C 型肝炎基因序列片段間連動關係之研究

The study of relationship of amino acid sequences and nucleotide sequences between HCV strains

孫光天	施秀娟	楊孔嘉	鄭婷方
Koun-Tem Sun	Shiou-Jiuan Shr	Kung-Chia Young	Ting-Fang cheng
國立臺南大學 數位學習科技系	國立臺南大學 數位學習科技系	國立成功大學 醫學院 醫技系	國立成功大學 醫學院 分子醫學所
ktsun@mail.nutn.edu.tw	*dtg@ms85.url.com.tw*	*t7908077@mail.ncku.edu.tw*	*t1695406@mai.ncku.edu.tw*

摘要

C 型肝炎病毒(HCV)是造成輸血後肝炎的主要病原，其感染分佈現已成全球化趨勢。疫苗的發展亦是困難重重，導因於病毒基因的突變率太高，甚至無法保護再次受到同一病毒感染。目前認為病毒的複製需要其他細胞及病毒因子共同協助，故各片段在 HCV 病毒複製上扮演什麼角色，是 C 型肝炎相關研究者積極探討之問題。本研究由「Los Alamos HCV sequence database」所提供的 225 株全長 C 型肝炎病毒序列資料。以資料探勘技術-關連法則演算法，來探討轉譯區中 NS2、NS3、NS5B 胺基酸片段與非轉譯區中 3'UTR、5'UTR 核苷酸片段在病毒的複製中是否存在協同變異的連動關係。透過關連法則所探勘出的連動關係，與近幾年來文獻所指出與 HCV 病毒複製相關的區域或位點對應，皆有相同結果，故本研究之方法，對未來尋求片段間連動關係，提供一快速分析技術。

關鍵字： C 型肝炎、資料探勘、關連法則

Abstract

A major etiological agent of post-transfusion hepatitis is caused by HCV, which has been widespread globally. Vaccine development has been hampered by a high degree of antigenic variation and a lack of protection against viral reinfection. At present, some researches on the virus replication consider some cells and virus factors assisting together. What role does it play on each virus replication in HCV is the most important issue. In our research, we apply data mining technology-「Association rule」 to find the covariation relationship between coding (NS2、NS3、NS5B) and non-coding (3'UTR、5'UTR) region from 225 full-length hepatitis C virus sequences which come from 「Los Alamos HCV sequence database」.We locate some covariation relationships from association rule technology, and they are corresponded to the regions about HCV replications which we survey from literatures. In a summary, the contribution in our research is that we design a fast analytical technology for finding some covariation relationship between coding and non-coding regions

Keywords: HCV、Data mining 、Association rule

1、前言

C 型肝炎病毒學的分類是被歸類在 Flaviviridae 科（family）之 hepacivirus 屬（genus），為一具有套模的 RNA 病毒。其基因體由一條單股正向的 RNA 組成，全長約 9.6 kb，包含單一個開放讀架，可產生至少十個結構與非結構性病毒蛋白質。在 C 型肝炎病毒中，某個基因片段的某個核苷酸發生突變，不只影響此基因的功能，也連帶對其他相關基因的功能造成互補的改變。傳統都是利用選殖方式透過病毒細胞培養進行實驗，但很少文獻中探討利用資料探勘技術來挖掘基因序列位置變異之關聯性。

目前資料探勘技術已被廣泛應用於企業界，且亦有應用於生物醫學領域上的許多實例，這其中包括疾病的

診斷或是預後的評估。資料探勘可從大量的資料中，發掘潛在有用資訊，得到事前從未知曉的重要知識、規則，能夠輔助決策人員作決策時重要的參考。資料探勘技術中最為成熟和利用最多就是關連法則(association rule)[10]，它是使用既有的資料建立關連法則並找出在資料集之中一起頻繁出現的屬性值(attribute-value)；在本研究中，應用關連分析探勘的理論與技術，挖掘出序列中的頻繁集合（例如使用Apriori）。這種觀點在資料探勘界被廣泛接受[4][16]。

