UPPSALA
UNIVERSITET

UPTEC X 19015
Examensarbete 30 hp
Juni 2019

Validation of a new software for detection of resistance associated substitutions in Hepatitis C-virus

Caitlin Vigetun Haughey

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Validation of a new software for detection of resistance associated substitutions in Hepatitis C-virus

Caitlin Vigetun Haughey

Hepatitis C infection is a global disease that causes an estimated 399,000 deaths per year. Treatment has improved dramatically in recent years through the development of direct acting antivirals that target specific regions of the Hepatitis C virus (HCV). Unfortunately the virus can have a preexisting resistance or become resistant to these drugs by mutations in the genes that code for the target proteins. These mutations are called resistance-associated substitutions (RASs). Since RASs can cause treatment failure for patients, resistance detection is performed in clinical practice to select the ideal regimen. Currently RASs are detected by using Sanger sequencing and a partly manual workflow that can discriminate the presence of a RAS if it is present in 15-20% of viruses in a patients blood. A new method with the capacity to detect lower ratios of RASs in HCV sequences was developed, which utilizes Pacific Biosciences' (PacBio's) sequencing and a bioinformatics analysis software called CLAMP. To validate this new approach, 123 HCV patient samples were sequenced with both methods and then analyzed. The RASs detected with the new method were congruent to what was found with the Sanger-based workflow. The new approach was also shown to correctly genotype the virus samples, identify any co-existing mutations on the same sequences, and detect if there were any mixed genotype infections in the samples. The new procedure was found to be a valid replacement for the Sanger based workflow, with the possibility to perform additional analyses and perform automated and time efficient RAS detection.

Handledare: Kåre Bondeson
Ämnesgranskare: Claes Ladenvall
Examinator: Jan Andersson
ISSN: 1401-2138, UPTEC X 19015

Populärvetenskaplig sammanfattning

En ny metod för att identifiera resistensmutationer i Hepatit C virus

Hepatit C virusinfektion är en leversjukdom som orsakas av Hepatit C viruset (HCV) och vanligen överförs mellan personer via blod intravenöst. De vanligaste sättet att smittas är via infekterade sprutor, så som vid droganvändning. Mindre vanligt är att att smittas sexuellt, via blodtransfusioner och icke steriliserad medicinsk utrustning. Sjukdomen är globalt sett väldigt utbredd, med ungefär 130-170 miljoner smittade människor. Utav dessa personer så lider majoriteten av dem med vad som kallas kronisk hepatit C, vilket innebär att infektionen finns kvar i kroppen under en lång tid. Lindriga eller inga symptom upplevs då de första åren, men viruset orsakar i många fall slutligen skrumplever och levercancer.

För att behandla HCV användes länge ett modifierat typ av protein som celler producerar när ett virus angriper, interferon, tillsammans med det antivirala läkemedlet Ribaravin. Denna typ av medicinering gav vanligtvis patienter biverkningar och ledde endast till förbättring i 50-60% av fallen. Behandling av HCV har dock förbättrats avsevärt de senaste åren i samband med att nya läkemedel började framställas. Dessa läkemedel är så kallade ”direkt agerande antiviraler” och deras höga verkningsgrad beror till stor del på att de riktar in sig på väldigt specifika måltavlor i virussekvensen och inaktiverar dessa. Detta leder i sin tur att viruset förlorar förmågan att utföra nödvändiga steg när det ska replikera sig i den infekterades celler. Denna nya behandlingsmetod med dessa läkemedel har visat sig vara verkningsfulla i så många som 96% av fallen där de har använts.

Trots dessa nya effektiva mediciner så finns det fortfarande risk för att behandlingar mot HCV misslyckas. Detta beror på att Hepatit C viruset kan utveckla resistens genom mutationer i de delar av sin arvsmassa som utgör måltavlor för läkemedlen. Dessa mutationer finns ibland naturligt i virusens populationer, men de kan också uppstå under behandling på grund av naturligt urval. De virus som inte har resistensmutationen dör då ut när läkemedlet tas, medan de överlevande resistenta virusen kan fortsätta att föröka sig och infektionen fortskrider.

För att undvika att behandlingar misslyckas så utreds det om en patient bär på ett HCV med resistensmutationer innan man valt vilken typ av terapi som ska ges. Detta gör man genom att ta prover från patienten, rena fram viruset, och sedan använder man sekvensering för att läsa av virusets arvsmassa. På så sätt kan man studera positionerna i sekvensen och urskilja huruvida viruset bär på några mutationer som kan göra det resistent mot vissa läkemedel. I samband med detta kan man även klassificera viruset till olika grupper som kallas genotyper. Genotyperna är baserade på hur procentuell likhet mellan dom olika virustyperna, där de som är mer lika tillhör samma genotyp. Detektion av mutationer och klassificering av viruset sker

oftast genom användning av så kallad Sangersekvensering, en typ av sekvenseringsmetod som döptes efter en av skaparna, Frederick Sanger. Denna metod är gammal och etablerad med hög noggrannhet, men med dålig känslighet. Sangersekvensering kan endast hitta mutationer om de förekommer i minst 15-20% av alla virussekvenser, vilket gör att mutationer som finns i lägre antal inte kommer att kunna identifieras.

Ett sätt att undvika detta problem är att använda sig av en sekvenseringsteknik med bättre känslighet, vilket är något som testats av en forskningsgrupp på SciLifeLab i Uppsala. 123 patientprover med Hepatit C har då sekvenserares på en ny typ av teknologi från Pacific Biosciences (PacBio) som använder sig av det som kallas ”long-read sequencing”, en process som är känsligare än Sanger och gör det möjligt att läsa av längre segment av genetiskt material. En bioinformatisk analys framställdes för att kunna gå igenom PacBio proverna och utföra sökningen efter mutationer. De 123 proverna sekvenserares även med Sanger och analyserades i ett befintligt delvis manuellt arbetsflöde för förekomst av resistenskopplade mutationer för att göra det möjligt att jämföra de två metoderna.

Det projekt som beskrivs i denna rapport var att utföra denna jämförelse av den konventionella Sanger-baserade metoden och den nya bioinformatiska analysen, beträffande de två metodernas förmåga att hitta resistensmutationer och bestämma virusens genotyper för de 123 HCV-proverna. För att den nya metoden ska anses vara ett tänkbart alternativ för dessa typer av analyser i klinisk verksamhet, så måste genotyperna för proverna stämma överens med klassificeringen från den Sanger-baserade metoden, och alla de mutationer som identifierats med hjälp av den gamla metoden också kunna hittas med den nya tekniken.

Resultaten från projektet visade på att den nya metoden fann alla mutationer i proverna som referensmetoden hittade, och att virusens genotyper stämde fullständigt överens mellan metoderna. Utöver detta så kunde den PacBio baserade metoden hitta mutationer som förekom i färre antal, så få som i endast 0,5% av alla virussekvenser. Den nya analysen kunde även studera om vissa mutationer förekom tillsammans ofta i ett prov, och avgöra om det fanns några patienter som hade blandade viruspopulationer av olika genotyper. Resultaten för denna nya analysmetod överensstämmer helt med förväntade resultat. Metoden är enkel att använda, kräver väldigt lite manuellt arbete av användaren, och gör det möjligt att studera Hepatit C infektioner på helt nya sätt.

Table of Contents

Abbreviations

CCS	Circular Consensus Sequence
CLR	Continuous Long Read
DAA	Direct Acting Antiviral
dNTP	Deoxynucleotidetriphosphates
ddNTP	Dideoxynucleotidetriphosphates
GT	Genotype
HCV	Hepatitis C Virus
IRES	Internal Ribosome Entry Site
NGS	Next Generation Sequencing
NS3	Non-structural protein 3
NS5A	Non-structural protein 5A
NS5B	Non-structural protein 5B
NTR	Non-translated Region
ORF	Open Reading Frame
PacBio	Pacific Biosciences
PID	Personal Identifier
PCR	Polymerase Chain Reaction
RAS	Resistance Associated Substitution
RAV	Resistance Associated Variant
SMRT	Single Molecule Real Time
SVR	Sustained Virological Response
ZMW	Zero-mode Waveguide

1. Introduction

Hepatitis C is a liver disease caused by the Hepatitis C virus (HCV). The infection can cause both acute and chronic infection, with the later making up for 8 out of 10 cases. For a chronic HCV infection the virus is then present in the bloodstream for years and causes liver cirrhosis for 15–30% of infected persons within 20 years. The disease is present globally with an estimated 71 million people infected worldwide. The virus is blood borne and causes approximately 399 000 deaths every year due to cirrhosis and liver cancer (World Health Organisation (WHO) 2018).

Treatment of Hepatitis C has been revolutionized through the development of direct acting antivirals (DAA), drugs that target the HCV proteins specifically with limited side effects (Jakobsen et al. 2017). Despite this treatment failures still occur due to resistance development. These mutations are known as resistance-associated substitutions (RASs) and occur in the genes that encode for the proteins targeted by the drugs. The substitution rates in these genes are relatively high, resulting in HCV being divided into 7 distinct genotypes each with a number of subtypes. These genotypes are associated with naturally occurring RASs and thus require different treatment regimens (Houghton 2016).

Usually HCV is genotyped and analyzed for resistance mutations in patients before initializing treatment, and the conventional practice is to use Sanger sequencing. Sanger can detect RASs that are present in at least 15-20% of the viruses in the sample (Rohlin *et al.* 2009). More recently Next generation sequencing (NGS) technologies have emerged as an alternative method for resistance detection, which are able to find viral variants at a much lower prevalence, as low as 0,5-1% (Sarrazin 2016). However NGS technologies have certain drawbacks such as short read lengths and relying on clonal amplification, which introduces PCR related errors.

A bioinformatics analysis software called CLAMP was created to screen for RASs in HCV samples sequenced with Pacific Biosciences' (PacBio) long-read technology (Bergfors *et al.* 2016). The software is written in R and Perl and can perform both HCV genotyping and resistance detection. For this new automated bioinformatics procedure to be deemed a valid alternative to a Sanger based partly manual workflow used in clinical practice, it had to meet certain requirements such as overall cost and amount of work. Most importantly, the results would have to meet a 95% agreement of RAS detection with the Sanger based analysis workflow as the reference. This master's project compared HCV infected patients data from both Sanger sequencing and PacBio SMRT sequencing to evaluate this new method as a possible replacement method for HCV detection.

2. Background

Hepatitis C is an infectious disease caused by the Hepatitis C virus (HCV), which primarily affects the patient's liver. Approximately 71 million people have chronic Hepatitis C infection globally, and in 2015 it was estimated that 1,75 million new HCV infections occurred, amounting to 23,7 new HCV infections per 100 000 people. The disease causes an estimated 399 000 deaths every year (World Health Organisation (WHO) 2018). The disease presents itself either as an acute- or a chronic infection, with the later accounting for roughly 80% of HCV infections. Acute Hepatitis C infections only make up for approximately 15% of cases (Maheshwari *et al.* 2008). Acute HCV is often symptomless and 10-50% of these infections resolve themselves spontaneously (Shiffman 2011). More commonly the infection will turn chronic, with the virus being present and replicating in the patient's liver over a long period of time. During the first few years of a chronic infection the patient will have mild to no symptoms, however after several years the virus may cause liver cirrhosis and cancer (Westbrook & Dusheiko 2014). The virus is blood borne and is in most cases spread through intravenous drug use, transfusion of unscreened blood or through poorly sterilized medical equipment. The disease can also be spread through sexual contact and from a HCV positive mother to her child at birth, but these cases are less common (World Health Organisation (WHO) 2018).

2.1 Hepatitis C Virus - Virology

HCV belongs to the *Hepacivirus* genus and it is a small single-stranded RNA virus (55-65 nm). The genome of the virus is positive-stranded RNA with a size of approximately 9,600 nucleotides and a single long open reading frame (ORF). This ORF is translated to produce an amino acid polyprotein of around 3000 bases that is then post-translationally processed into a number of mature structural and non-structural proteins (Paul *et al.* 2014). The ORF is flanked at both the 5' and 3' ends by non-translated regions (NTR). The 5' NTR functions as an internal ribosome entry site (IRES). Binding of a ribosome to the IRES initiates the translation of the 3kb long polyprotein. The 3' NTR contains a conserved sequence essential for the replication of the HCV genome (Lohmann *et al.* 1999). The final mature protein products from the translated polyprotein are from start to finish; C-E1-E2-p7-NS2-NS3-NS4A-NS4B-NS5A-NS5B. C, E1 and E2 are structural proteins that make up the virus particle. p7 is a viroprotein that together with nonstructural protein 2 (NS2) facilitate virus assembly. The remaining nonstructural proteins NS3, NS4A, NS4B, NS5A and NS5B, form a replicase complex for replication of HCV (De Francesco 1999) (Figure 1).

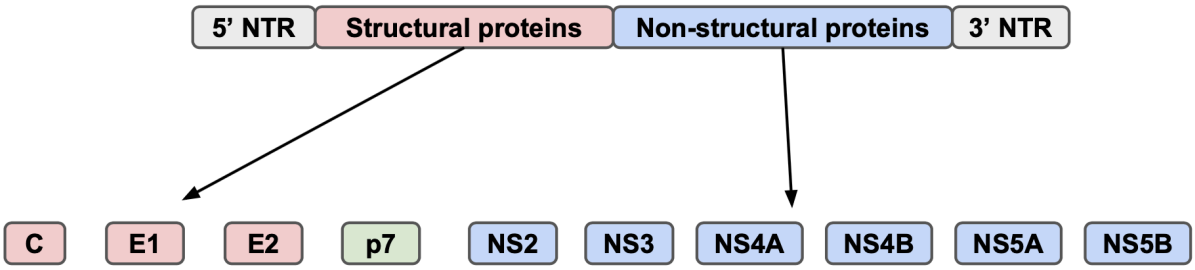

Figure 1. Genome organisation of the Hepatitis C virus. The RNA is translated to produce a single polyprotein, which is further processed to produce active structural and non-structural proteins. The structural proteins include Core protein (C), E1 and E2. The nonstructural proteins are NS2, NS3, NS4A, NS4B, NS5A, and NS5B, and p7 is a viroprotein.

2.1.1 NS3, NS5A and NS5B

The NS3, NS5A and NS5B proteins are targeted by modern day HCV medicines known as direct acting antivirals (DAA's). NS3 is a protease that forms a complex with the NS4A cofactor transmembrane protein to perform protein cleavage. NS5B is an RNA-dependent RNA-polymerase involved in the replication of the HCV's genome (Paul *et al.* 2014). For the last of the protein targets, NS5A, our understanding of structure and function is limited. It is however generally accepted that NS5A plays an important role in replication of RNA and through interaction and binding with other non-structural proteins (Macdonald & Harris 2004).

2.1.2 HCV genotypes and subtypes

Because the RNA-dependent RNA-polymerase NS5B lacks proof-reading ability, mutations occur often through each replication cycle of the virus. This results in HCV genomes being highly heterogenous and are thus classified into 7 distinct genotypes (GT1-GT7), each with its own set of subtypes (Houghton 2016). HCV genotypes are denoted by "GT" and the number of the genotype, and subtypes are designated with a lower case letter (a-z). Viruses within a genotype have an estimated sequence similarity of 60-70%, while within a subtype the similarity is approximately 75-85% (Table 1).

Table 1. Grouping of HCV types based on percentage sequence identity. Performed using phylogenetic tree analysis and pair-wise comparison of HCV sequences (Nakano *et al.* 2012).

Group	% sequence identity	Denoted number/letter
Genotypes	60-70%	1-7
Subtypes	75-85%	a-z

HCV genotypes and subtypes differ in geographical distribution, with GT1 being the most common genotype worldwide, making up for approximately 46% of HCV cases. The second most common is GT3, with 30% of cases (Messina *et al.* 2015). Due to the high percentage of sequence discrepancy between groups, and the fact that there are more than 80 confirmed HCV subtypes (Smith *et.al.* 2018), each virus type requires a particular treatment approach. In some rare cases the patient may be infected with more than one HCV type at once, something that is more commonly encountered in patients that undergo repeated risks for exposure such as intravenous drug users (Pham *et al.* 2010).

2.3 Treatment of Hepatitis C

The main goal of HCV treatment is to cure the infection, which is determined by whether or not sustained virological response (SVR) is achieved. SVR is defined as undetectable levels of viral RNA in the patient's bloodstream for 12 weeks (SVR12) or 24 weeks (SVR24) after the conclusion of treatment (Pawlotsky *et al.* 2018). Treatment has advanced greatly during the last few years, mostly due to the development of the so called direct acting antivirals (DAA) that target one of the NS3, NS5A and NS5B proteins specifically and that have limited side effects (Jakobsen et al. 2017). DAA's are divided into four categories depending on targeted protein and the binding site of the drug: (i) NS3/4A protease inhibitors; (ii) NS5A inhibitors, (iii) nucleoside NS5B polymerase inhibitors, and (iv) non-nucleoside NS5B polymerase inhibitors (De Clercq 2014). Before DAA's were developed, HCV patients were treated with a combination of Interferon alpha and Ribavarin. This regimen was only effective in around 50% of cases and was highly associated with severe side effects (Manns et al. 2001, Sung et al. 2011). In comparison, DAA's have been shown to be effective in up to 96 % of cases (Zhang *et al.* 2016). The current standard treatment for HCV is a combination of DAA's that target different non-structural proteins, but the specific choice of these DAA's is based on the genotype of the virus, the current state of the patient's liver, and the potential pre-existence of resistance mutations in the viruses NS3, NS5A and NS5B regions. Some HCV genotypes, such as GT3a, are known to be more difficult to treat. Cases of HCV with genotype 3a are commonly linked to low SVR rates and DAA treatment failure, in particular if liver cirrhosis has developed (Kanwal *et al.* 2014). Regardless of the genotype, treatment failure regularly ensues if the virus has pre-existing resistance associated substitutions (Pawlotsky *et al.* 2018).

