

中国科学技术大学

硕士学位论文



从局部到整体：一种新的 乙肝病毒分型框架

作者姓名: 周 鑫

学科专业: 生物信息学

导师姓名: 吴家睿 教授

梁治 研究员

完成时间: 二〇一三年十月

University of Science and Technology of China
A dissertation for master degree



From local to global:a new
perspective of Hepatitis B virus
Genotyping framework

Author : Xin Zhou
Speciality : Bioinformatics
Supervisor : Prof.Jiarui Wu
Dr. Zhi Liang
Finished Time : 10, 2013

从局部到整体：一种新的乙肝病毒分型框架

八系

周鑫

中国科学技术大学

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：_____ 签字日期：_____

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构递交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

公开 保密 _____ 年

作者签名：_____ 导师签名：_____

签字日期：_____ 签字日期：_____

摘要

在慢性乙型肝炎病毒患者的阿德福韦酯抗病毒治疗过程中，我们观察（数据来自重庆医科大学）到，38/200 的病人样本在为期一年的治疗之后出现基因型分布转化现象。同时，我们通过二代测序方法（包括 roche454 焦磷酸测序，solexa 测序）发现这些病人的抗药性增强与其血清中乙型肝炎病毒的亚型分布发生转换之间存在这关联。

我们从乙型肝炎病毒 NCBI 基因组集合出发，基于基因组集合上体现的各个位点之间的相关性，同时考虑到病毒基因组上大部分冗余保守位点对于病毒区分过程的可忽略性，使用独立成分分析等机器学习方法找到合适用于表征乙型肝炎病毒分类的特征片段，同时研究不同乙型肝炎病毒亚型之间的相互依赖情况。最后，完成了乙型肝炎病毒的基于短序列分型框架的建立。

然后，我们使用建立好的短序列分型框架对我们在合作中获得的病人样本进行分型工作，并且在使用短序列分型框架之前对获得的二代测序原始数据进行特定片段选择，不同样本的标签设计以及测血结果的噪音消除工作。最后通过我们设计的分型框架，我们进一步确认了病毒阿德福韦耐药性的成因以及环境压力和病毒亚型反转之间的相关性。

基于不同乙型肝炎病毒亚型的差异，对阿德福韦酯的敏感性分析，以及二代测序的高分辨率，我们给慢性乙型肝炎患者长期治疗后出现耐药现象给出了一个合理的解释，也进一步为未来个性化医疗提供了一套型的病毒分型方法。同时，我们首次使用的独立成分分析方法实现对乙型肝炎病毒的各个亚型的完整分类，同时也提出了寻找病毒重组推断的新方法。

关键词： HBV(乙型肝炎病毒)，基因型，慢性病毒性肝炎，抗病毒治疗，深度测序，独立成分分析

ABSTRACT

ABSTRACT

38 chronic hepatitis B virus (CHB) patients of 200 CHB patients were investigated HBV subtype shift during adefovir dipivoxil(ADV)' s antiviral therapy. At the same time, we have detected that these patients' ADV antiviral efficiency decreased as well as their CHB virus' genotype has shifted from C type domination to B type domination, which is less sensitive to ADV via HBV genome analysis, via next generation sequencing techniques (including 454 pyro-sequencing and solexa).

Based on the genome dataset of hepatitis B virus(HBV) downloaded from NCBI and the correlation represented among different locus of HBV genome and the fact that there are many evolutionary conserved locus across whole HBV genome, we constructed a machine learning method derived from ICA(independent component analysis) to classify 8 HBV subtype by short segments. At last, we built a framework of HBV categorization and analyzed the details of mutual dependence among different subtype.

After we constructed the HBV classification framework, we cooperated with Chongqing First Affiliated Hospital to categorize different CHB patients' serum virus' subtype. Besides the construction of our HBV classification framework, we selected different short segments from HBV genome for Chongqing First Affiliated Hospital, designed barcodes for different patient' s segments and processed the original sequencing data, such as noise removal of NGS data. Finally, we convinced our assumption that CHB' s resistance and subtype shift are caused by HBV mix infection but not mutation.

In conclusion, taking into account that genome variance among different HBV subtype and their sensitivity to ADV treatment and 2nd sequencing technique' s ultra-high resolution for DNA density. We have provided a rational explanation for

ABSTRACT

HBV's subtype-shift and ADV's decreased antiviral efficiency during treatment, as well as built up a framework of virus genotype for personal medical treatment in the future. Additionally, this is the first time that we applied ICA (independent component analysis) to HBV subtype classification. ICA is a highly accurate method for recovering HBV subtype distribution and we can even retrieve the HBV genome's accurate recombination pattern vi our ICA method.

Keywords: Chronic hepatitis B, Deep sequencing, Genotype shift, Antiviral therapy, ICA

目 录

摘要	I
ABSTRACT	III
目录	V
第1章 引言	1
1.1 乙型肝炎病毒 (HBV) 以及慢性乙肝	1
1.2 乙型肝炎病毒亚型 (Hepatitis B virus(HBV)'s subtypes)	2
1.3 乙型肝炎病毒结构 (DNA structure)	3
1.4 乙型肝炎病毒预分型技术 (REGA)	5
1.4.1 REGA 分型方法	6
1.5 独立成分分析 (ICA)	7
1.5.1 动机	7
1.5.2 主成分分析 (PCA)	9
1.5.3 独立成分分析 (ICA)	10
1.5.4 独立和不相关	13
1.6 二代测序技术	14
1.6.1 454 测序方法系统误差的解释	14
1.7 EM 算法矫正二代测序荧光数据	16
1.7.1 EM 算法	16
1.7.2 EM 算法过程	16
1.8 HBV 基因组分型的模型抽象	18

第 2 章 材料, 方法和应用	19
2.1 HBV 的基于窗口分型框架	19
2.2 从 NCBI 获取 HBV 序列	19
2.3 NCBI Meta Data 预处理	20
2.4 REGA 分型数据的处理	20
2.4.1 HBV 标准序列集合构建 (Gold Standard Gene Set)	21
2.4.2 HBV 基因组位点偶联分析	22
2.4.3 基于 HBV 基因组亚型特征的序列分型	46
2.5 基于短序列窗口 HBV 分型平台构建及其应用	58
2.5.1 病人实际样本的获得	59
2.5.2 Solexa 测序结果还原原始序列比例的探究	60
2.5.3 窗口分型能力评价函数的确定	61
2.5.4 利用 RT 区进行确定窗口分型能力的参数	63
2.5.5 对 454RT 区测序结果进行荧光强度的校正	64
2.5.6 消除 polyhomer 读数的随机波动带来的实际判断误差	64
2.5.7 原始荧光数据的校正	66
2.5.8 HBV 分型 Solexa 窗口扩增的设计	67
2.5.9 HBV 测序结果的分型	69
2.6 HBV 短序列分型结果分析	70
2.6.1 二代测序还原 HBV 基因型亚型的标准曲线	70
2.6.2 二代测序结果和患者混合感染情况分析	72
2.6.3 基因型反转现象的解释	75
参考文献	79
附录	85
2.7 454 测序和 Solexa 测序后分型结果的对比数据	85
2.8 Solexa 测序 Barcode 及组合设计	86

目 录

致 谢	91
在读期间发表的学术论文与取得的研究成果	93

第 1 章 引言

1.1 乙型肝炎病毒 (HBV) 以及慢性乙肝

乙肝病毒是一种小型的部分双链 DNA 病毒¹。虽然一般来说，大部分患者都感染的是急性得 HBV 感染，但是，这个世界上仍然存在多过 350,000,000 得患者遭受得是慢性乙型肝炎得感染，同时，这些慢性乙肝患者最终发展为肝硬化或者肝癌得几率似乎更高 [来自:[WHO. Hepatitis B](#)]。对 HBV 病毒来说，一般存在这 HBeAg, HBsAg 和 HBcAg 三种抗原，分别对应着 HBV 病毒颗粒上的不同的位置。HBcAg 位于 HBV 颗粒最靠近 DNA 核心的位置，主要用于表明 HBV 的病毒复制能力是激活的。HBcAg 的存在意味着 HBV 病毒具有着传播感染的能力，相对处于中间位置的 HBeAg 则被认为是 HBcAg 的胞外形式，所以，一般来说医院对 HBV 病毒的监测都主要针对的是 HBeAg。同时，抗体也可以以 HBeAg 作为靶点，来发挥免疫应答作用。这两类蛋白其实都来自于 HBV DNA 的 Core 开放阅读框和'Pre C' 开放阅读框，但是基于某种原因前者成为的‘微粒状’，而后者则倾向于非颗粒的分泌形式。最后，HBsAg 主要是 HBV DNA 上'Pre S1', 'Pre S2' 和'S' 开放阅读框的翻译产物，分布在 HBV 的包膜上，可以被认为是 HBV 病毒感染的标志。

从上面的抗原属性，我们可以知道一般来说 HBeAg 主要就是作为 HBV 监测感染的标志之一，可是，对于慢性乙型肝炎患者来说，被感染者的 e 表面抗原 (HBeAg) 既可能是阳性，也可能是阴性。e 表面抗原 (HBeAg) 的血清转化点 (HBeAg 降低到免疫系统无法响应产生抗体) 会意味着 HBV 感染者从慢性 HBV 感染转化为乙肝病毒的非活性携带状态，这些病人常会表现为：表面抗原阴性，抗 -HBe 阳性以及高水平的 HBV DNA 滴度。这往往是由于他们感染的 HBV 不持续表达 HBeAg 的缘故。因此，这可能就会导致这些 HBV 携带者产生更加严重的肝脏损伤。

而目前我们主流的针对 HBeAg 阳性和 HBeAg 阴性 HBV 感染的治疗方法

遵循的原则是：抑制病毒复制，防止进一步的损伤。所以，为了维持这个长期的病毒抑制，对慢性乙型肝炎携带者的长期的治疗是必须的。否则，一旦停止治疗，HBV 的 DNA 滴度常常会迅速的反弹。对于 HBeAg 阳性患者，我们常常使用 HBeAg 的血清转化位点作为疗效评估的参数。

血清转换 (seroconversion) 其实就是一个集体免疫对 HBV 病毒的表面抗体敏感性的一个阈值估计，这个阈值估计体现了生物体的鲁棒性，但是也就回导致集体对 HBeAg 阴性的 HBV 病毒不敏感。一般来说，血清转化与谷丙转氨酶 (ALT) 表达水平，HBV DNA 滴度以及肝脏的组织学特性都存在相关性。

所以对于 HBeAg 阴性患者，我们无法在使用血清转化作为参数，一般直接使用肝脏组织学参数来作为判据和参数。

目前针对慢性乙型肝炎患者的主流治疗方法分为如下两种种：1，干扰素 $IFN - \alpha$ ；2，拉米夫定 (*lamivudine*) 或者阿德福韦酯 (*Adefovir Dipivoxil*) 等核酸类似物抗病毒药物，对病毒 DNA 链的合成和延伸产生竞争性抑制。这两种方法都能够在一定比例的患者人群中实现较好的疗效。但是， $IFN - \alpha$ 存在者不良反应剂量限制，同时摄入必须要求肠道摄入，而拉米夫定则过于容易诱导病毒产生抗药性。

相较而言，虽然阿德福韦酯也存在拉米夫定同样的问题，但是产生抗药性的几率相对小了一些，所以，我们实际上在现在的治疗过程中往往更倾向与使用阿德福韦酯定为治疗方案。但是，实际上虽然阿德福韦酯可以在短期迅速降低病人的 HBV 抗原数量，并且降低 HBV 的病毒滴度，但是由于监测本身的盲区属性，导致了我们的慢性乙肝患者往往在接受治疗很长事件之后仍然出现了病毒的药物耐受而进一步反弹的现象。这个现象也是我们此次论文中发起讨论的基本动机。

1.2 乙型肝炎病毒亚型 (Hepatitis B virus(HBV)'s subtypes)

我们为了能够对于为何慢性乙肝患者在接受抗病毒治疗很长一段时间之后还出现了反弹的现象进行研究，一个比较直接的想法就是在治疗过程中，出现了新的 HBV 的基因型差异，因为新的基因型差异对应的 HBV 病毒实际上是对现有的抗病毒药物不敏感而导致了病毒感染在治疗一段时间之后出现了反弹

的迹象。于此同时, Liu 和 Mirandola²⁻³ 在他们的报道中暗示了 HBV 亚型中的 A 亚型实际上存在着对于核苷酸类似物形式的抗病毒药物存在着低敏感性, 所以, 这暗示了我们有可能可以从 HBV 病毒的基因型角度入手来研究关于 HBV 对抗病毒药物的耐药性变化的相关性。

由于本身我们可以针对 HBV 分型为 8 种到 10 种不同的亚型, 每一个亚型内部大概存在 8% 以下的差异, 虽然有科学家已经报道了 A 型基因存在对类似核酸类似物药物不敏感的现象了, 但是, 我们对于中国大陆的患者来说, 主要的感染病毒亚型主要是 B 和 C 亚型, 所以, 我们希望能够从 HBV 的序列结构入手, 通过高精度的测序, 来验证是否类似的 HBV 亚型转化也存在与我们治疗的患者种, 并且希望能够给这种现象一个合理的解释;

我们把分析动机划分作两个类型: 1, 就是 HBV 的亚型转化是在药物作用下发生突变而产生的吗? 2, 或者是本身 HBV 在感染的过程种就是混合感染, 只是由于普通医疗技术有限, 没能够监测出实际上的不同的 HBV 亚型在治疗之前实际上就已经存在了。基于这个假设, 实际上我们通过 NCBI 采样数据集合的方式进行了深度分析, 并且最终排除了实际上不同亚型会通过进化而产生亚型转换的可能, 因为实际上我们发现 HBV 亚型间的进化是相互独立的, 所以, 对于现有问题唯一的解释便是 HBV 出现了混合感染而最终导致在治疗的后期出现了 HBV 基因型反转式的耐药性突变。

1.3 乙型肝炎病毒结构 (DNA structure)

由于我们需要研究 HBV 基因组上位点之间的相互约束与对应蛋白质之间的关系, 所以, 我们需要先大致了解 HBV 整个基因组表达蛋白的实际情况 [图:1.1]。

乙肝病毒主要包含 4 个主要的开放阅读框 (open reading frame, 后面缩写为: ORF), 并且他们全部是由负链来负责编码蛋白。他们满足以下特点:

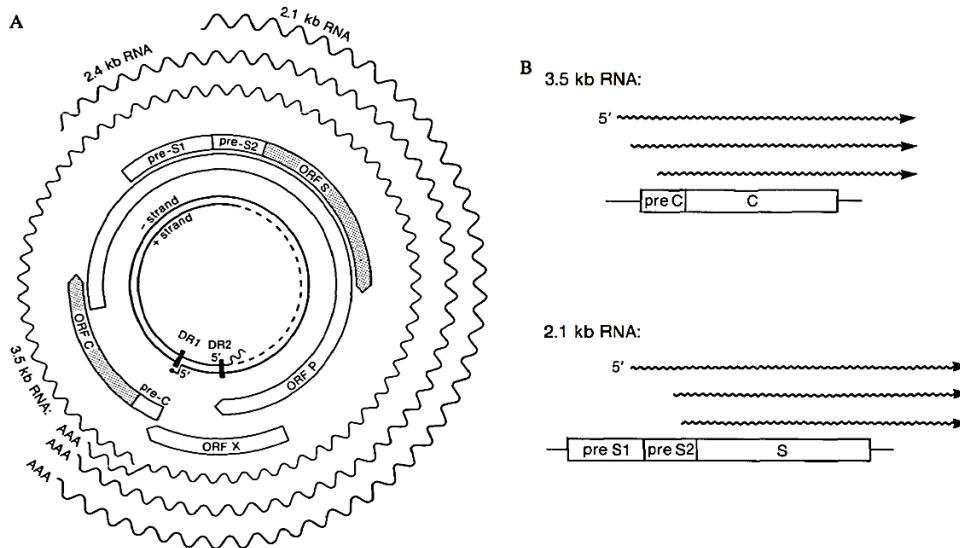
1. 病毒表面抗原基因的复杂性非常高。
2. HBsAg 的开放阅读框 (叫做 ORF-S)。

3. ORF-S 上游是一个 in-phase 阅读框：ORF Pre-S. 这个 Pre-S 里面有两个保守的 ATG 密码子，这个密码子主要用于引导额外的 HBsAg 相关蛋白的合成。
4. 3 中说到的两个 ATG 起始密码子，把整个 ORF Pre-S 区分成了两个子区 =Pre-S1 + Pre-S2。生成的蛋白产物属于病毒的衣壳类。
5. 还存在 ORF-C 区域：编码 HBcAg。
6. 同理存在 Pre-ORF-C 区域。生成：HBcAg 相关多肽。
7. 还有 ORF-X 与 ORF-C 两个区域。
8. 还有与前面几个开放阅读框产生重叠的区间 ORF-P，编码病 HBV 病毒的相关聚合酶。
9. 最后 HBV 基因组开放阅读框对应的的产物 mRNA 还会通过 RNA splicing 等工作之后才能够完成并进行翻译。

从上面的 HBV DNA 结构图和分析我们可以看到，在乙肝病毒基因组里面，每一个碱基都会被归属到至少一个编码蛋白的开放阅读框里面，同时，有超过 50% 的碱基还会同时归属到大于等于 2 个开放阅读框里面，甚至，在不同的开放阅读框里，不同的 ATG 其实密码子也会对应着不同的蛋白表达。正是由于 HBV 病毒的如此特性，我们才会认为 HBV 基因组上面的每一个碱基都会基于蛋白功能的受到进化上的约束压力，使得全局的碱基在进化过程中没有非编码基因和编码基因之分。但是，与此同时，HBV 基因组里面也仍然存在着保守性序列片段的存在。

此外，HBV 基因组还有一些小的特性需要注意：

1. 有一个 11nt 的短序列片段，在 HBV 基因组里面重复出现两次。（DR1,DR2 重复出现在 +/- 链上面），都在 5' 端，也就是说，即使是一个开放阅读框，HBV 也可能对应翻译出两个不同的对应蛋白质。
2. 还有在核心的抗原编码序列 (core antigen coding sequence) 的 5' 端会出现 TATAAA 保守。是 cleavage/polyadenylation 的信号序列。

图 1.1 HBV 的 DNA 结构示意图 (图片来源:⁴)

1.4 乙型肝炎病毒预分型技术 (REGA)

由于乙型肝炎病毒存在着 8 种不同的基因型，所以，如何很好的区分开这 8 个亚型就成为了一个很重要的 HBV 病毒研究工作，也是很多其他 HBV 相关工作的基础。

由于 HBV 基因型的进化树分型聚类之后呈现出来的亚团体效应，所以，我们只需要确定对不同 HBV 基因组之间的距离度量的定义是一致的（满足汉明距离的基本要求），我们就可以通过不同的分类器方法来确定 HBV 的 8 个亚型，基于 neighbor join method 的进化树分析方法是层次聚类算法的一种。

