chips development for GWAS

20K SNP Illumina Infinium II Array (reseq of 16 Apple cultivars, Illumina 30x)

487K SNP Affymetrix Axiom Array (reseq of 63 Apple cultivars, Illumina 20-30x)
600K SNP Affymetrix Axiom Array Walnut (reseq. 18 cultivars, Illumina 80x)
70K SNP Affymetrix Axiom Array Pear (reseq. 55 cultivars, Illumina ~5x)

Task:
Analyze 500,000 of these...
(1 x SNP)

parents

AA   AB   BB   genotypes

ASSIsT
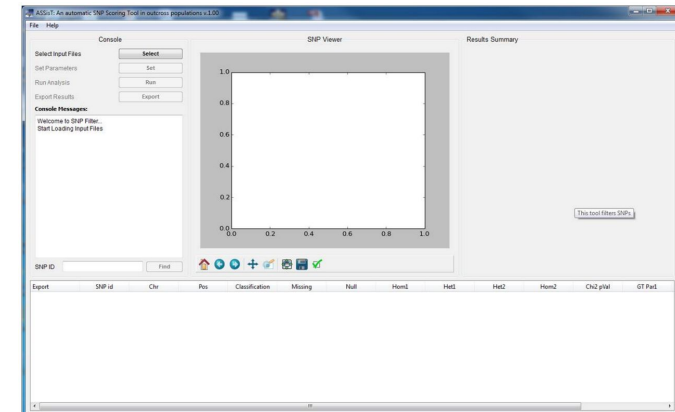
[Di Guardo et al., Bioinformatics, 2015]

Bix @FEM - Examples

RNAseq data analysis with Pathway Inspector

[Bianco et al., Bioinformatics, 2017]

https://pathwayinspector.fmach.it

Bix @FEM - Examples

RNAseq data analysis with Pathway Inspector

https://pathwayinspector.fmach.it

Bix @FEM - Examples

**Sequencing and assemblying of Sars-Cov-2 samples
from the Province of Trento (sponsored by Fondazione VRT)**

First 72 samples
assembled...
... 240 more to go!

Artic networks V3 protocol
amplification & sequencing of 98,
~400bps long sequences
overlapping by ~20bps

basecalling &
de-multiplexing
with guppy

"quasi"-real time
sequencing run
checking with
rampart

RAMPART

ARTICnetwork

post-processing and
consensus calling with
artic networks pipeline
and IRMA assembler

Nextstrain

down-stream analysis:
multiple-sequence alignment,
nextstrain, deepVariant,...
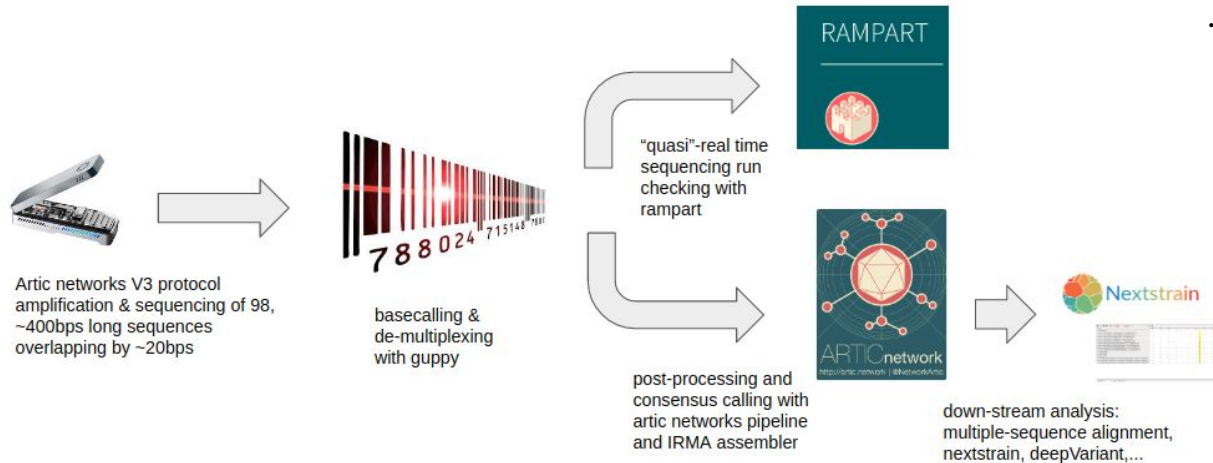
Bix @FEM - Examples

