Universitat Oberta de Catalunya

uoc.edu

Optimització d'un *pipeline* bioinformàtic per a l'anàlisi de qualitat i la classificació genotípica del virus de l'Hepatitis B (VHB) a partir de dades de seqüenciació NGS.

Alicia Aranda Fernandez

Màster en Bioinformàtica i Bioestadística

Àrea 2

Jose Luis Mosquera Mayo

(Carles Ventura Royo)

2 de juny de 2022

FITXA DEL TREBALL FINAL

Títol del treball:	Optimització d'un *pipeline* bioinformàtic per a l'anàlisi de qualitat i la classificació genotípica del virus de l'Hepatitis B (VHB) a partir de dades de seqüenciació NGS.
Nom de l'autor/a:	Alicia Aranda Fernandez
Nom del consultor/a:	Jose Luis Mosquera Mayo
Nom del PRA:	Carles Ventura Royo
Data d'entrega (mm/aaaa):	06/2022
Titulació:	Màster en Bioinformàtica i Bioestadística
Àrea del Treball Final:	Treball de Fi de Màster (Àrea 2)
Idioma del treball:	Català
Nombre de crèdits:	15
Paraules clau	VHB, anàlisi de qualitat, genotipat

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball.*

El virus de l'Hepatitis B (VHB) és un agent infecciós de la família *Hepadnaviridae*, la qual integra diversos agents vírics amb genoma de DNA bicatenari retrotranscrit causants d'infeccions hepàtiques en aus i mamífers. Actualment és un problema important de salut global tot i disposar d'una vacuna efectiva. Tenint en compte les implicacions clíniques i epidemiològiques dels diferents genotips existents del VHB, és important la seva determinació, ja que poden esdevenir molt útils en la predicció del risc de morbiditat i l'orientació en el tractament de la infecció. En el present projecte s'ha elaborat un paquet de R a partir de la revisió i simplificació d'un *pipeline* existent dissenyat especialment per a l'anàlisi de qualitat de dades de seqüenciació massiva i el genotipat del VHB. L'aplicació del paquet construït i del *pipeline* original sobre un mateix conjunt de dades ha permès corroborar la correcta implementació de les funcions definides, ja que els resultats obtinguts en ambdues aproximacions han estat idèntics. A més, la simplificació realitzada permet als usuaris sense formació en programació dur a terme fàcilment diversos anàlisis rellevants en l'àmbit de la recerca. El paquet obtingut facilita l'anàlisi de qualitat de dades provinents de seqüenciació amb Illumina, de manera que no només es pot realitzar el genotipat posterior del VHB a partir de dues regions del seu genoma, si no que addicionalment es podrien realitzar altres estudis, per exemple de la diversitat de les quasiespècies.

Abstract (in English, 250 words or less):

Hepatitis B virus (HBV) is an infectious agent of the *Hepadnaviridae* family, which includes reverse transcribing double-stranded DNA viruses that cause liver infections in birds and mammals. It is currently a major global health problem despite having an effective vaccine. Given the clinical and epidemiological implications of different HBV genotypes, it is important to determine it in infected patients, as it can be useful in predicting the risk of morbidity and guiding infection treatment. An R-package has been developed in the present project based on the revision and simplification of an existing pipeline specifically designed for the quality analysis of next-generation sequencing raw data and HBV genotyping. The application of the built-in package and the original pipeline on the same dataset has corroborated the correct implementation of the defined functions since the results obtained are identical in both approaches. In addition, the provided simplification allows non-programming users to easily carry out relevant research studies. The package obtained facilitates Illumina next-generation sequencing data quality analysis, so that not only the subsequent genotyping can be performed from two regions of HBV genome, but also other studies could be performed, i.e., quasispecies diversity.

Índex

Llista de figures

Llista de taules

1 Resum

El virus de l'Hepatitis B (VHB) és un agent infecciós de la família *Hepadnaviridae*, la qual integra diversos agents vírics amb genoma de DNA bicatenari retrotranscrit causants d'infeccions hepàtiques en aus i mamífers. Actualment és un problema important de salut global tot i disposar d'una vacuna efectiva. Tenint en compte les implicacions clíniques i epidemiològiques dels diferents genotips existents del VHB, és important la seva determinació, ja que poden esdevenir molt útils en la predicció del risc de morbiditat i l'orientació en el tractament de la infecció. En el present projecte s'ha elaborat un paquet de R a partir de la revisió i simplificació d'un *pipeline* existent dissenyat especialment per a l'anàlisi de qualitat de dades de seqüenciació massiva i el genotipat del VHB. L'aplicació del paquet construït i del *pipeline* original sobre un mateix conjunt de dades ha permès corroborar la correcta implementació de les funcions definides, ja que els resultats obtinguts en ambdues aproximacions han estat idèntics. A més, la simplificació realitzada permet als usuaris sense formació en programació dur a terme fàcilment diversos anàlisis rellevants en l'àmbit de la recerca. El paquet obtingut facilita l'anàlisi de qualitat de dades provinents de seqüenciació amb Illumina, de manera que no només es pot realitzar el genotipat posterior del VHB a partir de dues regions del seu genoma, si no que addicionalment es podrien realitzar altres estudis, per exemple de la diversitat de les quasiespècies.

2 Introducció

2.1 Context i justificació del treball

La infecció per part del virus de l'Hepatitis B (VHB) és un problema important de salut global. Tot i que existeix una vacuna segura que proporciona una protecció de la malaltia d'un 98-100%, al 2019 296 milions de persones patien la infecció crònica per VHB, la qual comporta un elevat risc de desenvolupar malalties hepàtiques com cirrosi i carcinoma hepatocel·lular (CHC) que donen lloc a una important càrrega de morbimortalitat (820.000 defuncions al 2019) [1].

El VHB és un agent infecciós representant de la família vírica *Hepadnaviridae*, la qual integra diversos virus amb genoma de DNA bicatenari retrotranscrit causants d'infeccions hepàtiques en aus i mamífers [2]. Presenta un genoma de 3,2 kb de DNA parcialment de doble cadena, amb una cadena negativa completa i una cadena positiva de longitud variable que conformen una estructura circular relaxada (rcDNA) [2]. Aquest genoma està constituït per 4 marcs oberts de lectura (ORFs) altament solapats entre ells (fins a un 67%) [2, 3]:

- **P** → codifica la polimerasa viral, que presenta activitat DNA polimerasa, transcriptasa inversa i RNasa H.
- **S** → codifica 3 glicoproteïnes de superfície, que formen l'envolta i l'antigen de superfície (HBsAg) i permeten l'entrada del virus als hepatòcits.
- **C** → codifica la proteïna o antigen *core* (HBcAg), component estructural de la nucleocàpsida, i l'antigen e (HBeAg) immunomodulador, que no forma part de la partícula viral.
- **X** → codifica la proteïna multifuncional i reguladora X (HBx), la qual no constitueix la partícula viral.

L'elevada variabilitat genètica que presenta el VHB és deguda a fenòmens de substitució nucleotídica i recombinació, els quals venen donats en part pel complex mecanisme de replicació que presenta [2]. Un cop les partícules virals s'uneixen de forma específica als hepatòcits, es produeix la seva endocitosi seguida de l'eliminació de l'envolta, de manera que la nucleocàpsida es transporta al porus nuclear. El genoma en forma de rcDNA del virus s'allibera al nucleoplasma, on diversos factors de reparació cel·lular el transformen en el que es coneix com DNA circular covalentment tancat (cccDNA) [2, 4]. D'aquesta forma, el cccDNA persisteix al nucli de la cèl·lula infectada associat a histones de l'hoste i proteïnes virals, en forma de minicromosoma episomal [2, 4]. Aquest actua com a motlle a partir del qual la RNA polimerasa II genera diferents RNAs virals: RNA pre-genòmic (pgRNA) i preCore (que donarà lloc al HBeAg), els quals són més llarg que el propi genoma víric, i altres mRNAs subgenòmics [2].

El pgRNA resulta clau en el procés de replicació del VHB, ja que permet generar les proteïnes *core* i la polimerasa viral que juntament amb el mateix pgRNA formen les nucleocàpsides víriques [2]. Un cop encapsidat, es dona la retrotranscripció del pgRNA a rcDNA per part de la polimerasa; aquest procés resulta complex i implica diverses translocacions de la polimerasa al llarg del motlle, de manera que es poden produir diversos errors. Concretament, el pgRNA pot ser retrotranscrit a una altra forma del genoma viral, DNA lineal de doble cadena (dslDNA), que sol integrar-se en el genoma de l'hoste sense produir pgRNA però sí altres RNAs virals [2].

A més, la polimerasa viral no disposa d'activitat *proof reading* i presenta una taxa de substitucions per posició i any de l'ordre de 10^{-5} en la retrotranscripció del pgRNA, comparable a la d'alguns virus de RNA com els retrovirus [5]. Això implica que en un mateix hoste infectat es forma un conjunt de poblacions víriques estretament relacionades però no idèntiques, constituïdes per variants genètiques que defineixen una quasiespècies viral. Aquestes variants, per tant, estan sotmeses a un procés continu de variació genètica, competició i selecció [4]. Les diferents poblacions de variants genètiques es poden diferenciar en grups filogenètics, anomenats genotips; s'han descrit fins a 9 genotips diferents ben caracteritzats (A-I) més un desè putatiu (J), els quals divergeixen en més d'un 7,5% en la seva seqüència genòmica. A més, els genotips es poden subdividir en més de 35 subgenotips, que presenten una divergència d'entre un 4-7,5% en el seu genoma [2, 6, 7].

D'altra banda, es poden produir fenòmens d'intercanvi de DNA o recombinació entre seqüències víriques de genotips diferents que infecten un mateix hoste. Aquest fet dificulta la correcta classificació genotípica del VHB, ja que en molts casos les soques víriques recombinants s'introdueixen incorrectament com a nous genotips o subgenotips. Els punts de recombinació entre genotips poden ser molt variables, però en alguns estudis s'ha vist que més del 60% de les recombinacions detectades es donen entre les posicions 1640-1900 del genoma del VHB, i d'altres s'han identificat entre les posicions 3150-830 (gen *S*) [7]. De fet, la regió genòmica 1600-2000 nt mostra una densitat de punts de recombinació 5 vegades major que la resta del genoma [8].

La distribució geogràfica i ètnica es mostra diferent entre els genotips del VHB, i a més aquests es correlacionen amb la gravetat de les malalties hepàtiques [6]. De fet, les principals guies clíniques per al tractament de l'hepatitis B crònica recomanen determinar el genotip del VHB només per tal de seleccionar els pacients que rebran tractament antiviral amb interferó pegilat, atès que els genotips A i B s'associen a taxes més altes de pèrdua d'HBeAg o HBsAg que els

genotips C i D [9, 10, 11]. Tot i això, a part del tractament amb interferó, altres estudis han demostrat l'existència de diferències entre els genotips pel que fa a l'estat de seroconversió del HBeAg (antigen soluble en sèrum i índex de la replicació viral, infectivitat, inflamació, resposta a tractament antiviral, etc.), el pronòstic i progrés de la malaltia (cronificació), la resposta a la vacunació [7,12], etc. Tenint en compte les implicacions clíniques i epidemiològiques dels genotips del VHB, cal remarcar la importància de la determinació del genotip en pacients infectats i de les possibles barreges i recombinacions entre genotips, ja que poden esdevenir molt útils en la predicció del risc de morbiditat i l'orientació en el tractament de la infecció per VHB.

2.2 Objectius del treball

Objectius generals

1. Anàlisi i simplificació d'un *pipeline* de genotipat de VHB ja existent mitjançant la creació d'un paquet de R.

Objectius específics

1. Revisar, documentar i definir l'estructura completa del *pipeline* disponible.
2. Implementar diverses funcions en un paquet que simplifiqui el processament de les dades de seqüenciació.
3. Descriure els resultats obtinguts i seleccionar els més rellevants.

2.3 Enfoc i mètode

Actualment, existeixen diverses aproximacions que permeten realitzar una classificació genotípica del VHB més o menys precisa. En el cas del present treball, es disposa d'un *pipeline* que permet realitzar l'anàlisi de qualitat dels *reads* obtinguts per seqüenciació massiva (NGS) mitjançant la plataforma MiSeq (Illumina Inc., San Diego, CA, EUA) i obtenir la classificació genotípica del VHB a partir de la seqüenciació de dos amplicons diferents del genoma. No obstant, aquest *pipeline* presenta una complexitat elevada, en especial a l'hora d'aplicar-se per part d'usuaris no experts en programació; per aquest motiu es fa notòria la necessitat de simplificar els *scripts* disponibles mitjançant la creació d'un paquet de R.

Cal remarcar que l'estratègia escollida no es basa en crear un *pipeline* de nou si no en revisar i simplificar un d'existent, ja que el temps requerit en el primer cas supera l'abast del treball. A més, donada l'extensió del codi proporcionat i tenint en compte que part del procés de classificació genotípica ja es troba adaptada en un paquet de Bioconductor desenvolupat pel mateix grup de recerca [13], la revisió i simplificació proposada no cobreix la totalitat del *pipeline* original, si no que es basa en la implementació de funcions per tal de realitzar únicament l'anàlisi de qualitat de les dades fins a l'obtenció d'haplotips consens. Aquests haplotips corresponen a les seqüències úniques de les quasiespècies que cobreixen l'amplicó sencer i són comunes per ambdues cadenes *forward* i *reverse*, a les quals se'ls assigna una freqüència calculada a partir la suma dels *reads* d'ambdues cadenes. Així doncs, cada haplotip correspon a una variant de quasiespècies i tots ells suposen la base de diferents procediments que es poden realitzar *a posteriori*, entre les quals s'inclou el genotipat de les seqüències víriques, però també d'altres com l'anàlisi de la diversitat de les quasiespècies. Per aquest motiu, l'optimització del *pipeline* proporcionat no inclourà la classificació genotípica del VHB.

2.4 Planificació del treball

2.4.1 Tasques

La planificació mostrada a continuació recull les diferents entregues associades al treball de final de màster, les quals es realitzaran paral·lelament a les tasques exposades a continuació. Entre parèntesi es mostren les hores assignades a cadascuna de les tasques:

- Recerca bibliogràfica (10 h)
- Plantejament dels objectius (10 h)
- Revisió dels *scripts*
    - Descripció dels components del *pipeline* (60 h)
    - Anàlisi de punts forts i febles (10 h)
    - Generació d'un *flow chart* (20 h)
- Creació del paquet de R (135 h totals)
    - Disseny i implementació de noves funcions
    - Documentació de les funcions creades
    - Empaquetat de les funcions
- Descripció dels resultats obtinguts
    - Aplicació del *pipeline* original i modificat amb dades reals de NGS (10 h)
    - Comparació dels resultats d'ambdues aproximacions (12 h)
- Redacció de la memòria (50 h)
- Elaboració de la presentació (20 h)

2.4.2 Calendari

Les tasques indicades a la secció anterior han estat organitzades segons la temporització indicada a la figura 1. A més, s'inclouen les diverses entregues associades al present Treball de Fi de Màster, que han permès l'assoliment de les fites presentades més endavant.

