Dynamique de l'évolution intra-patient du VHC par NGS

Alban Caporossi

AVERTISSEMENT

LIENS

UNIVERSITE GRENOBLE ALPES

UFR DE MEDECINE DE GRENOBLE

Année : 2017

Dynamique de l'évolution intra-patient du VHC par NGS

THESE

PRESENTEE POUR L'OBTENTION DU TITRE DE DOCTEUR EN MEDECINE

DIPLOME D'ETAT

Alban Caporossi

[Données à caractère personnel]

THESE SOUTENUE PUBLIQUEMENT A LA FACULTE DE MEDECINE DE GRENOBLE

Le : 24/10/2017

DEVANT LE JURY COMPOSE DE

Président du jury :
Monsieur le Professeur Alexandre MOREAU-GAUDRY
Membres :
Monsieur le Professeur Pascal POIGNARD (directeur de thèse)
Monsieur le Professeur Olivier EPAULARD
Monsieur le Professeur Olivier FRANCOIS

UFR de Médecine

UNIVERSITÉ Grenoble Alpes

Doyen de la Faculté : **Pr. Jean Paul ROMANET**

Année 2017-2018

ENSEIGNANTS A L'UFR DE MEDECINE

CORPS	NOM-PRENOM	Discipline universitaire
PU-PH	**ALBALADEJO Pierre**	**Anesthésiologie réanimation**
PU-PH	**APTEL Florent**	**Ophtalmologie**
PU-PH	**ARVIEUX-BARTHELEMY Catherine**	**Chirurgie générale**
PU-PH	**BAILLET Athan**	**Rhumatologie**
PU-PH	**BARONE-ROCHETTE Gilles**	**Cardiologie**
PU-PH	**BAYAT Sam**	**Physiologie**
PU-PH	**BENHAMOU Pierre Yves**	**Endocrinologie, diabète et maladies métaboliques**
PU-PH	**BERGER François**	**Biologie cellulaire**
MCU-PH	**BIDART-COUTTON Marie**	**Biologie cellulaire**
MCU-PH	**BOISSET Sandrine**	**Agents infectieux**
PU-PH	**BONAZ Bruno**	**Gastro-entérologie, hépatologie, addictologie**
PU-PH	**BONNETERRE Vincent**	**Médecine et santé au travail**
PU-PH	**BOREL Anne-Laure**	**Endocrinologie, diabète et maladies métaboliques**
PU-PH	**BOSSON Jean-Luc**	**Biostatistiques, informatique médicale et technologies de communication**
MCU-PH	**BOTTARI Serge**	**Biologie cellulaire**
PU-PH	**BOUGEROL Thierry**	**Psychiatrie d'adultes**
PU-PH	**BOUILLET Laurence**	**Médecine interne**
PU-PH	**BOUZAT Pierre**	**Réanimation**
MCU-PH	**BRENIER-PINCHART Marie Pierre**	**Parasitologie et mycologie**
PU-PH	**BRICAULT Ivan**	**Radiologie et imagerie médicale**
PU-PH	**BRICHON Pierre-Yves**	**Chirurgie thoracique et cardio- vasculaire**
MCU-PH	**BRIOT Raphaël**	**Thérapeutique, médecine d'urgence**
MCU-PH	**BROUILLET Sophie**	**Biologie et médecine du développement et de la reproduction**
PU-PH	**CAHN Jean-Yves**	**Hématologie**
PU-PH	**CARPENTIER Françoise**	**Thérapeutique, médecine d'urgence**
PU-PH	**CARPENTIER Patrick**	**Chirurgie vasculaire, médecine vasculaire**
PU-PH	**CESBRON Jean-Yves**	**Immunologie**
PU-PH	**CHABARDES Stephan**	**Neurochirurgie**
PU-PH	**CHABRE Olivier**	**Endocrinologie, diabète et maladies métaboliques**
PU-PH	**CHAFFANJON Philippe**	**Anatomie**
PU-PH	**CHARLES Julie**	**Dermatologie**
PU-PH	**CHAVANON Olivier**	**Chirurgie thoracique et cardio- vasculaire**
PU-PH	**CHIQUET Christophe**	**Ophtalmologie**

PU-PH	**CHIRICA Mircea**	**Chirurgie générale**
PU-PH	**CINQUIN Philippe**	**Biostatistiques, informatique médicale et technologies de communication**
MCU-PH	**CLAVARINO Giovanna**	**Immunologie**
PU-PH	**COHEN Olivier**	**Biostatistiques, informatique médicale et technologies de communication**
PU-PH	**COURVOISIER Aurélien**	**Chirurgie infantile**
PU-PH	**COUTURIER Pascal**	**Gériatrie et biologie du vieillissement**
PU-PH	**CRACOWSKI Jean-Luc**	**Pharmacologie fondamentale, pharmacologie clinique**
PU-PH	**CURE Hervé**	**Oncologie**
PU-PH	**DEBILLON Thierry**	**Pédiatrie**
PU-PH	**DECAENS Thomas**	**Gastro-entérologie, Hépatologie**
PU-PH	**DEMATTEIS Maurice**	**Addictologie**
MCU-PH	**DERANSART Colin**	**Physiologie**
PU-PH	**DESCOTES Jean-Luc**	**Urologie**
MCU-PH	**DETANTE Olivier**	**Neurologie**
MCU-PH	**DIETERICH Klaus**	**Génétique et procréation**
MCU-PH	**DOUTRELEAU Stéphane**	**Physiologie**
MCU-PH	**DUMESTRE-PERARD Chantal**	**Immunologie**
PU-PH	**EPAULARD Olivier**	**Maladies Infectieuses et Tropicales**
PU-PH	**ESTEVE François**	**Biophysique et médecine nucléaire**
MCU-PH	**EYSSERIC Hélène**	**Médecine légale et droit de la santé**
PU-PH	**FAGRET Daniel**	**Biophysique et médecine nucléaire**
PU-PH	**FAUCHERON Jean-Luc**	**Chirurgie générale**
MCU-PH	**FAURE Julien**	**Biochimie et biologie moléculaire**
PU-PH	**FERRETTI Gilbert**	**Radiologie et imagerie médicale**
PU-PH	**FEUERSTEIN Claude**	**Physiologie**
PU-PH	**FONTAINE Éric**	**Nutrition**
PU-PH	**FRANCOIS Patrice**	**Epidémiologie, économie de la santé et prévention**
MCU-MG	**GABOREAU Yoann**	**Médecine Générale**
PU-PH	**GARBAN Frédéric**	**Hématologie, transfusion**
PU-PH	**GAUDIN Philippe**	**Rhumatologie**
PU-PH	**GAVAZZI Gaétan**	**Gériatrie et biologie du vieillissement**
PU-PH	**GAY Emmanuel**	**Neurochirurgie**
MCU-PH	**GILLOIS Pierre**	**Biostatistiques, informatique médicale et technologies de communication**
MCU-PH	**GRAND Sylvie**	**Radiologie et imagerie médicale**
PU-PH	**GRIFFET Jacques**	**Chirurgie infantile**
PU-PH	**GUEBRE-EGZIABHER Fitsum**	**Néphrologie**
MCU-PH	**GUZUN Rita**	**Endocrinologie, diabétologie, nutrition, éducation thérapeutique**
PU-PH	**HAINAUT Pierre**	**Biochimie, biologie moléculaire**
PU-PH	**HENNEBICQ Sylviane**	**Génétique et procréation**
PU-PH	**HOFFMANN Pascale**	**Gynécologie obstétrique**
PU-PH	**HOMMEL Marc**	**Neurologie**
PU-MG	**IMBERT Patrick**	**Médecine Générale**
PU-PH	**JOUK Pierre-Simon**	**Génétique**
PU-PH	**JUVIN Robert**	**Rhumatologie**

PU-PH	**KAHANE Philippe**	**Physiologie**
MCU-PH	**KASTLER Adrian**	**Radiologie et imagerie médicale**
PU-PH	**KRACK Paul**	**Neurologie**
PU-PH	**KRAINIK Alexandre**	**Radiologie et imagerie médicale**
PU-PH	**LABARERE José**	**Epidémiologie ; Eco. de la Santé**
MCU-PH	**LABLANCHE Sandrine**	**Endocrinologie, diabète et maladies métaboliques**
MCU-PH	**LANDELLE Caroline**	**Bactériologie - virologie**
MCU-PH	**LAPORTE François**	**Biochimie et biologie moléculaire**
MCU-PH	**LARDY Bernard**	**Biochimie et biologie moléculaire**
MCU-PH	**LARRAT Sylvie**	**Bactériologie, virologie**
MCU - PH	**LE PISSART Audrey**	**Biochimie et biologie moléculaire**
PU-PH	**LECCIA Marie-Thérèse**	**Dermato-vénéréologie**
PU-PH	**LEROUX Dominique**	**Génétique**
PU-PH	**LEROY Vincent**	**Gastro-entérologie, hépatologie, addictologie**
PU-PH	**LEVY Patrick**	**Physiologie**
PU-PH	**LONG Jean-Alexandre**	**Urologie**
PU-PH	**MAGNE Jean-Luc**	**Chirurgie vasculaire**
MCU-PH	**MAIGNAN Maxime**	**Thérapeutique, médecine d'urgence**
PU-PH	**MAITRE Anne**	**Médecine et santé au travail**
MCU-PH	**MALLARET Marie-Reine**	**Epidémiologie, économie de la santé et prévention**
MCU-PH	**MARLU Raphaël**	**Hématologie, transfusion**
MCU-PH	**MAUBON Danièle**	**Parasitologie et mycologie**
PU-PH	**MAURIN Max**	**Bactériologie - virologie**
MCU-PH	**MC LEER Anne**	**Cytologie et histologie**
PU-PH	**MERLOZ Philippe**	**Chirurgie orthopédique et traumatologie**
PU-PH	**MORAND Patrice**	**Bactériologie - virologie**
PU-PH	**MOREAU-GAUDRY Alexandre**	**Biostatistiques, informatique médicale et technologies de communication**
PU-PH	**MORO Elena**	**Neurologie**
PU-PH	**MORO-SIBILOT Denis**	**Pneumologie**
PU-PH	**MOUSSEAU Mireille**	**Cancérologie**
PU-PH	**MOUTET François**	**Chirurgie plastique, reconstructrice et esthétique ; brûlologie**
MCU-PH	**PACLET Marie-Hélène**	**Biochimie et biologie moléculaire**
PU-PH	**PALOMBI Olivier**	**Anatomie**
PU-PH	**PARK Sophie**	**Hémato - transfusion**
PU-PH	**PASSAGGIA Jean-Guy**	**Anatomie**
PU-PH	**PAYEN DE LA GARANDERIE Jean-François**	**Anesthésiologie réanimation**
MCU-PH	**PAYSANT François**	**Médecine légale et droit de la santé**
MCU-PH	**PELLETIER Laurent**	**Biologie cellulaire**
PU-PH	**PELLOUX Hervé**	**Parasitologie et mycologie**
PU-PH	**PEPIN Jean-Louis**	**Physiologie**
PU-PH	**PERENNOU Dominique**	**Médecine physique et de réadaptation**
PU-PH	**PERNOD Gilles**	**Médecine vasculaire**
PU-PH	**PIOLAT Christian**	**Chirurgie infantile**
PU-PH	**PISON Christophe**	**Pneumologie**

PU-PH	**PLANTAZ Dominique**	**Pédiatrie**
PU-PH	**POIGNARD Pascal**	**Virologie**
PU-PH	**POLACK Benoît**	**Hématologie**
PU-PH	**POLOSAN Mircea**	**Psychiatrie d'adultes**
PU-PH	**PONS Jean-Claude**	**Gynécologie obstétrique**
PU-PH	**RAMBEAUD Jacques**	**Urologie**
PU-PH	**RAY Pierre**	**Biologie et médecine du développement et de la reproduction**
PU-PH	**REYT Émile**	**Oto-rhino-laryngologie**
PU-PH	**RIGHINI Christian**	**Oto-rhino-laryngologie**
PU-PH	**ROMANET Jean Paul**	**Ophtalmologie**
PU-PH	**ROSTAING Lionel**	**Néphrologie**
MCU-PH	**ROUSTIT Matthieu**	**Pharmacologie fondamentale, pharmaco clinique, addictologie**
MCU-PH	**ROUX-BUISSON Nathalie**	**Biochimie, toxicologie et pharmacologie**
MCU-PH	**RUBIO Amandine**	**Pédiatrie**
PU-PH	**SARAGAGLIA Dominique**	**Chirurgie orthopédique et traumatologie**
MCU-PH	**SATRE Véronique**	**Génétique**
PU-PH	**SAUDOU Frédéric**	**Biologie Cellulaire**
PU-PH	**SCHMERBER Sébastien**	**Oto-rhino-laryngologie**
PU-PH	**SCHWEBEL-CANALI Carole**	**Réanimation médicale**
PU-PH	**SCOLAN Virginie**	**Médecine légale et droit de la santé**
MCU-PH	**SEIGNEURIN Arnaud**	**Epidémiologie, économie de la santé et prévention**
PU-PH	**STAHL Jean-Paul**	**Maladies infectieuses, maladies tropicales**
PU-PH	**STANKE Françoise**	**Pharmacologie fondamentale**
MCU-PH	**STASIA Marie-José**	**Biochimie et biologie moléculaire**
PU-PH	**STURM Nathalie**	**Anatomie et cytologie pathologiques**
PU-PH	**TAMISIER Renaud**	**Physiologie**
PU-PH	**TERZI Nicolas**	**Réanimation**
MCU-PH	**TOFFART Anne-Claire**	**Pneumologie**
PU-PH	**TONETTI Jérôme**	**Chirurgie orthopédique et traumatologie**
PU-PH	**TOUSSAINT Bertrand**	**Biochimie et biologie moléculaire**
PU-PH	**VANZETTO Gérald**	**Cardiologie**
PU-PH	**VUILLEZ Jean-Philippe**	**Biophysique et médecine nucléaire**
PU-PH	**WEIL Georges**	**Epidémiologie, économie de la santé et prévention**
PU-PH	**ZAOUI Philippe**	**Néphrologie**
PU-PH	**ZARSKI Jean-Pierre**	**Gastro-entérologie, hépatologie, addictologie**

PU-PH : Professeur des Universités et Praticiens Hospitaliers
MCU-PH : Maître de Conférences des Universités et Praticiens Hospitaliers
PU-MG : Professeur des Universités de Médecine Générale
MCU-MG : Maître de Conférences des Universités de Médecine Générale

Remerciements

Aux Prs Patrice Morand et Pascal Poignard pour m'avoir accueilli dans le service de Virologie et permis de réaliser ce travail.

Au Pr Olivier François pour son expertise en biologie évolutive et en analyse de données et pour le temps qu'il m'a consacré.

Au Pr Alexandre Moreau-Gaudry pour son soutien dans mon projet qu'il a su intégrer aux projets à venir au niveau de l'établissement.

Au Pr Olivier Epaulard pour son expertise de clinicien et son ouverture sur l'analyse de données.

A mes maîtres de stage et notamment les Prs Jean-Luc Bosson et Patrice François pour avoir compris mon projet, pour leur soutien et pour m'avoir permis de me former en biologie.

A tout le staff de Virologie et notamment à Katia qui a permis de générer les données sur lesquelles sont basées ce travail.

A Om Kulkarni pour son intervention dans ce travail notamment sur la partie analyse de la coalescence.

A mes parents, sans lesquels tout cela n'aurait pas été possible.

Aux parents de Sylvie, pour leur soutien entier et constant.

A Sylvie, présente à mes côtés au quotidien qui comble mon existence et rend possible de nombreux projets.

A Maxime, Laetitia et Antoine, mes enfants, qui font la fierté de leur père.

Table des matières

TABLE DES FIGURES

I

Contexte et Problématiques

1

Introduction

1.1 Virus de l'Hépatite C (VHC) et résistance au traitement

Le Virus de l'Hépatite C (VHC) est un virus à ARN hépatotrope qui occasionne des dommages progressifs au foie pouvant être responsables de cirrhose hépatique et de carcinome hépatocellulaire. Globalement, 71.1 millions de personnes ont une infection chronique [Blach et al., 2017]. Dans les pays à forte prévalence du VHC, les facteurs de risque majeurs de cette infection virale liée au sang sont l'usage sans précautions de drogues injectées et les actes médicaux non stériles (infection iatrogène). La procédure diagnostique virologique inclut la recherche d'anticorps anti-VHC dans le sérum, la détection d'ARN VHC, la détermination du génotype et du sous-type viral et, plus récemment, l'évaluation de la présence de mutations de résistance (substitutions nucléotidiques dans le génome viral entraînant la résistance au traitement). Divers agents antiviraux directs (Direct-acting Antiviral Agents (DAAs)) ont été développés pour cibler 3 protéines impliquées dans les étapes cruciales du cycle de vie du VHC : la protéase NS3/4A, la protéine NS5A et la protéine NS5B, ARN polymérase ARN-dépendante. La combinaison de deux ou trois de ces DAAs peut éradiquer l'infection à VHC (définie par une réponse virologique (absence d'ARN) soutenue 12 semaines après traitement) dans plus de 90% des patients, en incluant les populations qui ont été difficiles à traiter par le passé. Avant l'arrivée des agents antiviraux directs après 2011, l'association Interféron pegylé - Ribavirine était le

standard de traitement. Parmi les patients suivis pour une hépatite C chronique, 40% de ceux présentant un génotype 1 et 80% de ceux présentant un génotype 2 avaient une réponse virologique soutenue sous ce traitement au prix d'effets indésirables importants [Manns et al., 2017].

Malgré cette proportion élevée de patients guéris par une combinaison de DAAs, la résistance aux agents antiviraux directs anti-VHC est une cause importante d'échec de traitement (jusqu'à 15% des patients selon le groupe du patient et son traitement [Buti et al., 2016]). La présence de variants viraux résistants aux inhibiteurs de NS5A en pré-traitement est associée à de plus faibles taux de guérison virologique dans certains groupes de patients, notamment ceux qui présentent un génotype 1a ou 3a, une cirrhose et/ou qui étaient non-répondeurs aux traitements de première génération (à base d'interféron pégylé). Les virus résistants aux DAAs dominent généralement au moment de l'échec virologique (le plus souvent à la rechute). Les virus résistants aux inhibiteurs de protéase NS3-4A disparaissent du sang périphérique en quelques mois alors que les virus résistants aux inhibiteurs de NS5A persistent pendant des années. Il existe des traitements de seconde ligne mais les stratégies de traitement de première ligne devraient être optimisées pour prévenir efficacement l'échec de traitement dû à une résistance aux anti-VHC [Pawlotsky, 2016] & [Pawlotsky, 2016, Table 2 - Liste des substitutions associées à une résistance connue (substitutions d'acides aminés signalées pour réduire la susceptibilité de différents génotypes ou sous-types du VHC aux DAAs)].

1.2 La notion de quasi-espèce

Chez un patient, le VHC présente une distribution de population virale dite «quasi-espèce» virale. Introduite en 1977 [Eigen et al., 1977], cette notion de quasi-espèce désigne une distribution bien définie de mutants générés par un processus de mutation/sélection. La sélection n'agit pas sur un seul mutant mais sur la quasi-espèce dans sa globalité. Les populations virales comme le VHC et le VIH (Virus de l'Immunodéficience Humaine)

ont été appelées quasi-espèces pour souligner leur hétérogénéité génétique étendue. La propriété de quasi-espèce de ces virus a formé la base d'un modèle qui permet d'expliquer le mécanisme de pathogenèse de ces virus chez l'humain [Nowak, 1992].

Pour Eigen, une quasi-espèce est définie par la distribution de mutants -ou «espèces» moléculaires étroitement liées- en équilibre générés par un processus spécifique de mutation/sélection décrivant la réplication erronée d'acides nucléiques (en l'occurrence ici l'ARN, cf. Figure 1.1).