目前對於感染 C 型肝炎的患者並沒有可以徹底治療的方法或藥物，對於病毒複製的機轉也不夠清楚。目前認為病毒的複製需要其他細胞及病毒因子共同協助。在分析 C 型肝炎病毒複製過程中，因受限於個體間差異以及可供實驗的個數，而往往無法得到代表性的結果[12][17]。此特點可透過大量已知的 HCV 基因體序列資料，利用關連法則加以克服。關連法則在生物醫學領域的應用，而且都有很高的預測正確率。若推廣至 C 型肝炎生物醫學領域的應用，可建立出病毒基因體結構的「關連法則」或「頻繁集合」，作為生物學家研究 C 型肝炎病毒轉譯與複製的考參。

因此在本研究中嘗試運用資料探勘技術，以關連法則演算法，透過大量已知 HCV 基因體的序列資料，來分析和探討轉譯區與非轉譯區在病毒的複製中是否存在協同變異的連動關係。

2、研究方法

2.1 資料來源

本研究相關資料來源，由「Los Alamos HCV database」(http://hcv.lanl.gov/content/hcv-db/index)，所提供之完整C型肝炎病毒序列(全長 9000 以上)共計 225 筆（可分為 1a、1b、2a、2b四個亞型），依研究需取拮取C型肝炎病毒核苷酸序列的 3’UTR、5’UTR等片段及蛋白質序列的NS2、NS3、NS5 等片段。本研究資料處理分成三部份進行，如下：

第一部份為跨亞型(Inter-subtype)，依 225 株病毒序列樣本中的蛋白質序列(NS2、NS3、NS5)與核苷酸序列(3’UTR、5’UTR)，找出共同突變特徵集合規則。

第二部份找出亞型內(Intra-subtype)，各亞型群中之各別的共同突變特徵集合規則，依亞型內(Intra-subtype)區分符合條件者共 4 型，其中 1a 有 27 筆，1b 有 126 筆，2a 有 30 筆，2b 有 23 筆，其餘零散者為 28 筆。

第三部份找出亞型內(Intra-subtype)中以標準株為擷取特徵標準，是否有與標準株共同突變特徵不同之集合規則。本研究將分析的資料做前置處理，主要是為了平衡標準株和非標準株的有效樣本比例，進而製作非標準株與標準株之樣本群。依亞型區分為四群，分別為:1a、1b、2a、2b，針對 1:1 方式進行樣本數放大，以非標準株樣本數為基礎，標準株數本放大為非標準株樣本的倍數為 1 倍。

2.2 資料前處理

由於標準株為該子型被共認為最初的病毒序列，我們希望能找到與標準株不同之改變位點，所以需要作第三部份－“找出亞型內(Intra-subtype)中以標準株做為擷取特徵之標準，是否有與標準株共同突變特徵不同之集合規則”，在此部份我們希望能凸顯出與標準株的不同之處，所以需要將標準株的實驗數量放大，讓實驗過程不會因為標準株的數量過於小，而找不到與標準株不同的部份，本研究採用（標準株：非標準株）1：1 的放大，以求實驗結果不會過於偏向標準株或非標準株的改變情形。

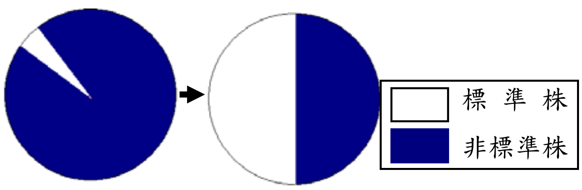

圖 1：標準株放大示意圖

為了避免找到沒有意義的法則(指不會變動的片段或位點)，所以在找關聯法則之前，我們先用 GeneDoc 軟體去除相同的部份，當標準株放大成“標準株：非標準株＝1：1”時，當標準株與其他序列不同時，只會去除標準株的部份，非標準株的部份會全部保留下來，所以就可以找到與標準株不同的位點。

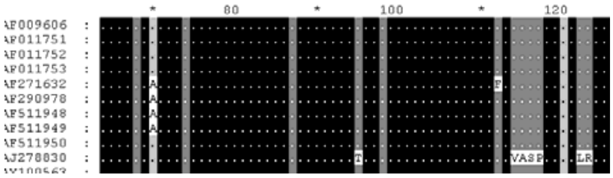

圖 2：標準株放大前

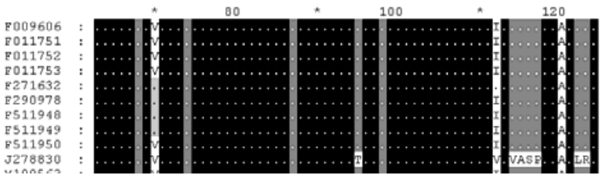

圖 3：標準株放大後

由圖 2 與圖 3 可知當標準株放大前，122 這個位點為”A”的序列為大多數，所以被 GeneDoc 去除了，但當標準株放大時，由於 122 位點，標準株為”I”，所以其他為”A”的序列，則可保留下來，作為找尋關聯法則的位點集合。

