MSc DISSERTATION ASSESSMENT COVERSHEET

1ST (SUPERVISOR) MARKER

Name of student:	Juan Ledesma Moreno
MSc Programme:	MSc in Bioinformatics (part-time)
Title of Dissertation:	Integration of clinical and next generation sequencing data into a database system to improve the management of Hepatitis C virus information in a national reference microbiology service unit
Year of Submission:	2021
Project Supervisor:	**Conrad Bessant**
Declaration:	**'This research dissertation is submitted for the MSc in Bioinformatics at Queen Mary, University of London'**
Data sharing (delete appropriately)	I **do** allow consent for my project to be shared with future cohorts of students on MSc programmes in the School of Biological and Chemical Sciences and on institutional repositories and websites.

<u>To be completed by Supervisor</u>:

Second Marker Name	
Turnitin score if higher than 17%	

Section A. Continuous Assessment Mark (For supervisor only to assess the student's performance during the practical stages of the project. Note this scheme is inclusive of project type, for example, including field-based, laboratory, modelling and meta-analytic studies). Please highlight the words or phrases that are appropriate to justify the grade.

Mark (%)	**Criteria**
95 – 100%	Outstanding technical capacity, went beyond expectation in developing protocols, or analytical tools. The engagement was outstanding contributing to lab meetings, and the life of the hosting research group.
85 – 94 %	Exceptional performance showing outstanding technical ability, originality and initiative, high levels of commitment and application, ability to plan and organise the research programme and to contribute substantially to the development of the work.
70 – 84 %	Excellent performance, showing clear evidence of originality, initiative and ability to contribute to the development of the programme.
60 – 69 %	Good performance, notable for steady commitment, sound technical ability and some evidence of initiative and originality. Some contribution to the development of the project but mainly following the advisor's suggestions.
50 – 59 %	Performance generally satisfactory but with some deficiencies in technical ability and/or limited levels of commitment and application to the project. Little or no contribution to the development of the work and very little if any initiative and originality.
40 – 49 %	Performance somewhat weak, characterised by poor technical ability and low levels of commitment and application. Poor understanding of the project and effectively no contribution to the planning and organisation of the work.
20 – 39 %	Unsatisfactory performance. Very poor technical ability amounting to inability to perform routine tasks reliably. Very low levels of commitment and application including unacceptably low attendance.
0 – 19 %	Very poor performance! Careless and totally disorganised, and non-existent technical skills. Unacceptably low attendance.

Please comment in the box below on the student's coursework below, focusing on the student's performance productivity, attitude, commitment, timeliness, originality, organization, technical ability.

Section A. Supervisor's Mark for continuous assessment =_______________%

Please complete the dissertation marking (to follow)

Section B (1st marker). Assessment of dissertation report

Please highlight the points that best describe the project

Mark (%)	**Criteria**
95 - 100%	Outstanding performance in producing a document which could be submitted as it for publication from a technical, analytical and editorial perspective.
85 – 94%	Exceptional project report showing very broad understanding of the project area and outstanding knowledge of the relevant literature. Exceptional presentation and analysis of results, logical organisation and ability to evaluate critically and discuss results with insight and originality.
70 – 84 %	An excellent project report showing evidence of wide reading, with clear presentation and thorough analysis of results and an ability to evaluate critically and discuss research findings. Clear indication of insight, understanding and originality. An extremely competent and well-presented report overall, excellent in most aspects.
60 – 69 %	A good project report which shows a clear understanding of the problem and sound knowledge of the relevant literature. Sound presentation and analysis but perhaps not exploiting the results to the full. Relevant interpretation and critical evaluation of results, though with some limitations regarding the scope. Good general standard of presentation and organisation.
50 – 59 %	A satisfactory project report which shows some understanding of the problem but limited knowledge and appreciation of the relevant literature. Presentation, analysis and interpretation of the results at a basic level and showing little originality or critical evaluation. Some weaknesses in the organisation and presentation of the report.
40 – 49 %	A weak project report showing only limited understanding of the problem and superficial knowledge of the relevant literature. Results presented in a somewhat confused or inappropriate manner and incomplete or erroneous analysis in places. Discussion and interpretation of results severely limited, including some basic misunderstandings, and with very little originality or critical evaluation. General standard of presentation weak.
20 – 39 %	An unsatisfactory project report containing substantial errors and omissions. Very limited understanding, or in some cases misunderstanding, of the problem and very restricted and superficial appreciation of the relevant literature. Very poor, confused and, in some cases, incomplete presentation of the results and limited analysis of the results including some serious errors. Severely limited discussion and interpretation of the results revealing little or no ability to relate experimental results to the existing literature. Very poor overall standard of presentation.

Mark (%)	Criteria
0 – 19 %	A very bad project report containing many errors and faults. Virtually no real understanding of the problem and of the literature pertaining to it. Haphazard presentation of results, and in some cases incompletely presented and virtually non-existent or inappropriate or plainly wrong analysis. Discussion and interpretation seriously confused or wholly erroneous revealing basic misconceptions.

Please comment on the student performance in the dissertation, with reference, as necessary, to each of the component parts, abstract, introduction, results, discussion, references (appropriateness and accuracy).

Section B. Supervisor's Mark (written report) = _______________%

Section B. Second Marker (written report)* = _________________%
***from 2nd Marker marksheet**

Agreed mark Section B (written report) =* _____________*%

If the two marks for section B. (written report) differ by more than 10% a third marker will be needed to adjudicate and for a dissertation mark to be agreed between all three.

(if necessary third markers name: _________________)

Overall mark:

= (0.75 x Agreed Mark Section B) + (0.25 x Supervisor's Mark Section A*) *from page 3

= ________________

NB Masters students' overall dissertation marks relate to the University descriptive categories as follows:

Mark (%)	**Grade**
70.0% or above	Distinction
60.0 - 69.9%	Merit
50.0 - 59.9%	Pass
Below 49.9%	Fail

Please note the dissertations are second (blind) marked. Markers may mark the work in sequence, one after the other, or in parallel. If they differ in their mark for section B by more than 10%, then a third marker will be needed.

ABSTRACT

Background: Hepatitis C virus (HCV) shows a high degree of genetic diversity having been classified into 8 genotypes and more than 90 subtypes distributed worldwide. Recombination is a rare event but has also been described. The acute infection is usually asymptomatic but can lead to chronic hepatitis. Current drugs used to treat HCV infection show a high rate of success but viral genotyping and screening for resistance mutations are recommended to select the best treatment for the patients.

Aim: The Antiviral Unit performs viral characterization on samples from patients infected with HCV by means of using next generation sequencing (NGS). The objective of this study is to implement the use of the software GLUE in order to store NGS sequences and the related clinical and epidemiological information and test a specific method provided by GLUE to perform the genotyping.

Methods: A GLUE project containing sequence data from 224 reference strains and 9 recombinant viruses was initially developed. A selection of 460 sequences generated in the laboratory by NGS were added to the project and tested for genotyping using the Maximum Likelihood Clade Assignment (MLCA) method. The data from two test SQLite databases, generated to recreate the laboratory information management system in PHE, was imported into custom tables in the database GLUE holds. The correct population of the database was tested by querying for known inputs/outputs.

Results: The MLCA method was able to correctly identify the set of recombinant viruses used in the project and classify the NGS sequences into genotypes and subtypes. Inter and intra-genotype recombination forms consisting on 2k/1b, 1a/3a, 1c/1a and 1c/1/1a were detected in sequences but further analysis must be carried out to confirm these findings. However, the tests carried out on the GLUE database showed issues with the custom tables used to store clinical and epidemiological data related to the sequences.

Conclusion: GLUE was able to provide an automated method to identify the genotypes and subtypes of HCV sequences from clinical samples but the software had limitations when working with custom tables to accommodate new data, which potentially can affect the management of the GLUE project.

Abstract word count: 350

Main text word count: 6739

ACKNOWLEDGEMENTS

I would like to thank Public Health England (PHE) for the financial support for my training on Bioinformatics, finally completed with this dissertation.

Special thanks to David Williams and Ana Penedos, both from PHE, not only for the excellent supervision of this work but also for the good advice on Bioinformatics, your help and support during all this time. It is always a pleasure to work with you and to chat about life.

Special thanks to Tamyo Mbisa, head of the Antiviral Unit (AVU) in PHE, for his involvement in this project and giving feedback for the dissertation. It would have been impossible to start this journey in Bioinformatics without his help during all these years, firstly with his support for me to start this MSc in Bioinformatics and recently for giving me the opportunity to keep developing my skills in Virus Reference Department Bioinformatics. My total gratitude too to all the members of the AVU for the understanding and support during these two tough years.

Thanks to Conrad, from Queen Mary University of London, for the advice, supervision and organising the MSc during these times of pandemic. Muchas gracias por todo. It has been a pleasure.

Thanks to all my friends in England, Spain and around the world for being always there. Wherever you are I always can count on you.

Many thanks to my family for the love that you have always given me during all these years and teaching me the importance of fighting for what I believe. Os quiero mucho.

Finally, huge thanks to my partner, Diana. Your support and love during all these years is priceless. It has been hard to achieve this and other things in my life but I have always had your support during this time even when I felt like giving up. Thanks for your love and for being my friend. May the 4th being with you! Te quiero Dianita.

Salud!

LIST OF CONTETS

LIST OF ALL FIGURES

LIST OF TABLES

Appendices

INTRODUCTION

Hepatitis C virus (HCV) is a positive single strand RNA virus belonging to the *Hepacivirus* genus, Family *Flaviviridae*. Its genome, which is about 9.6 kilobases long, consists of an open reading frame that codes for a precursor polyprotein flanked by two non-coding regions at 5' and 3' ends (Moradpour and Penin, 2013). The precursor polyprotein is cleaved into 10 mature proteins, 3 structural and 7 non-structural (NS) proteins. The structural proteins Core and envelope glycoproteins E1 and E2 are involved in the formation of the viral nucleocapsid and the entry into the cell and assembly of the progeny virion, respectively (Dubuisson and Cosset, 2014; Shi and Suzuki, 2018). Proteins NS3, NS4A, NS4B, NS5A and NS5B play an important role in the replication of the virus whereas protein p7 and NS2 contribute to the viral assembly process (Dubuisson and Cosset, 2014; Gosert et al., 2003; Suzuki, 2012). NS2 and N-terminal of NS3 is an autoprotease that cleaves the region between NS2 and NS3, an essential step in RNA replication (Kim and Chang, 2013). NS3 acts as serine protease and produces the cleavage of NS4A, NS4B, NS5A and NS5B whereas its C-terminal end is an RNA helicase. NS4A works as a cofactor of NS3 by creating a complex of both enzymes needed for NS3 to carry out its function (Lindenbach, 2013). NS4B and NS5A are involved in the modulation of the RNA replication and the assembly of the virus while NS5B is the RNA dependent RNA polymerase (Dubuisson and Cosset, 2014).

HCV, as other RNA viruses, shows a great degree of genetic diversity due to the non-proofreading RNA polymerase. HCV has been classified so far into 8 genotypes according to the genetic differences of up to 35% at nucleotide level (Borgia et al., 2018; Smith et al., 2014). In addition genotypes can be sub-divided into multiple subtypes that may differ up to 15% at nucleotide level having been described with 90 or more subtypes (Hedskog et al., 2019; Smith et al., 2019). Genotype 1, highly distributed in Central and East Asia, Europe, Western Africa and the American continent, is the most prevalent genotype being responsible for almost 50% of infection followed by genotype 3 with an estimated rate of infection of up to 30% and mainly found in South and Central Asia, Australasia, Eastern and Western Europe and Tropical Latin America (Messina et al., 2015; Petruzziello et al., 2016). Genotypes 2 and 4 account for 17-27% of the global infections, mostly in West Africa and Pacific Asia and Central and North Africa and Middle East, respectively. The estimation for the infection due to genotypes 6 and 5 ranged between 3 and 6% and their distribution is restricted to South East and East Asia and Southern Africa. Genotypes 7 and 8 are not very frequent and have been associated with few cases from the Democratic Republic of Congo and India, respectively (Borgia et al., 2018; Murphy et al., 2015). Recombination, which requires the infection of the same cell with two different virus strains simultaneously, can also contribute to viral genetic diversity. Although some studies have documented a degree of prevention from superinfection by a second virus in cells already infected with HCV or in patients after liver transplantation

(Ramírez et al., 2010; Tscherne et al., 2007), natural HCV recombination events have been reported. Inter-genotype recombinant viruses has been described with forms such as 2a/1b, 2b/1b, 2b/1a, 2a/1a, 3a/1b, 2b/6w, 2i/6h, 2i/6p 2k/1b and genotypes 2 and 5 in Japan, Philippines, Taiwan, Austria, Italy, Russia, Uzbekistan or the United States of America (Bhattacharya et al., 2011; Hoshino et al., 2012; Kageyama et al., 2006; Kalinina et al., 2002; Kurbanov et al., 2008; Lee et al., 2010; Legrand-Abravanel et al., 2007; Noppornpanth et al., 2006; Okada et al., 2018; Paolucci et al., 2017a; Stelzl et al., 2017). In addition intra-genotype recombination, consisting of mixed genomes from 1b and 1a or 4d and 4a, has been previously found in France, Uruguay, Portugal, Peru and Brazil (Almeida Calado et al., 2011; Colina et al., 2004; Gaspareto et al., 2016; Morel et al., 2016; Moreno et al., 2009).

The main transmission route of HCV to humans is through direct exposure to contaminated blood that occurs during transfusion of unscreened blood, unsafe medical procedures or use of recreational injected drugs (Aceijas & Rhodes, 2007; Hajarizadeh et al., 2013; Lanini et al., 2019; Spearman et al., 2019). Although sexual transmission seems not to be a high risk of infection between heterosexual individuals (Terrault et al., 2013), the infection has increased among men that have sex with men infected by human immunodeficiency virus (HIV), possibly due to injuries during unprotected sex, sexual practices including fisting or the use of toys or sex groups (Chan et al., 2016; Jordan et al., 2016). Vertical transmission has also been considered as a possible route with factors like maternal viral load or certain obstetric procedures that can increase the risk of transmission between mothers infected with the virus and their newborn babies (Benova et al., 2014; Mavilia and Wu, 2017).

Acute infection is usually asymptomatic and, although some individuals are able to clear the virus, the progression of the infection leads to chronic hepatitis in most of the patients (Lingala and Ghany, 2015). Up to 20% of the individuals suffering HCV chronic infection can develop cirrhosis and hepatocellular carcinoma (Spearman et al., 2019; Zoulim et al., 2003), which depends on host factors like age, gender and genetics, individual behaviour like alcohol consumption or other co-infection with HIV or hepatitis B virus (Lingala and Ghany, 2015; Missiha et al., 2008).

The current treatments for HCV infections are based on the combination of direct-acting antivirals (DAA) that target the NS3/NS4A complex, the NS5A protein and the RNA polymerase NS5B (Preciado et al., 2014; Spearman et al., 2019). The therapy using DAAs has showed high rate of success at treating patients with chronic infections using specific combinations of drugs and regimens to treat mostly all the genotypes effectively (Baumert et al., 2019; González-Grande et al., 2016). However, the appearance of resistance mutations in the targets for those drugs (Di Stefano et al., 2021; Palanisamy et al., 2018) can lead to failure of the therapy. Consequently screening for resistance mutations and performing genotyping are recommended steps to perform in order to select the best drug and the specific regimen to treat infections in naive patients and those with treatment failure (Di Stefano et al., 2021; Lagging et al., 2018; Paolucci et al., 2017b).