而为了能够在提高 HBV 不同亚型的分型效率，之后⁵等人开始引入 PSSM 对 HBV 进行基于位置权重矩阵定义的中心的聚类。由于使用不同 HBV 亚型的 PSSM 矩阵，不仅可以加快新加入的 HBV 基因组序列的分型，同时还可以保留下来每一个位点在进化上在全亚型内部的差异性，使得我们更容易的观察到整个 HBV 基因组上，不同位点之间的相关系数（协方差）。

Tulio de Oliveira *et al.* 于 2005 年发布了 REGA 分型工具⁶，通过 REGA，用户可以实现片段长度大于 800bp 的 HBV 基因组片段的分型。

1.4.1 REGA 分型方法

REGA 是基于进化树 (NJ and Bootstrap) 实现的长片段病毒基因型分类工具。

REGA 基于的 Bootstrap 方法是：对于某一个固定长度的序列，将其与事先准备好的参考序列 (reference sequences) 进行序列对其，然后对其使用 Bootstrap 方法，重新独立可重复 (independently repeated) 的排列现有的这一条 query sequence 的每一个列，再对新的重排产生的序列构建行的进化树⁷。再新的构建的进化树上，如果序列局部的分支的拓扑结构可以重现的话，用户就可以对于整个进化树层层的拓扑结构的重现率做出一个相应的统计：Bootstrap 的做法的目的是为了得到一个物种之间相关的估计。这里假设实际的相关度是 θ ，那么估计可以被写作 $\hat{\theta}$ 。在这文章⁷ 里面，作者使用数值化 ATCG 的方法，使用每一个 Bootstrap 得到的 X(sequence) subsets 来进行进化树的构建工作。

Bootstrap 方法也同样被使用在了 REGA 分型方法之中，具体的 Bootstrap 计算方法如下⁷：

首先，对基因组上的 A, T, C, G 进行数值化处理；然后，使用每一个 Bootstrap 得到的子集序列来进行进化树的构建，如此，便可以统计在不同的 Bootstrap 子集种出现的与原有的进化树拓扑结构类似的子区域的信息。

同时，由于每一个 Bootstrap 序列的出现都会满足某一个出现频率：

$$\hat{\pi}_k = \text{Column}_{X_k=x} / n \hat{\pi}^* \rightarrow D \rightarrow \text{TREE} \quad (1.1)$$

D 是一个度量 Bootstrap 得到的不同的序列之间的距离的矩阵：

$$D_{ij} = \left\{ \sum_k \pi_k (X_{ki} - X_{kj})^2 \right\}^{\frac{1}{2}}$$

通过 Bootstrap 得到的 X_i ，我们就可以完成不同的 Bootstrap 的进化树的构建工作。虽然通过 Bootstrap 采样得到的序列分布全局之间会存在差异，但是，我们只考虑得到的单个 Bootstrap 序列 X_i 和 X_j 之间的 distance 时，所有的 Bootstrap 子集在这个监测基底之下被认为是满足独立同分布 (i.i.d) 的，所以，我们通过比较不同的 Bootstrap 的子拓扑结构，找出合适的总是重复的结构，作为基于进化树对整个序列集合分型结构可靠性的证据。

REGA(1.2a, 1.2b) 方法本身可以被认为是决策树，使用了基于滑动窗口的 Bootstrap 和 Bootscan 判据，一次来证明提供的目的序列的每一个 400bp 滑动窗口是不是可以找到类似结构的 HBV 亚型把对应的窗口归属到对应的亚型里面 (如果 Bootstrap> 阈值)。

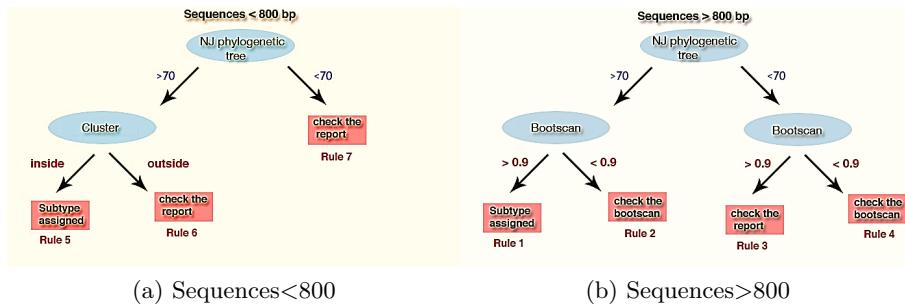


图 1.2 REGA 分型示意图 (Oxford HBV REGA Decision Trees)

虽然 REGA 分型在短序列分型的应用上受到自身框架的限制，但是，由于我们需要首先构建一个分型准确的 HBV 标准序列集合，使用窗口滑动的 Bootscan 方法确定的给定 HBV 基因组 (全长) 的乙型肝炎病毒亚型。除了使用 Bootstrap 来确定目标序列的基因亚型归属之外，REGA 方法还基于 400bp 的窗口滑动的方式，在全局设定一个较为严格的阈值来区分用户提供的目标序列是否是嵌合体。所以 REGA 在分型规则 ($\text{Bootstrap} > 70\%$) 的基础上还引入了滑动窗口打分机制 Bootscan 来消除嵌合体在 HBV 分型过程里的影响。短窗口的 Bootscan 方法，克服了 Bootstrap 对于小片段嵌合体的不敏感的问题，所以 REGA 分型方法，可以准确的确定出 HBV 全基因组上的发生嵌合位点，也就能够为我更好的排除掉重组问题对我们建立 HBV 标准全基因组集合的干扰。

1.5 独立成分分析 (ICA)

1.5.1 动机

我们通过 REGA 方法从原始数据库种，获得了非嵌合的，单一 HBV 亚型数据，如果基于 HBV 本身物种的假设，由于其自身可能会受到环境压力，功能保守性，宿主选择，表达蛋白活性等多方面的原因，整个 HBV 基因组上的位点是否为相互独立的，或者两两之间是否存在进化上的约束，就成为了 HBV 基因

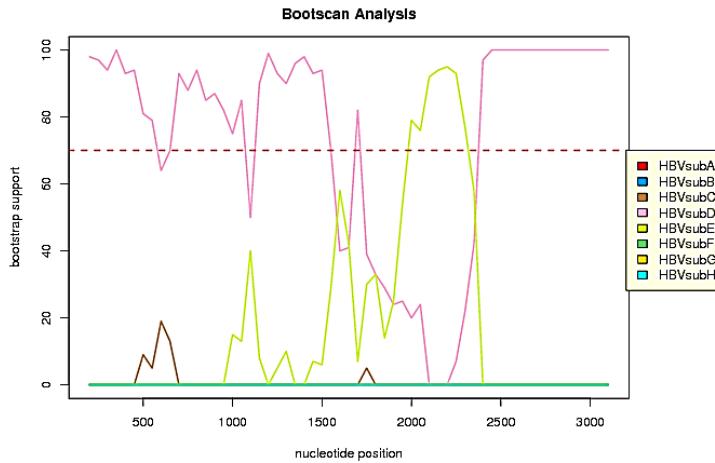


图 1.3 HBV 序列: gi|164509319|emb|AM494716.1|, length: 3182 bp 使用 REGA 分型结果, 可以找到全基因组上面, 该序列发生重组的位点 (来源: [Oxford HBV Automated Subtyping Tool \(Version 1.0\)](#))

组, 以及 HBV 的不同亚型在进化上表现出来的重要特征。由于受到现有的二代测序技术序列读长的约束, 对于序列长度 $> 800\text{bp}$ 的基因组序列, 我们很难直接得到完整的对应基因组概貌, 同时, 如果使用基于测序短片断结果来进行全基因组的拼接可以本身又是一个 NP hard 的问题, 虽然近些年来, de Bruijn Graph 等方法的出现, 基于贪心策略, 可以从二代测序的基因组片段中得到全局基因组的一个解, 但是, 由于这种贪心策略本身不一定代表者生物体功能的本质, 同时我们也知道整个基因组上不同片段在对基因描述时的重要性权重也不完全相同, 所以, 我们希望能够从序列本身的内部结构出发, 基于问题的确定更为高效的测序方法来还原 HBV 基因组的亚型信息。同时, 从个性化医疗的发展角度来看, 如果我们能够把对病患血清 HBV DNA 的全基因组蛮力 (brute force) 测序, 转变为定点的监测少量位置, 对于技术的推广, 帮助医务工作者的角度上来说也会优于获得高度冗余的全局海量全基因组数据。

由于 HBV 序列本身的高度冗余性, 同时考虑到统一物种基因在进化中受到的相互约束, 我们希望能把整个 HBV 的全基因组数据, ~ 3000 位点意味着极高的基因组空间维度, 考虑到 HBV 全基因组本身的高度保守性 ($\sim 8\%$ 序列差异), 通过统计学习方法, 压缩到我们可计算, 易于通过二代测序技术获取的片段空间内。基于这个目标, 我们希望能够从 HBV 全基因组空间, 线性变

化到某一个低维度子空间里，而这个低维度的序列子空间 S （意味着 HBV 基因组上的某些片段），既可以代表不同亚型间的差异，有能够保证与 $\mathbb{C}S$ 的序列存在着显著的相关性，这样我们就可以保证 S 内部的位点尽可能的相互独立， $P(S) = \prod P(S_i)$ ，同时，又可以保证 S 与 $\mathbb{C}S$ 的位点具有显著的关联，这样，我们就可以使用少量的 HBV 序列子空间（短片段），来对 HBV 实现分型。

1.5.2 主成分分析 (PCA)

主成分分析 (PCA) 就是一种常用的数据降维方法，在实际操作中，PCA 就是一个关于原数据集合的一个线性变换，可以说是基于对奇异值分解方式的一种变体，对于一个普通的矩阵 X ，我们可以对他进行奇异值分解得到 $X = U\Sigma V^T$ ，但是，实际上，如果我们对于 $X^T X$ 的相关矩阵来分解的话（具体操作需要保证量纲一致，并且标准化），实际上我们就能够得到一个对于这个相关矩阵的方差分布情况的描述，对应的 $\Sigma^T \Sigma$ （特征值）从大到小排列实际上表述的是重新调整变换过后的坐标对 X 内数据间方差的解释能力，同时，我们可以认为不同的特征向量之间是线性无关的，也就是 $Y = AX + b$ 是不满足的，所以可以说明不同的特征向量之间某种程度上较为互不依赖性，但是实际上，真正的完全独立 (Independent) 和线性无关之间还存在着差异，这个问题我们在下面的工作里会有进一步讨论。

1.5.2.1 主成分的一般定义

存在随机变量 X_1, X_2, \dots, X_n ，样本标准差记为 S_1, S_2, \dots, S_n ，首先作标准变换：

$$C_j = a_{j1}x_1 + a_{j2}x_2 + \dots + a_{jp}x_p, j = 1, 2, \dots, p \quad (1.2)$$

我们有如下定义：

1. $C_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$ ，满足 $\text{Var}(C_1)$ 最大，称 C_1 是第一主成分；
2. $C_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p$ ，满足 $\text{Var}(C_2)$ 第二大，且 $(a_{11}, a_{12}, \dots, a_{1p}) \perp (a_{21}, a_{22}, \dots, a_{2p})$ ；
3. $p \in (1, 2, \dots, p)$ ，各个主成分向量的属性依次类推。

1.5.2.2 主成分的性质

1. 各个主成分之间各不相关, 但是注意, 这里所说的线性相关, 即 $Y=AX+C$, 所以我们可以定义相关系数 $\text{Corr}(C_i, C_j) = p(C_i)p(C_j) - p(C_i, C_j) = 0$;
2. 组合系数 $(a_{i1}, a_{i2}, \dots, a_{ip})$ 构成的向量是单位向量;
3. 各主成分的方差是依次递减的, 即 $\text{Var}(C_1) > \text{Var}(C_2) > \dots > \text{Var}(C_p)$;
4. 总体方差不变
5. 原变量和主成分的相关系数 $\text{Corr}(C_i, x_j) = a_{ji} = a_{ij}$
6. 基于上面的假设, 我们可以得到 X_1, X_2, \dots, X_p 这 n 个随机变量的相关矩阵 R , $(a_{i1}, a_{i2}, \dots, a_{ip})$ 就相当于是这个相关矩阵的第 i 个特征向量 (Eigenvector), 特征值 λ_i 就是 i th 主成分的方差, 由 (3) $\Rightarrow \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$

1.5.2.3 主成分分析计算过程

1. 原数据集标准化之后, 为: $X=(X_1, X_2, \dots, X_p)^T$, 每一维 $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$, 所以, 标准化结果 $Z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$
2. 计算相关矩阵 $R = \frac{Z^T Z}{n-1}$
3. 求解 (2) 中 R 的特征值, X 中的各个主成分。并且, 一般保留前 k 个特征值, 使得 $\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i} \geq 0.75$, 或者更大, 说明只需要使用 top k 个 eigenvector 就可以解释所有的随机变量带来的方差了, 所以, PCA 将随机变量的维数从 p 压缩到 k ($k < p$)

1.5.3 独立成分分析 (ICA)

虽然通过主成分分解, 我们可以找到一组线性无关的成分, 但是, 在实际情况中, 往往各个信号之间不是简单的不满足线性无关就独立了, 信号数据的原始分布也不一定满足高斯分布, 这时候我们使用 PCA 分析就不一定能够从混合信号中分离出不同的信号。

以我们本次工作中分析的 HBV 基因组为例进行分析，因为整个 HBV 全基因组，由于对应的开放阅读框（ORF）的覆盖区域有所差别，有可能会意味着，最后 HBV 的 Genome 差异来源于很多不同的位点所代表的信号，这样，我们就可以把这些独立的代表这不同可能进化原因或者进化压力的 HBV 碱基集合定义为 n 个信号源 $s = (S_1, S_2, \dots, S_n)^T$ ，而实际上我们看到的现有 HBV 基因池中的所有的 HBV 单体型，实际上是这些不同信号源的一个线性组合。定义矩阵 A，我们采集到的 HBV 基因池为 $x = (X_1, X_2, \dots, X_n)^T$ ，那么： $x = As$

在整个过程中，我们只获得了实际的采样数据 x，而源信号 S 和信号的组合方式 A 都是未知的，这个从采样数据推测原信号的过程，我们称之为盲信号分离。取 $W = A^{-1} \rightarrow s = A^{-1}x$

$$W = \begin{bmatrix} \dots & w_1^T & \dots \\ \dots & w_2^T & \dots \\ \dots & \vdots & \dots \\ \dots & w_n^T & \dots \end{bmatrix} \quad (1.3)$$

如此，每一个独立信源都可以被表述为采样矩阵的线性组合： $s_j^i = w_j^T x^i$

1.5.3.1 密度函数与线性变换

假设我们的随机变量 s 有概率密度函数 $p_s(s)$ （连续值是概率密度函数，离散值是概率）。为了简单，我们再假设 s 是实数，还有一个随机变量 $x=As$ ，A 和 x 都是实数。令 p_x 是 x 的概率密度

令 $W = A^{-1}$ ，首先将式子变换成 $s = Wx$ ，

$$F_x(x) = P(X \leq x) = P(S \leq Wx) = F_s(Wx) \quad (1.4)$$

$$\frac{\partial F_x(x)}{\partial x} = \frac{\partial F_s(Wx)}{\partial x} = \frac{\partial F_s(Wx)}{\partial Wx} |W| \quad (1.5)$$

$$p_x(x) = p_s(Wx) |W| \quad (1.6)$$

1.5.3.2 ICA 算法

考虑到我们最后在 HBV Genome 上使用独立成分分析方法，使用所有来自不同信源的碱基位点子集，来估计全局对应的 HBV 病毒基因亚型。我们使用的 ICA 解释是基于 Bell 和 Sejnowski⁸ 的最大似然估计的解释方式。

我们假定每个 s_i 有概率密度 p_s , 那么给定原始信号的联合分布就是 $p(s) = \prod_{i=1}^n p_s(s_i)$, 这基于不同信源之间统计上独立的假设，如此这些找到的信源的独立性不在限定于相关 (correlation)，而是任意函数形式下的不独立 (dependent) . 基于 $p_x(x) = p_s(Wx) |W|$, 我们可以求得:

$$p(x) = p_s(Wx) |W| = |W| \prod_{i=1}^n p_s(w_i^T x) \quad (1.7)$$

公式 [1.7] 中，左边是采样结果，是 HBV 实际呈现出来的基因组序列分布情况的描述，右边则是实际上每一个信源对应概率乘以 $|W|$

但是，在没有先验知识的情况下，我们无法得到 W 和 s ，因此，我们需要知道 $p_s(s_i)$, 所以，Bell 和 Sejnowski 按照一定的性质设计了一个概率密度函数给各个信源。并且，这个信源的概率密度函数应满足两个性质：单调递增和 $[0, 1]$ 。

所以， s 的累积概率密度分布函数定义为：

$$g(s) = \frac{1}{1 + e^{-s}} \quad (1.8)$$

求导数之后，得到：

$$p_s(s) = g'(s) = \frac{e^s}{(1 + e^s)^2} \quad (1.9)$$

得到了 $p_s(Wx)$ 的概率描述之后，对于已知的采样结果 $(x_1^1, x_2^2, \dots, x_n^i)$ 来说，我们使用对数似然估计重新表达 $\prod_{i=1}^n x_i$, 满足:

$$l(W) = \sum_{i=1}^m \left(\sum_{j=1}^n \log g'(w_j^T x^{(i)}) + \log |W| \right) \quad (1.10)$$

公式 [1.10] 中，对应的大括号里面是 $\log(p_x(x^i))$ 。所以，为了最大化对应的 W 的似然函数，基于 $\nabla_W |W| = |W| (W^{-1})^T$ 和 $\log g'(s) = 1 - 2g(s)$. 我们可以

使用梯度方式迭代求解 W .

$$W := W + \alpha \begin{pmatrix} 1 - 2f(w_1^T x^{(i)}) \\ 1 - 2f(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2f(w_n^T x^{(i)}) \end{pmatrix} (x^{(i)})^T + (W^T)^{-1} \quad (1.11)$$

W 迭代至收敛，既可用于 $s = Wx$ 来推断产生采样信号的独立信源。

1.5.4 独立和不相关

使用 ICA 的原因是我们要从实际的序列中，找出独立的 HBV 子序列集合，但是，之前的如 PCA 等方法，能够找到的仅仅只是线性无关的位点子集。我们可以定义：

- 1, 如果随机变量 X_i 和 X_j 独立，当且仅当 $p(X_i, X_j) = p(X_i)p(X_j)$;
- 2, 如果随机变量 X_i 和 X_j 线性无关，当且仅当 $E(X_i, X_j) = E(X_i)E(X_j)$;