2.3、關連規則

在資料探勘的方法當中，關聯法則挖掘是最普遍被使用的方法之一。其中關聯規則(Association Rules) 最早由 Agrawal, Imieliński, and Swami 於 1993 年提出，而關聯法則(Association Rules)中最具代表性的方法是由 Agrawal 等學者在 1994 年所提出的 Apriori 演算法，許多推導關聯法則技術的相關演算法，都是以 Apriori 為基礎加以改良或延伸。此演算法主要是從資料庫中各個交易(transactions)的項目(items)中找出最大項目集合(large itemset)，然後依據所得結果中之所有最大項目集合，再逐一產生關聯規則，步驟說明如下：

計算資料庫中所有單一項目集合(1-itemsets)的個數並找出符合最小支持值者，此結果即是長度為 1 的最大項目集合(large 1-itemsets)。

由長度為 1 的最大項目集合產生長度為 2 的候選項目集合(candidate2-itemsets)，然後計算並決定長度為 2 的最大項目集合(large2-itemsets)。

從所有長度為 2 的最大項目集合中，產生符合最小信賴度的關聯規則。

重複步驟二、步驟三，依序產生所有長度為 k 的最大項目集合(large k-itemsets)，直到無法再產生更大項目集合為止。

其中最小支持值(minimum support)是指項目集合(itemsets)在資料庫中出現的比例，亦可稱為出現頻率，所有產生出來的最大項目集(Large Itemset)皆須滿足此最小限制值。最小信賴值(minimum confidence)是指兩個項目集合間的信賴程度，例如關聯規則 X => Y，則其最小信賴度代表在 X 出現下，出現 Y 之機率，所有從最大項目集合產生出來的關聯規則皆須滿足此最小限制值。

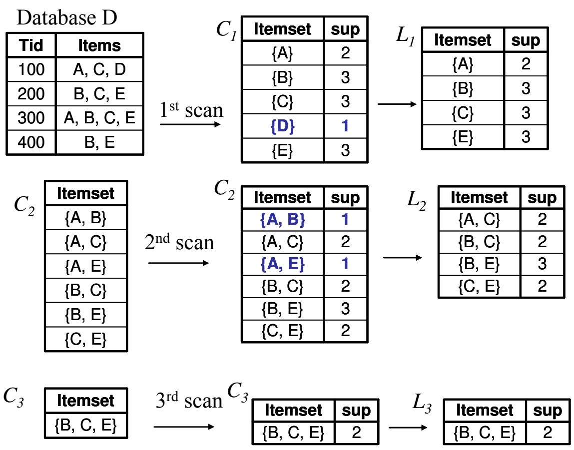

圖 4：產生頻繁項目集及候選項目集

2.4、運作流程

Nowak(1995)提到後基因體時代(Post-Genomics Era)，最需要思考的重要問題是如何從大量的序列資料發掘出其中隱含的資訊。因此在複雜的病毒序列中，欲尋找基因變化之相關性規則，必須透過資訊技術來分析這些序列資料，方能從中獲得有意義的資訊及知識。本研究將應用生物資訊軟體工具以及專為基因序列位置的相關性研究，所設計之「基因序列格式轉換系統」，並且透過 Weka (Waikato Environment for Knowledge Analysis)資料探勘工具中的 Apriori 演算法則來挖掘出病毒序列突變位置相關性之頻繁集合，分析探勘出的集合在 C 型肝炎病毒複製上扮演的角色。本研究探勘之運作流程如下圖所示：