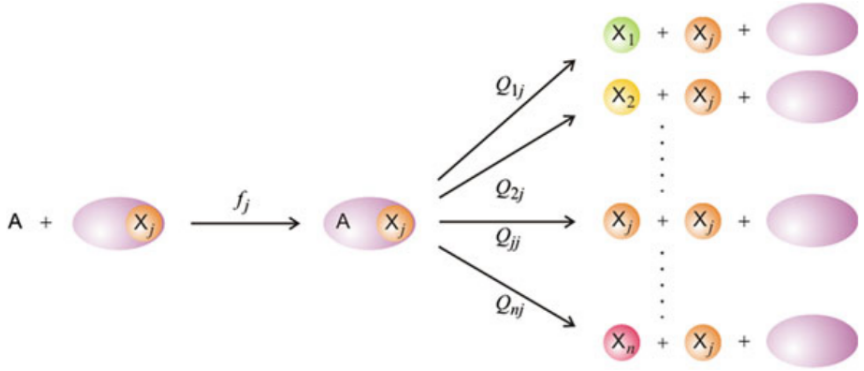


Figure 1.1 – Une vue moléculaire de la réplication et de la mutation. Le système de réplication **E** (*violet*), communément une simple molécule réplicase -comme dans la réaction en chaîne par polymérase (PCR)- ou un complexe multienzymatique se lie à la molécule d'ADN ou ARN matrice ($\mathbf{X}_j$, *orange*) pour former un complexe de réplication $\mathbf{E} \cdot \mathbf{X}_j$ et se réplique à un taux f_j. Lors du processus de copie de la matrice, des erreurs de réplication mènent à des mutations. La réaction mène à une copie correcte à la fréquence Q_{jj} et à un mutant $\mathbf{X}_k$ à la fréquence Q_{kj}. Communément, nous avons $Q_{jj} \gg Q_{kj}$ pour tout $k \neq j$. En d'autres termes, la réplication correcte l'emporte sur la formation de mutants. La stœchiométrie de la réplication nécessite $\sum_{i=1}^{n} Q_{ij} = 1$, étant donné que le produit doit être soit correct soit incorrect. La réaction est terminée par la dissociation complète du complexe de réplication. La somme de tous les monomères activés est désignée par **A**. Une conséquence du modèle est la factorisation des contributions du taux de réplication et de la fréquence des mutants : $w_{kj} = Q_{kj} \cdot f_j$ [Domingo et al., 2015]

Soit n différentes séquences d'acides nucléiques $I_1, I_2, ..., I_n$ qui peuvent servir de matrices pour la réplication. Chaque variant est caractérisé par une séquence nucléotidique spécifique. Cette séquence nucléotidique peut déterminer le taux de réplication d'un variant donné. Les taux de réplication des variants $I_1, I_2, ..., I_n$ peuvent être exprimés par $a_1, a_2, ..., a_n$. Ces quantités représentent les valeurs sélectives de chaque mutant (*fitness*). En l'absence de mutation, le variant avec le taux de réplication/*fitness* le plus élevé croîtra

le plus rapidement et constituera à terme de façon homogène toute la population.

Le résultat de la sélection dans ce monde sans erreur est donc une population homogène constituée par le variant se répliquant le plus rapidement. Cependant, la réplication n'est pas sans erreur. Par conséquent, il est nécessaire de définir la probabilité Q_{ij} que la réplication (erronée) de la matrice I_j ait pour résultat la production de la séquence I_i. Les quantités Q_{ij} pour $i = 1, 2, ..., n$ et $j = 1, 2, ..., n$ forment ce que l'on appelle la matrice de mutations.

Un système d'équations différentielles ordinaires décrit l'évolution temporelle de la population de ces séquences d'acides nucléiques. Le taux de croissance d'un variant spécifique, par exemple I_1, peut être écrit comme

$$dx_1/dt = a_1 Q_{11} x_1 + a_2 Q_{12} x_2 + ... + a_n Q_{1n} x_n$$

avec $x_1, x_2, ..., x_n$ les tailles de populations de variants $I_1, I_2, ..., I_n$

Les nouvelles particules de variant I_1 peuvent être formées par réplication sans erreur de I_1. Cela se produit au taux de réplication a_1 et à la probabilité Q_{11}. Le taux global est par conséquent donné par $a_1 Q_{11} x_1$. La réplication avec erreurs de tous les autres mutants $I_2, ..., I_n$ peut aussi conduire à de nouvelles particules I_1. Cela est représenté par les termes de croissance $a_2 Q_{12} x_2 + ... + a_n Q_{1n} x_n$ dans la précédente équation. De la même façon, nous pouvons écrire le taux de production de n'importe lequel des autres variants pour obtenir le système complet d'équations différentielles suivant

$$dx_i/dt = \sum_{j=1}^{n} a_j Q_{ij} x_j - x_i \overline{a}$$

$$\text{avec} \begin{cases} i = 1, ..., n \\ \overline{a} = \sum_{k=1}^{n} a_k x_k \Big/ \sum_{k=1}^{n} x_k \end{cases}$$

$\overline{a}$ est le taux de réplication/*fitness* moyen de la population et le terme $-x_i\overline{a}$ est là pour conserver une taille de population $c = \sum_{i=1}^{n} x_i$ constante ($dc/dt = \sum_{i=1}^{n} dx_i/dt = 0$).

Dans ce contexte, la population ne va plus seulement être constituée par la séquence croissant le plus rapidement mais par l'ensemble global de mutants présentant des taux de

réplication différents. Cet ensemble de mutants est la quasi-espèce.

On peut réécrire le système d'équations différentielles précédent comme suit

$d\vec{x}/dt = W\vec{x} - f(\vec{x})\vec{x}$

avec $\vec{x}$ le vecteur contenant les densités de population des séquences individuelles

$\vec{x} = (x_1, x_2, ..., x_n)$

La matrice W contient les taux de réplication et les probabilités de mutation

$$W = \begin{pmatrix} a_1Q_{11} & a_2Q_{12} & \cdots & a_nQ_{1n} \\ a_1Q_{21} & a_2Q_{22} & \cdots & a_nQ_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_1Q_{n1} & a_2Q_{n2} & \cdots & a_nQ_{nn} \end{pmatrix}$$

et $f(\vec{x}) = \sum_{k=1}^{n} a_k x_k \Big/ \sum_{k=1}^{n} x_k$

En négligeant le terme de conservation de taille de population, la distribution de mutants en équilibre peut ainsi être déduite en calculant les valeurs et vecteurs propres de W, problème classique en algèbre linéaire

$W\vec{x} = \lambda\vec{x}$

La quasi-espèce peut ainsi être définie mathématiquement par le vecteur propre dominant $\vec{x} = (x_1, x_2, ..., x_n)$ correspondant à la plus grande valeur propre λ_{max} de la matrice W. Ce vecteur propre $\vec{x}$ (conservation de la «direction géométrique» à l'équilibre) décrit la structure de population exacte de la quasi-espèce ; chaque mutant I_i est présent dans la quasi-espèce avec une fréquence x_i. La plus grande valeur propre est exactement le taux de réplication moyen de la quasi-espèce

$\lambda_{max} = \sum a_i x_i$

avec $\sum x_i = 1$

La fréquence d'un variant donné dans la quasi-espèce ne dépend pas de son taux de réplication seulement mais aussi de la probabilité qu'il soit produit par réplication avec erreurs des autres matrices et leurs fréquences dans la distribution de quasi-espèce. Par

conséquent, la séquence individuelle I_i avec son taux de réplication a_i n'est plus la cible de la sélection. La quasi-espèce elle-même est la cible de la sélection dans un processus de mutation/sélection. Malgré ce que l'on considère habituellement, une quasi-espèce est en mesure de guider les mutations. Bien que l'acte de mutation reste intrinsèquement stochastique, la sélection opère sur la structure de la quasi-espèce globale laquelle est adaptée au «*fitness landscape*» (relation entre les génotypes et les taux de réplication). Par conséquent, l'évolution peut être guidée vers les pics de cette *fitness landscape* simplement par le fait que les mutants à haut taux de réplication auront plus de descendants que les mutants à bas taux de réplication.

1.2.0.1 Seuil d'erreur de réplication

Une autre notion importante dans la théorie des quasi-espèces est le seuil d'erreur de réplication. Si la réplication était sans erreur, aucun mutant ne serait produit et aucune évolution ne serait possible. Toute évolution serait aussi compromise si le taux d'erreur de la réplication était trop élevé (seules quelques mutations mènent à une amélioration de l'adaptation, la plupart menant à une détérioration). La théorie des quasi-espèces nous permet de quantifier la précision minimale de réplication qui maintienne l'adaptation (cf. Figure 1.3).

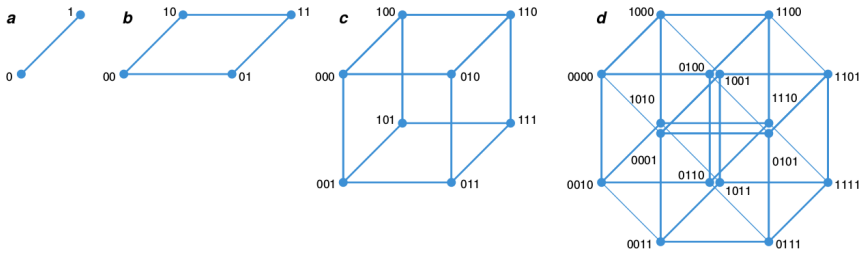


Figure 1.2 – Exemple d'espace de séquences - chaque point représente une séquence unique et le degré de séparation entre points reflète leur degré de dissimilarité - dans cet exemple, nous considérons de petites séquences de valeurs binaires (0 ou 1) de longueur 1 (a), 2 (b), 3 (c) et 4 (d, hypercube de 4 dimensions). Les espaces de séquences pour les génomes viraux sont bien plus complexes car ils impliquent des séquences de milliers de positions chacune pouvant accueillir 1 sur 4 nucléotides différentes. [Eigen, 1993]

Nous considérons pour cela un ensemble d'approximations afin d'obtenir les expressions

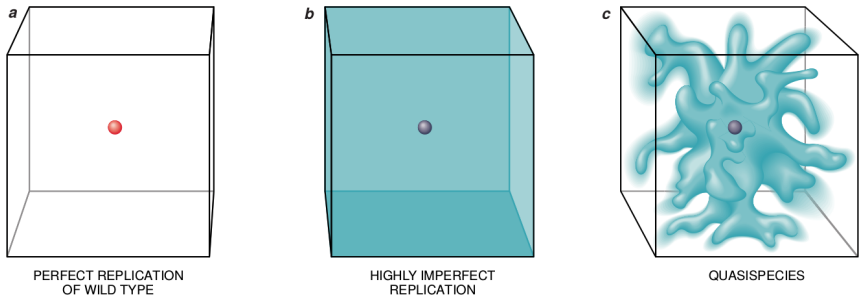


Figure 1.3 – Impact du taux d'erreur du processus de réplication sur la dynamique de population d'un virus - Représentation d'une population virale dans l'espace de séquences (cf. Figure 1.2) - réplication sans erreur (a), réplication à taux d'erreurs élevé (b), réplication à taux d'erreurs intermédiaire (c) [Eigen, 1993]

d'une solution analytique suffisamment précise pour les applications les plus courantes [Domingo et al., 2015]. La première approximation consiste à ne considérer que la génération des mutants à partir de la souche sauvage ou la séquence la plus fréquente («séquence maître») et à négliger toute génération retour du mutant au sauvage/à la séquence la plus fréquente ainsi que le passage d'un mutant à un autre («*zero mutational backflow*»). La seconde approximation consiste à relâcher la contrainte sur la taille de population constante et à ne pas compenser toute modification du taux de mutation. On désignera cette approche avec ces approximations par «approche phénoménologique».

Le taux de mutation p pour chaque nucléotide est considéré indépendant de la position dans la séquence («taux d'erreur uniforme»). La probabilité de réplication sans erreur est la même pour toute séquence X_k et est exprimée par

$$Q_{kk} = Q = (1-p)^l$$

pour tout $k = 1, ..., n$; avec l la longueur de la séquence polynucléotidique

Ainsi, pour une réplication sans erreur ($p = 0$), toutes les copies sont correctes ($Q = 1$). Dans le cas contraire ($p > 0$), la fraction des réplicats corrects diminue de façon monotone avec l'augmentation du taux de mutation p (cf. Figure 1.4).

La fréquence des copies avec erreurs $\mathbf{X}_j$ s'exprime par

$$Q_{jk} = (1-p)^{l-d_{jk}} p^{d_{jk}} = Q\varepsilon^{d_{jk}}$$

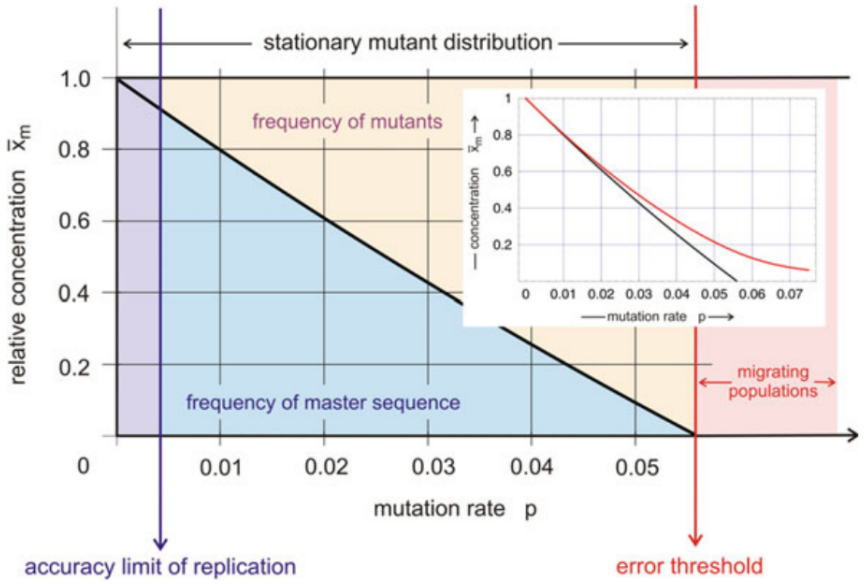


Figure 1.4 – Le seuil d'erreur. La fréquence stationnaire de la séquence maître $\mathbf{X}_m$ est fonction du taux de mutation local p. Dans le cadre de l'approximation «*zero mutational backflow*», la fonction $\overline{x_m}(p)$ est pratiquement linéaire dans l'exemple particulier montré ici. Dans l'insert, l'approximation (*noir*) est montrée avec la solution exacte (*rouge*). Le taux d'erreur p a deux limitations naturelles : (i) La limite de précision physique du processus de réplication prévoit une limite inférieure pour le taux de mutation, et (ii) le seuil d'erreur définit une précision de réplication minimale requise pour supporter la transmission de l'information génétique et fixe une limite supérieure pour le taux de mutation. [Domingo et al., 2015]

$$\text{avec} \begin{cases} \varepsilon = p/1 - p \\ d_{jk} \text{ la distance de Hamming entre 2 séquences } \mathbf{X}_j \text{ et } \mathbf{X}_k \end{cases}$$

La distance de Hamming est le nombre (minimum) de positions auxquelles les 2 séquences diffèrent. Avec ces approximations et notations, il est simple de calculer la concentration stationnaire de la séquence maître $\mathbf{X}_m$ que nous représentons par $\hat{x}_m^{(0)}$:

$$\hat{x}_m^{(0)} \propto Q - \sigma_m^{-1}$$

$$\text{avec} \begin{cases} \sigma_m = \dfrac{f_m}{\overline{f}_{-m}} \\ \overline{f}_{-m} = \dfrac{\sum_{i=1, i\neq m}^{n} f_i \hat{x}_i^{(0)}}{\hat{c} - \hat{x}_m^{(0)}} \\ \hat{c} = \sum_{i=1}^{n} \hat{x}_i^{(0)} \end{cases}$$

Nous précisons des concentrations stationnaires par un chapeau et l'approximation *zero mutational backflow* par un exposant «(0)». Le *fitness* moyen des mutants est désigné par

$\overline{f}_{-m}$ et par conséquent la supériorité du maître vis-à-vis des mutants en terme de *fitness* par le rapport σ_m. De la même manière, nous obtenons pour les mutants $\mathbf{X}_j$:

$$\hat{x}_j^{(0)} = \varepsilon^{d_{jm}} \frac{f_m}{f_m - f_j} \hat{x}_m^{(0)} \propto Q - \sigma_m^{-1}$$

En substance, la fréquence à laquelle un mutant est présent dans la quasi-espèce dépend de 2 quantités :

— la distance de Hamming d_{jm} entre la séquence $\mathbf{X}_j$ et le maître $\mathbf{X}_m$ (plus la séquence est proche du maître, plus sa part dans la distribution stationnaire est élevée)

— la différence de *fitness* entre $\mathbf{X}_m$ et $\mathbf{X}_j$ (plus le *fitness* du mutant est élevé, plus sa fréquence dans la quasi-espèce est élevée)

En conséquence, une quasi-espèce n'est pas une collection arbitraire de variants mais une distribution très ordonnée avec une séquence maître au centre entourée d'un nuage de mutants dans l'espace de séquences.

Dans le cadre de l'approche phénoménologique, la concentration stationnaire de la séquence maître ainsi que celles des mutants présentent un facteur $(Q - \sigma_m^{-1})$ qui exprime la dépendance des concentrations sur le taux de mutation p. Cette dépendance est levée si la condition $(Q = \sigma_m^{-1})$ est remplie. Le taux de mutation p_{max} correspondant est calculé simplement :

$$Q = (1 - p_{max})^l = \sigma_m^{-1} \quad \text{ce qui donne} \quad p_{max} \approx \frac{ln\,\sigma_m}{l} \text{ ou } l_{max} \approx \frac{ln\,\sigma_m}{p}$$

Effectivement, une quasi-espèce conventionnelle existe seulement dans l'intervalle

$$0 \leq p < p_{max}$$

A des taux de mutation plus élevés que la valeur seuil, nous n'obtenons aucune information sur la nature de la solution à long terme du système de réplication-mutation dans le cadre de l'approche phénoménologique. Le phénomène du taux de mutation maximum tel que décrit par l'équation précédente a été appelé le «seuil d'erreur» : afin d'assurer une stabilité au cours de l'évolution de l'information génétique stockée dans les séquences d'acide nucléique, l'inexactitude de la réplication ne doit pas excéder une certaine valeur critique définie par la longueur de la séquence l et la supériorité de la séquence maître

σ_m. Alternativement, pour une précision de la réplication donnée, le seuil d'erreur définit une certaine longueur de chaîne polynucléotidique l_{max} qui ne peut pas être dépassée sans mettre en péril l'héritage de l'information génétique.

Il s'agit d'ailleurs du mode d'action supposé mais non démontré de la ribavirine sur les virus à ARN, utilisée en particulier dans le traitement du VHC : elle aurait une activité mutagénique en contribuant notamment à une augmentation du taux de mutation [Crotty et al., 2001, Cuevas et al., 2009, Feigelstock et al., 2011, Dietz et al., 2013, Ortega-Prieto et al., 2013].

1.3 EVOLUTION DES TECHNIQUES DE SÉQUENÇAGE

Pour évaluer cette diversité de population modélisée par une distribution en quasi-espèce, nous avons besoin de techniques de mesure adaptées à cette problématique. L'information génétique est stockée dans les séquences nucléotidiques de l'ADN ou l'ARN d'un organisme. Le processus de détermination de l'ordre correct des nucléotides dans un fragment donné (gène, groupe de gènes, chromosome ou génome complet) est appelé séquençage des nucléotides. Il existe différentes méthodes de séquençage. Le séquençage première génération développé par Frederick Sanger en 1977 a été largement utilisé jusqu'à l'arrivée du séquençage dernière génération (séquençage haut débit, «*High-Throughput Sequencing*» (HTS) ou «*Next-Generation Sequencing*» (NGS)) à partir de 2005.

Le séquençage Sanger a pour principe la terminaison de synthèse de brin par incorporation sélective de didésoxyribonucléotides (ddNTPs, nucléotides modifiés) terminateurs de chaînes tels que ddGTP, ddCTP, ddATP et ddTTP par l'ADN polymérase durant la réplication de l'ADN. Cette méthode est ainsi appelée méthode de séquençage par terminaison de chaîne. Les nucléotides normaux ont des groupements 3'OH pour la formation d'une liaison phosphodiester entre nucléotides adjacents pour continuer la formation du brin. Au contraire, les ddNTPs ne présentent pas ce groupement 3'OH et sont donc incapables de former des liaisons phosphodiester entre nucléotides. Par conséquent, l'élongation de

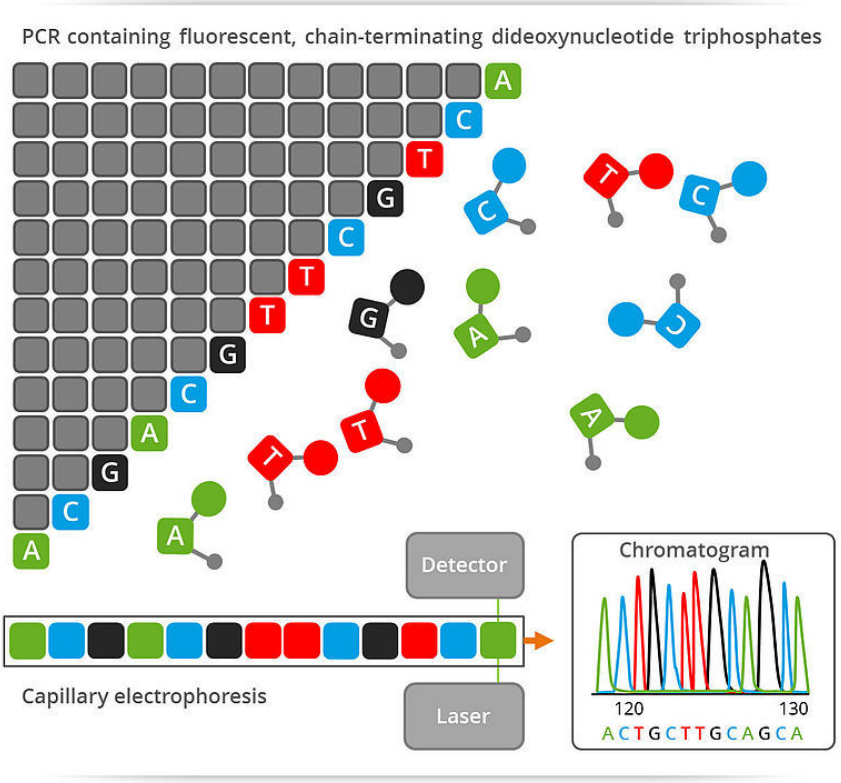


FIGURE 1.5 – Illustration du séquençage Sanger (gatc-biotech.com, 2017)

la chaîne est arrêtée.