线性无关是一个相对于独立更弱的条件， X 的概率密度函数不同，可能会存在这线性相关但是不独立的现象。所以，ICA 的盲信号分析领域的一个强有力方法，也是求非高斯分布数据隐含因子的方法。从之前我们熟悉的样本 - 特征角度看，我们使用 ICA 的前提条件是，认为样本数据由独立非高斯分布的隐含因子产生，隐含因子个数等于特征数，我们要求的是隐含因子。而 PCA 认为特征是由 k 个正交的特征（也可看作是隐含因子）生成的，我们要求的是数据在新特征上的投影。同是因子分析，一个用来更适合用来还原信号（因为信号比较有规律，经常不是高斯分布的），一个更适合用来降维（避免维数灾难， k 个正交的即可）。有时候也需要组合两者一起使用。

而实际在 HBV 基因组的各个亚型的生成过程中，本身位点上的选择一方面由于突变发生的概率低，另外一方面由于位点的便宜受到环境压力，ORF 编译蛋白质的活性等诸多原因约束，很难能够朝着体系最大熵值的方向演化，所以，我们在例如 MCMC，Gibbs 采样中使用的最大熵值假设也就不能够满足，位点的相关分布也就很可能不会简单的满足高斯混合模型。但是，我们仍然认为进化上相互独立的位点集合之间应当存在这互信息最小化来表征实际情况中的相互独立的位点集合以及序列集合。

基于上面的分析我可以这么认为，我们可以通过不同寻找独立成分的方式，来找出不同的独立 HBV 特征序列，这样的话，实际上这些独立开来的序列本身之间应当难以相互转化，至少可以认为在过去的历史记录里这样的事情没有发生过，基于这个想法，我们可以进一步找到一种 HBV 的序列特征划分，与此同时，我的工作中将会进一步证明，实际上不同的 HBV 亚型实际上就是 HBV 基因组上不同的信源 ('Physical Source')，所以，这个证明实际上也暗示我们实际上所有的基因型反转现象都不会来自于相互独立的信源之间的相互转化，而是来源于实际上监测分辨率不足而导致的把混合感染误认为是单一 HBV 亚型感染了。所以，我使用独立成分分析的方法，也就解决了关于之前慢性 HBV 感染在阿德福韦治疗过程出现亚型反转时候提出的两个假设机制。

1.6 二代测序技术

基于我们的动机，因为我们希望完成的是一个病毒分型的工作来解释 HBV 基因型反转对于抗病毒药物耐受性的影响，同时，我们又希望能够使用短片段来表征实际上的 HBV 亚型间差异，所以，我们为了能够进一步验证我们的想法，设计了不同长度的短片断窗口来进行实际上的分型工作，以希望能够做到相互验证的效果，并且进一步找出实际上最合适完成短片断分型的窗口长度范围。一般来说，不同的测序方法实际上对应了不同的最合适测序长度，例如 Solexa[图：1.4] 一般最合适的测序长度大约为上下游各 70bp，共计 140bp 左右的短片断测序工作，但是如果要大于这个长度的话，我们一般使用的是罗氏的 454 测序方法⁹，一般来说这种测序方法能够实现 400bp 以下的片段测序工作，最长设计能够达到 800bp，但是由于 454 测序本身存在这 Polyhomer 这类的问题，所以，我们在考虑使用 454 测序时，往往还需要引入预处理来排查方法本身的系统误差。

1.6.1 454 测序方法系统误差的解释

我们通过 454 测序的原理可以了解到实际上对于准备好的序列测序模版来说，该方法使用的是游离的 dNTP 与模版链反应来发出荧光的方式作为信号转化，光信号经过标准化处理之后就转化成了对应的序列。但是，由于在 dNTP 以

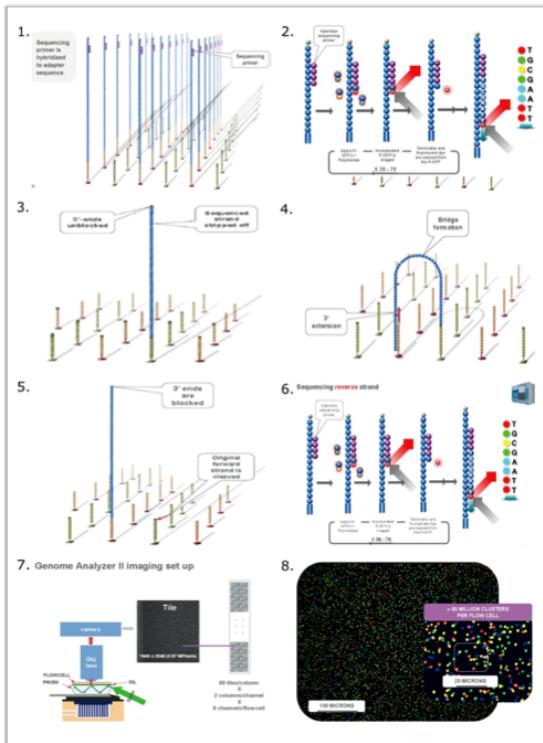


图 1.4 Solexa 测序原理示意图（来源：Illumina 公司）

(TACG) 流的方式输入到反应体系中的时候，实际上对于'AAAA' 或者'TTTT' 这类的 Polyhommer 来说，实际上是直接所有完全反应的，这时候，就会出现问题，例如对于'TTATT' 来说，实际上根本就没有 4 个'T' 连接在一起，却由于靠的太近，之前的 A 没有洗脱 100% 等原因，仍然显示出了大于'TT' 的效果，所以，为了克服这个现象，我们需要设计新的处理流程来矫正序列。与此同时，Solexa 测序方法则不会带来这个问题 [1.4]，因为实际上，Solexa 每一次完成一个位置的测序之后都使用了 Terminator 阻断后续反应，这样的话，我们就可以避免 Polyhommer 带来的读数不准的现象。而对于 454 方法来说，我们在本次试验中使用了 EM 算烦和 K-means 方法联合的方式来进行对 454 Reads 的矫正工作，以此尽量的改正系统误差带来的问题。

1.7 EM 算法矫正二代测序荧光数据

1.7.1 EM 算法

对于上述 454 测序矫正方法中使用到的 EM 算法，我们做一个基本的解释。EM 算法，即最大期望算法，在统计中被用于寻找依赖于不可观察的隐性变量的概率模型中的最优解，实际上是一个对参数集合的最大似然估计。一般来说，这些模型除了有未知参数和已知观测数据外，还有一些隐含变量。为了消除 454 焦磷酸测序过程中，由于系统对测序片段上 Polyhomer 位置的荧光强度读数偏差，我们使用期望最大化算法来还原实际上在 454 测序过程中得到的组成荧光强度数据的不同 Polyhomer 数目的混合模型。来矫正系统读数偏差带来的 Polyhomer 读数误差。

在 HBV 的 454 测序的 Flow 数据中，我们甚至都不知道最大的 Polyhomer 数是多少，所以，我们只能基于类似 Gaussian Mix Model 这类的假设，来构建出一个隐含 HBV Polyhomer 读数的采样函数，基于这个函数，我们可以通过最大似然估计的方式，迭代收敛出实际上的隐变量的最优解。同时，幸运的是，454 测序的数据结果，我们在预处理过程中发现实际上测序荧光数据实际上是对应着 Polyhomer 中连续碱基数的 log-normal 分布，基于这个假设，我们来学习出实际上 HBV 序列测序中对应的各个 Polyhomer 读数统计中心，实际上也就应该对应着最后的真实可能的 Polyhomer 读数。

1.7.2 EM 算法过程

假设现在有一个参数估计的问题，现在有一个训练集 $\{x^{(1)}, \dots, x^{(m)}\}$ ，它包含了 m 个独立的样本。若想将模型 $p(x, z)$ 中的参数去拟合这些数据，则似然性由以下公式给出：

$$\begin{aligned} l(\theta) &= \sum_{(i=1)}^m \log_p(x; \theta) \\ &= \sum_{(i=1)}^m \log \sum_z p(x, z; \theta) \end{aligned} \quad (1.12)$$

但是，如何找到合适的参数 θ 的值，使估计的似然性最大，可能是非常困难的。上面的公式中， $z^{(i)}$ 是隐含随机变量，大多数情况下，如果 $z^{(i)}$ 能够被观测到，那么最大似然估计就会变得很容易。

但是在 $z^{(i)}$ 无法观测的情况下，EM 算法就能够很有效地去实现最大似然估计。EM 算法的策略就是迭代。从偏理论一点的角度来讲，EM 迭代的过程每一步都会对结果有所改进，除非已经达到了一个（至少是局部的）最优解。

现在引入一个关于隐含变量的分布 $q(Z)$ 。注意到对于任意的 $q(Z)$ ，都可以对 Log-likelihood Function 作如下分解：

$$\begin{aligned} \log p(X | \theta) &= \sum_z q(Z) \log (p(X, Z | \theta)) / q(Z) + (-\sum_z q(Z) \log (p(Z | X, \theta)) / q(Z)) \\ &\stackrel{\triangle}{=} l(q, \theta) + KL(q || p) \end{aligned} \quad (1.13)$$

其中， $KL(q || p)$ 是分布 $q(Z)$ 和 $p(Z | X, \theta)$ 之间的 Kullback-Leibler divergence。由于 Kullback-Leibler divergence 是非负的，并且只有当两个分布完全相同的时候才会取到零^{10,11}，因此，可以得到关系， $l(q, \theta) \leq \log_p(X | \theta)$ ，亦即 $l(q, \theta)$ 是 $\log_p(X | \theta)$ 的一个下界。

现在考虑 EM 的迭代过程，记上一次迭代得出的参数为 θ^{old} ，现在要选取 $q(Z)$ 以令 $l(q, \theta^{old})$ 最大，由于 $l(q, \theta^{old})$ 并不依赖于 $q(Z)$ ，因此 $l(q, \theta^{old})$ 的上限（在 θ^{old} 固定的时候）是一个定值，它取到这个最大值的条件就是 Kullback-Leibler divergence 为零，亦即 $q(Z)$ 等于后验概率 $p(Z | X, \theta^{old})$ 。把它带入到 $l(q, \theta^{old})$ 的表达式中可以得到：

$$\begin{aligned} l(q, \theta) &= \sum_Z p(Z | X, \theta^{old}) \log_p(X, Z | \theta) - \sum_Z p(Z | X, \theta^{old}) \log_p(X, Z | \theta^{old}) \\ &= Q(\theta, \theta^{old}) + \text{const} \end{aligned} \quad (1.14)$$

其中， const 是常量，而 $Q(\theta, \theta^{old})$ 则正是之前所得到的“同时包含了 sample 和隐含变量的 Log-likelihood function 关于后验概率的期望”，因此对这个对应到 EM 中的“E-step”。

在接下来的 $M-step$ 中，固定住分布 $q(Z)$ ，再选取合适的 θ 以将 $l(q, \theta)$ 最大化，这次其上界 $\log_p(X | \theta)$ 也依赖于变量 θ ，并会随着 $l(q, \theta)$ 的增大而增大（因为有前面的不等式成立）。一般情况下 $\log_p(X | \theta)$ 增大的量会比 $l(q, \theta)$ 要多一些，这时候 Kullback-Leibler divergence 在新的参数 θ 下又不为零了，因此可以进入下一轮迭代，重新回到“E-step”去求新的 $q(Z)$ ；另一方面，如果这里 Kullback-Leibler divergence 在新的参数下还是等于 0，那么说明已经达到了一个（至少是局部的）最优解，迭代过程可以结束了。

上面的推导中可以看到每一次迭代 $E-step$ 和 $M-step$ 两个步骤都是在对解进行改进，因此迭代的过程中得到的 likelihood 会逐渐逼近（至少是局部的）最优值。另外，在 $M-step$ 中除了用最大似然之外，也可以引入先验使用 Maximum a Posteriori(MAP) 的方法来做。还有一些很困难的问题，甚至在迭代的过程中 $M-step$ 一步也不能直接求出最大值，这里通过把要求放宽——并不一定要求出最大值，只要能够得到比旧的值更好的结果即可，这样的做法通常称作 Generalized EM (GEM)。

1.8 HBV 基因组分型的模型抽象

以 NCBI Genome Dataset 和实际情况中测得的额外 HBV 序列片段共同作为 Target Sequence Database，通过 REGA 分型的 NCBI 的 HBV 序列可以被认为的是 Labeled Sequences。也就是在序列空间 $(S_1, S_2, \dots, S_k, S_{k+1}, \dots, S_n)$ 中，找到已经存在了的基于全序列方法定义的部分标签 $Labels(Y_1, Y_2, \dots, Y_k) (k < n)$ 所对应的序列空间的情况下，估计剩下的未标记序列的问题。这是一个可以被认为是一个半监督学习的问题，基于已被标记的部分序列的标签信息，我们确定出剩余部分序列的标签结构，并且推测出全体序列对应的标签 (Label) 情况。

第 2 章 材料, 方法和应用

2.1 HBV 的基于窗口分型框架

由于全序列测序本身涉及到序列的拼装问题是一个 NP hard 问题, 所以, 为了在过程中不使用序列拼装全基因组, 我希望找到 HBV 序列的特征局部结构 (local structure), 找到可以代表全局一致 (local and global consistency) 的基因型。我把 HBV NCBI Genome 库当作是前人在科学的研究过程中对 HBV Genome 分布的采样数据, 以此来找到 HBV 基因组内部的统计规律, 找到可以显著区分且代表 $\sim 3200bp$ 的 HBV 序列亚型的 HBV 核苷酸子集 S 。

我要求可区分 HBV 亚型的序列子集 S 必须满足:

1. 子集内部序列的进化关联尽可能小
2. 子集内部序列通过进化上的耦合尽可能的覆盖到足够完备的外部序列
3. 子集内部的位点在基因组坐标上尽可能的连续

为了找到满足这些特定约束下的位点序列集合, 我使用 NCBI Hepatitis B virus genome dataset 作为 HBV 基因的各亚型完备采样。并且使用位点耦合分析方法 (Site Coupled Analysis) 来找到 HBV 基因组全局上的位点进化耦合子集 S 。然后, 基于位点耦合子集和单一位点的可分型特征共同确定 HBV 基因组上的可分型片段。

2.2 从 NCBI 获取 HBV 序列

在[NCBI/nuccore](#)页面上的 Search 下拉列表中选择”Genome”, 然后在搜索内容中填写 “Hepatitis B virus[organism]”; 点击返回的 HBV reference genome (NC_003977); 在新打开的页面中将 Display 下拉框中选择为”Other genomes for species”; 这样得到的搜索结果有 2588 个 HBV 基因组序列. 共获得序列 2588 条。(获取数据的时间节点: 2011.6.20)。

2.3 NCBI Meta Data 预处理

在获得了 2588 条 HBV 原始数据之后, 我完成以下 2 步数据预处理工作, 然后把 HBV 基因组预处理数据进行 REGA 全基因组分型。

1. REGA 输入数据要求获得原始数据集中的序列 dbj ID 与 fasta sequence 的唯一对应。经过初筛的 2588 条序列被存储于 RS(Raw HBV genome sequences).fasta 文件中
2. 1 中获得的 RS(Raw HBV genome sequences).fasta 基于 REGA tools 的最大输入限制被划分做 4 个 subRS 文件, 使用 REGA web server 对其中所有的序列完成分型工作.

2.4 REGA 分型数据的处理

REGA web server 对输入的 HBV 原始序列进行分型, 并针对 RS.fasta 文件中每一条序列提供相应的 xml 格式分型结果。REGA 输出结果:

```
<sequence name="AY330915.1" length="3182">
<nucleotides>...</nucleotides>
<result id="pure">...</result>
<result id="scan">(Matrix) </result>
<assigned>...</assigned>
```

基于'assigned' 标签和 Bootstrap support 标签可以得到相应的 Bootstrap 分值和基因型的归属。

从窗口的角度上说, 可以认为对于其中的很多序列, 本身存在着前后序列完全不一致的问题, 即重组问题。也就是从我的 Profile 上来看: <profile id="assigned"> = HBVsubB - HBVsubC - HBVsubB, 这可能在暗示我提供的目标序列可能存在嵌合体的现象, 但是, 也暗示我使用一个 400bp 长度的窗口, 是具有区分开得到其对应亚型的能力。

REGA 分型结果统计, REGA 的分型结果来看 (图:2.1), 有大于 1/2 的样本序列都获得了 Bootstrap score = 1 的高分, 也就是说其分型非常确定, 而那些 Bootstrap score ≤ 0.7 的样本序列, 则认为其分型结果可靠度不高。

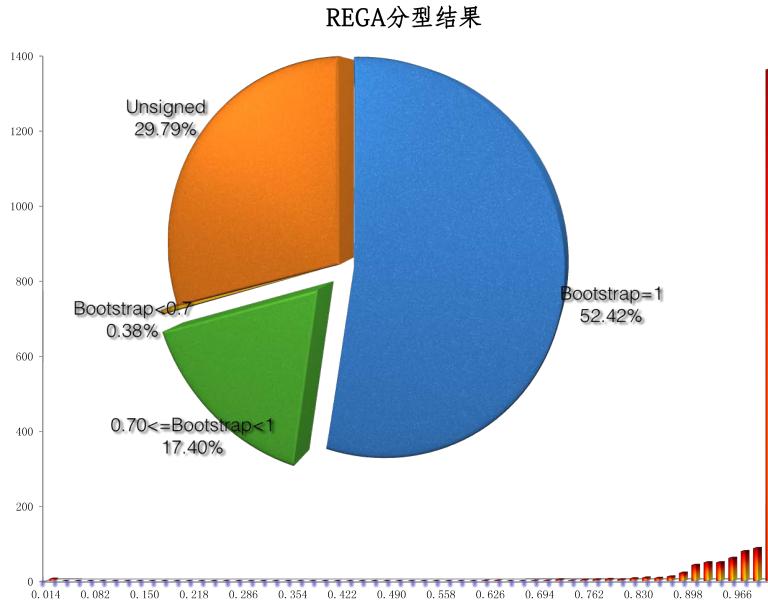


图 2.1 REGA 分型结果 support score 统计数据示意图

2.4.1 HBV 标准序列集合构建 (Gold Standard Gene Set)

基于 REGA tool 的输出, 我可以得到目标 HBV 序列的 Bootstrap 分数, Bootscan 分数和对应的 Genotype. 为了不引入任何的可能嵌合体 (Chimera), 我只选择 Assigned Support = 100%(100% 的窗口都是单一, 非混合基因组) 的序列作为黄金标准序列来进行进化偶联子集 S 的归属和固定长度可分型片段的搜索工作。因为只有选择 Assigned Support = 100% 的窗口, 也就是所有的滑动窗口都被归属到同一亚型的情况下, 我才可以完全确定这条序列的所有信息都可以被用来代表这一个 HBV 病毒亚型的每一个局部特征。而不会在统一基因型下混入亚型的片段信息。