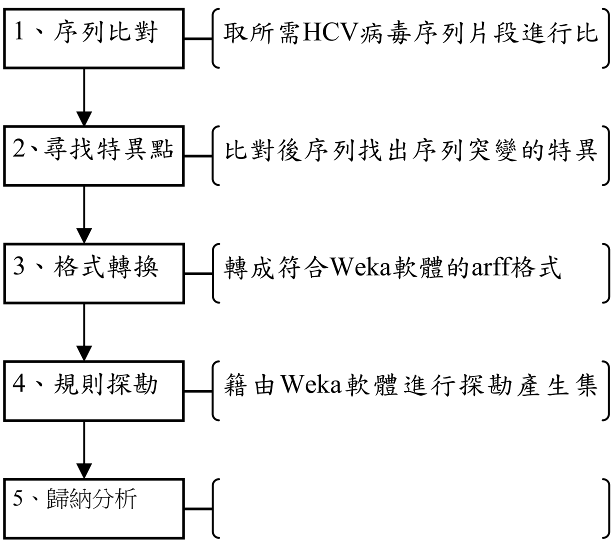

圖 5：HCV 序列探勘之運作流程圖

首先，將序列總長度大於 9000 以上的 225 株 HCV 病毒序列進行排序定位。經由此比對過程，我們可以找出序列間的共同區域，並同時可以辨別出序列間的差異。方法是先將序列資料轉換為 Fasta 格式，再透過 Clustal X 軟體進行比對。

第二步驟，將比對後序列，透過多序列編輯分析軟體 Gene Doc 進行編輯，將各序列中特異點的部分挑選出來，並刪除所有序列相同的部分，大部分相同時，則相同的部分以點的方式顯示。

第三步驟，經過上述處理程序之後，將此資料以 CSV 檔案格式透過專為基因序列位置相關性之研究，所設計之「基因序列格式轉換系統」，轉換為便於探勘技術進行之資料格式。

第四步驟，將轉換完成的 arff 格式檔案的資料，透過 Weka 軟體進行規則探勘。選擇 Apriori 關聯演算法，設定關聯分析相關數值後，進行關聯規則探勘。

最後，利用本研究撰寫的「分析連動關係相關性的系統」，可輕易的找出探勘出的結果在原始樣本中的樣貌。且可查詢欲知連動關係所符合的代碼，及位點左右的資訊以利分析，再透過文獻所指出與 HCV 病毒複製相關的區域或位點，加以分析其扮演的角色為何。

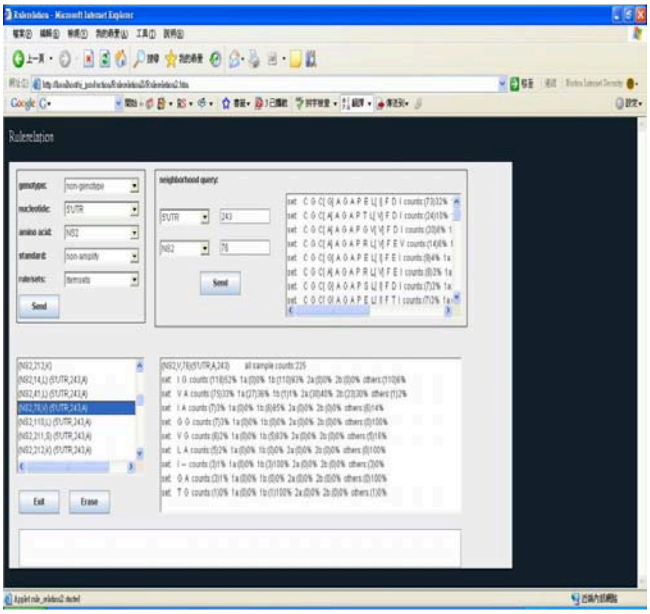

圖 6：分析連動關係相關性的系統畫面

3、結果與討論

3.1、C 型肝炎單點分析

本研究採用 225 株跨亞型的 C 型肝炎病毒株 (全長 9000 以上)，分別對其核苷酸序列的 3’UTR、5’UTR 等片段及蛋白質序列的 NS2、NS3、NS5 等片段做分析。使用 Weka 軟體探勘不同片段之間集合的相關性，再以文獻或臨床實驗中所指出與 HCV 病毒複製相關的區域或位點，加以分析其扮演的角色為何。