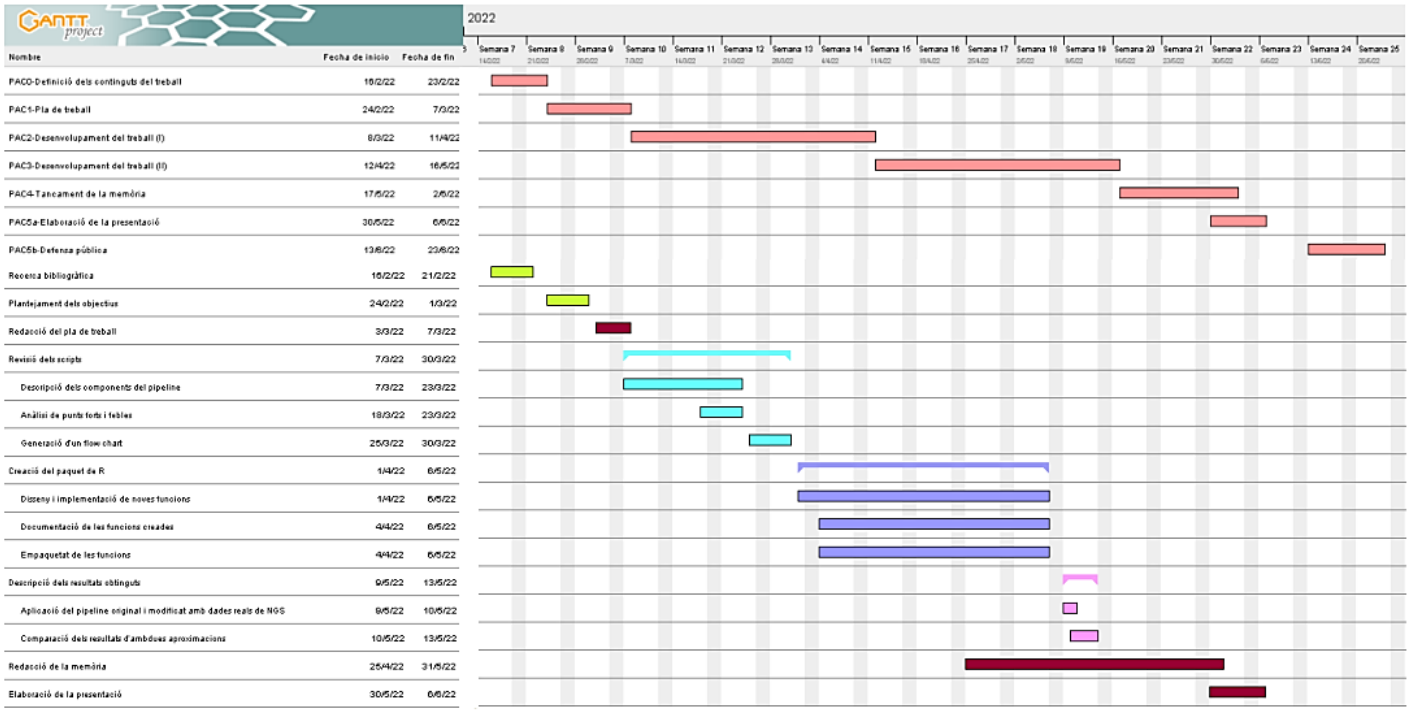


Figura 1. Diagrama de Gantt amb la planificació plantejada en el present treball.

Fites

Les fites més importants que han determinat la consecució de les tasques plantejades son les següents, en ordre de prioritat:

1. Obtenció del *flow chart* i revisió completa dels *scripts*.
2. Obtenció del paquet en R amb les funcions implementades.

2.5 Breu resum de les contribucions

El present projecte ha permès l'obtenció d'un pla de treball òptim per l'assoliment dels objectius plantejats i una memòria final que inclou una descripció detallada del procediment i resultats obtinguts. La realització del treball finalitza amb la realització d'una presentació en format virtual on s'exposaran els aspectes més rellevants.

Addicionalment, com a producte final s'ha generat un paquet de R amb funcions implementades que simplifiquen part del *pipeline* disponible per al genotipat del VHB, així com un *pipeline* simplificat en format d'*script* de R que permet aplicar les funcions definides sobre dades de NGS.

2.6 Breu descripció dels altres capítols de la memòria

- **Capítol 3. Estat de l'art**: revisió dels diferents mètodes disponibles actualment per tal de dur a terme la classificació genotípica del VHB, així com les possibles eines per tal de realitzar un anàlisi de qualitat complet a partir de dades NGS.
- **Capítol 4. Metodologia**: inclou una descripció detallada dels passos seguits per a la consecució del projecte i l'assoliment dels objectius inicials.
- **Capítol 5. Resultats**: presentació dels resultats més rellevants, associats a l'anàlisi de qualitat i el genotipat d'un conjunt de dades NGS del VHB.
- **Capítol 6. Discussió**: contrast entre els resultats obtinguts a partir del *pipeline* original i l'aplicació del paquet de R generat.

- **Capítol 7. Conclusions**: avaluació del compliment dels objectius inicials i establiment de futures línies de treball.
- **Capítol 8. Glossari**: descripció dels acrònims emprats en la redacció del treball.
- **Capítol 9. Bibliografia**: referències bibliogràfiques consultades.

3 Estat de l'art

Al llarg dels anys s'han desenvolupat diversos mètodes per tal de dur a terme l'anàlisi del genotip del VHB [3, 7]. En general, la majoria de tècniques moleculars existents es basen en l'estudi del gen *S* o la regió preS, ja que sol estar més conservada en un mateix genotip que altres parts del genoma víric [7,14,15,16]. Tot i que moltes d'aquestes aproximacions no es solen fer servir en l'actualitat, es presenta a continuació un resum de les seves característiques:

- **Hibridació inversa (INNO-LiPA)**: es tracta d'un mètode en el qual el DNA d'estudi (normalment la regió preS/gen *S*) s'amplifica per PCR emprant encebadors biotinilats [3]. A continuació, els productes de PCR són confrontats amb sondes específiques immobilitzades en diverses tires de nitrocel·lulosa, de manera que hibridaran a la regió on s'hagin immobilitzat les sondes específiques del seu genotip [3, 17]. Aquest mètode permet detectar infeccions per part d'un o múltiples genotips, és relativament econòmic i específic, però resulta insensible en l'anàlisi de molècules gèniques amb polimorfismes d'un sol nucleòtid (SNPs) o delecions que poden hibridar inespecíficament amb les sondes, donant lloc a falsos positius o negatius pels diferents genotips en revelar aquestes tires mitjançant una reacció colorimètrica [3].
- ***Restriction fragment polymorphism*** (RFLP): l'estratègia emprada en aquest cas consisteix en l'amplificació per PCR del fragment desitjat, normalment el gen *S* del VHB, seguida d'una digestió del DNA amb enzims de restricció, que actuen en dianes de restricció específiques de cada genotip, i la separació dels fragments resultants per electroforesi [3]. Així doncs, l'elecció dels enzims de restricció ve determinada segons les seqüències dels diversos genotips del VHB [3]. No obstant, aquesta tècnica pot donar lloc a indeterminacions, i cal tenir en compte que qualsevol mutació en les dianes de restricció dels enzims poden afectar a la sensibilitat de la prova [3, 7].
- ***Multiplex* PCR**: en aquest cas es realitzen dues rondes de PCR per tal d'analitzar el gen *S* o la regió preS1 del genoma. La primera PCR es realitza utilitzant *primers* universals dissenyats d'acord a regions nucleotídiques conservades, mentre que a la segona PCR s'empren *primers* específics dels diferents genotips, en base a regions conservades dins d'un mateix genotip però sense homologia respecte la resta [3, 7]. La sensibilitat del mètode és major respecte el RFLP i es poden detectar subgenotips i infeccions mixtes, però també presenta el problema derivat de la presència de SNPs en les regions d'amplificació dels *primers* [3, 7].
- ***Oligonucleotide microarray chips***: la classificació genotípica del virus també es pot realitzar mitjançant *microarrays*, de manera que els amplicons de la regió preS marcats amb Cy5 s'hibriden sobre un xip de DNA on hi ha immobilitzades diverses sondes específiques de genotip [3]. La sensibilitat d'aquests *microarrays* resulta ser bastant elevada i a més permeten una millor detecció d'infeccions mixtes, però el cost resulta elevat i la sensibilitat pot veure's afectada en presència de SNPs o delecions en les seqüències [3, 7].

- ***Flow-through reverse dot blot*** (FT-RDB): aquesta tècnica consisteix en amplificar per PCR una regió conservada del gen *S* emprant un *primer reverse* marcat amb biotina en l'extrem 5', de manera que els amplicons s'afegeixen a una membrana de niló amb sondes específiques de genotip. A continuació, es realitza la detecció afegint un conjugat d'estreptavidina-peroxidasa amb un cromogen [3]. Tot i ser un mètode econòmic, ràpid, sensible i que permet detectar infeccions mixtes, presenta una efectivitat limitada en presència de mutacions gèniques [3, 7].
- ***Restriction fragment mass polymorphism*** (RFMP): de la mateixa manera que el RFLP, aquesta tècnica requereix de la digestió del fragment de DNA desitjat per enzims de restricció, per la qual cosa presenta la mateixa limitació referent a les mutacions localitzades en les dianes de restricció [3]. A més, en aquest cas els fragments obtinguts s'analitzen emprant un equip d'espectrometria de masses MALDI-TOF, de forma que es poden determinar variants resistents a alguns fàrmacs, però presenta el desavantatge d'emprar un sistema costós i que requereix de personal especialitzat [3, 7].
- **PCR *Invader assay***: aquest mètode es basa en l'addició del que es coneix com encebadors *Invader* sobre els amplicons de la regió desitjada (gen *S* o regió *core* del VHB), en conjunt amb unes sondes específiques de genotip. En funció del genotip de la mostra, es dona una ruptura de la sonda específica, que mitjançant un sistema FRET permetrà la generació d'una senyal fluorescent [3, 18]. De la mateixa manera que altres tècniques presentades, la seva sensibilitat és molt elevada, però pot veure's afectada en funció de les mutacions presents a les seqüències analitzades [3].
- ***Real-time* PCR**: l'amplificació en temps real de regions concretes del genoma del VHB permeten alhora la quantificació del nivell de DNA en sèrum i la classificació genotípica [3]. Això és possible mitjançant l'anàlisi de la temperatura de *melting* (T_m), que varia entre els diferents genotips en funció del contingut en GC de la seqüència. Aquesta aproximació és molt sensible, d'alt rendiment i evita la contaminació creuada, però presenta dificultats a l'hora de diferenciar entre genotips amb valors de T_m molt propers [3].

D'altra banda, malgrat que els mètodes moleculars poden presentar certs avantatges, l'aproximació més concloent i fiable és l'**anàlisi filogenètic de la seqüència del genoma complert** del VHB [3, 7]. Tot i això, aquesta no sol realitzar-se en el diagnòstic rutinari donat que requereix una inversió important de temps i diners, i a més és difícil d'abordar emprant tècniques de seqüenciació de segona generació [3]. Per aquest motiu, es sol realitzar l'anàlisi filogenètic a partir de gens o regions individuals del genoma, encara que en aquest cas els resultats no són útils per a determinar el subgenotip [3, 7]. A més, les aproximacions en les que es seqüencia únicament un fragment només permeten determinar el genotip d'aquest (no del genoma sencer), i tampoc identifiquen les possibles recombinacions entre genotips [7, 14]. No existeix un consens pel que fa a quina regió del genoma del VHB seria més adient per aquesta tasca, però els mètodes de genotipat per seqüenciació han mostrat una major precisió respecte els mètodes moleculars, per exemple [7, 12].

Una de les estratègies plausibles per tal de dur a terme el genotipat del VHB consisteix en seqüenciar diverses regions del genoma de forma paral·lela, tal i com es planteja en el grup de recerca en Malalties Hepàtiques de l'Institut de Recerca Vall d'Hebron (VHIR) [12]. Actualment els amplicons escollits es seqüencien mitjançant la tecnologia MiSeq d'Illumina, i corresponen a les

regions situades entre les posicions aproximades 2844-3276[1] (amplicó **preS1**), que poden variar en funció del genotip, i les posicions 1255-1611 (amplicó **5'X** o 5pX). Aquesta aproximació permet detectar de forma més precisa les possibles recombinacions existents entre els genotips del VHB, ja que ambdues regions es troben a banda i banda dels punts on es donen la majoria de recombinacions (veure figura 2).

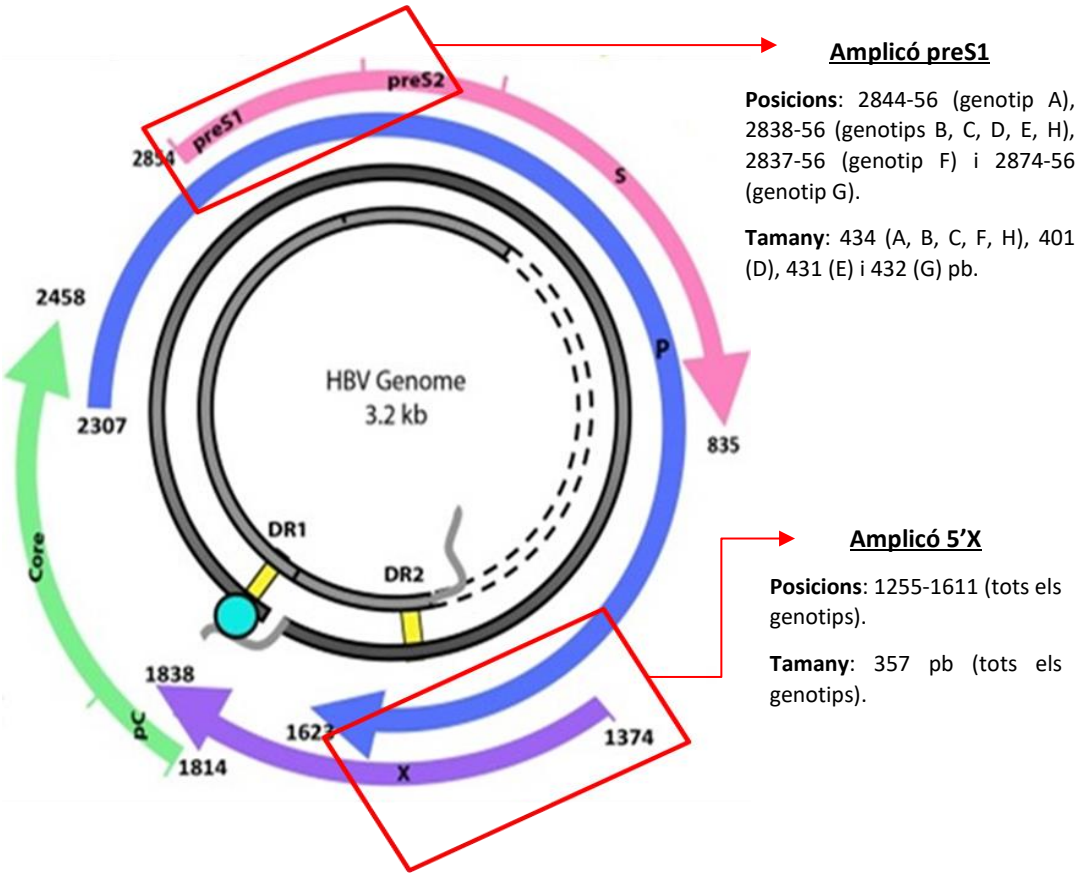


Figura 2. Representació del genoma del VHB amb els corresponents ORFs i posicions nucleotídiques. S'indiquen en vermell les regions seleccionades en el grup de recerca en Malalties Hepàtiques del VHIR per tal de realitzar el genotipat del virus, amplicons preS1 i 5'X (les posicions i tamanys no inclouen els *primers* específics emprats). Imatge adaptada a partir de la publicació de Gonzàlez, C. *et al*. (2018) [19], i cedida pel mateix grup de recerca.