Sequencing methods are routinely used in clinical Virology. The application of next generation sequencing (NGS) technology together with the reduction of its cost in the last years has generated a huge amount of data to deliver more efficiently common proceedings in this field such as diagnosis, viral characterization, phylogenetic analysis and discovery of new viruses (Barzon et al., 2011; Chiara et al., 2021; Parikh et al., 2017; Parker and Chen, 2017; Popescu et al., 2018; Thorburn et al., 2015). Bioinformatics software has been developed in parallel to viral sequencing to cover the increase of its needs. Online specific virus-driven tools are commonly used to share and access sequence data for influenza, HIV, HCV or severe acute respiratory syndrome coronavirus 2 (Bogner et al., 2006; Kuiken et al., 2005, 2003; Maxmen, 2021; Zhang et al., 2017) as well as to perform sequence analysis like subtyping and analysis of recombination, antiviral resistance genotyping or phylogenetic analysis (Kalaghatgi et al., 2016; Liu and Shafer, 2006; Neher and Bedford, 2015; Pickett et al., 2012; Schultz et al., 2009). A recently developed tool, called Genes Linked by Underlying Evolution (GLUE) (Singer et al., 2018), provides a resource to store sequence data and perform standard analysis. The core schema of GLUE is designed to store the information in a fixed relational database managed through specific objects (Singer et al., 2018). The software can be locally installed and, as it is not focused on a particular virus, the schema can used to accommodate specific viral sequences of interest (Dennis et al., 2019; Singer et al., 2020, 2019) and be extended by adding new fields and custom tables for clinical data. Projects, focused on evolutionary characteristics of the viruses they are designed for, usually include a set of consensus reference sequences, which provides the genomic regions of interest, and a genetic classification of the sequences based on their phylogenetic relationships. Nucleotide alignments are used in GLUE to define groups of related sequences based on their genetic homology. Although GLUE has its specific command lines and hierarchical modes to interact with the database and the objects, scripts can be created to allow the users an easy way to access and manipulate the data.

The Antiviral Unit (AVU) at Public Health England routinely receives samples from patients infected with HCV to test for the antiviral resistance profile and perform the genotyping and subtyping of the viruses. By means of an NGS method the samples are processed and whole genome sequences of HCV are generated and used to characterise the virus. The results of the antiviral resistance genotyping and the identification of the subtype for each sample are then shared with clinicians from hospitals so decisions on best treatment for the patients can be made.

The aim of this study is to develop a GLUE project to store HCV whole genome sequences generated in AVU, integrate related clinical and epidemiological data in the database and test a genotyping tool before GLUE is fully deployed for routine use in the lab.

MATERIALS AND METHODS

DATA

Next Generation Sequencing data

A total of 460 sequences generated in 9 different NGS experiment runs were used to test the project. The sequences were generated using a Sequence Capture enrichment NGS protocol using Illumina MiSeq instrument and MiSeq Reagent Kit V2 300 cycles according to the previously reported (Manso et al., 2020). Each FASTQ files was labelled with the sample identifier. After human read removal, the reads from FASTQ were processed using a combination of *de novo* assembly and reference mapping as described previously (Manso et al., 2020). The resulting BAM file was then analysed using a PHE C++ software, QuasiBAM (Penedos et al., 2015) that generated a consensus sequence stored in a FASTA file and a tabular file which recorded nucleotide frequency and additional metrics for each nucleotide contained in the sequence.

The sequences were generated from 376 clinical samples and 20 positive controls. Sixty of these clinical samples were repeated for sequencing and used to test the database for duplicated information. The remaining 316 samples were processed in single NGS experiments, which generated 320 sequences. The FASTQ files of 313 of these samples produced single consensus sequences using the default bioinformatics pipeline described above. A bioinformatics pipeline under development called genomancer, using similar approach as the default one but able to distinguish coinfection due to different viral strains (unpublished), was applied on the FASTQ files for 2 of samples and generated 2 sequences per sample. In addition, the paired FASTQ files from one sample were analysed using the default pipeline and genonamcer, which produced 1 and 2 sequences, respectively.

NCBI-GenBank Sequences

Reference sequences from the latest HCV classification performed in 2019 (Smith et al., 2019) were downloaded from International Committee on Taxonomy of Viruses[1] (ICTV). The set consisted of 224 sequences which included at least 1 representative for each subtype confirmed by ICTV (n=176) as well as sequences with identified genotype but non-assigned subtype so far (n=48) (Table A1). A CSV file containing the identifier of these sequences and the HCV classification according to genotype and subtype was generated. The sequences were aligned using HCVAlign tool[2] and MAFFT v7 (Katoh et al., 2002) with the codon-alignment set to 5 codons. Afterwards, the alignment was visually inspected and manually edited in regions where some nucleotides were misaligned. A phylogenetic tree was then generated (newick encoding) using FastTree v2.1.11 (Price et al., 2010) with substitution model GTR and gamma distributed rate variation across sites.

A set of nine FASTA files corresponding to nucleotide sequences from recombinant viruses previously described (AB622121, AB677527, AB677530, AM408911, AY587845, DQ155560, DQ364460, EU643835 and JF779679) was obtained from the database NCBI-GenBank to test a genotyping in the GLUE project.

Clinical and epidemiological data

PHE's laboratory information management system (LIMS), Management Of Laboratory Information System (Molis v4.4), manages the booking of external samples sent to PHE, recording patient information and assigning an unique identifier (or Molis number) to each received sample. The Molis number is used during sample processing in PHE to track the sample status and identify the results for each sample. An additional database system, managed by a team of epidemiologists at PHE, is used to collect patient and epidemiological information not included in Molis.

Although these two databases were intended to be used, due to limitations to access to these data at the time of the developing of the project, two testing databases (molis.db and epi_database.db) were created using SQLite v3.33.0[3] and Python package SQLite v2.6.0 to emulate the relational tables and the data contained in the PHE data systems. These databases contained specific information (sample ID, sample and reception date, previous HCV genotyping result, patient ID, hospital, date of birth, age at the time of diagnosis, country and nationality, ethnicity, city of residence, gender, treatment and antiviral drugs used for the treatments) intended to be used in the GLUE project.

HCV GLUE AVU PROJECT

Project directory

A structured directory was created to store the data to be used in the project and allow GLUE easy access to it. The main directory consisted of a parent directory called *HCV-GLUE-AVU* that contained the additional subdirectories:

- **sources/**, used to store the sequence data in GenBank XML and/or FASTA format. The sequences were organised in this subdirectory according to particular characteristics they shared. Sequences downloaded from NCBI-GenBank were stored in a specific folder called "*ncbi-refseqs*" while those sequences generated by Illumina sequencing in the same experiment were stored in the same folder, after being processed as described below.
- **glue/**, used to store the scripts with GLUE commands to interact with the project and allow access to the data and the creation of the objects in the database. This folder was also used to store JavaScript programs, used for the data population in the project.
- **modules/**, used to store the configuration of the modules (tools that allow functionalities in GLUE) in XML format needed to be used for the project.

- **alignments/**, contained the alignment of the reference sequences generated in section 3.1.2 to be imported in the project and used in several steps of the project building.
- **trees/**, used to store a phylogenetic tree with the genetic distances and bootstrap values generated in section 3.1.2. This subdirectory also contained other files needed to be used in the process of transferring of the data to the project in order to provide total functionality to the phylogeny reconstruction.
- **tabular/**, contained data stored in CSV files to be transferred to the fixed GLUE tables and custom tables created in project. The synthetic databases created in section "3.1.3. Clinical and epidemiological data" were saved in this subdirectory for convenience. Specific folders (i.e. table_sequence, table_patient, table_sample, epi_data and who_countries) were created to organise the data according to their nature.
- **sqlite3/**, contained the Python scripts and data used to create the synthetic databases used in the project.
- **queries/**, used to store the information retrieved after querying the database of the project.
- **zDrive/**, used to store the original fasta sequences generated by NGS to test the project.

The parent directory also contained the following files that will be discussed in the next sections: *avuHcvProject.glue*, *import_NGSseqs_to_GLUE.py*, *populate_metadata_from_molis.py*, *add_Epidata_to_Patient_and_Treatment.py*, *genotyping_of_NGS_sequences.py*, *query_recombinant_viruses.py, query_NGS_genotyping.py* and *query_info_NGS_run.py*. The project is available on the open repository https://github.com/juanledesma78/HCV-GLUE-AVU.

Building the project

A project called *hcv_glue_avu* was created using GLUE v1.1.107 and the dependencies Java OpenJDK Runtime Environment v1.8.0_282, MySQL v8.0.23, Blast-2.2.31, Mafft v7.475 and RAxML 8.2.12 on a machine Intel® Core™ i5-8265U CPU @ 1.60GHz × 8 with 16 GB RAM using Ubuntu 18.04.5 as OS. Python 3.8 was also used to perform specific steps of the project. The script *avuHcvProject.glue*, located in the parent directory, worked as a master build file and guided all the steps needed to develop the project:

A) **Extension of the core schema** (steps 1 and 2 in Figure 1). All the scripts needed for this process, located in subdirectory glue/, were executed under the specific GLUE mode *schema-project*. The script *SchemaExtensions.glue* added new fields to the table *sequence* in order to capture metadata from those sequences retrieved from the database NCBI-Genbank and the sequences generated by NGS. An additional field *phylogeny* was also created in the table *alignment* to enable the transfer of the phylogenetic data to the database. This script also invoked two more glue scripts so 9 custom tables were created to store geographical details according to the World Health Organization for the reference sequences and clinical and epidemiological data from the sequences generated in AVU. At this stage, the links between custom tables and default GLUE tables were established.

B) **Creation of modules** (step 3 in Figure 1). Certain functionalities in GLUE projects (i.e. handling FASTA files, enabled the data transfer from CSV files or working with alignments and phylogeny data) are achieved by means of the use of the 57 modules available in GLUE (see documentation[3]). The modules and the specific parameters must be defined before being created in the project. The same module can be set with different configuration so they can interact with different data. Once the modules are created in the project, they are called in GLUE to be used in those phases of the project when their function is needed. Specific configurations saved in module/ in XML documents were used to create 41 modules in the project *hcv_glue_avu* (Table 1), process led by the script *Module.glue* after being invoked by the master file. The creation of modules was performed under the mode *project*, which was used for the remaining phases of the building project.

C) **Importing the reference sequences** (steps 4 and 5 in Figure 1). An initial step to obtain the sequence data of the reference agreed in the latest HCV classification from the database NCBI-GenBank was performed by the module *NcbiRefSeqsImporter_HCVSmith2019*, which contained the accession numbers of the reference sequences and those recombinant viruses selected for the project (3.1.2 NCBI-Genbank Sequences). The sequences and other metadata were saved in GenBank XML format in a folder *sources/ncbi-refseqs* created by the module. The master file *avuHcvProject.glue* then transferred the sequence data to the project by using the file name as identifier to record the new sequences (*sequenceID* in *sequence*) and by importing the nucleotide sequences from the specific field from the XML file. At this stage the notation *ncbi-refseqs* was added to the field *source_name* in objects *sequence* and *source*.

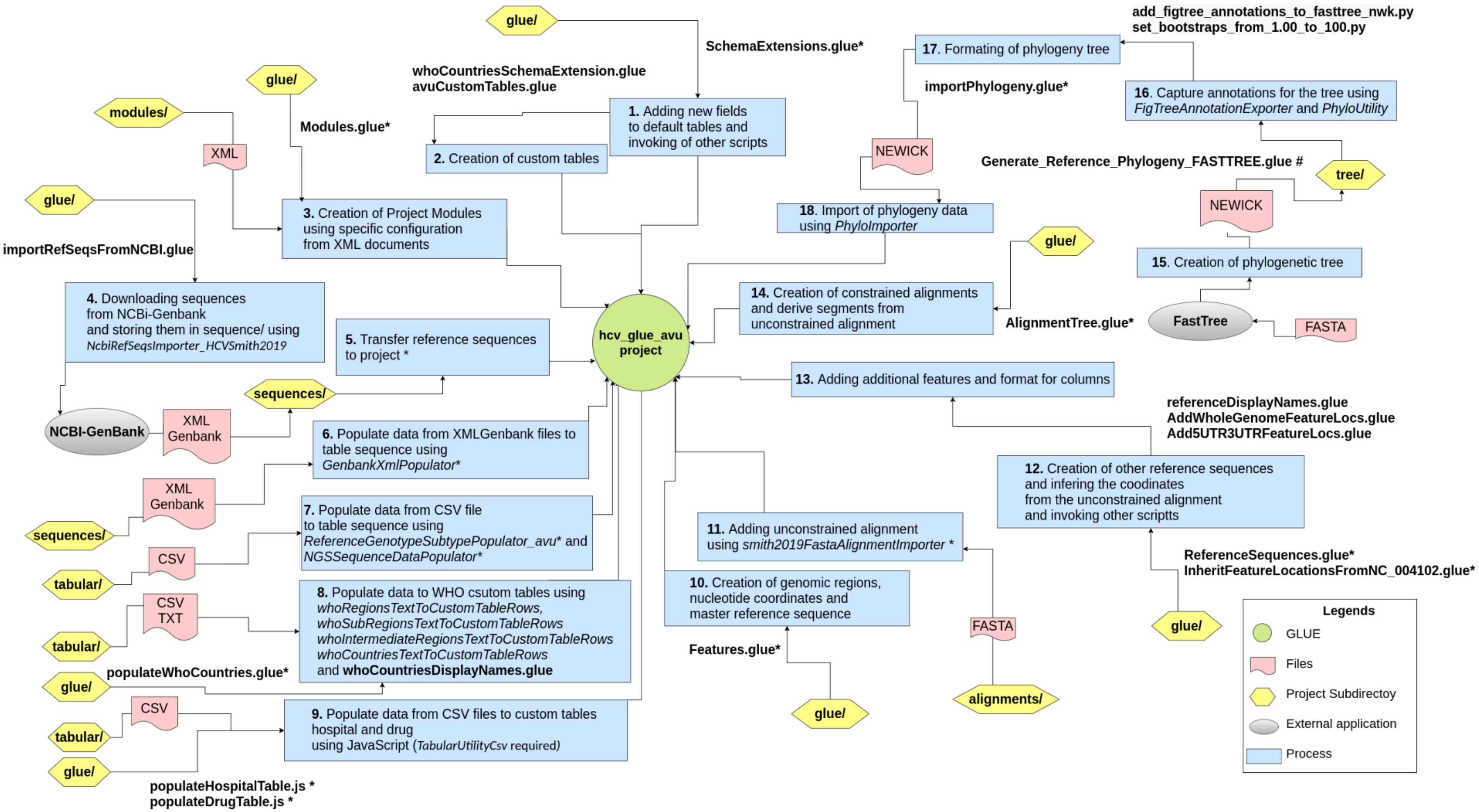
glue/
SchemaExtensions.glue*
whoCountriesSchemaExtension.glue
avuCustomTables.glue
1. Adding new fields to default tables and invoking of other scripts
2. Creation of custom tables
glue/
Modules.glue*
modules/
XML
3. Creation of Project Modules using specific configuration from XML documents
glue/
importRefSeqsFromNCBI.glue
4. Downloading sequences from NCBi-Genbank and storing them in sequence/ using NcbiRefSeqsImporter_HCVSmith2019
5. Transfer reference sequences to project *
NCBI-GenBank
XML Genbank
sequences/
6. Populate data from XMLGenbank files to table sequence using GenbankXmlPopulator*
XML Genbank
sequences/
7. Populate data from CSV file to table sequence using ReferenceGenotypeSubtypePopulator_avu* and NGSSequenceDataPopulator*
CSV
tabular/
CSV TXT
tabular/
8. Populate data to WHO csutom tables using whoRegionsTextToCustomTableRows, whoSubRegionsTextToCustomTableRows whoIntermediateRegionsTextToCustomTableRows whoCountriesTextToCustomTableRows and whoCountriesDisplayNames.glue
populateWhoCountries.glue*
glue/
CSV
tabular/
9. Populate data from CSV files to custom tables hospital and drug using JavaScript (TabularUtilityCsv required)
glue/
populateHospitalTable.js *
populateDrugTable.js *
hcv_glue_avu project
17. Formating of phylogeny tree
add_figtree_annotations_to_fasttree_nwk.py
set_bootstraps_from_1.00_to_100.py
importPhylogeny.glue*
16. Capture annotations for the tree using FigTreeAnnotationExporter and PhyloUtility
NEWICK
Generate_Reference_Phylogeny_FASTTREE.glue #
tree/
18. Import of phylogeny data using PhyloImporter
NEWICK
glue/
15. Creation of phylogenetic tree
14. Creation of constrained alignments and derive segments from unconstrained alignment
AlignmentTree.glue*
FastTree
FASTA
13. Adding additional features and format for columns
referenceDisplayNames.glue
AddWholeGenomeFeatureLocs.glue
Add5UTR3UTRFeatureLocs.glue
12. Creation of other reference sequences and infering the coodinates from the unconstrained alignment and invoking other scriptts
11. Adding unconstrained alignment using smith2019FastaAlignmentImporter *
ReferenceSequences.glue*
InheritFeatureLocationsFromNC_004102.glue*
10. Creation of genomic regions, nucleotide coordinates and master reference sequence
FASTA
glue/
Features.glue*
glue/
alignments/
Legends
GLUE
Files
Project Subdirectoy
External application
Process

Figure 1. Procedures performed during the project building. Bold characters are used to identify the glue scripts and JavaScript programs and italic characters are used to identify the modules used in each step. The elements of the workflow invoked by the master file avuHcvProject.glue for execution are labelled with an asterisk.