**Sequencing and assemblying of Sars-Cov-2 samples
from the Province of Trento (sponsored by Fondazione VRT)**

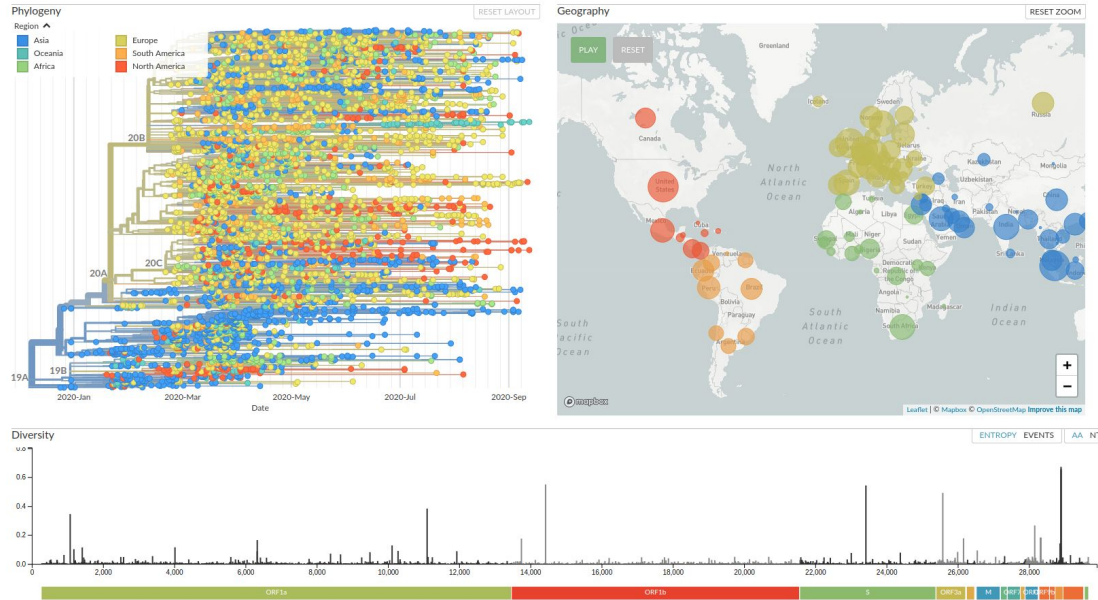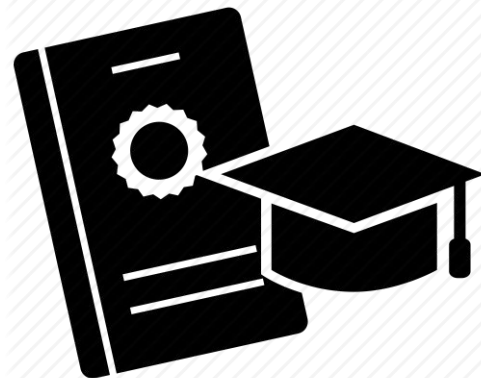# Opportunities @FEM

MSc External thesis

Are you interested in a bioinformatics project in NGS data analysis, RNA Seq, data integration?