Diversos autors han proposat eines bioinformàtiques útils per tal de determinar el genotip del VHB un cop realitzada la seqüenciació del genoma sencer o d'una fracció d'aquest. Per exemple, l'algoritme BLAST incorporat al *National Center for Biotechnology Information* (NCBI) permet la identificació de les seqüències més properes a la mostra d'estudi mitjançant la comparació d'aquesta amb el conjunt disponible en el repositori GenBank [7]. A més, l'NCBI també ha desenvolupat una eina específica que facilita el genotipat de diversos virus, anomenada NCBI *genotyping tool* [7, 15]. Altres mètodes disponibles en línia inclouen la base de dades HBVdb i els algoritmes HBV STAR, *BioAfrica-Oxford Automated Subtyping Tool* i HepSEQ *Genotyper* [7, 15], alguns dels quals permeten tant el genotipat del virus com la detecció de recombinants i mutacions amb característiques clíniques rellevants [7, 15].

[1] Aquestes coordenades s'han establert per part del grup de recerca per tal de facilitar les anàlisis bioinformàtiques, però realment segons la nomenclatura establerta aquestes posicions serien 2844-56.

És important remarcar, però, que totes les eines de genotipat mencionades requereixen com a *input* les seqüències a analitzar en format FASTA. En canvi, les lectures de seqüència generades per la tecnologia MiSeq es retornen en arxius FASTQ comprimits (fastq.gz), on cadascun dels *reads* presenta els valors de qualitat associats per base [20] i informació rellevant sobre l'experiment de seqüenciació [21]. Per tant, és necessari realitzar un processament i filtrat de les dades previ a qualsevol altre anàlisi, com seria la classificació genotípica del VHB. A més, l'elevada variabilitat genètica del VHB i la possible recombinació intergenotípica impliquen que les anàlisis posteriors a la seqüenciació de regions genòmiques del VHB han de considerar aquests fenòmens, per tal d'evitar classificacions errònies. A més, les eines bioinformàtiques emprades per a aquesta tasca han de permetre que el personal científic i assistencial pugui comprendre fàcilment els resultats obtinguts, per tal de prendre decisions adequades en funció d'aquests.

El procés de *Quality Assessment* (QA) de les dades crues es pot realitzar de múltiples maneres, en funció del mètode experimental emprat i de la tecnologia de seqüenciació utilitzada [21]. Una de les eines més utilitzades és FastQC, que presenta de forma senzilla diverses característiques relacionades amb la qualitat per posició de seqüència, el contingut en bases, la longitud dels *reads*, etc [21, 22]. També es pot obtenir una idea sobre la qualitat de les dades emprant eines com NGS-QC *toolkit*, SolexaQA, PIQA i FastX, entre d'altres [21, 23]. A banda d'això, per tal de procedir al filtrat de les dades crues i l'eliminació dels adaptadors emprats en la seqüenciació, existeix un gran conjunt d'eines disponibles com Trimmomatic, AdapterRemoval, Scythe/Sickle, Cutadapt, Fastx-Toolkit o Reaper [24]. Tenint en compte l'elevada quantitat de metodologies disponibles, cada grup d'investigació pot emprar diverses combinacions d'aquestes.

En el cas del grup de recerca en Malalties Hepàtiques del VHIR, tant el processat de les dades crues de seqüenciació com el genotipat del VHB es realitzen a partir de *pipelines* escrits en llenguatge R, de manera que es realitzen els passos de forma conjunta sense requerir múltiples eines computacionals externes. A més, es realitzen passos addicionals de control de qualitat per minimitzar la incidència en el resultat final de *reads* provinents d'artefactes de PCR i que continguin errors de seqüenciació. No obstant, la possibilitat de simplificar almenys un dels *pipelines* en un paquet de R permetria facilitar la seva implementació per part de personal no especialitzat en bioinformàtica (veure secció 2.3), de manera que els resultats necessaris es puguin obtenir de manera ràpida i senzilla.

4 Metodologia

4.1 Revisió i descripció del *pipeline* original

El *pipeline* original proporcionat per a la realització del present treball va ser desenvolupat en el grup de recerca en Malalties Hepàtiques de l'Institut de Recerca Vall d'Hebron (VHIR) per part del Dr. Josep Gregori i Font [25]. Està dissenyat per analitzar dades NGS generades amb MiSeq (Illumina) a partir de dues regions diferents del genoma del VHB (amplicons 5'X i preS1), mitjançant l'aplicació de funcions dels paquets *Biostrings* [26] i *APE* [27] entre d'altres. A més, aquest es basa en una aproximació *haplotype-centric*, de manera que els *reads* finalment seleccionats han de cobrir completament l'amplicó desitjat i estar presents en ambdues cadenes

del DNA [25]. Globalment, la secció del *pipeline* revisada (escollida segons la informació detallada a la secció 2.3) permet realitzar els passos següents [25, 28, 29]:

1. **Solapament de *reads* aparellats**: si les dades provenen d'un experiment de seqüenciació de lectures aparellades (*paired-end sequencing*), s'empra l'eina FLASH (*Fast Length Adjustment of SHort reads*), per tal de detectar de forma precisa i ràpida el solapament correcte entre els *reads* d'extrems aparellats i estendre la seva longitud [30]. En aquest cas s'estableix un rang de solapament dels *reads* aparellats d'entre 20-300 parells de bases (pb) i com a màxim un 10% de *mismatches* a la regió solapada.
2. **Control de qualitat dels arxius**: un cop obtinguts els arxius amb els *reads* aparellats, es realitza un control de qualitat (QC) de les dades crues i de les obtingudes després d'aplicar el pas anterior, incloent informació de la qualitat per posició, la longitud dels *reads* i altres paràmetres d'interès.
3. **Filtrat per *Phred Score***: es descarten tots aquells *reads* que disposin d'un 5% o més de les seves bases per sota d'un *Phred Score* (Q-*score*) de 30. Una base amb un valor de Q30 implica que presenta una probabilitat de 10^{-3} d'haver estat assignada erròniament [20]. També es realitza un anàlisi de QC sobre els *reads* filtrats.
4. **Demultiplexat dels *reads***: tenint en compte les seqüències d'oligonucleòtids incloses en els *reads* per tal de dur a terme la creació dels amplicons de les dues regions analitzades, en aquest procés es realitza la cerca d'aquests identificadors per tal d'associar els *reads* a les mostres corresponents, dins d'un rang de posicions esperades en les lectures. En primer lloc, es localitzen uns oligonucleòtids de 10 parells de bases anomenats MIDs (*Multiplex IDentifiers*), situats a l'extrem 5' dels amplicons, els quals permeten distingir entre les mostres provinents de diferents pacients i orígens; en aquest cas només s'accepta un *mismatch* entre les seqüències. En segon lloc, s'identifiquen les seqüències corresponents als oligonucleòtids que hibriden amb una diana específica del genoma viral, els *primers* específics (20-30 pb), que permetran identificar les regions d'estudi del genoma del VHB per cadascuna de les cadenes *forward* i *reverse*, acceptant un màxim de 3 *mismatches*. Finalment, es retallen les seqüències identificades i s'obté un fitxer FASTA per cada combinació de MID (identificador de mostra), *primer* i cadena de DNA. Cal remarcar que aquests fitxers contenen un conjunt d'haplotips resultants de col·lapsar els *reads* en seqüències úniques, indicant les freqüències corresponents.
5. **Alineament d'haplotips**: per cadascun dels fitxers FASTA obtinguts, els haplotips s'alineen respecte la seqüència màster, que correspon a l'haplotip més abundant dins del fitxer (amb major nº de *reads*). En aquest pas es descarten tots aquells haplotips que no cobreixin la totalitat de l'amplicó o bé presentin més de dues indeterminacions, tres *gaps* o 250 mutacions respecte l'haplotip màster.
6. **Intersecció d'haplotips**: a partir dels haplotips provinents de la mateixa mostra però cadenes de DNA diferents, es realitza primer un filtrat per tal de descartar tots aquells que presentin una abundància per sota del 0.2%. A continuació, es duu a terme la intersecció dels haplotips d'ambdues cadenes per tal de seleccionar únicament els que estiguin presents en ambdues cadenes, obtenint per cadascuna de les mostres els anomenats **haplotips consens**.

7. **Resultats finals**: a partir dels resultats generats al llarg de l'anàlisi, es realitza un control del rendiment individual per cadascun dels passos i del rendiment global, així com la determinació d'insercions i delecions al llarg de les seqüències dels haplotips consens obtinguts. També es genera un informe on s'inclou un resum de les dades emmagatzemades en els fitxers FASTA finalment obtinguts, per a cadascuna de les regions analitzades i les diferents mostres processades per seqüenciació massiva.

Per tal de comprendre els passos del *pipeline* a nivell bioinformàtic abans de realitzar la seva simplificació i la implementació de funcions en un paquet de R, s'ha procedit a la inspecció completa dels diferents *scripts* mitjançant la creació d'un projecte amb control de versions Git, allotjat en un repositori GitHub (veure materials suplementaris). Aquest inclou tots els fitxers del *pipeline*, però només aquells escollits per a la consecució del projecte disposen del codi completament documentat. En total s'han revisat 15 *scripts*, dels quals 3 es troben al directori global de treball i 12 estan inclosos en una subcarpeta específica (veure secció 4.1.1). Donat que els resultats obtinguts en aquesta fase del treball esdevenen essencials per tal de comprendre les etapes posteriors, es presenta en aquest apartat la informació extreta arran de la revisió completa dels *scripts* indicats.

4.1.1 Estructura del directori

En essència, el *pipeline* està format per dues seccions diferenciades, que corresponen a l'anàlisi de qualitat de les dades (QA) i al procés de genotipat del VHB. Aquestes parts es descriuen en els fitxers de codi R anomenats *Miseq_QA_Pipeline-v2.5.R* i *HBV_Genotype_MySeq_Pipeline-v1.3.R* respectivament, en els quals s'implementen els diferents passos del processament bioinformàtic a partir de l'execució d'altres subfitxers de codi. D'altra banda, el document *HBV_nt_gaps_pars.R* permet definir el conjunt de directoris i variables de paràmetres requerides per a l'execució posterior del genotipat del VHB. La figura 3 mostra l'estructura del directori de treball necessari per a l'aplicació del *pipeline*, el qual inclou un total d'11 carpetes internes descrites a continuació:

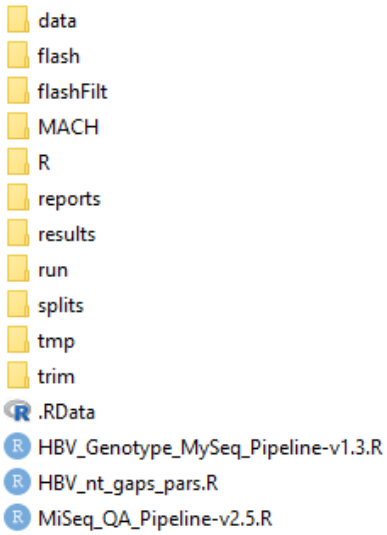


Figura 3. Estructura global del directori de treball per a l'execució del *pipeline* original d'anàlisi de qualitat i genotipat del VHB.

- **/run**: conté els arxius comprimits .fastq.gz de les mostres que s'analitzaran, resultants de la seqüenciació.
- **/data**: conté els arxius següents necessaris per l'anàlisi de qualitat i demultiplexat:
  - **primers.csv**: informació de seqüència i posició de mapeig al genoma del VHB dels *primers forward* i *reverse* per a les regions 5'X i preS1.
  - **samples.csv**: arxiu de mostres que inclou l'identificador de cada pacient, el dels adaptadors MID emprats i les posicions al genoma del VHB de la regió amplificada corresponent.
  - **mids.csv**: inclou la seqüència de tots els adaptadors MID i els identificadors que s'associaran a cada pacient de l'arxiu samples.csv.

Addicionalment, en aquesta carpeta s'inclouen dos arxius que amb les seqüències de referència d'ambdues regions del VHB, així com un altre que disposa del codi genètic, necessaris per realitzar el procés de genotipat posteriorment.

- **/R**: conté els *scripts* de R que s'executen al llarg del *pipeline* (es criden des dels fitxers de codi del directori global).
- **/flash**: emmagatzema els fitxers FASTQ que es generen en executar el programa FLASH [30] per estendre els *reads* amb extrems aparellats (arxius R1 i R2 provinents de la carpeta /run per a les dues regions del VHB avaluades).
- **/flashFilt**: inclou els fitxers FASTQ que es generen després d'eliminar els *reads* resultants de FLASH que presenten un 5% o més de les seves bases per sota de Q30.
- **/splits**: conté els fitxers FASTA (.fna) obtinguts en realitzar el demultiplexat per MIDs dels *reads*. Donat que s'analitzen dues regions del genoma viral per a cadascun dels pacients, s'assigna un adaptador MID diferent a les respectives mostres avaluades. Així doncs, en aquesta carpeta es guarda un fitxer per cada MID (indicant a quina regió correspon, segons l'arxiu de mostres) que inclou les corresponents seqüències identificades, així com dos fitxers anomenats MID0 (un per cada regió) on s'inclouen els *reads* que no han pogut assignar-se a cap MID indicat a l'arxiu de mostres.
- **/trim**: emmagatzema els fitxers FASTA generats després de retallar dels *reads* les seqüències entre els adaptadors MID i els *primers* específics. Per fer-ho, s'empra cadascun dels fitxers de la carpeta /splits i es realitza el demultiplexat per *primers* per tal d'identificar les seqüències corresponents a les cadenes *forward* o *reverse*, de manera que es duplica el nombre de fitxers respecte la carpeta anterior.
- **/tmp**: guarda dos fitxers FASTA temporals generats durant l'execució del programa d'alineament múltiple MUSCLE [31]. Per a cada mostra avaluada es crea el fitxer *muscleInFile.fna*, amb les seqüències d'entrada per a l'alineament múltiple, i el fitxer *muscleOutFile.fna*, on es desa el resultat del procés que es copiarà a la carpeta /MACH.
- **/MACH**: conté dos tipus diferents d'arxius FASTA per a cadascuna de les mostres avaluades. El primer tipus, que inclou el terme *MAfwrv,* inclou els resultats de l'alineament múltiple amb MUSCLE (provinents de la carpeta /tmp); el segon tipus de fitxer, identificat amb el terme *MACHpl02,* conté els haplotips consens generats a partir de la intersecció entre les cadenes *forward* i *reverse*.
- **/reports**: inclou tots els arxius amb informes generats al llarg dels diferents passos del procés d'anàlisi de qualitat, com són gràfics, taules i fitxers .RData.
- **/results**: inclou els fitxers obtinguts com a resultat del procés de genotipat, així com l'informe final d'anàlisi de qualitat de les dades de seqüenciació.

4.1.2 Anàlisi de punts forts i febles

A continuació es presenten les taules 1 i 2, que resumeixen els diferents punts forts i febles detectats en els *scripts* revisats, respectivament:

Taula 1. Punts forts identificats durant la revisió del *pipeline* original.