首先, 为了后期工作的准确性, 并且避免模糊碱基带来的未知权值变量的干扰, 首先, 我删除了 1362 条分型确定的序列中, 包含不准确碱基信息 (W 或者 N 或者 R 或者 S 或者 Y 或者 K 或者 M 或者 V) 的序列。同时, 由于 REGA 自身提供基因型判别的 Bootscan 阈值提供了 70% 做为是否把序列判别为单一基因型的判据¹², 所以, 我把 Assigned Support $\in [70\%, 100\%)$ 的其他序列标记为一个外部测试集合。所以, 通过标准序列集合筛选, 我得到了: 1, HBV 标准序列集合, 1109 条序列; 2, HBV 外部测试集合, 331 条序列。

在后续的序列片段分型效果验证过程中, 我么可以不仅使用 $10\times$ 交叉验证来验证我选择的短序列片段分型能力的准确性和泛化能力, 同时也使用这个样本容量 331 的外部测试集来确定相应短序列片段的分型能力。

2.4.2 HBV 基因组位点偶联分析

2.4.2.1 多重序列对齐

为了能够保证 HBV 标准序列集合中的采样结果可以分析 HBV 基因组上每一个特定位点在不同序列间进化情况, 我首先通过多重序列对齐把所有的 HBV 标准序列集合中的序列投影到一个共同的序列坐标系下。这个坐标系下, HBV 标准序列集合中的每一条序列都被对齐到了一个固定的长度 ($3674bp$) 范围下。即, 多重序列对齐的结果保证了标准序列集合下的所有序列的所有位点都可以在同一个尺度下面讨论。

实验中分别尝试了 MAFFT 多重序列比对, ClustalW 多重序列比对及 T-coffee 多重序列比对。最后权衡了时间代价 (T-coffee 最劣) 和插入 gap 数目 (MAFFT 结果中插入了最多 gaps), 我最终采用了 CLustalW 的多重序列对齐结果。基于 ClustalW 序列对齐结果把所有标准序列集合中的序列投影到了 $3674bp$ 的序列框架下。

2.4.2.2 HBV 标准序列集合基于序列相似的 pattern 分析 (Similarity)

REGA tools 提供了所有 HBV 标准序列集合中对应的亚型, 同时, HBV 的亚型在这个 1109 条序列组成的标准序列集合中并不是均匀分布的:

表 2.1 HBV 标准序列集合亚型分布

基因亚型	A	B	C	D	E	F	G	H
序列数目	142	31	752	75	40	31	19	19

显然, 直接从 NCBI 上获得的 HBV 标准序列集合不是 $A \sim H$ 8 个亚型的均匀分布, 由于标准序列集合对不同 HBV 亚型的采样不平衡现象, 在我们处理了 HBV 标准序列集合过程中, 首先我需要消除序列间不平衡带来的各个 HBV 亚型对位点进化分析中贡献权值的不同, 以避免在不同 HBV 基因型亚型的特

征提取过程中出现序列样本不平衡带来的 bias pattern。为了在后续的序列进化特征提取过程中尽可能多的保留除了 C 亚型 HBV 基因之外的其他 HBV 亚型基因的位点特征, 我需要调整整个 HBV 标准序列集合中各个基因亚型对于 HBV 基因组位点进化离散差异的贡献, 以此来找到所有 HBV 亚型间的基因型间位点差异。2009 年, Batuwita 等人¹³ 将不平很分类问题的解决方案使用到了生物数据领域. 而在这之前, 对于各种数据出现的标记不平衡现象, 也提出了很多的方法:

1. 重采样方法:
2. SMOTE 方法:
3. 使用 Hellinger Distance, 以及 HDDT 算法来完成对特征的标准化。(消除不平衡带来的特征对应值偏差)
4. ROSE: 随机过采样方法 (Random Over Sampling Example)。

在这些方法中, 下重采样或者上重采样本身存在着或多或少的问题: 1, 下重采样需要删除掉一些 majority subset 的样本点, 这样的话可能会存在被删除掉的数据一些涵盖有重要的 Data Concept 的节点的风险。2, 直接基于原始数据的复制的上重采样, 则有可能会导致出现某一个样本进行大量复制而出现内部耦合加强的现象, 如果这个样本本身存在着大量噪音的话, 学习过程中就会出现过拟合 (overfitting) 的现象。

而小样本合成的重采样方法 (SMOTE:Synthetic Minority Over-sampling Technique) 是相对简单重采样来说保留最多数据信息的方法。因为我要区分出的是决定不同 HBV 亚型之间的耦合位点集合, 所以, 我因该在尽可能的保持 HBV 标准序列集合中的不同亚型之间全序列模式差异的前体条件下, 最大可能的标准序列们保留亚型内模式。同时, 我首先对不平衡的原始 HBV 标准序列集合进行处理:

从 HBV 标准序列集合的 gaps 分布情况来看, HBV 的大部分位点的 gap 频率都集中在极少 ($\leq 8.20\%$) 和极多 ($\geq 95.80\%$) 之间, 所以, 我任意选择这两个边界值范围之间的 cutoff 对于裁减之后的 HBV 标准序列集合坐标都不会有

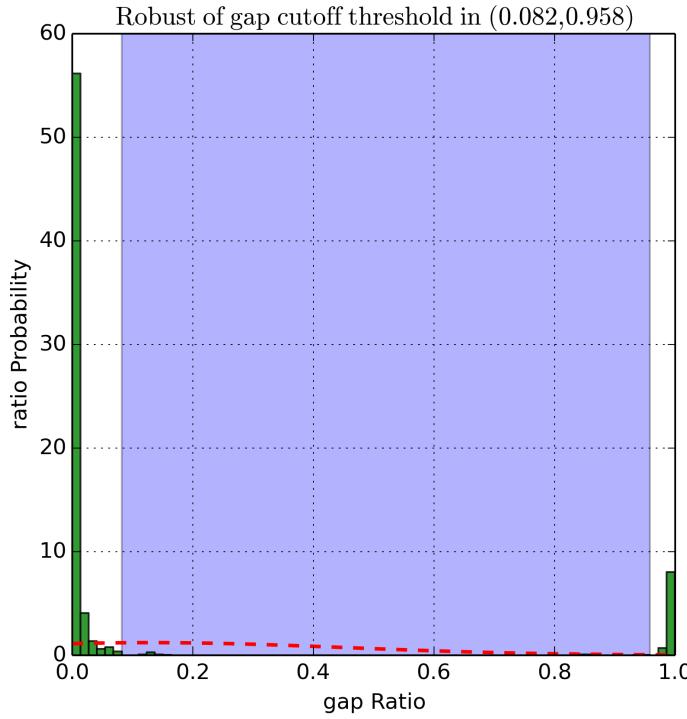


图 2.2 HBV 多重序列对齐结果的 gap 分布情况

影响, 这样保证了我后续讨论的算法在坐标系体系上存在这健壮性 (Robust)。所以, 我的序列耦联位点分析工作可以使用 (8.20%, 95.80%) 间的任意数值。故:

1. 使用 gap cutoff 设置为 0.2, 来对现有的序列集合进行裁剪
 2. 计算整个 N 条序列集合中序列两两之间的相似性 (pairwise similarity),

$$\text{similarity}(S_i, S_j) = \frac{\sum_{k=1}^L \delta(S_i(k), S_j(k))}{L}$$
 - 3.
- $$\delta(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases} \quad (2.1)$$

我可以得到实际上的 HBV 未修饰标准序列集合内部序列相似性:

从标准序列的序列间相似性热图 (pairwise similarity heatmap)(图:2.2) 上看, 我可以比较容易的区分出不同的亚型, 但是, 从序列集合的平均相似性集

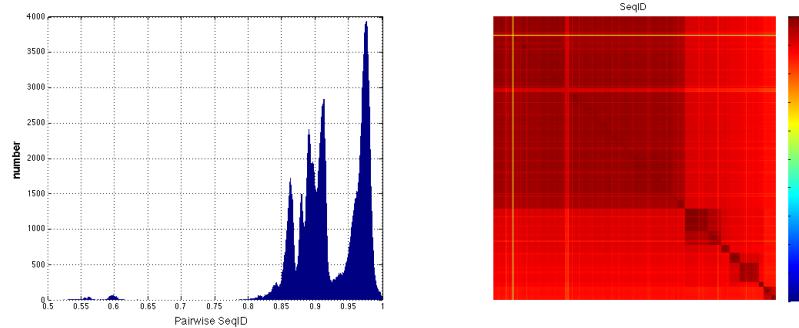


图 2.3 HBV 多重序列对齐结果的 gap 分布情况

中到了 95% 附近, 这比实际上的 HBV 亚型平均相似性高很多, 所以, 需要使用 SMOTE 消除整个 HBV 标准序列集合的亚型序列不平衡现象。

为了验证 SMOTE 方法的结果是否健壮, 我遍历并观察 $k \in [1, 10]$ 合成的 SMOTE 序列集合对序列相似性.

```

input :  $k, S$ 
output:  $new_S$ 

Function:  $SMOTE(k, S)$ 
1 从每一个亚型序列子集中随即抽取一条序列;
2 定 KNN 算法的 K, 然后找到每一个对应抽取的序列的 K 个最近邻
(K-nearest-neighbors) ;
3 从这 K 个最近邻 (nearest neighbors) 里选择出 1 个  $\hat{X}$ , 从  $[0, 1]$  中随机
抽取一个  $\lambda$ , 生成出一条合成序列;
4 把新的合成序列加入到序列集合 S 中 return  $new_S \leftarrow S$ 

```

算法 2.1: SMOTE algorithm

通过比较, 我发现选择不同的 k 值 (图:2.5), 在平衡标准序列集合的条件下, 各个序列之间的 pairwise 差异类似, 基于 similarity 的各个亚型间的序列相似性差异也很明显。

SMOTE 具体操作:

1. 使用 Maximum Likelihood (极大似然方法) 的 FastTree 构建进化树, 构建的进化树的同时得到了两两标准序列在进化上的距离 (序列间汉明距离)

2. 基于已知的距离矩阵, 进行 SMOTE 的分析, 其中, 每一个亚型子矩阵都可以人工合成序列了。
3. SMOTE 合成采样的序列数目基于最大分类 C 类的数目来设定, 所以 HBV 又重新一共合成了 $A \sim H$ 一共 4907 条序列。满足平均每一个亚型都获得 752 条序列 (等价与现有采样中最大数目亚型 C 亚型的序列数目)
4. $k \in [1, 5, 10]$, 检验每一个生成的新的平衡的序列集合的 threshold cutoff 的敏感性 (图:2.4)

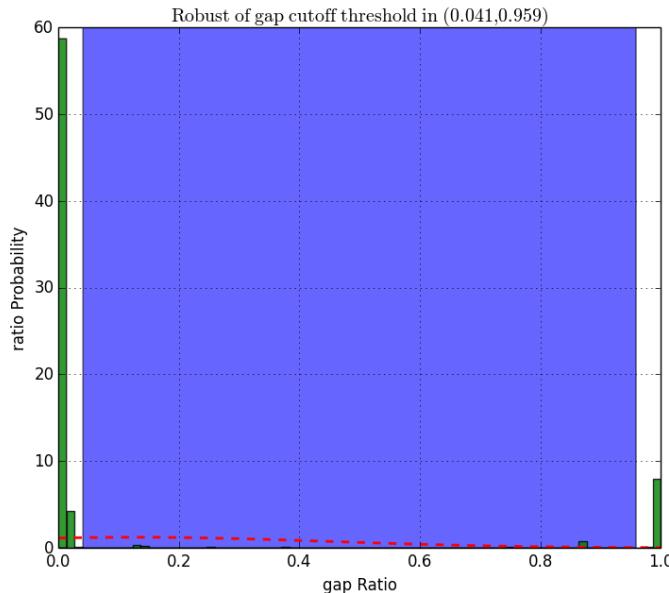


图 2.4 HBV SMOTE 集合多重序列对齐结果的 gap 分布情况;gap frequency 是集中分布在极高频率和极低频率之间 (只集中在头尾小部分区间内)

通过比较序列的整体相似性的变化量, 我发现不同的 SMOTE 参数带来的序列平均差异是健壮的, 所以我可以把 SMOTE 的 K-NN 中的参数 K 固定下来。然后我们发现, 对于不同的 K 对应的 SMOTE 算法, 新合成的序列给原始序列带来的序列差异是一致的。从 (图:2.6) 图中序列相似性 pairwise 矩阵的颜色深浅可以很容易的分别存在的八个 HBV 亚型之间的显著区别 (类内距离 \leq 类间距离)。

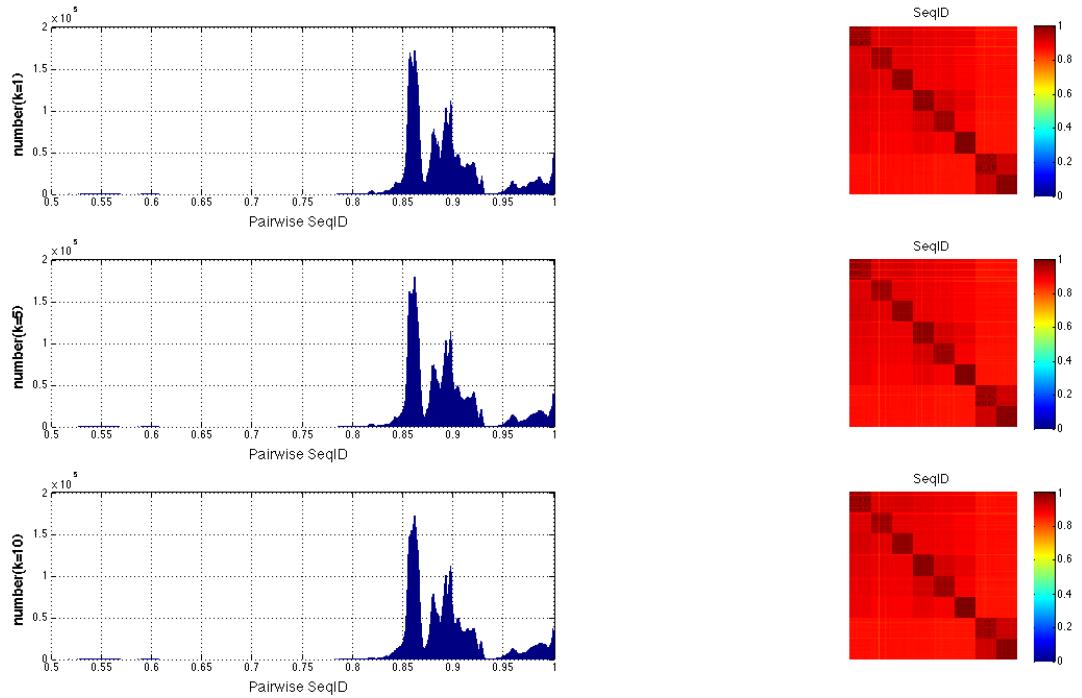


图 2.5 从上到下, 分别为 SMOTE 函数中最近邻数目为 1,5,10 时, 得到的 HBV 全基因组序列 pairwise 相似性分析

比较不同 k —SMOTE 下带来的相似性 (similarity) 分布和 pairwise diatance 矩阵, 我发现 (图:2.6):

再对比 (2.3) 和 (2.5), 我发现, SMOTE 即平衡了实际上各个 HBV 亚型之间的序列数目差异, 进一步让 HBV 全基因组序列上的序列间两两相似性 (pairwise similarity) 趋向于 $E_{sim} = 89.00\%$; $Std_{sim} = 4.06\%$, 结果与实际文献中报道的基因组亚型的序列差异吻合。

2.4.2.3 SMOTE 平衡标准序列集合的位点保守性 (Conservation)

我需要从 HBV 基因组序列集合中找出能够一致性表述全局 HBV 基因组亚型信息的局部片段, 这主要是基于 2.4.2.2 中, 对于完整序列集合内两两相似性的描述, 因为所有的 HBV 基因采样集合中, 大家的平均相似程度 (similarity $\sim 89.00\%$), 所以, 实际上, HBV 序列上能够用来描述 HBV 序列差异

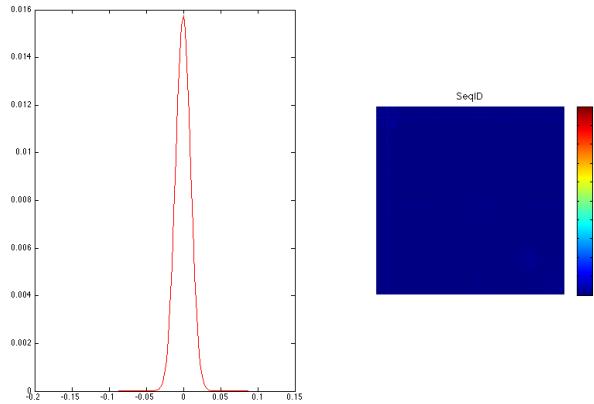


图 2.6 比较不同 k 带来的序列相似性差异; 相似性差异满足分布: $E = -0.000296129$; $Std = 0.0101335$

的碱基数目平均值约为 300bp, 而这个数目, 是我现有的二代测序技术能够准确, 单次完全覆盖的序列片段长度。

所以, 基于 HBV 基因组序列的大部分位点倾向于完全保守的推论, 我使用保守性计算方法, 排除掉 HBV 基因组对应的保守序列背景骨架, 并且使用统计耦合分析找到剩余不保守位点之间的相关情况, 通过:

1. HBV 基因组上大部分位点都倾向于保守
2. 其他非保守位点在进化过程中相互之间存在这非线性相关约束

以上两个推论, 我构建了基于 HBV 基因组的位点相关性分析框架, 这个框架曾被用于股市分析和蛋白质异构调控位点的分类工作中¹⁴, 通过这类方法, 科学家们可以找到在蛋白质异构调控中相互联通却又不再氨基酸序列上连续的氨基酸集合, Rama Ranganathan et.al 将这类序列位点的子集定义做蛋白质上的小集团位点 (Protein Sector). 而我这里使用类似的方法, 来寻找 HBV 基因组全局上相关联的碱基位点集合, 这些子集中的碱基位点在进化上受到类似的压力选择等信号源决定因素。为了方便操作, 我先对 HBV 的多重序列对齐文件中的 (A,T,C,G,-) 做因子处理, 然后在使用 KL 散度 (Kullback–Leibler divergence) 来表示不同位置的碱基保守性问题 (基于全局分析), HBV SMOTE 序列集合的位点保守性实现:

1. 使用 SMOTE 合成的新平衡全基因组序列作为输入, 得到不同碱基在全局条件下的分布 Q ; Q 满足 $(q^A, q^T, q^C, q^G) = (0.227 : 0.283 : 0.270 : 0.220)$, 也就是说实际情况中, HBV 序列上各个碱基使用频率接近于 $1 : 1 : 1 : 1$ (也存在一定的嘧啶倾向选择性)。
2. 同时, 记录下多重序列对齐 (MSA) 后得到的 $3674bp$ 比例尺下, 每一个位置上的 (A, T, C, G) 的碱基概率分布。设定: 位置 i 上, 出现碱基 $(a \in (A, T, C, G))$ 和 (包含 gap) 的概率为: $(f_i^{(a)} | i \in (A, T, C, G, -))^*$.
3. 定义对于任意某一位点, 如果四个碱基的分布比例接近于 $1:1:1:1$, 也就是说这个位置的碱基选择可以被认为是没有受到外界干扰, 而是随机的散落在整个样本空间里 (满足最大熵原理), 此时, 对于每一个位点对应的碱基分布 $f \rightarrow f_i^{(a)}$, 我定义其 KL 散度, 相当于该碱基位点的 f 要得到最优编码实际上相对于均匀分布 U 要额外增加多少比特位的信息。而这些信息主要应该包含的是对该位点的额外约束描述。所以, 我认为应该包含 HBV 基因组上为点之间的相关约束信息。

所以, 这里我定义 HBV 基因组上每一个位置的每一个碱基的保守性:

$$D_i^{(a)} = f_i^{(a)} \ln \frac{f_i^{(a)}}{q^{(a)}} + (1 - f_i^{(a)}) \ln \frac{1 - f_i^{(a)}}{1 - q^{(a)}} a \in (A, T, C, G, -) \quad (2.2)$$

$D_i^{(a)}$ 也被看作是序列对齐坐标上每一个位置的碱基频率相对于背景碱基频率分布 $q^{(a)}$ 的偏差程度, 如果位点趋向于保守, 他的自身的熵值就会有下降, D 就会上升。而只有在亚型之间不保守的序列位点, 才具有作为分型的表征位点的可能性。

对于 (A, T, G, C) 四个碱基来说, 每一个碱基在 i_{th} 位置的频率被定义做 $f_i^{(a)}$. 所以, 实际上通过二项分布定义每一个 $f_i^{(a)}$ 在 Mbp 长的对齐结果中出现的概率:

$$P_M(f_i^{(a)}) = \frac{M!}{(Mf_i^{(a)})!(M(1-f_i^{(a)}))!} (q^{(a)})^{(Mf_i^{(a)})} (1 - q^{(a)})^{(M(1-f_i^{(a)}))}$$

当 M 很大的时候, 是用 stirling equation 代替 $M!$ $\Rightarrow P_M(f_i^{(a)}) \sim e^{-MD_i^{(a)}}$

^{*} $f_i^{(a)} = \langle x_{i,s}^{(a)} \rangle$

所以, 保守性 (conservation) 可以用来描述实际位点上的碱基分布相对于背景碱基分布的离差程度

我注意到, 在每一个位置的碱基的保守性 $D_i^{(a)}$ 中, D 是一个非线性函数, 当 f 趋近与 1 的时候他的增速很快。为了能够之间体现每一个位置的保守性 D_i , 我使用每一个位置的所有保守性 $D_i^{(a)}$ 来计算他们在整个基因组上的平均值, 并最后代表该位点的表征保守性 (conservation)。越保守的位点, 存在的位点间相关越显著, 所以, 不同于传统的 HBV 亚型分类方法, 我们在考虑位点间或者序列间耦合时, 把每一个位点的保守权重都考虑到序列进化带来的差异和模式中。

保守性在原始 HBV 标准序列集合和 HBV SMOTE 序列集合上的差异也不大 (图:2.7):

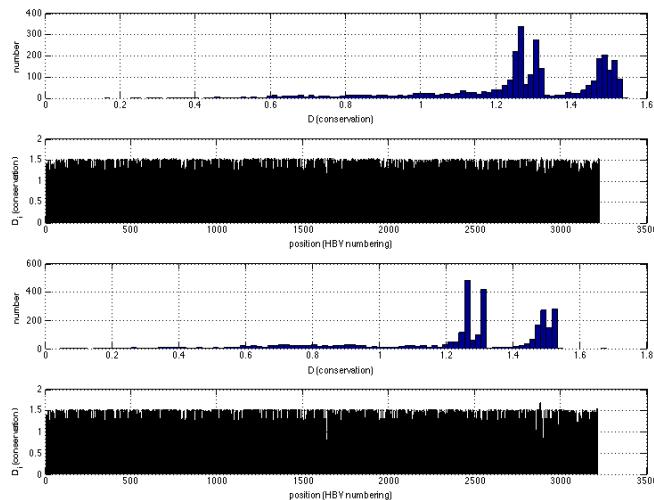


图 2.7 原始序列集合 (裁减高于 gap 频率高于 cutoff 的位点, $cutoff=0.2$) 与 SMOTE($k=5$) 序列集合 (裁减高于 gap 频率高于 cutoff 的位点, $cutoff=0.2$) 的保守性比较

2.4.2.4 HBV 碱基位点之间的耦合 (相关) 度量

HBV 基因组每一个重新调整坐标 (rescaled) 后的位置上的不同碱基的出现的概率分布, 使得我可以使用他们来计算两个碱基之间的协方差, 基于协方差我可以进一步推算位点间相关系数等更多信息。

所以, 定义相关系数 C_{ij}^{ab} :

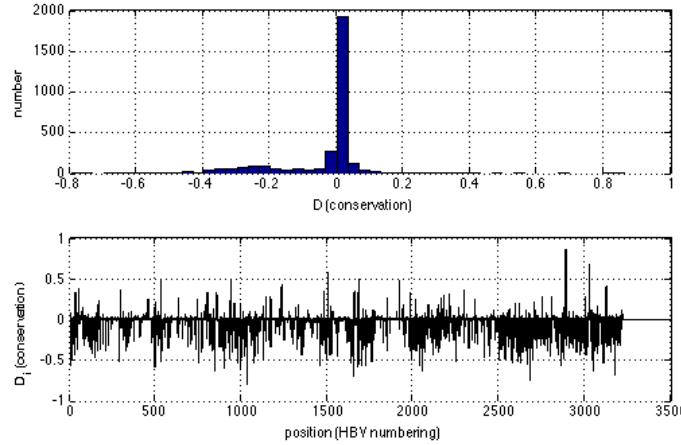


图 2.8 原始序列集合与 SMOTE($k=5$) 序列集合中共同包含的位置 (对应到 3674bp 坐标系下) 的保守性差别; $E = -0.0478696$; 这说明实际上 SMOTE 并没有改变位点对应的保守性, 同时也说明大部分位点的保守性, 这导致了序列的 SMOTE 没有引入过多的序列的位点间差异

$$C_{ij}^{ab} = f_{ij}^{ab} - f_i^{(a)} f_j^{(b)} \quad (2.3)$$

考虑到实际上, 每一个位点的 KL 散度系数, 表明了这个位点是不是容易发生变化, 这样, 我希望能够在每一个 C_{ij}^{ab} 的基础上考虑变化难易程度的权值因子, 这样的话, 才能够找出那些相对保守, 但是实际上显著相关的位点。

然后, 在计算两两位点两两碱基间相关性时, 在 C_{ij}^{ab} 上乘以其中每一个位点关于这个碱基的 KL 散度的一阶导数, 相当于说额外的 bit 数都是有一致的相关性, 越保守的位点之间, 存在的相关就越可靠可以认为 (因为 KL 散度的一阶导数越大, 该位置碱基发生变换的概率也就减小, 同时发生的概率就更小了, 所以, 位点相关的显著性上升)。

$$\hat{C}_{ij}^{ab} = \phi(D_i^{(a)}) \phi(D_j^{(b)}) C_{ij}^{ab} \quad (2.4)$$

其中, $\phi(D) = \frac{\partial D}{\partial f}$, 表示位点保守性 (Conservation) 的梯度。

在实际的操作中, 把计算两个位点的相关性模糊掉碱基维的特性, 同时是用把 $f_i^{(a)}$ 表示为 $X3d_{si}^{(a)}$ 的均值之后,

$$\hat{C}_{ij} = \phi_i \phi_j (\langle X3d_{si} X3d_{sj} \rangle_s - \langle X3d_{si} \rangle_s \langle X3d_{sj} \rangle_s) \quad (2.5)$$

所以, 在我最后得到的 \hat{C}_{ij} 中, 对角线元素不是 1, 而相当于 $\phi_i\phi_j$, 也就是说, 我可以直接做 \hat{C}_{ij} 的对角元素度量相关位点对的 conservation 程度。

2.4.2.5 基于 HBV 碱基间耦合进行位点的统计耦合分析 (SCA)

1. 使 A, T, C, G, 一对应为不同的因子, 然后把 HBV SMOTE 标准序列集合表示为一个 3 维张量, $X3d_{s \times p \times a}$, 其中 s 表示 HBV SMOTE 标准序列集合的第 s 条序列, i 表示的是 MSA 序列中的第 ith 位置的碱基, 而 a 表示的是哪一个碱基 (A,T,C,G 拥有字母对应的特殊的数字索引 $A : 0, T : 1, C : 2, G : 3, - : 4$). 在 $X3d$ 中, 如果满足 HBV SMOTE MSA 数据集中的第 s 条序列的第 i 位置坐标上的碱基 x, 怎对应的 $X3d[s][i][x] = 1$, 其他的碱基索引对应的 $X3d[s][i][x] = 0$. 在 $X3d$ 矩阵中, 只存在 1/0 两个选项。
2. 基于 $X3d$ 矩阵, 我针对每一个位置, 每一个碱基都存在对应的保守性函数的一阶导数: $\phi(D_i^{(a)}) = \frac{\partial D_i^a}{\partial f_i^a} = \ln \left[\frac{f_i^a(1-q^a)}{(1-f_i^a)q^a} \right]$, 所以, 我可以得到 2 维矩阵: $D_{p \times a}$.
3. 通过 2, 我找到每一个基因组坐标上对应碱基的 KL 散度的导数, 来表明对应 HBV 在该位置上发生在这个位点上的碱基突变的难易程度。同时, 我参照 (Efron. and Tibshirani, 1994) 相关矩阵的近似, 使得 $\hat{C}_{ij}^{ab} \sim \frac{1}{M^2} \tilde{C}_{ij}^{ab}$.
4. 通过 $\frac{W_{(i,a)} P_{(i,a)}}{norm}$ 的值, 抹除了 $X3d$ 的碱基维, 把 $X3d$ 投影到 $M \times L$ 的 2 维的 X 投影矩阵上。所以, 我可以使用标准序列集合 3 维张量来确定 HBV 基因组位点间相关性和序列间相关性。

所以, 基于 (4) 中有 $X3d$ 投影得到的矩阵 X , 可以定义 HBV 基因组位点间相关性矩阵 \tilde{C} 和 M 条标准序列集合的序列相关矩阵 \tilde{S} ;

$$\tilde{C} = \frac{1}{M} \tilde{X}^T \tilde{X} \quad (2.6)$$

$$\tilde{S} = \frac{1}{L} \tilde{X} \tilde{X}^T \quad (2.7)$$

2.4.2.6 提取显著耦联位点和显著耦联序列

1. 我分别对 SMOTE 前后的两个矩阵的位点之间的相关性进行分析, 也可以用来分析序列之间的相似性 $C(M \times M)$ 位点之间相关观测矩阵; $S(L \times L)$ 序列之间相关 (Correlation) 测量矩阵。
2. 这样我可以得到要从中观察出实际上的位点之间的相关程度在不同位点对之间得差异 (Variation) 的情况。我需要使用对每一个位点随机 (Random) 洗乱的算法想办法抹除掉位点之间的相关差异, 然后在来评估实际上的位点之间的相关的显著性。
3. 对每一个位点, 在 $f_i^{(a)}$ 不变的情况下, 随机洗乱 (Random Permutation) 每一列的碱基顺序, 以达到抹除实际上 HBV 基因组上面不同位点在进化上的耦联模式。在 100 次随机化试验之后, 我得到了随机矩阵相关位点分值分布 $Cr_{L \times L \times 100}$, 基于这个标准化分布, 我可以标准化我的位点相关矩阵。

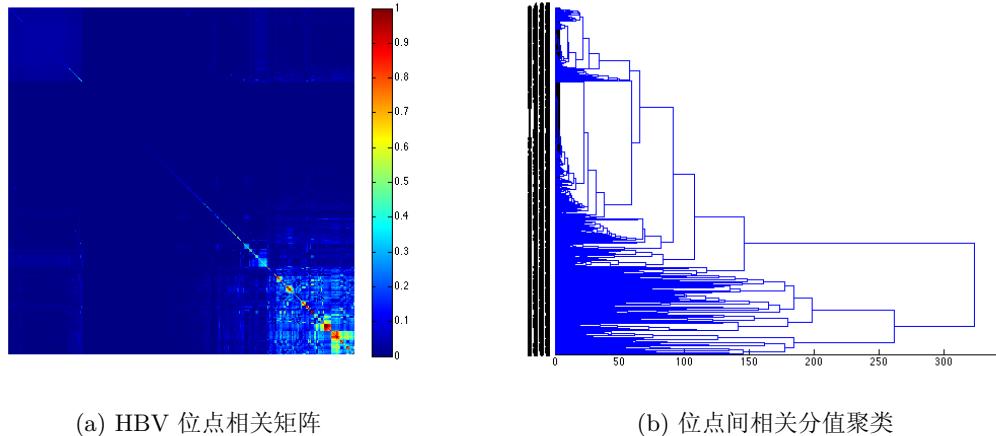


图 2.9 左图: SMOTE HBV 标准序列集合得到的位点相关全局描述 (Landscape), 从数值上说, 相关分值高的位点不多, 但是由于我了解到实际上 HBV 的位点相关矩阵等价于实际相关系数 \times 位点对保守程度, 所以, 不同位点对间的相关分值不只该进行简单的全局对比, 因为各自相关系数的缩放因子不同; 右图: HBV 位点相关的分值的层次聚类, 没有显著的相关分值聚类现象

由于不同位点对间的缩放因子 $\phi_i\phi_j$ 不同, 所以, 我在实验中引入关于 HBV

标准训练集合的 H_0 (零假设 H_0) 来进行位点相关矩阵的标准化工作, 如此之后, 我才能认为得到的标准位点相关矩阵 Z 才具有位点对之间的可比性(即在位点对之间有相较于随即洗乱的 H_0 的显著性).

$$Z_{ij} = \frac{C_{ij} - \langle Cr \rangle_{100}}{\sigma(Cr)_{100}} \quad (2.8)$$

通过上面的方法, 我计算出对因基于位点洗乱方法的标准化位点相关矩阵和标准化序列相关矩阵。

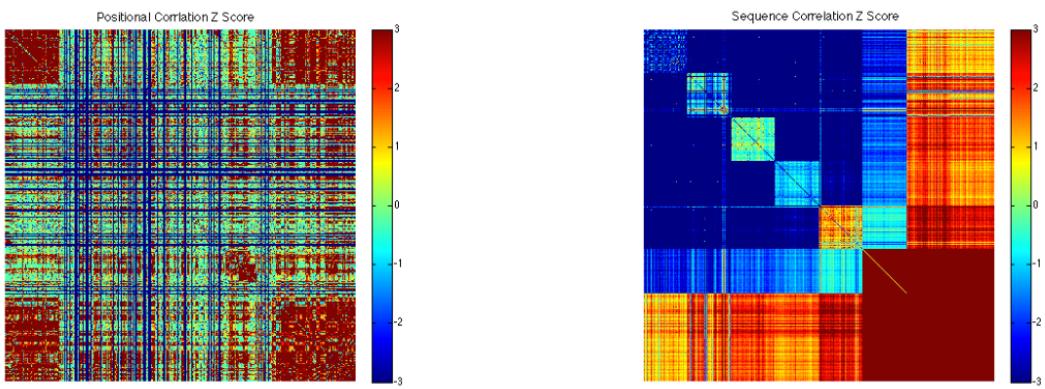


图 2.10 标准化相关矩阵: 左图, 位点的标准化相关矩阵, 在对角线和反对角线上都显示着显著相关的位点对集合的存在; 右图, 序列相关标准化矩阵, 由于序列的保守性本来就很 $\sim 89.0\%$, 所以我没有找到单纯基于位点的重排不能够完全消除掉位点间的相关性。

我只选择位点相关标准矩阵中显著相关的位点对。所以是用 $Z \leq 3.0$ 作为阈值 (threshold), 定义两个相关的位点对作为位点相关网络的一条边. 作为位点相关网络 (*Position,Correlation*) $\sim (V, E)$ 把位点相关矩阵转化为一个位点相关邻接矩阵。存在位点间的显著相关, 说明进化上两个位点在 HBV 基因组上的演化轨迹存在模式依赖, 我可以通过一个端点的变化情况推断另一个端点的变化。相关如此, 我可以通过相关网络的度分布 (图:2.11) 观察每一个位点在计划过程中作为模式决定的输入。另外, 我可以通过寻找网络中的联通份量来确定位点相关模式的传递范围。

2.4.2.7 提取位点相关网络联通分量

无向图的谱聚类使用拉普拉斯矩阵 $L = D - A$.

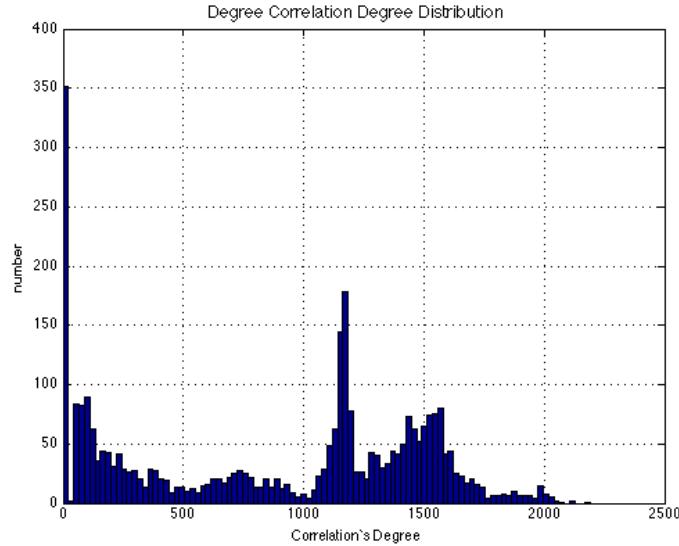


图 2.11 标准化 ($Z \text{ score} \leq 3$) 位点相关邻接矩阵描述的位点相关网络中, HBV 中各位点的度分布。整个位点相关网络的平均度: average degree = 875

D 是度矩阵, 是对角阵。 A 是邻接矩阵。是 Z 标准矩阵的 0/1 模糊矩阵。所以, $D = \text{tr}(Z^T Z)$.