2.4 Resistance development

HCV can become resistant to antiviral drugs through mutations in the specific genes. These mutations are called resistance-associated substitutions (RASs) and result in an amino acid change in the non-structural proteins targeted by drugs, the NS3, NS5A and NS5B genes. These RAS's lead to a decrease in the sensitivity of the virus to the effects of the drug. A virus strain that has one or more *in-vitro* confirmed RASs is known as a resistance associated variant (RAV). In some cases resistance substitutions can pre-exist as naturally occurring variants of the virus, or they can be developed during treatment as a result of drug selection pressure (Wyles & Luetkemeyer 2017). For an infected patient a number of genetically different HCV quasi-species exist simultaneously due to the high error-rate of the viruses replication. These subpopulations of the virus have different levels of viral replication fitness, with the mutated versions of the virus generally having a reduced fitness compared to the wild-type virus. However when a DAA is administered this induces a positive selection for the mutated viruses, which allows the resistant variants to take over as the majority HCV strain (Pawlotsky 2016). If a resistance to drug emerges in a patient that is treatment-naive, the infection can be due to a naturally occurring resistance associated variant of the virus.

2.5 Resistance detection and HCV genotyping

Resistance in HCV first started to gain clinical relevance when DAA's were approved for use and became the recommended treatment regimen. Naturally occurring resistance variants can have a significant effect on whether or not the first course of treatment is successful, while attained variants that cause treatment failure will impact the choice of a new regimen. Moreover, each of the 7 HCV genotypes have different possible RASs, naturally occurring ones and those developed during therapy. Thus each genotype has its own specific resistance profile to certain antiviral drugs (Patiño-Galindo *et al.* 2016). HCV is currently genotyped and screened for resistance mutations using either population-based Sanger sequencing or Next generation sequencing (NGS), with the former being the more commonly used approach (Chevaliez *et al.* 2012).

2.5.1 Sanger sequencing

Detection of RASs is based around DNA sequencing, were the most conventional of the approaches is to utilize so called population-based Sanger sequencing. This method was developed in 1977 by Frederick Sanger and colleagues (Sanger *et al.* 1977), and is based on a so called "chain termination in DNA replication" process of sequencing. This chain termination method of sequencing requires a number of components, a single-stranded DNA template, normal deoxynucleotidetriphosphates (dNTPs), and modified

dideoxynucleotidetriphosphates (ddNTPs), a DNA polymerase to synthesize a copied DNA from the nucleotides, and lastly a DNA primer to initiate synthesis. Because the ddNTPs lack the 3'-OH group that is needed to form a phosphodiester bond between nucleotides, elongation of the DNA strand is effectively terminated when a ddNTP has been incorporated. The four different ddNTPs (ddATP, ddGTP, ddCTP and ddTTP) are labelled with specific fluorescent dyes, which allows a detector to identify which of the nucleotides that was incorporated. The Sanger process results in a pool of copied DNA segments of different lengths terminated at the 3' end (Sanger & Coulson 1975, Sanger *et al.* 1977). Sanger sequencing in general has a low sensitivity in regards to detecting single polymorphisms such as RASs in samples, with an estimated detection threshold of 15-20% (Rohlin *et al.* 2009). This means that if a resistance mutation is present in less than 15% of all virus sequences, it will not be detected by Sanger. The method does however remain the "golden standard" for genotyping and RAS detection in HCV, due to the fact that drug-specific RASs are considered to be clinically relevant if they have been proven to reduce likelihood of SVR. And this has been shown to only occur for current DAA regimens if the mutation is present in at least 15% of the viruses (Pawlotsky 2016).

2.5.2 Next Generation sequencing

Deep sequencing approaches using NGS technologies are a viable alternative for detection of resistance mutations, and are able to detect viral variants with a sensitivity of down to 0,5-1% (Sarrazin 2016). The deep sequencing method thus has the capacity to detect minor variants below Sanger's limit. When this approach is used for research purposes a cut-off level of 1% is usually used for what is deemed significant, however for clinical usage NGS thresholds are often set to >10% due to the current standards of clinical relevance (AASLD-IDSA 2018). RASs that are found at a lower rate may not give the HCV population a sufficient resistance to reduce a patient's SVR with the currently available DAA regimens (Sarrazin *et al.* 2016). Though it can be argued that this cut-off for significant RASs should be questioned, as it was most likely set to its current rate simply because it is the limit of detection for the reference method of population sequencing. NGS methods are increasing in popularity and resistant variants present at lower frequencies (<15%) in HCV patients might be able to impact whether or not SVR is achieved, making it worth considering updating this cut-off value (Perales *et al.* 2018).

2.6 Long read sequencing

Even though deep sequencing methods using NGS have a lot of advantages, these short-read technologies do have some drawbacks when it comes to HCV resistance detection. Even though the throughput is high the length of the reads are considerably shorter than those from Sanger sequencing. Due to this NGS methods are often unable to detect multiple resistance mutations that co-exist in the same viral RNA sequence, which can lead to a high fold increase of resistance (Sorbo *et al.* 2018). NGS technology also relies on clonal amplification for sequencing, something that often introduces PCR related errors into the sequence products. In comparison, the new third generation sequencing methods have the ability to provide long reads of up to 100kb without any amplification (Pollard *et al.* 2018).

2.6.1 Sequencing with Pacific Biosciences

Pacific Biosciences (PacBio) sequencing technology is a long-read based approach which is based on so called single-molecule real-time (SMRT) sequencing, and can detect viral variants with high sensitivity. PacBio SMRT sequencing captures sequence information in real time during the replication process of the DNA molecule of interest. The application requires input material in the form of a library of at least 5 micrograms of double stranded DNA (Ardui et al. 2018). The library is constructed by ligating hairpin adapters onto the molecules, forming them into circular constructs known as SMRTbells. The sample of SMRTbells are then loaded to a chip called a SMRT cell, were the SMRTbell then diffuses into nanoscale observation chambers called a zero-mode waveguides (ZMWs). Each of these ZMW contains a single polymerase that binds to the hairpin adaptors of the SMRTbell and starts the replication by incorporating one out of four fluorescently labelled nucleotides. The fluorescence emitted by the nucleotides when incorporated is recorded by a camera as a movie of light pulses corresponding to a sequence of bases, which is called a continuous long read (CLR) (Rhoads & Au 2015). Since the SMRTbell has two hairpin adapters it forms a closed circle, the polymerase can continue sequencing the target multiple times. The CLR can then be split to subreads by cutting at the position of the adaptor sequences. The consensus sequence of reads then results in a circular consensus sequence (CCS) read with very high accuracy. The number of times the target will be sequenced depends on its length, and so if the target DNA is too long to be sequenced multiple times, a CCS read cannot be created and a single read becomes the final output. Long read sequencing technologies are known to have a relatively high error rate (3–15%) compared to short-read NGS technologies, however these errors are mostly insertions and deletions (indels) and occur at random positions in the sequence. Since the target is sequenced multiple times and the reads are overlapped into a CCS read the erroneous bases can be removed, resulting in a final product with high accuracy (Ameur et al. 2019).

2.7 Development of new method for HCV resistance detection

A new bioinformatics approach for genotyping HCV and detecting RASs was developed recently based on Pacific Biosciences' (PacBio) long-read sequencing. This new analysis procedure is performed by a software called CLAMP, which was written in the programming languages R and Perl by a team at the National Genomics Infrastructure (NGI) at SciLifeLab in Uppsala (Bergfors *et al.* 2016). CLAMP was first developed to assist in the treatment of chronic myeloid leukemia (CML), by checking the *BCR-ABL1* fusion transcript for mutations that may lead to the patient becoming resistant to tyrosine kinase inhibitor (TKI) based therapy (Cavelier *et al.* 2015). The software was then adapted to detect RASs in the NS5A and NS5B genes complemented with the capability to genotype HCV. CLAMP takes HCV samples sequenced with PacBio as input and starts by comparing the samples to reference sequences for each genotype to find the closest match. Once the virus has been genotyped the program uses lists of confirmed drug-specific RASs to check if any are present in the virus sample.

The modified version of CLAMP was designed to be used in a clinical setting, mainly aimed for RAS detection in the NS5A, at the Uppsala University Hospital department of clinical microbiology, virology (CMB). Since Sanger sequencing combined with a partially manual bioinformatic workflow is the reference method for HCV RAS detection, the objective of the new approach is that it will replace the Sanger based process. The performance of CLAMP has to be equivalent to the reference method, it has to correctly perform genotyping and ideally detect all RASs found by the Sanger based reference method in order to safely replace the currently used reference method. This master's project will focus on comparing HCV infected patients RASs detection data derived from both the Sanger sequencing and PacBio SMRT sequencing with the objective to evaluate CLAMP in accordance to ISO/EN15189 for use in clinical laboratory practice (International Organisation of Standardizations requirements for quality and competence in medical laboratories; ISO 15189 2012).

2.8 Project goal

The purpose of this project was to assess whether or not the new computational method is capable of both detecting all of the RASs found with the reference method based on Sanger sequencing, as well as lower prevalence RASs that cannot be detected with Sanger (<15%). This study evaluates if the PacBio based method can replace the Sanger based one on a set of criteria from the client at the Uppsala University Hospital such as overall cost, reliability, effectiveness and amount of work. To perform this comparison, 123 patient samples were sequenced with both approaches and then analyzed. The results of the comparison had to meet predetermined quality objectives established with the Sanger sequencing based routine method as the reference. These objectives included:

- A minimum of 500 reads for each sample.
- A cut-off value of 15 % for viral variants, i.e. the lowest proportion of variant reads that can be reported as a RAS.
- The agreement of the detected RASs between the two methods must be at least 95%.
- The genotypes determined by the new method are not allowed to deviate from that of the reference method.
- The new method of detection needs to be significantly more time efficient as compared to the reference method.

If these criteria are met, the new method can be regarded as a viable alternative for Hepatitis C resistance detection in clinical practice. The results of the project could also be of use to evaluate if lower prevalence resistance variants can be clinically relevant and thus endorse the need for a detection method with a lower detection limit. The results from the study are presented in this scientific report and as a computational procedure for detecting RASs from PacBio data.

3. Materials and Methods

The main goal of the project was to perform comparative analyses of two different computational methods based on PacBio and Sanger data respectively, and to evaluate the performance of PacBio for Hepatitis C virus RAS detection. To do this the project was divided into smaller milestones:

1. A literature study, gathering of all the necessary background information needed to perform the project. This step entailed the planning of the project by finding relevant references, choosing the methods to be used for the comparison, making sure all the required data was available, determining the cut-off value and other criteria. The literature study helped to develop an understanding of many important aspects of the project: resistance genes in HCV and how they mutate at different rates, how HCV is genotyped in practice, the possible drawbacks and limitations of certain methods et cetera.

2. Performing the comparative analyses of the Sanger data and the PacBio data to evaluate performance, the accuracy of the HCV genotyping and ability to detect RASs present in different frequencies. The analyses were done with programming in R, Perl and Python and the resources available for this step were computers at both NGI and CMB with access to all the relevant data and software, as well as the student's own private computer.

3. Assess whether or not the new method using PacBio is a valid alternative for clinical detection of RASs based on criteria such overall performance in detection, medical safety, time efficiency and economical benefits. The clinical relevance of the results were determined using predetermined cut-off values. The relative user friendliness of the method was also evaluated, since the intended users of the end product are not bioinformaticians.

3.1 Preparation of HCV samples and Sanger sequencing

128 HCV patient samples were selected by clinicians at the CMB department for the study, were 16 of these were taken in 2016 and the other 112 in 2017. The samples were selected at random without any correlation except for the fact that they all originated from patients examined at hospitals in Sweden. Out of the 128 HCV samples, five were excluded from the set and were not used in the project analyses. Three of these were removed due to not having been genotyped and screened at CMB, making it impossible to compare the results with those from PacBio. Another sample was removed because of the PacBio read files being empty, possibly due to some error in the sequencing process. Lastly one more sample was excluded

from the set because the Sanger sequence and the PacBio reads did not correspond at all to one another, leading us to believe that a mix-up had occurred at some point in the workflow preceding the sequencing. The final total of samples used for the Sanger and PacBio comparison was thus 123. All 123 samples and their results can be found in Appendices (Table A1).

3.1.1 PCR amplification and sequencing

The virus RNA was extracted, synthesized into cDNA and then amplified with a nested PCR approach according to the procedure devised by HCV specialists at CMB (Lindström *et al.* 2015). This procedure can be summarized as follows. The RNA was extracted from patient plasma or serum samples with the NucliSENS easyMAG system from the biotechnological company bioMérieux. cDNA was then synthesized by reverse transcription and the product used as the template during PCR amplification in one single step, using the GeneAmp PCR system 9700 and a primer pair. Next a nested PCR was run with Taqman Universal PCR Master Mix from Applied Biosystems and primers that target the NS5A region of the virus, which results in PCR amplicons of 636 basepairs in length. The PCR products were checked on a 2% agarose e-gel to verify that the amplification was successful, and lastly prepared and sent to Eurofins Genomics for sequencing using Sanger.

3.1.2 HCV genotyping (Sanger based results)

To perform genotyping with the Sanger based method, the non-structural gene NS5B is amplified with PCR and sequenced. A consensus sequence from the NS5B gene is created using Applied Biosystems SeqScape software. The consensus is run through the National Center for Biotechnology Information's (NCBI) viral genotyping tool which compares the sequence to all HCV genotypes and subtypes in the database by using BLAST. The tool returns the results in the form a score table of the sequence matched against all subtypes, with the highest score being the most probable type for the virus sample. The sequenced NS5B gene is also aligned against HCV genotype reference sequences using the MEGA software version 7 (Kumar *et al.* 2016). A phylogenetic tree is created from the alignment, containing the sample sequence along with the reference genotypes. The tree is then compared to the results from NCBI and the exact subtype is determined.

3.1.3 Resistance detection in NS5A with reference method

After genotyping the NS5A sample sequence is loaded into SeqScape and aligned. The sequence is visualized as a chromatogram were each position is manually analyzed for ambiguous double or triple peaks, indicating that some virus sequences in the sample have a nucleotides that deviate from the majority population. These additional peaks exhibit a significant signal and is included in the edited consensus sequence if it's amplitude corresponds to at least 20% of the larger peak. If these nucleotides are located in positions

that result in a change of amino acid that gives the virus resistance to certain DAA's in the patient, the sample is said to have a confirmed RAS in this position (Figure 2). Lastly the virus consensus sequence is imported into Geno2pheno for an additional resistance prediction, which is the MaxPlanck institute's online web based system for RAS detection and evaluation of resistance relative to currently used NS5A-inhibitors. (Kalaghatgi *et al.* 2016).