Dans cette méthode, l'ADN simple brin à séquencer sert de brin modèle pour la synthèse d'ADN in vitro. Les autres pré-requis sont des amorces oligonucléotidiques (sélection du fragment d'ADN à séquencer par *Polymerase Chain Reaction* (PCR)), des désoxyribonucléotides précurseurs de la nouvelle chaîne (dNTPs) et l'enzyme ADN polymérase nécessaire à la réplication de l'ADN. La polymérisation des nouveaux brins d'ADN se fait avec les dNTPs. L'incorporation, peu fréquente car en faible concentration, d'un ddNTP, marqué avec un fluorochrome propre selon la base concernée (A, T, G ou C), dans le brin d'ADN empêche la polymérisation de se poursuivre. De nombreux fragments d'ADN sont ainsi obtenus, de longueurs différentes mais finissant tous par un didésoxynucléotide. La séquence du fragment étudié est déduite de la succession des différentes fluorescences après séparation des fragments d'ADN par électrophorèse et lecture de la fluorescence par un détecteur après excitation par un faisceau laser (cf. Figure 1.5).

Le séquençage haut débit (HTS) quant à lui représente une famille de technologies de séquençage plus moderne. Elles sont, entre autres, représentées par 454 (pyroséquençage), SOLiD (séquençage par ligation), Ion Proton (séquençage par synthèse avec semi-conducteur), Illumina (séquençage par synthèse), Oxford Nanopore (séquençage par les nanopores) ou Pacific Biosciences (séquençage d'une seule molécule en temps réel). Elles sont capables de produire des millions de séquences en une analyse ou *run*. Pour cela, elles utilisent des approches massivement parallèles permettant de séquencer des centaines de milliers de fragments simultanément. Par conséquent, le HTS, tout en intégrant ses limites, permet d'échantillonner des populations de brins d'ADN ou ARN. Il se révèle alors un outil très précieux pour analyser une distribution de population virale telle qu'une quasi-espèce.

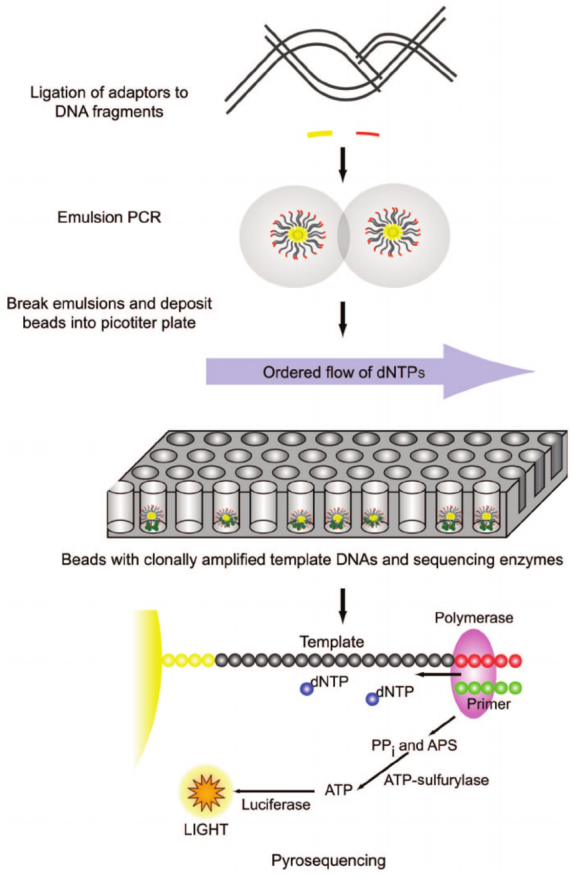


FIGURE 1.6 – Illustration du séquençage 454 [Voelkerding et al., 2009]

Dans ce travail, la technologie 454 (pyroséquençage) a été utilisée pour analyser notre distribution de population virale en quasi-espèce (cf. Figure 1.6). Cette technologie peut

produire de longues séquences de 700 à 1000 bases ce qui est proche de la longueur des séquences produites par séquençage Sanger (au maximum ~1 kb). Ces séquences sont produites de façon massivement parallèle par la lecture des signaux optiques générés par l'ajout des bases [Perry, 2012].

L'ADN ou ARN est fragmenté en des séquences plus petites jusqu'à 1 kb ou amplifié sous forme d'un fragment de la taille souhaitée. Des adaptateurs génériques sont ajoutés aux extrémités et ces derniers sont hybridés à des billes, un fragment d'ADN par bille. Les fragments sont ensuite amplifiés de façon clonale par PCR en utilisant des amorces spécifiques aux adaptateurs.

Chaque bille est ensuite placée dans un puits libre d'une plaque. Par conséquent, chaque puits ne contiendra qu'une seule bille, couverte d'une multitude de copies identiques d'un produit de PCR. Les puits contiennent aussi l'ADN polymérase et les buffers de séquençage (solutions de lavage et de dilution).

La plaque est recouverte avec l'une des 4 espèces de dNTPs. Là où ce nucléotide poursuit la séquence, il est ajouté à la séquence lue (*read*). Si cette base se répète, il en sera ajouté plus. Par exemple, si nous recouvrons la plaque avec des bases Guanine et la suite dans une séquence est G, un G sera ajouté. Cependant, dans le cas où la suite dans la séquence est GGGG, alors 4 Gs seront ajoutés.

L'ajout de chaque nucléotide génère un signal lumineux suite à la libération d'un pyrophosphate (PPi). Les localisations des signaux sont détectées et utilisées pour déterminer à quelles billes les nucléotides sont ajoutés.

Le mix de dNTP est lavé avant l'introduction du mix de dNTP suivant. Le processus est ainsi répété, bouclant circulairement sur les 4 dNTPs.

Ce type de séquençage génère des graphes pour chaque séquence lue (cf. Figure 1.7), montrant la densité du signal pour chaque lavage nucléotidique. La séquence peut ensuite être déterminée informatiquement à partir de la densité du signal dans chaque lavage.

Toutes les lectures de séquences que nous obtenons du 454 seront de longueurs différentes

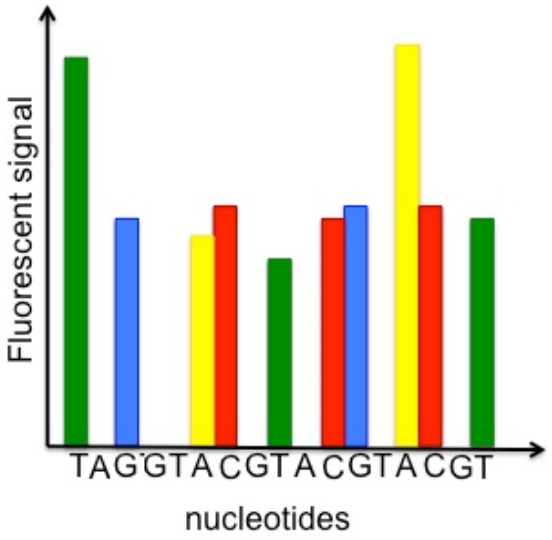


Figure 1.7 – Séquençage 454 - graphe pour la séquence lue «TTGACTCGAACT» [Perry, 2012]

car des nombres différents de bases seront ajoutés à chaque cycle.

En conclusion, seul le séquençage haut débit peut nous permettre d'analyser la distribution de population en quasi-espèce étant donné qu'il permet de générer une séquence pour tout brin d'ADN et, de manière massivement parallèle, de lire au moins des dizaines de millions de bases lors d'un seul *run* sur une seule plaque (cf. Figure 1.8). Hormis les techniques longues et fastidieuses utilisant le clonage, le séquençage Sanger n'a pas cette capacité d'échantillonnage, raison pour laquelle cette technologie est mise de côté dans ce domaine de recherche. Il est à noter cependant que les séquences individuelles sont moins précises dans le cas du HTS : le taux d'erreur propre à chaque technologie de HTS est alors à considérer dans les analyses (par exemple 0.0543 substitutions pour 100 bases pour la technologie 454 utilisée ici (GS Junior) [Jünemann et al., 2013]).

Dans ce travail, nous avons uniquement procédé à un séquençage haut débit d'amplicons, soit au séquençage ciblé de régions d'ARN amplifiées par RT-PCR.

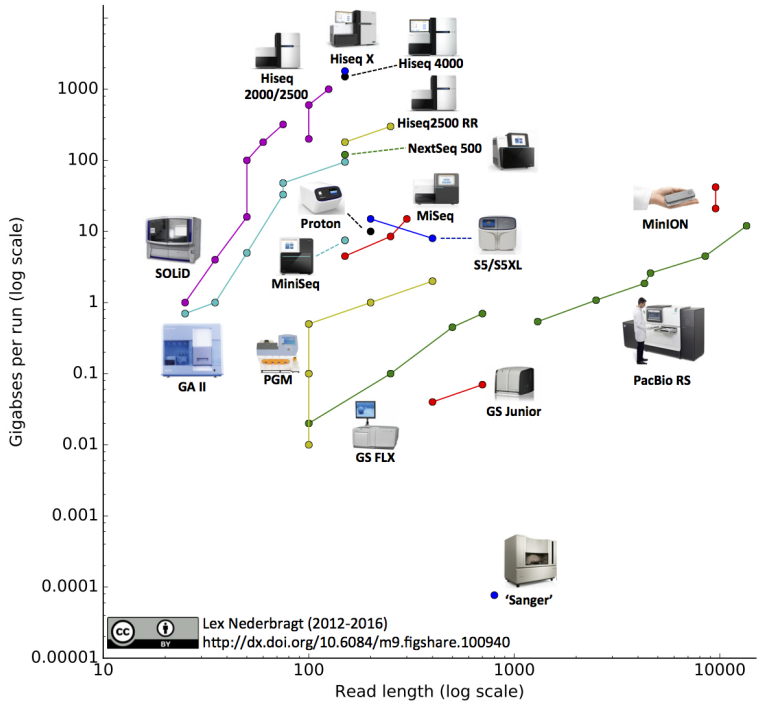


FIGURE 1.8 – Développements du séquençage haut débit : instruments, longueurs de *read*, débit [Nederbragt, 2016]

1.4 VHC ET NGS

Ces dernières années, les technologies de HTS ont permis d'étudier l'évolution de populations virales au cours d'une infection en autorisant l'acquisition rapide de milliers de courtes séquences d'ARN à partir d'échantillons temporels multiples [Nelson et al., 2015]. Effectivement, comme évoqué précédemment, le séquençage haut débit est devenu un outil essentiel dans la description de l'hétérogénéité génétique et dans la compréhension de la dynamique de l'évolution intra-patient des espèces virales [Beerenwinkel et al., 2012].

Plus spécifiquement, le HTS a été appliqué à la description de la diversité intra-hôte du VHC [Lauck et al., 2012, Li et al., 2011, Ramachandran et al., 2011], à la détection de goulots d'étranglement génétiques de transmission dans l'infection [Wang et al., 2010], à la prédiction de réponse à la trithérapie chez les patients infectés chroniques au VHC

et à l'évaluation de variants résistants aux inhibiteurs de protéase [Larrat et al., 2014], à l'évaluation de variants résistants aux inhibiteurs de polymérase [Donaldson et al., 2014], à l'étude de la réponse du VHC à la sélection imposée par les traitements anti-viraux [Ahmed et al., 2015] et à la discrimination entre un échappement au traitement et une réinfection [Abdelrahman et al., 2014].

Avant d'aborder le cœur de ce travail, une revue des différentes méthodes d'analyse utilisées est effectuée.

1.5 Méthodes d'analyse de données

1.5.1 Recherche dans les bases de données : BLAST

Le programme BLAST («*Basic Local Alignment Search Tool*») est un algorithme de recherche de similitudes locales (score d'alignement avec récompense/pénalité sur *match*/*mismatch* de base et coût d'ouverture/extension dans les cas d'insertions-délétions (indels) de bases) [Deléage et al., 2013].

La première étape consiste à établir la liste de mots exacts de longueur fixée ($W = 11$ pour les acides nucléiques) de la séquence requête. La deuxième étape consiste à établir la liste exhaustive des mots identifiés dans la base. Ensuite, pour chaque mot trouvé, l'algorithme étend progressivement de part et d'autre tant que le score sur le segment est supérieur à une valeur seuil.

Au final, à chaque séquence candidate de la base est attribuée une E-value $E()$ qui représente la probabilité d'obtenir par chance un score supérieur à la valeur seuil. Dans l'hypothèse d'une distribution des scores selon une loi de Poisson, on peut exprimer

$$\text{p-value} = 1 - e^{-E()}$$

Nous recherchons donc ici pour une séquence requête la séquence candidate de la base qui minimise la E-value correspondant à un score statistiquement significatif.

BLAST est accessible par exemple sur le site de l'EBI (Institut Européen de Bioinformatique, NCBI BLAST) et plus particulièrement en ce qui nous concerne sur le site du projet de base de données du VHC (HCV BLAST).

1.5.2 Estimation de la diversité nucléotidique

Estimer la diversité nucléotidique consiste à mesurer le degré de polymorphisme génétique au sein d'une population de séquences. C'est une mesure de variation génétique [Nei et al., 1979].

Pour cela, un alignement multiple des séquences sur lesquelles on souhaite estimer la diversité nucléotidique doit être réalisé. Effectivement, l'alignement multiple des séquences a classiquement pour but d'identifier les résidus (nucléotides ou acides aminés) essentiels qui ont été préservés au cours de l'évolution. De nombreuses approches ont été envisagées pour les alignements multiples dont la méthode d'alignement progressif. Brièvement, il s'agit de décomposer la question de l'alignement simultané de N séquences en $N(N-1)/2$ alignements (nombre d'alignements de 2 séquences d'un ensemble à N séquences) de toutes les paires possibles puis de grouper les paires en partant tout d'abord des paires les plus proches pour ensuite diverger. D'autres méthodes plus récentes utilisent un alignement itératif, dans lequel l'alignement est construit de manière répétée en optimisant progressivement le score global. Contrairement à l'alignement progressif (algorithme de type glouton («*greedy*»)), ces méthodes ont l'avantage de ne pas dégrader progressivement l'alignement en incorporant de plus en plus d'indels du fait de la réévaluation des alignements déjà effectués pendant l'agrégation.

Le logiciel MAFFT permet de réaliser des alignements multiples de séquences.

Cet alignement permet l'estimation de la diversité nucléotidique suivante

$$\pi = \sum_{ij} x_i x_j \pi_{ij} = 2 * \sum_{i=2}^{n} \sum_{j=1}^{i-1} x_i x_j \pi_{ij}$$

avec
$$\begin{cases} x_i \text{ et } x_j \text{ les fréquences respectives des séquences } i \text{ et } j \\ \pi_{ij} \text{ le nombre de différences nucléotidiques par site nucléotidique entre les séquences } i \text{ et } j \\ n \text{ le nombre de séquences dans l'échantillon} \end{cases}$$

Cela revient donc à mesurer le nombre moyen de différences nucléotidiques par site entre 2 séquences dans toutes les paires possibles de l'échantillon.

Le package pegas sous R permet d'estimer la diversité nucléotidique.

1.5.3 Analyse en Composantes Principales (ACP)

Cette approche permet de définir un certain nombre de variables représentatives qui collectivement expliquent la majeure partie de la variabilité dans le jeu de données initial [James et al., 2013]. Les directions des composantes principales sont les directions dans l'espace des caractéristiques selon lesquelles les données originales sont très variables. Ces directions définissent aussi les lignes et les sous-espaces qui sont aussi proches que possible du nuage de données.

L'analyse en composantes principales (ACP) consiste à calculer les composantes principales et à utiliser ces composantes pour comprendre les données. Il s'agit d'une approche non supervisée car elle ne fait intervenir que des jeux de caractéristiques $X_1, X_2, ..., X_p$ sans réponse associée à expliquer ou à prédire. L'ACP est utilisée ici en tant qu'outil de visualisation de données dans le cadre d'une analyse exploratoire des données.

L'ACP trouve une représentation à basse dimension du jeu de données qui contient autant de variation que possible. L'idée est que chacune des n observations évolue dans un espace de dimension p mais toutes ces dimensions n'ont pas la même importance. L'ACP cherche un petit nombre de dimensions pertinentes définies par la quantité de variabilité que les observations ont selon chacune d'elles. Chaque dimension ou composante principale définie par l'ACP est une combinaison linéaire des p caractéristiques.

La première composante principale d'un jeu de caractéristiques $X_1, X_2, ..., X_p$ est la combinaison linéaire normalisée des caractéristiques

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p$$

qui a la plus grande variance. Par normalisée, nous signifions $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Les éléments $\phi_{11}, ..., \phi_{p1}$ sont les poids de la première composante principale ; ensemble, les poids définissent le vecteur de poids de la composante principale $\phi_1 = (\phi_{11}\phi_{21}...\phi_{p1})^T$. La contrainte sur les poids (somme des carrés égale à 1) permet de prévenir une variance arbitrairement élevée.

Soit $\mathbf{X}$ un jeu de données $n \times p$, centré (intérêt dans la variance uniquement : les moyennes

des colonnes de $\mathbf{X}$ (variables) sont nulles). Nous regardons ensuite la combinaison linéaire des valeurs des caractéristiques de l'échantillon de la forme

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + ... + \phi_{p1}x_{ip}$$

qui a la variance d'échantillon la plus grande, sachant que $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Autrement dit, sachant le jeu de données centré, le vecteur de poids de la première composante principale résout le problème d'optimisation

$$\underset{\phi_{11},...,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1}x_{ij} \right)^2 \right\} \text{ sachant } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Effectivement, à partir de l'équation précédente, nous pouvons écrire l'objectif d'optimisation en fonction de $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$. Etant donné que $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ (jeu de données centré), la moyenne des $z_{11}, ..., z_{n1}$ sera aussi nulle. Par conséquent, l'objectif d'optimisation que nous maximisons est simplement la variance de l'échantillon des n valeurs de z_{i1}. Nous appelons *scores* de la première composante principale les valeurs $z_{11}, ..., z_{n1}$. Le problème d'optimisation peut être résolu via une décomposition en éléments propres, une technique standard d'algèbre linéaire.

Il existe une interprétation géométrique pertinente pour la première composante principale. Le vecteur de poids ϕ_1 avec les éléments $\phi_{11}, \phi_{21}, ..., \phi_{p1}$ définit une direction dans l'espace des caractéristiques selon laquelle les données varient le plus. Si nous projetons les n points de données $x_1, ..., x_n$ sur cette direction, les valeurs projetées sont elles-mêmes les *scores* de la composante principale $z_{11}, ..., z_{n1}$ (cf. Figure 1.9).