同时我们可以得知 L 满足: 1, L 矩阵是一个半正定矩阵。2, L 矩阵存在着特征值为 0 的, 且 0 特征值对应的是不同的联通子图。

计算拉普拉斯矩阵 L 的零特征值特征向量*

$[V1 \ Q1] = \text{eigs}(Lp, Dp, 1500, 'SM', \text{opt});$ 对应对矩阵是否实对称阵的描述 opt

通过拉普拉斯矩阵的方法, 我一共找到了 15 个特征值为 0 特征向量, 分别对应表示这 HBV Genome 内部的 15 个联通分量 2.12。

2.4.2.8 HBV 位点关联矩阵和序列关联矩阵的特征向量分析

\tilde{C}, \tilde{S} 都是通过 $X3d$ 在序列维 M 和碱基位点维 L 上的投影 $X \ D_i^{(a)}$ 而来, 所以, 我通过 \tilde{C}, \tilde{S} 的数学描述, 我还可以等价的认为位点相关矩阵 \tilde{C} 和序列

*同时注意, 在 matlab 进行关于特征值的计算过程中, 可能会遇到关于的矩阵 Matrix 是奇异。解决方案: 由于当 A 是奇异阵, 胆识 $I+A$ 不是, eigrnvalue-1 就可以。这样的话, 对应的特征值相当于集体 +1。同时我要找的 Cluster 也会发生变化。(是不是不需要这个), 我只想要在进行一个统计来估计可分型位点的 average degree 就可以了, 最后我在最小的 1500 个特征值里面找到了相应的 0 特征值向量。

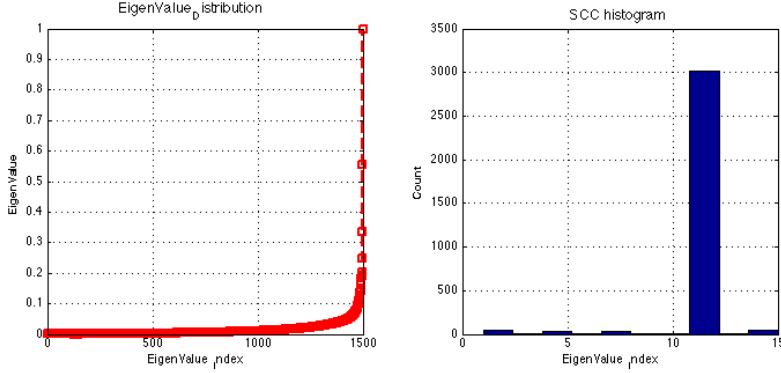


图 2.12 左图：通过特征值求解，我找到了 15 个特征值为 0 的特征向量，对应 $A+I$ 矩阵的 $131_{th} \sim 145_{th}$ 的特征向量；右图：在我找到的 15 个特征向量中，实际上聚类后得到： $\frac{2968}{3205} = 92.61\%$ 的位点都在 12_{th} cluster 中，而这部分位点是保守性显著较高的位点，相当于 HBV 基因组的框架，而分型的位点，很可能集中到其他的 14 个位点联通分量之中。另外，如果我是用 HBV 标准序列集合当作 HBV 对其基因型库的采样，我假定基因亚型作为不同基因采样的约束输入，Barabasi et.al 在最新的工作中发现，可以只使用强联通分量里的位点子集作为来实现对输入的观察。

相关矩阵 \tilde{S} 都来自 X 的分解矩阵。

$$X = U\Sigma V^T \quad (2.9)$$

可以得到 U 和 V 两个 $M \times M$ 和 $L \times L$ 两个酉矩阵，所以，我可以基于 X 的奇异值分解，得到对应的在 HBV 基因组序列维 M 和位点维 L 的特征值：因为：

$$\tilde{C} = \frac{1}{M}(U\Sigma\Sigma^T U^T) \quad (2.10)$$

$$\tilde{S} = \frac{1}{L}(V\Sigma^T\Sigma V^T) \quad (2.11)$$

因为 U, V 都满足 $U^T = U$ ，所以，我发现 $\frac{1}{M}\Sigma\Sigma^T$ 和 $\frac{1}{L}\Sigma^T\Sigma$ 分别对应了序列相关矩阵和位点相关矩阵的特征值。所以 U, V 对应了 \tilde{C} 和 \tilde{S} 各自的特征值。分别被表示为 $|U_1\rangle, \dots, |U_L\rangle$ 和 $|V_1\rangle, \dots, |V_M\rangle$ 。

因为，我们发现在 SMOTE 标准序列的最大的几个特征值处存在着最为显著的序列位点关系，所以，我们可以先取出 Top 3 的特征值对应的位点权值向量，Top1, Top2, Top3 特征值对应的特征向量坐标投影；我们发现，由于各自的

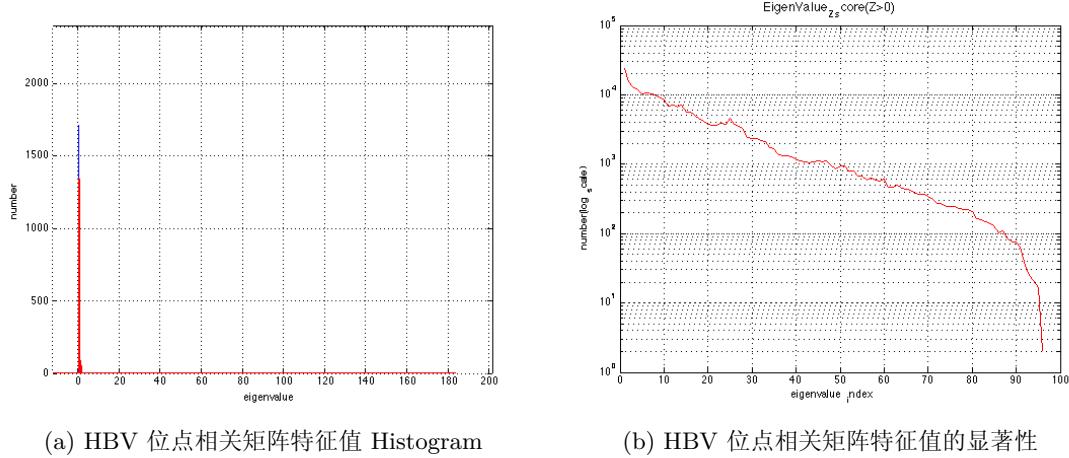


图 2.13 左图：以随机位点相关矩阵作为背景分布（蓝色），对比 HBV 采样集合的特征值分布与随机矩阵特征值分布的差异；右图：显著性上升的特征值，对于位点来说，前 95 个特征向量（Eigenvector）都具有统计显著性。其中每一个特征向量代表的位点权值说明了其中每一个组成位点的权值（与这个特征向量的偏相关系数），也表明不同的位点之间在每一个正交基底下的相关程度

权值关联位点自身的保守度，这会导致我看到的特征向量权值对应的实际数值很小，一方面，从很多的位点都集中在 0 附近表示大部分位点之间并不存在显著的关联。另外一方面，我们需要找到实际上的相关位点集合 [2.142.152.16]：

虽然在考虑了位点对 (site-pair) 对应位点的保守性之后，实际上各个特征向量对应的线性组合时位点的权值，较小，但是这只能说明 HBV 基因组相对来说全局结构较为保守，所以 $c\psi(D_i)$ 对位点相关矩阵的缩放不是标准的相关度量。

假设实际上 HBV 基因组中的各基因序列表现其实时由相互独立的 X 个位点集合作为独立信源组合的结果，所以我们得到的 HBV 标准序列集合所表现出的位点相关矩阵实际上时这些独立位点集合的线性组合：也就是我们由 $C = Ws$,

1. 基于位点相关矩阵的特征值分解，按照特征值从大到小排序 \tilde{C} 对应特征向量 $|C_1\rangle, \dots, |C_M\rangle$
2. 使用基于位点的随机排列方式，构建 HBV 位点相关矩阵的零假设 (Null Hypothesis, H_0)，然后选择处最显著的前 k 个特征向量： $|C_1\rangle, \dots, |C_k\rangle$, $k \leq M$ ，同时，top k 个特征向量可以解释尽可能多的特征差异。

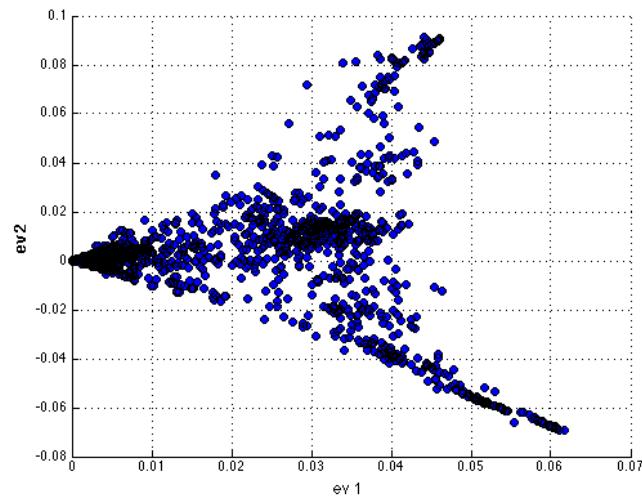


图 2.14 位点相关矩阵两维特征向量的各位点权重分布散点图, 第 1 特征向量 *v.s* 第 2 特征向量

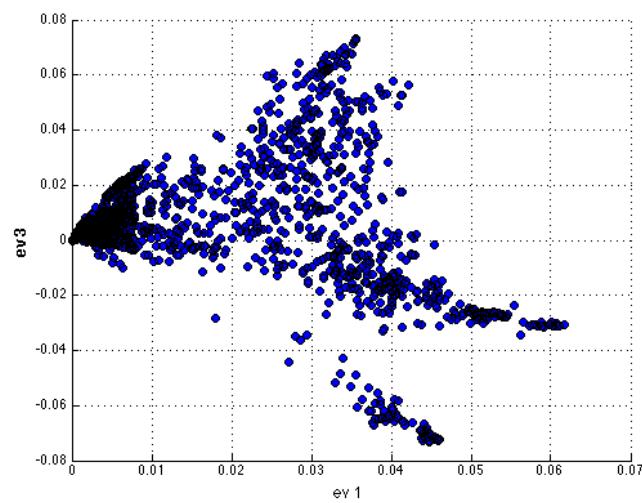


图 2.15 位点相关矩阵两维特征向量的各位点权重分布散点图, 第 1 特征向量 *v.s* 第 3 特征向量

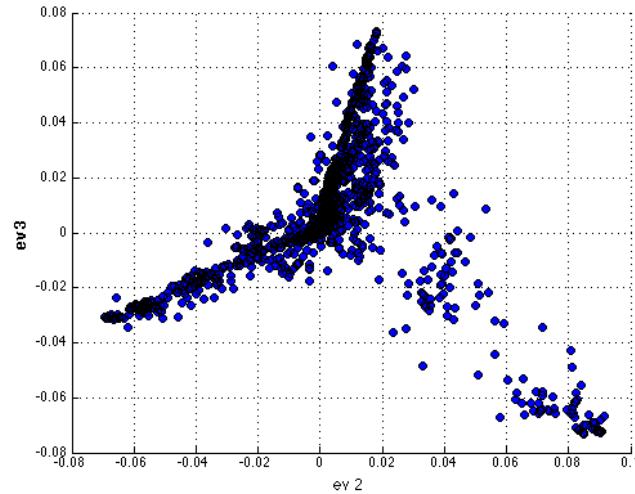


图 2.16 位点相关矩阵两维特征向量的各位点权重分布散点图, 第 2 特征向量 *v.s* 第 3 特征向量

3. HBV 位点相关矩阵的前 k 个特征向量组成 $k \times M$ 的混合信号矩阵 Z , 我们假设相关位点的信号, 这些相关位点的方差最大的 k 个特征向量实际上有相互独立的 k 个独立信号源通过线性变换得到的, 独立成分分析 (ICA) 可以通过迭代的方式学习得到信源向量的线性变换矩阵 W^{15} , 这个算法设定 $W_0 = I_k$ 作为去信号混合矩阵 (unmixing matrix) 的启发初始值, 然后, 我们按照 A. J. Bell et.al 提出的 W 更新方法, $\Delta W =$

$$\Delta W = \epsilon \left(I_k + \left(1 - \frac{2}{1 + \exp(-WZ)} \right) (WZ)^T \right) W \quad (2.12)$$

然后, 通过对最显著的 k 个特征向量的迭代独立成分学习, 我们可以找出表达为实际上位点相关的 k 个独立信源向量, 而在这 k 个独立信源向量里面, 每一个信源都是由 3205 个 HBV 基因组位点基于 W 线性组合形成的, 每一个位点对于每一个独立成分的贡献都可以被 s 对应的权重记录下来。基于每一个位点对独立信源的贡献值, 我们可以确定出整个 HBV 基因组位点中实际上作为 HBV 位点演化信号的位点集合。然后, 为了能够最大程度的还原实际采样的信号 Z 的同时, 只是用少量的实际上在信号源位点向量中不为 0 的位点即可。对 W 取逆元 $A = W^{-1}$, 我们可以得到任何模糊对应的, 对于 ICA 对应的原

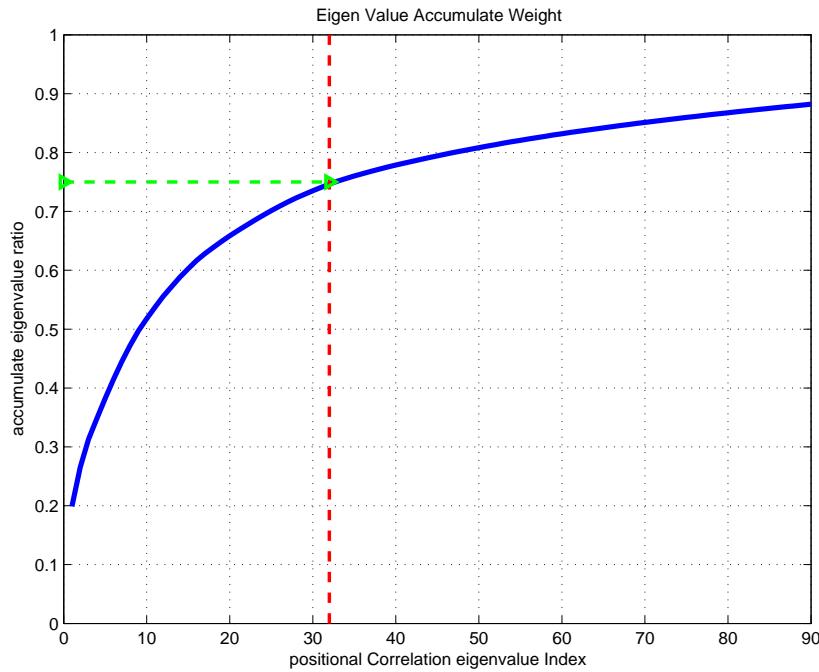


图 2.17 基于 HBV 标准序列集合得到的位点相关矩阵进行特征分解，我们发现 top 95 个特征向量都是显著的特征向量，但是实际上，使用相关矩阵来描述整个 HBV 基因组上位点之间的相关，我们可以只保留 $\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i} \geq 0.75$ ，甚至更高的位点，所以，在 HBV 位点相关序列集合中，我们可以把相关位点对应的正交基底规约到 32 个特征向量（以 0.75 作为规约的阈值）。

信号原，我们发现，只需要保留 $ICA\ weight \geq 7.900$ 或者 $ICA\ weight \geq 3.100$ 的位点对应不同信道的信息，我们就可以还原 85% or 95% (图:2.18) 的原始 HBV 标准序列集合对应的位点相关矩阵信息。所以，我们只需要保留在 ICA 还原的 S 矩阵中权值大于 7.900 的位点即可。基于对还原独立成分中信号权值的分类，我们可以从 HBV 的位点中提取出压缩部分的位点索引 (图:2.19)，并且，我们可以发现，完全还原 HBV 标准序列集合的位点，均匀的分布在 HBV 基因组的重新定义 (rescaled) 的坐标上。也就是说，确定 HBV 我们可以通过 $\frac{517}{3205}$ 的位点作为信号源完全还原 HBV 标准序列集合对应的完全位点相关矩阵。又因为位点相关矩阵和序列相关矩阵相当于都是来自与 $X3d$ 矩阵投影的 SVD 正交阵，所以，这也就告诉我们，使用 $\frac{517}{3205}$ 的位点就可以完全恢复 HBV 标准序列集合的序列相关矩阵，确定序列间差异。

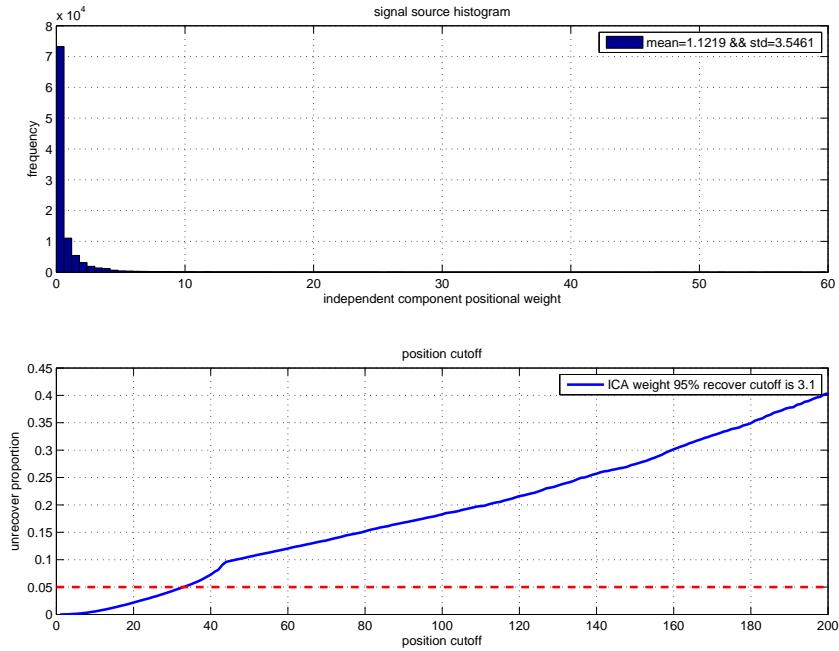


图 2.18 上图：发现 ICA 盲信号恢复后，发现大部分位点对独立成分（信源）的贡献均为 0，这说明表示大部分位点在信号识别过程中时冗余的；下图：我们取 W 的你矩阵，使用不同的阈值模糊化整体 ICA 信源信号组合矩阵，对独立信号的贡献权值 $\leq cutoff$ 的位点置 0(模糊处理)，重新定义坐标的其他位点对应的权值。实际信号的恢复程度，恢复信号与实际信号之间差异使用 2 阶范式 $\|X\|^2$ 表示，我们可以针对不同的要求，得到对应的对各个独立成分中基于权值的选择各个位点，最后，各个独立成分中只保留权重 \geq 某一个特定阈值的贡献位点，所以，我们可以发现保持位点间相关矩阵的信号，我们只需要使用全基因组上的少许位点即可。