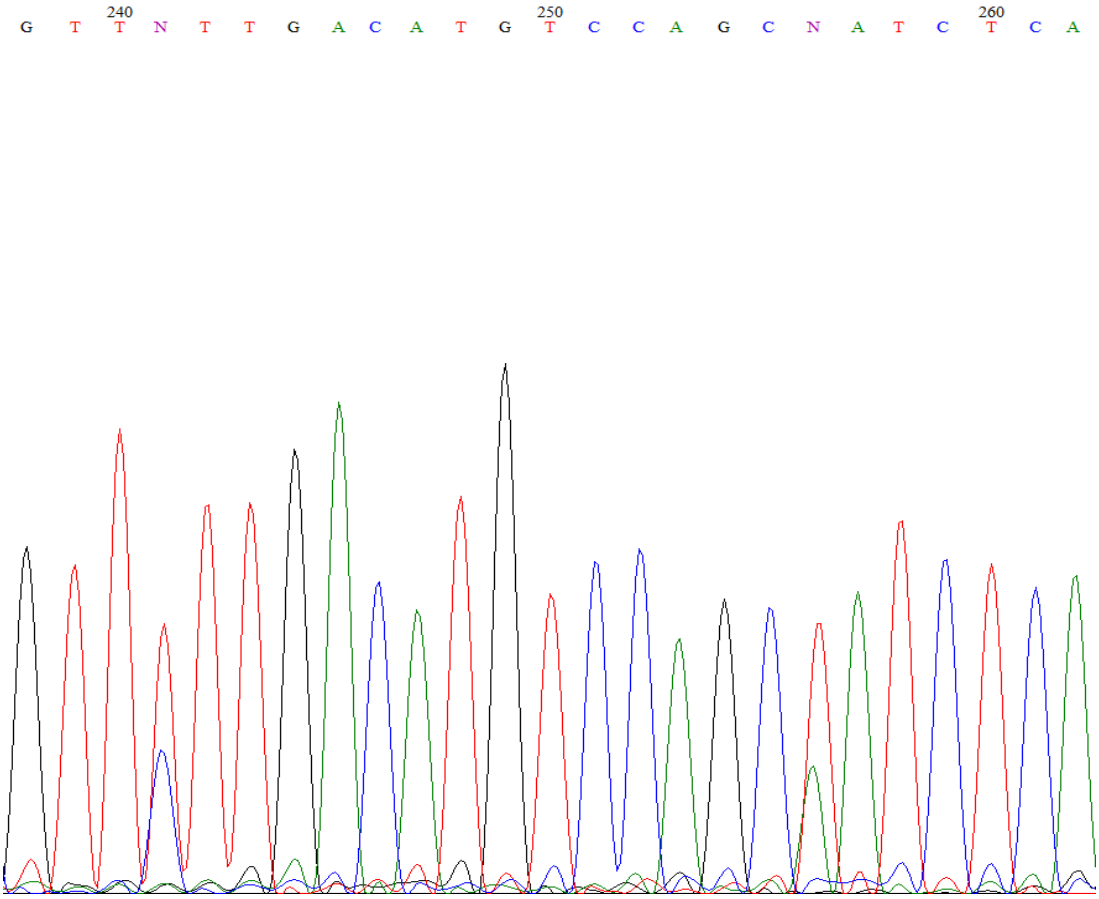

Figure 2. An example of a chromatogram from Sanger sequencing. The chromatogram displays nucleotide bases from position 238 to 262 in the NS5A gene sequence. Position 241 shows a double peak, which could be due to a mutation. The larger peak (red) equals the nucleotide T, and the second peak (blue) equals C. If the mutation results in a change of amino acid that gives the virus resistance, the sample has a confirmed RAS in this position.

3.2 PacBio sequencing and detection with CLAMP software

Along with sending the PCR products to Eurofins Genomics for Sanger sequencing, the samples were also sent to the Uppsala Genome Center at SciLifeLab for SMRT sequencing on the Pacific Biosciences RS II system. The NS5A sample sequences were generated into SMRTbells and barcoded to enable them to be pooled together 8 samples in one SMRTcell. This pooling of samples reduces the total cost of the long-read sequencing, making it a more appealing alternative for the HCV detection. Once the sequencing was finished, CCS reads were created by utilizing protocols on the SMRT portal, a web directory for secondary analysis of PacBio data. The CCS reads from the samples were retrieved as files in fastq format for further analysis. The read files are input into the CLAMP program and first go through a series of filtering processes before detection of resistance mutations. Sequences that are not from viruses NS5A gene are removed by checking each read for the primer pair in the beginning and end of the read. Only the reads containing both primers are saved and used for the analysis.

3.2.1 Genotyping of PacBio reads

The first step of the CLAMP analysis is the genotyping using SNP calling, which compares the sequence reads to each of the HCV subtype reference sequences and identifies bases that differ. The reference sequence with the highest coverage (least amount of mismatched bases) is selected as the optimal reference and the sample is denoted to this subtype/genotype. The NS5A reference sequences used for this study were the ones that had been classified for the sample data set by genotyping of the Sanger results, GT 1a, 1b, 2b, 3a and 4a. The references for the different subtypes were obtained from the NCBI database and their accession numbers are listed below (Table 2)

Table 2. The reference sequences used for HCV genotypes, listed by genotype and their GenBank accession numbers. Reference sequences were acquired from the National Center for Biotechnology Information (NCBI).

Genotype	Accession number
1a	NC_004102.1
1b	AJ238799.1
2b	AF238486.1
3a	NC_009824.1
4a	Y11604.1

3.2.2 RAS detection in NS5A with CLAMP

The sample reads are screened for RASs in the program using RAS (mutation) query sequences specific to each genotype. These tables contain lists of selected confirmed RASs, with the mutated position in the sequence flanked on both sides by 20 nucleotides. Each RAS is then labelled by the wild type amino acid, the position in the sequence, and the mutated aa. For example: if a resistance mutation occurs in position 30 of the sequence, and the amino acid is changed from an A to a K, the RAS is defined as ”A30K”. The RASs for each genotype are listed below (Table 3). The full RAS lists with the sequences of 20 bases around each side of the mutation site for each genotype can be found in Appendix B (Table B1).

Table 3. Resistance Mutations (RAS) in NS5A gene for each genotype. *In-vivo* confirmed clinically relevant RASs and non-confirmed RASs for the NS5A region, with the mutations listed in the columns by genotype. Each RAS is specified by the wild-type amino acid, the position in the aa chain, and the mutated aa. RASs that require two nucleotide changes for the causing amino acid change are shown in brackets. The non-RASs are denoted with an ”n” at the end. Non-RASs are unrelated to the objectives of defining resistance according to current standards, but may be used for exploratory purposes.

GT1a	GT1b	GT2b	GT3a	GT4a
Clinically relevant RASs				
M28T	L28P	Y93H	M28T	V28M
M28V	L28T (L28P+P28T)		A30T	L30I
M28A (M28V+V28A)	R30Q		A30K (A30T+T30K)	L30V
M28K	L31M		A30V	L30stop.1
M28N (M28K+K28N)	L31V		L31F	L30L
Q30E	L31I		L31I	L30stop.2
Q30H	L31F		L31V	Y93H
Q30K	P32L		L31M	Y93C
Q30R	Y93H		Y93H	Y93stop.1
L31M	Y93N		Y93C	Y93stop.2
L31V	Y93S		Y93N	
P32L	Y93C			
H58D				
H58Y				
H58R				

H58Q	
H58P	
A92T	
Y93C	
Y93H	
Y93N	
Y93S	
Y93F	
Non-RASs	
A25Gn	G33Rn
G33Rn	P35Tn
F36Ln	P58An
R44Kn	A62Sn
R48Qn	A62Tn
G51Cn	T64An
K68Rn	T64Sn
V75An	L74In
V75In	H85Yn
R78Kn	

CLAMP selects the mutation table that corresponds to the genotype assigned to the sample and loops through all reads for the sample to check if they contain the listed mutations. Each of the mutation sites are pinpointed in the CCS reads sequences by matching 20 bases flanking both sides of the site. A maximum of 9 mismatches in these flanking regions are allowed in comparison to the reference sequence, since the NS5A gene is known to be quite variable. The position of the mutation however has to match exactly for the read to be classified as having that specific RAS. Once all reads for a sample have been screened for mutations, the prevalence of each RAS in the sample is calculated as the number of reads out of the total number have the mutation in that position. A RAS prevalence of >1% of the reads is assigned as positive, though this cut-off can be adjusted by the user. A sample that is positive for more than one RAS will in turn be run through a clonal distribution step, which is done to detect if these resistance mutations co-exist in the same viral sequence or not. The software outputs a number of result files, with the most important ones being a table with

each screened RAS for and their frequencies in the sample, a QC plot with the coverage for each position of the sequence, and a plot displaying the clonal distributions.

3.3 Identification of mixed infections

As previously mentioned, patients may be infected with more than one HCV type at once, something that can affect the outcome of the treatment. However the reference detection method using Sanger sequencing does not have the capacity to identify mixed infections, the dominant genotype in the sample will be detected while the minor types go unnoticed (del Campo *et al.* 2018). So it was requested by the clients at Uppsala University Hospital that the new method using would include a function that can detect mixed infections in samples. This functionality does not exist in the CLAMP software, and so an additional step in the project was to write code to perform this task and run it for the samples to reveal if any of these have more than one HCV type. The mixed infection analysis part was written in Python and starts by using NCBI's local command-line version of the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.* 1990) known as BLAST+. These tools are available online for downloading and installation (Camacho *et al.* 2009) and allows the user to create their own local database of reference sequences to blast against. The tool was used to genotype the reads from the virus samples by matching each read against a local blast+ reference database containing the HCV genotype sequences. The reads were then grouped by their best blast hit according to the Expect value (E-value) and the number of reads matched to each genotype could be calculated. Any reads with an E-value greater than 0,001 are removed since these do not have a high enough significance to be classified to any genotype. Whether or not the patient sample contains a mixed infection is assessed by if the minority virus is present in more than 15% of all reads, a cut-off value that can be modified by the user.

The most common length out of the grouped reads is calculated, and then the reads are filtered by number of gaps and sequence length. The reads that have the same length as the most common read length and a maximum of 1 gap are saved. These filtered reads are then aligned using the Multiple Sequence Comparison by Log- Expectation (MUSCLE) software (Edgar 2004). Lastly a consensus sequence is created from the aligned reads, with a consensus threshold of >85% for a position to be unambiguously set to a certain character. For example, assuming a position contains 12 A's, 1 G's and 1 T and the consensus threshold is set to 85% or below, then the consensus for this position will be set to A. This step is performed using the AlignInfo module for Python and the *dumb_consensus()* method. The consensus sequence for the sample can then be used to verify the result of te CLAMP based RAS detection and genotyping by submitting it to the web-based RAS detection system, Geno2pheno. This ability can assist in keeping the list of clinically relevant RASs updated through an externally

curated list of RASs. The entirety of the two processes, the old method with Sanger and the new PacBio method, are visualized with a flowchart (Figure 3).

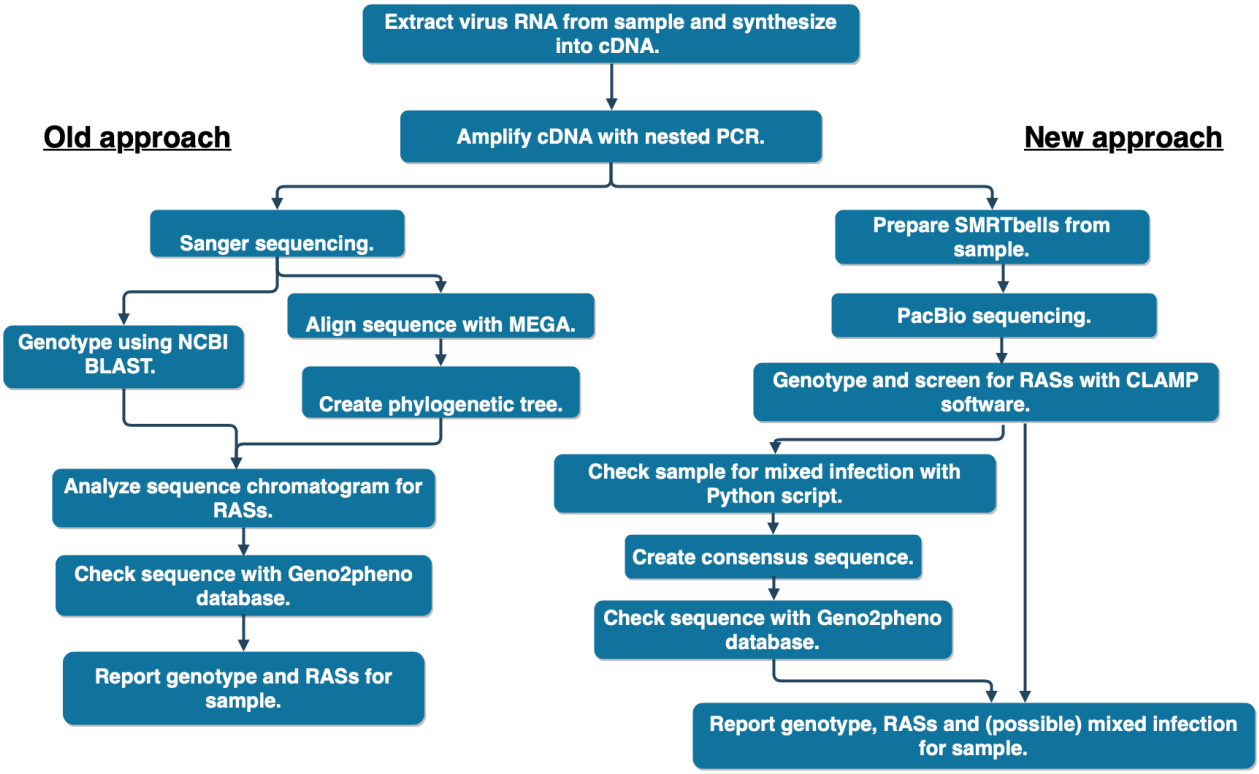

Figure 3. Flowchart displaying the process of preparing the HCV samples using both the old Sanger approach and the new PacBio approach. The first two steps of both the methods involve extraction of the virus RNA, synthesizing of cDNA and PCR amplification. The left part of the chart shows the old approach, which uses Sanger for sequencing and then genotypes the consensus sequence with both BLAST and a phylogenetic tree. The sequence is manually screened for mutations by going through the sequence chromatogram and with Geno2pheno. The right part of the chart shows the new approach, which uses PacBio sequencing. For genotyping and resistance detection in the PacBio reads the software CLAMP is used. The sample reads are also analyzed with a Python script to detect any possible mixed genotype infections. This script also creates a consensus sequence for the reads that can be input into Geno2pheno.

4. Results

4.1 Analysis of resistance mutations and genotyping

The screening of HCV is used by the clinicians at CMB to report the genotype, which RASs are present in the patient sample, and which drugs the virus is resistant against. This information is sent to the physician responsible for the patient's treatment so that an informed decision can be made regarding which DAA regimen to use. This project's main focus was thus to evaluate the CLAMP software's ability to detect RASs and genotype the virus in an efficient and reliant way. To determine this the new method has to comply with the results of the RAS detection and genotyping using the reference method of Sanger sequencing. The new method needs to have a 95% concordance with the reference method in regards to RASs detected above the 15-20% cut-off. We expect to find additional mutations with the new method below this cut-off since this is Sanger's detection limit whilst PacBio has been shown to be able to detect mutations down to 0,1% (Baybayan & Nolden 2017).

All samples were genotyped with full agreement to the reference method (Sanger based workflow). Further, all RASs found with the reference method were subsequently detected in the PacBio reads by CLAMP. 44 out of all the samples were positive for at least one RAS with a prevalence of 15% or above. For the samples that did not have a RAS above this cut-off, 14 of these had at least one RAS with a prevalence less than 15% (Table 4). 65 samples did not have any detected RASs. As predicted the new method was able to identify RASs below the 15% limit that were not detected with Sanger. The total list of samples with or without any mutations can be found in the Appendices (Table A1).

Table 4. Results from resistance detection analysis of PacBio reads using CLAMP. Showing the predicted genotype, RASs present at >15%, 10-15%, 5-10% and 1-5%. The RAS frequencies are shown in brackets.

Sample ID	GT[1]	RAS (>15%)	RAS (10-15%)	RAS (5-10%)	RAS (1-5%)
pb_440_1_bc1	3a	A30K (99,6%)			
pb_440_1_bc2	3a	Y93H (70,8%)			A30K (4,3%)
pb_440_1_bc3	1a	H58R (95,5%)			
pb_440_1_bc4	1a	M28V (58,7%)			
pb_440_1_bc5	1a			M28V (8,7%)	
pb_440_1_bc8	1a	Y93H (99,8%)			
pb_440_2_bc1	3a	H58P (99,7%)			
pb_440_2_bc2	3a	Y93H (100%)			
pb_440_2_bc3	1a	M28V (90,8%), Q30R (92,8%)		M28A (9,1%)	
pb_440_2_bc4	3a	A30K (99,5%)			
pb_440_2_bc5	1a	Y93H (100%)			
pb_440_2_bc6	3a				H58R (1,1%)
pb_440_2_bc8	3a	L31M (99,7%), Y93H (99,9%)			
pb_440_3_bc2	1a				M28V (1,3%)
pb_440_3_bc3	3a		M28V (11,8%)		
pb_440_3_bc5	1b	Y93H (100%)			
pb_440_3_bc6	3a	M28V (98,8%)	Q30R (13,3%)		
pb_440_3_bc7	1a	Y93H (45,5%)			
pb_440_4_bc5	1a	M28V (99,7%)			
pb_440_4_bc6	3a	L31M (99,6%)			
pb_440_5_bc2	1a	Y93H (99,2%)			
pb_440_5_bc4	3a	M28V (91,6%)			
pb_440_6_bc2	3a	Y93H (93,6%)			
pb_440_6_bc6	3a				M28V (2,2%)

pb_440_6_bc7	1a	M28V (71,5%)		
pb_440_6_bc8	1a	Y93H (99,9%)		
pb_440_7_bc7	1a		Y93H (6,4%)	A30V (1,3%)
pb_440_8_bc1	1a	H58Y (20,4%)		
pb_440_8_bc4	1b	M28V (99,7%)		
pb_440_8_bc7	3a			H58R (1,1%)
pb_440_8_bc8	1a	M28T (20,7%), Q30K (64,1%), ***Q30R (18,6%), L31M (19,3%)***[2]		Q30E (1%)
pb_455_1_bc10	1a	Q30R (91,2%)		
pb_455_1_bc11	3a			M28A (1,1%), L31M (1,4%)
pb_455_1_bc12	1a	H58P (99,7%)		
pb_455_1_bc15	1a	A30K (99,9%)		
pb_455_1_bc16	1a	H58Q or H58P (42,7%)[3]		
pb_455_2_bc9	1a			Q30R (2,5%)
pb_455_2_bc11	1a			H58Q (1,1%)
pb_455_2_bc16	4a			A30V (1%)
pb_455_3_bc14	3a	Y93H (99,7%)		
pb_455_3_bc15	1b	M28V (70,7%)		
pb_455_4_bc11	1a	M28V (100%)		L31M (2,8%)
pb_455_4_bc14	1a	Y93H (99,8%)		
pb_455_4_bc16	3a	M28V (99,9%)		
pb_455_5_bc11	1a	Q30H (99,7%), Y93H (100%)		
pb_455_5_bc13	1a	A30K (99,7%)		
pb_455_5_bc14	3a	Q30R (92,2%), Y93F (66,6%)		
pb_455_6_bc10	1a	A30K (99,7%)		
pb_455_6_bc11	3a			M28V (1,3%)

pb_455_6_bc12	1a	Q30H (99,9%), Y93H (99,9%)		
pb_455_6_bc15	3a			H58R (1,2%)
pb_455_7_bc9	3a	A30V (99,7%)		
pb_455_7_bc14	3a	Y93F (96,6%)		Q30K (4,1%), Q30R (1,3%), Y93H (3,9%)
pb_455_7_bc15	3a	H58P (100%)		
pb_455_7_bc16	1a	H58Y (96,1%)		M28V (1,7%), H58R (1,6%)
pb_455_8_bc9	1a	A92T (95,1%)		Y93N (1,7%)
pb_455_8_bc11	1a	M28V (100%), Q30K (99,9%), H58Q (99,5%)		Q30R (1,4%), H58P (1,4%)
pb_455_8_bc15	1a		H58P (5,1%)	

[1] Genotype classified with both Sanger and PacBio sequencing.
[2] RASs Q30R and L31M not identified with Sanger sequencing.
[3] Neither Sanger nor PacBio could discern whether the mutated amino acid in this position was Proline (P) or Glutamine (Q).