Après avoir déterminé la première composante principale des caractéristiques Z_1, nous pouvons chercher la seconde composante principale Z_2. La seconde composante principale est la combinaison linéaire de $X_1, ..., X_p$ qui présente la variance maximale parmi toutes les combinaisons linéaires non corrélées avec Z_1. Les *scores* de cette seconde composante principale $z_{12}, z_{22}, ..., z_{n2}$ prennent la forme suivante

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + ... + \phi_{p2}x_{ip}$$

avec ϕ_2 le vecteur de poids de la seconde composante principale, avec les éléments $\phi_{12}, \phi_{22}, ..., \phi_{p2}$. Contraindre Z_2 d'être non corrélé avec Z_1 est équivalent à contraindre

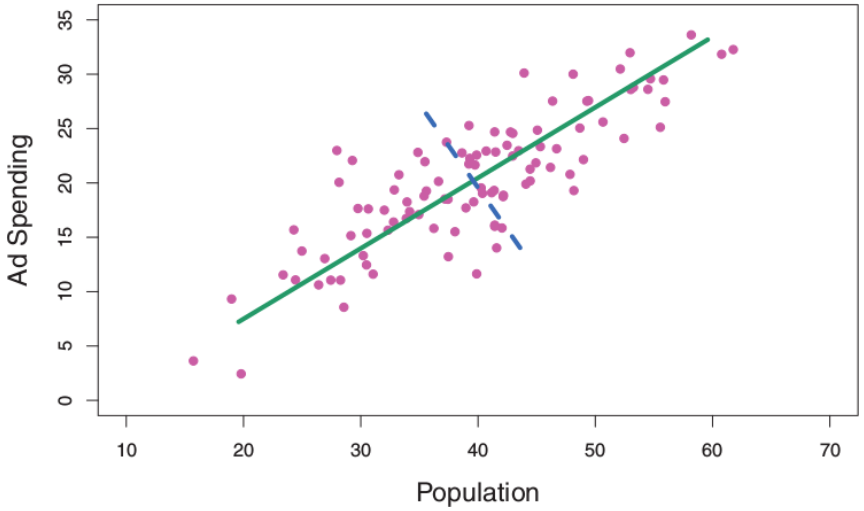


FIGURE 1.9 – La taille de population et la dépense en publicité pour 100 villes différentes sont représentées en points violets. La ligne continue verte représente la première composante principale et la ligne pointillée bleue la seconde composante principale [James et al., 2013]

la direction de ϕ_2 d'être orthogonale (perpendiculaire) à la direction de ϕ_1. Pour trouver ϕ_2, nous devons résoudre le même problème d'optimisation que précédemment avec ϕ_2 remplaçant ϕ_1 et avec la contrainte supplémentaire que ϕ_2 est orthogonale à ϕ_1.

Les composantes principales calculées, nous pouvons les représenter sous forme de graphes l'une en fonction de l'autre pour produire des visualisations à basse dimension des données. Par exemple, nous pouvons représenter sous forme de graphes le vecteur de *scores* Z_1 en fonction de Z_2, Z_1 en fonction de Z_3, Z_2 en fonction de Z_3 et ainsi de suite. Géométriquement, cela revient à projeter les données originales dans l'espace défini par ϕ_1, ϕ_2 et ϕ_3 et représenter sous forme de graphe les points projetés.

Le programme prcomp du package stats de R permet d'effectuer une analyse en composantes principales.

1.5.4 PHYLOGÉNIE MOLÉCULAIRE

La phylogénie moléculaire permet de reconstruire l'arbre phylogénétique d'un ensemble de séquences nucléotidiques d'intérêt. Les séquences étudiées doivent être toutes homologues, c'est-à-dire dériver d'une séquence ancestrale commune. Elles doivent aussi

avoir été alignées au préalable (cf. § sur l'alignement multiple des séquences dans 1.5.2). L'arbre phylogénétique est constitué de nœuds, qui représentent des organismes ancestraux partagés, de feuilles, qui sont les organismes étudiés, et de branches reliant les nœuds entre eux et aux feuilles qui représentent les lignées évolutives d'ancêtre à descendant.

Les branches des arbres représentent l'évolution moléculaire qui s'est produite entre un organisme ancestral et un de ses descendants. Une valeur précise est attribuée le plus souvent à la longueur de chaque branche. Cette valeur, toujours exprimée en substitutions/site ou colonne de l'alignement, est une quantité d'évolution moléculaire. Elle est égale au nombre moyen de remplacements d'un résidu (nucléotide ou acide aminé) par un autre qui se sont produits à chaque site de la molécule étudiée le long de cette branche.

1.5.4.1 Modèle markovien de l'évolution moléculaire

La plupart des méthodes utilisées en phylogénie moléculaire font l'hypothèse que le processus de substitution des bases est déterminé par un modèle probabiliste qui fixe à tout instant la probabilité par unité de temps qu'une substitution $x \rightarrow y$ se soit produite en un site donné et dans une lignée donnée. Les modèles les plus courants sont homogènes : ils supposent que ces probabilités sont les mêmes pour toutes les lignées et à tout instant. Les modèles les plus simples supposent aussi que ces probabilités sont les mêmes pour tous les sites. Mais il est plus courant d'utiliser une distribution qui modélise la variation des probabilités de substitution entre sites. Le processus évolutif est un processus aléatoire qui se produit à un instant donné dans une population de génomes d'une même espèce, sans que les organismes ancestraux à cette espèce puissent interférer avec le processus du moment. En conséquence, les modèles probabilistes de Markov (la prédiction du futur à partir du présent n'est pas rendue plus précise par des éléments du passé, notion d'absence de «mémoire») s'appliquent parfaitement à la modélisation du processus évolutif au niveau moléculaire. Un modèle évolutif probabiliste markovien s'exprime tout entier par une matrice carrée Q, dite matrice des taux instantanés de substitution, dont les lignes et les colonnes représentent les états possibles (4 pour l'ADN) et les termes non diagonaux

sont les taux de substitution par unité de temps.

$$Q = \begin{pmatrix} -\lambda_A & m_{AT} & m_{AC} & m_{AG} \\ m_{TA} & -\lambda_T & m_{TC} & m_{TG} \\ m_{CA} & m_{CT} & -\lambda_C & m_{CG} \\ m_{GA} & m_{GT} & m_{GC} & -\lambda_G \end{pmatrix}$$

avec $\begin{cases} m_{ij} \text{ le taux de substitution } i \rightarrow j \text{ par site par unité de temps} \\ \lambda_i = \sum_{j \neq i} m_{ij} \text{ le taux par unité de temps avec lequel la base } i \text{ est modifiée} \end{cases}$

La variation de la composition en bases d'un site d'une séquence au cours du temps s'exprime donc

$$F(t + dt) = F(t) + Q^T F(t) dt$$

avec $\begin{cases} F(t) = (A(t)\, T(t)\, C(t)\, G(t))^T \\ X(t) \text{ la fréquence de la base } X \text{ au temps } t \end{cases}$

D'où la dérivation de F par rapport au temps

$$dF(t)/dt = F^{'}(t) = Q^T F(t)$$

Ce qui nous permet de définir les fréquences d'équilibre de Q : fréquences des bases qui restent inchangées par le processus markovien (d'où $F^{'}(t) = Q^T F(t) = 0$). N'importe quelle fréquence initiale $F(0)$ tend vers la fréquence d'équilibre F_{eq} quand t tend vers $+\infty$ si le processus de substitution (la matrice Q) est constant. Biologiquement, la fréquence d'équilibre est une composition en bases caractéristique d'une séquence qui reste inchangée malgré la modification continue de la séquence par des substitutions.

Un processus évolutif qui applique une matrice de taux Q à des séquences est dit stationnaire si les fréquences des résidus de la séquence ancestrale sont les fréquences d'équilibre de Q. Ces fréquences resteront inchangées au cours de l'évolution.

Nous pouvons maintenant définir la matrice de transition P donnant les probabilités conditionnelles de changement au cours du temps :

$$P_{ij}(t) = proba(j \text{ à l'instant } t \,|\, i \text{ à l'instant } 0)$$

Les quantités $P_{ij}(t)$ donnent la probabilité de tout état à un instant futur à partir de tout état initial, en intégrant tous les états intermédiaires possibles. P s'exprime en fonction de Q

$$P(t) = e^{Qt}$$

Nous pouvons aussi introduire le modèle nucléotidique de Markov utilisé dans notre travail : le modèle HKY. Ce modèle représente une réalité biologique qui est que les transitions -substitutions entre 2 purines ($A \leftrightarrow G$) ou entre 2 pyrimidines ($C \leftrightarrow T$)- sont observées plus fréquemment dans les séquences que les transversions -substitutions entre purine et pyrimidine. Ainsi, ce modèle admet l'existence de 2 taux de substitutions, celui des transitions a et celui des transversions b avec $a > b$ (excès de transitions par rapport aux transversions). Enfin, il permet d'introduire en plus n'importe quel vecteur de fréquences d'équilibre ($\pi_A\, \pi_T\, \pi_C\, \pi_G$).

$$Q = \begin{pmatrix} -\lambda_A & \pi_T b & \pi_C b & \pi_G a \\ \pi_A b & -\lambda_T & \pi_C a & \pi_G b \\ \pi_A b & \pi_T a & -\lambda_C & \pi_G b \\ \pi_A a & \pi_T b & \pi_C b & -\lambda_G \end{pmatrix}$$

Considérons une branche le long de laquelle s'applique le processus évolutif markovien associé à la matrice Q pendant t unités de temps. Supposons de plus que l'on se trouve à l'état stationnaire pour lequel le résidu i a la probabilité π_i. Etant donné que λ_i est le taux par unité de temps avec lequel une base i est modifiée, on en déduit que le taux par unité de temps avec lequel un site est modifié est $\sum_i \pi_i \lambda_i$ puisque π_i est la probabilité pour qu'un site soit dans l'état i. La longueur de la branche considérée, soit le nombre attendu de substitutions par site le long de cette branche, est alors exprimée par

$$l = \left(\sum_i \pi_i \lambda_i \right) t$$

Par convention, on normalise les taux m_{ij} de façon à ce que $\sum \pi_i \lambda_i = 1$. Alors, $l = t$ et le temps est mesuré par les longueurs des branches.

A l'inspection visuelle d'un alignement multiple de séquences homologues un peu divergentes, on découvre immédiatement que le taux d'évolution varie fortement entre sites

d'une même molécule.

L'approche couramment utilisée pour prendre en compte cette réalité consiste à faire l'hypothèse que les différents sites d'une même molécule évoluent selon une matrice des taux vQ, où Q est une matrice de taux constante, et v un nombre positif qui varie entre sites selon une distribution de probabilités appelée la distribution gamma et qui est de moyenne 1. Les sites pour lesquels v a une valeur élevée sont les sites d'évolution rapide. La Figure 1.10 présente plusieurs formes possibles de la distribution gamma de moyenne 1 qui dépendent d'un paramètre noté α. Sa variance vaut $1/\alpha$. Ainsi, les faibles valeurs de α modélisent les molécules dont les sites évoluent à vitesse très variable. La limite de ce modèle de variabilité du taux d'évolution entre sites quand α tend vers $+\infty$ est le modèle à taux constant entre sites. La distribution gamma est utilisée par commodité technique pour modéliser la variation des taux d'évolution entre sites, sans qu'elle n'ait de validité biologique particulière.

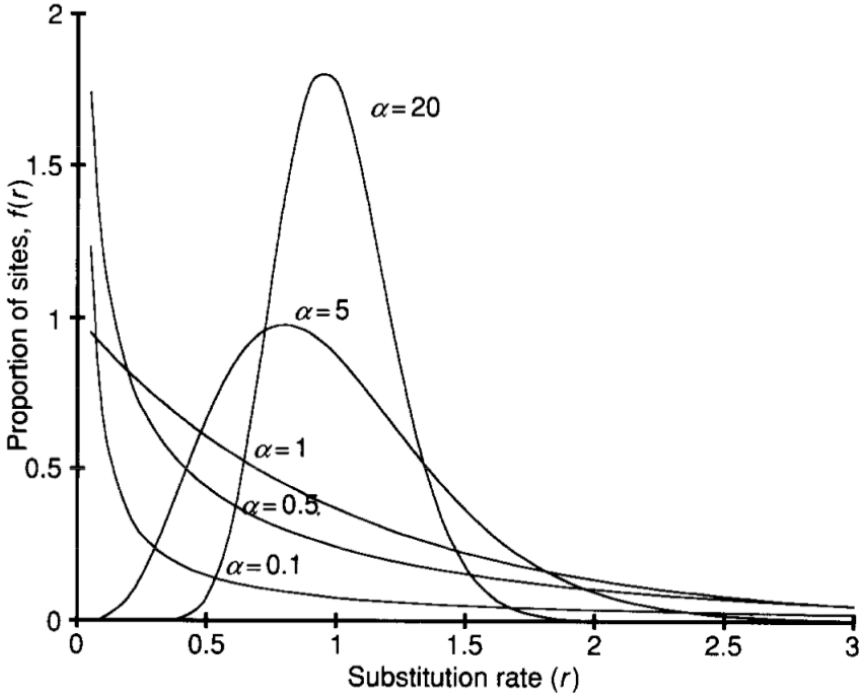


FIGURE 1.10 – La fonction de densité de probabilité, $f(r)$, de la distribution gamma selon les taux de substitution des sites (r) [Yang, 1996]

1.5.4.2 Approche par maximum de vraisemblance

Concernant la méthode de calcul d'arbre, l'approche au maximum de vraisemblance est plus précise que les approches par les méthodes de distances (dont la méthode Neighbor-Joining) [To et al., 2015]. Nous avons donc utilisé dans ce travail l'approche par maximum de vraisemblance.

L'approche par maximum de vraisemblance consiste à probabiliser entièrement le processus évolutif (une probabilité est attribuée à tous les événements évolutifs possibles depuis n'importe quelle séquence ancestrale jusqu'aux feuilles), et à trouver quel est le scénario évolutif, constitué d'une topologie (forme) d'arbre et des longueurs de ses branches, qui a la plus forte probabilité d'avoir donné naissance aux séquences analysées.

Le modèle probabiliste utilisé au maximum de vraisemblance comporte les paramètres suivants :

- un arbre phylogénétique raciné arbitrairement et ses longueurs de branches
- une matrice des taux Q normalisée commune à toutes les branches (qui implique des fréquences d'équilibre des résidus)
- une valeur α qui détermine la variation des taux d'évolution entre sites modélisée par une distribution gamma

Ce modèle prévoit que les données observées, un alignement de séquences homologues ($S1$, $S2$, ...), ont été engendrées par l'arbre à partir d'une séquence ancestrale inconnue, le processus Q ayant été appliqué sur chaque branche. On considère aussi que la séquence ancestrale, et donc toutes les autres, a une composition égale à la fréquence d'équilibre de la matrice Q (processus stationnaire). Enfin, on suppose que les sites de la molécule évoluent indépendamment les uns des autres (discutable mais nécessaire pour la faisabilité du calcul).

L'objectif est de calculer la vraisemblance du modèle, soit la probabilité des données pour les valeurs des paramètres du modèle. L'indépendance entre les processus évolutifs aux divers sites entraîne que la vraisemblance du modèle est le produit de la vraisemblance

pour chaque site. En pratique, on calcule le logarithme naturel de la vraisemblance (pour un problème de représentation informatique d'un nombre très faible)

$$log(L) = \sum_{sites} log(L(site))$$

La vraisemblance en un site se calcule à partir de la matrice de transition P qui donne la probabilité de tout état en fin de branche, pour toute longueur de branche, et tout état en début de branche. On a vu que, pour une branche de longueur l, $P = e^{Ql}$, puisque, Q étant normalisée, une longueur de branche est équivalente à une durée évolutive.

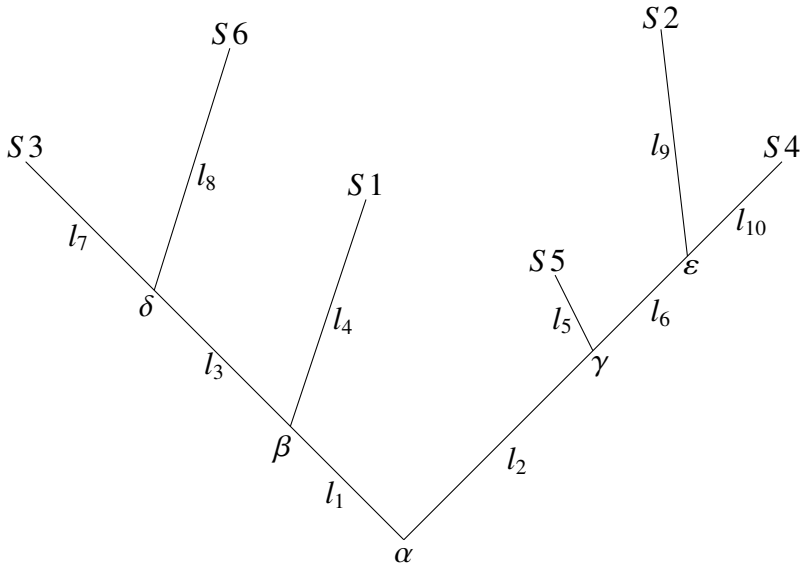


FIGURE 1.11 – Exemple d'arbre phylogénétique - $S\,i$: séquences homologues, l_i : longueurs des branches (subst./site), $\alpha, \beta, \gamma, \delta, \varepsilon$: résidus ancestraux inconnus dont toutes les combinaisons de valeurs sont envisagées [Deléage et al., 2013]

La vraisemblance en un site est la somme des probabilités de toutes les histoires évolutives possibles qui se terminent par les résidus observés dans les séquences au site traité. Dans l'exemple de la Figure 1.11, cela s'exprime par

$$L(site) =$$

$$\sum_{\alpha}\sum_{\beta}\sum_{\gamma}\sum_{\delta}\sum_{\varepsilon} \pi_\alpha P_{\alpha\beta}(l_1)P_{\alpha\gamma}(l_2)P_{\beta\delta}(l_3)P_{\beta S1}(l_4)P_{\delta S3}(l_7)P_{\delta S6}(l_8)P_{\gamma S5}(l_5)P_{\gamma\varepsilon}(l_6)P_{\varepsilon S2}(l_9)P_{\varepsilon S4}(l_{10})$$

où les sommes portent sur tous les résidus possibles (4 pour l'ADN), $P_{xy}(l)$ étant le terme x, y de la matrice P pour une branche de longueur l et S_z le résidu présent dans la séquence

z au site considéré. Le terme π_α représente la fréquence du résidu ancestral α inconnu.

L'algorithme de Felsenstein (non détaillé ici) permet de calculer progressivement la vraisemblance d'un site, des feuilles vers la racine, pour des arbres de bien plus grande taille où ce calcul de cumul de probabilités de tous les scénarios ancestraux deviendrait trop lourd.

La prise en compte de la variabilité des taux d'évolution entre sites selon la distribution gamma impose de discrétiser cette distribution en calculant les quantiles $1/K, 2/K, ..., (K-1)/K$ puis les moyennes r_i de cette distribution dans les K intervalles inter-quantiles. Ces valeurs r_i dépendent de la valeur du paramètre α qui contrôle la variance de la distribution. Les taux relatifs d'évolution des sites sont alors supposés prendre chaque valeur r_i avec la probabilité $1/K$. Comme on ne sait pas à quel taux chaque site évolue, on calcule la moyenne pondérée des vraisemblances de ce site comme s'il avait évolué à chacun des K taux. L'attribution du taux r_i à un site consiste à multiplier toutes les longueurs de branches de l'arbre par r_i, et à calculer la vraisemblance comme précédemment. Finalement, la vraisemblance d'un site s'exprime par

$$L(site) = \frac{1}{K} \sum_{i=1}^{K} L(site \mid r_i)$$

où $L(site \mid r_i)$ signifie que les longueurs des branches on été multipliées par r_i

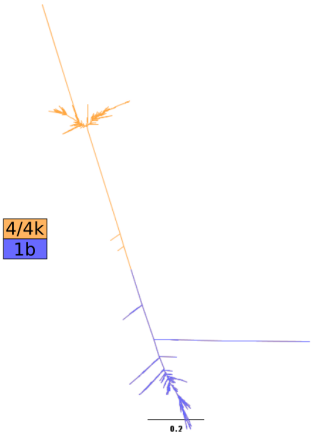
Une valeur généralement adéquate pour K, le nombre de valeurs que peut prendre la distribution gamma discrétisée, est 4.

Au final, étant donné l'expression de la vraisemblance du modèle pour des valeurs données de tous les paramètres du modèle, la méthode de reconstruction phylogénétique au maximum de vraisemblance consiste à rechercher quelles sont les valeurs des paramètres pour lesquels cette vraisemblance est maximale. La maximisation de la vraisemblance est réalisée en considérant une topologie d'arbre courante et en recherchant les valeurs des paramètres numériques (longueurs des branches, termes de la matrice Q, paramètre α) qui maximisent la vraisemblance pour cette topologie, puis en recommençant avec une topologie voisine de la topologie courante, et ceci tant que la vraisemblance augmente. L'arbre

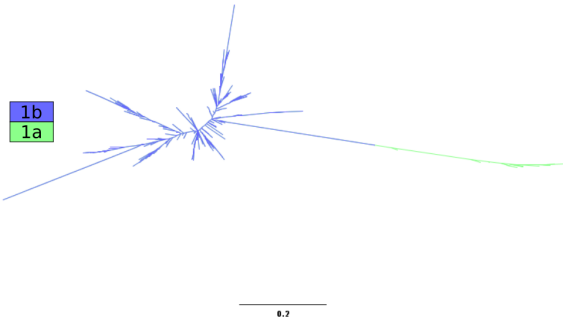
de départ est obtenu avec une méthode rapide, comme la méthode Neighbor-Joining.