Talk to me or email me at:

luca.bianco@fmach.it
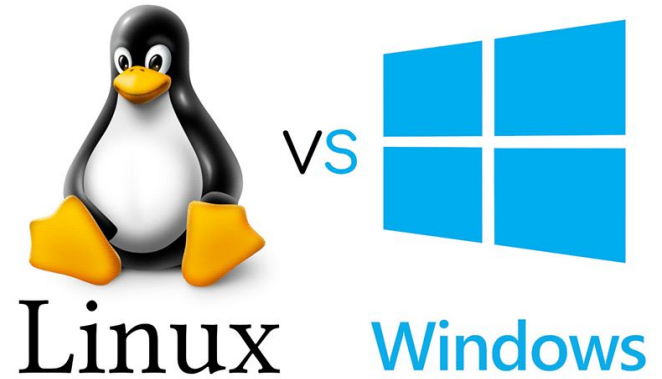
Scientific Programming Practical

Back to business now!

Linux or Windows?

vs

Up to you, as far as this course is concerned...

but, if you are looking for a career in bioinformatics, I think it would be a good idea to get familiar with Linux

Two options:
- Linux on windows (via virtualization software)
- Dual boot system (decide which to use at boot)

In the description of the practical you have some instructions on how to do the two things.

Think about the two options today and install Linux in the next few days...

Scientific Programming Practical

In this practical you will

1. Install Python 3.x (and pip)
2. Install Visual Studio Code
3. Get familiar with the Python console
4. Start using Visual Studio Code and advanced features (like debugging)
5. End the session with some exercises

Scientific Programming Practical

Console VS. Integrated Development Environment (IDE)

```
biancol@bluhp:~$ python3
Python 3.5.2 (default, Aug 18 2017, 17:48:00)
[GCC 5.4.0 20160609] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

Python is an **interpreted** language, therefore we can **interact directly** with the interpreter typing python code in the **console**

```
>>> print("Hi there")
Hi there
>>> print("{} + {} = {}".format(10,20, 10+20))
10 + 20 = 30
>>>
```

Scientific Programming Practical

Console VS. Integrated Development Environment (IDE)

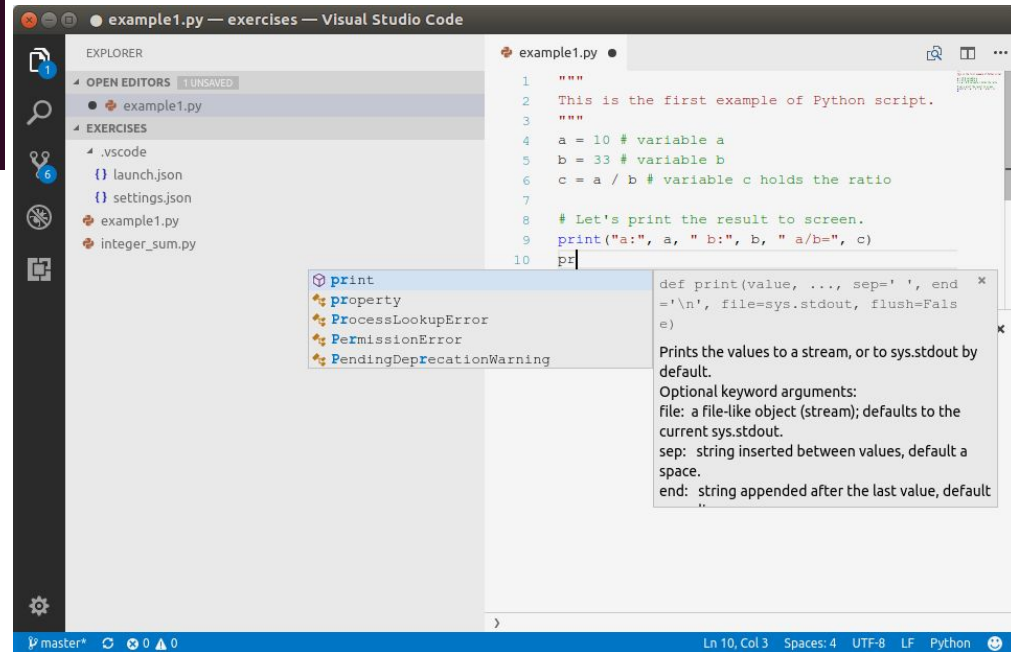Console: very convenient in some occasions for small things you do not do often, or for learning purposes...

...but we want to write code that we can save and reuse (i.e. **modules**)

Scientific Programming Practical

Console VS. Integrated Development Environment (IDE)

Scientific Programming Practical

Console VS. Integrated Development Environment (IDE)