Table 1. Modules used in the project *hcv_glue_avu*. The project modules were created based on the modules available in GLUE but specific configurations were set for certain modules (italic) to perform their function in the project. For additional details on the parameters used see XML documents in subdirectory modules/ at https://github.com/juanledesma78/HCV-GLUE-AVU).

GLUE module used to create the project module	Project Module name	Function
ncbiImporter	NcbiRefSeqsImporter_HCVSmith2019	Imports the HCV reference sequences and the metadata from NCBI-GenBank. This module was used in an independent step to the project so the sequences were available before building the project
textFilePopulator	*ReferenceGenotypeSubtypePopulator_avu*, *NGSSequenceDataPopulator*	Populates genotyping data of the reference sequences obtained from NCBI-GenBank into fields *genotype* and *subtype* in table *sequence* Populates data from NGS sequences into fields *sample*, *hcv_wg_pipeline* and *pipeline_version* in table *sequence*
tabularUtility	TabularUtilityCsv	Allows the population of custom tables for AVU use with data contained in CSV files.
blastFastaAlignmentImporter	smith2019FastaAlignmentImporter	Imports the unconstrained alignment containing HCV reference sequences recommended by Smith 2019
kuiken2006CodonLabeler	kuiken2006CodonLabeler	Labels codons (Kuiken et al., 2006)
freemarkerTextToGlueTransformer	*WhoRegionsTextToCustomTableRows,* *whoSubRegionsTextToCustomTableRows,* *whoIntermediateRegionsTextToCustomTableRows,* *whoCountriesTextToCustomTa*	Populates information about WHO geographical distribution into the custom tables

	bleRows	
*GenbankXmlPopulato*r	*GenbankXmlPopulato*r	Populates table *sequence* with metadata of reference sequences from GenBank XML files
RaxmlPhylogenyGenerator	*RaxmlPhylogenyGenerator*	Generates a newick tree using RAxML software. However, FastTree2 was the software used to create the final phylogeny data used in the project
alignmentColumnsSelector	*PhylogenyColumnsSelector*	Selects the regions or block of nucleotides and the feature to be used to construct a phylogeny
FigTreeAnnotationExporter	*FigTreeAnnotationExporter*	Captures the IDs from the sequences in the database and exports the annotations to the tree generated in the phylogeny
PhyloUtility	PhyloUtility	Provides essential functionalities to phylogenetic processes
PhyloExporter	PhyloExporter	Exports a phylogenetic tree stored in table *alignment*
PhyloImporter	PhyloImporter	Imports a phylogenetic tree to be stored in table *alignment*
MaxLikelihoodPlacer	*MaxLikelihoodPlacer,* *MaxLikelihoodPlacerCore,* *MaxLikelihoodPlacerE1,* *MaxLikelihoodPlacerE2,* *MaxLikelihoodPlacerP7,* *MaxLikelihoodPlacerNS2,* *MaxLikelihoodPlacerNS3,* *MaxLikelihoodPlacerNS4A,* *MaxLikelihoodPlacerNS4B,* *MaxLikelihoodPlacerNS5A,* *MaxLikelihoo*dPlacerNS5B	Sets the object *alignment* and the substitution model and runs Maximum-likelihood Clade Assignment method using those parameters.
MaxLikelihoodGenotyper	*MaxLikelihoodGenotyper,* *MaxLikelihoodGenotyperCore,* *MaxLikelihoodGenotyperE1,* *MaxLikelihoodGenotyperE2,* *MaxLikelihoodGenotyperP7,* *MaxLikelihoodGenotyperNS2,*	Sets the cut-offs for the classification performed with the Maximum-Likelihood Clade Assignment method and gives the final classification of the

	MaxLikelihoodGenotyperNS3, *MaxLikelihoodGenotyperNS4A,* *MaxLikelihoodGenotyperNS4B,* *MaxLikelihoodGenotyperNS5A,* *MaxLikelihoodGenotyperNS5B*	sequences.
fastaExporter	fastaExporter	Exports a selection of nucleotide data into a FASTA file.
fastaAlignmentExporter	fastaAlignmentExporter	Exports a nucleotide alignment from a FASTA file into GLUE.

D) **Data population of table sequence and custom tables** (steps 6-9 in Figure 1). During this step, only data intended to be populated in a one-time process was transferred to the project database. Three different ways were used to carry out the data population. A first method used only modules to transfer the metadata from GenBank XML documents or from a CSV file (3.1.2 NCBI-Genbank Sequences) to table *sequence*. The data from this CSV file was needed as the genotype/subtype information for the reference sequences downloaded from the database NCBI-GenBank could not match the latest HCV classification. A second method, to populate the data for the custom tables WHO, combined the use of modules, a glue script and CSV and TXT files. JavaScript programs were used for a third method in order to populate the data containing a list of submitting hospitals and information about generic drugs used for treatments in custom tables *hospital* and *drug*.

E) **Definition of genomic regions and reference sequences** (steps 10-13 in Figure 1). The genomic regions (object *feature* in the project) to use in the project were defined according to the structure of the genome of HCV. A total of 16 genomic regions were defined for the project. An initial region, which consisted of the whole genome was used to define the non-coding regions 5' UTR and 3' UTR as well as the precursor polyprotein. Two additional features were created derived from the precursor protein, structural and non-structural proteins. These new features were then used to define the coding regions for the proteins Core, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A and NS5B. The script *Feature.glue*, which guided this process, also provided the specific nucleotide coordinates for the coding regions using the sequence NC_004102 as reference for the numbering. Additional references were created in the project using the other sequences recommended for the HCV classification. To avoid adding new coordinates for each new reference, the nucleotides positions for the new sequences were inferred using an alignment containing all the sequences and the reference NC_004102. In those cases where the sequences presented frame shifts due to insertion and deletions (indels), the coordinates for the blocks of aligned nucleotides (object *segment*) between sequences were derived manually and the problematic indels omitted from the segments and recorded in the script *InheritFeatureLocationsFromNC_004102.glue*.

F) **Generation of Alignment tree** (step 14-18 in Figure 1). The alignment tree is a specific structure in GLUE that groups phylogenetically related sequences and captures the homologies between blocks of aligned nucleotides from these sequences (Singer et al., 2018). GLUE uses two different alignments to handle the information on the sequences: constrained and unconstrained. A constrained alignment is generated using a specific sequence as reference to provide the coordinates for the other sequences in the alignment. Blocks of aligned nucleotides, which identify the homologies between sequences, are stored in columns but no column is used for those insertions that are not present in the reference. By contrast, unconstrained alignment does not use a reference to number the nucleotides position for the sequences in the alignment. Thus, blocks of aligned nucleotides as well as all the insertions found in each sequence will be stored in columns, which can lead to large number of columns as the number of sequences and the differences among them increase. In order to solve this potential issue, GLUE balances the use of both to store the information on nucleotides homologies in the projects. In order to create the alignment tree, an initial object *alignment* which contains all the sequences belonging to the same species is used to derive different genetic clades and create new children objects *alignments*. Each *alignment* contains several *members* or sequences, one of them being the proposed reference sequence (*refName*) for the sub-classification. The organisation of the *alignments* is stored in the relational database so that every child *alignment* is connected to its parent *alignment* and as are all the members. In the project *hcv_glue_avu*, the initial *alignment*, called *AL_MASTER*, was created by including all the HCV reference sequences. New *alignments* (i.e. *AL_1, AL_2, AL_1a, AL_1b, etc*) were derived from the master alignment by selecting the sequences belonging to specific genotypes and subtypes based on the HCV classification and a reference is assigned for each nascent *alignment*. At this stage, the nucleotide sequences for each *alignment member* were not aligned as they were transferred from object *sequence*. The external unconstrained alignment, imported in the previous step to define the coordinates for the reference sequences, was then used to define the homology blocks (aligned nucleotides) between sequences in the same region. This process was carried out by *AlignmentTree.glue*. A final step to capture the phylogenetic data from a NEWICK file derived from the alignment was performed by using the scripts *Generate_Reference_Phylogeny_FASTTREE.glue* and *importPhylogeny.glue*.

Adding Antiviral Unit data

The script *import_NGSseqs_to_GLUE.py* was used to upload the sequences generated by NGS to the project. The script took as an argument the path in zDrive/ where the original FASTA files were stored and created a new subdirectory for each experiment in sources/. The FASTA header and FASTA file name of each sequence were formatted to use the same ID (Figure 2) according to the following convention: sequence id, number(s) of the consensus sequence identified in the sequence analysis, alternative bioinformatics pipeline and id of the NGS experiment for the sequence. The edited sequences were transferred to the respective folder in sources/ and then imported into the project using the subprocess module.

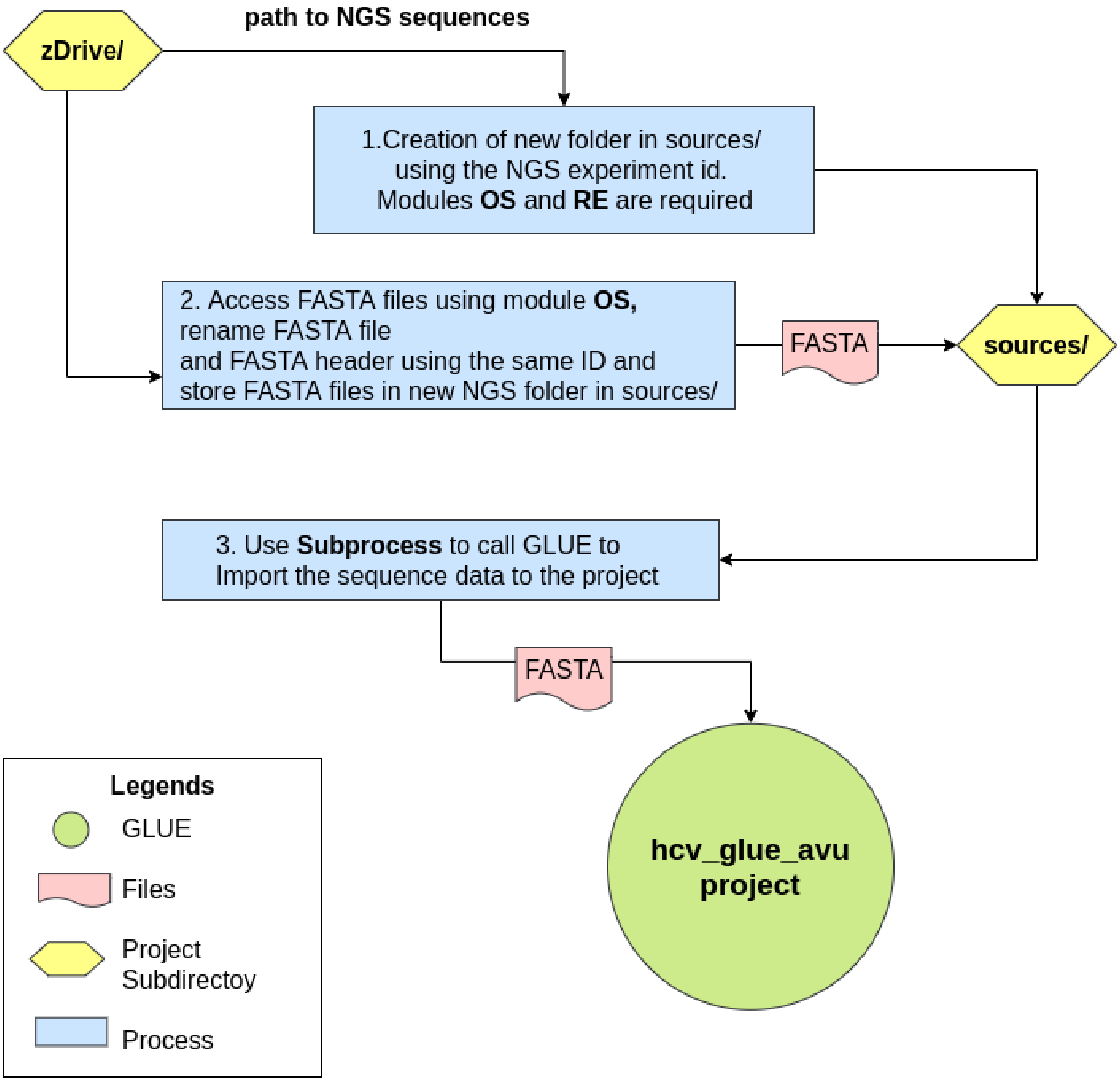
path to NGS sequences
zDrive/
1.Creation of new folder in sources/ using the NGS experiment id. Modules OS and RE are required
2. Access FASTA files using module OS, rename FASTA file and FASTA header using the same ID and store FASTA files in new NGS folder in sources/
FASTA
sources/
3. Use Subprocess to call GLUE to Import the sequence data to the project
FASTA
hcv_glue_avu project
Legends
GLUE
Files
Project Subdirectoy
Process

Figure 2. Processes to import the NGS sequences to the hcv_glue_avu project using the script import_NGSseqs_to_GLUE.py.

Clinical information retrieved from the synthetic database *molis.db* was imported to GLUE using the script *populate_metadata_from_molis.py*, which took as an argument the specific NGS experiment in sources/ to be processed. The script accessed and queried the database using SQLite3 and SQL commands and, by means of the Python Pandas library v1.2.5, stored the data needed for the custom tables *patient* and *sample* in CSV files, saved in tabular/table_patient/ and tabular/table_sample/, respectively (Figure 3). The script also accessed the FASTA files and generated a CSV file that contained the fields *sequence id*, *molis id*, *bioinformatics pipeline* and *version* derived from the headers of the sequence data. The data stored in the CSV files was populated to the corresponding tables in the GLUE project using *populatePatientTable.js*, *populateSampleTable.js* and the module *NGSSequenceDataPopulator*, after being invoked by the module subprocess.

Additional epidemiological data was imported to the project by means of *add_Epidata_to_Patient_and_Treatment.py*. The database *epi_database.db* was accessed using SQLite3 and the information about patient and their treatments were extracted using pandas by matching the field *patient id* contained in the CSV file in table_patient. Two CSV files for treatment and epidemiological information were created and stored in tabular/epi_data/. As in the previous python script, GLUE was called and ran the programs *populateEpiDataPatient.js* and *populateTreatmentTable.js* to add the new information to the database.

Genotyping

The Maximum Likelihood Clade Assignment (MLCA) method developed in GLUE (Singer et al., 2018) was tested for classifying the sequences in the project according to genotypes and subtypes using two different approaches.

An initial JavaScript program, called *genotyping_of_known_recombinant_viruses_test.js*, used 11 genomic regions (precursor polyprotein, Core, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A and NS5B) and other parameters required for MLCA and specified in the modules derived from *MaxLikelihoodGenotyper* and *MaxLikelihoodPlacer* to run (Table 1). Once the genotyping analysis on the sequences of known recombinant viruses obtained from NCBI-GenBank was completed, the results from each classification were then updated in table *sequence*. Identical principle was used to write another program called *genotyping_by_genes.js* to analyse batches of sequences from the same NGS experiment. However, the genotyping in this program was carried out using only the 10 genes derived from the precursor protein. After completion of the analysis, the results were updated in the corresponding fields in the table *sequence* and, when the results matched for the 10 genes, they were consolidated and uploaded in the fields *genotype* and *subtype* in table sequence. The execution of this JavaScript program was carried out by using the script *genotyping_of_NGS_sequences.py*.