4.1.1 Sanger vs. PacBio

The comparative analysis of the two different sequencing methods was based on evaluating if the new method would be able to perform HCV genotyping and resistance detection in congruence with the reference method with Sanger. Above the Sanger limit of 15-20%, the two methods needed to conform to at least 95% in regards to found RASs. Below this limit the new approach was expected to find mutations that the reference method could not. In all of the samples, a total of 54 RASs were found at a ratio of 15% or above. Out of all of the RASs identified at 15% or above, the two lowest prevalence RASs were not found in the consensus sequence obtained Sanger. Both of these mutations came from the same sample, RAS L31M at 19,3%, and Q30R at 18,6%. 29 RASs with prevalence below 15% were found in the samples in total, all of which could not be found in the Sanger consensus sequences. These differences in detected resistance mutations between the two methods effectively put Sanger sequencing's sensitivity at around 20%, since it could not find any mutations under this percentage. The ten RASs with the lowest prevalence above 20%, along with the ten RASs with the highest prevalence below 20% were plotted into a bar chart, displaying the cut-off point for when Sanger could no longer detect any mutations (Figure 4). The bar chart

containing all of the RASs found above and below 15% in all samples can be found in Appendices (Figure C1).

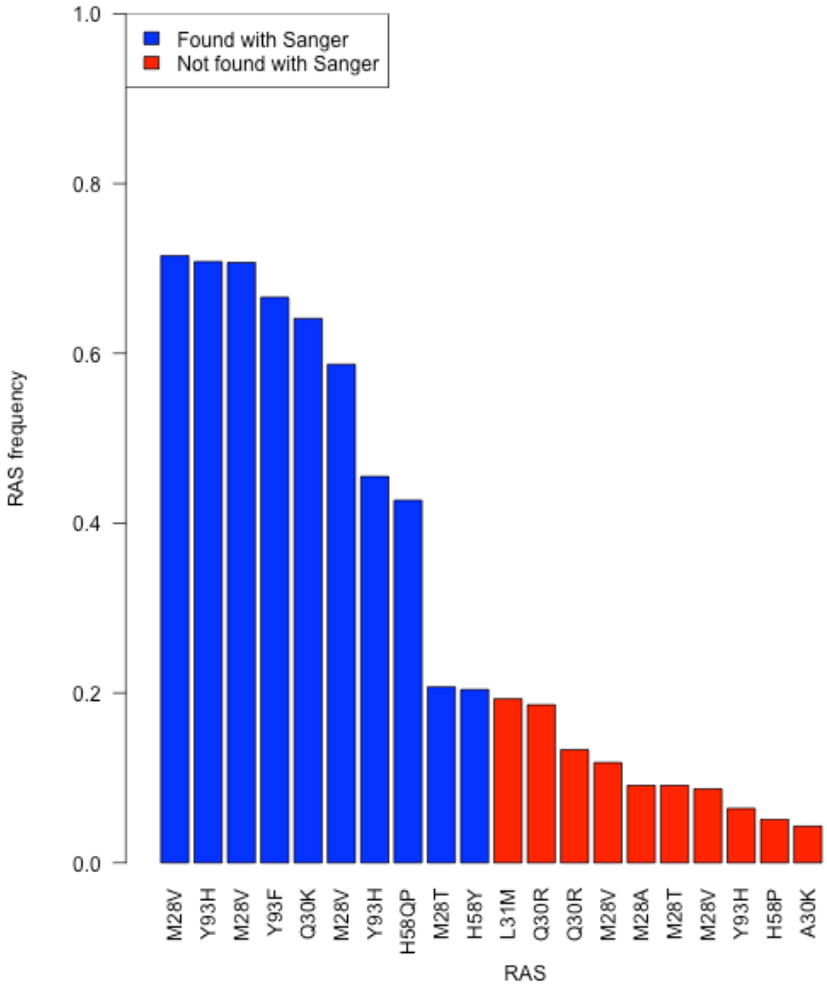

Figure 4. Bar chart displaying 10 RASs found in samples with prevalence above 20%, and 10 RASs below 20%. The bars in blue were found with both the PacBio and Sanger methods, while the bars in red were not found using Sanger. The cut-off between these groups correlate with a detection limit of 20% for Sanger sequencing.

4.1.2 Sensitivity and specificity

The sensitivity and specificity of the CLAMP/PacBio approach was calculated using the Sanger-based method as the reference and counting the number of true positives (TP) true negatives (TN), false positives (FP) and false negatives (FN). These measurements were calculated using:

$$sensitivity = TP/(TP + FN).$$

$$specificity = TN/(TN + FP).$$

The true positives (TP) in the samples were counted as RASs detected above the 20% limit by both approaches. True negatives (TN) were all of the possible RAS positions analysed in the samples that were not positive for a RAS above 20%. False positives (FP) were any RASs detected above 20% with the new method that were not detected with the reference method. And lastly the false negatives (FN) were any RASs not detected above 20% by the new approach, but were detected with the reference approach.

For the 123 samples, a total of 52 true positive (TPs) RASs were detected. The true negatives (TNs) were all of the possible RAS positions for all samples that were not positive for a >20% RAS, which amounted to 2387 TNs. The new approach did detect all RASs found with the reference method, and so there were 0 false negatives in the data set. However due to the fact that the CLAMP software is currently programmed to only check one codon position for any exact RAS match, the new method did result in certain RAS positive positions being matched against more than one specific RAS in the same read. This flaw of CLAMP can be corrected by altering the software to match all 3 positions of a codon exactly for a positive match, but for this project this was not done. And so 5 false positive RASs were detected in the 123 samples by the new method.

The sensitivity and specificity of the new CLAMP/PacBio approach were calculated as:

$$sensitivity = TP/(TP + FN) = 52/(52 + 0) = 1.$$

$$specificity = TN/(TN + FP) = 2387/(2387 + 5) = 0{,}9979097 \approx 0{,}997.$$

4.2 Investigating the existence of co-mutations

The samples were grouped into separate tables based in genotype to determine most common RAS for each type and to analyze if any mutations occur more frequently together in HCV patients. The three most common RASs for genotype 1a, 1b and 3a were counted (Table 5). The most common mutations were not calculated for genotype 2b and 4a, since GT2b only has one clinically relevant resistance mutation and there was only one GT4a sample.

Table 5. The three most common RASs for genotypes 1a, 1b and 3a. The number of samples that have the mutation at at least a 1% prevalence are shown in brackets. GT1b only had two RASs with 1% or above.

Most common RASs	GT1a	GT1b	GT3a
1st	M28V (17)	Y93H (2)	Y93H (11)
2nd	Q30R (8)	L31M (2)	A30K (6)
3rd	H58P (6)	-	A30V (3)

4.2.1 Heat maps

To establish the existence of any distinguishable co-mutations, the samples were grouped by genotype. The prevalence of the samples RASs were then plotted as heat maps. Since the 2b genotype only has one RAS on its list and the 4a type only had one sample in the entire data set, heat maps were only created for GT1a (Figure 5), GT1b (Figure 6) and GT3a (Figure 7). Due to the majority of the samples having only a few out of all the possible RASs, and more than 80% of these being in the top 85th and bottom 15th percentiles, the color key scale was set to accommodate this.

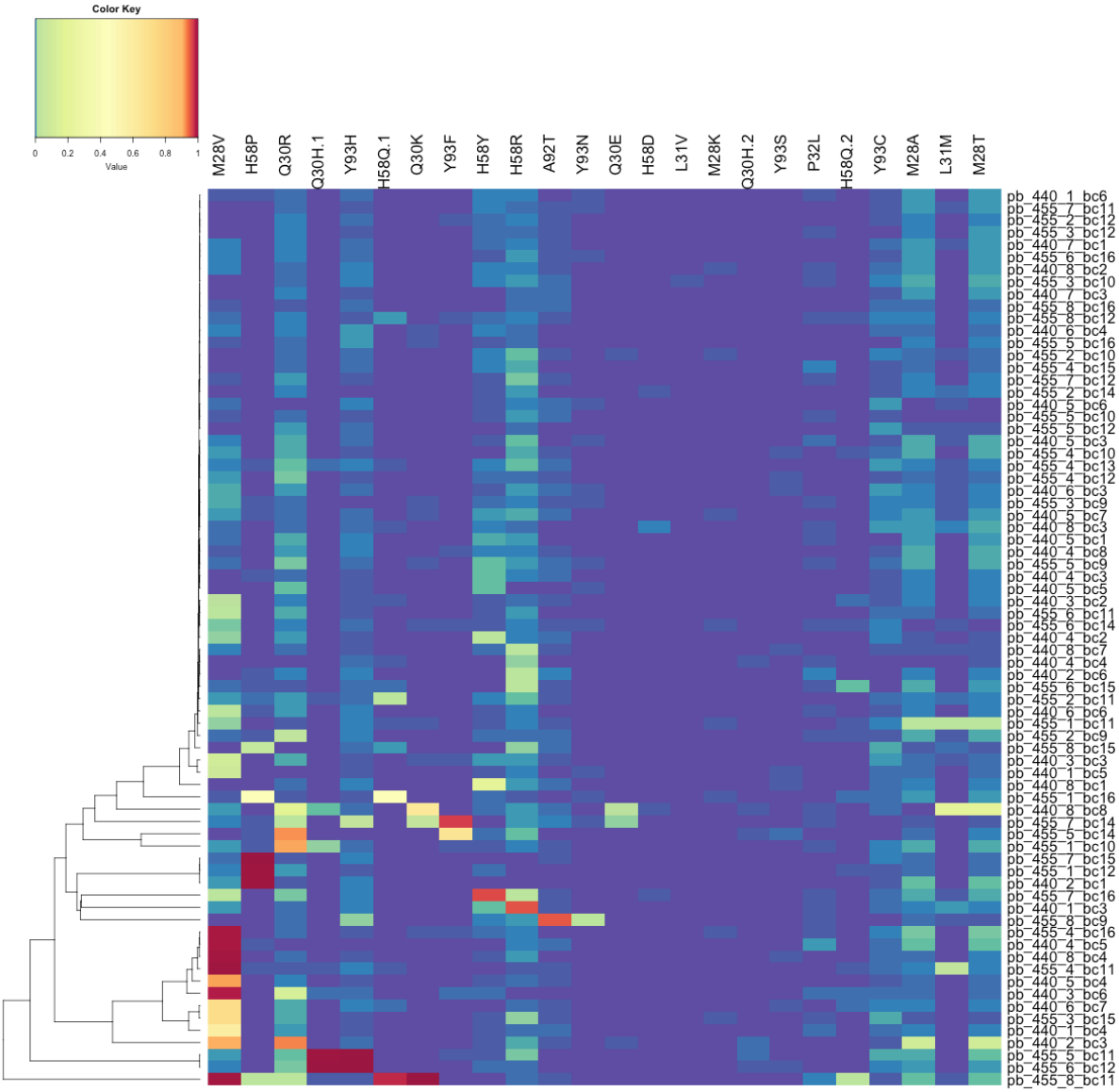

Figure 5: Heat Map of RASs for GT1a samples. Heat map showing the frequencies of resistance mutations for the 73 samples in the data set with genotype 1a. Sample IDs are listed by row, while resistance mutations are listed by column. The heatmap is clustered by rows, demonstrated as a dendrogram left of the plot. Mutation frequencies are coloured from purple/blue (low) to red (high). The majority of the samples were positive for one or more low level RASs (prevalence <1%) in position 28, and 18 with samples a RAS above 1% in this position (green/yellow colour cells in columns 1, 22 and 24). 8 samples had a prevalence of >1% for the Q30R RAS (green/yellow cells in column 3), and 5 samples had the RAS H58R at >1% (column 10).

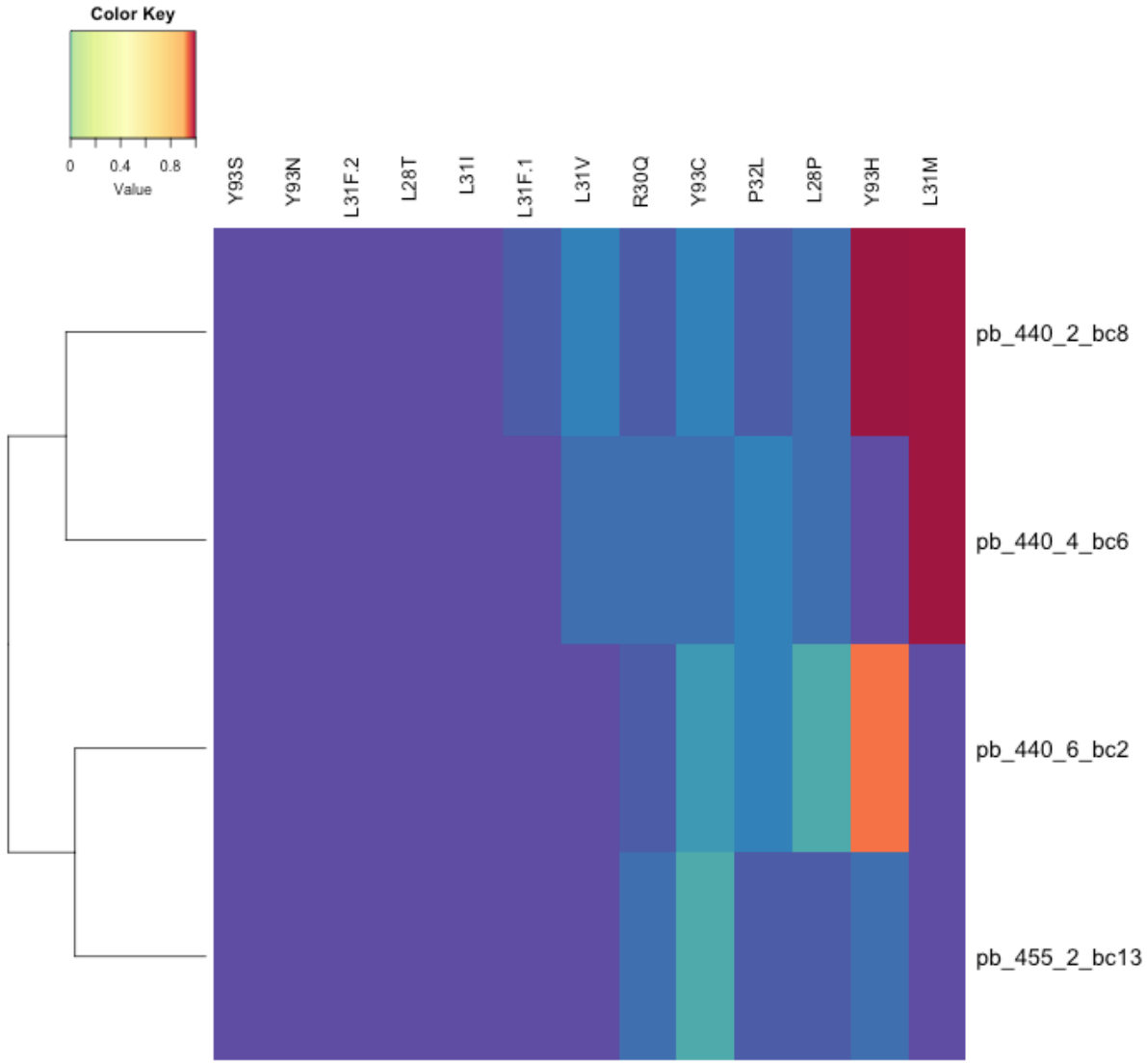

Figure 6: Heat Map of RASs for GT1b samples. Heat map showing the frequencies of resistance mutations for the 4 samples in the data set with genotype 1b. Sample IDs listed by row, resistance mutations listed by column. The heatmap is clustered by rows, demonstrated as a dendrogram left of the plot. Mutation frequencies are coloured from purple/blue (low) to red (high). 2 samples were highly positive (prevalence >90%) for the Y93H RAS (column 12) and 2 samples had the RAS L31M at frequencies above 99% (column 13).