在單點的部分，我們參考文獻中發現，探勘出的位點與文獻研究指出與病毒複製相關突變位點相符，本技術有很高的正確性，結果如下表所示。

表 1：各基因片段單點突變文獻整理

基因型	突變位點	分析	參考文獻
5’UTR (1b)	204 A 243 A	在亞型 1b 中與結果完全符合	[13]
5’UTR (1a)	G107A C204A G243A	在跨亞型中與結果完全符合，在亞型 1a 中與結果 C204A, G243A 符合	[9]
5’UTR (1a,	G34A A35G	在跨亞型中與結果完全符合	[11]

1b)			
5’UTR (1b,2c, 3a)	C183U C204U G350A	在亞型 1b 中與結果 C183U,C204U 符合	[14]
5’UTR (1b,2c, 3a)	A215G C340U	在亞型 1b 中與結果 C340U 符合	[14]
5’UTR (1a)	C183A A214G	在亞型 1a 中與結果完全符合	[5]
NS3	F418Y	在跨亞型中與結果完全符合	[3]
NS3	S297P I71V	在跨亞型中與結果完全符合	[18]
NS3	Q86R	在跨亞型中與結果完全符合	[1]
NS3 (1b)	A358T I386L Y418F R470 M	在亞型 1b 中與結果完全符合	[7]
NS3 (1a)	E176G	在亞型 1a 中與結果完全符合	[15]
NS3 (1a)	S332P	在亞型 1a 中與結果完全符合	[18]
3’UTR	G9T C12G	在跨亞型中與結果完全符合	[6]

3.2、C 型肝炎連動關係分析

在連動關係的部分，雖然臨床與文獻中鮮少探討，但就機率學的角度來看，我們可推得在 C 型肝炎病毒序列中，最有可能突變的位點及其組合。本研究分跨亞型(Inter-subtype)、亞型內(Intra-subtype)中不以標準株為擷取特徵標準及亞型內(Intra-subtype)中以標準株為擷取特徵標準三個部分來探討，結果如圖 7 所示。

在跨亞型部分可得知 NS2 及 5’UTR 兩片段產生關連，在 NS2 位點 76 為 V，5’UTR 位點 243 為 A。此關連性的位置之表現為 C 型肝炎病毒複製的之特色，此推論之正確性必須於臨床實驗進一部加以驗證。

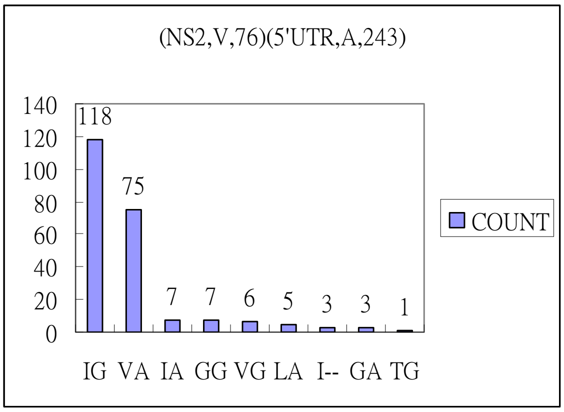

圖 6：跨亞型 NS2 位點 76 與 5’UTR 位點 243 關連

4、結論

本研究旨在探討運用資料探勘技術，以關連法則演算法，透過大量已知 HCV 基因體的序列資料，來分析和探討轉譯區中 NS2、NS3、NS5B 胺基酸片段與非轉譯區中 3’UTR、5’UTR 核苷酸片段在病毒的複製中是否存在共同協助的連動關係。故利用透過關連法則探勘出的連動關係，建立出病毒基因體結構的「關連法則」或「頻繁集合」，可作為生物學家研究 C 型肝炎病毒轉譯與複製的參考。

本研究在單點的部分，探討過去的文獻中發現，我們探勘出的位點與文獻研究指出與病毒複製相關突變位點相符，有很高的正確性。因此，透過本研究所設計的「分析連動關係相關性的系統」，可以快速找出位點或位點之間的所有鹼基組合，這對於「轉譯區與非轉譯區在病毒複製上是否有關連」、「病毒的複製需要那些病毒因子共同協助」、「建立代表性的結果」都是相當重要且有意義的參考分析資料，對生物資訊之處理，提供一有效工具。

參考文獻

[1] K. I. Abe, M. Ikeda, H. Dansako, K. Naka, and N. Kato, "Cell culture-adaptive NS3 mutations required for the robust replication of genome-length hepatitis C virus RNA," 2007.

