LauzHack 2019

The SOPHiA Challenge

November 16, 2019

Build a scalable service to identify pathogens and threats in hospital environments

Hi there,

we challenge you to build a service for a network of hospitals where various environments are tested for the presence of pathogenic bacteria and viruses. The hospitals intend to test environments such as patient rooms, operating rooms, nurse stations as well as the patients that reside in these environments. DNA samples are collected and go through high-throughput sequencing that produces a large data set where short DNA fragments from thousands of organisms are mixed together.

The service performs the analysis of the DNA data sets provided by the hospitals, determines the composition of species that are present in order to identify and characterize potential threats. These threats may include known pathogenic strains of microorganisms, bacteria that are resistant to antibiotics or sanitization procedures and others to be defined by the team. The final objective is to build a fast, accurate, scalable service that can efficiently help hospitals identify and react to such threats.

Metagenomic sequencing using the latest high-throughput technologies allows rapid DNA analysis of environmental samples. These sequencing techniques produce millions of short reads of 100-300 base pairs that describe the nucleotide sequences of DNA fragments from the mix of organisms that were present in the sample.

The two data sets that we provide were generated by whole genome metagenomic sequencing of real DNA samples from various hospital environments. They will let you focus on various aspects of the challenge, such as processing the data to identifying threats in a fast and accurate manner or presenting the results in ways that are most useful to hospitals in overcoming the threats.

Dataset 1: Brain infections

This publicly available dataset comprises fastq files with NGS reads from 11 different samples. The samples were taken from different parts of brain tissue from 10 different patients. For most patients you have only one sample which was taken after brain imaging found anomalies in certain parts that could be caused by infections. For PT8 there are two samples S1 and S2 which correspond to the outer and inner perimeter of a white spot seen on the MRI. To simplify your life a bit, we already provide the files where all reads that match the human reference genome have been removed. The remaining reads may originate from a variety of sources, the pathogens in the sample as well as false signals coming from contaminations from sample extraction or wetlab procedures. You can assume that each patient has a different disease. General idea: Given a set of reference genomes (bacteria, viruses, fungi, ...) you may be able to identify the pathogens for the different samples by sequence similarity.

Difficulties: Be aware, that many species share genomic sequences. The closer they are related, the more. Therefore, sequence comparison to a too small set of reference genomes may lead to misleading findings.

Dataset 2: Comparative metagenomics

This dataset is the output of metagenomic sequencing of DNA samples from various hospital environments from the US. Whole-genome DNA was extracted and sequencing was performed using the Ilumina HiSeq instrument in 125 bp x 125 bp paired end mode. You will be provided 200 samples (2 fastq.gz files each) that are in average 200-300 Mb in size and if interested we can give more details on the source of each sample.

With this dataset the challenge is more about efficient processing that allows fast and accurate identification of pathogens and scalable methods of comparing the findings between different environments. It would be very interesting to find ways of locating the source of infections by comparing environments that are known to be spatially close to each other.

Wish you fun,

Adam, Jonas and Sergei
Sophia Genetics

Data

The data is available from the following Azure blob storage account:

https://lauzhack2019.blob.core.windows.net/data/?sv=2019-02-02&ss=b&srt=sco&sp=rwdlac&se=2019-11-18T07:45:33Z&st=2019-11-15T23:45:33Z&spr=https&sig=tgeZyvviE6hxD%2F53ZYsAYpJPcdozHWjyC6q%2FSImm0NQ%3D

You can either use azcopy from Microsoft to list and download the samples or wget by including the files names in the URL, e.g.:

```
wget https://lauzhack2019.blob.core.windows.net/data/dataset_2/SRR5240636_1.fastq.gz?sv=2019-02-02&ss=b&srt=sco&sp=rwdlac&se=2019-11-18T07:45:33Z&st=2019-11-15T23:45:33Z&spr=https&sig=tgeZyvviE6hxD%2F53ZYsAYpJPcdozHWjyC6q%2FSImm0NQ%3D
```

The list of sample names will be provided on the slack channel. You can find reference genomes for a large variety of widely abundant species as well as pathogens here:

https://www.ncbi.nlm.nih.gov/refseq/