Basée sur les distances évolutives entre séquences, cette méthode est une approximation efficace de la méthode d'évolution minimale qui consiste à sélectionner l'arbre qui requière le nombre minimal d'événements évolutifs. Effectivement, on cherche dans ce cas un arbre qui induit des distances aussi proches que possible des distances évolutives mesurées entre toutes les paires de séquences de l'alignement.

Le résultat de l'analyse au maximum de vraisemblance est une topologie non racinée d'arbre phylogénétique avec ses longueurs de branches, et des estimations des termes de la matrice Q et du paramètre α. La Figure 1.12 présente des exemples d'arbres générés par maximum de vraisemblance dans le cadre de ce travail.



(a) Patient 1 - région NS5B



(b) Patient 2 - région NS5B

FIGURE 1.12 – Arbres calculés par maximum de vraisemblance et selon le modèle de substitution HKY

Le logiciel PhyML permet de calculer des arbres phylogénétiques selon l'approche au maximum de vraisemblance.

Après avoir vu le principe de la méthode de reconstruction phylogénétique au maximum de vraisemblance, nous pouvons maintenant aborder le principe de la méthode d'analyse utilisée pour estimer la date d'événements ancestraux.

1.5.4.3 Estimation de la date d'événements ancestraux

Pour estimer la date d'événements ancestraux à partir d'un arbre phylogénétique, nous avons utilisé l'algorithme LSD («*Least-Squares Dating*») [To et al., 2015]. Cet algorithme prend en entrée un arbre phylogénétique raciné R d'un jeu de n séquences datées en série avec des longueurs de branches connues. On indice les nœuds internes de R par $1, 2, ..., n-1$ et les feuilles par $n, n+1, ..., 2n-1$. Le nœud 1 est la racine de l'arbre. La date du nœud i est désignée par t_i. Par conséquent, $t_n, t_{n+1}, ..., t_{2n-1}$ sont connues. Les dates sont mesurées depuis l'origine ce qui entraîne $t_i \geq t_j$ quand i est plus récent que j.

Pour tout nœud i différent de la racine ($i = 1$), on considère $a(i)$ le nœud parent de i. b_i est la longueur de la branche $(i, a(i))$; c'est une estimation du nombre de substitutions par site qui se sont produites le long de cette branche de $t_{a(i)}$ à t_i. Avec une horloge moléculaire stricte, le taux de substitution (soit le nombre attendu de substitutions par site par unité de temps) le long de l'arbre est constant et est désigné par ω. Le but de l'algorithme LSD est d'estimer le taux de substitution et les dates de tous les nœuds internes soit $(\omega, t_1, ..., t_{n-1})$.

En supposant une horloge moléculaire stricte, la longueur attendue de la branche $E(b_i)$ est égale à ω multiplié par l'intervalle de temps $(t_i - t_{a(i)})$. Etant donné le bruit d'échantillonnage et les erreurs d'estimation, l'estimation de la longueur de la branche b_i (issue de l'arbre d'entrée R) peut être exprimée par

$$b_i = \omega(t_i - t_{a(i)}) + \varepsilon_i$$

où ε_i est le terme de bruit (erreur). Une approximation normale est utilisée pour la distribution de ce terme de bruit. Par conséquent, nous supposons

$\varepsilon_i \sim N(0, \sigma_i^2)$

où $N(0, \sigma_i^2)$ désigne la distribution normale de moyenne nulle et de variance σ_i^2. Une limite de ce modèle est que les branches courtes pourraient être négatives selon l'expression de b_i mais une contrainte de précédence temporelle impose des branches positives (cf. ci-dessous). Comme l'évolution est indépendante d'une branche à l'autre, nous supposons que les termes de bruit sont mutuellement indépendants. Le critère des moindres carrés pondérés à minimiser est donné par

$$\phi(\omega, t_1, ..., t_{n-1}) = \sum_{i=2}^{2n-1} \frac{1}{\sigma_i^2}(b_i - \omega(t_i - t_{a(i)}))^2$$

avec
$\hat{\sigma_i^2} = \frac{b_i + c/s}{s}$ l'estimation de σ_i^2 basée initialement sur $\hat{\sigma_i^2} = \frac{b_i}{s}$
$c = 10$ introduit pour contrecarrer les faibles valeurs de b_i (estimé après simulations)
c/s introduit pour contrecarrer les faibles valeurs de s
s la longueur de la séquence

Le taux de substitution ω est positif et la valeur minimale du taux estimé est fixée dans LSD à $\hat{\omega} \geq \omega_{min} > 0$ (par défaut, $\omega_{min} = 10^{-10}$ subst./site/unité temps). De plus, le temps est mesuré en avançant de la racine vers les feuilles de l'arbre, par conséquent il doit satisfaire la contrainte de précédence temporelle $t_i \geq t_{a(i)}$ pour tout nœud i qui n'est pas la racine de l'arbre ($i = 1$). En d'autres termes, tout nœud fille ($i > 1$) est plus récent que son nœud parent ($a(i)$). Cette exigence évidente est implémentée dans LSD alors que ce n'est pas le cas dans d'autres algorithmes de datation, ceci pour des raisons de coût computationnel.

Les estimations sont obtenues en minimisant la fonction objectif ϕ qui a un unique minimum. Le minimum de cette fonction est calculé en annulant la dérivée au premier ordre de ϕ relativement à chaque variable $t_1, ..., t_{n-1}$. Cela nous permet de déduire une expression de t_i en fonction de ω ($t_i(\omega)$) pour $i = 1, ..., n-1$. La réinjection de cette expression dans ϕ nous permet d'estimer la valeur $\hat{\omega}$ qui minimise cette fonction. Si $\hat{\omega} < \omega_{min}$ alors nous posons $\hat{\omega} = \omega_{min}$. L'injection de $\hat{\omega}$ dans $t_i(\omega)$ nous permet de calculer toutes les estimations de dates $\hat{t_i}$.

Cependant, rien ne garantit que les estimations de dates satisfont les contraintes de précédence temporelle. Une adaptation de l'algorithme permet de s'assurer du respect de ces contraintes. Elle est basée sur une approche de résolution de problèmes d'optimisation avec contraintes linéaires : l'algorithme active set.

Soit $x = (\omega, t_1, ..., t_{n-1})$; la fonction à minimiser est $\phi(x)$ définie ci-dessus, soumise aux contraintes $t_i - t_{a(i)} \geq 0$ pour $i = 2, ..., 2n - 1$. x est un point possible si et seulement si il satisfait toutes les contraintes. Une contrainte i est «active» à x si et seulement si $t_i = t_{a(i)}$. L'algorithme active set appliqué à notre problème peut se résumer ainsi : en partant d'un point x possible avec C le jeu de contraintes actives (initialisé avec les contraintes violées ($t_i - t_{a(i)} < 0$) déduites de l'algorithme précédent), nous calculons la solution qui minimise ϕ sachant C soit la solution minimale telle que $t_i = t_{a(i)}$ pout tout $i \in C$. Nous avons à calculer par conséquent le point stationnaire (x^*, λ^*) de la fonction de Lagrange :

$$\Gamma(x, \lambda) = \phi(x) - \sum_{i \in C} \lambda_i (t_i - t_{a(i)})$$

Effectivement, la méthode des multiplicateurs de Lagrange (λ_i) permet de trouver les points stationnaires (dont le maximum et le minimum) d'une fonction dérivable d'une ou plusieurs variables, sous contraintes.

Nous vérifions ensuite si (i) quelques contraintes sont violées dans x^* et (ii) toutes les contraintes dans C sont utiles. C est mis à jour en fonction en écartant la contrainte la plus inutile (correspondant à la valeur la plus négative des λ_i) et en ajoutant la contrainte la plus violée ($t_i - t_{a(i)}$ la plus négative). L'algorithme s'arrête quand toutes les contraintes dans C sont utiles et aucune contrainte n'est violée.

Enfin, une dernière adaptation de l'algorithme permet d'estimer la position de la racine pour les arbres non racinés. Elle est effectuée en cherchant le point dans l'arbre qui minimise la fonction objectif ϕ quand l'arbre est raciné à ce point.

Les auteurs ont rendu disponible leur code source sur le site de la plateforme bioinformatique ATGC pour mettre en œuvre l'algorithme LSD afin d'estimer la date d'événements ancestraux.

Une autre approche basée sur la théorie de la coalescence a aussi permis d'estimer la date d'événements ancestraux. La théorie de la coalescence est un modèle rétrospectif de génétique des populations qui a pour objectif de suivre l'évolution de tous les allèles d'un gène donné de tous les individus d'une population, jusqu'à une seule copie ancestrale, appelée ancêtre commun le plus récent [Wakeley, 2009]. Les relations d'hérédité entre les allèles sont représentées sous la forme d'un arbre similaire à un arbre phylogénétique. Cet arbre est aussi appelé coalescent, et la compréhension des propriétés statistiques du coalescent sous différentes hypothèses forme la base de la théorie de la coalescence. Le coalescent utilise des modèles de dérive génétique, en remontant le temps pour reconstruire la généalogie des ancêtres. Cette approche ne sera pas détaillée ici mais les résultats qui en sont issus dans le cadre de ce travail recoupent largement ceux issus de l'approche LSD basée sur l'horloge moléculaire.

II

Contribution

2

Dynamique de l'évolution intra-patient du VHC par NGS

Dans cet article, nous avons voulu étudier la relation entre l'absence de réponse au traitement associant l'Interféron pegylé à la Ribavirine et la variation et la dynamique de l'évolution de 2 régions du génome du VHC (Core et NS5B) au cours du traitement et des années qui ont suivi pour 2 patients suivis pour une hépatite C chronique. La région Core étant plus conservée et jouant un rôle de contrôle, la région NS5B est particulièrement étudiée car elle permet la génération de l'ARN polymérase ARN dépendante du VHC qui permet sa réplication.

Evolutionary dynamics of hepatitis C virus in patients with mixed infection over a 13-year follow-up period

Caporossi A[1,2*], Kulkarni O[1*], Blum MGB[1], Leroy V[3], Morand P[4,5], Larrat S[4,5], François O[1]

1 Université Grenoble Alpes, TIMC-IMAG/CNRS/UMR 5525, Grenoble, France

2 Pôle Santé Publique, Centre Hospitalier Universitaire Grenoble Alpes, Grenoble, France

3 Clinique d'Hépatologie, Centre Hospitalier Universitaire Grenoble Alpes, Grenoble, France

4 Institut de Biologie Structurale UMR2075 CEA-CNRS-UGA, Grenoble, France

5 Laboratoire de Virologie, Pôle Biologie, Centre Hospitalier Universitaire Grenoble Alpes, Grenoble, France

*These authors contributed equally to the study

CORRESPONDENCE : Alban Caporossi acaporossi@chu-grenoble.fr

Laboratoire TIMC-IMAG - Faculté de Médecine - Domaine de la Merci 38706 La Tronche cedex - France

KEYWORDS : Hepatitis C Virus ; Multiple infections ; High-throughput sequencing ; Longitudinal study ; Genetic heterogeneity ; Selective sweeps.

SHORT TITLE : Evolution of HCV in patients with mixed infections

ABSTRACT :

High throughput sequencing of rapidly evolving hepatitis C virus populations enables large-scale studies of within-host viral heterogeneity and resistance to antiviral treatment regimes that are commonly observed in clinical cases with chronic infections. While escape from direct antiviral drugs can be explained by resistance mutations in targeted genomic regions, the absence of response to dual therapies has remained less understood. In this study we implemented amplicon sequencing to survey genomic variation at the Core and NS5B regions of HCV over a period of 13 years for two patients followed for chronic hepatitis C at Grenoble-Alpes University Hospital. From samples obtained at several time points, we observed mixed infection by multiple HCV genotypes in patients. Genetic heterogeneity and sample composition analysis provided information about the changes in viral population over the course of clinical treatment, with NS5B experiencing a sharp increase in diversity after treatment initiation compared to baseline. Evidence for HCV population genetic structure was observed in each patient, occurring in divergent lineages in phylogenetic trees. These observations point towards diversifying selection occurring post-treatment, acting on standing genomic variation and maintaining high genetic heterogeneity during infection. Being associated with treatment efficiency, the results provide the first evidence for antiviral treatment inducing soft selective sweeps in chronically infected patients with multiple HCV genotypes.

Introduction

High genetic heterogeneity is a principal characteristic of hepatitis C viruses (HCV) within infected hosts, and it is often used to predict the disease outcome (Farci et al. 2000). In recent years, high-throughput sequencing (HTS) technologies have enabled studying the evolution of viral populations over the course of infection by allowing the rapid acquisition of thousands of short RNA sequences from multiple time-point samples (Nelson & Hughes 2015). HTS has become an essential tool in describing genetic heterogeneity and in understanding intra-patient evolutionary dynamics of viral species (Beerenwinkel et al. 2012). More specifically HTS has been applied to describe within-host diversity of HCV (Lauck et al. 2012 ; Li et al. 2011 ; Ramachandran et al. 2011), to detect transmission bottlenecks in infection (Wang et al. 2010), to predict response to triple therapy in chronic hepatitis C patients and evaluate resistance variants of protease inhibitors (Larrat, Kulkarni, et al. 2015), and to study the response of HCV to selection imposed by antiviral drugs (Ahmed & Felmlee 2015).

In spite of the importance of understanding within-host evolutionary dynamics of HCV, surveys of intra-host evolution of genotypes of chronically infected patients along extended periods of time have however remained relatively infrequent (Culasso et al. 2014, Raghwani et al. 2016). A better understanding of intra-patient evolution of HCV in populations is thus needed before accurate predictions of response to treatments can be made (Di Lello et al. 2015).

During the course of infection, HCV evolves rapidly and results in quasi-species (Eigen et al. 1988 ; Gregori et al. 2013). The quasi-species composition of HCV populations allows the virus to escape the immune response and develop resistance to treatment leading to chronic infection (Bull et al. 2011 ; Astrovskaya et al. 2011 ; Vermehren et al. 2012 ; Lauring et al. 2013 ; Larrat, Vallet, et al. 2015). HCV quasi-species have been classified into 7 genotypes, which differ from each other from 30% to 33% at the nucleotide level, and 65 subtypes, which differ from each other from 20% to 25% (Smith et al. 2014 ; Simmonds et al. 2005). The evolution of drug resistance in highly mutable RNA

viruses occurs by the fixation of specific drug-resistance mutations, and HCV genotypes and subtypes confer differential resistance to multiple therapies.

Considering a retrospective analysis of HCV evolutionary dynamics in non-responder patients, treatments were based on dual therapies with interferon-alpha and ribavirin (dates of treatment in the years 2003 and 2008). For these treatments, genotypes 1 and 4 were known to be more resistant than genotypes 2 and 3 to the antiviral action of interferon-alpha and ribavirin (Pawlotsky 2009). However the mechanisms by which resistance occurred in patients with mixed infections from multiple genotypes remain unknown (Schröter et al. 2003), and it has been hypothesised that resistance to treatment in RNA viruses could involve either hard or soft selective sweeps (Pennings et al. 2014 ; Feder et al. 2016). In a classic or 'hard' sweep, a resistance mutation arises in a single viral particle, and ultimately reaches fixation in the entire intra-patient population (Maynard Smith & Haigh 1974 ; Burke 2012). As this mutation spreads to the whole population, background mutations linked to it also increase in frequency, and reduce genetic diversity in the population. A 'soft' selective sweep, by contrast, describes the dynamics of selection acting on mutations present on many particles in a population with standing genetic variation (Hermisson & Pennings 2005). Resistance mutations originate multiple times on different genetic backgrounds, and selection can increase overall genetic diversity. Soft sweeps have been identified in HIV (Pennings et al. 2012 ; Messer & Petrov 2013), but this mode of selection has not been described in longitudinal studies of HCV patients.

In this study, we performed deep sequencing on two regions of the HCV genome, Core (464bp) and NS5B (381bp), for two chronically infected non-responder patients treated with dual therapy in the 2000's. Multiple genotypes and subtypes were detected in both patients. By using RNA sequences from multiple sampling time points, the evolution of HCV genetic heterogeneity was followed over a period of time greater than thirteen years. The evolutionary dynamics of viral populations was studied by using measures of nucleotide diversity, exploratory data analysis, coalescent modelling and phylogenetic

methods. Phylogenetic reconstruction on temporal samples pinpointed population structure and complex evolutionary dynamics of viral populations during infection. Our results provided evidence of soft selective sweeps underpinning the intra-host evolution of multiple genotypes of HCV along a 13-year follow-up period.

Materials & methods

Patients under study & sample preparation and sequencing

Two patients followed at Grenoble-Alpes University Hospital for a chronic hepatitis C virus infection were included in the study. The patients had a known date of infection because of an identified transmission event, transfusion or professional exposure. Dates of antiviral treatment were also available (years 2003 and 2008). The number of temporal samples ranged from 5 to 8 per patient, and a total of thirteen serum samples were available over a follow-up ranging from 10 to 13 years (Table 1). Informed consents for participating in research were obtained from the patients.

The samples were stored at $-80°C$ and were retrospectively analysed. HCV RNA was extracted from 1 ml of plasma using EasyMAG (bioMérieux, Marcy l'Étoile, France) with an elution volume of $25\mu l$. Extracted RNA ($15\mu l$) was purified using Turbo DNase Ambion (Life Technologies, Cergy Pontoise, France), and cDNA was synthesised with the AccuScript high-fidelity kit (Agilent, Garches, France) using random primers.

For nucleotide sequencing, the 454 GS Junior platform was used (Roche Diagnostics). The Core region (nucleotides [nt] 288 to 751 according to reference HCV strain H77) (Pham et al. 2009) and the NS5B (nt 8256-8636) (Sandres-Sauné et al. 2003) region were amplified using Phusion Hot Start II (Finnzyme, Illkirch, France) and primers described in (Pham et al. 2009 ; Sandres-Sauné et al. 2003) which were added universal M13 tails. The Polymerase Chain Reaction (PCR) cycling conditions were initial denaturation at $95°C$ for 15 min followed by 50 cycles of denaturation at $94°C$ for 30 s, annealing at $64°C$ for Core and $57°C$ for NS5B 30s and elongation at $72°C$ for 70s with a final step of extension at $72°C$ for 10 min. Sample-specific multiplex identifier (MID) sequences

associated with universal M13 sequence were added with 25 cycles of PCR (56°*C* 15s, 72°*C* 20s) using the same Taq polymerase on PCR products 1 : 50 diluted.

The amplification efficacy was assessed using the Agilent DNA 1000 reagent kit and the Agilent 2100 expert bioanalyzer (Agilent, Les Ulis, France). The expected amplicons were obtained for all the samples. The amplicons were pooled equimolarly and purified using Agencourt AMPure XP reagents (Beckman Coulter, Roissy, France).

A library on the amplicon pool was prepared using the GS Junior rapid library prep kit. An emulsion PCR was run with the GS Junior emPCR (Lib-L) kit. Sequencing was performed with a 454 GS Junior PicoTiterPlate to a target depth reading of 1,000X. For Core and NS5B regions respectively, a mean of 22,071 +/- 3,336 and 9,209 +/- 969 reads per nucleotide was obtained with a median length of 579 nt and 392 nt.

Sequence data preparation

The SFF files obtained from the 454 GS Junior were converted to the standard FASTQ format using the Roche supplied sffinfo tool, a part of the GS Data Analysis Package (http://www.454.com/products/analysis-software/). Cutadapt (Martin 2011) was used for adapter removal. Corrections of 454 error modes such as homopolymer indels and carry forward/incomplete extension (CAFIE) were performed using RC454 (Henn 2012) with a consensus assembly for each region (Core/NS5B) generated by VICUNA (Yang 2012). Additional filtering by read sequence length (300 bp) was performed. Duplicate sequences were removed from the final alignments, and their multiplicity values per patient and per time point were recorded. The Core region sequences, which have been previously reported as being highly conserved ($\mu = 0.28 - 0.43 \times 10^{-3}$ substitutions/site/year, Gray et al. 2011), were used as control and compared to GENBANK in order to estimate a rate of sequencing error.