```
biancol@bluhp:~$ python3
Python 3.5.2 (default, Aug 18 2017, 17:48:00)
[GCC 5.4.0 20160609] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

```
>>> print("Hi there")
Hi there
>>> print("{} + {} = {}".format(10,20, 10+20))
10 + 20 = 30
>>>
```
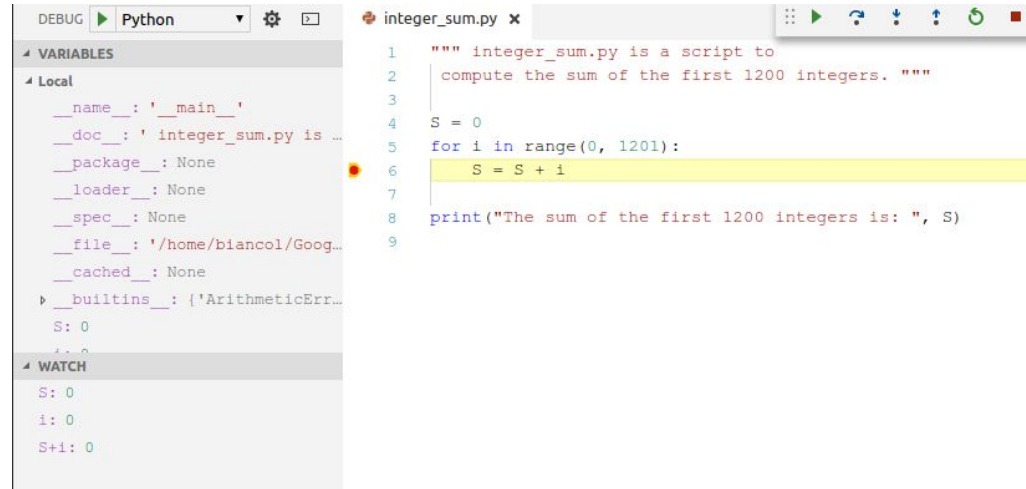
The debugger

```
DEBUG   Python                          integer_sum.py ×

▲ VARIABLES                             1  """ integer_sum.py is a script to
▲ Local                                 2   compute the sum of the first 1200 integers. """
    __name__: '__main__'                3
    __doc__: ' integer_sum.py is …     4  S = 0
    __package__: None                   5  for i in range(0, 1201):
    __loader__: None                    6      S = S + i
    __spec__: None                      7
    __file__: '/home/biancol/Goog…     8  print("The sum of the first 1200 integers is: ", S)
    __cached__: None                    9
  ▷ __builtins__: {'ArithmeticErr…
    S: 0

▲ WATCH
    S: 0
    i: 0
    S+i: 0
```

Notebooks and Jupyter

"Jupyter is a web-based interactive development environment for python/R.. notebooks, code, and data."

Notebooks contain both the **code**, some **text describing the code** and the **output of the code execution**,

