Pipeline used for the analysis of omics data

Project title: Application of animal genomics and data mining to predict and monitor novel coronavirus potential infections (VirAnimalOne)
Organization: University of Bologna (Italy)
PI: Luca Fontanesi (luca.fontanesi@unibo.it)

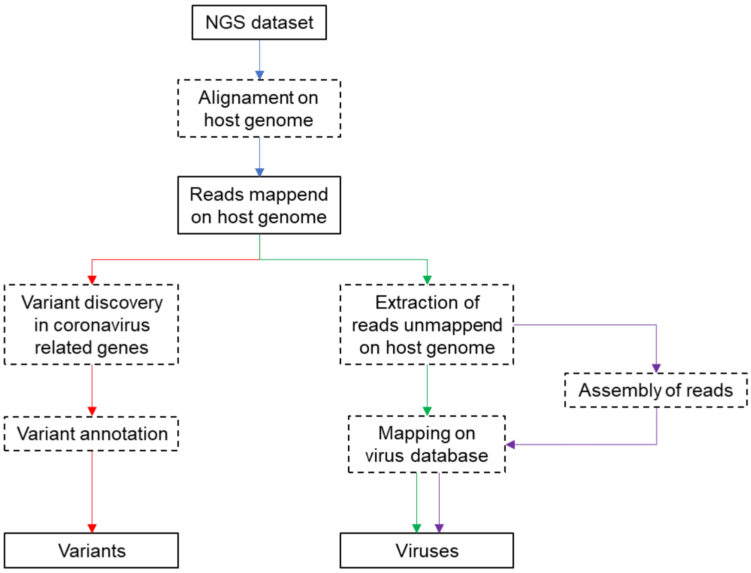
Short background

Public DNA sequence databases (European Nucleotide Archive, ENA: https://www.ebi.ac.uk/ena; Sequence Read Archive, SRA: https://www.ncbi.nlm.nih.gov/sra) contain large numbers of information derived from next generation sequencing (NGS) projects on a variety of animal species and produced for many different purposes that could be mined for other objectives. Therefore, we took advantage from these extensive resources to exploit them for a novel purpose which aimed to retrieve additional information useful in a context of a One Health general approach against new viral infections, involving domestic animals as potential hosts We followed two main avenues of exploitation of these resources and derived by the following concepts from our previous research results: we recently reported that these datasets can be mined to identify virus that have been involuntarily sequenced together with the animal genome or transcriptome, as they were in the tissues that served for nucleic acid isolation (Bovo et al., 2017, PLoS One 12, e0179462); variants of the host genome could confer resistance or susceptibility to virus infections as they might modify host protein receptors that are encoded at the DNA level. At present there are no extensive exploitation of the publicly available resources towards these two directions.

Applied Methodology

Therefore, in this project we applied large scale mining of publicly deposited genomic datasets derived from genome sequencing of pets, livestock and related wild animal species: (i) to identify unexpected coronavirus and/or other virus sequences, (ii) to mine the host animal genomes for potential variants that might confer resistance or susceptibility to SARS-CoV-2 and other coronaviruses known to infect both humans and animals and (iii) to evolutionarily and structurally evaluate host receptor conformations and infer potential animal susceptibility to coronavirus infections. For these purposes, the bioinformatics pipeline showed in **Figure 1** was applied. Details are given in next paragraphs.

Figure 1. Bioinformatics pipeline used in the project.


NGS dataset
Alignment on host genome
Reads mappend on host genome
Variant discovery in coronavirus related genes
Extraction of reads unmappend on host genome
Assembly of reads
Variant annotation
Mapping on virus database
Variants
Viruses

Animals and sequencing datasets

Genomic information (i.e. whole-genome sequencing data; WGS) produced from cattle, pigs, chickens, rabbits, related wild species (that can be useful from an evolutionary perspective) were retrieved from the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/). REST APIs available through the ENA taxonomy services were used to retrieve the deposited metadata information (e.g. taxonomy identifier, project identifier, sequenced bases, etc.) considering the following taxonomic elements and identifiers: (i) 9903 (genus *Bos;* cattle dataset), 9822 (genus *Sus;* pig dataset), (ii), 9331 (species *Gallus gallus;* chicken dataset), 9984 (genus *Oryctolagus*; rabbit dataset), 9665 (genus *Mustela*; ferret dataset) and 9865 (species *Felis catus*; cat dataset). For the final database, we retained only datasets presenting the following metadata tags: library_source = GENOMIC, library_layout = PAIRED and library_strategy = WGS. Then, based on the number of sequenced nucleotides and the size of the reference genome(s), only the subset of datasets presenting an estimated depth of sequencing against the corresponding host reference genome greater than 5× were retained for further analyses. Sequencing data were locally download via the Aspera ascp command line client. The datasets are presented in **Table S1-S6** provided as Excel files (please see the attached files).