Sample composition and genetic heterogeneity analysis

Sampled genotypes and subtypes were identified by using the HCV BLAST align tool on the HCV database web site (http://hcv.lanl.gov, Kuiken et al. 2005). For all

samples, the frequencies of identified genotypes or subtypes were computed from the BLAST outputs, and average similarity scores were reported for all the phylogenetic groups identified. Nucleotide Diversity Estimates (NDEs) were obtained for each region and each time point using the R package pegas (Paradis 2010 ; R Core Team 2016). NDEs were also computed after separating the identified genotypes or subtypes in patient samples. In addition, within-patient population genetic structure and temporal evolution were investigated by using principal component analysis (PCA) on polymorphic RNA sites. PCA were performed with the R program 'prcomp' (R Core Team 2016) after re-sampling the patient data in order to reduce the effect of an unbalanced design.

Phylogenetic trees

Global phylogenetic trees including all unique reads were obtained using the neighbour-joining (NJ) algorithm (Saitou & Nei 1987). Figtree was used to obtain graphical tree representations (http://tree.bio.ed.ac.uk/software/figtree). The NJ trees were used for checking the clock-likeness of each data set using Path-O-Gen/TempEst, which performs a regression of root-to-tip genetic distance on sampling times (Rambaut et al. 2016). Only the trees that passed the criterion of having a positive correlation between divergence and sampling time points were considered for assessing maximum clade credibility trees from subsequent coalescent analyses. For patient 1 who was infected by multiple genotypes, separate trees for each genotype were used for assessing the clock-likeness criterion.

Phylogenetic trees were also obtained using the maximum likelihood (ML) method. PhyML 3.1 was used to compute the trees with default parameters (Guindon et al. 2010). To minimise the effect of unbalanced samples, 100 unique reads per time point were analysed after multiple sequence alignment using MAFFT 7 with the option "adjust-direction" to comply with direction of nucleotide sequences (Katoh & Standley 2013).

Coalescent analysis

The BEAST software (v1.8.1) was used for a Bayesian Markov Chain Monte Carlo (MCMC) analysis of the demographic parameters of HCV samples for patient 1

and 2 (Drummond et al. 2012). Each replicate included 20 sequences from each time point. These sequences were obtained by subsampling unique reads for each time point. Subsampling was performed using the 'subsample_fasta.py' script in the QIIME package (Caporaso et al. 2010). The final sequence alignment for BEAST was obtained using MAFFT (Katoh & Standley 2013) on default settings.

Molecular substitution model selection was performed using jModelTest (Darriba et al. 2012) on default parameters. The models with the optimal values of the Bayesian information criterion were chosen as parameters in BEAST runs. The HKY+G substitution model was the most suitable model across all data sets (Hasegawa et al. 1985). Regarding the molecular clock model, relaxed models did not converge, and a strict model was used for all runs. For computing demographic parameters (TMRCA, Time since the most recent common ancestor), the tips were calibrated using the 'Guess dates' option (Drummond et al. 2005). All priors of the Bayesian analysis were set at their default values except for the clock rate parameter, which was assigned a weakly informative gamma distribution with shape 0.001 and scale 1,000.

Each replicate was assigned a chain length of 20 million steps. When MCMC runs did not achieve convergence, the chain lengths were doubled. Replicates achieving no convergence were discarded. Tracer v1.5 was used for post-run analysis. Convergence was checked via posterior effective sample size values being greater than 200.

Molecular-clock analysis

In addition to the coalescent analysis, LSD a simple and fast molecular-clock model based on a Gaussian model was applied to date ancestral events by using least squares criterion (To et al. 2016). We used the ML trees as input of LSD as well as the sampling dates of the sequences for temporal constraints on the tips. We performed the analysis with the temporal precedence constraint version of the algorithm to impose the constraint that the date of every node is equal or more ancient than the dates of its descendants. We computed confidence intervals for mutation rates and for dates by replicating the analysis on 100 trees.

Results

Sample composition and genetic diversity analysis

Using high-throughput sequencing, a total number of 15,234 RNA sequences for the Core and NS5B regions were obtained at distinct time points of HCV infection for two patients. For each patient sample, the allele frequency spectrum exhibited a large excess of uniquely represented sequences (90%), and a total number of 14,494 reads (95%) occurred less than twice in each sample.

For each sample, similarity scores were computed for the NS5B and Core regions by using the BLAST align tool on the HCV database (Kuiken et al. 2005). The similarity scores ranged between 93.3% and 99.1% (Table 2 and Table 3), and all scores were above thresholds considered in subtype classification (Simmonds et al. 2005). For the control region (Core region), the average similarity score for genotype 1b sequences with sequences obtained from Sanger sequencing was 98.2%, showing that HTS sequencing errors occurred at a small rate in sample alignments.

For patient 1, the sample composition at NS5B indicated a co-occurrence of two genotypes, 1b and 4/4k (unclassified subtype 4 and subtype 4k) (Figure 1C, Table 2). The frequency of genotype 4/4k increased from 0% in 2002 (before treatment) to 68.89% during the year of treatment. After treatment failure, genotype 4 and subtype 4k gradually increased in frequency to reach 99.1% in 2014, but those genotypes were not found in the 2010 sample. The results for the Core region showed a pattern similar to NS5B (Figure 1A, Table 2). Consistently with the fact that Core is less divergent in terms of genotype separation than is NS5B, variation was less pronounced for the Core region. The frequency of genotype 4/4k increased from 0% in 2002 (before treatment) to 4.97% during the year of treatment. After treatment failure, genotype 4 and subtype 4k also increased in frequency at the Core region, and reached 55.8% in the 2014 sample. Nucleotide diversity estimates (NDEs) for the Core region remained at a constant level during infection (NDE = 0.029, sd = 0.014, Figure 1B). In contrast, the NDE values for the NS5B region exhibited a sharp increase at the onset of treatment with dual therapy during the year 2003

(NDE = 0.14, sd = 0.0009, Figure 1D). This sudden increase was followed by a gradual decrease after treatment (NDE = 0.023, sd = 0.006 in 2014). Overall, the results supported a multiple infection hypothesis in which patient 1 was infected from two genetically distinct genotypes.

For patient 2, the results at NS5B detected the presence of 2 subtypes of genotype 1, subtypes 1a and 1b (Figure 2C, Table 3). The frequency of subtype 1a increased from 0% in 2005 (before treatment) to 18.4% in 2010/2014 after treatment. The results for Core indicated that this region was similar to subtype 1b, except for the 2010 sample for which the frequency of subtype 1a was around 18.2% (Figure 2A, Table 3). The results supported a multiple infection hypothesis in which patient 2 was infected from subtype 1a and subtype 1b HCV strains. Genetic heterogeneity at the Core region remained at a constant level during infection (NDE = 0.011, sd = 0.003, Figure 2B), except for the sample from year 2010. For the 2010 sample, substantial levels of NDE were observed (Figure 2B, NDE = 0.093, sd = 0.001). For the NS5B region, genetic heterogeneity increased after treatment with dual therapy from NDE = 0.026 (2008, sd = 0.001) to NDE = 0.10 (2010, sd = 0.02), and it decreased to a lower level in 2014 (Figure 2D).

Population structure analysis

RNA sequences from the NS5B region were used to produce principal component (PC) plots for all time samples after correction for uneven sampling (Figure 3). For patient 1 and patient 2, the PC plots reproduced the temporal evolution of genotype and subtype frequencies obtained in Figure 1 and Figure 2 respectively. For patient 1, the first axis clearly separated genotype 1 strains from genotype 4 strains. Axis position in each time sample strongly agreed with sample composition. The second axis separated samples according to their sampling dates. For patient 2, sequences obtained prior to treatment were grouped together, and clustered separately from sequences obtained post-treatment. The first axis separated the strains from subtype 1a and subtype 1b.

Phylogenetic analysis

Global NJ trees and coalescent trees for two patients were obtained from the NS5B

and Core sequences sampled during the course of infection. Results for patients 1 and 2 are reported in Figure 4 (NS5B region) and in supplementary Figure S1 (Core region). ML trees for a sample consisting of 100 unique reads per time point for the same two patients were also computed for the NS5B and Core regions. Overall, the results supported a hypothesis of diversifying selection acting on genome-wide standing variation after treatment initiation in patients 1 and 2 (Hermisson & Pennings 2005).

For patient 1, the global phylogenetic NJ tree for NS5B exhibited two major lineages corresponding to the co-existence of genotype 1 and genotype 4/4k (Figure 4A). Trees of genotype 4 sequence alignments exhibited a divergent clade that contained sequences sampled in year 2007 and all sequences sampled in years 2008 and 2014. The average divergence between 2003 genotype 4 strains and 2014 strains was about 21%. A coalescent analysis of genotype 4 sequence alignments estimated the date of the most recent ancestor for this genotype during the year 1999 (Figure 4C, Figure S2). This date had a 95% highest posterior density interval equal to I = (1998.1, 2001.6), indicating that secondary infection by genotype 4 after the initial infection by genotype 1 in year 1979 is likely to predate the year 2001. The molecular-clock analysis using LSD estimated the date of the most recent ancestor of the genotype 4 strains during the year 1995 with a 95% confidence interval (CI) equal to (1989.9, 1997.2) (Figure S3).

For patient 2, the BEAST coalescent tree provided results consistent with a global phylogenetic NJ tree reconstructed from all sequences from the NS5B region (Figure 4B-D). Following treatment initiation, two divergent lineages were observed in the years 2010 and 2014 corresponding to the HCV subtypes 1a and 1b (Figure 4D, Figure S2). The average difference between the 2005 strains (all 1b) and subtype 1a strains sampled in 2014 was about 16.5%, and were of the same order of divergence as in patient 1 genotype 4 strains. The date of the most recent ancestor for genotype 1b sequences was inferred during the year 1964. The large 95% highest posterior density interval for this estimate, I = (1940,1984), included the known infection date (year 1970). The date of the most recent ancestor for subtype 1a sequences was found during the year 1997 with a 95%

highest posterior density interval equal to I = (1994, 2003), indicating that secondary infection by subtype 1a likely occurred in the nineties (Figure 4D). The molecular-clock analysis estimated the date of the most recent ancestor for genotype 1b during the year 1966 with a 95% CI equal to (1954.7, 1976.2). The date of the most recent ancestor for subtype 1a sequences was found during the year 1996 with a 95% CI equal to (1992.5, 1998.9) (Figure S3).

Discussion

Viral populations consist of 'quasi-species' representing a collection of multiple strains of the same virus with small changes in their genomic sequence (Astrovskaya et al. 2011). Quantifying viral population diversity within a host requires HTS technology capable of sampling a large number of individual sequences accurately (Bernini et al. 2011 ; Culasso et al. 2014 ; Nelson & Hughes 2015). Our study described the evolution of the Core and NS5B regions of the HCV genome in two non-responder patients before and after their treatment with dual therapy in the 2000's. Based on multiple time samples and high-throughput amplicon sequencing, we conducted a retrospective analysis of within-host HCV genetic heterogeneity in response to treatment. For both patients, sample composition, population structure analysis, and nucleotide diversity estimates provided consistent descriptions of the viral population dynamics. Similarity and phylogenetic analyses revealed the existence of multiple lineages of HCV corresponding to distinct genotypes (patient 1) or subtypes (patient 2). The BEAST and LSD analyses of the NS5B region provided consistent estimates of the date of secondary infection for both patients, with higher accuracy for LSD due to a larger sample size.

The Core region of the HCV genome is known to correspond to a highly conserved structural gene, and this region has low evolutionary rate compared to the rest of the genome (Gray et al. 2011). In agreement with the prediction for this region, genetic heterogeneity remained at a low level during infection in each patient. Nevertheless a punctual increase of diversity was found in patient 2 at one time point in year 2010. Although this

last event remains uncharacterised, it was consistent with the co-occurrence of distinct viral subtypes. Differences in HCV variants of highly conserved regions have been observed between plasma and peripheral blood cells during chronic infection (Roque-Afonso et al. 2005). In addition, phylogenies reconstructed from HCV genomes sampled from multiple time points have frequently revealed two or more genetically distinct co-existing viral lineages that are not detected at all sampling times (Ramachandran et al. 2011). We provided evidence that HCV exhibits within-host population genetic structure, and those findings supported the hypothesis that the virus is able to circulate between distinct cell tissues (Gray et al. 2012 ; Raghwani et al. 2016).

Distinct genomic compartments of HCV can exhibit highly different evolutionary trajectories (Culasso et al. 2014). In agreement with this observation, trajectories of the NS5B region differed from those of the Core region substantially. In each patient, genetic heterogeneity at the NS5B region exhibited a sharp increase, and reached exceptional levels after treatment (Raghwani et al. 2016). The diversity peak was followed by a long period during which genetic diversity decreased gradually. We interpret the results as evidence of soft selective sweeps occurring at the NS5B or at a linked region in the genome, and mixed infection maintained by balancing selection during viral evolution.

Soft sweeps differ from classic sweeps as they occur when one or many resistant strains are present in the standing genetic variation prior to treatment, possibly at low frequency (Pennings et al. 2012). Classic selective sweeps assume that resistance mutations occur in the viral genome de novo, and evolve to fixation in the entire population (Messer & Petrov 2013). After treatment, the frequencies of genotypes or subtypes carrying resistance mutations increased in the viral population. Parallel selection on resistant strains from diverging lineages resulted in an increase of genetic heterogeneity. Following the diversity burst, genotypes with lower adaptive rates were slowly replaced by genotypes with higher adaptive rates, generating a gradual decrease of genetic diversity post treatment. The BLAST and phylogenetic analyses on the NS5B region confirmed our hypothesis and revealed complex evolutionary dynamics in a structured viral population.

In conclusion, evolution of the Core and NS5B regions from HCV populations in two patients with mixed infection was studied using amplicon sequencing. Analysing sample composition and measures of genetic diversity, contrasting patterns were reported for the Core and NS5B regions. These patterns could be explained by complex population structure and by diversifying selection acting on distinct genotypes over the course of infection. While it is important in the evolution of drug resistance for HIV (Pennings et al. 2014), evolution through soft sweeps has not been previously reported for HCV. According to Feder et al. (2016), the evidence for soft sweeps occurring in the HCV genome after treatment is consistent with a low efficiency of the medical treatment. Dual therapy was based on treatments for which the occurrence of non-responder patients has not been clearly explained by viral resistance (Larrat, Vallet, et al. 2015). In contrast, more efficient direct anti-viral treatments have many naturally occurring resistance mutations, and these treatments could generate harder selective sweeps (Feder et al. 2016). We interpret the observed patterns of genetic diversity and the shape of phylogenetic trees as the first evidence for evolution of resistance to treatment through soft selective sweeps in patients with chronic multiple HCV infection. This result is of crucial importance to our understanding of the evolution of HCV infection and the efficiency of past and new treatments.

Acknowledgements, funding : Om Kulkarni was funded by the Marie-Curie Initial Training Network INTERCROSSING.

References

Ahmed, A. & Felmlee, D.J., 2015. Mechanisms of hepatitis C viral resistance to direct acting antivirals. Viruses, 7(12), pp.6716–6729.

Astrovskaya, I. et al., 2011. Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC bioinformatics, 12 Suppl 6(Suppl 6), p.S1.

Beerenwinkel, N. et al., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Frontiers in Microbiology, 3, p.329.

Bernini, F. et al., 2011. Within-host dynamics of the hepatitis C virus quasispecies population in HIV-1/HCV coinfected patients J. Tavis, ed. PLoS ONE, 6(1), p.e16551.

Bull, R.A. et al., 2011. Sequential bottlenecks drive viral evolution in early acute hepatitis c virus infection J. J. Ou, ed. PLoS Pathogens, 7(9), p.e1002243.

Burke, M.K., 2012. How does adaptation sweep through the genome ? Insights from long-term selection experiments. Proceedings. Biological sciences / The Royal Society, 279(1749), pp.5029–38.

Caporaso, J.G. et al., 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7(5), pp.335–336.

Cock, P.J.A. et al., 2009. Biopython : freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford, England), 25(11), pp.1422–3.

Culasso, A.C.A. et al., 2014. Intra-host evolution of multiple genotypes of hepatitis C virus in a chronically infected patient with HIV along a 13-year follow-up period. Virology, 449, pp.317–327.

Darriba, D. et al., 2012. jModelTest 2 : more models, new heuristics and parallel computing. Nature Methods, 9(8), pp.772–772.

Drummond, A.J. et al., 2005. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Molecular Biology and Evolution, 22(5), pp.1185–1192.

Drummond, A.J. et al., 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29(8), pp.1969–1973.

Eigen, M., McCaskill, J. & Schuster, P., 1988. Molecular quasi-species. The Journal of Physical Chemistry, 92(24), pp.6881–6891.

Farci, P. et al., 2000. The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. Science (New York, N.Y.), 288(5464), pp.339–44.

Feder, A.F. et al., 2016. More effective drugs lead to harder selective sweeps in the evolu-

tion of drug resistance in HIV-1. eLife, 5.

Gouy, M., Guindon, S. & Gascuel, O., 2010. SeaView Version 4 : A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Molecular Biology and Evolution, 27(2), pp.221–224.

Gray, R.R. et al., 2012. A new evolutionary model for hepatitis C virus chronic infection G. F. Rall, ed. PLoS Pathogens, 8(5), p.e1002656.

Gray, R.R. et al., 2011. The mode and tempo of hepatitis C virus evolution within and among hosts. BMC Evolutionary Biology, 11(1), p.131.

Gregori, J. et al., 2013. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants O. Schildgen, ed. PLoS ONE, 8(12), p.e83361.

Gregori, J. et al., 2016. Viral quasispecies complexity measures. Virology, 493, pp.227–237.

Guindon, S. & Gascuel, O., 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology, 52(5), pp.696–704.

Guindon, S. et al., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies : Assessing the Performance of PhyML 3.0. Systematic Biology, 59(3), pp.307–321.

Hall, M.D., Woolhouse, M.E.J. & Rambaut, A., 2016. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods : A simulation study. Virus Evolution, 2(1).

Hasegawa, M., Kishino, H. & Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution, 22(2), pp.160–174.

Henn, M.R., et al., 2012. Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. PLoS Pathogens 8(3)

Hermisson, J. & Pennings, P.S., 2005. Soft sweeps : molecular population genetics of

adaptation from standing genetic variation. Genetics, 169(4), pp.2335–52.

Katoh, K. & Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7 : Improvements in Performance and Usability. Molecular Biology and Evolution, 30(4), pp.772–780.

Kuiken, C. et al., 2005. The Los Alamos hepatitis C sequence database. Bioinformatics, 21(3), pp.379–384.

Larrat, S., Vallet, S., et al., 2015. Naturally Occurring Resistance-Associated Variants of Hepatitis C Virus Protease Inhibitors in Poor Responders to Pegylated Interferon-Ribavirin Y.-W. Tang, ed. Journal of Clinical Microbiology, 53(7), pp.2195–2202.

Larrat, S., Kulkarni, O., et al., 2015. Ultradeep Pyrosequencing of NS3 To Predict Response to Triple Therapy with Protease Inhibitors in Previously Treated Chronic Hepatitis C Patients M. J. Loeffelholz, ed. Journal of Clinical Microbiology, 53(2), pp.389–397.

Lauck, M. et al., 2012. Analysis of Hepatitis C Virus Intrahost Diversity across the Coding Region by Ultradeep Pyrosequencing. Journal of Virology, 86(7), pp.3952–3960.

Lauring, A.S., Frydman, J. & Andino, R., 2013. The role of mutational robustness in RNA virus evolution. Nature Reviews Microbiology, 11(5), pp.327–336.

Di Lello, F.A., Culasso, A.C.A. & Campos, R.H., 2015. Inter and intrapatient evolution of hepatitis C virus. Annals of Hepatology, 14(4), pp.442–449.

Li, H. et al., 2011. Genetic Diversity of Near Genome-Wide Hepatitis C Virus Sequences during Chronic Infection : Evidence for Protein Structural Conservation Over Time N. H. Shoukry, ed. PLoS ONE, 6(5), p.e19562.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), p.10.

Maynard Smith, J. & Haigh, J., 1974. The hitch-hiking effect of a favourable gene. Genetical research, 23(1), pp.23–35.

Messer, P.W. & Petrov, D.A., 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology & Evolution, 28(11), pp.659–669.

Nelson, C.W. & Hughes, A.L., 2015. Within-host nucleotide diversity of virus populations : Insights from next-generation sequencing. Infection, Genetics and Evolution, 30, pp.1–7.

Paradis, E., 2010. pegas : an R package for population genetics with an integrated-modular approach. Bioinformatics, 26(3), pp.419–420.

Pawlotsky, J.-M., 2009. Therapeutic implications of hepatitis C virus resistance to antiviral drugs. Therapeutic advances in gastroenterology, 2(4), pp.205–19.