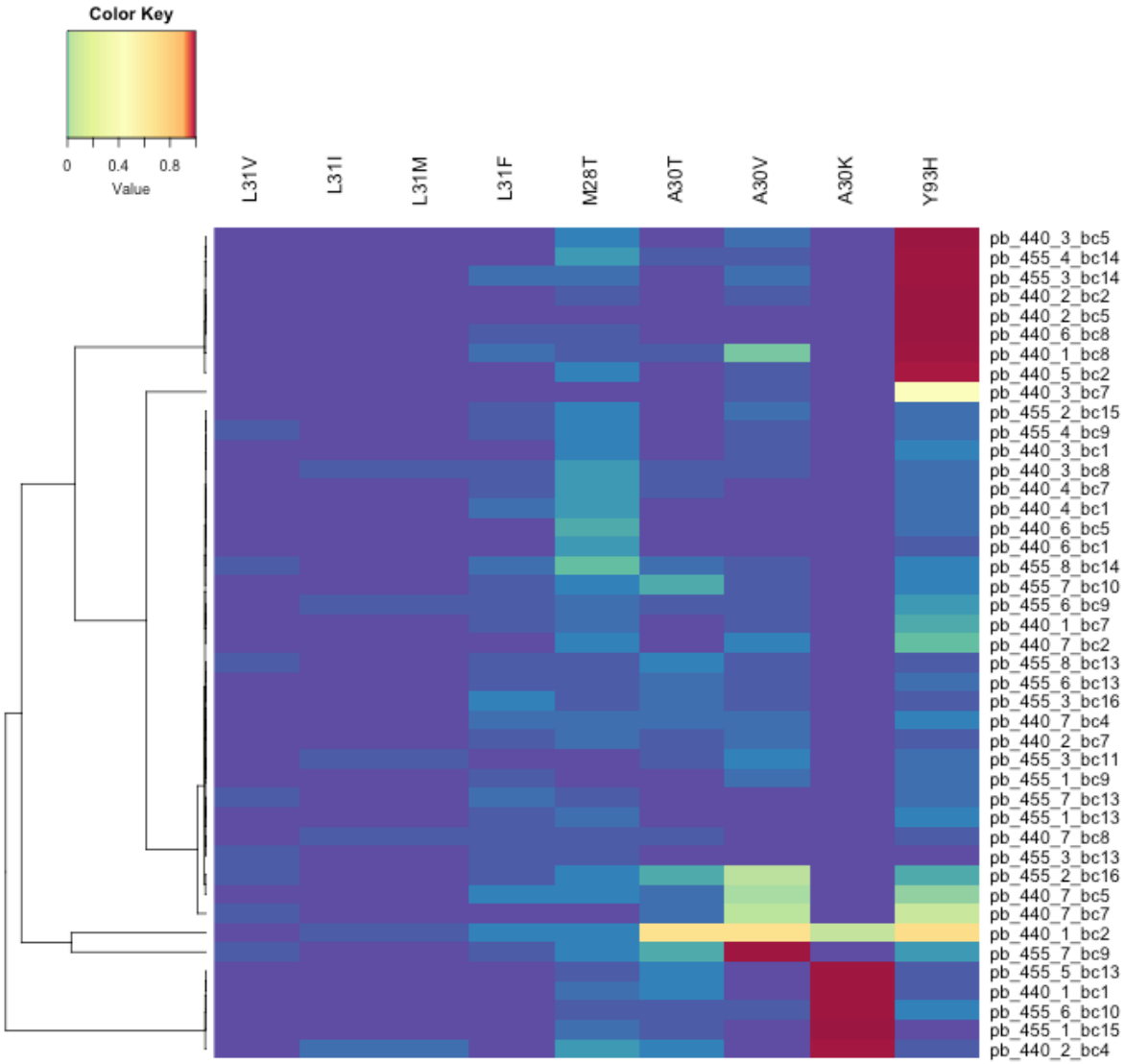

Figure 7: Heat Map of RASs for GT3a samples. Heat map resistance mutation frequencies for the 43 samples with genotype 3a. Sample IDs listed by row, resistance mutations listed by column. The heatmap is clustered by rows, demonstrated as a dendrogram left of the plot. Mutation frequencies coloured from purple/blue (low) to red (high). 11 samples had a prevalence at >1% for the RAS Y93H (column 9), 6 samples with A30K at >1% (column 8), and 4 samples with A30V at >1% (column 7).

4.2.2 Assess Clonal distributions/Viral variants

Sequencing with a long-read technology such as PacBio and the formation of CCS reads allows the user to determine if any RASs are present on the same virus sequence. This can be of great interest to clinicians since some co-existing mutation patterns have been linked to high numbers of fold-change for resistance to certain DAA drugs (Fridell *et al.* 2010, Sorbo *et al.* 2018, Smith *et al.* 2019) So for the patient samples that scored positive for more than one resistance mutation, a clonal distribution table was created by the CLAMP software. These distribution tables were then illustrated in pdf plots of the viral clones, along with the number of reads in the sample with these mutation patterns. In this analysis non-RAS were included to illustrate the potential of exploring possible new relationships that can give high resistance. Two NS5A samples that contained viral variants with mutation combinations that have been shown to increase fold change in research studies, pb_440_8_bc8 (Figure 8) and pb_455_8_bc11 (Figure 9).

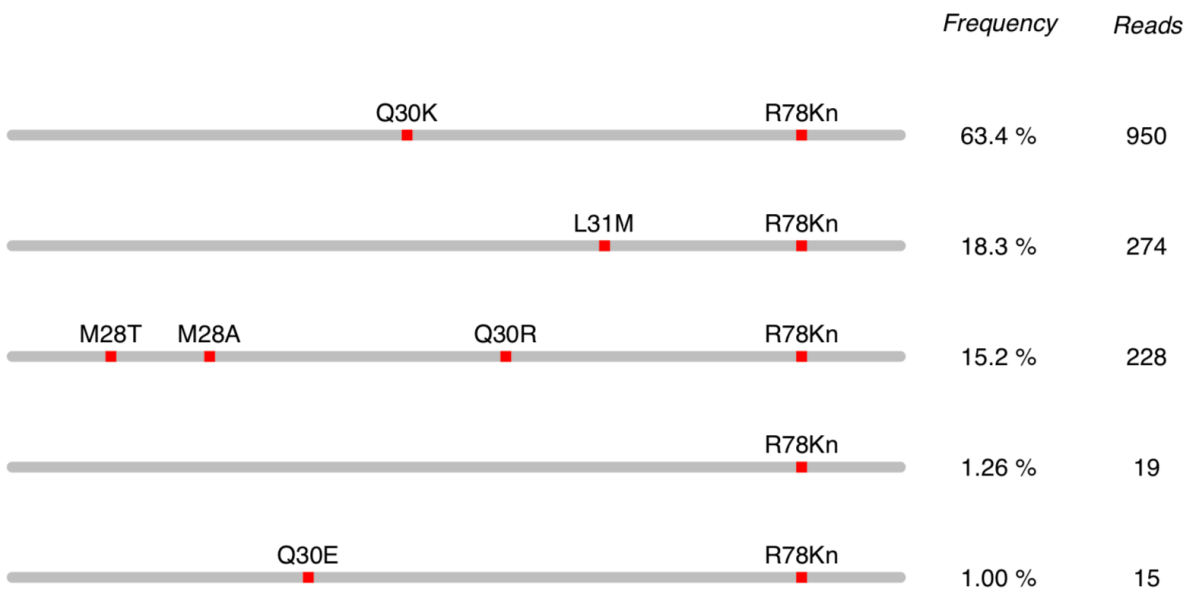

Figure 8. Clonal distribution result from patient sample pb_440_8_bc8, genotype 1a. Showing ratio and read counts of the viral clones in the sample present at at least >1%.'n' is assigned to non-confirmed RAS. The mutation pattern (M28T +Q30R) exists in the sample at a prevalence of 15,2%. This pattern has been associated with a fold-change of 8462 for the NS5A inhibitor Daclatasvir and 3 537 179 for Ombitasvir (Ng *et al.* 2018).

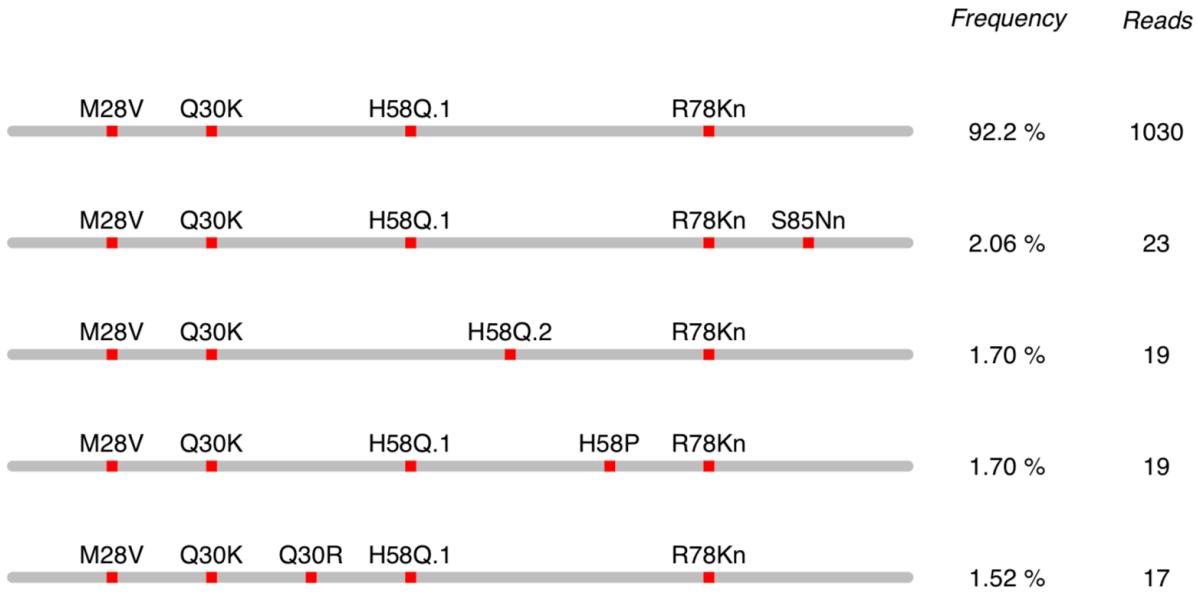

Figure 9. Clonal distribution result from patient sample pb_455_8_bc11, genotype 1a. Showing prevalence and read counts of the viral clones in the sample present at at least >1%. 'n' is assigned to non-confirmed RAS. The mutation pattern (M28V + Q30K) exists in the majority of the variants, while the pattern (M28V + Q30R) exists in the sample at a prevalence of 1,52%. The (M28V + Q30R) pattern has been associated with a fold-change of 350 for the NS5A inhibitor Daclatasvir (Fridell *et al.* 2011).

4.3 Mixed Infections

The last one of the analyses for the project was to investigate whether any of the 123 HCV patient samples contained so called mixed infections of more than one HCV genotype. As previously described this was done using a separate Python script written by the student that utilises NCBI's local command-line version of BLAST, multiple sequence alignment using MUSCLE, and Python module AlignInfo. For each sample the reads are matched against HCV genotype references and grouped by which reference the read has closest similarity to. The reads are then filtered, aligned and a consensus sequence is created for the sample. Out of all samples in the data set, only two had reads originating from more than one genotype, pb_455_8_bc9 and pb_455_8_bc13. For the first one of these samples, the majority genotype was GT1a with 1368 reads matching this type, and the minority was GT3a with 4 reads. Giving the sample a distribution of 99,7% GT1a and 0,3% GT3a. The second mixed infection had the majority genotype GT3a with 1531 reads, and the minority genotype GT1a with 2 reads. The distribution for this sample was 99,87% for the dominant type and 0,13% for the minority type (Figure 10). All genotyping results using the BLAST method corresponded to those from the Sanger approach, except for the mixed samples for which only the majority

types 1a and 3a were identified using the reference method. For all 123 samples, all reads matched to a genotype with an e-value of 0,001 or lower.

Due to the fact that the conventional method used as a reference for the new approach does not have the ability to discern if samples contain mixed HCV infections, some other measure for validating the results of the mixed infection analysis was needed. Therefore 13 mixed infections were artificially constructed *in silico* from the existing samples by sampling a specific number of reads from different genotypes. The artificial samples consisted of between 100 to 1500 reads each and the reads came from two or more genotypes that were present in the original dataset: GT1a, GT1b, GT2b, GT3a and GT4a. These constructed mixed infections were then run with the Python code to see if the program would correctly classify the reads to the different genotypes. The validation was successful for all 13 manufactured samples, every read was correctly assigned to genotype and the distributions of the different genotypes per sample were accurately determined.

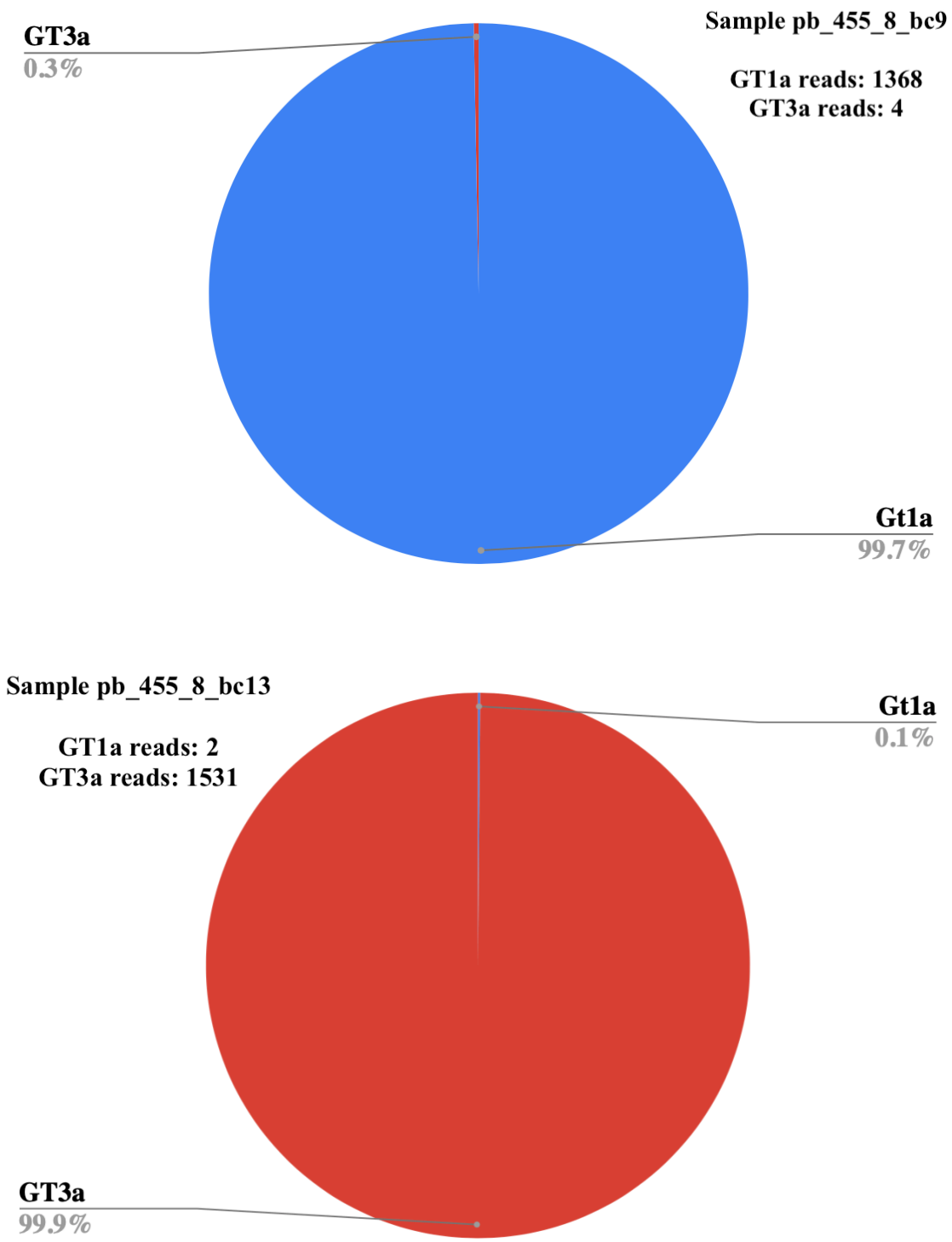

Figure 10. Results from mixed genotype detection analysis of samples pb_455_8_bc9 and pb_4550_8_bc13 (top to bottom). Percentage of reads in a sample matching each genotype (for reads with E-value < 0.001), using Python code, the BLAST+ tool and HCV reference sequences (Table 2). The first sample had the dominant genotype 1a with 1368 reads, and 3a as the minority genotype with 4 reads. The second sample had 3a as the majority genotype (1531 reads) and 1a as the minority (2 reads).