Jupyter is becoming the de-facto standard for writing technical documentation.
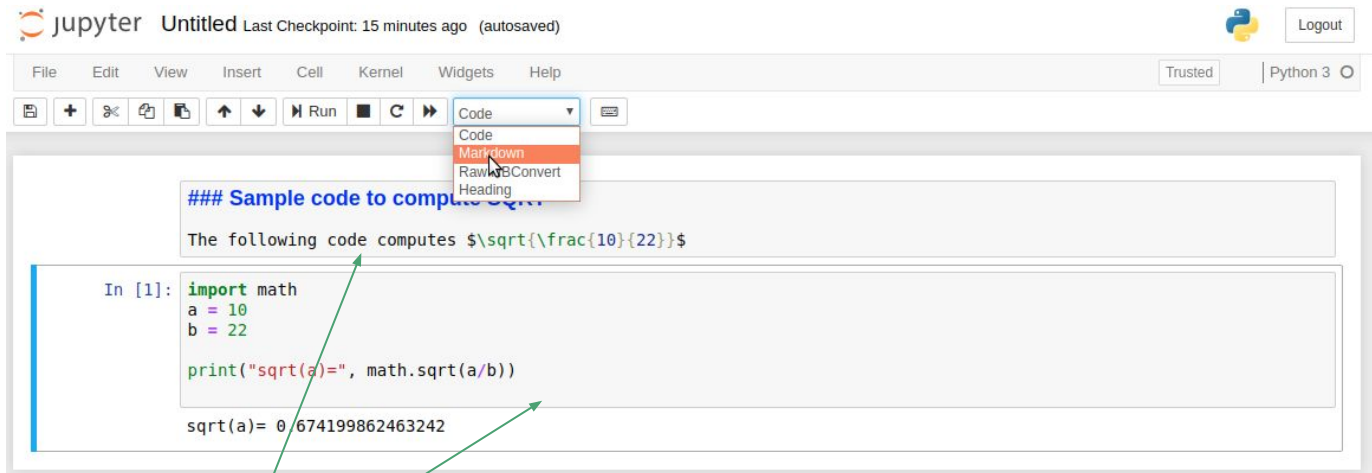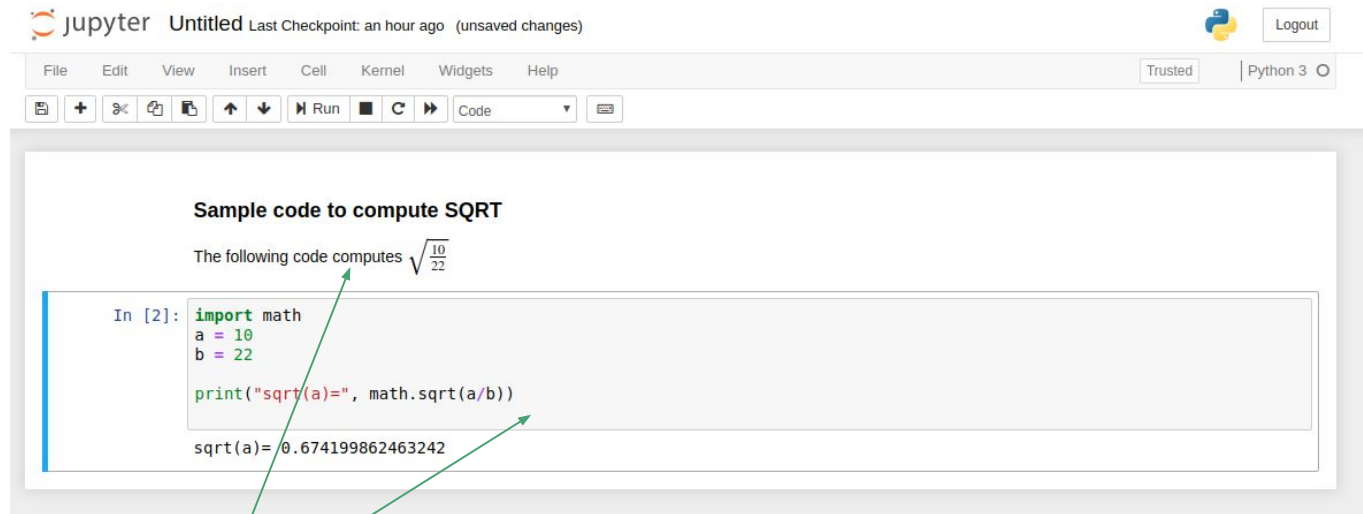

Cells

Notebooks and Jupyter

Notebooks contain both the **code**, some **text describing the code** and the **output of the code execution**,

Jupyter is becoming the de-facto standard for writing technical documentation.

A cell can be executed by clicking on **Run**

Cells
(after Run)

Resources

All material regarding practicals will be found here:

http://qcbsciprolab2020.readthedocs.io

@

luca.bianco@fmach.it

Scientific Programming for QCB

Download: PDF EPUB HTML

General Info

The contacts to reach me can be found at this page.