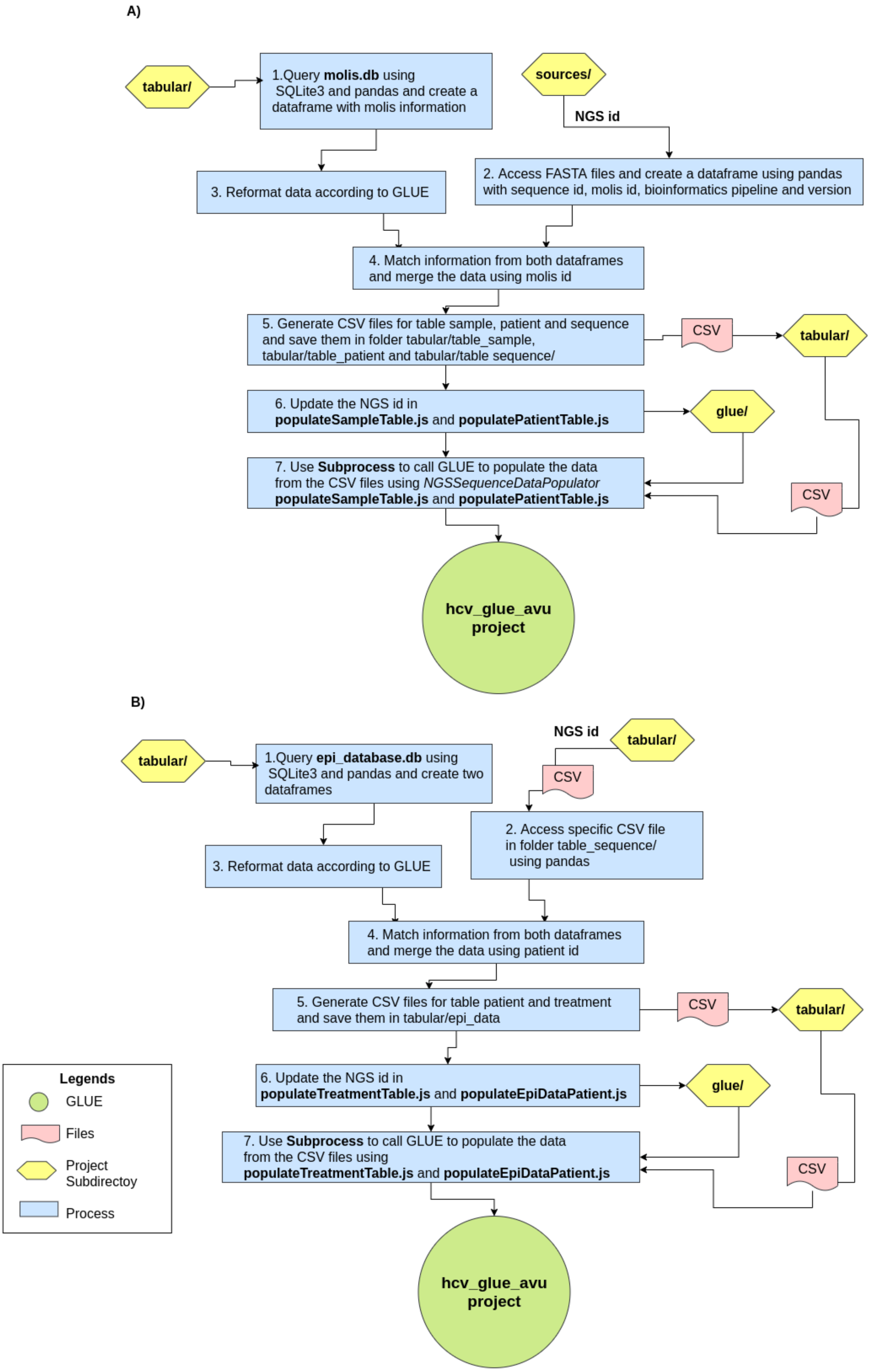
A)
tabular/
1.Query **molis.db** using SQLite3 and pandas and create a dataframe with molis information
sources/
NGS id
2. Access FASTA files and create a dataframe using pandas with sequence id, molis id, bioinformatics pipeline and version
3. Reformat data according to GLUE
4. Match information from both dataframes and merge the data using molis id
5. Generate CSV files for table sample, patient and sequence and save them in folder tabular/table_sample, tabular/table_patient and tabular/table sequence/
CSV
tabular/
6. Update the NGS id in **populateSampleTable.js** and **populatePatientTable.js**
glue/
7. Use **Subprocess** to call GLUE to populate the data from the CSV files using *NGSSequenceDataPopulator* **populateSampleTable.js** and **populatePatientTable.js**
CSV
hcv_glue_avu project
B)
tabular/
1.Query **epi_database.db** using SQLite3 and pandas and create two dataframes
NGS id
tabular/
CSV
2. Access specific CSV file in folder table_sequence/ using pandas
3. Reformat data according to GLUE
4. Match information from both dataframes and merge the data using patient id
5. Generate CSV files for table patient and treatment and save them in tabular/epi_data
CSV
tabular/
6. Update the NGS id in **populateTreatmentTable.js** and **populateEpiDataPatient.js**
glue/
7. Use **Subprocess** to call GLUE to populate the data from the CSV files using **populateTreatmentTable.js** and **populateEpiDataPatient.js**
CSV
hcv_glue_avu project
Legends
GLUE
Files
Project Subdirectoy
Process

Figure 3. Processes to import data from the testing databases molis.db (A) and epi_data.db (B) into the GLUE project.

Query tests

Queries were used to test the performance of the project database and interrogate for results of analysis done by MLCA. A first approach consisted of obtaining the information using the specific GLUE command lines in order to

a) test the custom tables and the links with the default tables

```
‘list sequence --whereClause "source.name='NGS91'" sequenceID
sample.id sample.sample_date sample.patient.id
sample.patient.treatment’
```

b) retrieve results for the genotyping on recombinant viruses,

```
‘list sequence ---whereClause "reference_status ='recombinant'"
sequenceID genotype subtype genotype_core subtype_polyprotein
genotype_e1 subtype_e1 genotype_e2 subtype_e2 genotype_p7 subtype_p7
genotype_ns2 subtype_ns2 genotype_ns3 subtype_ns3 genotype_ns4a
subtype_ns4a genotype_ns4b subtype_ns4b genotype_ns5a subtype_ns5a
genotype_ns5b subtype_ns5b‘
```

c) obtain data for an NGS experiment of interest.

```
‘list sequence --whereClause "source.name like 'NGS%' and sequenceID
not like 'PC%'" sequenceID source.name genotype subtype
genotyping_core genotyping_e1 genotyping_e2 genotyping_p7
genotyping_ns2 genotyping_ns3 genotyping_ns4a genotyping_ns4b
genotyping_ns5a genotyping_ns5b hcv_wg_pipeline pipeline_version
sample.id sample.sample_date sample.reception_date
sample.initial_genotype sample.hospital.hospital_name
sample.patient.id sample.patient.city_of_residence
sample.patient.country_of_birth sample.patient.date_of_birth
sample.patient.diagnosis_date sample.patient.ethnicity
sample.patient.gender sample.patient.nationality
sample.patient.hiv_infection sample.patient.treatment.regime
sample.patient.treatment.treatment_date sample.patient.treatment.id
sample.patient.treatment.drug.id
sample.patient.treatment.drug.manufacturer
sample.patient.treatment.drug.therapy_class ’
```

A second approach was performed using Python library Subprocess to access the project and Pandas to process results. The query to retrieve the genotyping results on recombinant viruses was split in two smaller queries and the data then stored in two CSV files in queries/tmp (Figure 4A). These files were then merged in a final CSV file by means of Pandas, saved in queries/. These processes were carried out by the script *query_recombinant_viruses.py*. Another script, *query_NGS_genotyping.py*, was created to retrieve the information from the NGS sequences after analysis with MLCA. Similar proceeding was used to retrieve the information of those sequences from a specific NGS experiment using *query_NGS_genotyping.py.* In this case, the query was split in 4 smaller queries processed in GLUE through the use of Subprocess and the CSV files, created by Pandas, were consolidated in a final CSV file that contained all the information requested (Figure 4B).

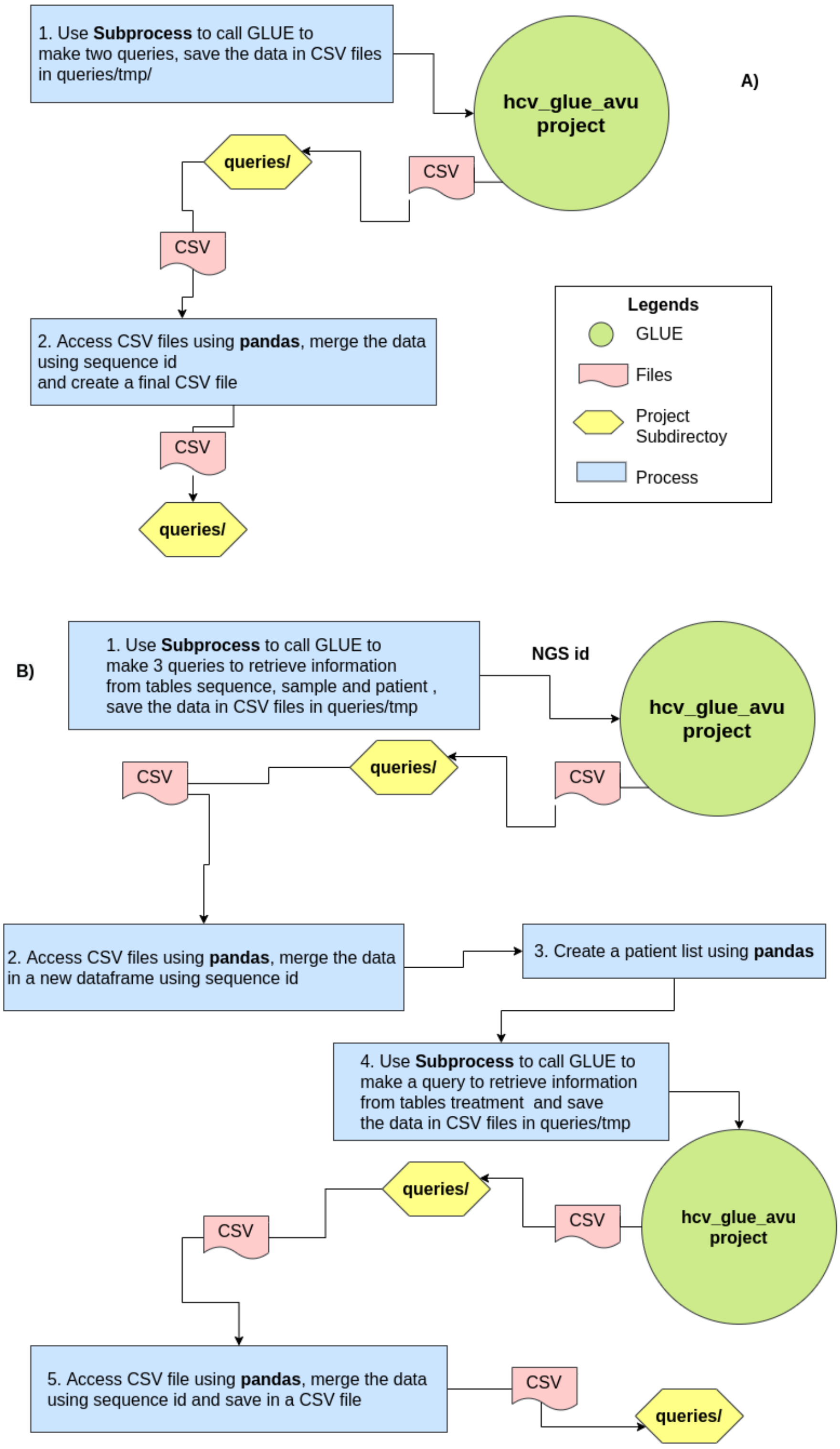
A)
1. Use **Subprocess** to call GLUE to make two queries, save the data in CSV files in queries/tmp/
hcv_glue_avu project
CSV
queries/
CSV
2. Access CSV files using **pandas**, merge the data using sequence id and create a final CSV file
CSV
queries/
Legends
GLUE
Files
Project Subdirectoy
Process
B)
1. Use **Subprocess** to call GLUE to make 3 queries to retrieve information from tables sequence, sample and patient , save the data in CSV files in queries/tmp
NGS id
hcv_glue_avu project
CSV
queries/
CSV
2. Access CSV files using **pandas**, merge the data in a new dataframe using sequence id
3. Create a patient list using **pandas**
4. Use **Subprocess** to call GLUE to make a query to retrieve information from tables treatment and save the data in CSV files in queries/tmp
hcv_glue_avu project
CSV
queries/
CSV
5. Access CSV file using **pandas**, merge the data using sequence id and save in a CSV file
CSV
queries/

Figure 4. Processes to query the project hcv_glue_avu to retrieve information about the genotyping results (A) and all the information available for a specific NGS experiment (B)

RESULTS AND DISCUSSION

The project *hcv_glue_avu* contains a database with 26 tables, 17 of them corresponding to the fixed GLUE tables (Figure 5). While the table *alignment* was modified just to add a field to capture the phylogeny data, the table *sequence* was extended with 42 new fields for the metadata of the reference sequences obtained from NCBI-GenBank (n=24), for external information not contained in the GenBank XML files of the reference sequences (n=2), metadata generated from the NGS sequences (n=3) and for results after the genotyping analysis (n=11). The extension of the schema was performed successfully by adding two sets of custom tables (4 tables referred as *who* and 5 tables specifically created to accommodate clinical and epidemiological data associated with the NGS sequences for the project). However, when a diagram was created to visualise the final schema of the project (Figure 5), links displayed the relationships between fixed GLUE tables but no visual relationship was observed between the custom tables or between these tables and *sequence*. Interestingly, when the custom tables were defined in the project, those fields that were used as foreign keys in child tables to connect to parent tables (i.e. *patient_id* in table *sample* to connect to *patient*, see *avuCustomTables.glue*) had to be omitted to allow a successful extension of the schema. Although the links between custom tables were defined (i.e. using commands like ‘`create link sample patient --multiplicity MANY_TO_ONE`’), the final relationship seemed to be created by specifying the rows of the table child and parent at the time of the data population (see lines 36-43 of *populateSampleTable.js*). The transfer of data from CSV files, which contained those fields not defined in the schema, was successfully carried out and it appears that GLUE’s implementation of the ORM using Apache Cayenne limits the setup of relationships with multiple custom table links. Other potential processes for the extension of database, such as adding multiple constrains in a custom table, may need to be tested by modifying the JavaScript programs used for the data population or, more likely, by accessing directly the MySQL database in GLUE in order to change the relationships. This last option could be a limitation as the management of the database that GLUE holds may need to be externally modified rather than using the software itself.