5. Discussion

The use of direct acting antivirals (DAA's) has revolutionised treatment of Hepatitis C leading to SVR in up to 96% of cases (Zhang *et al.* 2016, Bachofner *et al.* 2018) with limited side effects, transforming it into a curable disease. However due to the relatively high price of DAAs, treatment might be deferred until later stages of the infection in high income countries, or not be available at all in lower-middle income countries (Mantovani *et al.* 2016). Resistance development remains an issue which needs handling, resistance development causing treatment failure greatly prolongs treatment and further increases cost. Sanger Sequencing has long been the preferred method for RAS detection, though it can be argued that a new method is needed that has a higher sensitivity. This project was conducted to demonstrate the advantages of using a fully automated bioinformatic approach and long-read sequencing for HCV resistance detection. Using long-reads and the analysis software CLAMP, we have the ability to find RASs present in ratios below the Sanger detection limit, identifying co-mutations and mutation patterns in the same viral clone, and detecting the presence of mixed genotype infections in patient samples.

5.1 Resistance detection

All resistance mutations detected using Sanger sequencing could also be found using the new method, which was expected since PacBio reads generally have a high consensus accuracy due to the creation of circular consensus sequences (CCSs) (Nakano *et al.* 2017). Using the new detection method, a total of 54 RASs were identified in the 123 samples at a prevalence of 15% or above. Two of these RASs were not found in the consensus sequence of the sample given by the reference method using Sanger. These two mutations were found in the same sample (pb_440_8_bc8) located in positions 30 and 31 (Q30R, L31M) at frequencies 18,6% and 19,3% respectively (Table 4), (Figure 4). The sensitivity of Sanger sequencing is generally recognised as 15-20% (Rohlin *et al.* 2009), and so this discrepancy between the two methods is most likely due to Sanger having a limit of detection at approximately 20% for this type of analysis. The lowest prevalence RAS detected by Sanger for this study was a H58Y mutation, which was detected at 20,4% with the new method (Table 4). This further affirms a 20% limit of detection using the Sanger method. The sample with the two discrepant mutations came from a genotype 1a virus and had two other RASs that were detected with Sanger, M28T (20,7%), and Q30K (64,1%). The presence of one majority variant with Q30K at 64,1% most likely prevented the signal of the minority Q30R variant (18,6%) from being detected with the Sanger approach. Both amino acid changes in position 30 are considered clinically relevant for genotype 1a as they cause resistance *in-vitro* to certain DAAs, however

the Q30K RAS is considered susceptible to Elbasvir while Q30R is not (Pawlotsky *et al.* 2018, Sorbo *et al.* 2018). Since the RAS M28T was also positive in this sample, Elbasvir would not have been recommended for treatment either way in this case since M28T is considered resistant to this drug (Pawlotsky *et al.* 2018). It is important to note however that if M28T had not been present in the sample, a regimen might have been chosen based solely on the resistance profile of Q30K, which could have resulted in treatment failure due to the Q30R mutation passing by undetected when using Sanger sequencing.

Another case where this problem of Sanger's limited sensitivity becomes apparent is for patient sample pb_440_3_bc6, for which the RAS M28V (98,8%) was found by both approaches, but a second RAS Q30R with lower ratio (13,3%) was not detected in the consensus sequence from Sanger (Table 4). This sample was genotyped as GT1a and as previously mentioned the Q30R is associated with resistance against many different DAAs for this genotype and may very well affect the success of treatment, while the M28V mutation is considered susceptible to most DAA regimens and does not in general have much bearing on the choice of treatment. The Q30R RAS in this sample has a prevalence just below the 15% limit for what is considered a clinically relevant abundance of a RAS (Pawlotsky 2016), yet the Q30R RAS might very well have increased for this patient since treatment was based on the assumption that M28V was the only RAS present in the sample. This could ultimately lead to treatment failure and cause a relapse of the infection, with the Q30R now being the dominant variant of the virus. Due to the patient data being anonymised before the samples were made available for the project, no follow up of the treatment progress was possible and it could not be verified if the Q30R RAS influenced the outcome.

This study established a detection limit for RAS detection with Sanger at approximately 20%, which is relatively high considering that resistance mutations are considered relevant down to 15% and the limited knowledge of the relevance of lower prevalence mutations. This lack of knowledge stems from the fact that NGS methods and long-read sequencing alternatives have as of yet not been used for HCV resistance detection in clinical practice. Using an approach that allows for detection of variants below the 15% cut-off could prove to be beneficial for RAS detection and HCV treatment, as well as provide the possibility to study how low prevalence viral variants in the NS5A region impact SVR rates *in-vitro* for different treatments. It is important to note that the RAS detected with the new method that were present in abundances below the 15% could not be verified since the reference method used has a detection limit of 15-20%. To completely verify these low abundance RASs, another method for validation would have to be used.

5.2 Mutation patterns and co-mutations

As well as performing genotyping and resistance detection, the CLAMP software performs an analysis step to determine if multiple RASs are present on the same virus sequence. Certain co-existing mutations have been known to greatly increase fold resistance for the viral variant (Fridell *et al.* 2011, Sorbo *et al.* 2018). In this study there were 16 samples with more than one RAS, and for these samples clonal distribution tables were generated. The tables display the viral variants, their frequencies and number of reads containing these mutations. For most of these samples, the detected combinations of mutations have not been linked to any increase in fold resistance.

Two samples were found to contain significant co-mutations, sample pb_440_8_bc8 and pb_455_8_bc11 . The combination found in the first sample was (M28T+Q30R) with a prevalence of 15,2% (Figure 8), a mutation pattern that increases fold resistance against NS5A inhibitor Daclatasvir to 8462 and Ombitasvir to 3 537 179 (Ng et al. 2018). As previously mentioned, the Q30R mutation was not detected in the sample by the Sanger method, while M28T and Q30K were detected. In this particular instance the mutations on their own are considered resistant to most DAA's for the genotype, and so the mutation pattern (M28T+Q30R) on the same sequence would not have affected the choice of regimen either way. For the other sample the co-existing mutation (M28V+Q30R) was found at 1,52% (Figure 9), which causes a fold-change of 350 against Daclatasvir (Fridell *et al.* 2011). Daclatasvir would however not have been chosen for treatment since the single mutation Q30R is not susceptible to the drug, and the mutation Q30K was found at a high prevalence and is also resistant to Daclatasvir.

To further investigate if any mutation patterns occur more often together for certain genotypes, heat maps were created for the samples displaying the intensity of all the RAS frequencies to help discover any trends in the data. Genotype 1a was the most common type among the samples, along with having the longest list of possible resistance mutations. In the heat map for genotype 1a, it could be observed that almost all samples had at least one low (<1%) RASs in position 28, the Q30R RAS, as well as H58Y (Figure 5). Though for the GT1a samples with mutations with higher prevalence RASs, there are no evident patterns of mutations regularly appearing together. The second heat map contains the samples of genotype 1b, for which there were also no apparent patterns in the mutations (Figure 6). The last heat map for genotype 3a did contain a possible trend in mutations co-occurring in low to intermediate frequencies (0,5-70%), the RAS A30V and Y93H (Figure 7). Though for the same two RASs, it can also be seen that for the samples where the mutations reach higher frequencies they are not found together. This might indicate that these RASs are not prone to

occur together on the same viral sequence, but to draw any such conclusions about the (A30V+Y93H) pattern more GT3a samples with one or both of these mutations would need to be studied.

Two cases of co-existing mutations on the same viral sequence were identified in the data set that had the potential of increasing the fold resistance, although they would not have affected the treatment regimens in their samples. These cases do however show the importance of checking for viral variants with multiple RASs, since there are many recorded cases were mutations greatly increase HCV resistance when occurring together in the same viral sequence. Detecting these co-mutations is made possible by the long CCS reads from PacBio that are able to cover the entire NS5A region for each virus sequence multiple times, which gives it a clear advantage over other NGS technologies that have the same sensitivity but shorter reads.

5.3 Mixed genotype infections

The determined genotypes of the PacBio results using CLAMP were coherent with the genotyping done with the Sanger approach, which was one of the quality objectives set before the start of the project to validate the new method. Apart from the standard genotyping, using long-read sequencing and CCS reads allows for detection of mixed HCV genotype infections. This was performed using a Python script and local BLAST against reference sequences for the genotypes. Two of the samples used in the study were found to contain reads from more than one genotype, pb_455_8_bc9 and pb_455_8_bc13. The first sample had the majority type GT1a and four reads of GT3a, and the second had the majority type GT3a and two GT1a reads (Figure 10). These minor HCV subtypes were not found with the old approach, since the result of Sanger is given in the form of a consensus sequence that only conveys the majority variant of the virus. Given that the reference method could not be used to validate the mixed infection analysis, the procedure had to be tested with artificially constructed mixed infections data sets using reads from different genotypes. The analysis successfully genotyped all the mixed reads without failure. This level of effectiveness for the coinfection analysis was expected, since HCV genotypes and subtypes are highly heterogeneous with relatively low sequence similarities (Table 1), and the BLAST algorithm is a well established tool with high accuracy for sequence comparison (Altschul *et al.* 1990).

A mixed infection could prove problematic for the treatment of a patient, since the chosen DAA regimen would be based solely based on the majority type. This would allow the minority virus to take over once the majority type has been successfully treated and cause a relapse of the illness with a new genotype (Pham *et al.* 2010, McNaughton *et al.* 2014). By

using this analysis the user can detect if there is more than one genotype present in an HCV infection, and the analysis also enables the user to identify in which specific subpopulation any RASs are present. Even though HCV coinfections have been documented to occur, very little research has been done on whether or not minority types cause DAA failure and patient relapse (Blackard & Sherman 2007). For this project a cut-off value of 15% was set for the minority type in the infections to it to be considered a true mixed infection, and so the two samples found with reads from different genotypes were not classified as true coinfections. This cut-off was not chosen with any basis in established clinical significance, due to the lack of studies on the subject. The cut-off can be altered by the user, and ultimately it is up to the clinician analysing the results to determine what classifies as a coinfection. Using a method that makes detection of mixed infections possible can be very beneficial since it would not only help clinicians select the best treatment, but also to verify if minority genotypes actually have the ability to cause reinfections.

5.4 Efficiency of the new bioinformatics approach

The full method of analysis is to sequence using PacBio, genotype and screen the resulting reads with CLAMP program, and then use the Python script to find mixed infections and create a consensus sequence. The software is relatively easy to run and non-laborious since it only requires the user to input the sample ID's of interest and to start the programs with specifically chosen or default parameters. The program then performs the analyses automatically and the user can return later to go through the results. When using the old method of resistance detection, it had to be done manually by analysing each peak in the sequence chromatogram and determining whether or not multiple peaks correspond to a mutation in that position of the sequence. The new method would automize this process and be a great improvement in regards to the amount of work for the clinician responsible for the HCV genotyping and RAS detection of patient samples.

An advantage with Sanger Sequencing is that it costs less than NGS technologies. PacBio sequencing and other long-read technologies cost more than using NGS and significantly more than Sanger, which can make it difficult to argue for replacing Sanger with PacBio. One way to reduce the cost of PacBio is by multiplexing a number of samples into one batch using barcodes, which allows multiple samples to be loaded into a single SMRTcell. For this study samples were pooled into batches of 8, which reduces the cost to an eighth of the full price. The most significant cost saving factor of the new method for analysis is the time efficiency and the reduced workload of the automized procedure, resulting in fewer man-hours spent on HCV analysis. The new approach could also prove useful for avoiding treatment failures,

something that would help decrease the cost of having to perform additional investigations and a second line of treatment.

The new analysis has a clear drawback in that each of the mutations in the RAS lists are screened for by checking only one position in the NS5A gene. Changes in amino acids that lead to HCV resistance sometimes require more than one change in the codon, which requires the user to check that both of these positions have mutated to establish that the RAS is present in the sample. This restriction of the analysis to check only one nucleotide position at a time can also cause some ambiguity as to which amino acid the mutation has caused. This problem can be solved by altering the CLAMP software to screen entire codons for mutations, although this task would require time and effort. Since the main scope of the project was to evaluate the efficiency of the new CLAMP based workflow in comparison to the old approach, there was not enough time to modify the software to fix this issue. The analysis also requires the lists of RASs to be consistently updated to reflect which mutations are recognised as clinically relevant by current standards. This step was made slightly easier by addition of the function that creates a consensus sequence of the PacBio reads, making it possible for the user to run the sample against the web-based Geno2pheno to check for new RASs.

6. Conclusion

The aim of this project was to perform a comparative analyses of HCV NS5A sequence data from patient samples sequenced with both Sanger and PacBio SMRT sequencing, to establish if the new method using the CLAMP software is a valid alternative for HCV resistance detection in clinical practice. Each sample was screened for resistance associated substitutions (RASs) and genotyped with both the Sanger and PacBio data, and the results of the two methods were compared. The analysis concluded that the new method was able to correctly genotype the viruses and identify all RASs that the Sanger method had found. Additional RASs were detected using PacBio that Sanger had not found, and all of these mutations were present at frequencies below 20%, placing the limit of detection for Sanger at approximately 20%. The new workflow allowed for RASs to be detected in the sample down to 0,5%, which could help clinicians select the most effective treatment for patients as well as lead to a better understanding of low prevalence viral variants and their relevance in regards to resistance development. Along with performing resistance detection and genotyping, the new method is able to identify the co-existence of mutations on the same virus sequence as well as detecting if a patient sample contains a mix of multiple HCV genotypes. These applications are not possible to perform using the Sanger based method, and they make it possible to find variants

with mutations that together can greatly increase the viruses resistance to certain drugs, and to detect mixed infections that could possibly cause relapse with a new virus genotype.

Using PacBio for sequencing HCV would be more expensive than using the old method of Sanger sequencing. However, the user-friendly and automated process of the PacBio based approach would substantially reduce the number of work hours required to analyse each sample, which would ultimately decrease overall costs. The added possibility of a more accurate detection of co-infections made possible with the new approach improves medical safety, which undoubtedly makes it worth considering as an alternative to the reference method.

7. Acknowledgements

I would like to thank both of my supervisors Adam Ameur and Kåre Bondeson for all of their aid and support in performing this project. Adam for all his advice and guidance concerning bioinformatics and programming issues, and Kåre for sharing his vast knowledge on the subject matter. Their advice through meetings and discussions throughout the course of the project has been invaluable. I also want to thank Midori Kjellin for helping me find crucial information and understanding the routines and procedures, and Irma De La Cruz for performing all of the PCR work that gave us our data sets. Lastly I wish to thank Johan Lennerstrand and Anders Bergqvist for always being available with their teachings as well as to answer my questions.

8. Supplementary materials

The source code for the identify mixed genotype infections analysis can be found in the Github repository: https://github.com/caithaughey/Identify_mixed_HCV_genotype_infections.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

American Association for the Study of Liver Diseases & Infectious disease association of America (AASLD & IDSA). 2018. HCV Guidance: Recommendations for Testing, Managing, and Treating Hepatitis C: HCV Resistance Primer. WWW-document 2018-05-24: https://www.hcvguidelines.org/evaluate/resistance. Accessed 2019-05-07.

Ameur A, Kloosterman WP, Hestand MS. 2019. Single-Molecule Sequencing: Towards Clinical Applications. Trends in Biotechnology 37: 72–85.

Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Research 46: 2159–2168.

Bachofner J, Valli PV, Bergamin I, Kröger A, Künzler P, Baserga A, Braun DL, Seifert B, Moncsek A, Fehr J, Semela D, Magenta L, Müllhaupt B, Terziroli Beretta-Piccoli B, Mertens J, The Swiss Hepatitis C Cohort Study null. 2018. Excellent outcome of direct antiviral treatment for chronic hepatitis C in Switzerland. Swiss Medical Weekly 148: w14560.

Baybayan P, Nolden L. 2017. Abstract 5366: Detection of low-frequency somatic variants using single-molecule, real-time sequencing. Cancer Research 77: 5366–5366.

Bergfors A, Leenheer D, Bergqvist A, Ameur A, Lennerstrand J. 2016. Analysis of hepatitis C NS5A resistance associated polymorphisms using ultra deep single molecule real time (SMRT) sequencing. Antiviral Research 126: 81–89.