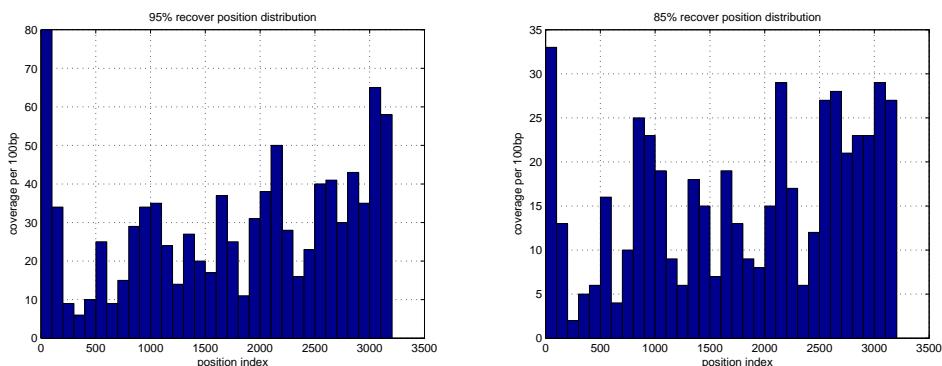


图 2.19 分别时恢复 95% 和 85% 原始序列集合相关信息时，必要的 HBV 基因组位点分布，整体上代表位点均匀的分布在整個 HBV Genome 上，平均每 100bp 存在 25 ~ 30bp 的信源位点，只要监测到这些位点，就可以完全恢复 HBV 序列集合的相关信息

所以, 通过 HBV 基因组的位点关联分析, 我们可以发现, HBV 基因组的差异信息实际上被包含在了少数独立的信源相关位点里面, 而这些位点可能存在同一个位点同时包含了多个来自不同信道的信源的线性组合。也就是说, 在整个 HBV 基因组上决定每一条序列构成的不同信源只需要使用整个 HBV 基因组上 25% ~ 30% 的位点。所以, 只要我们能够单独的检测出这些位点的实际信息, 我们就可以完整的判断整个 HBV 基因组的序列差异信息。不同的序列位点信源的组合还将决定不同的序列信源。

因为位点相关矩阵 C , 和序列相关矩阵 S 都是来自 $X3d$ 矩阵在 $L \times M$ 平面上的投影矩阵 X , 所以, 除了了解 C 表示了进化上位点之间的相互约束信息之外, S 矩阵应当描述了 HBV 基因组序列的不同信源信息; 我们使用相同的方法分析 HBV 序列相关矩阵 S :

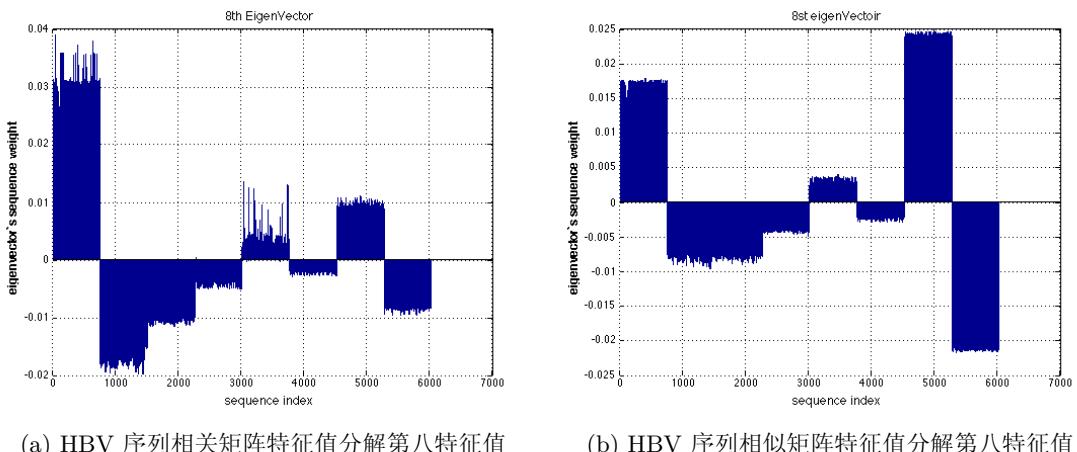


图 2.20 直接对 HBV 序列相关矩阵进行特征之分解, 我们发现起前 46 个特征值相对于随机 H_0 假设背景显著增强。同时, 我们观察各个特征向量下各序列对应特征向量的权值, 发现 HBV 序列相关矩阵在第 8 特征值精细结构上存在这优良的 HBV 亚型分型能力, 但是如果我们将位点的 KL 相关熵直接区分序列相似矩阵, 则 B 亚型和 C 亚型之间难以区分; 从左往右, 不同的特征向量权重内对应的是不同的 HBV 亚型, 依次为: $A \rightarrow C \rightarrow B \rightarrow E \rightarrow D \rightarrow G \rightarrow F \rightarrow H$

受 HBV 序列相关矩阵的启发, 我认为, 这解释了为什么 B, C 亚型存在难以区分和易于被认为是混合的情况, 因为传统的分型方法都是基于序列相似矩阵, 而不考虑位点本身在进化过程中表现出来的保守程度。这样的情况下, 我们没有办法很好的区分 B, C 亚型, 但是如果我们将每一个位点之间的耦合关

系以及对应的 KL 熵, 作为位点在分型过程中对序列差异的贡献权重, 就能够将 B, C 亚型区分开来; 同时, 从 C 可以显著的区分开 HBV 基因的不同亚型来看, 我认为不同的亚型可能就是不同的信源, 本身对 HBV 序列集合基于序列方式的采样就是 8 个正交的亚型信源采样混合的结果; 所以我们通过 ICA 分析进行 C 和 S 的 posICA 和 seqICA 分析*。发现, 不同的 HBV 基因型亚型在原始信源分析中会被正交的划分开来。这也就是说, 实际上不同的 HBV 亚型就是不同的序列差异信源, 代表着各个不同的进化上的“Physical Source”。

同时, 对比使用 SMOTE 消除标准序列几个各个不同亚型之间的不平衡和不消除 HBV 各个亚型间不平衡的结果, 我们发现, 如果不消除不同的 HBV 序列亚型间的不平衡, 大部分亚型之间仍然可以被区分出来 (2.22)。不同的 HBV 亚型在不同的网格线内部对应不同的独立成分向量上的显著的权值。说明只要我们能够认定来源的 HBV 的数目, 即实际上独立的 HBV 亚型有多少种, 就可以还原出不同的 HBV 亚型, 从我们的分析中也可以得出: HBV 的不同亚型之间是相互独立的。

同时, 我们还发现, 使用独立成分分析, 我们可以确定不同的 HBV 亚型可以看作是独立的信号源和信道。基于不同的 HBV 亚型序列实际上在进化上相互正交, 同时, 考虑到我们分析发现, 实际上, 整个序列上相关的位点是均匀分布的, 所以, 我们通过随机生成 *breakpoint* 的方式拼装了来自不同亚型的序列, 因为实际上存在 8 个亚型, 所以, 我们一共可以生成 $\binom{8}{2}$ 对亚型配比, 每一种组合生成 100 种序列组合。我们随机在之前生成的平衡 SMOTE_k_5 序列集合里面等概率的抽出各种亚型对应序列, 然后在随机在基因组上抽取一个重组位点生成新的重组序列 (Chimera), 基于这个假设, 猜想很可能实际上每一个不同的嵌合体都可以被表述作不同的‘Physical Source’的线性组合; 我们使用我们人工随机合成的 $\binom{8}{2} \times 100$ 的人工合成序列, 并且把这些序列对齐到标准序列集合所对应的坐标系下;

使用独立成分分析所有合成序列对应的独立信道的具体情况:

IC_P1: 都包含 B 亚型; IC_P2: 都包含 H 亚型; IC_P3: 都包含 G 亚型;
IC_P4: 都包含 D 亚型; IC_P5: 都包含 F 亚型; IC_P6: 都包含 C 亚型; IC_P7:
都包含 E 亚型; IC_P8: 都包含 A 亚型

* posICA: 位点间独立成分分析; seqICA: 序列间独立成分分析。

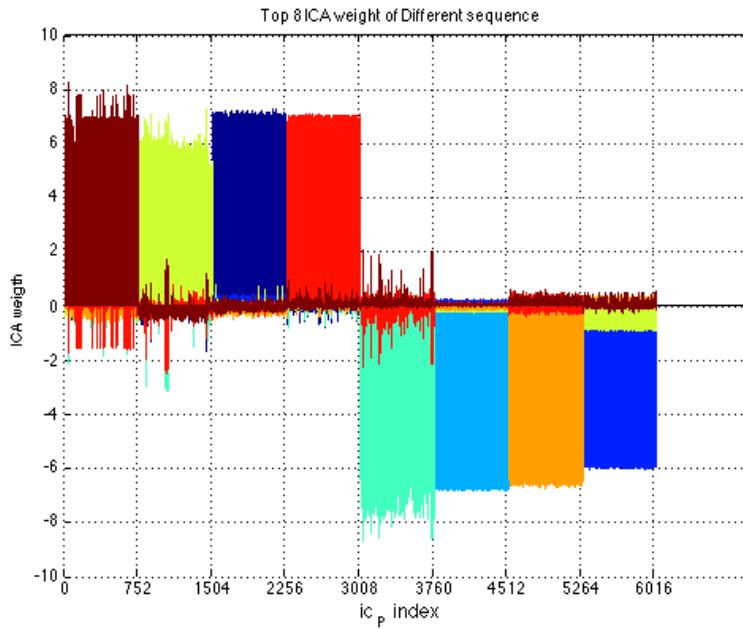


图 2.21 使用独立成分分析可以逼近一个 W 来线性组合不同的信道, 我使用 ICA 盲信号处理方法恢复出来的 k 个 ($k \geq 8$) 个独立信源, 总是可以被表示为 8 个 HBV 亚型作为主导的相互独立的信源。所以, 认为, 亚型分类代表了不同的物理信源, 不同的亚型在进化上相对独立。对应相同投影矩阵推算出的位点相关矩阵, seqICA 的主要解释位点意味着在 HBV 分型上区分不同的 HBV 亚型, 从左往右, 不同的颜色代表了不同的独立成分向量, 也是不同的 HBV 基因亚型, 他们分别是不同的独立成分内部的权值, 上图中, 从左往右, 不同的 ICA 独立分量权重刚好对应的是不同的 HBV 亚型, 依次为: $A \rightarrow C \rightarrow B \rightarrow E \rightarrow D \rightarrow G \rightarrow F \rightarrow H$

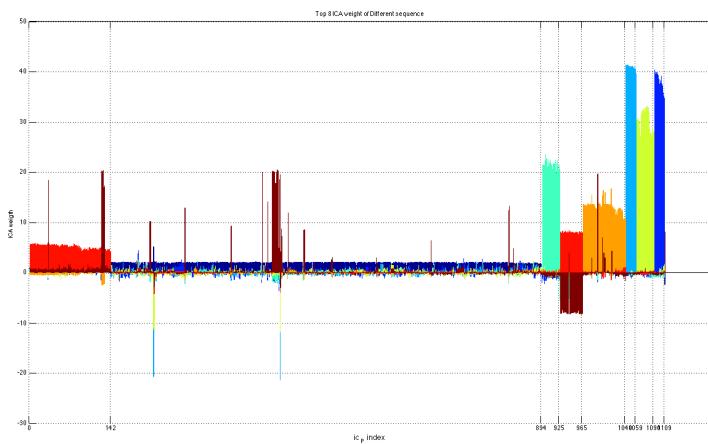


图 2.22 在不平衡的随机采样序列集合上, 对整个 HBV 亚型进行分类的结果, 不同的颜色对应不同的 HBV 亚型分类信号

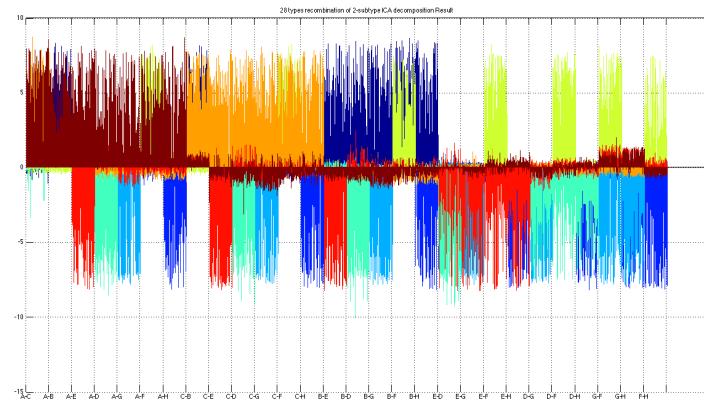
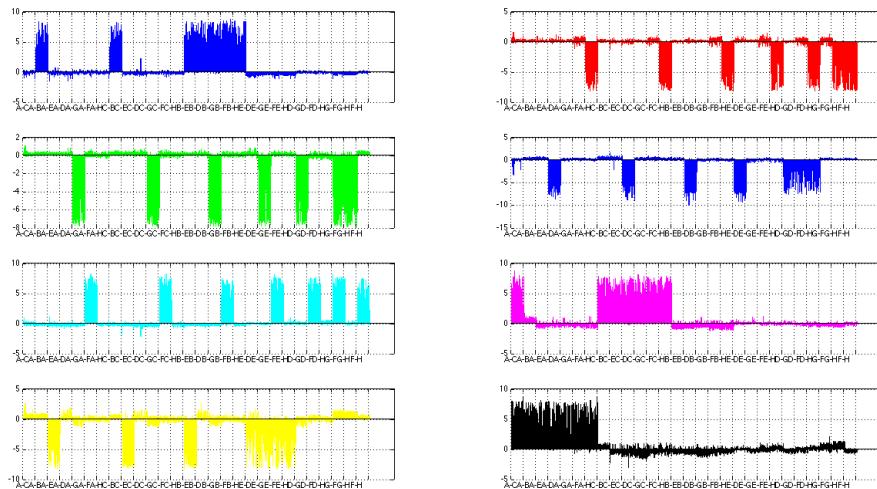

 图 2.23 $k=8$ 分离出所有序列对应 8 个独立信道权值分布


图 2.24 对 2800 条人工合成的序列进行独立成分分析之后得到的正交独立成分具体分布情况；实际上，我们发现，对于不同的亚型之间的重组思想来说，每一个两亚型重组序列都还原出两组较为显著的权值分布来。也就是说，对于各个亚型来说，重组嵌合体实际上是各个亚型独立信源的线性组合的假设与实际情况互相吻合。各个信道表示的是各个亚型重合，实际上表示的是：从上至下，从左至右的顺序各自描述的亚型分别是 $B \rightarrow H \rightarrow G \rightarrow D \rightarrow F \rightarrow C \rightarrow E \rightarrow A$

通过(图:2.24),可以发现,从上至下,从左至右,我们可以发现,每一个Independent Component 都单独代表一个HBV 亚型的特征,即使这个位置对应的是一个HBV 重组型的一个主要成分(从上至下,从左至右的顺序各自描述的亚型分别是 $B \rightarrow H \rightarrow G \rightarrow D \rightarrow F \rightarrow C \rightarrow E \rightarrow A$)。所以,我们可以得到,通过位点相关矩阵的投影,我们可以得到不同的重组型的 Parents Information source 信息。同时,每一个HBV 序列分解后对应的 IC 的权值(weight)实际对应了组成不同HBV 亚型带来的亚型间在HBV 基因组上的比例。由于合成序列过程中是使用不同的序列随机选择 breakpoint 而合成的,所以,我们也可以说明,对应的HBV 亚型的可分型位点在整个HBV 基因组上是随机分布的。而重组,就表示为在这个序列上,两个正交的独立信源对于这条序列在序列相关矩阵的中都有贡献。并不是单纯的落在这个亚型空间正交基底的各个亚型信道基底上。

同时,由于实际上我们可以完全的恢复出每一条合成序列实际对应的各个信道的组合,我们发现,使用欧式距离作为每一个信道对各条序列权值的度量,我们可以完全通过各个亚型的贡献,来推算出实际上重组发生的大致位置。我们完全恢复了 81.07% 的重组序列的亚型分布情况。同时我们发现,恢复的各个重组位点的权值分布与实际的各个亚型的贡献之间本身存在这显著的线性关系(图:2.25),也就是说,我们可以通过 ICA 恢复出来的各个亚型信道对各序列的贡献,来估计实际发生重组的位点。同时,我们发现剩余的 18.93% 未恢复的序列中,亚型在重组中的各自贡献有显著差异,并且几乎所有的序列位点都显著的来自于单一亚型(主导重组亚型占比例: 92.73%)(图: 2.26)。所以,使用 ICA 的方式可以很好的还原出 HBV 各亚型形成的重组体。

2.4.3 基于 HBV 基因组亚型特征的序列分型

使用 SMOTE HBV 标准序列集合可以找到位点相关矩阵, 和序列相关矩阵, 并且能够通过序列相关矩阵还原 HBV 的各个亚型, 所以, 基于 HBV 标准序列集合的先验信息, 我们可以构建一套只依赖于来自独立信道的位点组成的短片段的 HBV 亚型分型模式。从 [2.4.2.8] 的独立成分分析可以看出, 完全能够保持序列间进化约束信息只需要 HBV 全基因组中 $\sim 16\%$ 的序列位点, 同时, 如果我们实际考虑的序列差异程度不是基于准种, 而只是亚型的话, 实际上需

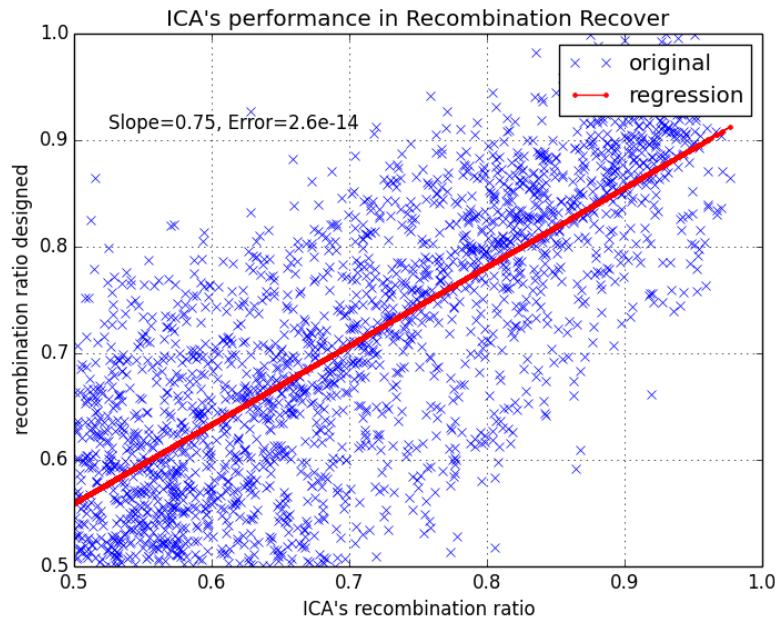


图 2.25 独立信道分析提取出的各个亚型权值暗示这亚型重组配比的信息，实际重组位点(序列长度配比)于恢复出的各信道的权值比例呈现显著的线性相关

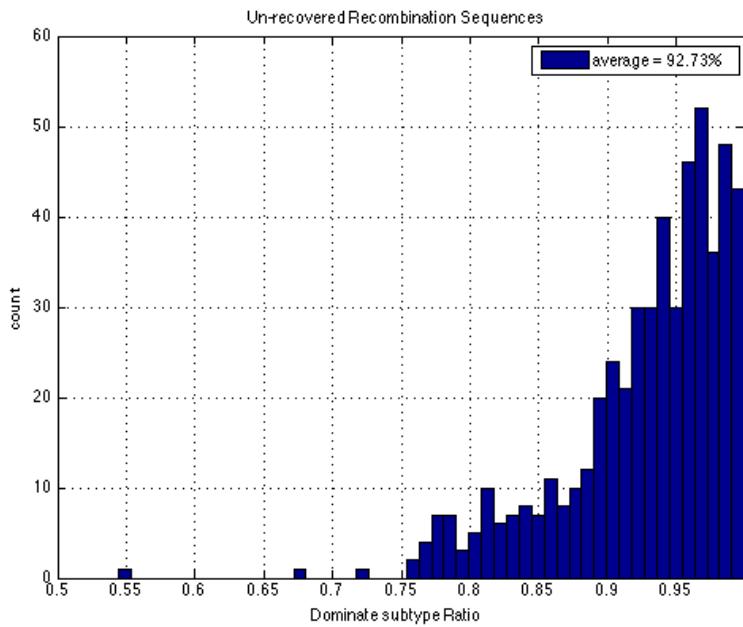


图 2.26 未能恢复出重组的合成亚型序列 (18.93% 输入重组序列) 中，主导序列在全长中的比例，实际上重组 90% 的偏离向主要亚型导致了无法区分出该合成序列对应的第二正交信道。

要使用恢复亚型的位点应当 $\leq 16\%$ 全基因组序列。

又因为我们之前发现 (图:2.21) 可以通过 C 矩阵相关的序列关联矩阵 S 分离不同的亚型为独立信源, 所以我们认为分离独立信源的位点是有限的, 同时, 由于序列独立成分矩阵实际上反过来也可以看作一个特征向量的线性组合, 因为相关矩阵的特征值分解 (PCA) 本质上等价于 (K -means) 聚类方法⁷, 所以, 我们在进行外部序列归属时, 可以等价于扩展 K -means 聚类方法, 这就等价于我们观察不同的目标序列和不同独立信道 (亚型) 中心的距离。

所以, 我使用了 PSSM 位置特异矩阵, 来表示每一个 HBV 亚型的中心.