Timetable and lecture rooms

Due to the current situation regarding the Covid-19 pandemic, Practicals will take place ONLINE this year. They will be held on **Mondays from 14:30 to 16:30** and on **Wednesdays from 11:30 to 12:30**.

Practicals will use the Zoom platform (https://zoom.us/) and the link for the connection will be published on the practical page available in this site a few minutes before the start of the session.

This first part of the course will tentatively run from **Wednesday, September 23rd, 2020 to Monday, November 2nd, 2020**.

Moodle

In the moodle page of the course you can find announcements and videos of the lectures. It can be found here.

Zoom links

The zoom links for the practicals can be found in the Announcements section of the moodle web page. To get you started quickly, I report them here:

Join Zoom Meeting https://unitn.zoom.us/j/97253388646

Meeting ID: 972 5338 8646 Passcode: 794500

Slides

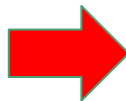Slides of the practicals will be available on the top part of each practical page.

Timetable

Mondays:

ONLINE: 15,30 - 17,30

Wednesdays:

ONLINE: 11,30 - 13,30

**!!! please write these details down,
I will remove them from the site !!!
(they will be on moodle)**

http://qcbsciprolab2020.readthedocs.io

Timetable and lecture rooms

Due to the current situation regarding the Covid-19 pandemic, Practicals will take place ONLINE this year. They will be held on **Mondays from 14:30 to 16:30** and on **Wednesdays from 11:30 to 12:30**.

Practicals will use the Zoom platform (https://zoom.us/) and the link for the connection will be published on the practical page available in this site a few minutes before the start of the session.

This first part of the course will tentatively run from **Wednesday, September 23rd, 2020 to Monday, November 2nd, 2020.**

Zoom links

The zoom links for the practicals can be found in the Announcements section of the moodle web page. To get you started quickly, I report them here:

Join Zoom Meeting https://unitn.zoom.us/j/97253388646

Meeting ID: 972 5338 8646 Passcode: 794500

@
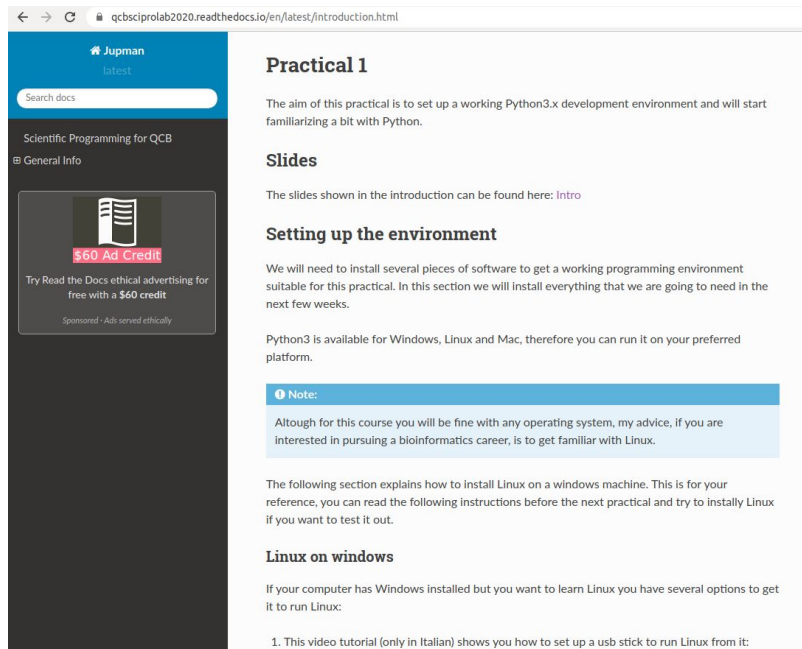
luca.bianco@fmach.it

Any questions?

If not, please go to:

https://qcbsciprolab2020.readthedocs.io/latest/introduction.html

@

luca.bianco@fmach.it