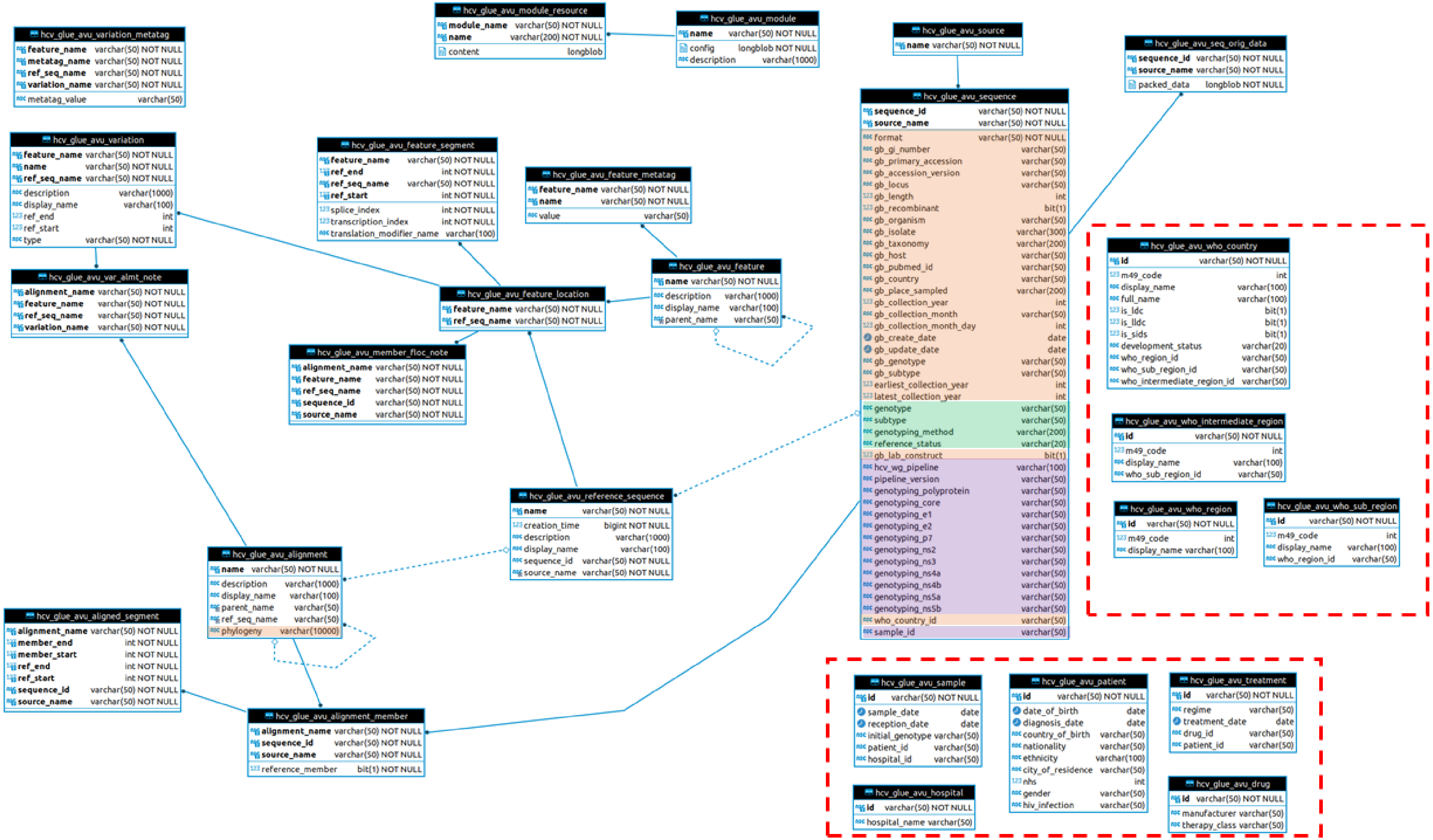
hcv_glue_avu_variation_metatag
feature_name varchar(50) NOT NULL
metatag_name varchar(50) NOT NULL
ref_seq_name varchar(50) NOT NULL
variation_name varchar(50) NOT NULL
metatag_value varchar(50)
hcv_glue_avu_module_resource
module_name varchar(50) NOT NULL
name varchar(200) NOT NULL
content longblob
hcv_glue_avu_module
name varchar(50) NOT NULL
config longblob NOT NULL
description varchar(1000)
hcv_glue_avu_source
name varchar(50) NOT NULL
hcv_glue_avu_seq_orig_data
sequence_id varchar(50) NOT NULL
source_name varchar(50) NOT NULL
packed_data longblob NOT NULL
hcv_glue_avu_variation
feature_name varchar(50) NOT NULL
name varchar(50) NOT NULL
ref_seq_name varchar(50) NOT NULL
description varchar(1000)
display_name varchar(100)
ref_end int
ref_start int
type varchar(50) NOT NULL
hcv_glue_avu_feature_segment
feature_name varchar(50) NOT NULL
ref_end int NOT NULL
ref_seq_name varchar(50) NOT NULL
ref_start int NOT NULL
splice_index int NOT NULL
transcription_index int NOT NULL
translation_modifier_name varchar(100)
hcv_glue_avu_feature_metatag
feature_name varchar(50) NOT NULL
name varchar(50) NOT NULL
value varchar(50)
hcv_glue_avu_sequence
sequence_id varchar(50) NOT NULL
source_name varchar(50) NOT NULL
format varchar(50) NOT NULL
gb_gi_number varchar(50)
gb_primary_accession varchar(50)
gb_accession_version varchar(50)
gb_locus varchar(50)
gb_length int
gb_recombinant bit(1)
gb_organism varchar(50)
gb_isolate varchar(300)
gb_taxonomy varchar(200)
gb_host varchar(50)
gb_pubmed_id varchar(50)
gb_country varchar(50)
gb_place_sampled varchar(200)
gb_collection_year int
gb_collection_month varchar(50)
gb_collection_month_day int
gb_create_date date
gb_update_date date
gb_genotype varchar(50)
gb_subtype varchar(50)
earliest_collection_year int
latest_collection_year int
genotype varchar(50)
subtype varchar(50)
genotyping_method varchar(200)
reference_status varchar(20)
gb_lab_construct bit(1)
hcv_wg_pipeline varchar(100)
pipeline_version varchar(50)
genotyping_polyprotein varchar(50)
genotyping_core varchar(50)
genotyping_e1 varchar(50)
genotyping_e2 varchar(50)
genotyping_p7 varchar(50)
genotyping_ns2 varchar(50)
genotyping_ns3 varchar(50)
genotyping_ns4a varchar(50)
genotyping_ns4b varchar(50)
genotyping_ns5a varchar(50)
genotyping_ns5b varchar(50)
who_country_id varchar(50)
sample_id varchar(50)
hcv_glue_avu_var_almt_note
alignment_name varchar(50) NOT NULL
feature_name varchar(50) NOT NULL
ref_seq_name varchar(50) NOT NULL
variation_name varchar(50) NOT NULL
hcv_glue_avu_feature_location
feature_name varchar(50) NOT NULL
ref_seq_name varchar(50) NOT NULL
hcv_glue_avu_feature
name varchar(50) NOT NULL
description varchar(1000)
display_name varchar(100)
parent_name varchar(50)
hcv_glue_avu_member_floc_note
alignment_name varchar(50) NOT NULL
feature_name varchar(50) NOT NULL
ref_seq_name varchar(50) NOT NULL
sequence_id varchar(50) NOT NULL
source_name varchar(50) NOT NULL
hcv_glue_avu_who_country
id varchar(50) NOT NULL
m49_code int
display_name varchar(100)
full_name varchar(100)
is_ldc bit(1)
is_lldc bit(1)
is_sids bit(1)
development_status varchar(20)
who_region_id varchar(50)
who_sub_region_id varchar(50)
who_intermediate_region_id varchar(50)
hcv_glue_avu_who_intermediate_region
id varchar(50) NOT NULL
m49_code int
display_name varchar(100)
who_sub_region_id varchar(50)
hcv_glue_avu_reference_sequence
name varchar(50) NOT NULL
creation_time bigint NOT NULL
description varchar(1000)
display_name varchar(100)
sequence_id varchar(50) NOT NULL
source_name varchar(50) NOT NULL
hcv_glue_avu_who_region
id varchar(50) NOT NULL
m49_code int
display_name varchar(100)
hcv_glue_avu_who_sub_region
id varchar(50) NOT NULL
m49_code int
display_name varchar(100)
who_region_id varchar(50)
hcv_glue_avu_alignment
name varchar(50) NOT NULL
description varchar(1000)
display_name varchar(100)
parent_name varchar(50)
ref_seq_name varchar(50)
phylogeny varchar(10000)
hcv_glue_avu_aligned_segment
alignment_name varchar(50) NOT NULL
member_end int NOT NULL
member_start int NOT NULL
ref_end int NOT NULL
ref_start int NOT NULL
sequence_id varchar(50) NOT NULL
source_name varchar(50) NOT NULL
hcv_glue_avu_alignment_member
alignment_name varchar(50) NOT NULL
sequence_id varchar(50) NOT NULL
source_name varchar(50) NOT NULL
reference_member bit(1) NOT NULL
hcv_glue_avu_sample
id varchar(50) NOT NULL
sample_date date
reception_date date
initial_genotype varchar(50)
patient_id varchar(50)
hospital_id varchar(50)
hcv_glue_avu_patient
id varchar(50) NOT NULL
date_of_birth date
diagnosis_date date
country_of_birth varchar(50)
nationality varchar(50)
ethnicity varchar(100)
city_of_residence varchar(50)
nhs int
gender varchar(50)
hiv_infection varchar(50)
hcv_glue_avu_treatment
id varchar(50) NOT NULL
regime varchar(50)
treatment_date date
drug_id varchar(50)
patient_id varchar(50)
hcv_glue_avu_hospital
id varchar(50) NOT NULL
hospital_name varchar(50)
hcv_glue_avu_drug
id varchar(50) NOT NULL
manufacturer varchar(50)
therapy_class varchar(50)

Figure 5. Entity Relationship Diagram showing the schema of the project and the relationships between tables. Discontinuous red lines are used to identify custom tables. Fields highlighted in orange, green and purple correspond to those used to store metadata from the sequences downloaded form NCBI-Genbank, populate data from CSV files/glue scripts and store the result from the genotyping determined by using Maximum Likelihood Clade Assigment method (MLCA)

GenBank XML and FASTA were the main formats used to transfer sequence data in the glue project. Sequences generated by NGS were added into the project using FASTA files after making two modification form the original files. GLUE prevents import of any sequence data in FASTA format if the file name and the FASTA identifier do not match. As a consequence, all sequences were modified to use a unique ID for these parameters and then imported in the project by transferring the file name, the NGS experiment and the contents of the FASTA file into fields *sequence_id*, *source_name* and *packed_data*, respectively, in the table *seq_orig_data* (Figure 5). Another restriction in GLUE is that two sequences are not allowed to have the same ID if they are located in the same object *source* (Singer et al., 2018). Although this constraint did not affect the duplicated sequence data generated from re-sequencing of 60 samples as they belonged to different experiments, the inclusion of the NGS ID in the header of the sequences not only provided more detailed information to describe the data but also was used to avoid potential issues with external alignment software if the data needed to be processed before being imported into GLUE. Duplicated sequences from the same sample were used in the project to test the population of clinical information in the database, which was successfully transferred to GLUE by using JavaScript programs that dealt with duplications. Other metadata generated from the sequences and stored in a CSV file was populated to table *sequence* by means of a module (Figure 1, step 7). By contrast, the metadata from reference sequences of the project were imported into the same table using GenBank XML format, which also was used to transfer the nucleotide sequence (Figure 1, steps 5 and 6). As the original NGS sequence data must be modified to assign the same ID before importing the sequences, the creation of the XML document may be tested in the future to consolidate the metadata and the nucleotide sequence from NGS FASTA files in a single file that could be used to perform two different steps of the project, reducing the number of files in the project directory and simplifying the process.

GLUE software is designed to provide tools to perform common operations on sequences such as alignments or phylogenetic tress, using for example MAFFT or RaxML through the specific modules *maffAligner* and *raxmlPhylogenyGenerator* (see section Module type reference of the GLUE documentation[3]). However, the user is also allowed to use alternative software packages to perform the same operations. In the project *hcv_glue_avu*, the choice of using the online tool HCVAlign[2] created an alignment with all the reference sequences in frame as gaps are introduced to compensate potential frame shifts. This approach made it easier to identify indels present in some reference sequences and the modification of the blocks of nucleotides that were going to be used to infer the feature locations from the master reference sequence (see *InheritFeatureLocationsFromNC_004102.glue*). FastTree2, used to generate the phylogenetic tree, was an additional modification to the standard processing in GLUE. Although the annotations in the tree needed to be modified for the data to work in GLUE (Figure 1, steps 16 and 17), the process is less time consuming than using RaxML and it is routinely used for the work on HCV and other viruses in AVU (Bradshaw et al., 2021; Ledesma et al., 2019; Mbisa et al., 2019). A script, called *Generate_Reference_Phylogeny_RAXML.glue*, which used the standard operations using RaxML has however been kept in the project, in case other usera prefer the default option.

Table 2. Genotyping result of the selected recombinant viruses using the MLCA method and 11 genomic regions. No assignment of a genotype/subtype is represented by an empty field. The nucleotide positions covered for each gene analysed (displayed in brackets) and the coordinates of the breakpoints are in reference to the sequence H77.

ID	Recombinant form	Reference	Breakpoints	Precursor Polyprotein	Core	E1	E2	P7	NS2	NS3	NS4A	NS4B	NS5A	NS5B
AB622121	2b/1b	(Yokoyama et al., 2011)	3432	1	2b (342-914)	2b (915-1490)	2b (1491-2591)	2b (2592-2780)	2b (2781-3431)	1b (3432-5324)	1b (5325-5486)	1b (5487-6269)	1b (6270-7610)	1b (7611-9386)
AB677527	2b/1b	(Hoshino et al., 2012)	3286-3293	1	2b (342-914)	2b (915-1490)	2b (1491-2591)	2b (2592-2780)	2b (2781-3431)	1b (3432-5324)	1b (5325-5486)	1b (5487-6269)	1b (6270-7610)	1b (7611-9386)
AB677530	2b/1b	(Hoshino et al., 2012)	3286-3293	1b	2b (271-843)	2b (844-1419)	2b (1420-2520)	2b (2521-2709)	2b (2710-3360)	1b (3361-5253)	1b (5254-5415)	1b (5416-6198)	1b (6199-7539)	1b (7540-9315)
AM408911	2/5	(Legrand-Abravanel et al., 2007)	3366-3389		2j (319-891)	2 (892-1467)	2 (1468-2568)	2_unassigned_JF735117 (2569-2757)	2 (2758-3408)	(3409-5301)	5 (5302-5463)	5 (5464-6246)	5 (6247-7599)	5 (7600-9246)
AY587845	2k/1b	(Kalinina et al., 2002, 2004; Kurbanov et al., 2010)	3186	1b	2k (269-841)	2k (842-1417)	2k (1418-2518)	2k (2519-2707)	2 (2708-3358)	1b (3359-5251)	1b (5252-5413)	1b (5414-6196)	1b (6197-7537)	1b (7538-9313)
DQ155560	2i/6p	(Noppornpanth et al., 2006)	3405-3464	6	2i (324-896)	2i (897-1475)	2i (1476-2582)	2i (2583-2771)	2i (2772-3422)	6p (3423-5315)	6p (5316-5477)	6p (5478-6260)	6p (6261-7616)	6p (7617-9360)
DQ364460	2b/1b	(Noppornpanth et al., 2006)	3456	1	2b (274-846)	2b (847-1422)	2b (1423-2523)	2b (2524-2712)	2b (2713-3363)	1b (3364-5256)	1b (5257-5418)	1b (5419-6201)	1b (6202-7542)	1b (7543-9315)
EU643835	2b/6w	(Lee et al., 2010)	3429	6w	2b (66-638)	2b (639-1214)	2b (1215-2315)	2b (2316-2504)	2b (2505-3155)	6w (3156-5048)	6w (5049-5210)	6w (5211-5993)	6w (5994-7346)	6w (7347-9122)
JF779679	2b/1a	(Bhattacharya et al., 2011)	3429-3440	1a	2b (74-646)	2b (647-1222)	2b (1223-2323)	2b (2324-2512)	2b (2513-3163)	1a (3164-5056)	1a (5057-5218)	1a (5219-6001)	1a (6002-7345)	1a (7346-9028)

The analysis using MLCA method on selected recombinant viruses showed that whereas the results of the classification using the coding regions for the 10 mature proteins were concordant with the identification of recombinant forms in the publications (Table 2), the use of the precursor polyprotein did not result in a correct characterization of the genotypes and subtypes of the viruses. For instance, the method did not classify the sequence AM408911 using the whole precursor but it did assign genotype 5 and 2 for the same sequence using individual the set of genes, as initially reported (Legrand-Abravanel et al., 2007). In the same sequence, the Core gene was subtyped as 2j but the method could not characterise the NS3 gene (Table 2). Although the MLCA method provided by GLUE, with a customed configuration of the JavaScript program, works at detecting recombination in genomes, inferring the recombination breakpoints will still need to be confirmed by other software packages such as SimPlot or Recombination Analysis Too (Etherington et al., 2005; Lole et al., 1999; Salminen et al., 2009).

The genotyping of the NGS sequences showed that 355 sequences had matching results using 10 genes, the most frequent subtypes being 1a and 3a (Table 3). An additional 35 sequences presented a conclusive clade assignation. However, the classification for some sequences was not based on the entire set of genes being missing for some genomic regions (see 1c, 2k, 3a or 3h in Table 3). Forty sequences showed a sub-classification matching most of the genes but with other regions only being classified at the level of genotype, which matched the results for subtyping in the other genes (see for instance 1a1 or 3a3 in Table 3). The presence of undetermined nucleotides and ambiguities in sequences can affect the final outcome of phylogenetic analysis (Lemmon et al., 2009), which could explain the results for those sequences with regions failing to be classified or just identified at the genotype level. All the genotyping results were consolidated in the fields *genotype* and *subtype* in table *sequence*, being a comment added for the problematic sequences so that they can be further investigated and validated.