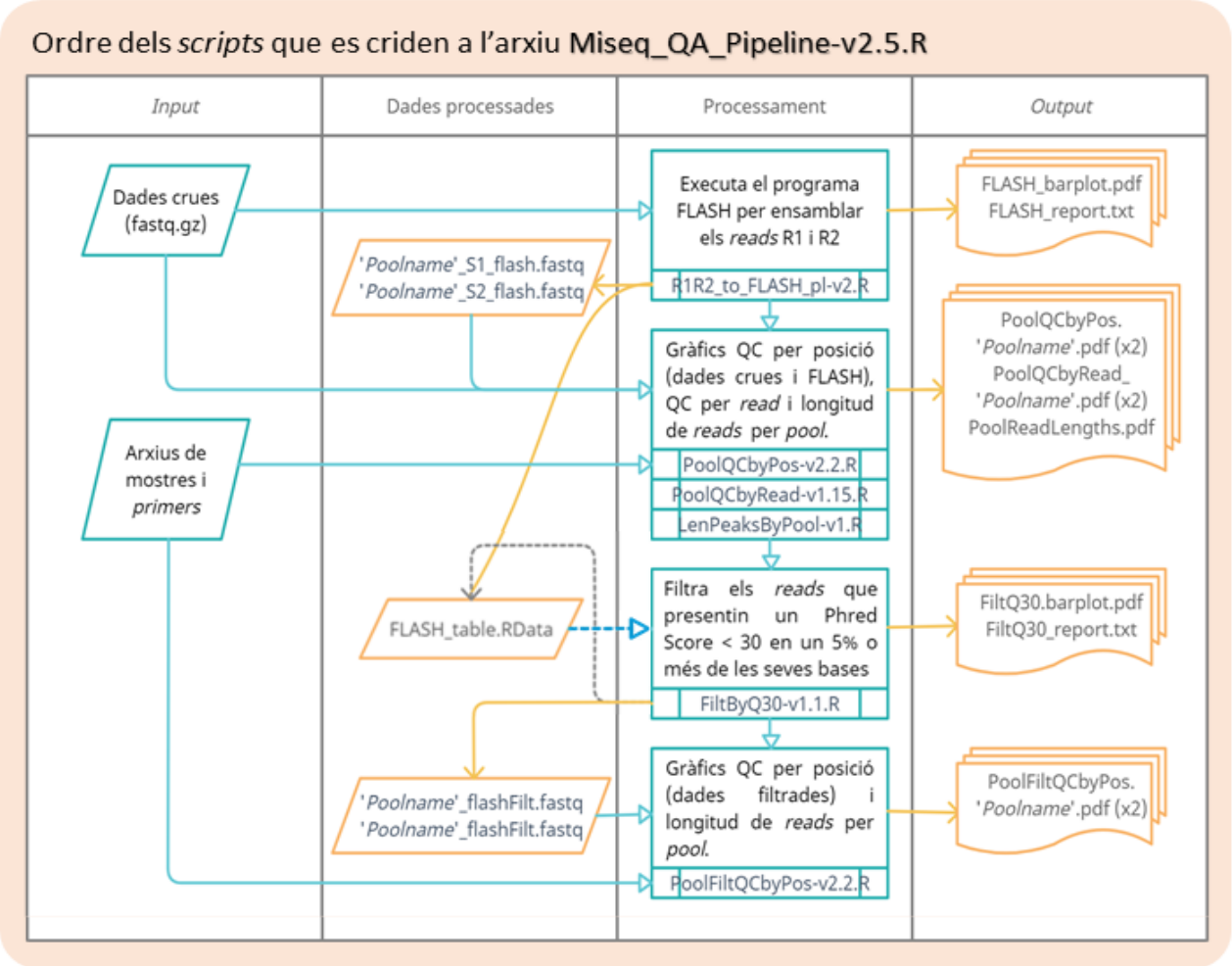
PUNTS FORTS	
✓	L'aproximació *haplotype-centric* suposa una avantatge pel que fa a l'anàlisi dels haplotips consens generats, ja que el fet que sempre cobreixin la longitud sencera de l'amplicó permet evitar la pèrdua d'informació biològica rellevant per a anàlisis posteriors.
✓	A partir de les dades crues de seqüenciació el *pipeline* implementa de forma interna l'eina FLASH [30], que s'executa de manera automàtica dins del *pipeline* requerint únicament la instal·lació de l'executable a l'ordinador de treball.
✓	Per tal de generar els haplotips consens a partir de la intersecció entre les cadenes s'implementa l'eina MUSCLE [31], que permet realitzar l'alineament múltiple de totes les seqüències de la mostra avaluada. De la mateixa manera que amb el programari FLASH, aquesta eina s'executa de manera interna al *pipeline*, requerint només la instal·lació de l'executable.
✓	Part del processament de les dades es realitza emprant funcions incloses en diversos paquets de Bioconductor, que permeten la lectura i manipulació de les seqüències, l'alineament d'aquestes i la cerca de patrons (adaptadors, *primers*...), de manera senzilla i reproduïble [32].
✓	L'estructuració del directori en diverses carpetes permet localitzar de forma ràpida la ubicació dels diferents arxius, ja siguin d'entrada (dades crues i fitxers necessaris per al processament) o de sortida (gràfics, taules de resultats i dades processades).
✓	Tot i que alguns dels passos consumeixen molta memòria a nivell computacional, es pot realitzar un anàlisi complet de la qualitat de les dades en relativament poc temps.

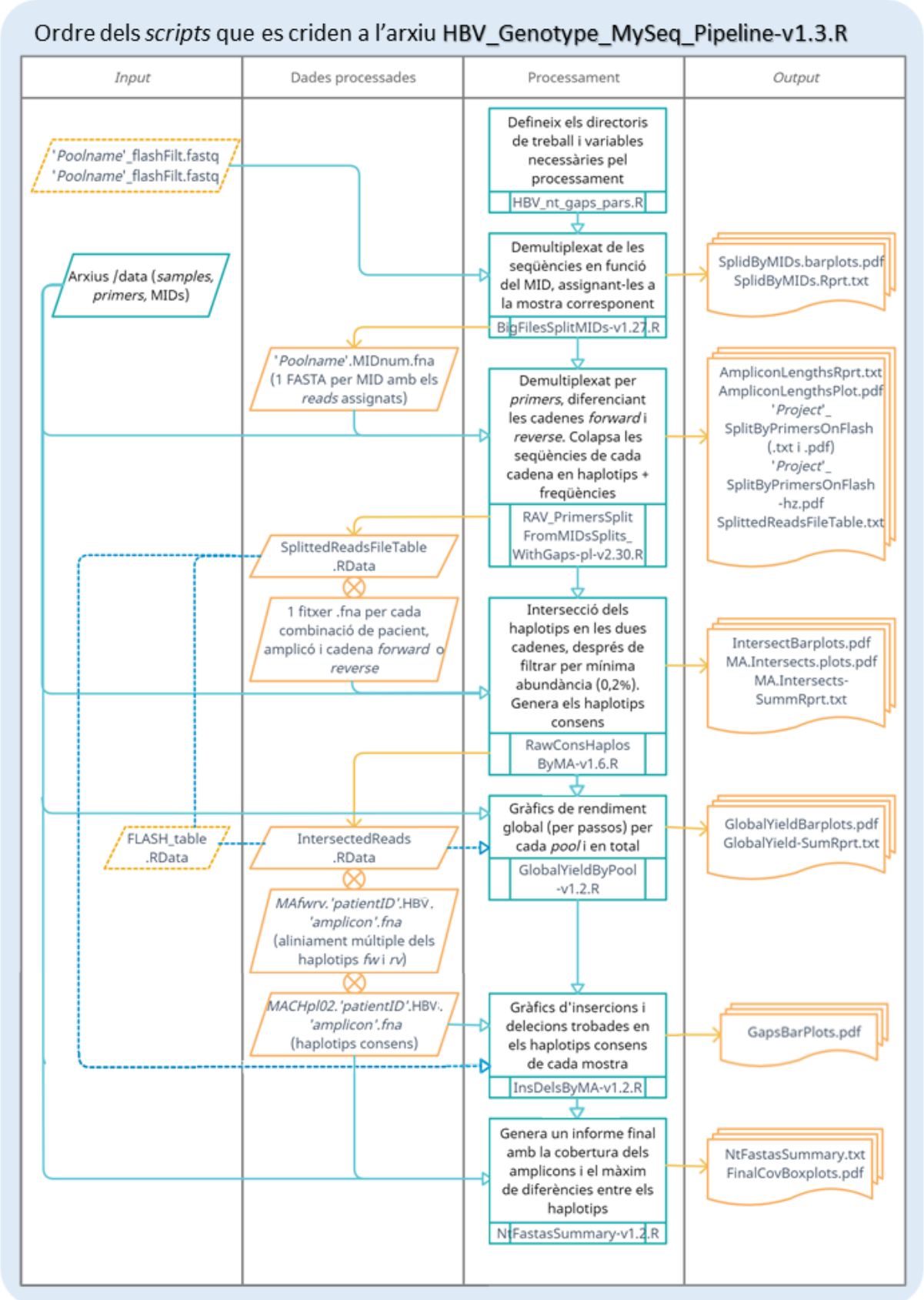
Taula 2. Punts febles identificats durant la revisió del *pipeline* original.

PUNTS FEBLES	
×	Els paquets requerits es carreguen múltiples cops en els diferents *scripts*, augmentant el temps d'execució i generant redundància.
×	Els directoris de treball i variables s'especifiquen reiteradament en els diversos *scripts*. A més, algunes de les variables es redefineixen successivament, complicant la comprensió i seguiment del flux de treball.
×	Algunes de les funcions locals definides reben el mateix nom en diferents *scripts* tot i presentar algunes diferències, generant confusió en la seva interpretació. Addicionalment, es defineixen funcions basades únicament en aplicar d'altres ja incloses en l'entorn de R.
×	Algunes de les variables i funcions locals definides no són requerides en el processament, de manera que augmenta la complexitat i el temps d'execució innecessàriament.
×	Hi ha parts del *pipeline* que inclouen bucles molt extensos, la qual cosa alenteix el procés i pot provocar problemes d'execució.
×	Es genera una gran quantitat de resultats, alguns dels quals contenen informació redundant i poden resultar difícils d'interpretar pel personal sanitari i assistencial.

4.1.3 Diagrama de flux

A partir de l'anàlisi exhaustiu de la secció del *pipeline*, ha estat possible detallar la seva estructura lògica mitjançant la creació d'un *flow chart* (figura 4), tot especificant els diferents fitxers d'entrada i sortida del procés.




Ordre dels *scripts* que es criden a l'arxiu HBV_Genotype_MySeq_Pipeline-v1.3.R
Input
Dades processades
Processament
Output
'Poolname'_flashFilt.fastq
'Poolname'_flashFilt.fastq
Arxius /data (samples, primers, MIDs)
Defineix els directoris de treball i variables necessàries pel processament
HBV_nt_gaps_pars.R
Demultiplexat de les seqüències en funció del MID, assignant-les a la mostra corresponent
BigFilesSplitMIDs-v1.27.R
SplidByMIDs.barplots.pdf
SplidByMIDs.Rprt.txt
'Poolname'.MIDnum.fna (1 FASTA per MID amb els reads assignats)
Demultiplexat per primers, diferenciant les cadenes forward i reverse. Colapsa les seqüències de cada cadena en haplotips + freqüències
RAV_PrimersSplit FromMIDsSplits_ WithGaps-pl-v2.30.R
AmpliconLengthsRprt.txt
AmpliconLengthsPlot.pdf
'Project'_ SplitByPrimersOnFlash (.txt i .pdf)
'Project'_ SplitByPrimersOnFlash -hz.pdf
SplittedReadsFileTable.txt
SplittedReadsFileTable .RData
1 fitxer .fna per cada combinació de pacient, amplicó i cadena forward o reverse
Intersecció dels haplotips en les dues cadenes, després de filtrar per mínima abundància (0,2%). Genera els haplotips consens
RawConsHaplos ByMA-v1.6.R
IntersectBarplots.pdf
MA.Intersects.plots.pdf
MA.Intersects-SummRprt.txt
FLASH_table .RData
IntersectedReads .RData
Gràfics de rendiment global (per passos) per cada pool i en total
GlobalYieldByPool -v1.2.R
GlobalYieldBarplots.pdf
GlobalYield-SumRprt.txt
MAfwrv.'patientID'.HBV.'amplicon'.fna (aliniament múltiple dels haplotips fw i rv)
MACHpl02.'patientID'.HBV.'amplicon'.fna (haplotips consens)
Gràfics d'insercions i delecions trobades en els haplotips consens de cada mostra
InsDelsByMA-v1.2.R
GapsBarPlots.pdf
Genera un informe final amb la cobertura dels amplicons i el màxim de diferències entre els haplotips
NtFastasSummary-v1.2.R
NtFastasSummary.txt
FinalCovBoxplots.pdf

Llegenda dels símbols emprats:

	Procés o pas del *pipeline*. Inclou una breu descripció del procés que s'està executant.
	Procés pre-definit. Indica els arxius de codi R que s'executen en el pas indicat i que es van cridant al *pipeline*.
	Fitxers d'entrada/sortida. Corresponen als arxius d'entrada necessaris per al desenvolupament del *pipeline* (blau) o bé als fitxers de dades obtingudes del pas de processament, que serveixen d'entrada per a passos posteriors (taronja). Les línies discontínues indiquen que l'arxiu provè del diagrama de flux anterior.
	Documents generats en el procés indicat, en concret gràfics (.pdf) o taules de resultats (.txt).
	Símbol AND. Indica que els fitxers s'obtenen del mateix pas de forma conjunta.
	Direcció dels passos del processament i dels corresponents fitxers d'entrada.
	Fletxa que indica la generació de resultats de qualsevol tipus.
	Actualització del fitxer indicat.
	Fletxa que indica la direcció d'entrada exclusivament dels fitxers amb extensió .RData, unint aquells que s'utilitzin de manera conjunta en el mateix procés.

Figura 4. Diagrama de flux que inclou l'estructura ordenada d'execució dels *scripts* revisats en el present treball. Es representen per separat els passos realitzats en cadascuna de les parts definides en el *pipeline* global, així com la llegenda dels símbols mostrats en el diagrama.

4.2 Implementació de funcions

Un cop revisats els fitxers de codi escollits en el present projecte, s'ha procedit a la seva simplificació mitjançant la definició de diverses funcions, per tal de disminuir al màxim els punts febles detectats en el *pipeline*. És important esmentar que la simplificació no s'ha realitzat sobre els 15 fitxers indicats, ja que tal i com s'ha mencionat a la secció 4.1.1, els 3 arxius presents al directori global de treball només permeten executar els diversos *scripts* i definir els paràmetres d'inicialització necessaris. Per tant, l'estratègia general plantejada es basa en l'optimització de cadascun dels 12 *scripts* (allotjats al subdirectori /R) en una sola funció, establint els paràmetres requerits com arguments de les funcions corresponents. D'aquesta manera, l'usuari pot modificar les variables de forma senzilla sense requerir de cap fitxer addicional.

Convé subratllar, però, que aquells passos de l'anàlisi prou similars a nivell de codi s'han pogut realitzar mitjançant la creació d'una sola funció, establint arguments condicionals que permeten realitzar els deguts canvis segons s'indiqui a cada pas. És el cas dels *scripts* que faciliten el control

de qualitat per posició de les dades abans del filtrat per *Phred Score* i també després, els quals es trobaven en arxius diferents i han pogut optimitzar-se en una única funció.

D'altra banda, en el cas dels arxius de codi massa extensos per ser descrits en una única funció de R, ha estat necessària la definició de funcions addicionals, anomenades *helper*. Aquestes permeten executar part de l'*script* original, ja sigui per tal de substituir els extensos bucles presents en aquest o bé per evitar la presència de codi repetitiu en diverses funcions parentals. Cal remarcar, però, que les funcions *helper* no poden ser aplicades per l'usuari de forma individual, ja que han estat dissenyades únicament per facilitar l'execució del procés global.

