Improving Integrated Molecular Surveillance of Bad Bugs with Variation Graphs

Integrated molecular surveillance (IMS) is a method for tracking the spread of infectious diseases caused by microorganisms such as bacteria, viruses, and fungi by using molecular techniques. Whereas traditional epidemiological data can be incomplete and work at a limited resolution, molecular data give unbiased insights such for identification and contact tracing. It has become a essential component of pathogen surveillance (Knyazev et al 2022). The generic goal of IMS is to understand the genetic characteristics of a pathogen, as well as to track its spread and identify potential outbreaks (Fig. 1).

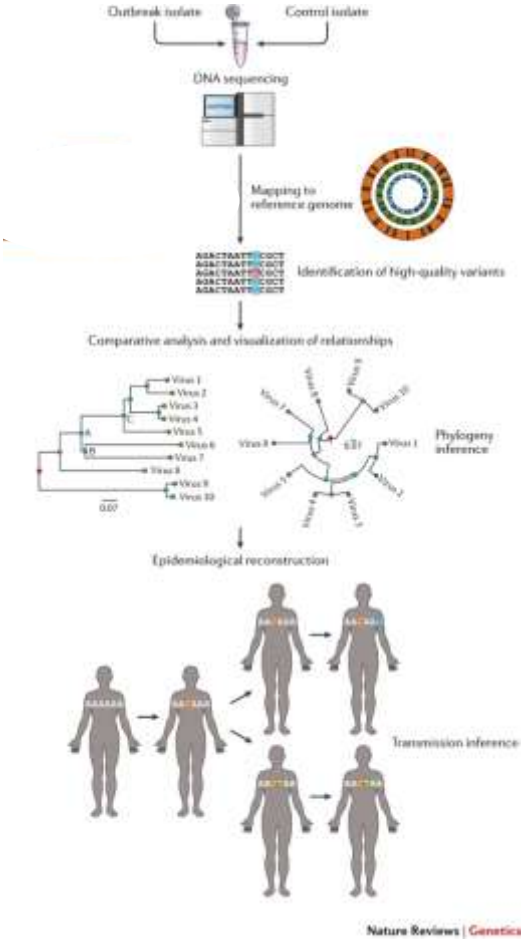

Figure 1: (adapted from Gardy and Loman, 2018), Genomic approaches to identifying transmission events typically involve four steps. In **step 1**, outbreak isolates, and often non-outbreak control isolates, are sequenced and their genomes mapped against a single reference genome. In **step 2**, the genomic differences between the sequences are identified — depending on the pathogen and the scale of the outbreak, these may include features such as genetic variants, insertions and deletions or the presence or absence of specific genes or mobile genetic elements. In **step 3**, these features are examined to infer the relationships between the isolates from whence they came — a variant common to a subset of isolates, for example, suggests that those cases are epidemiologically linked. At **step 4**, the genomic evidence for epidemiological linkages is reviewed in the context of known epidemiological information, such as social contact between two cases or a common location or other exposure. Recently, automated methods for inferring potential epidemiological linkages from genomic data alone have been developed, greatly facilitating large-scale genomic epidemiological investigations

The two first steps of integrated molecular surveillance correspond to strain typing and identification. This information can then be used for tracking the spread of the pathogen, identify potential outbreaks or assess its evolution over time (step 3 and 4).

Whole genome sequencing (WGS) gives access to an organism's entire genome. This can provide detailed information about the pathogen's genetic makeup, including the presence of virulence factors or antibiotic resistance genes.

Even in the case of WGS, the use of a single reference genome makes it challenging to accurately identify the differences between strains. Genomic information that is not present in

the reference strain could be virtually lost. Given that pathogens are highly diverse (two strains of the same bacteria *E. Coli* can have up to 40% difference in their genomic content), this can have dramatic impacts on the subsequent steps. Sequence variation graphs, that summarise the genomic diversity across a set of strains of the same species, have been proposed to overcome those limitations (Fig. 2). However, the technology of variation graph is still in its infancy and it is not clear nowadays how much they can be used in practice in an IMS setup.