Blackard JT, Sherman KE. 2007. Hepatitis C Virus Coinfection and Superinfection. The Journal of Infectious Diseases 195: 519–524.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC bioinformatics 10: 421.

Cavelier L, Ameur A, Häggqvist S, Höijer I, Cahill N, Olsson-Strömberg U, Hermanson M. 2015. Clonal distribution of BCR-ABL1 mutations and splice isoforms by single-molecule long-read RNA sequencing. BMC Cancer, doi 10.1186/s12885-015-1046-y.

Chevaliez S, Rodriguez C, Pawlotsky J-M. 2012. New Virologic Tools for Management of Chronic Hepatitis B and C. Gastroenterology 142: 1303-1313.e1.

De Clercq E. 2014. Current race in the development of DAAs (direct-acting antivirals) against HCV. Biochemical Pharmacology 89: 441–452.

De Francesco R. 1999. Molecular virology of the hepatitis C virus. Journal of Hepatology 31: 47–53.

del Campo JA, Parra-Sánchez M, Figueruela B, García-Rey S, Quer J, Gregori J, Bernal S, Grande L, Palomares JC, Romero-Gómez M. 2018. Hepatitis C virus deep sequencing for sub-genotype identification in mixed infections: A real-life experience. International Journal of Infectious Diseases 67: 114–117.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797.

Fridell RA, Qiu D, Wang C, Valera L, Gao M. 2010. Resistance Analysis of the Hepatitis C Virus NS5A Inhibitor BMS-790052 in an In Vitro Replicon System. Antimicrobial Agents and Chemotherapy 54: 3641–3650.

Fridell RA, Wang C, Sun J-H, O'Boyle DR, Nower P, Valera L, Qiu D, Roberts S, Huang X, Kienzle B, Bifano M, Nettles RE, Gao M. 2011. Genotypic and phenotypic analysis of variants resistant to hepatitis C virus nonstructural protein 5A replication complex inhibitor BMS-790052 in Humans: In Vitro and In Vivo Correlations. Hepatology 54: 1924–1935.

Houghton M. 2016. Towards the Control of Hepatitis C. In: Miyamura T, Lemon SM, Walker CM, Wakita T (ed.). Hepatitis C Virus I: Cellular and Molecular Virology, pp. 3–14. Springer Japan, Tokyo.

International Organization for Standardization. ISO 15189:2012 Medical laboratories - Requirements for quality and competence. WWW-document 2012: https://www.iso.org/standard/56115.html. Accessed 2019-05-09.

Jakobsen JC, Nielsen EE, Feinberg J, Katakam KK, Fobian K, Hauser G, Poropat G, Djurisic S, Weiss KH, Bjelakovic M, Bjelakovic G, Klingenberg SL, Liu JP, Nikolova D, Koretz RL, Gluud C. Direct−acting antivirals for chronic hepatitis C. Cochrane Database of Systematic Reviews 2017, Issue 9.

Kalaghatgi P, Sikorski AM, 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.

Kanwal F, Kramer JR, Ilyas J, Duan Z, El−Serag HB. 2014. HCV genotype 3 is associated with an increased risk of cirrhosis and hepatocellular cancer in a national sample of U.S. Veterans with HCV. Hepatology 60: 98–105.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution 33: 1870–1874.

Lindström I, Kjellin M, Palanisamy N, Bondeson K, Wesslén L, Lannergard A, Lennerstrand J. 2015. Prevalence of polymorphisms with significant resistance to NS5A inhibitors in treatment-naive patients with hepatitis C virus genotypes 1a and 3a in Sweden. Infectious Diseases 47: 555–562.

Lohmann V, Korner F, Koch J-O, Herian U, Theilmann L, Bartenschlager R. 1999. Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. Science; Washington 285: 110–3.

Macdonald A, Harris M. 2004. Hepatitis C virus NS5A: tales of a promiscuous protein. Journal of General Virology 85: 2485–2502.

Maheshwari A, Ray S, Thuluvath PJ. 2008. Acute hepatitis C. The Lancet 372: 321–332.

Manns MP, McHutchison JG, Gordon SC, Rustgi VK, Shiffman M, Reindollar R, Goodman ZD, Koury K, Ling M-H, Albrecht JK. 2001. Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. The Lancet 358: 958–965.

McNaughton AL, Thomson EC, Templeton K, Gunson RN, Leitch ECM. 2014. Mixed Genotype Hepatitis C Infections and Implications for Treatment. Hepatology (Baltimore, Md) 59: 1209.

Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, Pybus OG, Barnes E. 2015. Global distribution and prevalence of hepatitis C virus genotypes. Hepatology 61: 77–87.

Nakano T, Lau GMG, Lau GML, Sugiyama M, Mizokami M. 2012. An updated analysis of hepatitis C virus genotypes and subtypes based on the complete coding region. Liver International 32: 339–345.

Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, Shinzato M, Minami M, Nakanishi T, Teruya K, Satou K, Hirano T. 2017. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Human Cell 30: 149–161.

Ng TI, Pilot-Matias T, Tripathi R, Schnell G, Krishnan P, Reisch T, Beyer J, Dekhtyar T, Irvin M, Lu L, Asatryan A, Campbell A, Yao B, Lovell S, Mensa F, Lawitz EJ, Kort J, Collins C. 2018. Resistance Analysis of a 3-Day Monotherapy Study with Glecaprevir or Pibrentasvir in Patients with Chronic Hepatitis C Virus Genotype 1 Infection. Viruses, doi 10.3390/v10090462.

Patiño-Galindo JÁ, Salvatierra K, González-Candelas F, López-Labrador FX. 2016. Comprehensive Screening for Naturally Occurring Hepatitis C Virus Resistance to Direct-Acting Antivirals in the NS3, NS5A, and NS5B Genes in Worldwide Isolates of Viral Genotypes 1 to 6. Antimicrobial Agents and Chemotherapy 60: 2402–2416.

Paul D, Madan V, Bartenschlager R. 2014. Hepatitis C Virus RNA Replication and Assembly: Living on the Fat of the Land. Cell Host & Microbe 16: 569–579.

Pawlotsky J-M. 2016. Hepatitis C Virus Resistance to Direct-Acting Antiviral Drugs in Interferon-Free Regimens. Gastroenterology 151: 70–86.

Pawlotsky J-M, Negro F, Aghemo A, Berenguer M, Dalgard O, Dusheiko G, Marra F, Puoti M, Wedemeyer H. 2018. EASL Recommendations on Treatment of Hepatitis C 2018. Journal of Hepatology 69: 461–511.

Perales C, Chen Q, Soria ME, Gregori J, Garcia-Cehic D, Nieto-Aponte L, Castells L, Imaz A, Llorens-Revull M, Domingo E, Buti M, Esteban JI, Rodriguez-Frias F, Quer J. 2018. Baseline hepatitis C virus resistance-associated substitutions present at frequencies lower than 15% may be clinically significant. Infection and Drug Resistance 11: 2207–2210.

Pham ST, Bull RA, Bennett JM, Rawlinson WD, Dore GJ, Lloyd AR, White PA. 2010. Frequent multiple hepatitis C virus infections among injection drug users in a prison setting. Hepatology 52: 1564–1572.

Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. 2018. Long reads: their purpose and place. Human Molecular Genetics 27: R234–R241.

Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics 13: 278–289.

Rohlin A, Wernersson J, Engwall Y, Wiklund L, Björk J, Nordling M. 2009. Parallel sequencing used in detection of mosaic mutations: Comparison with four diagnostic DNA screening techniques. Human Mutation 30: 1012–1020.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology 94: 441–448.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74: 5463–5467.

Sarrazin C. 2016. The importance of resistance to direct antiviral drugs in HCV infection in clinical practice. Journal of Hepatology 64: 486–504.

Sarrazin C, Dvory-Sobol H, Svarovskaia ES, Doehle BP, Pang PS, Chuang S-M, Ma J, Ding X, Afdhal NH, Kowdley KV, Gane EJ, Lawitz E, Brainard DM, McHutchison JG, Miller MD, Mo H. 2016. Prevalence of Resistance-Associated Substitutions in HCV NS5A, NS5B, or NS3 and Outcomes of Treatment With Ledipasvir and Sofosbuvir. Gastroenterology 151: 501-512.e1.

Shiffman ML. 2011. Chronic Hepatitis C Virus: Advances in Treatment, Promise for the Future. Springer Science & Business Media. p. 14.

Smith, D.B., Bukh, J., Kuiken, C., Muerhoff, A.S., Rice, C.M., Stapleton, J.T., Simmonds, P. International Committee on Taxonomy of Viruses (ICTV). HCV Classification. 2018. WWW-document: https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/56/hcv-classification. Accessed 2019-04-24.

Smith D, Magri A, Bonsall D, Ip CLC, Trebes A, Brown A, Piazza P, Bowden R, Nguyen D, Ansari MA, Simmonds P, Barnes E, STOP-HCV Consortium. 2019. Resistance analysis of genotype 3 hepatitis C virus indicates subtypes inherently resistant to nonstructural protein 5A inhibitors. Hepatology (Baltimore, Md) 69: 1861–1872.

Sorbo MC, Cento V, Di Maio VC, Howe AYM, Garcia F, Perno CF, Ceccherini-Silberstein F. 2018. Hepatitis C virus drug resistance associated substitutions and their clinical relevance: Update 2018. Drug Resistance Updates 37: 17–39.

Sung H, Chang M, Saab S. 2011. Management of Hepatitis C Antiviral Therapy Adverse Effects. Current Hepatitis Reports 10: 33–40.

Westbrook RH, Dusheiko G. 2014. Natural history of hepatitis C. Journal of Hepatology 61: S58–S68.

Wilkins T, Malcolm JK, Raina D, Schade RR. 2010. Hepatitis C: Diagnosis and Treatment. American Family Physician 81: 1351–1357.

World Health Organisation (WHO). 2018. Hepatitis C. WWW-document 2018-07-18: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c. Accessed April 17, 2019.

Wyles DL, Luetkemeyer AF. 2017. Understanding Hepatitis C Virus Drug Resistance: Clinical Implications for Current and Future Regimens. Topics in Antiviral Medicine 25: 103–109.

Zhang J, Nguyen D, Hu K-Q. 2016. Chronic Hepatitis C Virus Infection: A Review of Current Direct-Acting Antiviral Treatment Strategies. North American journal of medicine & science 9: 47–54.

Appendices

Appendix A.

Table A1. All 123 HCV samples used in the study. The first column is the ID number for the samples when sequenced with PacBio. The second column shows the genotype of the samples, and the third column is the resistance mutations found using Sanger. The fourth column shows the RASs found above 15% for that sample with PacBio, and the fifth column shows RASs found below 15% prevalence. The last column displays all the Non-RASs identified.

PacBio ID	GT[1]	RASs (Sanger)	RASs PacBio (>15%)	RASs PacBio (<15%)	Non-RASs
pb_455_1_bc11	1a			M28V, L31M, M28T, M28A	R78Kn, S85Nn, R44Kn, V75An
pb_440_1_bc6	1a				R78Kn, R48Qn, R44Kn
pb_440_4_bc2	1a				V75An, R78Kn
pb_440_1_bc1	3a	A30K	A30K		
pb_440_2_bc1	1a	H58P	H58P		R44Kn
pb_440_3_bc1	3a				A62Sn, H85Yn
pb_440_4_bc1	3a				A62Sn
pb_440_5_bc1	1a				R78Kn, R44Kn
pb_440_6_bc1	3a				A62Sn, H85Yn
pb_440_7_bc1	1a				S85Nn, R44Kn, R78Kn
pb_440_8_bc1	1a	H58Y	H58Y		V75An, R78Kn, S85Nn
pb_440_5_bc2	3a	Y93H	Y93H		
pb_440_1_bc2	3a	Y93H	Y93H	A30K	A62Sn, H85Yn
pb_440_2_bc2	3a	Y93H	Y93H		A62Sn, T64Sn, H85Yn, T64An
pb_440_3_bc2	1a			M28V	S85Nn, R78Kn, V75An
pb_440_6_bc2	1b	Y93H	Y93H		
pb_440_7_bc2	3a				A62Sn
pb_440_8_bc2	1a				R78Kn, V75An
pb_440_1_bc3	1a	H58R	H58R		R78Kn, R44Kn
pb_440_2_bc3	1a	M28V, Q30R	M28V, Q30R	M28A, M28T	R78Kn, F36Ln
pb_440_3_bc3	1a			M28V	A92Pn, V75An, F36Ln, R44Kn, S85Nn

pb_440_4_bc3	1a				V75An, S85Nn, R44Kn
pb_440_5_bc3	1a				V75An, R44Kn
pb_455_2_bc11	1a			H58Q	R78Kn, R44Kn
pb_440_6_bc3	1a				R78Kn, R44Kn
pb_440_7_bc3	1a				R78Kn, V75An, R81Wn, R44Kn
pb_440_8_bc3	1a				R78Kn, R44Kn, S85Nn
pb_440_1_bc4	1a	M28V	M28V		R44Kn, R78Kn
pb_440_2_bc4	3a	A30K	A30K		T64An, A62Sn
pb_440_4_bc4	1a				R78Kn, R48Qn, R44Kn
pb_440_1_bc5	1a			M28V	R78Kn, R48Qn
pb_440_2_bc5	3a	Y93H	Y93H		T64An, A62Tn
pb_440_3_bc5	3a	Y93H	Y93H		A62Sn, H85Yn
pb_440_4_bc5	1a	M28V	M28V		R78Kn, R48Qn
pb_440_5_bc5	1a				R78Kn, R44Kn, S85Nn
pb_440_6_bc5	3a				T64An, H85Yn, A62Sn
pb_440_7_bc5	3a				A62Sn, H85Yn
pb_440_5_bc4	1a	M28V	M28V		R78Kn, S85Nn
pb_440_6_bc4	1a				R78Kn
pb_440_7_bc4	3a				T64Sn, A62Sn, H85Yn
pb_440_8_bc4	1a	M28V	M28V		R78Kn
pb_440_2_bc6	1a			H58R	R48Qn, R78Kn
pb_440_3_bc6	1a	M28V	M28V	Q30R	R78Kn, R48Qn, V75An
pb_440_4_bc6	1b	L31M	L31M		
pb_440_5_bc6	1a				F36Ln, R78Kn
pb_440_6_bc6	1a			M28V	S85Nn, R78Kn
pb_440_1_bc7	3a				A62Sn, H85Yn
pb_440_2_bc7	3a				T64An, A62Tn, H85Yn, A62Sn
pb_440_3_bc7	3a	Y93H	Y93H		A62Sn, H85Yn
pb_440_4_bc7	3a				T64Sn, H85Yn, A62Tn

pb_440_5_bc7	1a				R44Kn, R78Kn
pb_440_6_bc7	1a	M28V	M28V		R44Kn, R78Kn
pb_440_7_bc7	3a			Y93H, A30V	A62Sn, H85Yn
pb_440_8_bc7	1a			H58R	R78Kn, V75An
pb_440_1_bc8	3a	Y93H	Y93H		H85Yn, T64An
pb_440_2_bc8	1b	Y93H, L31M	Y93H, L31M		
pb_440_3_bc8	3a				H85Yn, T64An, A62Tn
pb_440_4_bc8	1a				R78Kn, S85Nn, V75An, R44Kn
pb_440_5_bc8	4a				
pb_440_6_bc8	3a	Y93H	Y93H		T64An, A62Sn
pb_440_7_bc8	3a				A62Sn
pb_440_8_bc8	1a	Q30K, M28T	Q30K, M28T, L31M, Q30R	Q30E	R78Kn
pb_455_1_bc9	3a				H85Yn, A62Sn
pb_455_2_bc9	1a			Q30R	R44Kn, R78Kn, V75An
pb_455_3_bc9	1a				R78Kn
pb_455_4_bc9	3a				H85Yn, T64An, A62Tn
pb_455_5_bc9	1a				R44Kn, R78Kn, V75An
pb_455_6_bc9	3a				A62Sn, H85Yn
pb_455_7_bc9	3a	A30V	A30V		A62Sn
pb_455_8_bc9	1a	A92T	A92T	Y93N	R78Kn, R44Kn
pb_455_1_bc10	1a	Q30R	Q30R		R78Kn, R48Qn, R81Wn, S85Nn
pb_455_2_bc10	1a				V75An, S85Nn
pb_455_3_bc10	1a				
pb_455_4_bc10	1a				R78Kn, R44Kn, R48Qn, F36Ln, V75An
pb_455_5_bc10	1a				V75An, S85Nn
pb_455_1_bc12	1a	H58P	H58P		R78Kn, R44Kn
pb_455_6_bc10	3a	A30K	A30K		A62Sn
pb_455_7_bc10	3a				T64An, A62Sn, A62Tn
pb_455_8_bc10	2b				