Pennings, P.S. et al., 2014. Loss and Recovery of Genetic Diversity in Adapting Populations of HIV C. Fraser, ed. PLoS Genetics, 10(1), p.e1004000.

Pennings, P.S. et al., 2012. Standing Genetic Variation and the Evolution of Drug Resistance in HIV R. J. De Boer, ed. PLoS Computational Biology, 8(6), p.e1002527.

Pham, D.A. et al., 2009. High prevalence of Hepatitis C virus genotype 6 in Vietnam. Asian Pac J Allergy Immunol, 27(2–3), pp.153–160.

R Core Team, 2016. R : A Language and Environment for Statistical Computing.

Raghwani, J., et al., 2016. Exceptional Heterogeneity in Viral Evolutionary Dynamics Characterises Chronic Hepatitis C Virus Infection. PLoS Pathogens 12(9) :e1005894.

Rambaut, A., et al., 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evolution, 2 (1) vew007 ; DOI : 10.1093/ve/vew007

Ramachandran, S. et al., 2011. Temporal Variations in the Hepatitis C Virus Intrahost Population during Chronic Infection. Journal of Virology, 85(13), pp.6369–6380.

Roque-Afonso, A.-M. et al., 2005. Compartmentalization of Hepatitis C Virus Genotypes between Plasma and Peripheral Blood Mononuclear Cells. Journal of Virology, 79(10), pp.6349–6357.

Saitou, N. & Nei, M., 1987. The neighbor-joining method : a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4), pp.406–25.

Sandres-Sauné, K. et al., 2003. Determining hepatitis C genotype by analyzing the sequence of the NS5b region. Journal of Virological Methods, 109(2), pp.187–193.

Schröter, M. et al., 2003. Multiple infections with different HCV genotypes : prevalence and clinical impact. Journal of Clinical Virology, 27(2), pp. 200–204.

Simmonds, P. et al., 2005. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. Hepatology, 42(4), pp.962–973.

Smith, D.B. et al., 2014. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes : Updated criteria and genotype assignment web resource. Hepatology, 59(1), pp.318–327.

To, T.H. et al., 2016. Fast Dating Using Least-Squares Criteria and Algorithms. Systematic Biology, 65(1), pp.82–97.

Vermehren, J. et al., 2012. The role of resistance in HCV treatment. Best Practice & Research Clinical Gastroenterology, 26(4), pp.487–503.

Wang, G.P. et al., 2010. Hepatitis C Virus Transmission Bottlenecks Analyzed by Deep Sequencing. Journal of Virology, 84(12), pp.6218–6228.

Yang, X. et al., 2012. De novo assembly of highly diverse viral populations. BMC Genomics 13 :475.

Tables and Figures.

Patient id	Infection year*	Treatment duration**	Number of samples	Range of samples
1 DEHE	1979	Jan 2003 June 2003	8	2002-2014
2 AMCA	1970	Mar 2008 Oct 2008	5	2005-2014

* Infection by genotype 1 from blood transfusion or professional exposure.
** Dual therapy: pegylated interferon and ribavirin.

TABLE 1 – Patient and treatment information.

Year	Number of reads	% Genotype		% Nucleotide similarity (Var)**			
		1b	**4/4k**	**1b**		**4/4k**	
NS5B							
2002*	622	**99.9**	--	95.64	(0.96)	--	
2003*	1031	31.10	**68.89**	95.88	(1.31)	93.80	(0.22)
2005*	927	51.82	48.01	96.06	(0.98)	93.69	(0.25)
2006*	626	**82.58**	17.25	95.85	(0.70)	93.85	(0.16)
2007	257	15.95	**84.05**	95.73	(0.70)	93.38	(0.51)
2008	239	8.78	**91.22**	95.61	(1.94)	93.30	(0.41)
2010	240	**100**	--	96.10	(0.40)	--	
2014	242	0.08	**99.10**	95.50	(0.50)	92.77	(0.22)
Core							
2002*	542	**99.44**	--	98.87	(0.50)	--	
2003	181	**95.02**	4.97	98.32	(0.70)	97.33	(1.00)
2005	610	**100**	--	99.06	(0.58)	--	
2006	510	**100**	--	98.46	(0.47)	--	
2007	595	**97.82**	2.18	98.89	(0.66)	96.30	(1.56)
2008	396	**86.62**	13.38	98.16	(1.17)	97.24	(0.61)
2010*	432	**99.76**	--	97.89	(0.64)	--	
2014	523	44.16	**55.83**	97.80	(2.39)	96.94	(0.78)

* This sample had unclassified genotypes (<2%) not shown.
** Similarity score obtained from BLAST in HCV align tool.

TABLE 2 – Genotype composition for the NS5B and Core regions (Patient 1).

Year	Number of reads	% Subtype		% Nucleotide similarity (Var)**			
		1a	**1b**	**1a**		**1b**	
NS5B							
2005	194	--	100	--	--	96.63	(0.30)
2007*	152	--	98.02	--	--	95.80	(1.05)
2008	93	--	100	--	--	96.11	(0.32)
2010*	683	18.44	81.25	96.20	(2.70)	94.96	(1.63)
2014*	420	18.33	81.4	96.22	(2.85)	95.46	(1.22)
Core							
2005	525	--	100	--	--	98.84	(0.41)
2007	494	--	100	--	--	99.01	(0.18)
2008	116	--	100	--	--	98.45	(0.56)
2010*	699	18.16	81.54	96.22	(2.36)	95.17	(1.76)
2014	397	--	100	--	--	98.04	(0.46)

* This sample had unclassified genotypes (<5%) not shown.
** Similarity score obtained from BLAST in HCV align tool

TABLE 3 – Genotype composition for the NS5B and Core regions (patient 2).

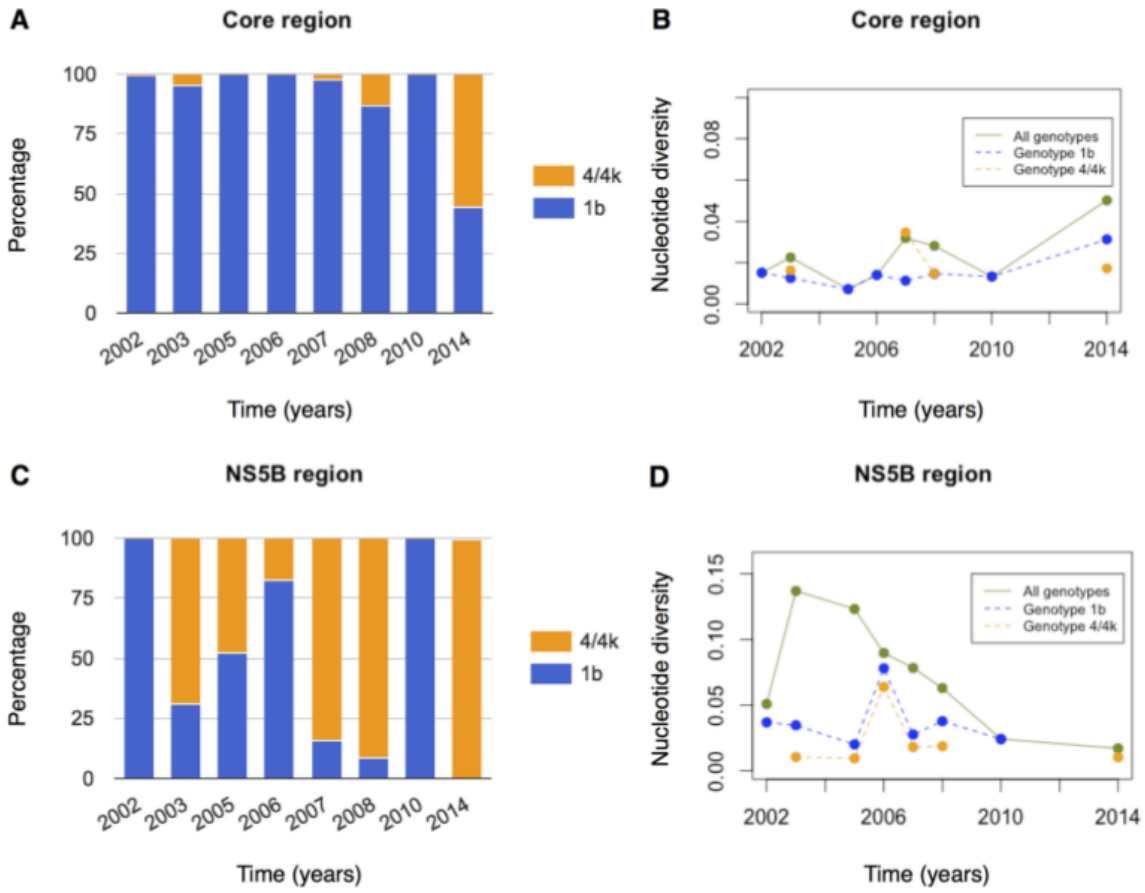


FIGURE 1 – Sample composition and genetic heterogeneity for patient 1.
(A-B) Core region : Percentage of genotypes 4 and 4k and nucleotide diversity estimates for each time sample. (C-D) NS5B region : Percentage of genotypes 4 and 4k and nucleotide diversity estimates for each time sample.

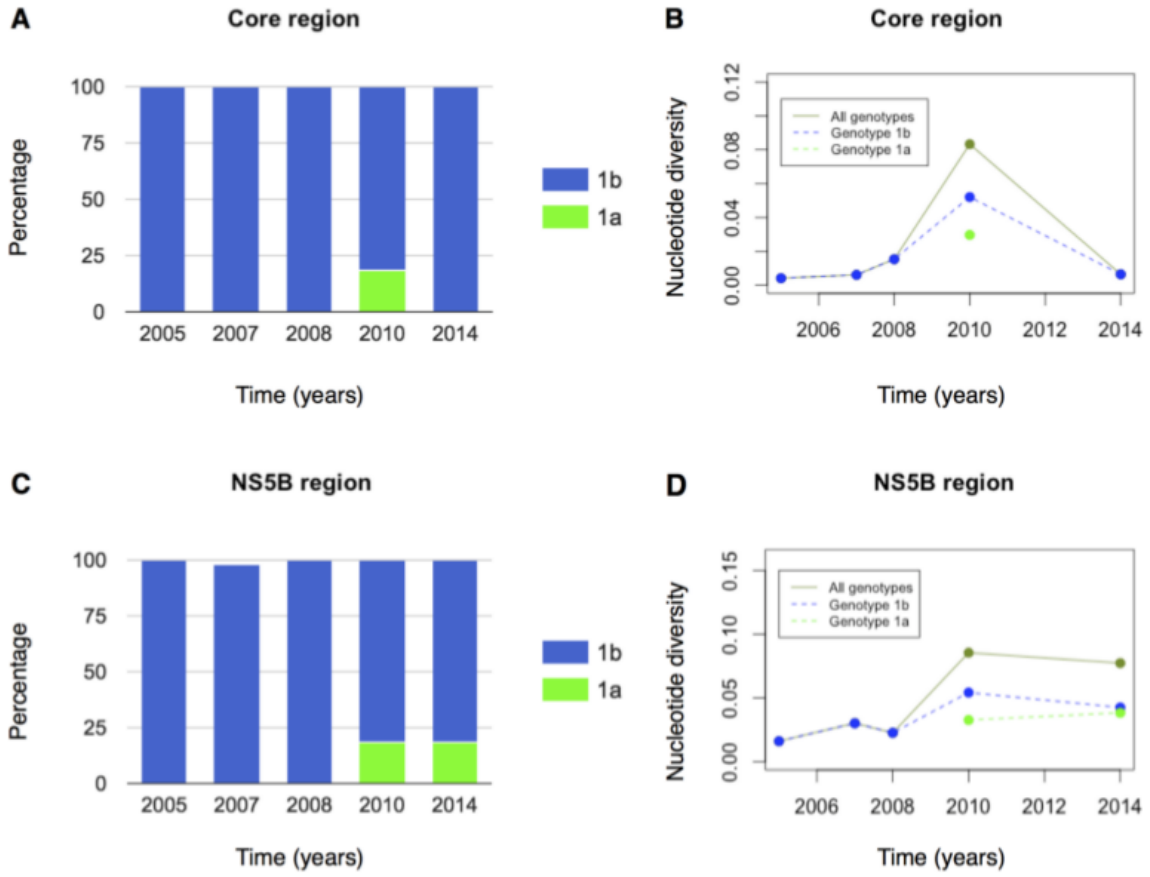


FIGURE 2 – Sample composition and genetic heterogeneity for patient 2.
(A-B) Core region : Percentage of genotype 1b and nucleotide diversity estimates for each time sample. (C-D) NS5B region : Percentage of genotype 1b and nucleotide diversity estimates for each time sample.

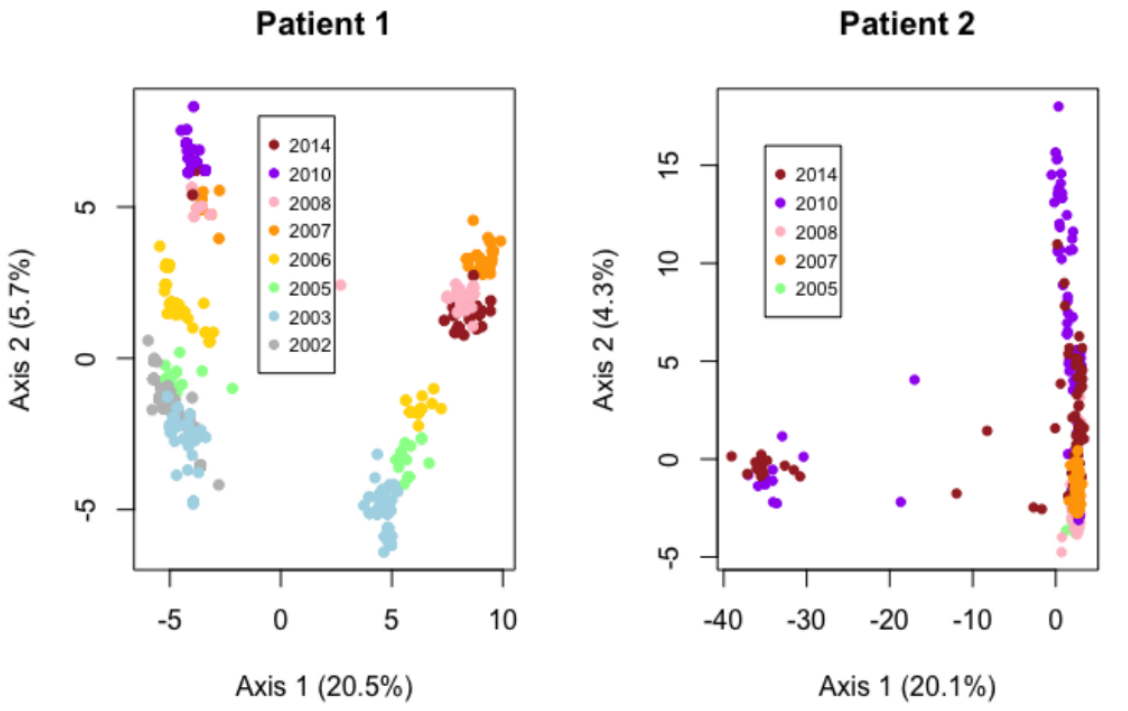


FIGURE 3 – Principal component plots for temporal samples from the NS5B region. Patient 1 : The left cluster consists of genotype 1 particles, whereas the right cluster consists of genotype 4 particles. Patient 2 : Subtype 1b particles cluster to the right, whereas subtype 1a particles correspond to the left part of the figure.

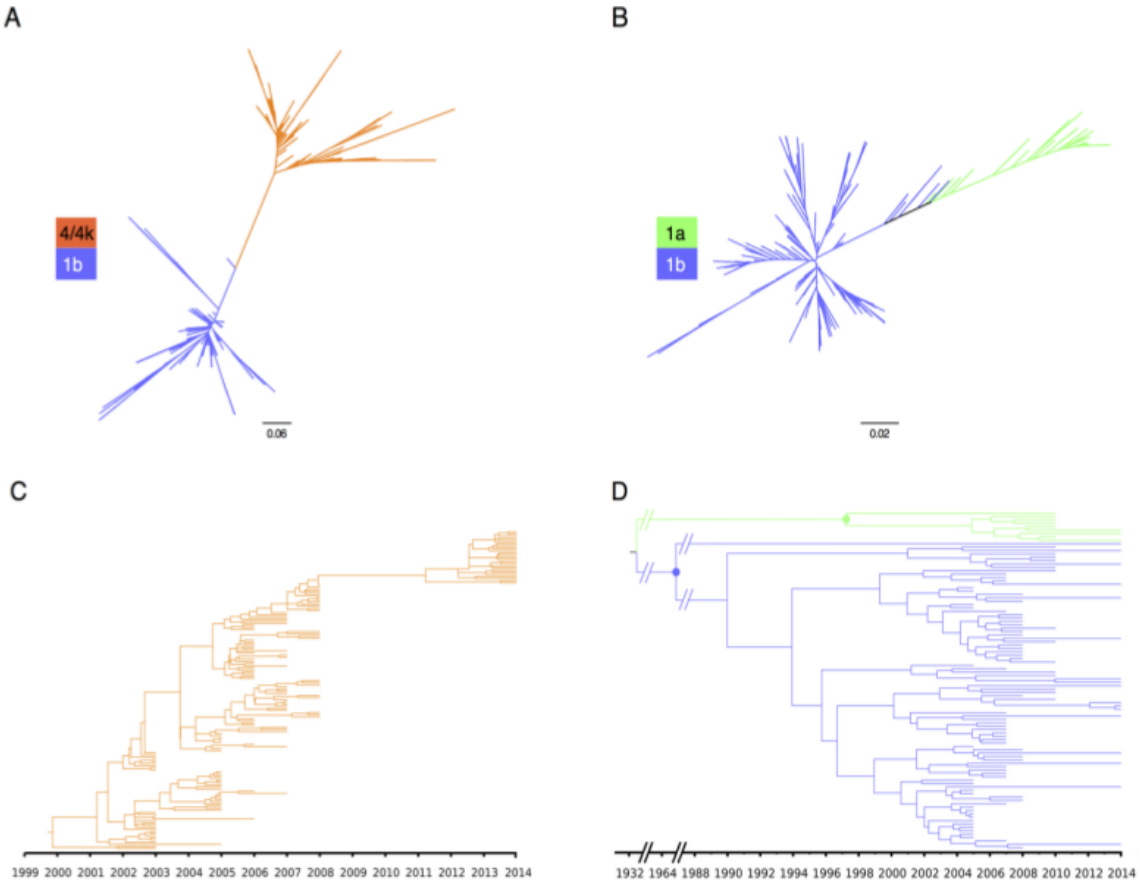


FIGURE 4 – Phylogenetic trees for the NS5B region.
(A) Unrooted NJ tree reconstructed from all sequences from patient 1. (B) Unrooted NJ tree reconstructed from all sequences from patient 2. (C) Maximum clade credibility trees obtained from BEAST by randomly subsampling twenty genotype 4 sequences from 8 time points (patient 1). (D) Maximum clade credibility trees obtained from BEAST by randomly subsampling 25 sequences from 5 time points (patient 2). Dates for the most recent ancestor of subtype 1a (green) and 1b (blue) were shown as dots. Long internal tree branches are represented by '//' symbols.

