**Motivation and objectives of the research in this field**

The pandemic outbreak of the coronavirus disease COVID-19, caused by the virus species SARS-CoV2, has created unprecedented attention towards the genetics mechanisms of coronaviruses. We respond to such urgent need by means of a novel research project, so as to facilitate current and future research studies. We are driven by a conceptual model for virus sequences, https://www.biorxiv.org/content/10.1101/2020.04.29.067637v3, and by our previous experience on search systems for human genomics, http://geco.deib.polimi.it/genosurf/.

In our broad vision, the availability of open-source databases for both human genomics and viruses will provide important opportunities for research, amplified to the maximum when human and viral sequences will be interconnected by the same human being, playing the role of host of a given virus sequence as well as provider of genomic and phenotype information. During the PhD, the student will add more discovery-oriented entities to the model, that could be of use in a future scenario, e.g., a new pandemic offspring. A user researching on diagnosis could ask, for example, what sequence patterns are unique to the whole or sub-part of the database (i.e., do not appear in viruses within the database). Whereas, a user working on vaccine development could be interested in what are the epitopes (i.e., antigen parts to which
Methods and techniques that will be developed and used to carry out the research

The student will use a data science approach, going through the phases of: *Data collection and preparation*. Data will be identified, cleaned and integrated. *Data exploration*. The prepared data will be analyzed with unsupervised techniques (e.g. clustering, association rules extraction) to gain insight on relevant relationships among data instances and/or features. *Model generation*. Several models will be designed for the relevant variables (e.g., risk level). Appropriate tuning for each algorithm will be performed. *Decision making*. Insight gained by all the analysis steps will be discussed with domain experts. Particularly relevant in this stage is the interpretability of the obtained results.

Educational objectives

The student will learn how to compare, classify and adapt available methods for data extraction, analysis and visualization in genomics (and, more generally, in data science); the student will also learn how to master genomic data and how to use for answering research questions related to viral infections.

Job opportunities

There is a huge market for data analysis experts who are specialized in genomic research, both in Academia and in industry. The field is exponentially growing.

Composition of the research group

1 Full Professors
2 Associated Professors
1 Assistant Professors
10 PhD Students

Name of the research directors

Stefano Ceri

Contacts

stefano.ceri@polimi.it
http://www.bioinformatics.deib.polimi.it/geco/
### Additional support - Financial aid per PhD student per year (gross amount)

<table>
<thead>
<tr>
<th>Housing - Foreign Students</th>
<th>1st year</th>
<th>2nd year</th>
<th>3rd year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1500.0 € per student</td>
<td>1000.0 € per student</td>
<td>1000.0 € per student</td>
</tr>
</tbody>
</table>

Max number of financial aid available: 4, given in order of merit.

### Housing - Out-of-town residents
(more than 80Km out of Milano)

--

### Additional information: educational activity, teaching assistantship, computer availability, desk availability, any other information

**LIST OF UNIVERSITIES, COMPANIES, AGENCIES AND/OR NATIONAL OR INTERNATIONAL INSTITUTIONS THAT ARE COOPERATING IN THE RESEARCH:** UFL One Health Center; NUS - Singapore; L3S - Hannover

**INCREASE IN THE SCHOLARSHIP FOR STAYS ABROAD:** Euro 566.36 per month, for up to 6 months

**EDUCATIONAL ACTIVITIES** (purchase of study books and material, including computers, funding for participation in courses, summer schools, workshops and conferences): financial aid per PhD student per year

2nd year: euros per student (1534)

3rd year: euros per student (1534)

**TEACHING ASSISTANSHIP:** (availability of funding in recognition of supporting teaching activities by the PhD student)

There are various forms of financial aid for activities of support to the teaching practice. The PhD student is encouraged to take part in these activities, within the limits allowed by the regulations.

**COMPUTER AVAILABILITY:**

1st year: individual use
2nd year: individual use
3rd year: individual use

**DESK AVAILABILITY:**

1st year: individual use
2nd year: individual use
3rd year: individual use

The student will use the Data Science and Bioinformatics lab at DEIB