Interestingly, 5 recombinant viruses were detected using the classification method. Two inter-genotype recombinant forms were detected; 1a/3a/3, in sequence H211900006-1_NGS99 , and 2k/2/1b, in H211040784-1_NGS95 and H210980721-1_NGS95. Although infections with HCV recombinant viruses are rare, the form 2k/1b has been frequently found in Europe in countries like Italy Ireland, Cyprus, Estonia or Russia (Demetriou et al., 2011; Kalinina et al., 2002; Kurbanov et al., 2008; Moreau et al., 2006; Paolucci et al., 2017a; Tallo et al., 2007; Viazov et al., 2010) The recombinant form 1a/3a seems to be very infrequent with a single case having been detected in Russia (Viazov et al., 2010). Additionally, intra-genotype recombinant forms, 1c/1a and 1c/1/1a, were also detected in sequences H191760561-1_NGS93 and H191760561-1_NGS94, respectively. A recombinant virus 1a/1c with several breakpoints in its full length genome has previously been described in a patient from India (Ross et al., 2008). However, the identification of the recombinant 1c/1a described in the current work was based only on 5 and 6 genes out of the total set used for the analysis. This sample should be re-sequenced to confirm the results.

Furthermore, a full genetic characterization of all the recombinant viruses detected in this project should be performed to find the breakpoints of the recombination and the clinical and the epidemiological information should also be collected. The full description of the results of the genotyping on the 440 sequences can be found in the file NGSsequences_genotyping_results.csv at https://github.com/juanledesma78/HCV-GLUE-AVU/tree/main/queries.

Table 3. Genotyping result of the NGS sequences using the MLCA method and 10 genomic regions. Empty fields were used for those cases were no classification was achieved. Dots were used to identify similar subtype or genotype assignments for each gene.

Genotyping according to genes	Frequency	Core	E1	E2	P7	NS2	NS3	NS4A	NS4B	NS5A	NS5B
1a	176	1a	1a	1a	1a	1a	1a	1a	1a	1a	1a
1a	1	·	·	·	·	·	·	·		·	·
1a	5	·	·	·		·	·	·	·	·	·
1a	1	·	·		·	·	·	·	·	·	·
1a	1	·		·	·		·	·	·	·	·
1a	1	·		·		·	·	·	·	·	·
1a	1	·		·			·	·		·	·
1a	1	·			·	·	·	·	·	·	·
1a	1	·				·	·			·	·
1a	1		·	·	·	·	·	·	·	·	·
1a	1		·	·		·	·	·	·	·	·
1a	2			·	·	·	·	·	·	·	·
1a	1			·		·	·			·	·
1a	1						·	·		·	·
1a 1	1				·	1	·	·		·	·
1a 1	1				·	·	·	·		1	1
1a 1	1						·	·	·	1	1
1a 1	2							·	1		
1a 1	1	·		1		·	·	·	·	·	·
1a 1	1	·	1	·	·	·	·	·	·	·	·
1a 1	1	·	·	·	1	1	·	·		·	·
1a 1	1	·	·	·	1	·	·	·	·	·	·
1a 1	5	·	·	·	·	·	·	1	·	·	·
1a 1	1		·	·	·	·	·	·	1	·	·
1 1a	1		1		·	1	·	·		·	·
1 1a	1	1	·	·	·	·	·	1	·	·	·
1 1a	9	1	·	·	·	·	·	·	·	·	·
1b	12	1b	1b	1b	1b	1b	1b	1b	1b	1b	1b
1b	1		·	·	·	·	·	·	·	·	·
1b	1				·	·	·	·	·	·	·
1b	1						·	·			
1b 1	1	·	·	·	·	·	·	1	·	·	·
1 1b	5	1	·	·	·	·	·	·	·	·	·
1c	1	1c	1c	1c	1c	1c	1c	1c	1c	1c	
1c 1	1	·	·	1	1	1	1	1	·	·	·
1o 1	1	1o	1	1o	1	1o	1	1o	1o	1o	1o
2a	1	2a	2a	2a	2a	2a	2a	2a	2a	2a	2a
2b	12	2b	2b	2b	2b	2b	2b	2b	2b	2b	2b

2b	1		·		·	·	·	·	·		
2c	1	2c	2c	2c	2c	2c	2c	2c	2c	2c	2c
2k	1	2k	2k								
3a	137	3a	3a	3a	3a	3a	3a	3a	3a	3a	3a
3a	1	·		·	·	·	·	·	·	·	·
3a	1	·		·	·		·	·		·	·
3a	1	·		·		·	·	·	·	·	·
3a	2		·	·	·	·	·	·	·	·	·
3a	1			·	·	·	·		·	·	·
3a	1			·						·	·
3a	1				·			·	·		
3a	1					·		·			
3a	1								·		
3a 3	1	·	3	·	·	·	·	·	·	·	·
3a 3	1	·	·	3	·	·	·		·	·	·
3a 3	1	·		·	·	·	·	3	·	·	·
3a 3	1		·	3		·	·		·	·	·
3a 3	1				·		3	·	·	·	·
3 3a	1			3		·	·		·	·	·
3b	2	3b	3b	3b	3b	3b	3b	3b	3b	3b	3b
3h	1	3h		3h			3h	3h	3h	3h	3h
4	1	4	4	4	4	4	4	4	4	4	4
4d	3	4d	4d	4d	4d	4d	4d	4d	4d	4d	4d
4k	2	4k	4k	4k	4k	4k	4k	4k	4k	4k	4k
4r	2	4r	4r	4r	4r	4r	4r	4r	4r	4r	4r
4w	1	4w	4w	4w	4w	4w	4w	4w	4w	4w	4w
5a	2	5a	5a	5a	5a	5a	5a	5a	5a	5a	5a
6f	2	6f	6f	6f	6f	6f	6f	6f	6f	6f	6f
6n	1	6n	6n	6n	6n	6n	6n	6n	6n	6n	6n
RF 1a 3a 3	1	1a	1a	3a	3a	3	1a	1a	3a	3a	3a
RF 1c 1 1a	1	1c		1			1a		1a	1a	1a
RF 1c 1a	1	1c					1a		1a	1a	1a
RF 2k 2 1b	2	2k	2k	2k	2k	2	1b	1b	1b	1b	1b
No results	5										
Total	440										

Limitations were found for the GLUE database. Exceptions were returned when the specific GLUE commands were used to query the database (Table A2), two of them happening when multiple custom tables that contained clinical and epidemiological data and MANY-TO-ONE relationships were queried simultaneously. Although the schema was properly defined according to the GLUE documentation[3] by the use of the lines ‘`create link sequence sample --multiplicity MANY_TO_ONE`’, ‘`create link sample patient --multiplicity MANY_TO_ONE`’, ‘`create link sample hospital --multiplicity MANY_TO_ONE`’, ‘`create link treatment drug --multiplicity MANY_TO_ONE`’ and ‘`create link patient treatment --multiplicity ONE_TO_MANY`’ in the script *avuCustomTables.glue*, the exceptions were only returned when three consecutive related tables were queried, using, for example, the GLUE command `sample.patient.treatment` to retrieve information of the treatment of the patients. Therefore, the use of python scripts had to be adapted to test the queries and retrieve the data needed from the GLUE project (see folder queries/). This limitation on the use of custom tables, confirmed by the developers after consultations, constitutes a potential issue if additional custom tables related to the previous ones are required to extend the project *hcv_glue_avu* in the future. A solution to explore could be the creation of an independent MySQL database, using SQL Alchemy, that defines all the custom tables needed for the project and respective relationships. The base tables in this database could then be accessed by GLUE, while full queries that link the data from the custom tables and the fixed GLUE tables could be conducted from SQL Alchemy.

CONCLUSION

An initial GLUE project has been tested to perform routine analysis on sequence data generated in AVU and to store important clinical and epidemiological information for patient management. The project provides a powerful tool to perform the subtyping of the sequences and can be used in parallel with the routine methods to deliver those duties in the lab. The *hcv_glue_avu* project, which is intended to be a dynamic tool requiring updates with new sequences as the HCV classification changes and the development of a web interface, is planned to be extended with another tool for the antiviral resistance genotyping, developed as a result of the collaboration between PHE and the University of Glasgow[4]. However, further work must also be explored as the current project has identified limitations when using the database in GLUE to store clinical and epidemiological data for the use in the laboratory.

Footnotes

1. https://talk.ictvonline.org/ictv_wikis/flaviviridae/hepacivirus/m/hepacivirus-files/8280
2. https://hcv.lanl.gov/content/sequence/VIRALIGN/viralign.html
3. https://sqlite.org/index.html
4. http://glue-tools.cvr.gla.ac.uk
5. https://github.com/giffordlabcvr/PHE-HCV-DRUG-RESISTANCE

REFERENCES

Aceijas, C., Rhodes, T., 2007. Global estimates of prevalence of HCV infection among injecting drug users. Int. J. Drug Policy 18, 352–358. https://doi.org/10.1016/J.DRUGPO.2007.04.004

Almeida Calado, R., Rocha, M.R., Parreira, R., Piedade, J., Venenno, T., Esteves, A., 2011. Hepatitis C virus subtypes circulating among intravenous drug users in Lisbon, Portugal. J. Med. Virol. 83, 608–615. https://doi.org/10.1002/JMV.21955

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., Palù, G., 2011. Applications of Next-Generation Sequencing Technologies to Diagnostic Virology. OPEN ACCESS Int. J. Mol. Sci 12, 12. https://doi.org/10.3390/ijms12117861

Baumert, T.F., Berg, T., Lim, J.K., Nelson, D.R., 2019. Status of Direct-Acting Antiviral Therapy for Hepatitis C Virus Infection and Remaining Challenges. Gastroenterology 156, 431–445. https://doi.org/10.1053/J.GASTRO.2018.10.024

Benova, L., Mohamoud, Y.A., Calvert, C., Abu-Raddad, L.J., 2014. Vertical Transmission of Hepatitis C Virus: Systematic Review and Meta-analysis. Clin. Infect. Dis. 59, 765–773. https://doi.org/10.1093/CID/CIU447

Bhattacharya, D., Accola, M.A., Ansari, I.H., Striker, R., Rehrauer, W.M., 2011. Naturally occurring genotype 2b/1a hepatitis C virus in the United States. Virol. J. 8. https://doi.org/10.1186/1743-422X-8-458

Blach, S., Zeuzem, S., Manns, M., Altraif, I., Duberg, A.S., Muljono, D.H., Waked, I., Alavian, S.M., Lee, M.H., Negro, F., Abaalkhail, F., Abdou, A., Abdulla, M., Abou Rached, A., Aho, I., Akarca, U., Al Ghazzawi, I., Al Kaabi, S., Al Lawati, F., Al Namaani, K., Al Serkal, Y., Al-Busafi, S.A., Al-Dabal, L., Aleman, S., Alghamdi, A.S., Aljumah, A.A., Al-Romaihi, H.E., Andersson, M.I., Arendt, V., Arkkila, P., Assiri, A.M., Baatarkhuu, O., Bane, A., Ben-Ari, Z., Bergin, C., Bessone, F., Bihl, F., Bizri, A.R., Blachier, M., Blasco, A.J., Brandao Mello, C.E., Bruggmann, P., Brunton, C.R., Calinas, F., Chan, H.L.Y., Chaudhry, A., Cheinquer, H., Chen, C.J., Chien, R.N., Choi, M.S., Christensen, P.B., Chuang, W.L., Chulanov, V., Cisneros, L., Clausen, M.R., Cramp, M.E., Craxi, A., Croes, E.A., Dalgard, O., Daruich, J.R., De Ledinghen, V., Dore, G.J., El-Sayed, M.H., Ergor, G., Esmat, G., Estes, C., Falconer, K., Farag, E., Ferraz, M.L.G., Ferreira, P.R., Flisiak, R., Frankova, S., Gamkrelidze, I., Gane, E., Garcia-Samaniego, J., Khan, A.G., Gountas, I., Goldis, A., Gottfredsson, M., Grebely, J., Gschwantler, M., Guimaraes Pessoa, M., Gunter, J., Hajarizadeh, B., Hajelssedig, O., Hamid, S., Hamoudi, W., Hatzakis, A., Himatt, S.M., Hofer, H., Hrstic, I., Hui, Y.T., Hunyady, B., Idilman, R., Jafri, W., Jahis, R., Janjua, N.Z., Jarčuška, P., Jeruma, A., Jonasson, J.G., Kamel, Y., Kao, J.H., Kaymakoglu, S., Kershenobich, D., Khamis, J., Kim, Y.S., Kondili, L., Koutoubi, Z., Krajden, M., Krarup, H., Lai, M.S., Laleman, W., Lao, W.C., Lavanchy, D., Lazaro, P., Leleu, H., Lesi, O., Lesmana, L.A., Li, M., Liakina, V., Lim, Y.S., Luksic, B., Mahomed, A., Maimets, M., Makara, M., Malu, A.O., Marinho, R.T., Marotta, P., Mauss, S., Memon, M.S., Mendes Correa, M.C., Mendez-Sanchez, N., Merat, S., Metwally, A.M., Mohamed, R., Moreno, C., Mourad, F.H., Mullhaupt, B., Murphy, K., Nde, H., Njouom, R., Nonkovic, D., Norris, S., Obekpa, S., Oguche, S., Olafsson, S., Oltman, M., Omede, O., Omuemu, C., Opare-Sem, O., Ovrehus, A.L.H., Owusu-Ofori, S., Oyunsuren, T.S., Papatheodoridis, G., Pasini, K., Peltekian, K.M., Phillips, R.O., Pimenov, N., Poustchi, H., Prabdial-Sing, N., Qureshi, H., Ramji, A., Razavi-Shearer, D., Razavi-Shearer, K., Redae, B., Reesink, H.W., Ridruejo, E., Robbins, S., Roberts, L.R., Roberts, S.K., Rosenberg, W.M., Roudot-Thoraval, F., Ryder, S.D., Safadi, R., Sagalova, O., Salupere, R., Sanai, F.M., Sanchez Avila, J.F., Saraswat, V., Sarmento-Castro, R., Sarrazin, C., Schmelzer, J.D., Schreter, I., Seguin-Devaux, C., Shah, S.R., Sharara, A.I., Sharma, M., Shevaldin, A., Shiha, G.E., Sievert, W., Sonderup, M., Souliotis, K., Speiciene, D., Sperl, J., Starkel, P., Stauber, R.E., Stedman, C., Struck, D., Su, T.H., Sypsa, V., Tan, S.S., Tanaka, J., Thompson, A.J., Tolmane, I., Tomasiewicz, K., Valantinas, J., Van Damme, P., Van Der Meer, A.J., Van Thiel, I., Van Vlierberghe, H., Vince, A., Vogel, W., Wedemeyer, H., Weis, N., Wong, V.W.S., Yaghi, C., Yosry, A., Yuen, M.F., Yunihastuti, E., Yusuf, A., Zuckerman, E., Razavi, H., 2017. Global prevalence and genotype distribution of hepatitis C virus infection in 2015: A modelling study. Lancet Gastroenterol. Hepatol. 2, 161–176. https://doi.org/10.1016/S2468-1253(16)30181-9

Bogner, P., Capua, I., Lipman, D.J., Cox, N.J., 2006. A global initiative on sharing avian flu data. Nat. 2006 4427106 442, 981–981. https://doi.org/10.1038/442981a

Borgia, S.M., Hedskog, C., Parhy, B., Hyland, R.H., Stamm, L.M., Brainard, D.M., Subramanian, M.G., McHutchison, J.G., Mo, H., Svarovskaia, E., Shafran, S.D., 2018. Identification of a novel hepatitis C virus genotype from Punjab, India: Expanding classification of hepatitis C virus into 8 genotypes. J. Infect. Dis. 218, 1722–1729. https://doi.org/10.1093/infdis/jiy401

Bradshaw, D., Bibby, D.F., Manso, C.F., Piorkowska, R., Mohamed, H., Ledesma, J., Bubba, L., Chan, Y.T., Ngui, S.L., Carne, S., Mbisa, J.L., 2021. Clinical evaluation of a Hepatitis C Virus whole-genome sequencing pipeline for genotyping and resistance testing. Clin. Microbiol. Infect. 0. https://doi.org/10.1016/J.CMI.2021.06.042

Chan, D.P.C., Sun, H.-Y., Wong, H.T.H., Lee, S.-S., Hung, C.-C., 2016. Sexually acquired hepatitis C virus infection: a review. Int. J. Infect. Dis. 49, 47–58. https://doi.org/10.1016/J.IJID.2016.05.030

Chiara, M., D’Erchia, A.M., Gissi, C., Manzari, C., Parisi, A., Resta, N., Zambelli, F., Picardi, E., Pavesi, G., Horner, D.S., Pesole, G., 2021. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Brief. Bioinform. 22, 616–630. https://doi.org/10.1093/BIB/BBAA297

Colina, R., Casane, D., Vasquez, S., García-Aguirre, L., Chunga, A., Romero, H., Khan, B., Cristina, J., 2004. Evidence of intratypic recombination in natural populations of hepatitis C virus. J. Gen. Virol. 85, 31–37. https://doi.org/10.1099/VIR.0.19472-0

Demetriou, V.L., Kyriakou, E., Kostrikis, L.G., 2011. Near-full genome characterisation of two natural intergenotypic 2k/1b recombinant hepatitis c virus isolates. Adv. Virol. 2011. https://doi.org/10.1155/2011/710438

Dennis, T.P.W., de Souza, W.M., Marsile-Medun, S., Singer, J.B., Wilson, S.J., Gifford, R.J., 2019. The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. Virus Res. 262, 15–23. https://doi.org/10.1016/j.virusres.2018.03.014

Di Stefano, M., Faleo, G., Mohamed Farhan Mohamed, A., Morella, S., Rita Bruno, S., Tundo, P., Ramon Fiore, J., Antonia Santantonio, T., 2021. Resistance Associated Mutations in HCV Patients Failing DAA Treatment. New Microbiol. 44, 12–18.