Addicionalment, per tal de reduir el temps global d'execució de l'anàlisi s'ha implementat en algunes de les funcions el que es coneix com paral·lelització, una forma de computació en la qual diverses instruccions s'executen de forma simultània en diferents ordinadors o processadors [33, 34]. Aquest mètode de computació es pot realitzar de diverses maneres en R, però una de les més senzilles i òptimes és la utilització de la funció `mclapply` (paquet *parallel*) [34]. No obstant, aquesta funció fa servir una aproximació coneguda com *forking*, la qual no funciona en sistemes Windows [34]. Per tant, ha estat necessari adaptar localment una funció disponible públicament anomenada `mclapply.hack`, que funciona de la mateixa manera que la funció del paquet *parallel* però en Windows [35]. Dins del conjunt de funcions definides, aquesta última també es pot considerar com a *helper*, ja que només s'ha definit en el paquet de R per tal d'incloure's en 3 de les funcions generades que consumien una important quantitat de temps.

4.3 Creació del paquet de R

A partir de les 15 funcions finalment definides ha estat possible la generació d'un paquet local en R, anomenat *QApckg*, mitjançant la utilització dels paquets *devtools* i *roxygen2* [36]. D'aquesta manera, s'ha procedit a la documentació de les funcions de forma simultània a la seva creació, ja que gràcies als paquets mencionats, la informació necessària pot editar-se en els mateixos arxius .R on es troba inclòs el codi de les corresponents funcions; aquesta documentació s'emmagatzema automàticament en un nou arxiu d'extensió .Rd (*R documentation file*). Per tal de portar a terme un correcte seguiment de l'evolució del paquet i les diverses funcions, també s'ha creat un projecte amb control de versions Git, allotjat en un repositori GitHub (veure materials suplementaris).

Finalment, un cop definida l'estructura final del projecte, la funció `devtools::build_manual()` ha permès l'obtenció del manual de referència del paquet generat, el qual inclou la informació detallada en els fitxers de documentació de totes les funcions. Com a conseqüència, els usuaris disposen de tots els detalls precisos per comprendre l'anàlisi realitzat en cadascuna de les funcions.

4.4 Comprovació del funcionament del paquet

A fi de verificar l'assoliment dels objectius inicials plantejats en el treball, ha estat necessari constatar l'eficàcia i funcionament del paquet de R creat, per tal d'inferir si els canvis efectuats respecte el *pipeline* proporcionat generen resultats erronis o diferents als originals. Per fer-ho, s'ha seleccionat un conjunt de dades òptim (projecte anomenat VHBass3) derivat d'un experiment de seqüenciació amb MiSeq (Illumina), realitzat per part del grup de recerca en Malalties Hepàtiques del VHIR a partir de les mostres de 12 pacients amb VHB. Per tant, es

disposa de 4 arxius comprimits .fastq.gz, corresponents als arxius R1 i R2 (*reads* d'extrems aparellats) dels dos amplicons analitzats, 5'X i preS1. Tot i això, és important mencionar que aquestes dades no es troben disponibles públicament, donats els interessos d'investigació que presenten per part del grup.

Partint dels arxius mencionats corresponents als 24 amplicons (2 per pacient, tenint en compte les dues regions avaluades), s'ha procedit a l'aplicació del *pipeline* original i del simplificat generat a partir del paquet de R. Aquest últim s'ha implementat mitjançant la creació d'un únic *script* de R, en el qual s'apliquen les diverses funcions del paquet *QApckg* per tal de realitzar l'anàlisi de qualitat de les dades i l'obtenció dels haplotips consens de forma senzilla i automatitzada.

D'altra banda, cal tenir en compte que el present projecte planteja l'optimització d'una part d'un *pipeline* destinat al genotipat del VHB. En conseqüència, a part dels resultats generats mitjançant ambdues aproximacions en relació a l'anàlisi de qualitat de les dades, també ha estat necessari contrastar els resultats del genotipat a partir dels haplotips consens obtinguts, que haurien de ser idèntics en ambdós casos. És a dir, tot i que la simplificació realitzada no inclogui la secció corresponent a la classificació genotípica del VHB, es fa notòria la necessitat de verificar que l'aplicació del paquet de R permet obtenir els mateixos resultats finals que el *pipeline* original. Així doncs, s'ha aplicat la part del *pipeline* original per al genotipat del VHB partint de les dades obtingudes mitjançant ambdues aproximacions.

5 Resultats

5.1 *QApckg*: paquet per a l'anàlisi de qualitat de dades NGS

L'optimització de la secció del *pipeline* escollida ha resultat en la creació d'un paquet de R que inclou un total de 15 funcions, les quals permeten realitzar un anàlisi de qualitat complet i generar un conjunt d'haplotips consens a partir de dades NGS provinents de dos amplicons del VHB. Tot i això, només 9 de les funcions del paquet conformen el *pipeline* simplificat obtingut, ja que la resta corresponen a funcions *helper*. A nivell general, l'estructura interna del paquet es descriu a continuació, indicant l'ordre necessari d'execució de les funcions:

1. `R1R2toFLASH`: aquesta funció resulta de la simplificació de l'*script R1R2_to_FLASH_pl-v2.R* del *pipeline* original (veure figura 4). A partir de les dades crues de seqüenciació i de l'executable del programa FLASH, indicant els paràmetres de solapament necessaris, permet realitzar l'extensió dels *reads* aparellats i generar diversos informes de resultats.
   - `executeFLASH`: funció *helper* per tal d'estendre amb FLASH els *reads* d'extrems aparellats corresponents a un dels *pools* experimentals (regions del VHB avaluades). D'aquesta manera es redueix l'extensió de la funció parental en aplicar aquesta sobre els dos *pools* establerts a l'experiment.
2. `PoolQCbyPos`: correspon a la simplificació dels *scripts PoolQCbyPos-v2.2.R* i *PoolFiltQCbyPos-v2.2.R*. Per tant, la funció avalua si els fitxers d'entrada són els obtinguts abans del filtrat per Q-*score* o després, per tal de realitzar els gràfics de QC per posició corresponents.
   - `QCscores`: funció *helper* que permet calcular els valors de qualitat (Q-*scores*) per cada posició dels *reads* o pel conjunt d'aquests, segons els arguments condicionals indicats.

Així doncs, està inclosa en les funcions parentals `PoolQCbyPos` i `PoolQCbyRead`, facilitant la seva execució sobre les dades corresponents a les dues regions del VHB.

- `QCplot`: funció *helper* dissenyada per tal de representar els valors de qualitat per posició computats. Inclou un argument condicional per tal d'obtenir un gràfic diferent segons si les dades provenen directament de la seqüenciació o de l'extensió amb FLASH. A més, mitjançant un altre argument es pot escollir que el gràfic generat inclogui els perfils de qualitat per *Sliding Window*; en aquest cas es calculen els valors de qualitat mitjana per cada interval de 10 bases, que es desplacen al llarg de tota la seqüència movent-se una sola posició entre una finestra de 10 bases i la següent.
- `mclapply.hack`: funció interna que permet aplicar el procés de paral·lelització en Windows per tal d'agilitzar l'execució de les funcions parentals on s'aplica (veure secció 4.2). En aquest cas, facilita la lectura dels arxius d'entrada i l'aplicació de la funció `QCscores` sobre aquests. Està inclosa en les funcions parentals `PoolQCbyPos`, `PoolQCbyRead` i `FiltbyQ30`.

3. `PoolQCbyRead`: correspon a la simplificació dels *scripts PoolQCbyRead-v1.15.R* i *LenPeaksByPool-v1.R*. Avalua les dades provinents de l'extensió amb FLASH per a cadascun dels *pools* avaluats i facilita l'obtenció de diversos gràfics de longitud i qualitat dels *reads*. Aquests representen la fracció de bases amb una qualitat inferior a Q30 de les lectures, de manera que els resultats són útils a l'hora d'aplicar el filtre de qualitat en la funció posterior.
4. `FiltbyQ30`: resulta de la simplificació de l'*script FiltByQ30-v1.1.R* i està dissenyada per tal de filtrar les lectures provinents de l'extensió per FLASH en funció de la qualitat de les seves bases. Aquest filtre sempre es realitza per Q30 donat que resulta adequat per obtenir dades òptimes, però l'usuari pot decidir mitjançant un argument específic el percentatge màxim de bases permeses per sota d'aquest valor de qualitat. Després d'aplicar aquesta funció, és recomanable tornar a executar la funció `PoolQCbyPos` introduint únicament els arxius resultants d'aquest pas.
5. `demultiplexMID`: funció resultant de simplificar el fitxer de codi anomenat *BigFilesSplitMIDs-v1.27.R*, que permet identificar les seqüències dels MIDs (oligonucleòtids de 10 bases situats a l'extrem 5' dels amplicons) per tal d'assignar els *reads* a les mostres a les que corresponguin. En aquest cas la seqüència d'aquests oligonucleòtids no es retalla dels *reads*, únicament es generen diversos arxius FASTA on s'emmagatzemen els *reads* associats al MID i regió del VHB corresponent.
6. `demultiplexPrimer`: funció dissenyada per a l'optimització de l'*script RAV_PrimersSplitFromMIDsSplits_WithGaps-pl-v2.30.R*. Per a cadascun dels arxius obtinguts amb la funció anterior detecta les seqüències dels *primers* específics segons la regió del VHB indicada, de manera que diferencia aquells *reads* corresponents a les cadenes *forward* i *reverse* (els quals es guarden en arxius FASTA diferents). La funció també retalla totes les seqüències d'adaptadors (incloent-hi els MIDs) deixant únicament la seqüència que s'analitzarà posteriorment, i genera diversos gràfics de resultats.
7. `primermatch`: funció *helper* definida per tal de substituir l'extens bucle `for` de la funció parental, el qual s'aplicava sobre el conjunt de mostres assignades a cadascun dels *pools* avaluats. En aquesta funció és on realment es detecten i es retallen les seqüències dels *primers*, per tant no pot executar-se individualment. Cal remarcar que es retorna la seqüència reversa complementària dels *reads* assignats a la cadena *reverse*, i que després

de la detecció de les seqüències tots els *reads* es col·lapsen en haplotips únics amb les respectives freqüències. D'aquesta manera es facilita el procés d'intersecció del pas posterior.

8. `ConsHaplotypes`: funció corresponent a la simplificació del fitxer *RawConsHaplosByMA-v1.6.R*, que permet realitzar la intersecció dels haplotips *forward* i *reverse* per a cadascuna de les regions avaluades (2 per pacient) i obtenir diversos gràfics i informes de resultats. A partir d'aquesta funció s'obtenen els haplotips consens necessaris per anàlisis posteriors.
   - `SaveHaplotypes`: funció *helper* que facilita la generació i emmagatzament dels haplotips consens, mitjançant l'ordenació d'aquests segons els nº de mutacions que presenten respecte l'haplotip màster (amb major freqüència).
9. `GblYield`: correspon a la simplificació de l'*script GlobalYieldByPool-v1.2.R* i parteix de les dades obtingudes en els passos/funcions 4, 6 i 7. Aquesta funció permet recopilar totes les dades del procés per tal de generar gràfics de rendiment global i per passos per cadascuna de les regions del VHB avaluades.
10. `PlotInDels`: funció per a l'optimització de l'*script InsDelsByMA-v1.2.R*, la qual permet inspeccionar les seqüències dels haplotips consens obtinguts per tal d'informar sobre les insercions i delecions detectades en aquests.

La documentació completa que descriu el funcionament i els arguments necessaris de cadascuna de les funcions es pot consultar en el manual de referència del paquet (veure materials suplementaris). En comparació amb la secció corresponent del *pipeline* original (descrita a la secció 4.1), *QApckg* permet obtenir un conjunt d'haplotips consens per a cada amplicó de forma senzilla i automatitzada, i a més garanteix que l'usuari disposi de tota la documentació necessària per tal d'aplicar les funcions desitjades i comprendre els diferents passos realitzats. La taula 3 resumeix la comparativa entre el nou paquet de R i el codi original revisat, en concret pel que fa als aspectes de la simplificació realitzada.

Taula 3. Aspectes diferencials entre la secció revisada del *pipeline* proporcionat i el paquet *QApkcg* generat.

	Secció revisada del *pipeline* original	***QApckg***
Carpetes inicials necessàries	11	2
***Scripts* totals requerits**	15	1
Càrrega dels paquets	Es carreguen múltiples cops en els *scripts*.	Es carreguen automàticament en inicialitzar *QApckg*.
Executables externs	2	1
Definició dels paràmetres necessaris	Variables definides en un *script* de R extern.	S'introdueixen com arguments de les funcions corresponents.

Total de funcions locals/*helper* definides	32	6
Format de les taules emprades en passos posteriors	S'emmagatzemen en arxius .RData que es guarden a la subcarpeta *reports*.	Es guarden com a variables en l'entorn global d'execució.
Inclou *checkpoints* sobre els arxius d'entrada	No	Sí
Inclou paral·lelització	No	Sí

És important aclarir, però, que la reducció del nombre de carpetes requerides inicialment és deguda a que les pròpies funcions de *QApckg* van generant la resta de subdirectoris, per tal d'emmagatzemar els diversos resultats obtinguts al llarg del procés. Per tant, només cal disposar de la carpeta /run amb els arxius FASTQ comprimits i la carpeta /data amb els fitxers de dades necessaris, detallades a la secció 4.1.. D'altra banda, tal i com s'ha indicat a la secció 4.4, les funcions del paquet generat s'executen mitjançant la creació d'un únic *script* de R (veure materials suplementaris), de manera que no cal disposar de fitxers de codi addicionals on realitzar el processament.

Una de les simplificacions aplicades en la creació del paquet ha consistit a substituir algunes de les funcions locals definides en el *pipeline* original per funcions ja disponibles en el paquet *QSutils* [13] o altres paquets de Bioconductor. Un exemple és la funció `muscle` inclosa en el paquet amb el mateix nom [37], que permet realitzar alineaments múltiples sense haver de descarregar l'executable extern. Addicionalment, s'han eliminat els arxius d'extensió .RData generats al llarg de l'execució del *pipeline*, reduint així el nombre de fitxers resultants. Aquesta simplificació consisteix en què les variables que quedarien emmagatzemades en aquests arxius ara són retornades per les funcions del paquet de R, de manera que es poden fer servir d'*input* en passos posteriors sense requerir la càrrega addicional de fitxers.

5.2 Resultats obtinguts en ambdues aproximacions

A partir del conjunt de dades del projecte VHBass3 (dades assistencials) s'han pogut aplicar les funcions definides en el paquet de R generat, així com la secció corresponent del *pipeline* original per tal de comparar els resultats obtinguts. L'aplicació de *QApckg* sobre les dades va requerir un temps aproximat de 3 minuts, mentre que el *pipeline* original va resultar en un temps de 5 minuts.

Tenint en compte que l'aplicació del paquet de R resulta en l'obtenció d'un total de 25 arxius (incloent taules .txt i informes en .pdf), es mostren a continuació únicament alguns dels gràfics més representatius, per tal de comparar-los amb els corresponents resultats del *pipeline* original. La resta de resultats obtinguts es poden consultar des dels repositoris de GitHub creats al llarg del projecte (veure materials suplementaris).

La figura 5 mostra els gràfics de QC per posició obtinguts per a les dades crues de seqüenciació (R1 i R2), les resultants de l'extensió amb FLASH (*Flash reads*) i també de les filtrades per Q30, únicament per a l'amplicó 5'X com a exemple. Es pot observar com la qualitat dels *reads* és considerablement millor després de filtrar per Q-*score*.