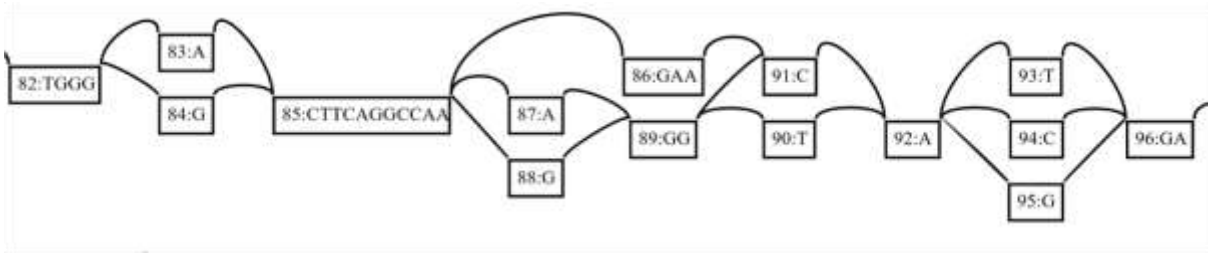

Figure 2: Toy example of a sequence variation graph. Sequences are represented as nodes, which are connected by edges

The goal of master project will be to setup an evaluation framework to investigate:

- How well variation graphs can perform compared to reference-based genotyping?
- Which strategy is the most suitable to construct such graph structures?
- Which strains should be considered when constructed a variation graph (thousands of strains are available for each species in public databases)?

To do so, we will prototype a benchmark tool on a few classical viruses and bacteria, and compare the performance of the different approaches for the IMS tasks of genotyping and of variant identification. How much can variation graphs improve strain genotyping and variants identification? How can results of different approaches be integrated?

This master project is done in collaboration with the Robert Koch Institute (RKI). The RKI is in charge of organising the country molecular surveillance for a range of pathogens such as SARS-CoV2, Tuberculosis or Gonorrhoea. The results from the comparison will be used to decide which potential variation graphs have for providing more accurate genomic information and for tracking the spread of pathogens. It can have a direct impact on IMS in Germany for the next epidemics.

Why should you join this project?

Students interested in the project should have good programming skills and experience with a previous programming project. We do not require any prior knowledge in genomics, biology, or epidemiology (but of course, it does not hurt having it).

The scientific project is suited for a large range of competencies and areas of interest. If you are interested in applying your programming and modeling skills to public health issues, or if you like to understand challenging algorithmic and data representation problems applied to large graph structures, this project is for you. We can adapt the project to your interest.

Look forward to an engaged supervision by our teaching team Please do not hesitate to contact us in case of questions or step by in building K.

Contacts:
Dr. Hugues Richard (hugues.richard@hpi.de / RichardH@rki.de)
Jens-Uwe Ulrich (jens-uwe.ulrich@hpi.de)
Prof. Dr. Bernhard Renard (bernhard.renard@hpi.de)

Further readings

The importance of molecular data for pathogen surveillance (Gardy & Loman 2018, Knyazev et al 2022).
The principles of **Integrated Molecular Surveillance** for viruses (Wohl et al 2016)
The concept of **pangenome variation graphs** (Eizenga et al 2020)
Advantages of full genome information for strain typing (Dudas & Bedford 2019)

1. Gardy JL, Loman NJ. 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet.;19(1):9-20. https://www.nature.com/articles/nrg.2017.88
2. Knyazev, S., Chhugani, K., Sarwal, V. *et al.* 2022. Unlocking capacities of genomics for the COVID-19 response and future pandemics. *Nat Methods* **19**, 374–380. https://doi.org/10.1038/s41592-022-01444-z
3. Wohl, Shirlee, Stephen F. Schaffner, and Pardis C. Sabeti. 2016. Genomic Analysis of Viral Outbreaks. PMID: 27501264, Annual Review of Virology 3 (1):173–195. https://doi.org/10.1146/annurev-virology-110615-035747
4. Jordan M. Eizenga, Adam M. Novak, *et al* .2020. Pangenome Graphs *Annual Review of Genomics and Human Genetics* 21:1, 139-162. https://www.annualreviews.org/doi/abs/10.1146/annurev-genom-120219-080406
5. Dudas, Gytis, and Trevor Bedford. 2019. The ability of single genes vs full genomes to resolve time and space in outbreak analysis. BMC Evolutionary Biology 19 (1):232. https://doi.org/10.1186/s12862-019-1567-0.