Dubuisson, J., Cosset, F.L., 2014. Virology and cell biology of the hepatitis C virus life cycle - An update. J. Hepatol. https://doi.org/10.1016/j.jhep.2014.06.031

Etherington, G.J., Dicks, J., Roberts, I.N., 2005. Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. Bioinformatics 21, 278–281. https://doi.org/10.1093/BIOINFORMATICS/BTH500

Gaspareto, K.V., Ribeiro, R.M., de Mello Malta, F., Gomes-Gouvêa, M.S., Muto, N.H., Mendes-Correa, M.C., Rozanski, A., Carrilho, F.J., Sabino, E.C., Pinho, J.R.R., 2016. HCV inter-subtype 1a/1b recombinant detected by complete-genome next-generation sequencing. Arch. Virol. 161, 2161–2168. https://doi.org/10.1007/S00705-016-2889-5

González-Grande, R., Jiménez-Pérez, M., Arjona, C.G., Torres, J.M., 2016. New approaches in the treatment of hepatitis C. World J. Gastroenterol. 22, 1421. https://doi.org/10.3748/WJG.V22.I4.1421

Gosert, R., Egger, D., Lohmann, V., Bartenschlager, R., Blum, H.E., Bienz, K., Moradpour, D., 2003. Identification of the Hepatitis C Virus RNA Replication Complex in Huh-7 Cells Harboring Subgenomic Replicons. J. Virol. 77, 5487–5492. https://doi.org/10.1128/JVI.77.9.5487-5492.2003

Hajarizadeh, B., Grebely, J., Dore, G.J., 2013. Epidemiology and natural history of HCV infection. Nat. Rev. Gastroenterol. Hepatol. https://doi.org/10.1038/nrgastro.2013.107

Hedskog, C., Parhy, B., Chang, S., Zeuzem, S., Moreno, C., Shafran, S.D., Borgia, S.M., Asselah, T., Alric, L., Abergel, A., Chen, J.J., Collier, J., Kapoor, D., Hyland, R.H., Simmonds, P., Mo, H., Svarovskaia, E.S., 2019. Identification of

19 novel hepatitis C virus subtypes-further expanding HCV classification. Open Forum Infect. Dis. 6. https://doi.org/10.1093/ofid/ofz076

Hoshino, H., Hino, K., Miyakawa, H., Takahashi, K., Akbar, S.M.F., Mishiro, S., 2012. Inter-genotypic recombinant hepatitis C virus strains in Japan noted by discrepancies between immunoassay and sequencing. J. Med. Virol. 84, 1018–1024. https://doi.org/10.1002/jmv.23300

Jordan, A.E., Perlman, D.C., Neurer, J., Smith, D.J., Jarlais, D.C. Des, Hagan, H., 2016. Prevalence of hepatitis C virus infection among HIV+ men who have sex with men: a systematic review and meta-analysis: http://dx.doi.org/10.1177/0956462416630910 28, 145–159. https://doi.org/10.1177/0956462416630910

Kageyama, S., Agdamag, D.M., Alesna, E.T., Leaño, P.S., Heredia, A.M.L., Abellanosa-Tac-An, I.P., Jereza, L.D., Tanimoto, T., Yamamura, J., Ichimura, H., 2006. A natural inter-genotypic (2b/1b) recombinant of hepatitis C virus in the Philippines. J. Med. Virol. 78, 1423–1428. https://doi.org/10.1002/JMV.20714

Kalaghatgi, P., Sikorski, A.M., Knops, E., Rupp, D., Sierra, S., Heger, E., Neumann-Fraune, M., Beggel, B., Walker, A., Timm, J., Walter, H., Obermeier, M., Kaiser, R., Bartenschlager, R., Lengauer, T., 2016. Geno2pheno[HCV] – A Web-based Interpretation System to Support Hepatitis C Treatment Decisions in the Era of Direct-Acting Antiviral Agents. PLoS One 11, e0155869. https://doi.org/10.1371/JOURNAL.PONE.0155869

Kalinina, O., Norder, H., Mukomolov, S., Magnius, L.O., 2002. A Natural Intergenotypic Recombinant of Hepatitis C Virus Identified in St. Petersburg. J. Virol. 76, 4034–4043. https://doi.org/10.1128/JVI.76.8.4034-4043.2002

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066. https://doi.org/10.1093/NAR/GKF436

Kim, C.W., Chang, K.-M., 2013. Hepatitis C virus: virology and life cycle. Clin. Mol. Hepatol. 19, 17–25. https://doi.org/10.3350/CMH.2013.19.1.17

Kuiken, C., Combet, C., Bukh, J., Shin-I, T., Deleage, G., Mizokami, M., Richardson, R., Sablon, E., Yusim, K., Pawlotsky, J.-M., Simmonds, P., 2006. A comprehensive system for consistent numbering of HCV sequences, proteins and epitopes. Hepatology 44, 1355–1361. https://doi.org/10.1002/HEP.21377

Kuiken, C., Korber, B., Shafer, R.W., 2003. HIV Sequence Databases. AIDS Rev. 5, 52.

Kuiken, C., Yusim, K., Boykin, L., Richardson, R., 2005. The Los Alamos hepatitis C sequence database. Bioinformatics 21, 379–384. https://doi.org/10.1093/BIOINFORMATICS/BTH485

Kurbanov, F., Tanaka, Y., Avazova, D., Khan, A., Sugauchi, F., Kan, N., Kurbanova-Khudayberganova, D., Khikmatullaeva, A., Musabaev, E., Mizokami, M., 2008. Detection of hepatitis C virus natural recombinant RF1_2k/1b strain among intravenous drug users in Uzbekistan. Hepatol. Res. 38, 457–464. https://doi.org/10.1111/J.1872-034X.2007.00293.X

Lagging, M., Wejstål, R., Duberg, A.-S., Aleman, S., Weiland, O., Westin, J., Group, ; for the Swedish Consensus, 2018. Treatment of hepatitis C virus infection for adults and children: updated Swedish consensus guidelines 2017. https://doi.org/10.1080/23744235.2018.1445281 50, 569–583. https://doi.org/10.1080/23744235.2018.1445281

Lanini, S., Ustianowski, A., Pisapia, R., Zumla, A., Ippolito, G., 2019. Viral Hepatitis: Etiology, Epidemiology, Transmission, Diagnostics, Treatment, and Prevention. Infect. Dis. Clin. North Am. 33, 1045–1062. https://doi.org/10.1016/J.IDC.2019.08.004

Ledesma, J., Williams, D., Stanford, F.A., Hewitt, P.E., Zuckerman, M., Bansal, S., Dhawan, A., Mbisa, J.L., Tedder, R., Ijaz, S., 2019. Resolution by deep sequencing of a dual hepatitis E virus infection transmitted via blood components. J. Gen. Virol. 100. https://doi.org/10.1099/jgv.0.001302

Lee, Y.-M., Lin, H.-J., Chen, Y.-J., Lee, C.-M., Wang, S.-F., Chang, K.-Y., Chen, T.-L., Liu, H.-F., Chen, Y.-M.A., 2010. Molecular epidemiology of HCV genotypes among injection drug users in Taiwan: Full-length sequences of two new subtype 6w strains and a recombinant form_2b6w. J. Med. Virol. 82, 57–68. https://doi.org/10.1002/JMV.21658

Legrand-Abravanel, F., Claudinon, J., Nicot, F., Dubois, M., Chapuy-Regaud, S., Sandres-Saune, K., Pasquier, C., Izopet, J., 2007. New Natural Intergenotypic (2/5) Recombinant of Hepatitis C Virus. J. Virol. 81, 4357–4362. https://doi.org/10.1128/JVI.02639-06

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference. Syst. Biol. 58, 130–145. https://doi.org/10.1093/sysbio/syp017

Lindenbach, B.D., 2013. Virion assembly and release. Curr. Top. Microbiol. Immunol. 369, 199–218. https://doi.org/10.1007/978-3-642-27340-7_8

Lingala, S., Ghany, M.G., 2015. Natural History of Hepatitis C. Gastroenterol. Clin. North Am. 44, 717–734. https://doi.org/10.1016/J.GTC.2015.07.003

Liu, T.F., Shafer, R.W., 2006. Web Resources for HIV Type 1 Genotypic-Resistance Test Interpretation. Clin. Infect. Dis. 42, 1608–1618. https://doi.org/10.1086/503914

Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., Ray, S.C., 1999. Full-Length Human Immunodeficiency Virus Type 1 Genomes from Subtype C-Infected Seroconverters in India, with Evidence of Intersubtype Recombination. J. Virol. 73, 152–160. https://doi.org/10.1128/JVI.73.1.152-160.1999

Manso, C.F., Bibby, D.F., Lythgow, K., Mohamed, H., Myers, R., Williams, D., Piorkowska, R., Chan, Y.T., Bowden, R., Ansari, M.A., Ip, C.L.C., Barnes, E., Bradshaw, D., Mbisa, J.L., 2020. Technical Validation of a Hepatitis C Virus Whole Genome Sequencing Assay for Detection of Genotype and Antiviral Resistance in the Clinical Pathway. Front. Microbiol. https://doi.org/10.3389/fmicb.2020.576572

Mavilia, M.G., Wu, G.Y., 2017. Mechanisms and Prevention of Vertical Transmission in Chronic Viral Hepatitis. http://www.xiahepublishing.com/ 5, 119–129. https://doi.org/10.14218/JCTH.2016.00067

Maxmen, A., 2021. One million coronavirus sequences: popular genome site hits mega milestone. Nature 593, 21. https://doi.org/10.1038/D41586-021-01069-W

Mbisa, J.L., Kirwan, P., Tostevin, A., Ledesma, J., Bibby, D.F., Brown, A., Myers, R., Hassan, A.S., Murphy, G., Asboe, D., Pozniak, A., Kirk, S., Gill, O.N., Sabin, C., Delpech, V., Dunn, D.T., Database, U.H.D.R., Asboe, D., Pozniak, A., Cane, P., Chadwick, D., Churchill, D., Clark, D., Collins, S., Delpech, V., Douthwaite, S., Dunn, D., Fearnhill, E., Porter, K., Tostevin, A., Stirrup, O., Fraser, C., Geretti, A.M., Gunson, R., Hale, A., Hué, S., Lazarus, L., Leigh-Brown, A., Mbisa, T., Mackie, N., Orkin, C., Nastouli, E., Pillay, D., Phillips, A., Sabin, C., Smit, E., Templeton, K., Tilston, P., Volz, E., Williams, I., Zhang, H., Dunn, D., Fairbrother, K., Fearnhill, E., Porter, K., Tostevin, A., Stirrup, O., Dawkins, J., O'Shea, S., Mullen, J., Smit, E., Mbisa, T., Cox, A., Tandy, R., Fawcett, T., Hopkins, M., Tilston, P., Booth, C., Garcia-Diaz, A., Renwick, L., Schmid, M.L., Payne, B., Chadwick, D., Hubb, J., Dustan, S., Kirk, S., Gunson, R., Bradley-Stewart, A., 2019. Determining the Origins of Human Immunodeficiency Virus Type 1 Drug-resistant Minority Variants in People Who Are Recently Infected Using Phylogenetic Reconstruction. Clin. Infect. Dis. 69, 1136–1143. https://doi.org/10.1093/CID/CIY1048

Messina, J.P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G.S., Pybus, O.G., Barnes, E., 2015. Global distribution and prevalence of hepatitis C virus genotypes. Hepatology 61, 77–87. https://doi.org/10.1002/hep.27259

Missiha, S.B., Ostrowski, M., Heathcote, E.J., 2008. Disease Progression in Chronic Hepatitis C: Modifiable and Nonmodifiable Factors. Gastroenterology 134, 1699–1714. https://doi.org/10.1053/J.GASTRO.2008.02.069

Moradpour, D., Penin, F., 2013. Hepatitis C virus proteins: From structure to function. Curr. Top. Microbiol. Immunol. 369, 113–142. https://doi.org/10.1007/978-3-642-27340-7_5

Moreau, I., Hegarty, S., Levis, J., Sheehy, P., Crosbie, O., Kenny-Walsh, E., Fanning, L.J., 2006. Serendipitous identification of natural Intergenotypic recombinants of hepatitis C in Ireland. Virol. J. 2006 31 3, 1–7. https://doi.org/10.1186/1743-422X-3-95

Morel, V., Ghoubra, F., Izquierdo, L., Martin, E., Oliveira, C., François, C., Brochot, E., Helle, F., Duverlie, G., Castelain, S., 2016. Phylogenetic analysis of a circulating hepatitis C virus recombinant strain 1b/1a isolated in a French hospital centre. Infect. Genet. Evol. 40, 374–380. https://doi.org/10.1016/J.MEEGID.2015.09.030

Moreno, P., Alvarez, M., López, L., Moratorio, G., Casane, D., Castells, M., Castro, S., Cristina, J., Colina, R., 2009. Evidence of recombination in Hepatitis C Virus populations infecting a hemophiliac patient. Virol. J. 2009 61 6, 1–9. https://doi.org/10.1186/1743-422X-6-203

Murphy, D.G., Sablon, E., Chamberland, J., Fournier, E., Dandavino, R., Tremblay, C.L., 2015. Hepatitis C Virus Genotype 7, a New Genotype Originating from Central Africa. https://doi.org/10.1128/JCM.02831-14

Neher, R.A., Bedford, T., 2015. nextflu: real-time tracking of seasonal influenza virus evolution in humans. Bioinformatics 31, 3546–3548. https://doi.org/10.1093/BIOINFORMATICS/BTV381

Noppornpanth, S., Lien, T.X., Poovorawan, Y., Smits, S.L., Osterhaus, A.D.M.E., Haagmans, B.L., 2006. Identification of a Naturally Occurring Recombinant Genotype 2/6 Hepatitis C Virus. J. Virol. 80, 7569–7577. https://doi.org/10.1128/JVI.00312-06

Okada, M., Hai, H., Tamori, A., Uchida-Kobayashi, S., Enomoto, M., Kumada, H., Kawada, N., 2018. Successful direct-acting antiviral treatment of three patients with genotype 2/1 recombinant hepatitis C virus. Clin. J. Gastroenterol. 2018 123 12, 213–217. https://doi.org/10.1007/S12328-018-0922-9

Palanisamy, N., Kalaghatgi, P., Akaberi, D., Lundkvist, Å., Chen, Z. wei, Hu, P., Lennerstrand, J., 2018. Worldwide prevalence of baseline resistance-associated polymorphisms and resistance mutations in HCV against current direct-acting antivirals. Antivir. Ther. 23, 485–493. https://doi.org/10.3851/IMP3237