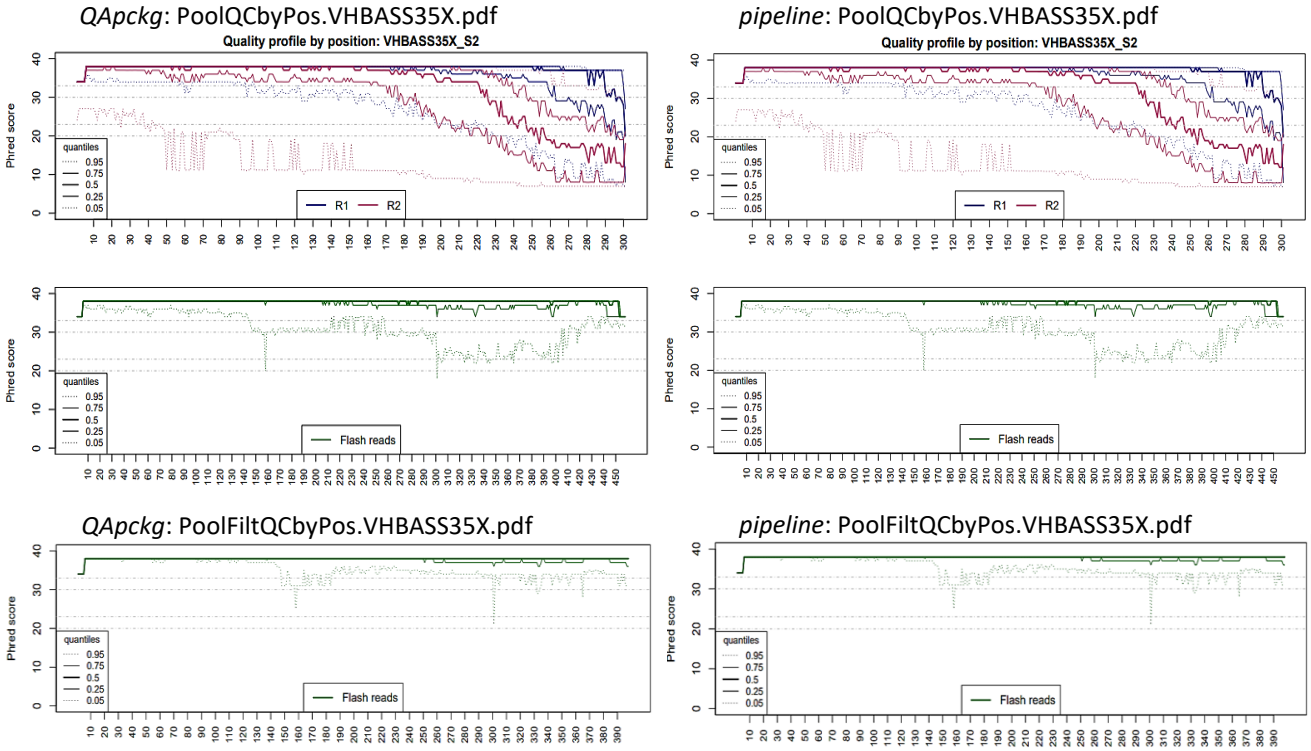
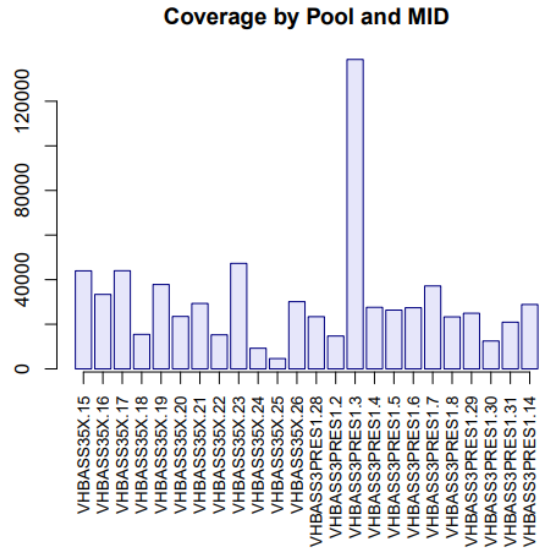


Figura 5. Gràfics de QC per posició obtinguts abans i després del filtrat per Q30, emmagatzemats als arxius indicats sobre cada gràfic. Es mostra únicament un dels *pools* avaluats com a exemple. A l'esquerra es mostren els resultats d'aplicar el paquet de R, i a la dreta els obtinguts a partir del *pipeline* original.

A la figura 6 es mostren, igual que a la figura 5, els resultats del paquet de R en contraposició amb els del *pipeline* original. En aquest cas es representen els *reads* assignats a cadascun dels identificadors MID indicats a l'arxiu original de mostres, després de realitzar el demultiplexat per MIDs. Ressalta la mostra identificada amb el MID 3, la qual presenta un nº de *reads* considerable en comparació a la resta.

D'altra banda, la figura 7 inclou els resultats obtinguts en ambdues aproximacions després del demultiplexat per *primers*. Cal mencionar que en aquest cas el nom de l'arxiu obtingut s'ha modificat en el paquet de R, per tal de reduir la seva extensió. Per a cada combinació de pacient i regió avaluada, s'identifica el nº de *reads* assignats a cadascuna de les cadenes de DNA. Aquest gràfic es correlaciona amb el de la figura 6, ja que l'ordre de disposició de les mostres és la mateixa; en general, el nº de *reads* assignats a les cadenes és proporcional als identificats per cadascun dels MIDS.

QApckg: SplidByMIDs.barplots.pdf



pipeline: SplidByMIDs.barplots.pdf

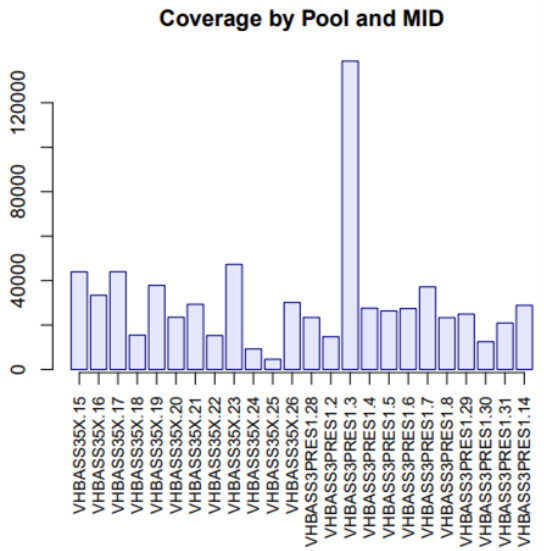
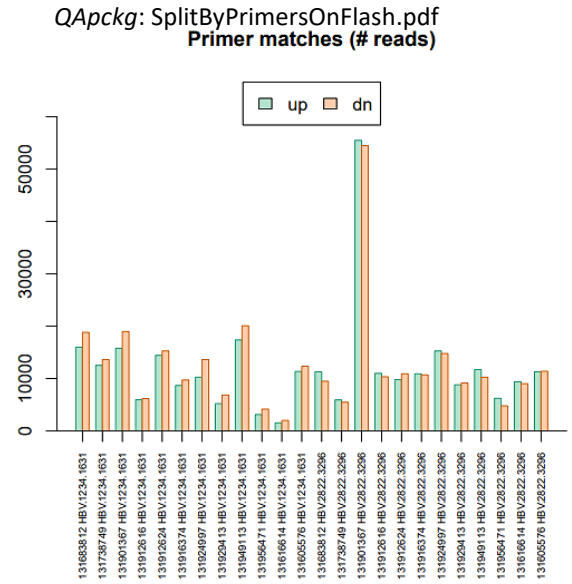


Figura 6. Gràfics de barres representant el nombre de *reads* assignats a cadascun dels adaptadors MID (eix X), indicant la regió a la qual correspon cadascun (eix Y). A l'esquerra es mostren els resultats d'aplicar el paquet de R, i a la dreta els obtinguts a partir del *pipeline* original.

QApckg: SplitByPrimersOnFlash.pdf



pipeline: VHBass3_SplitByPrimersOnFlash.pdf

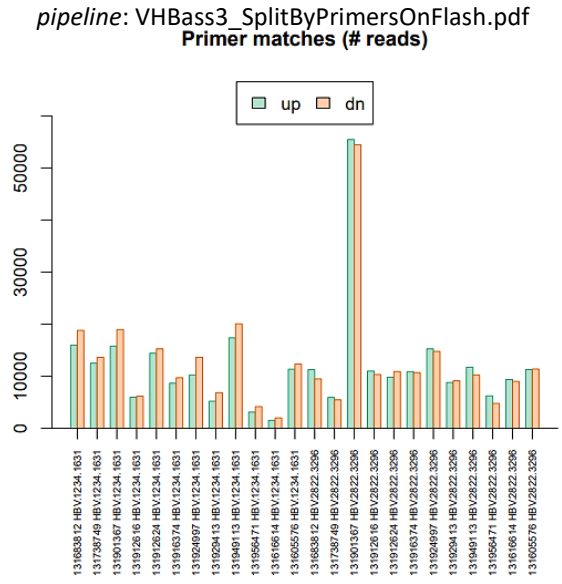
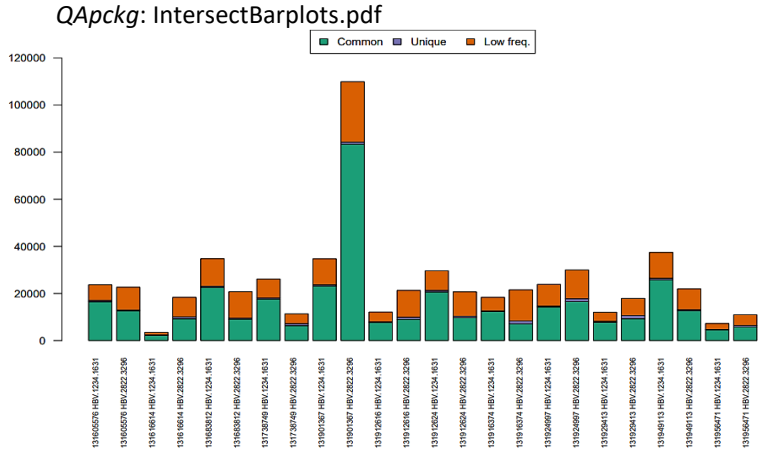


Figura 7. Gràfics de barres representant el nombre de *reads* assignats a les cadenes *forward* (up) i *reverse* (dn) per a cadascun dels 2 amplicons avaluats per pacient (eix Y). A l'eix X s'indica l'identificador de pacient seguit de les posicions de l'amplicó corresponent. A l'esquerra es mostren els resultats d'aplicar el paquet de R, i a la dreta els obtinguts a partir del *pipeline* original.

La figura 8 mostra un dels gràfics generats amb el paquet de R y el *pipeline* original després d'obtenir els haplotips consens a partir de la intersecció de les cadenes *forward* i *reverse*, on es representa el nº de *reads* comuns, els únics de cadena i els descartats per presentar una freqüència inferior al límit establert de 0.2%.

QApckg: IntersectBarplots.pdf



pipeline: IntersectBarplots.pdf

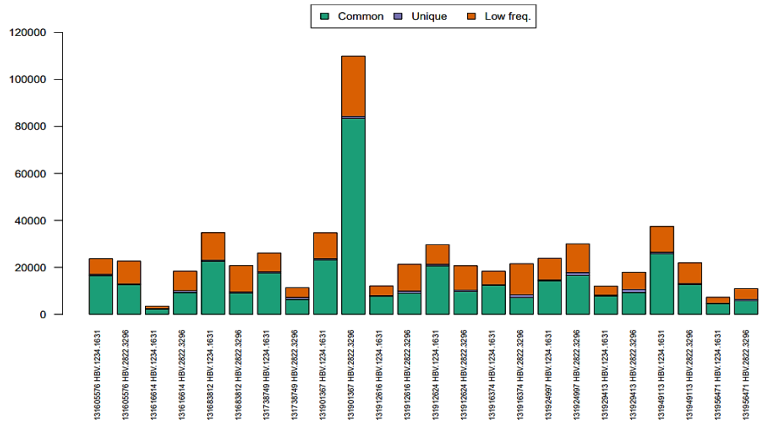
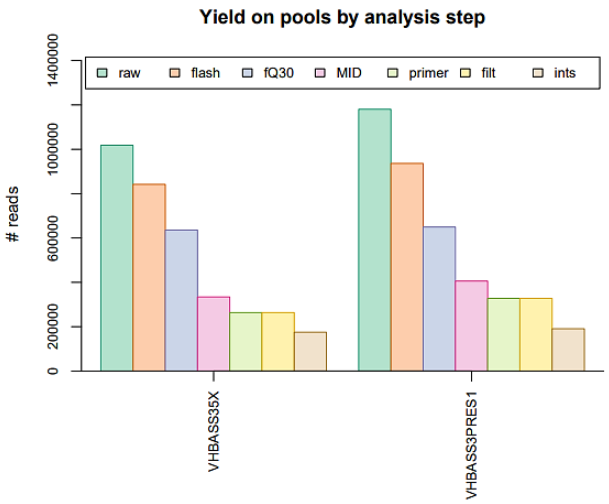


Figura 8. Gràfics de barres representant el nombre de reads (eix Y) comuns en ambdues cadenes (en verd), únics d'una cadena (lila) i descartats per baixa freqüència establint un mínim de 0.2% (taronja), per a cadascun dels amplicons avaluats. A l'eix X s'indica l'identificador de pacient seguit de les posicions de l'amplicó corresponent. A dalt es mostren els resultats d'aplicar el paquet de R, i a sota els obtinguts a partir del pipeline original.

QApckg: GlobalYieldBarplots.pdf



pipeline: GlobalYieldBarplots.pdf

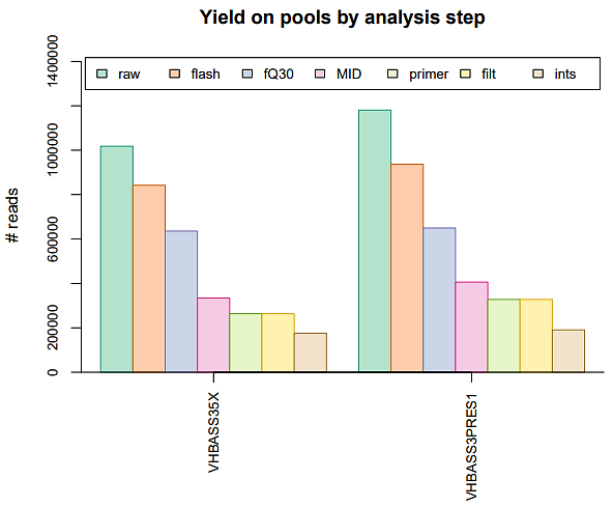


Figura 9. Gràfics de barres representant el rendiment de cadascun dels passos de l'anàlisi de qualitat. Per cadascun dels *pools* avaluats, s'indica el nº de *reads* de les dades crues de seqüenciació, extensió amb FLASH, filtrat per Q30, demultiplexat per MID i *primer*, filtrat per abundància i intersecció d'haplotips. A l'esquerra es mostren els resultats d'aplicar el paquet de R, i a la dreta els obtinguts a partir del *pipeline* original.