Sequence alignment, variant detection and annotation from sequencing data

Downloaded sequencing data were initially quality checked with FASTQC v.0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Then, reads were mapped on the host reference genome using the BWA-MEM algorithm v.0.7.17 and the parameters for paired-end data. Picard v.2.1.1 (https://broadinstitute.github.io/picard/) was used to remove duplicated reads. The following host reference genomes were used: ARS-UCD1.2 (cow), Sscrofa11.1 (pig), GRCg6a (chicken), OryCun2.0 (rabbit), Felis_catus_9.0 (cats), MusPutFur1.0 (ferrets). Detection of variants on aligned reads was carried out using the GATK v.4.1.8.1 (https://gatk.broadinstitute.org/hc/en-us) HaplotypeCaller. A joint variant calling approach with GATK4 GenotypeGVCFs was applied to identify variants affecting genes involved in coronavirus infections (please see the next paragraph: coronavirus related genes). Only biallelic variants, covered by at least 3 reads, were retained. Polymorphisms were detected considering a region spanning 5 kbp upstream and 5 kbp downstream the corresponding gene location as reported in Ensembl (or in NCBI if not available). Variants were annotated using the Variant Effect Predictor (VEP) v.95.0 (https://www.ensembl.org/info/docs/tools/vep/index.html), by predicting with SIFT v.5.2.2 (https://sift.bii.a-star.edu.sg/) their impact to the protein function.

Coronavirus related genes

A comprehensive list of genes involved in coronavirus infections was compiled based on a literature survey (**Table S7**). We initially identified a set of "core" genes resulting essential for the virus entry, including those genes coding for proteins involved in the viral surface recognition process (i.e. *ACE2*, *ANPEP*, *CEACAM1* and *DPP4*; Li, 2015 Receptor recognition mechanisms of coronaviruses: a decade of structural studies. J Virol 89, 1954-1964. https://doi.org/10.1128/JVI.02615-14) and genes encoding protein priming proteases (i.e. *FURIN* and *TMPRSS2*; e.g. Hou et al. 2020, New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis. BMC Med. 18, 216. https://doi.org/10.1186/s12916-020-01673-z). The gene set was that expanded by including 46 genes coding for proteins having host-virus interactions conserved between coronaviruses (e.g. the host protein targets the same viral protein in different coronaviruses) and 65 genes interacting with several coronaviruses, but not necessarily targeted by the same viral proteins (Perrin-Cocon et al., 2020 The current landscape of coronavirus-host protein–protein interactions. J Transl Med 18, 319. https://doi.org/10.1186/s12967-020-02480-z). Other 11 gene were further included based on the most recent studies about SARS-CoV-2 (e.g. Wang et al. 2021, Genetic screens identify host factors for SARS-CoV-2 and common cold coronaviruses. Cell 184, 106-119. https://doi.org/10.1016/j.cell.2020.12.004). The final gene set, evaluated in all the studied organisms, included a total of 128 genes. The porcine gene set was further expanded by adding other 10 protein coding genes (interacting with porcine coronaviruses (Perrin-Cocon et al., 2020).

Building a comprehensive virus database

The accession list of all known viral genomes was retrieved from the NCBI (https://www.ncbi.nlm.nih.gov/genome/viruses/; September 2020). The DNA sequences of each listed reference viral genome ad its strains was download in fasta format using the Entrez Programming Utilities (*esearch* and *efetch* utilities). Sequence redundancy was allowed in order to maximize the preferential sequence alignment of reads with specific virus strains. The final dataset included 238353 viral genomes (including different strain for each reference genome).

In silico identification of viruses

The virome was explored by adopting two strategies (**Figure 1**): (i) read mapping and (ii) assembly. Both procedures started with a step that implied the extraction of sequencing reads not mapped on the host reference genome. Unmapped reads pairs were subsequently extracted with Samtools v1.10 (Li et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352) with the options -f 12 and -F 256.

Read mapping – Unmapped reads were initially mapped over the virus database. In this initial step, BWA was used to speed up the identification process and reads pairs were separated and treated as single entities to maximize the mapping. To refine the results, mapped reads were aligned back over the same resource (virus database) by using BLAST + v.2.7.1 (algorithm *blastn*). As several strains are available for a given reference genome, for each read we retained all the alignments presenting an E-value ≤ 0.01, sequence coverage ≥ 97% and a sequence identity ≥ 75%. Reads mapped simultaneously over strains belonging to different reference genomes were discarded. Read pairs were interrogated for pointing out the same reference viral genome: pairs mapped on different reference genomes were discarded. Only viral genomes having at least three mapped pairs were considered as characterizing the sample.

Assembly – Following our previous work (Bovo et al., 2020. Shotgun sequencing of honey DNA can describe honey bee derived environmental signatures and the honey bee hologenome complexity. Scientific Reports 10, 9279. https://doi.org/10.1038/s41598-020-66127-1), unmapped reads were assembled with MEGAHIT v.1.1.3 (https://github.com/voutcn/megahit) with default parameters except the "meta-large" option that forced the usage of a k-mer list equal to 27,37,47,57,67,77,87. For each assembled contig, we retained all the alignments presenting an E-value ≤ 0.01, sequence coverage ≥ 50% and a sequence identity ≥ 75%. Also in this case, contigs mapped simultaneously over strains belonging to different reference genomes were discarded. The detection of one viral DNA sequence (contig) was considered sufficient for declaring the presence of a given virus.

Fundings:

EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-05-2018-2019, grant Agreement number 831644.