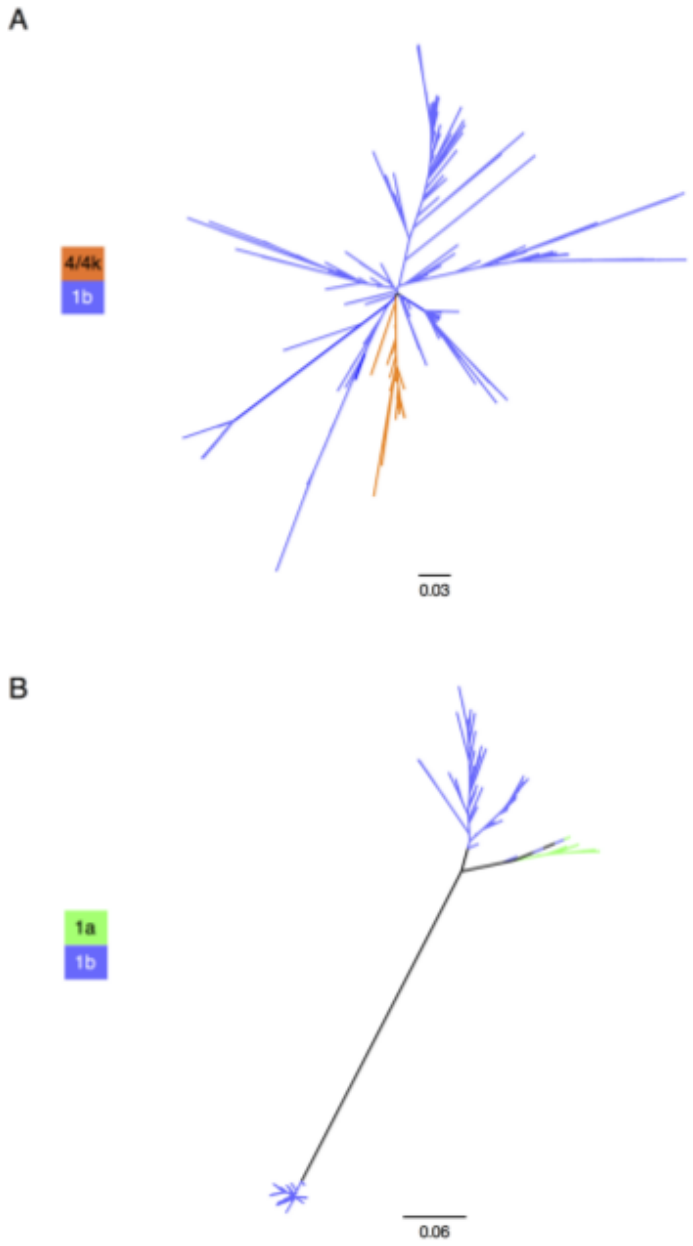


FIGURE S1 – Phylogenetic trees for the Core region.
(A) Unrooted NJ trees for patient 1 and (B) patient 2 with clustering by genotype/subtype seen in both patients.

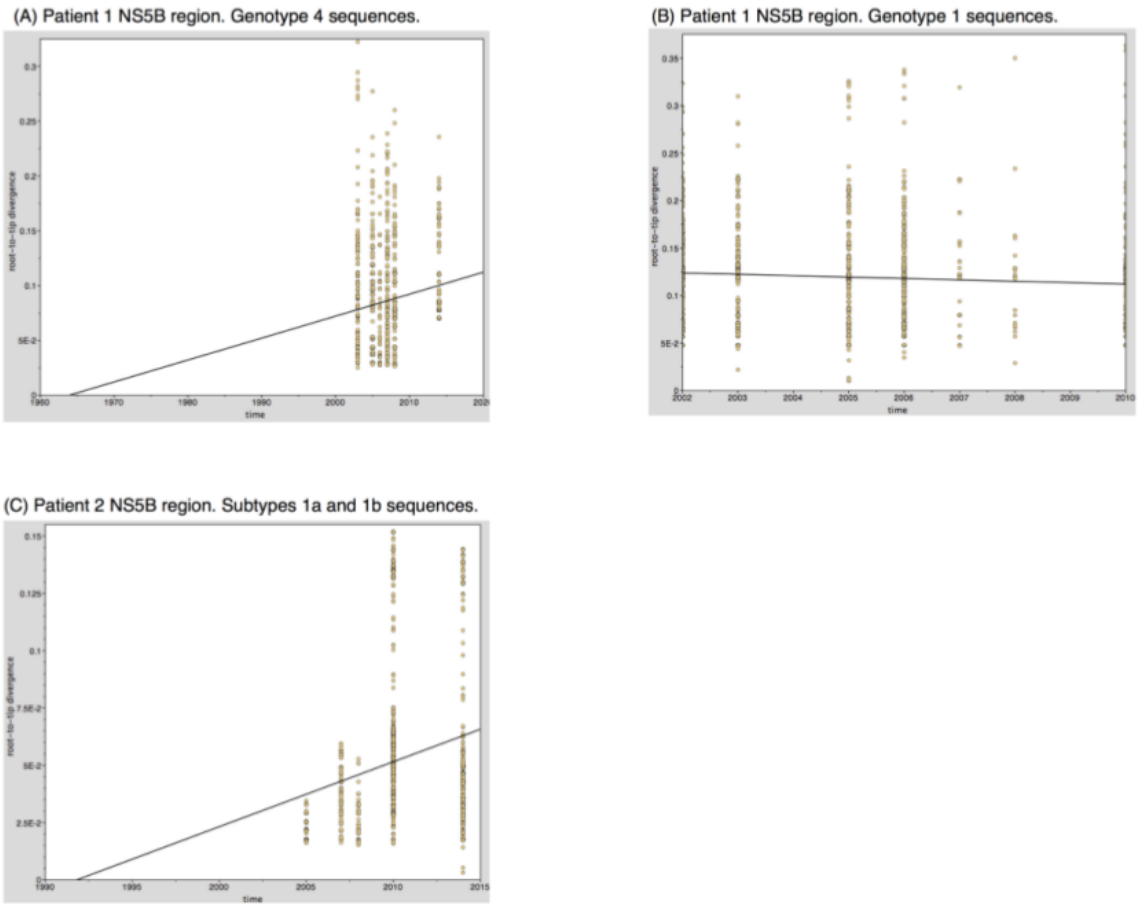


FIGURE S2 – TempEst outputs for various patients and genetic regions.
Cases from patient 1 (A) and 2 (C) with a positive slope were considered for building MCC trees after BEAST analysis, while patient 1 genotype 1b only (B) were found to be unsuitable for further BEAST analysis.

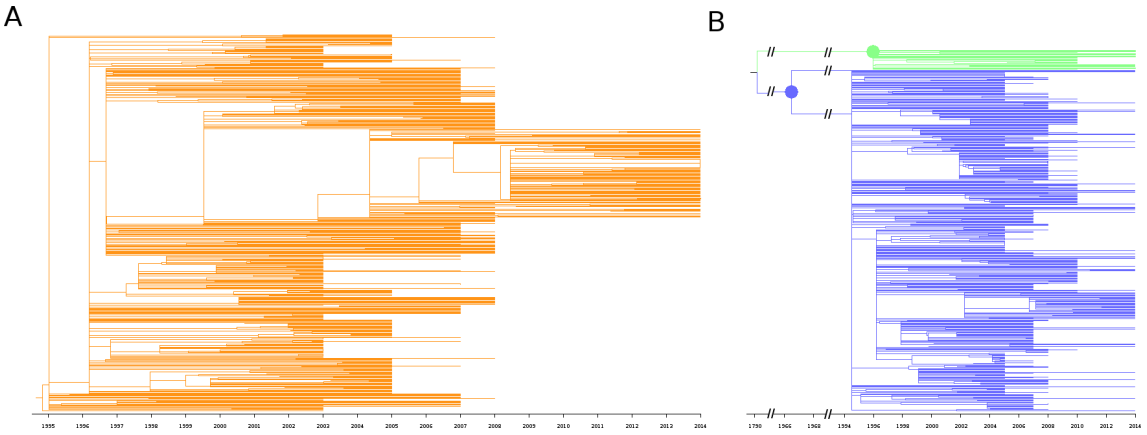


FIGURE S3 – Phylogenetic trees for the NS5B region.
(A) Molecular-clock analysis using LSD to estimate the date of the most recent ancestor of the genotype 4 strains (patient 1). (B) Molecular-clock analysis using LSD to estimate the dates of the most recent ancestors of the subtypes 1a (green dot) and 1b (blue dot) (patient 2).

THESE SOUTENUE PAR : Alban CAPOROSSI

TITRE : Dynamique de l'évolution intra-patient du VHC par NGS

CONCLUSION

Le séquençage haut débit des populations à évolution rapide du virus de l'hépatite C (VHC) permet des études à grande échelle de l'hétérogénéité virale intra-hôte et de la résistance aux schémas de traitement anti-viral communément observée dans les cas cliniques d'infection chronique. Alors que l'échappement au traitement anti-viral direct peut être expliqué par des mutations de résistance dans des régions génomiques ciblées, l'absence de réponse à la bithérapie (interféron pegylé et ribavirine) reste moins comprise. Dans cette étude, nous avons réalisé un séquençage par amplicon pour étudier la variation génomique des régions Core et NS5B du VHC sur une période de 13 ans pour deux patients suivis pour hépatite C chronique à l'Hôpital Universitaire Grenoble Alpes. A partir d'échantillons obtenus à différents points temporels, nous avons observé chez ces patients des infections mixtes composées de plusieurs génotypes du VHC. L'hétérogénéité génétique et l'analyse de la composition des échantillons a fourni des informations sur les changements dans la population virale au cours du traitement, avec NS5B qui a connu une forte augmentation de diversité après l'initiation du traitement comparativement à sa diversité pré-traitement. Des données supportant une structure génétique de population du VHC ont été observées chez tous les patients, faisant apparaître des lignées divergentes dans les arbres phylogénétiques. Ces observations orientent vers une sélection diversifiante se produisant après traitement, agissant à variation génomique constante et maintenant une grande hétérogénéité génétique durant l'infection. Associés avec l'efficacité du traitement, ces résultats fournissent la première preuve à l'occasion d'un traitement anti-viral d'induction de «réduction de variation sélective douce» («soft selective sweeps») chez des patients infectés chroniques avec plusieurs génotypes du VHC.

VU ET PERMIS D'IMPRIMER
Grenoble, le 12/10/17

LE DOYEN

Pour la Présidente
et par délégation
Le Doyen de la Médecine
Pr. Jean-Paul ROMANET

PROFESSEUR J.P. ROMANET

LE PRESIDENT DE LA THESE

MOREAU-GAUDRY

PROFESSEUR A. MOREAU-GAUDRY

BIBLIOGRAPHIE

[Abdelrahman et al., 2014] Abdelrahman, T., Hughes, J., Main, J., et al. (2014). **Next-generation sequencing sheds light on the natural history of hepatitis c infection in patients who fail treatment**. *Hepatology*, 61(1) :88–97.

[Ahmed et al., 2015] Ahmed, A., et Felmlee, D. (2015). **Mechanisms of hepatitis c viral resistance to direct acting antivirals**. *Viruses*, 7(12) :6716–6729.

[Beerenwinkel et al., 2012] Beerenwinkel, N., Günthard, H. F., Roth, V., et al. (2012). **Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data**. *Frontiers in Microbiology*, 3.

[Blach et al., 2017] Blach, S., Zeuzem, S., Manns, M., et al. (2017). **Global prevalence and genotype distribution of hepatitis c virus infection in 2015 : a modelling study**. *The Lancet Gastroenterology & Hepatology*, 2(3) :161–176.

[Buti et al., 2016] Buti, M., et Esteban, R. (2016). **Management of direct antiviral agent failures**. *Clinical and Molecular Hepatology*, 22(4) :432–438.

[Crotty et al., 2001] Crotty, S., Cameron, C. E., et Andino, R. (2001). **RNA virus error catastrophe : Direct molecular test by using ribavirin**. *Proceedings of the National Academy of Sciences*, 98(12) :6895–6900.

[Cuevas et al., 2009] Cuevas, J. M., Gonzalez-Candelas, F., Moya, A., et al. (2009). **Effect of ribavirin on the mutation rate and spectrum of hepatitis c virus in vivo**. *Journal of Virology*, 83(11) :5760–5764.

[Deléage et al., 2013] Deléage, G., et Gouy, M. (2013). **Bioinformatique - Cours et cas pratique**. Dunod.

[Dietz et al., 2013] Dietz, J., Schelhorn, S.-E., Fitting, D., et al. (2013). **Deep sequencing reveals mutagenic effects of ribavirin during monotherapy of hepatitis c virus genotype 1-infected patients**. *Journal of Virology*, 87(11) :6172–6181.

[Domingo et al., 2015] Domingo, E., et Schuster, P. (2015). **What is a quasispecies ? historical origins and current scope**. Dans *Current Topics in Microbiology and Immunology*, pages 1–22. Springer International Publishing.

[Donaldson et al., 2014] Donaldson, E. F., Harrington, P. R., O'Rear, J. J., et al. (2014). **Clinical evidence and bioinformatics characterization of potential hepatitis c virus resistance pathways for sofosbuvir**. *Hepatology*, 61(1) :56–65.

[Eigen, 1993] Eigen, M. (1993). **Viral quasispecies**. *Scientific American*, 269(1) :42–9.

[Eigen et al., 1977] Eigen, M., et Schuster, P. (1977). **The hypercycle. a principle of natural self-organization. part a : Emergence of the hypercycle**. *Die Naturwissenschaften*, 64(11) :541–565.

[Feigelstock et al., 2011] Feigelstock, D. A., Mihalik, K. B., et Feinstone, S. M. (2011). **Selection of hepatitis c virus resistant to ribavirin**. *Virology Journal*, 8(1) :402.

[James et al., 2013] James, G., Witten, D., Hastie, T., et al. (2013). **An Introduction to Statistical Learning**. Springer New York.

[Jünemann et al., 2013] Jünemann, S., Sedlazeck, F. J., Prior, K., et al. (2013). **Updating benchtop sequencing performance comparison**. *Nature Biotechnology*, 31(4) :294–296.

[Larrat et al., 2014] Larrat, S., Kulkarni, O., Claude, J.-B., et al. (2014). **Ultradeep pyrosequencing of NS3 to predict response to triple therapy with protease inhibitors in previously treated chronic hepatitis c patients**. *Journal of Clinical Microbiology*, 53(2) :389–397.

[Lauck et al., 2012] Lauck, M., Alvarado-Mora, M. V., Becker, E. A., et al. (2012). **Analysis of hepatitis c virus intrahost diversity across the coding region by ultradeep pyrosequencing**. *Journal of Virology*, 86(7) :3952–3960.

[Li et al., 2011] Li, H., Hughes, A. L., Bano, N., et al. (2011). **Genetic diversity of near genome-wide hepatitis c virus sequences during chronic infection : Evidence for protein structural conservation over time**. *PLoS ONE*, 6(5) :e19562.

[Manns et al., 2017] Manns, M. P., Buti, M., Gane, E., et al. (2017). **Hepatitis c virus infection**. *Nature Reviews Disease Primers*, 3 :17006.

[Nederbragt, 2016] Nederbragt, L. (2016). **developments in ngs**.

[Nei et al., 1979] Nei, M., et Li, W. H. (1979). **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proceedings of the National Academy of Sciences*, 76(10) :5269–5273.

[Nelson et al., 2015] Nelson, C. W., et Hughes, A. L. (2015). **Within-host nucleotide diversity of virus populations : Insights from next-generation sequencing**. *Infection, Genetics and Evolution*, 30 :1–7.

[Nowak, 1992] Nowak, M. A. (1992). **What is a quasispecies ?** *Trends in Ecology & Evolution*, 7(4) :118–121.

[Ortega-Prieto et al., 2013] Ortega-Prieto, A. M., Sheldon, J., Grande-Pérez, A., et al. (2013). **Extinction of hepatitis c virus by ribavirin in hepatoma cells involves lethal mutagenesis**. *PLoS ONE*, 8(8) :e71039.

[Pawlotsky, 2016] Pawlotsky, J.-M. (2016). **Hepatitis c virus resistance to direct-acting antiviral drugs in interferon-free regimens**. *Gastroenterology*, 151(1) :70–86.

[Perry, 2012] Perry, E. (2012). **Ebi : Next generation sequencing practical course**.

[Ramachandran et al., 2011] Ramachandran, S., Campo, D. S., Dimitrova, Z. E., et al. (2011). **Temporal variations in the hepatitis c virus intrahost population during chronic infection**. *Journal of Virology*, 85(13) :6369–6380.

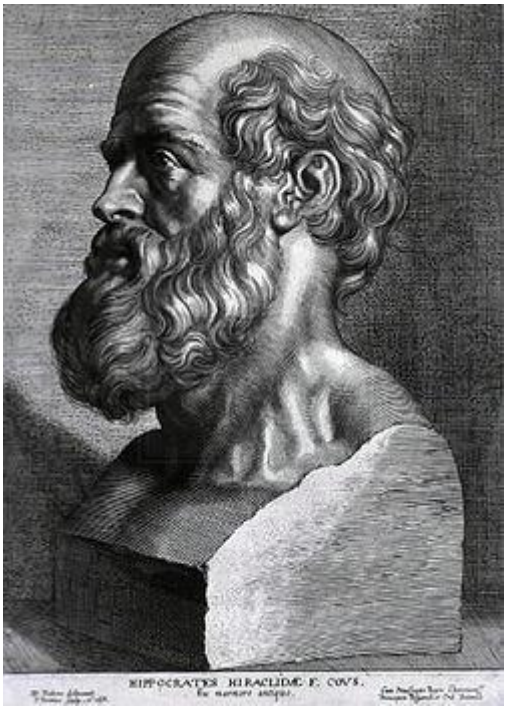
[To et al., 2015] To, T.-H., Jung, M., Lycett, S., et al. (2015). **Fast dating using least-squares criteria and algorithms**. *Systematic Biology*, 65(1) :82–97.

[Voelkerding et al., 2009] Voelkerding, K. V., Dames, S. A., et Durtschi, J. D. (2009). **Next-generation sequencing : From basic research to diagnostics**. *Clinical Chemistry*, 55(4) :641–658.

[Wakeley, 2009] Wakeley, J. (2009). **Coalescent Theory - An Introduction**. Macmillan Learning.

[Wang et al., 2010] Wang, G. P., Sherrill-Mix, S. A., Chang, K. M., et al. (2010). **Hepatitis c virus transmission bottlenecks analyzed by deep sequencing**. *Journal of Virology*, 84(12) :6218–6228.

[Yang, 1996] Yang, Z. (1996). **Among-site rate variation and its impact on phylogenetic analyses**. *Trends in Ecology & Evolution*, 11(9) :367–372.



SERMENT D'HIPPOCRATE

En présence des Maîtres de cette Faculté, de mes chers condisciples et devant l'effigie d'HIPPOCRATE,

Je promets et je jure d'être fidèle aux lois de l'honneur et de la probité dans l'exercice de la Médecine.

Je donnerais mes soins gratuitement à l'indigent et n'exigerai jamais un salaire au dessus de mon travail. Je ne participerai à aucun partage clandestin d'honoraires.

Admis dans l'intimité des maisons, mes yeux n'y verront pas ce qui s'y passe ; ma langue taira les secrets qui me seront confiés et mon état ne servira pas à corrompre les mœurs, ni à favoriser le crime.

Je ne permettrai pas que des considérations de religion, de nation, de race, de parti ou de classe sociale viennent s'interposer entre mon devoir et mon patient.

Je garderai le respect absolu de la vie humaine.

Même sous la menace, je n'admettrai pas de faire usage de mes connaissances médicales contre les lois de l'humanité.

Respectueux et reconnaissant envers mes Maîtres, je rendrai à leurs enfants l'instruction que j'ai reçue de leurs pères.

Que les hommes m'accordent leur estime si je suis fidèle à mes promesses.

Que je sois couvert d'opprobre et méprisé de mes confrères si j'y manque.

Résumé :

Le séquençage haut débit des populations à évolution rapide du virus de l'hépatite C (VHC) permet des études à grande échelle de l'hétérogénéité virale intra-hôte et de la résistance aux schémas de traitement anti-viral communément observée dans les cas cliniques d'infection chronique. Alors que l'échappement au traitement anti-viral direct peut être expliqué par des mutations de résistance dans des régions génomiques ciblées, l'absence de réponse à la bithérapie (interféron pegylé et ribavirine) reste moins comprise. Dans cette étude, nous avons réalisé un séquençage par amplicon pour étudier la variation génomique des régions Core et NS5B du VHC sur une période de 13 ans pour deux patients suivis pour hépatite C chronique à l'Hôpital Universitaire Grenoble Alpes. A partir d'échantillons obtenus à différents points temporels, nous avons observé chez ces patients des infections mixtes composées de plusieurs génotypes du VHC. L'hétérogénéité génétique et l'analyse de la composition des échantillons a fourni des informations sur les changements dans la population virale au cours du traitement, avec NS5B qui a connu une forte augmentation de diversité après l'initiation du traitement comparativement à sa diversité pré-traitement. Des données supportant une structure génétique de population du VHC ont été observées chez tous les patients, faisant apparaître des lignées divergentes dans les arbres phylogénétiques. Ces observations orientent vers une sélection diversifiante se produisant après traitement, agissant à variation génomique constante et maintenant une grande hétérogénéité génétique durant l'infection. Associés avec l'efficacité du traitement, ces résultats fournissent la première preuve à l'occasion d'un traitement anti-viral d'induction de «réduction de variation sélective douce» («*soft selective sweeps*») chez des patients infectés chroniques avec plusieurs génotypes du VHC.

Mots-clés : Virus de l'Hépatite C, Infection multiple, séquençage haut débit, Etude longitudinale, Diversité/Hétérogénéité génétique, Selective sweeps

UFR de Médecine

UNIVERSITÉ Grenoble Alpes