La figura 9 inclou els gràfics obtinguts després de computar el rendiment global dels passos de l'anàlisi de qualitat, des de l'extensió dels *reads* aparellats fins a l'obtenció dels haplotips consens. S'observa que en els primers 4 passos és on es perd un major nombre de *reads*, i en especial en el demultiplexat per MIDs. En general, el nombre de *reads* final és semblant entre els dos *pools* avaluats, tot i partir de més lectures de l'amplicó preS1.

5.3 Genotipat del VHB

Un cop obtinguts els haplotips consens mitjançant l'aplicació del *pipeline* original i simplificat sobre les dades del projecte VHBass3, ha estat necessari contrastar si els resultats de genotipat obtinguts resulten idèntics (justificat a la secció 4.4). De la mateixa manera que a la secció anterior, és important tenir en compte la gran quantitat d'arxius resultants de realitzar el genotipat (un total de 19), per la qual cosa es presenten a continuació únicament els gràfics més rellevants. La resta de materials obtinguts es poden consultar a través del repositori GitHub (veure materials suplementaris).

La figura 10 inclou els resultats de genotipat per a un dels amplicons d'estudi, en aquest cas 5'X, a mode d'exemple. Cal tenir en compte, però, que en totes figures presentades en aquesta secció el gràfic indicat com *QApckg* no ha estat generat mitjançant el paquet de R. Les comparacions presentades mostren els resultats d'aplicar el *pipeline* original (en concret la part del genotipat del VHB) sobre els haplotips consens obtinguts a partir de les dues aproximacions.

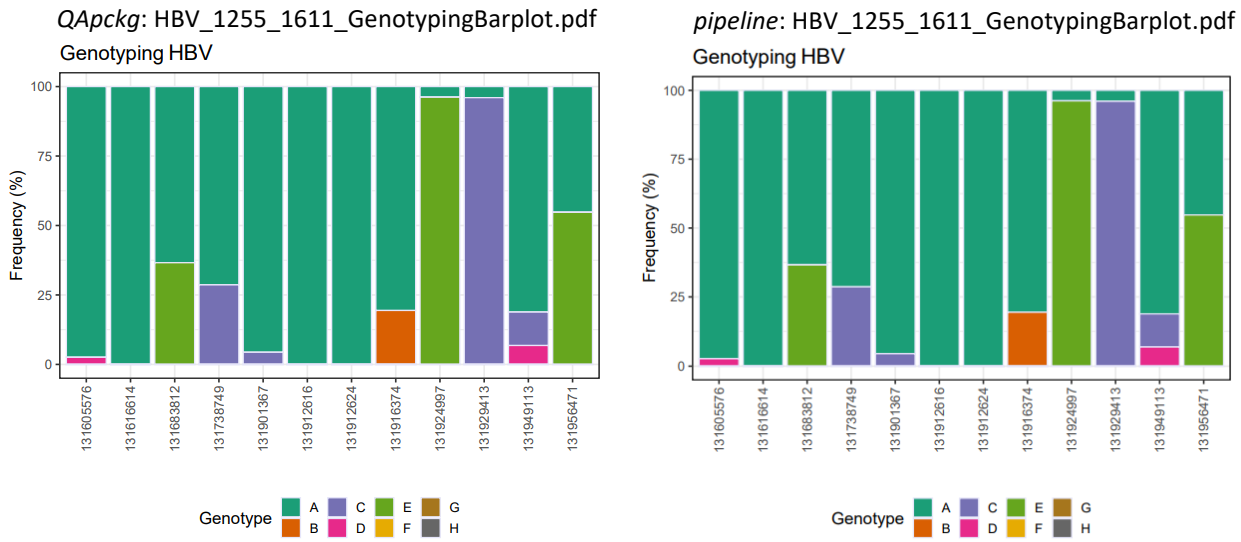


Figura 10. Gràfics representatius de la freqüència observada de cada genotip del VHB per a cadascun dels pacients de l'anàlisi. En aquest cas les dades només corresponen a l'amplicó 5'X (posicions 1255-1611). A l'esquerra es mostren els resultats obtinguts a partir dels haplotips del paquet de R, i a la dreta els obtinguts a partir del *pipeline* original.

La figura 11 inclou els resultats de genotipat dels diferents haplotips avaluats per un dels amplicons d'estudi (5'X). Aquest procés es realitza mitjançant un anàlisi discriminant basat en distàncies (*DB rule*) [38], de manera que per a cadascun dels haplotips, es computa la ràtio entre les distàncies als dos genotips més propers inferits per alineament múltiple. En aquest cas es considera que una ràtio major a 2 (és a dir, que la distància entre l'haplotip i el genotip més proper és més del doble que la distància amb el segon més proper) implica que el genotip inferit és prou fiable, la qual cosa succeeix en totes les mostres.

QApckg: HBV_1255_1611_HplGenotypingConfidence.pdf *pipeline*: HBV_1255_1611_ HplGenotypingConfidence.pdf

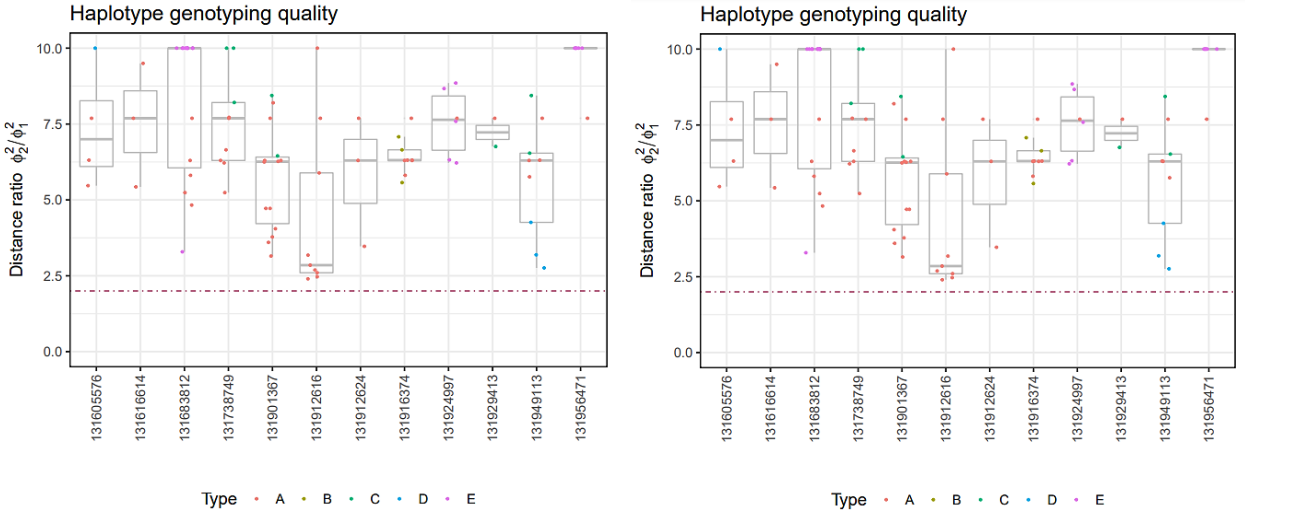
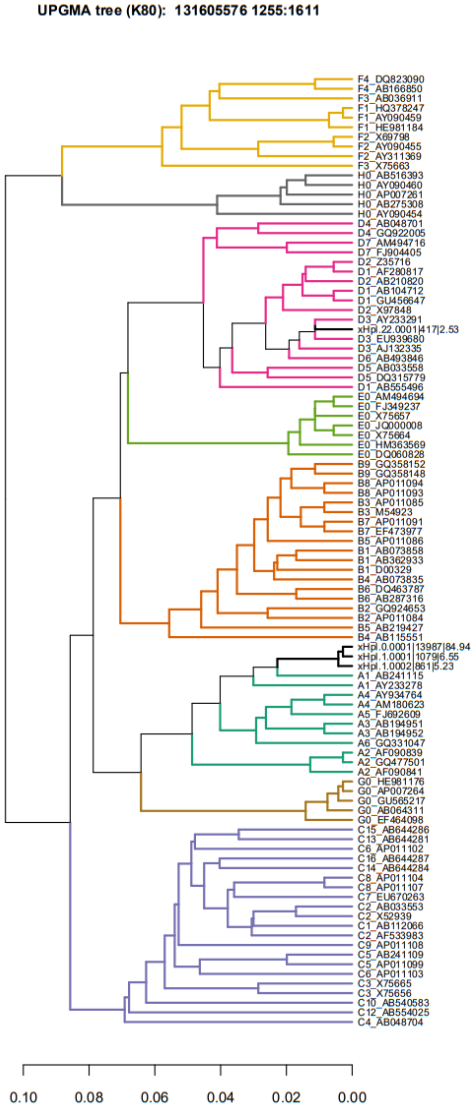


Figura 11. Representació de la ràtio de distàncies dels dos genotips més propers a cadascun dels haplotips consens avaluats. En aquest cas les dades només corresponen a l'amplicó 5'X (posicions 1255-1611). A l'esquerra es mostren els resultats obtinguts a partir dels haplotips del paquet de R, i a la dreta els obtinguts a partir del *pipeline* original.

La figura 12 presenta de forma específica els resultats de la reconstrucció filogenètica per a un únic amplicó i pacient (a mode d'exemple), realitzada a partir de l'alineament de les seqüències d'haplotips respecte un conjunt de seqüències de referència obtingudes de GenBank. En aquest cas es visualitzen 3 haplotips agrupats amb les seqüències del genotip A i un haplotip associat al genotip D.

QApckg: HBV_1255_1611_HplGenotyping.pdf



pipeline: HBV_1255_1611_HplGenotyping.pdf

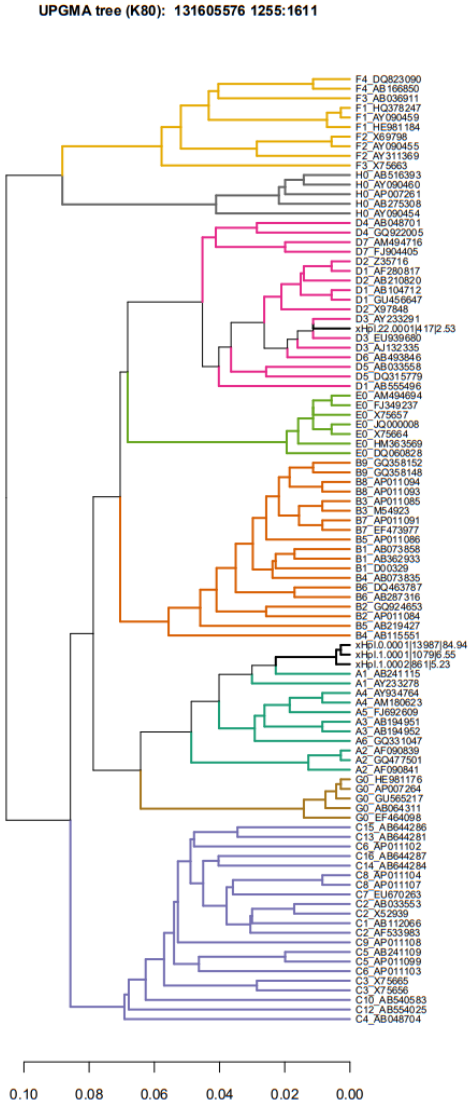


Figura 12. Arbre filogenètic UPGMA representant les relacions entre els haplotips consens obtinguts de l'amplicó 5'X per al pacient amb identificador 131605576. Les distàncies genètiques s'han computat amb el model de Kimura-80 [39]. A l'esquerra es mostren els resultats obtinguts a partir dels haplotips del paquet de R, i a la dreta els obtinguts a partir del *pipeline* original.

6 Discussió

Els resultats presentats a la secció anterior demostren que la simplificació realitzada a partir del paquet *QApckg* generat permet generar exactament els mateixos resultats que la secció corresponent del *pipeline* original. A més, els haplotips consens finalment obtinguts resulten ser els mateixos mitjançant ambdues aproximacions, ja que la realització del genotipat sobre els dos conjunts d'haplotips resulta en els mateixos gràfics. No obstant, cal mencionar que un dels arxius obtinguts en el procés d'anàlisi de qualitat, concretament en la intersecció dels haplotips d'ambdues cadenes, mostra certes diferències respecte els gràfics mostrats en aplicar el *pipeline* original. Aquest arxiu correspon a l'anomenat *MA.Intersects.plots.pdf*, el qual es presenta als materials suplementaris donada la seva extensió. Les diferències trobades són degudes al canvi

afegit en l'alineament múltiple dels haplotips *forward* i *reverse*, ja que en el paquet de R aquest es realitza mitjançant la funció `muscle` i no a partir de l'executable del programa. Tot i que l'algoritme d'aquesta funció i el de l'executable MUSCLE [31] funcionen igual, els resultats obtinguts varien en relació a la ordenació de les seqüències alineades. Malgrat això, s'ha corroborat en les gràfiques posteriors que aquest canvi no té cap impacte quant als resultats finals i els haplotips consens, ni tampoc sobre el genotipat realitzat.

La taula 3 proporciona una visió resumida d'algunes de les simplificacions que suposa l'aplicació del paquet *QApckg*, demostrant els avantatges que presenta respecte l'aplicació del *pipeline* original. Tenint en compte els punts febles descrits a la taula 2 (secció 4.1), es demostra que el paquet de R ha permès solucionar gran part d'aquests desavantatges; d'aquesta forma, l'usuari només requereix de la instal·lació del paquet local construït, les dades crues de seqüenciació i els arxius de dades, i a partir d'un únic *script* de R pot definir els paràmetres desitjats i realitzar un anàlisi de qualitat complet en un temps reduït. Tot i això, cal mencionar que en funció dels arxius FASTQ comprimits emprats en l'anàlisi, aquest pot comportar una inversió de temps major; per aquesta raó resulta important el procés de paral·lelització, ja que permet obtenir els resultats desitjats en una menor quantitat de temps.

Convé ressaltar que no tots els punts febles del *pipeline* original han pogut ser resolts, ja que la quantitat de resultats obtinguts en l'aplicació del paquet continua essent considerable, incloent alguns difícils d'interpretar. Aquest fet es podria solucionar mitjançant la creació d'un informe assistencial que reculli els resultats de l'anàlisi més rellevants a nivell clínic, la qual cosa es presenta com a línia de futur del treball (secció 7.2).

7 Conclusions

7.1 Conclusions

Tenint en compte els objectius plantejats inicialment, les conclusions del present treball es poden resumir en les següents:

1. El VHB presenta diversos reptes que dificulten el seu tractament, monitorització i predicció de l'evolució clínica actualment. Disposar d'un mètode senzill i ràpid per tal de realitzar el seu genotipat pot esdevenir crucial a l'hora de prendre decisions sobre el tractament i prevenció de la infecció.
2. La comunicació interdisciplinària entre el personal sanitari i bioinformàtic resulta imprescindible per tal d'assolir els objectius proposats en qualsevol projecte. És molt important que els investigadors coneguin els aspectes més rellevants de l'anàlisi bioinformàtic derivat de les seves dades, ja sigui a partir de la documentació dels *scripts* emprats o d'un informe assistencial que permeti entendre els resultats obtinguts.
3. En aquest projecte s'ha descrit i documentat l'estructura d'un *pipeline* per a l'anàlisi de qualitat i l'obtenció d'haplotips consens a partir de dades NGS derivades de la seqüenciació de dues regions del genoma del VHB. La creació d'un diagrama de flux permet descriure aquesta estructura de forma visual i organitzada.
4. La consecució del treball ha permès entendre el procediment global de creació i execució d'un paquet de R, des de la definició de les funcions incloses fins a la redacció de la

documentació associada, la qual ha de ser comprensible per als usuaris que vulguin inicialitzar el processament de dades NGS.