pb_455_3_bc11	3a				H85Yn, A62Sn
pb_455_4_bc11	1a	M28V	M28V	L31M	R78Kn, R48Qn, R44Kn
pb_455_5_bc11	1a	Y93H, Q30H	Y93H, Q30H		R78Kn, S85Nn
pb_455_6_bc11	1a			M28V	R78Kn, R48Qn
pb_455_7_bc11	1a				R78Kn, R81Wn
pb_455_8_bc11	1a	M28V, Q30K, H58Q	M28V, Q30K, H58Q	H58P, Q30R	R78Kn, S85Nn
pb_455_2_bc12	1a				R78Kn
pb_455_3_bc12	1a				R78Kn
pb_455_4_bc12	1a				R78Kn
pb_455_5_bc12	1a				R44Kn, R78Kn
pb_455_6_bc12	1a	Q30H, Y93H	Q30H, Y93H		R78Kn, S85Nn
pb_455_7_bc12	1a				R48Qn, R78Kn
pb_455_8_bc12	1a				R78Kn
pb_455_1_bc13	3a				T64An, A62Tn, H85Yn
pb_455_2_bc13	1b				
pb_455_3_bc13	3a				A62Sn
pb_455_4_bc13	1a				R78Kn
pb_455_5_bc13	3a	A30K	A30K		A62Sn, H85Yn
pb_455_6_bc13	3a				H85Yn, A62Sn
pb_455_7_bc13	3a				A62Tn
pb_455_8_bc13	3a				T64An, H85Yn, A62Sn
pb_455_1_bc14	2b				
pb_455_2_bc14	1a				R78Kn
pb_455_3_bc14	3a	Y93H	Y93H		A62Sn
pb_455_4_bc14	3a	Y93H	Y93H		H85Yn, A62Sn
pb_455_5_bc14	1a	Q30R, Y93F	Q30R, Y93F		R78Kn, R48Qn
pb_455_6_bc14	1a				R78Kn
pb_455_7_bc14	1a	Y93F	Y93F	Q30K, Y93H, Q30R	R78Kn, R44Kn, V75An
pb_455_8_bc14	3a				T64An, A62Sn, H85Yn
pb_455_1_bc15	3a	A30K	A30K		A62Sn
pb_455_2_bc15	3a				A62Sn, H85Yn

pb_455_3_bc15	1a	M28V	M28V		R78Kn, V75An, R44Kn, S85Nn
pb_455_4_bc15	1a				R44Kn
pb_455_6_bc15	1a			H58R	R78Kn, S85Nn
pb_455_7_bc15	1a	H58P	H58P		R78Kn, V75An, F36Ln
pb_455_8_bc15	1a			H58P	R44Kn, S85Nn, R78Kn, V75An
pb_455_1_bc16	1a	H58Q, H58P	H58Q, H58P		R78Kn, R48Qn, K68Rn, V75An
pb_455_2_bc16	3a			A30V	A62Sn, H85Yn
pb_455_3_bc16	3a				A62Tn
pb_455_4_bc16	1a	M28V	M28V		R44Kn, R78Kn
pb_455_5_bc16	1a				R44Kn, R78Kn, R48Qn
pb_455_6_bc16	1a				R44Kn, R78Kn, R48Qn
pb_455_7_bc16	1a	H58Y	H58Y	M28V, H58R	R78Kn, V75An
pb_455_8_bc16	1a				R78Kn

[1] Genotype classified with both Sanger and PacBio sequencing.

Appendix B.

Table B1. The mutation tables used in resistance detection for the HCV genotypes. Displaying clinically relevant RASs and non-RASs in the NS5A gene. The order on the mutation lists are GT1a, GT1b, GT2b, GT3a and lastly GT4a. The specific mutation site for the nucleotide is shown in the sequences inside hard brackets, were the first character is the wild-type/non-mutated nucleotide and the second is the mutated nucleotide. The mutation position is flanked by 20 bases left and right. The resulting RASs given by the nucleotide change are shown in the second column.RASs are labelled by first the symbol for the wild-type amino acid, then the amino acid position, and last the mutated amino acid. The RASs that occur due to combinations of two single mutations are shown in brackets. The non-confirmed RASs are denoted with an ”n” at the end.

Sequence	RAS
GT1a	
CTGGCTGAAAGCCAAGCTCA[T/C]GCCACAACTGCCTGGGATTC	M28T
CCTGGCTGAAAGCCAAGCTC[A/G]TGCCACAACTGCCTGGGATT	M28V
CTGGCTGAAAGCCAAGCTCG[T/C]GCCACAACTGCCTGGGATTC	M28A (M28V+V28A)
CTGGCTGAAAGCCAAGCTCA[T/A]GCCACAACTGCCTGGGATTC	M28K
TGGCTGAAAGCCAAGCTCAA[G/C]CCCACAACTGCCTGGGATTC	M28N.1 (M28K+K28N)
TGGCTGAAAGCCAAGCTCAA[G/T]CCCACAACTGCCTGGGATTC	M28N.2 (M28K+K28N)
TGAAAGCCAAGCTCATGCCA[C/G]AACTGCCTGGGATTCCCTTT	Q30E
AAAGCCAAGCTCATGCCACA[A/C]CTGCCTGGGATTCCCTTTGT	Q30H.1
AAAGCCAAGCTCATGCCACA[A/T]CTGCCTGGGATTCCCTTTGT	Q30H.2
TGAAAGCCAAGCTCATGCCA[C/A]AACTGCCTGGGATTCCCTTT	Q30K
GAAAGCCAAGCTCATGCCAC[A/G]ACTGCCTGGGATTCCCTTTG	Q30R
GCCAAGCTCATGCCACAACT[G/A]CCTGGGATTCCCTTTGTGTC	L31L
AAGCCAAGCTCATGCCACAA[C/A]TGCCTGGGATTCCCTTTGTG	L31M
AAGCCAAGCTCATGCCACAA[C/G]TGCCTGGGATTCCCTTTGTG	L31V
CAAGCTCATGCCACAACTGC[C/T]TGGGATTCCCTTTGTGTCCT	P32L
GCATTATGCACACTCGCTGC[C/G]ACTGTGGAGCTGAGATCACT	H58D
GCATTATGCACACTCGCTGC[C/T]ACTGTGGAGCTGAGATCACT	H58Y
CATTATGCACACTCGCTGCC[A/G]CTGTGGAGCTGAGATCACTG	H58R
ATTATGCACACTCGCTGCCA[C/A]TGTGGAGCTGAGATCACTGG	H58Q.1
ATTATGCACACTCGCTGCCA[C/G]TGTGGAGCTGAGATCACTGG	H58Q.2
CATTATGCACACTCGCTGCC[A/C]CTGTGGAGCTGAGATCACTG	H58P
GTGGGACGTTCCCCATTAAC[G/A]CCTACACCACGGGCCCCTGT	A92T
GACGTTCCCCATTAACGCCT[A/G]CACCACGGGCCCCTGTACTC	Y93C
GGACGTTCCCCATTAACGCC[T/C]ACACCACGGGCCCCTGTACT	Y93H

GGACGTTCCCCATTAACGCC[T/A]ACACCACGGGCCCCTGTACT	Y93N
GACGTTCCCCATTAACGCCT[A/C]CACCACGGGCCCCTGTACTC	Y93S
GACGTTCCCCATTAACGCCT[A/T]CACCACGGGCCCCTGTACTC	Y93F
CTTTAAGACCTGGCTGAAAG[C/G]CAAGCTCATGCCACAACTGC	A25Gn
AGCTCATGCCACAACTGCCT[G/C]GGATTCCCTTTGTGTCCTGC	G33Rn
CACAACTGCCTGGGATTCCC[T/C]TTGTGTCCTGCCAGCGCGGG	F36Ln
GTCCTGCCAGCGCGGGTATA[G/A]GGGGGTCTGGCGAGGAGACG	R44Kn
CGGGTATAGGGGGGTCTGGC[G/A]AGGAGACGGCATTATGCACA	R48Qn
GGGGGGTCTGGCGAGGAGAC[G/T]GCATTATGCACACTCGCTGC	G51Cn
TGAGATCACTGGACATGTCA[A/G]AAACGGGACGATGAGGATCG	K68Rn
AAACGGGACGATGAGGATCG[T/C]CGGTCCTAGGACCTGCAGGA	V75An
AAAACGGGACGATGAGGATC[G/A]TCGGTCCTAGGACCTGCAGG	V75In
GATGAGGATCGTCGGTCCTA[G/A]GACCTGCAGGAACATGTGGA	R78Kn

GT1b

CTGGCTCCAGTCCAAGCTCC[T/C]GCCGCGATTGCCGGGAGTCC	L28P
CCTGGCTCCAGTCCAAGCTC[C/A]CGCCGCGATTGCCGGGAGTC	L28T (L28P+P28T)
CCAGTCCAAGCTCCTGCCGC[G/A]ATTGCCGGGAGTCCCCTTCT	R30Q
TCCAAGCTCCTGCCGCGATT[G/A]CCGGGAGTCCCCTTCTTCTC	L31L
AGTCCAAGCTCCTGCCGCGA[T/A]TGCCGGGAGTCCCCTTCTTC	L31M
AGTCCAAGCTCCTGCCGCGA[T/G]TGCCGGGAGTCCCCTTCTTC	L31V
AGTCCAAGCTCCTGCCGCGA[T/A]TACCGGGAGTCCCCTTCTTC	L31I
TCCAAGCTCCTGCCGCGATT[G/T]CCGGGAGTCCCCTTCTTCTC	L31F.1
TCCAAGCTCCTGCCGCGATT[G/C]CCGGGAGTCCCCTTCTTCTC	L31F.2
CAAGCTCCTGCCGCGATTGC[C/T]GGGAGTCCCCTTCTTCTCAT	P32L
GAACATTCCCCATTAACGCG[T/C]ACACCACGGGCCCCTGCACG	Y93H
GAACATTCCCCATTAACGCG[T/A]ACACCACGGGCCCCTGCACG	Y93N
AACATTCCCCATTAACGCGT[A/C]CACCACGGGCCCCTGCACGC	Y93S
AACATTCCCCATTAACGCGT[A/G]CACCACGGGCCCCTGCACGC	Y93C

GT2b

GAACCTTCCCCATTAATTGC[T/C]ACACAGAAGGGCCTTGCGTG	Y93H

GT3a

ATGGCTCTCTGCTAAGATTA[T/C]GCCAGCGCTCCCTGGGCTGC	M28T
TCTCTGCTAAGATTATGCCA[G/A]CGCTCCCTGGGCTGCCCTTC	A30T

CTCTGCTAAGATTATGCCAA[C/A]GCTCCCTGGGCTGCCCTTCA	A30K (A30T+T30K)
CTCTGCTAAGATTATGCCAG[C/T]GCTCCCTGGGCTGCCCTTCA	A30V
CTGCTAAGATTATGCCAGCG[C/T]TCCCTGGGCTGCCCTTCATT	L31F
CTGCTAAGATTATGCCAGCG[C/A]TCCCTGGGCTGCCCTTCATT	L31I
CTGCTAAGATTATGCCAGCG[C/G]TCCCTGGGCTGCCCTTCATT	L31V
CTGCTAAGATTATGCCAGCG[C/A]TGCCTGGGCTGCCCTTCATT	L31M
GTACTTTCCCCATCAATGAG[T/C]ACACCACCGGACCCAGCACA	Y93H
AGATTATGCCAGCGCTCCCT[G/C]GGCTGCCCTTCATTTCCTGT	G33Rn
TGCCAGCGCTCCCTGGGCTG[C/A]CCTTCATTTCCTGTCAAAAG	P35Tn
GTGTGATGTCAACACGCTGT[C/G]CTTGCGGGGCAGCAATAACT	P58An
CACGCTGTCCTTGCGGGGCA[G/T]CAATAACTGGCCATGTGAAG	A62Sn
CACGCTGTCCTTGCGGGGCA[G/A]CAATAACTGGCCATGTGAAG	A62Tn
GTCCTTGCGGGGCAGCAATA[A/G]CTGGCCATGTGAAGAACGGG	T64An
GTCCTTGCGGGGCAGCAATA[A/T]CTGGCCATGTGAAGAACGGG	T64Sn
TGAAGAACGGGTCCATGCGG[C/A]TTGCAGGGCCGCGTACATGT	L74In
GTACATGTGCTAACATGTGG[C/T]ACGGTACTTTCCCCATCAAT	H85Yn
TACTTTCCCCATCAATGAGT[A/G]CACCACCGGACCCAGCACAC	Y93C
GTACTTTCCCCATCAATGAG[T/A]ACACCACCGGACCCAGCACA	Y93N
GT4a	
CGTGCTTAAAAGCAAAGTTC[G/A]TGCCCTTAATGCCAGGCATC	V28M
TAAAAGCAAAGTTCGTGCCC[C/A]TAATGCCAGGCATCCCCCTC	L30I
TAAAAGCAAAGTTCGTGCCC[C/G]TAATGCCAGGCATCCCCCTC	L30V
AAAAGCAAAGTTCGTGCCCT[T/A]AATGCCAGGCATCCCCCTCC	L30stop.1
TAAAAGCAAAGTTCGTGCCC[C/T]TAATGCCAGGCATCCCCCTC	L30L
AAAAGCAAAGTTCGTGCCCT[T/G]AATGCCAGGCATCCCCCTCC	L30stop.2
GAACCTTCCCCATCAATGCC[T/C]ACACCACAGGCCCTGGTGTA	Y93H
AACCTTCCCCATCAATGCCT[A/G]CACCACAGGCCCTGGTGTAC	Y93C
ACCTTCCCCATCAATGCCTA[C/G]ACCACAGGCCCTGGTGTACC	Y93stop.1
ACCTTCCCCATCAATGCCTA[C/A]ACCACAGGCCCTGGTGTACC	Y93stop.2

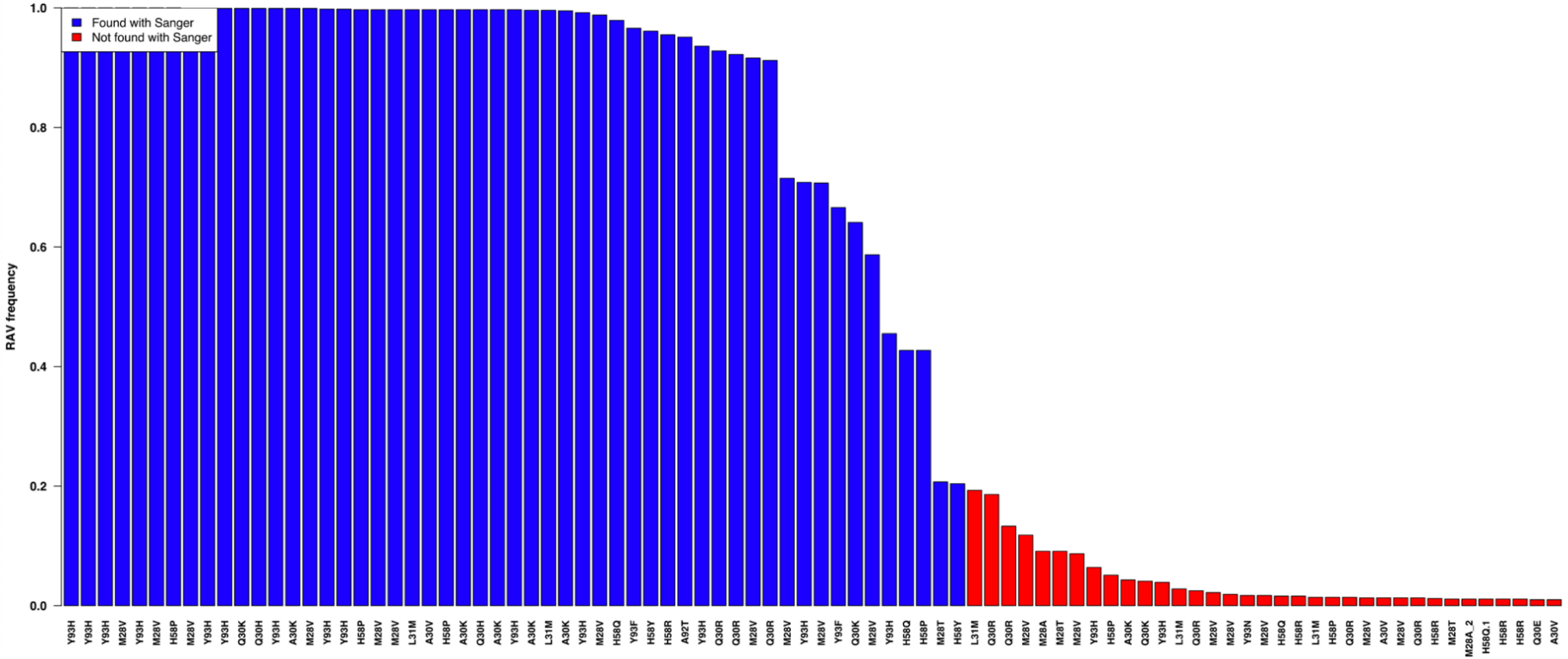

Figure C1. Bar chart displaying all RASs pooled together from the 123 samples above 1% frequency. The blue bars were found with both the PacBio and Sanger methods, while the bars in red were not found using Sanger. The cut-off between these groups correlate with a detection limit of 20% for Sanger sequencing. There are a total number of 88 RASs in the chart.