Paolucci, S., Premoli, M., Ludovisi, S., Mondelli, M.U., Baldanti, F., 2017a. HCV intergenotype 2k/1b recombinant detected in a DAA-treated patient in Italy. Antivir. Ther. 22, 365–368. https://doi.org/10.3851/IMP3130

Paolucci, S., Premoli, M., Novati, S., Gulminetti, R., Maserati, R., Barbarini, G., Sacchi, P., Piralla, A., Sassera, D., Marco, L. De, Girello, A., Mondelli, M.U., Baldanti, F., 2017b. Baseline and Breakthrough Resistance Mutations in HCV Patients Failing DAAs. Sci. Rep. 7. https://doi.org/10.1038/S41598-017-15987-1

Parikh, U.M., McCormick, K., Van Zyl, G., Mellors, J.W., 2017. Future technologies for monitoring HIV drug resistance and cure. Curr. Opin. HIV AIDS. https://doi.org/10.1097/COH.0000000000000344

Parker, J., Chen, J., 2017. Application of next generation sequencing for the detection of human viral pathogens in clinical specimens. J. Clin. Virol. 86, 20–26. https://doi.org/10.1016/J.JCV.2016.11.010

Penedos, A.R., Myers, R., Hadef, B., Aladin, F., Brown, K.E., 2015. Assessment of the Utility of Whole Genome Sequencing of Measles Virus in the Characterisation of Outbreaks. PLoS One 10, e0143081. https://doi.org/10.1371/JOURNAL.PONE.0143081

Petruzziello, A., Marigliano, S., Loquercio, G., Cozzolino, A., Cacciapuoti, C., 2016. Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes. World J. Gastroenterol. https://doi.org/10.3748/wjg.v22.i34.7824

Pickett, B.E., Sadat, E.L., Zhang, Y., Noronha, J.M., Squires, R.B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C.N., Dietrich, J., Klem, E.B., Scheuermann, R.H., 2012. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res. 40, D593–D598. https://doi.org/10.1093/NAR/GKR859

Popescu, B., Banica, L., Nicolae, I., Radu, E., Niculescu, I., Abagiu, A., Otelea, D., Paraschiv, S., 2018. NGS combined with phylogenetic analysis to detect HIV-1 dual infection in Romanian people who inject drugs. Microbes Infect. 20, 308–311. https://doi.org/10.1016/J.MICINF.2018.03.004

Preciado, M.V., Valva, P., Escobar-Gutierrez, A., Rahal, P., Ruiz-Tovar, K., Yamasaki, L., Vazquez-Chacon, C., Martinez-Guarneros, A., Carpio-Pedroza, J.C., Fonseca-Coronado, S., Cruz-Rivera, M., 2014. Hepatitis C virus molecular evolution: Transmission, disease progression and antiviral therapy. World J. Gastroenterol. https://doi.org/10.3748/wjg.v20.i43.15992

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One 5. https://doi.org/10.1371/journal.pone.0009490

Ramírez, S., Pérez-del-Pulgar, S., Carrión, J.A., Coto-Llerena, M., Mensa, L., Dragun, J., García-Valdecasas, J.C., Navasa, M., Forns, X., 2010. Hepatitis C virus superinfection of liver grafts: a detailed analysis of early exclusion of non-dominant virus strains. J. Gen. Virol. 91, 1183–1188. https://doi.org/10.1099/VIR.0.018929-0

Ross, R.S., Verbeeck, J., Viazov, S., Lemey, P., Ranst, M. Van, Roggendorf, M., 2008. Correspondence: Evidence for a Complex Mosaic Genome Pattern in a Full-length Hepatitis C Virus Sequence. Evol. Bioinforma. 4–249.

Salminen, M.O., Carr, J.K., Burke, D.S., McCutchan, F.E., 2009. Identification of Breakpoints in Intergenotypic Recombinants of HIV Type 1 by Bootscanning. https://home.liebertpub.com/aid 11, 1423–1425. https://doi.org/10.1089/AID.1995.11.1423

Schultz, A.-K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., Stanke, M., 2009. jpHMM: Improving the reliability of recombination prediction in HIV-1. Nucleic Acids Res. 37, W647–W651. https://doi.org/10.1093/NAR/GKP371

Shi, G., Suzuki, T., 2018. Molecular basis of encapsidation of Hepatitis C virus genome. Front. Microbiol. 9. https://doi.org/10.3389/FMICB.2018.00396

Singer, J., Gifford, R., Cotten, M., Robertson, D., 2020. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. Preprints 2020060225. https://doi.org/https://doi.org/10.20944/preprints202006.0225.v1

Singer, J.B., Thomson, E.C., Hughes, J., Aranday-Cortes, E., McLauchlan, J., Da Silva Filipe, A., Tong, L., Manso, C.F., Gifford, R.J., Robertson, D.L., Barnes, E., Ansari, M.A., Mbisa, J.L., Bibby, D.F., Bradshaw, D., Smith, D., 2019. Interpreting viral deep sequencing data with glue. Viruses 11. https://doi.org/10.3390/v11040323

Singer, J.B., Thomson, E.C., McLauchlan, J., Hughes, J., Gifford, R.J., 2018. GLUE: A flexible software system for virus sequence data. BMC Bioinformatics. https://doi.org/10.1186/s12859-018-2459-9

Smith, D.B., Bukh, J., Kuiken, C., Muerhoff, A.S., Rice, C.M., Stapleton, J.T., Simmonds, P., 2019. A Web Resource to Manage the Classification and Genotype and Subtype Assignments of Hepatitis C Virus [Online]. Int. Comm. Taxon. Viruses.

Smith, D.B., Simmonds, P., Jameel, S., Emerson, S.U., Harrison, T.J., Meng, X.J., Okamoto, H., Van der Poel, W.H.M., Purdy, M.A., 2014. Consensus proposals for classification of the family Hepeviridae. J. Gen. Virol. 95, 2223–2232. https://doi.org/10.1099/vir.0.068429-0

Spearman, C.W., Dusheiko, G.M., Hellard, M., Sonderup, M., 2019. Hepatitis C. Lancet. https://doi.org/10.1016/S0140-6736(19)32320-7

Stelzl, E., Haas, B., Bauer, B., Zhang, S., Fiss, E.H., Hillman, G., Hamilton, A.T., Mehta, R., Heil, M.L., Marins, E.G., Santner, B.I., Kessler, H.H., 2017. First identification of a recombinant form of hepatitis C virus in Austrian patients by full-genome next generation sequencing. PLoS One 12, e0181273. https://doi.org/10.1371/JOURNAL.PONE.0181273

Suzuki, T., 2012. Morphogenesis of infectious hepatitis C virus particles. Front. Microbiol. 3. https://doi.org/10.3389/fmicb.2012.00038

Tallo, T., Norder, H., Tefanova, V., Krispin, T., Schmidt, J., Ilmoja, M., Orgulas, K., Pruunsild, K., Priimägi, L., Magnius, L.O., 2007. Genetic characterization of hepatitis C virus strains in Estonia: Fluctuations in the predominating subtype with time. J. Med. Virol. 79, 374–382. https://doi.org/10.1002/JMV.20828

Terrault, N.A., Dodge, J.L., Murphy, E.L., Tavis, J.E., Kiss, A., Levin, T.R., Gish, R.G., Busch, M.P., Reingold, A.L., Alter, M.J., 2013. Sexual transmission of hepatitis C virus among monogamous heterosexual couples: The HCV partners study. Hepatology 57, 881–889. https://doi.org/10.1002/HEP.26164

Thorburn, F., Bennett, S., Modha, S., Murdoch, D., Gunson, R., Murcia, P.R., 2015. The use of next generation sequencing in the diagnosis and typing of respiratory infections. J. Clin. Virol. 69, 96–100. https://doi.org/10.1016/J.JCV.2015.06.082

Tscherne, D.M., Evans, M.J., Von Hahn, T., Jones, C.T., Stamataki, Z., McKeating, J.A., Lindenbach, B.D., Rice, C.M., 2007. Superinfection exclusion in cells infected with hepatitis C virus. J. Virol. 81, 3693–3703. https://doi.org/10.1128/JVI.01748-06

Viazov, S., Ross, S.S., Kyuregyan, K.K., Timm, J., Neumann-Haefelin, C., Isaeva, O. V., Popova, O.E., Dmitriev, P.N., Sharkawi, F. El, Thimme, R., Michailov, M.I., Roggendorf, M., 2010. Hepatitis C virus recombinants are rare even among intravenous drug users. J. Med. Virol. 82, 232–238. https://doi.org/10.1002/JMV.21631

Zhang, Y., Aevermann, B.D., Anderson, T.K., Burke, D.F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C.N., Lee, A.J., Li, X., Macken, C., Mahaffey, C., Pickett, B.E., Reardon, B., Smith, T., Stewart, L., Suloway, C., Sun, G., Tong, L., Vincent, A.L., Walters, B., Zaremba, S., Zhao, H., Zhou, L., Zmasek, C., Klem, E.B., Scheuermann, R.H., 2017. Influenza Research Database: An integrated bioinformatics resource for influenza virus research. Nucleic Acids Res. 45, D466–D474. https://doi.org/10.1093/NAR/GKW857

Zoulim, F., Chevallier, M., Maynard, M., Trepo, C., 2003. Clinical consequences of hepatitis C virus infection. Rev. Med. Virol. 13, 57–68. https://doi.org/10.1002/RMV.371

APPENDIX

***Table A 1**. HCV reference sequences used to create the project hcv_glue_avu in GLUE.*

Classification	Accession Number	Genotype and subtype status
1a	NC_004102, AF009606, M62321, M67463, HQ850279, EF407457	Confirmed
1b	D90208, M58335, EU781827, EU781828	Confirmed
1c	D14853, AY051292, AY651061	Confirmed
1d	KJ439768	Confirmed
1e	KC248194	Confirmed
1g	AM910652	Confirmed
1h	KC248198, KC248199	Confirmed
1i	KJ439772	Confirmed
1j	KJ439773	Confirmed
1k	KJ439774	Confirmed
1l	KC248193, KC248197, KC248196	Confirmed
1m	KJ439778, KJ439782	Confirmed
1n	KJ439781, KJ439775	Confirmed
1o	KJ439779, MH885469	Confirmed
1	HQ537007	Unassigned subtype
1	AJ851228	Unassigned subtype
1	KC248195	Unassigned subtype
1	KJ439780	Unassigned subtype

1	KJ439776	Unassigned subtype
1	KJ439777	Unassigned subtype
2a	D00944, AB047639, HQ639944	Confirmed
2b	D10988, AB030907, AB661388, AB661382	Confirmed
2c	D50409, JX227949	Confirmed
2d	JF735114	Confirmed
2e	JF735120	Confirmed
2f	KC844042, KC844050	Confirmed
2i	DQ155561	Confirmed
2j	HM777358, JF735113, HM777359	Confirmed
2k	AB031663, JX227953	Confirmed
2m	JF735111, JX227967	Confirmed
2q	FN666428, FN666429	Confirmed
2r	JF735115	Confirmed
2t	KC197238	Confirmed
2u	JF735112	Confirmed
2	JF735116	Unassigned subtype
2	JF735118	Unassigned subtype
2	JF735117	Unassigned subtype
2	JF735119	Unassigned subtype
2	JF735110	Unassigned subtype
2	KC197236	Unassigned subtype
2	KC197237	Unassigned subtype

2	KC197239	Unassigned subtype
3a	D17763, D28917, X76918, JN714194	Confirmed
3b	D49374, JQ065709	Confirmed
3d	KJ470619	Confirmed
3e	KJ470618	Confirmed
3g	JX227954, JF735123	Confirmed
3h	JF735126, JF735121	Confirmed
3i	FJ407092, JX227955	Confirmed
3k	D63821, JF735122	Confirmed
3	JF735124	Unassigned subtype
4a	Y11604, DQ988074, DQ418789	Confirmed
4b	FJ462435	Confirmed
4c	FJ462436	Confirmed
4d	DQ418786, FJ462437, EU392172	Confirmed
4f	EF589161, EU392175, EU392174	Confirmed
4g	FJ462432, JX227971	Confirmed
4g?	JX227963	Confirmed
4k	EU392173, FJ462438, EU392171	Confirmed
4l	FJ839870, JX227957	Confirmed
4m	FJ462433, JX227972	Confirmed
4n	FJ462441, JX227970	Confirmed
4o	FJ462440, JX227977	Confirmed
4p	FJ462431	Confirmed

4q	FJ462434	Confirmed
4r	FJ462439, JX227976	Confirmed
4s	JF735136	Confirmed
4t	FJ839869	Confirmed
4v	HQ537009, JX227959, HQ537008, JX227960	Confirmed
4w	FJ025855, FJ025856	Confirmed
4	FJ025854	Unassigned subtype
4	JX227964	Unassigned subtype
4	JF735127	Unassigned subtype
4	JF735132	Unassigned subtype
4	JF735131	Unassigned subtype
4	JF735130	Unassigned subtype
4	JF735129	Unassigned subtype
4	JF735138	Unassigned subtype
4	JF735135	Unassigned subtype
4	JF735134	Unassigned subtype
5a	AF064490, Y13184	Confirmed
5	KT595242	Unassigned subtype
6a	Y12083, AY859526, HQ639936, EU246930	Confirmed
6b	D84262	Confirmed
6c	EF424629	Confirmed
6d	D84263	Confirmed
6e	DQ314805, EU246932	Confirmed

6e?	EU246931	Confirmed
6f	DQ835760, EU246936	Confirmed
6g	D63822, DQ314806	Confirmed
6h	D84265	Confirmed
6i	DQ835770, DQ835762	Confirmed
6j	DQ835769, DQ835761	Confirmed
6k	D84264	Confirmed
6l	EF424628, JX183556	Confirmed
6m	DQ835767, DQ835766	Confirmed
6n	DQ278894, DQ835768, EU246938	Confirmed
6o	EF424627, EU246934	Confirmed
6p	EF424626	Confirmed
6q	EF424625	Confirmed
6r	EU408328	Confirmed
6s	EU408329	Confirmed
6t	EF632071, EU246939	Confirmed
6u	EU246940	Confirmed
6v	EU158186, EU798760, EU798761	Confirmed
6w	DQ278892, EU643834, EU643836	Confirmed
6xa	EU408330, EU408332, EU408331	Confirmed
6xb	JX183552, KJ567645	Confirmed
6xc	KJ567651	Confirmed
6xd	KM252789, KM252790, KM252791	Confirmed

6xe	JX183557, KM252792	Confirmed
6xf	KJ567646, KJ567647	Confirmed
6xg	MH492361, MH492360, MH492362	Confirmed
6xh	MG879000	Confirmed
6	DQ278891	Unassigned subtype
6	DQ278893	Unassigned subtype
6	JX183558	Unassigned subtype
6	JX183553	Unassigned subtype
6	JX183554	Unassigned subtype
6	JX183551	Unassigned subtype
6	JX183549	Unassigned subtype
6	JX183550	Unassigned subtype
6	KJ470620	Unassigned subtype
6	KJ470621	Unassigned subtype
6	KJ470622	Unassigned subtype
6	KJ470623	Unassigned subtype
6	KJ470624	Unassigned subtype
6	KJ470625	Unassigned subtype
6	KC844039	Unassigned subtype
6	KC844040	Unassigned subtype
6	KJ567652	Unassigned subtype
6	KJ567650	Unassigned subtype
6	KJ567649	Unassigned subtype

6	KJ567648	Unassigned subtype
6	KJ567644	Unassigned subtype
6	MG878999	Unassigned subtype
7a	EF108306	Confirmed
7b	KX092342	Confirmed
8a	MH590698, MH590699, MH590700, MH590701	Confirmed

***Table A 2**. Exceptions returned when queries using GLUE commands were used to retrieved information from the database.*

Query	GLUE Exception
Custom tables	`‘Exception in thread "main" java.lang.RuntimeException: Object of type: org.apache.cayenne.access.ToManyList cannot be put in a GLUE document’`
Genotyping results	`‘Exception in thread "main" java.lang.IllegalArgumentException: Height or width cannot be less than or equal to zero’`
All data from NGS	`‘Exception in thread "main" java.lang.RuntimeException: Object of type: java.util.ArrayList cannot be put in a GLUE document’`