5. S'han contrastat els resultats obtinguts amb la implementació del paquet creat amb els generats originalment amb el *pipeline* proporcionat, i tal com s'ha vist aquests han esdevinguts idèntics i s'han obtingut en un període de temps més curt, demostrant l'adequació del paquet *QApckg*. Ara bé, cal remarcar que la simplificació realitzada no ha estat una tasca senzilla, doncs l'edició del codi inicial pot propiciar l'aparició d'errors en els resultats.
6. La simplificació de diversos *scripts* bioinformàtics en un paquet funcional permetrà als usuaris/investigadors sense formació en programació realitzar fàcilment diversos anàlisis rellevants en l'àmbit de la recerca.

7.2 Línies de futur

El present treball ha permès la simplificació d'un *pipeline* bioinformàtic en un paquet de R útil per a l'anàlisi de qualitat derivat de dades NGS, el qual a més disposa de tota la documentació necessària per tal de ser aplicat per personal no especialitzat en bioinformàtica. No obstant, l'extensió del treball ha permès únicament simplificar part d'aquest *pipeline*, deixant de banda l'últim dels *scripts* revisats inicialment. Per tant, una de les principals línies de futur seria l'optimització de la resta del *pipeline*, incloent la part del genotipat del VHB, en el mateix o en un altre paquet de R, per tal de facilitar el procés complert de l'anàlisi.

Tanmateix, el paquet *QApckg* podria optimitzar-se encara més, mitjançant l'obtenció d'un informe assistencial que permeti comprendre els resultats més importants derivats de la seva execució, així com la implementació de la paral·lelització en totes o la majoria de les funcions del paquet per tal de reduir encara més el temps del processament bioinformàtic. Addicionalment, es podria millorar el paquet mitjançant programació orientada en objectes, concretament mitjançant l'ús d'objectes i classes S4 [40]. Aquest mètode suposa certs avantatges respecte la programació S3, tot i que requereix una major expertesa en l'àmbit de la programació.

D'altra banda, el paquet creat està dissenyat per processar dades NGS del VHB provinents de la plataforma MiSeq (Illumina), de manera que les funcions incloses s'han d'executar de forma seqüencial. Seria convenient comprovar l'aplicabilitat del paquet en qüestió en l'estudi d'altres virus o organismes, així com la possibilitat d'analitzar dades de NGS procedents de plataformes de seqüenciació diferents. A més, es podrien explorar els canvis requerits per poder prescindir d'algunes de les funcions del paquet en cas necessari, per exemple si no es fan servir els adaptadors MID i s'identifiquen els pacients amb algun sistema de marcatge alternatiu.

7.3 Seguiment de la planificació

La planificació plantejada en la secció 2.4 del present treball ha pogut seguir-se adequadament, malgrat presentar certes diferències respecte la temporització descrita inicialment al pla de treball.

Un dels riscos proposats al pla de treball tenia en compte la possible dificultat associada al procés de simplificació dels diversos *scripts* del *pipeline*, així com l'aparició de contratemps en la implementació de les funcions. Per tant, la quantitat d'hores invertides en les tasques de

creació de funcions i del paquet de R ha estat considerable, propiciant un canvi important en la planificació derivat de l'aplicació d'un dels plans de contingència previstos. La limitació temporal derivada de la dificultat del codi ha impedit la implementació de funcions per l'últim dels *scripts* del diagrama de flux (figura 4), en el qual es generava un informe final resumint la cobertura dels amplicons obtinguts. Cal tenir en compte, però, que la gran part de la informació retornada per aquest últim *script* es troba ja inclosa en els resultats obtinguts en passos anteriors, per la qual cosa restringir aquest pas no comporta una pèrdua de dades considerable.

Malgrat això, l'acció de mitigació més rellevant implementada va ser descartar la creació de l'informe assistencial a partir dels resultats obtinguts del *pipeline* simplificat. Això ha estat degut a que les hores previstes per a la generació d'aquest informe han estat destinades completament a la creació del paquet de R amb les diverses funcions, per la qual cosa no ha estat possible dissenyar l'informe.

8 Glossari

VHB	Virus de l'Hepatitis B
CHC	Carcinoma HepatoCel·lular
DNA	*Desoxyribonucleic Acid*
rcDNA	*relaxed-circular* DNA
ORF	*Open Reading Frame*
HBcAg	*Hepatitis B core Antigen*
HBeAg	*Hepatitis B e Antigen*
HBx	*Hepatitis B x protein*
cccDNA	*circular covalently closed circular* DNA
mRNA	*messenger Ribonucleic Acid*
pgRNA	*pregenomic* RNA
dslDNA	*double-stranded linear* DNA
HBsAg	*Hepatitis B surface Antigen*
NGS	*Next Generation Sequencing*
INNO-LiPA	*Innogenetics©- Line Probe Assay*
PCR	*Polymerase chain reaction*
SNP	*Single Nucleotide Polymorphism*
RFLP	*Restriction Fragment Length Polymorphism*
FT-RDB	*Flow-Through Reverse Dot Blot*
FRMP	*Restriction Fragment Mass Polymorphism*
MALDI-TOF	*Matrix-Assisted Laser Desorption/Ionization - Time-Of-Flight*
FRET	*Förster/Fluorescence resonance energy transfer*
VHIR	Vall d'Hebron Institut de Recerca
BLAST	*Basic Local Alignment Search Tool*

NCBI	*National Center for Biotechnology Intormation*
HBVdb	*Hepatitis B Virus data base*
HBV STAR	*Hepatitis B Virus Subtype Analyzer*
QA	*Quality Assessment*
QC	*Quality Control*
PIQA	*Pipeline Quality Analysis*
FLASH	*Fast Length Adjustment of SHort reads*
Q-*score*	*Quality score*
MID	*Multiplex IDentifier*
MUSCLE	*MUltiple Sequence Comparison by Log-Expectation*
DB *rule*	*Distance-Based rule*

9 Bibliografia

1. WHO. WHO | Hepatitis B [Internet]. World Health Organization; 2021 [Accessed 19 Feb 2022]. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-b
2. Caballero, A., Tabernero, D., Buti, M., *et al*. (2018). Hepatitis B virus: The challenge of an ancient virus with multiple faces and a remarkable replication strategy. *Antiviral Research*, *158*, 34-44. https://doi.org/10.1016/j.antiviral.2018.07.019
3. Guirgis, B. S. S., Abbas, R. O., & Azzazy, H. M. E. (2010). Hepatitis B virus genotyping: Current methods and clinical implications. *International Journal of Infectious Diseases*, *14*(11), e941-e953. https://doi.org/10.1016/j.ijid.2010.03.020
4. Rodriguez-Frias, F. (2013). Quasispecies structure, cornerstone of hepatitis B virus infection: Mass sequencing approach. *World Journal of Gastroenterology*, *19*(41), 6995. https://doi.org/10.3748/wjg.v19.i41.6995
5. Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rates of evolutionary change in viruses: Patterns and determinants. *Nature Reviews Genetics*, *9*(4), 267-276. https://doi.org/10.1038/nrg2323
6. Velkov, S., Ott, J., Protzer, U., & Michler, T. (2018). The Global Hepatitis B Virus Genotype Distribution Approximated from Available Genotyping Data. *Genes*, *9*(10), 495. https://doi.org/10.3390/genes9100495
7. Pourkarim, M. R. (2014). Molecular identification of hepatitis B virus genotypes/subgenotypes: Revised classification hurdles and updated resolutions. *World Journal of Gastroenterology*, *20*(23), 7152. https://doi.org/10.3748/wjg.v20.i23.7152
8. Araujo, N. M. (2015). Hepatitis B virus intergenotypic recombinants worldwide: An overview. Infection, *Genetics and Evolution, 36*, 500-510. https://doi.org/10.1016/j.meegid.2015.08.024
9. Sarin, S. K., Kumar, M., Lau, G. K., *et al* (2016). Asian-Pacific clinical practice guidelines on the management of hepatitis B: A 2015 update. *Hepatology International*, *10*(1), 1-98. https://doi.org/10.1007/s12072-015-9675-4
10. Terrault, N. A., Lok, A. S. F., McMahon, B. J., *et al* (2018). Update on prevention, diagnosis, and treatment of chronic hepatitis B: AASLD 2018 hepatitis B guidance. *Hepatology*, *67*(4), 1560-1599. https://doi.org/10.1002/hep.29800
11. European Association for the Study of the Liver. (2017). EASL 2017 Clinical Practice Guidelines on the management of hepatitis B virus infection. *Journal of hepatology*, *67*(2), 370–398. https://doi.org/10.1016/j.jhep.2017.03.021
12. Caballero, A., Gregori, J., Homs, M., *et al*. (2015). Complex Genotype Mixtures Analyzed by Deep Sequencing in Two Different Regions of Hepatitis B Virus. *PLOS ONE*, *10*(12), e0144816. https://doi.org/10.1371/journal.pone.0144816
13. Guerrero-Murillo M., Gregori i Font, J. (2022). *QSutils: Quasispecies Diversity*. R package version 1.14.0. (https://bioconductor.org/packages/release/bioc/html/QSutils.html)
14. Croagh, C. M. (2015). Genotypes and viral variants in chronic hepatitis B: A review of epidemiology and clinical relevance. *World Journal of Hepatology*, *7*(3), 289. https://doi.org/10.4254/wjh.v7.i3.289
15. Bell, T. G., & Kramvis, A. (2016). *Bioinformatics—Updated Features and Applications: The Study of Hepatitis B Virus Using Bioinformatics*. IntechOpen. https://doi.org/10.5772/63076

16. Ma, Y., Ding, Y., Juan, F., & Dou, X. G. (2011). Genotyping the hepatitis B virus with a fragment of the HBV DNA polymerase gene in Shenyang, China. *Virology Journal,* 8, 315. https://doi.org/10.1186/1743-422X-8-315
17. Osiowy, C., & Giles, E. (2003). Evaluation of the INNO-LiPA HBV genotyping assay for determination of hepatitis B virus genotype. *Journal of clinical microbiology, 41*(12), 5473–5477. https://doi.org/10.1128/JCM.41.12.5473-5477.2003
18. Tadokoro, K., Kobayashi, M., Yamaguchi, T., Suzuki, F., Miyauchi, S., Egashira, T., & Kumada, H. (2006). Classification of hepatitis B virus genotypes by the PCR-Invader method with genotype-specific probes. *Journal of Virological Methods*, *138*(1), 30-39. https://doi.org/10.1016/j.jviromet.2006.07.014
19. González, C., Tabernero, D., Cortese, M. F., *et al*. (2018). Detection of hyper-conserved regions in hepatitis B virus X gene potentially useful for gene therapy. *World Journal of Gastroenterology, 24*(19), 2095-2107. https://doi.org/10.3748/wjg.v24.i19.2095
20. Ewing, B., & Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research, 8*(3), 186-194. https://doi.org/10.1101/gr.8.3.186
21. Măndoiu, I., & Zelikovsky, A. (2016). *Computational Methods for Next Generation Sequencing Data Analysis*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119272182
22. Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data* [Online]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
23. Fotouhi, A., Majidi, M., Külekci, M.O. (2018). *Quality Assessment of High-Throughput DNA Sequencing Data via Range Analysis*. In: Rojas, I., Ortuño, F. (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2018. *Lecture Notes in Computer Science*, vol 10813. Springer, Cham. https://doi.org/10.1007/978-3-319-78723-7_37
24. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170
25. Soria, M. E., Gregori, J., Chen, Q., *et al.* (2018). Pipeline for specific subtype amplification and drug resistance detection in hepatitis C virus. *BMC Infectious Diseases*, *18*(1), 446. https://doi.org/10.1186/s12879-018-3356-6
26. Pagès, H., Aboyoun, P., Gentleman, R., *et al* (2022). *Biostrings: Efficient manipulation of biological strings*. R package version 2.64.0. https://bioconductor.org/packages/Biostrings
27. Paradis, E., Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526-528. https://CRAN.R-project.org/package=ape
28. Godoy, C., Sopena, S., Gregori, J., *et al.* (s.d.). 5. Hepatitis D Virus quasispecies study: Experimental and bioinformatic analysis by next generation sequencing methodology. 44.
29. Godoy, C., Tabernero, D., Sopena, S., *et al.* (2019). Characterization of hepatitis B virus X gene quasispecies complexity in mono-infection and hepatitis delta virus superinfection. *World Journal of Gastroenterology, 25*(13), 1566-1579. https://doi.org/10.3748/wjg.v25.i13.1566
30. Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957-2963. https://doi.org/10.1093/bioinformatics/btr507

31. Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113. https://doi.org/10.1186/1471-2105-5-113
32. Sepulveda, J. L. (2020). Using R and Bioconductor in Clinical Genomics and Transcriptomics. *The Journal of Molecular Diagnostics*, 22(1), 3-20. https://doi.org/10.1016/j.jmoldx.2019.08.006
33. Robey, R., & Zamora, Y. (2021). *Parallel and High Performance Computing*. Manning Publications Co. ISBN: 978-1-61729-646-8.
34. Peng, R. [Roger] (2016). *R Programming for Data Science: Parallel Computation*. Lulu.com. https://bookdown.org/rdpeng/rprogdatascience/parallel-computation.html
35. GitHub. (2014). *post-10-mclapply-hack.R* by VanHoudnos, N. https://github.com/nathanvan/mcmc-in-irt/blob/master/post-10-mclapply-hack.R
36. RStudio Team (2022). *RStudio Support: Package Development*. https://support.rstudio.com/hc/en-us/sections/200130627-Package-Development
37. Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32, 1792-1797. doi: 10.1093/nar/gkh340.
38. Cuadras C. (1989). *Distance analysis in discrimination and classification using both continuous and categorical variables*. In: Dodge Y, editor. *Statistical AData analysis and Interference*. Elsevier. pp. 459–473.
39. Kimura M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, *16*(2), 111–120. https://doi.org/10.1007/BF01731581
40. Genolini, C. (2008). *A (not so) short introduction to S4: Object Programming in R*. https://cran.r-project.org/doc/contrib/Genolini-S4tutorialV0-5en.pdf

Annexos

Llistat de materials suplementaris del treball

→ https://github.com/aliafdz/QA_genotipat_pipeline: Repositori GitHub on es troba documentat el codi del *pipeline* original.

- ***VHBass3_pipeline***: carpeta d'arxius (inclosa al repositori del punt anterior) on s'inclouen els gràfics i taules obtingudes després de l'aplicació del *pipeline* original sencer sobre les dades reals de NGS identificades amb el nom de projecte VHBass3.

→ https://github.com/aliafdz/QApckg: repositori GitHub on es troben tots els documents relatius al paquet de R generat, incloent el codi de les funcions, la seva documentació, el fitxer comprimit del paquet i el seu manual de referència.

- ***VHBass3_QApckg***: carpeta d'arxius (inclosa al repositori del punt anterior) on s'inclouen els gràfics i taules obtingudes després de l'aplicació de les funcions definides sobre les dades reals de NGS, identificades amb el nom de projecte VHBass3. També s'inclouen els resultats generats en aplicar la part de genotipat del *pipeline* original.

→ ***TFM_Alicia_PipelineSimplificat.R***: *script* de R en el qual s'apliquen les diverses funcions del paquet en l'ordre correcte per tal de realitzar el procés complet d'anàlisi de qualitat, des de l'extensió dels *reads* aparellats amb el programa FLASH fins a l'obtenció dels haplotips consens.