[2] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Conference, vol. 22: ACM Press New York, NY, USA, 1993, pp.

207-216.

[3] J. Bukh, T. Pietschmann, V. Lohmann, N. Krieger, K. Faulk, R. E. Engle, S. Govindarajan, M. Shapiro, M. St Claire, and R. Bartenschlager, "Mutations that permit efficient replication of hepatitis C virus RNA in Huh-7 cells prevent productive replication in chimpanzees," vol. 99: National Acad Sciences, 2002, pp. 14416-14421.

[4] A. J. Collier, J. Gallego, R. Klinck, P. T. Cole, S. J. Harris, G. P. Harrison, F. Aboul-ela, G. Varani, and S. Walker, "A conserved RNA structure within the HCV IRES eIF3-binding site," Nature Structural Biology, vol. 9, pp. 375-380, 2002.

[5] C. Creighton and S. Hanash, "Mining gene expression databases for association rules," Bioinformatics, vol. 19, pp. 79-86, 2003.

[6] A. T. Gates, R. T. Sarisky, and B. Gu, "Sequence requirements for the development of a chimeric HCV replicon system," vol. 100, 2004, pp. 213-22.

[7] D. J. Graham, M. Stahlhut, O. Flores, D. B. Olsen, D. J. Hazuda, R. L. Lafemina, and S. W. Ludmerer, "A genotype 2b NS5B polymerase with novel substitutions supports replication of a chimeric HCV 1b: 2b replicon containing a genotype 1b NS3-5A background," 2005.

[8] J. A. Grobler, E. J. Markel, J. F. Fay, D. J. Graham, A. L. Simcoe, S. W. Ludmerer, E. M. Murray, G. Migliaccio, and O. A. Flores, "Identification of a Key Determinant of Hepatitis C Virus Cell Culture Adaptation in Domain II of NS3 Helicase," vol. 278: ASBMB, 2003, pp. 16741-16746.

[9] L. H, S. YK, and L. SM., " Cell type–specific enhancement of hepatitis C virus internal ribosome entry site-directed translation due to 5_ nontranslated region substitutions selected during passage of virus in lymphoblastoid cells. ," J Virol vol. 74, pp. 7024–7031, 2000.

[10] J. Han and M. Kamber, "Data Mining:Concepts and Techiniques.," USA: Morgan Kaufmann., 2001.

[11] M. Honda, R. Rijnbrand, G. Abell, D. Kim, and S. M. Lemon, "Natural Variation in Translational Activities of the 5' Nontranslated RNAs of Hepatitis C Virus Genotypes 1a and 1b: Evidence for a Long-Range RNA-RNA Interaction outside of the Internal Ribosomal Entry Site," J. Virol. , vol. 73, pp. 4941-4951,1999.

[12] A. A. Kolykhalov, E. V. Agapov, K. J. Blight, K. Mihalik, S. M. Feinstone, and C. M. Rice, "Transmission of Hepatitis C by Intrahepatic Inoculation with Transcribed RNA," Science, vol. 277, pp. 570, 1997.

[13] J. Laporte, C. Bain, P. Maurel, G. Inchauspe, H. Agut, and A. Cahour, "Differential distribution and internal translation efficiency of hepatitis C virus quasispecies present in dendritic and liver cells," Blood, vol. 101, pp. 52 - 57, 2003.

[14] T. MA, D. E, and H. MN, " Lack of clinical significance of variability in the internal ribosome entry site of hepatitis C virus. ," J Med Virol, vol. 72, pp. 396-405, 2004.

[15] S. Rosenberg, "Recent advances in the molecular biology of hepatitis C virus," vol. 313, 2001, pp. 451-64.

[16] S. Stilou, P. D. Bamidis, N. Maglaveras, and C. Pappas, "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare," Medinfo, vol. 10, pp. 399-403, 2001.

[17] M. Yanagi, R. H. Purcell, S. U. Emerson, and J. Bukh, "Transcripts from a single full-length cDNA clone of hepatitis C virus are infectious when directly transfected into the liver of a chimpanzee," National Acad Sciences, 1997.

[18] Q. Zhu, J. T. Guo, and C. Seeger, "Replication of Hepatitis C Virus Subgenomes in Nonhepatic Epithelial and Mouse Hepatoma Cells," Journal of Virology, vol. 77, pp. 9204-9210, 2003.