2.4.3.1 PSSM 矩阵的构建

对于每一个位点, 我们完全的抹除不同的 HBV 亚型之间由于种群采样带来的位点碱基分布上贡献权值的差异, 我们只考虑各个亚型内部的, 在标准序列集合对齐坐标系下, 每一个位点的碱基的概率分布。然后建立出形如 (图:2.27) 的 3 维 PSSM 张量。

所以, PSSM 是 HBV 亚型内部碱基分布差异的表征方式。首先, 我们计算 PSSM 在不同的亚型带来差异, 这个不同亚型不同位点的碱基概率分布, 可以被认为就是实际上标准序列集合的一种平均中心表述方式。同时, 我们也可以发现存在一些位点, 他们的碱基频率分布会随着不同的亚型而显著不同, 而实际上目标序列和各个亚型 cluster 的距离主要由这些位点决定。已知, HBV 的序列相关矩阵下表示 HBV 总共由 8 个独立信道的信源, 所以, 我们认为如果要求区分开完整的 8 个 HBV 亚型, 我们的序列位点差异必须要能够在 $\binom{8}{2}$ 对 HBV 亚型对中表示亚型间可区分, 所以我们要能够归属一段序列的基因型, 至少需要 28 个位点的 28 对亚型碱基差异信息。

另外, 因为我们由 $PSSM_{L \times 8 \times 5}$ 矩阵, 8 表示由 $A \sim H$ 8 个 HBV 亚型, 5 表示对应的碱基空间 ($A, T, C, G, -$), 所以, 我们希望在每一个位点上, 对 C_8^2 对 HBV 亚型的碱基概率分布向量进行度量, 为了整个工作的前后一致, 我们使用欧式距离来度量同一个位置上, 不同亚型的碱基平均分布空间的差异。

$$E_{ab}^i = \sqrt{\sum_{j=1}^5 (PSSM_{(i,a,j)} - PSSM_{(i,b,j)})^2} \quad (2.13)$$

a, b 代表不同的 HBV 基因型, j 代表 PSSM 中碱基维的碱基索引。 i 表示 HBV 标准序列集合上对应的位置坐标。所以, 每一个位点相对于不同亚型带来的碱基空间概率中心差异, 可以作为评价该位点分型优良性质的度量。

[Position : 871]		A	T	C	G	-
Subtype						
A	0	0	0	1	0	
B	0	0	0	1	0	
C	0.00148368	0	0.00148368	0.9970326	0	
D	0	0	0	1	0	
E	0.05555556	0	0	0.9444444	0	
F	0	0	0.03571429	0.9642857	0	
G	0	0	0	1	0	
H	0	0	0	1	0	
[Position : 872]		A	T	C	G	-
Subtype						
A	0.00E+00	1	0.00E+00	0.00E+00	0.00E+00	
B	0.00E+00	1	0.00E+00	0.00E+00	0.00E+00	
C	0.00E+00	0.99406528	0.00593472	0.00E+00	0.00E+00	
D	0.00E+00	1	0.00E+00	0.00E+00	0.00E+00	
E	0.00E+00	1	0.00E+00	0.00E+00	0.00E+00	
F	0.00E+00	0.96428571	0.03571429	0.00E+00	0.00E+00	
G	0.00E+00	1	0.00E+00	0.00E+00	0.00E+00	
H	0.00E+00	1	0.00E+00	0.00E+00	0.00E+00	

图 2.27 每一个位点上, 八个 HBV 基因亚型的碱基分布, 即位置特异矩阵, 可以被认为是不同 HBV 亚型的小类的中心平均描述。

同时, 我们发现以 0.7 作为层次聚类的阈值得到的是健壮 (Robust) 的分型结果 [图:2.29]. 所以, 我们可以进一步观察可分型位点在 HBV rescaled 坐标系下的分布情况 [图:2.30]. 同时, 在通过位点 PSSM 碱基空间向量的欧式距离找出 889 个可分型位点。我们观察发现可分型位点的保守性在可分型位点和不可分型位点之间的差异 [图:2.31]

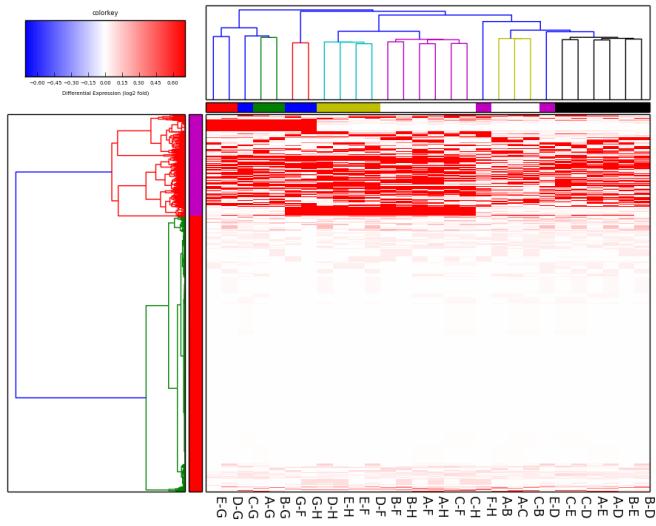


图 2.28 全 HBV 基因组上的位点在不同亚型碱基空间中心向量上由显著的分类面, 取 $0.7 \times \text{Max Distance}$ 为 flat clustering 的阈值, 我们发现可以正好可以把 HBV 基因组上的位点分作另个部分, 我认为这刚好就是可分型位点和不可分型位点的区别。

2.4.3.2 使用稀疏矩阵求解合适分型的短片断窗口

从 (图:2.28) 对应热图, 我们可以统计出每一对基因型通过 PSSM 向量的欧式距离判别为可基因型间可区别或不可区别的阈值:

所以, 我们定义隶属矩阵 $A_{28 \times 3674}$, $A_{ij} = 1$, 如果 PSSM 上的第 j 个位点在第 i 对亚型配对产生的碱基空间向量间欧式距离的平方 ≥ 0.9 , 否则, $A_{ij} = 0$. 使用这个方法, 我们得到的 A 矩阵描述着 HBV 基因组上所有位点的分型能力 (图:2.33)。

于是, 我们在 3674 个位点如果包含就在 x vector 的 $x_{i_} = 1, else = 0$, 这样的话, 我们可以找出合适的一个窗口 $(0, \dots, 0, 1, \dots, 1, 0, \dots, 0)$ 满足所有可以区分不同亚型的位点都被包含于其中。这里有点像是对第一步里面的那个 distance 做了模糊化之后的结果。 Ax 的结果是一个不含 0 的列向量。

为了找到特定的片段集合, 我们要求 $Ax = b^T$ 中的 b 列向量中不包含非 0 项, 所以, 可以假设 $b = (1, 1, \dots, 1)$, 因为 x 是 3674 维向量, 同时, $A_{28 \times 3674} : rank(A) = 28$, 所以, 我们这个方程存在着无穷多解. 如果我们找到一

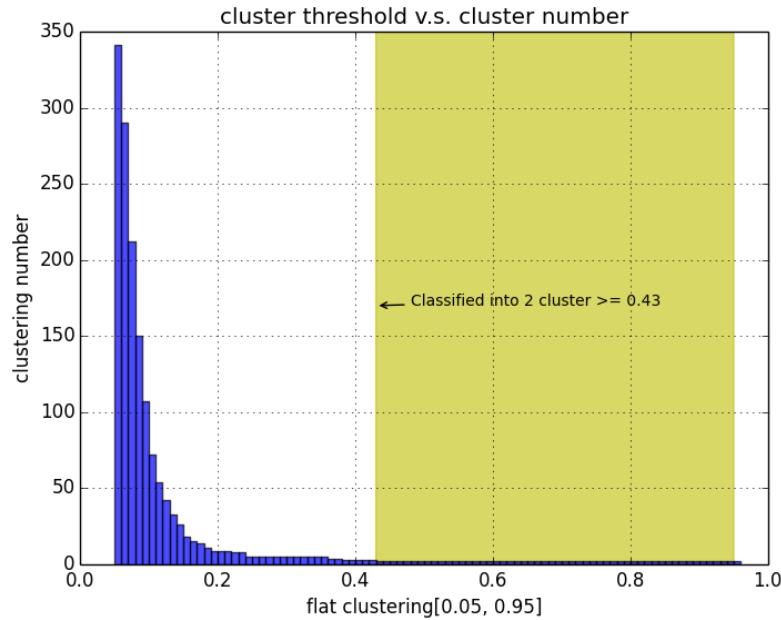


图 2.29 HBV 每一个位点的亚型间欧式距离可以被描述为层次聚类的输入向量, 同时, 我们控制不同的输入向量间距离作为 flat clustering 的阈值, 遍历 $cutoff \in (0, 1) \times Max\ Distance$, 我们发现 57% 的情况下, 都可以把位点集合作两类; 同时这也符合预先设定中的位点标签种类数 \in (可分型, 不可分型)

个满足条件的 x_0 , 那么 $Ax=b$ 的通解可以被表示为:

$$x = x_0 + x' \quad (2.14)$$

$$Ax' = 0 \quad (2.15)$$

但是, 在实际操作中, 我们需要的可分性片段除了满足 $Ax = b^T$ 之外, 还收到非 0 项在基因组坐标系下连续这个约束, 所以, 对于实际上我们设计不同的 x 信号采样向量, $x = (0, 0, 0, 1, 1, \dots, 1, 0, 0, \dots, 0)$; 对于不同的 HBV 片段长度, 我们可以直接通过暴力枚举的方式, 来确定 HBV 基因组上不同的片段的 8 亚型间的区别能力。设计不同的分型片段之间的间隔, 我们可以得到不同的窗口长度片段, 在 HBV 基因组全局下, 在 HBV 基因组不同的位置上的分型能力。由于窗口的长度增加, 会导致实际上 HBV 基因组上窗口数目减少, 所以我无法使用一个完整的 3 维张量来直接描述不同的片段长度下, HBV 基因组全局的分型能力 [图:2.34].

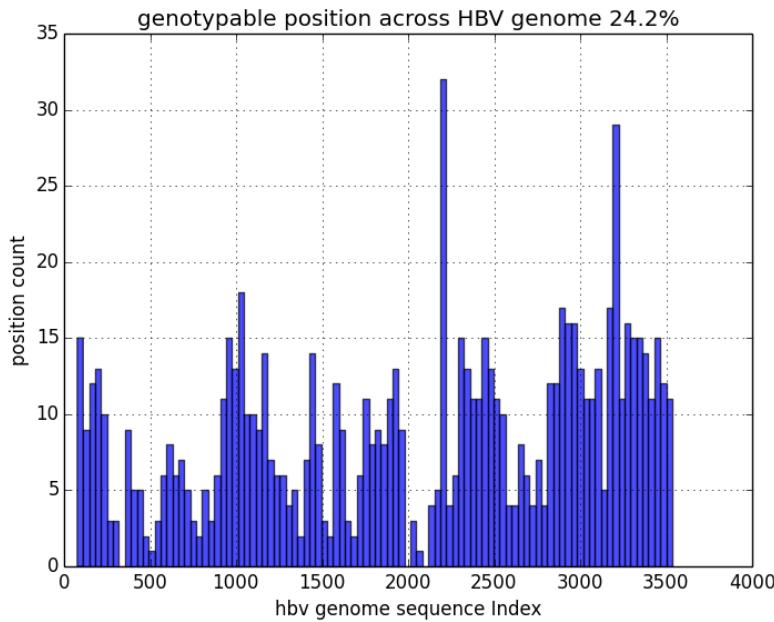


图 2.30 上图：我可以看到，存在某两种亚型间划分能力的位点均匀的分布在了 HBV 的全局基因组上，平均每 100bp 碱基中出现 24 个可分型位点；对其之后 HBV 基因组上，总共有 889 个位点具有 ($A \sim H$) 上的至少某一对基因型的区别能力。

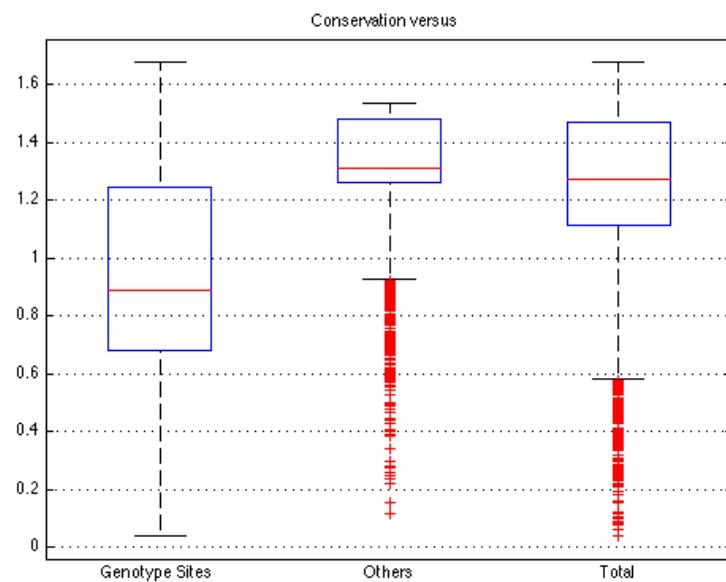


图 2.31 通过这种聚类方法找到的可分型位点，他们在保守性上与其他位点存在这显著差异，通过比较我们可以看出：可分型位点的保守性 $E = 0.9258 \sim 1.3023$, $p-value = 3.1851e^{-244}$

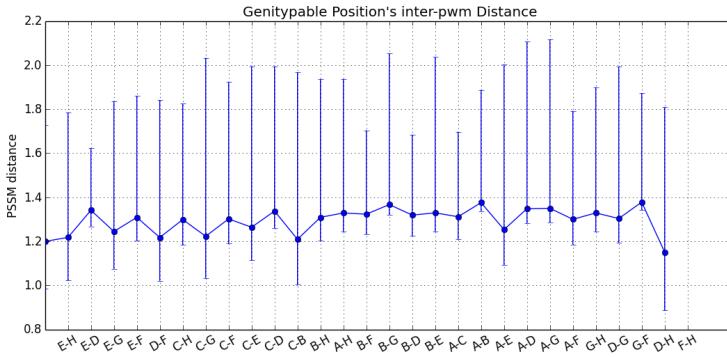


图 2.32 上图, 我分开统计了 28 对 HBV 亚型间欧式距离, 借此区分每一对亚型的可分型碱基概率中心欧式距离的下界; 我们得到可分型的亚型间碱基向量距离 ≥ 0.9 时, 这个位点可以被归属为在此对亚型下可区分的位点

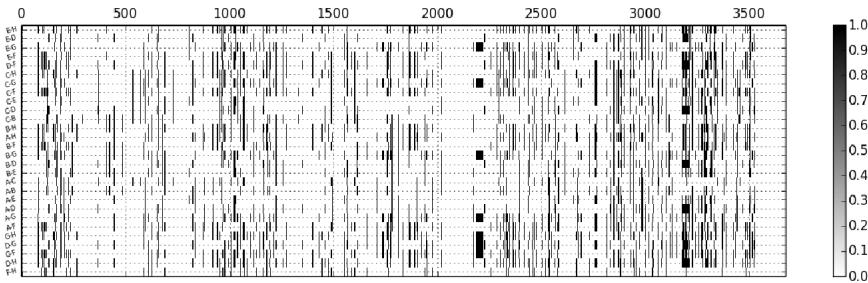


图 2.33 得到的矩阵 A 描述着 HBV 基因组上每一个位点的分型能力。我们需要从 3674 个位点中选择一个特定数目的位点子集合, 这个集合中的位点能够带来所有 HBV 亚型见的差异判别。上图是对 HBV 基因组上具有显著亚型间差异的位点描述矩阵, A 矩阵是一个 0,1 矩阵, 黑色代表亚型间可区分位点, 在矩阵中被为 1; 同理, 白色为 0. 基于这个矩阵也相当于是对测序采样位点进行线性变换, 作为亚型可分判别。

通过 (图 2.35) 可以得到, 当实际上的基因型检测窗口大于 100bp 的时候, 超过存在 80% 的片段可以包含 28 个 HBV 亚型的可分型位点。然后, 我们再使用实际上的滑动窗口来 HBV 基因组序列上的每一个窗口的分型能力进行评估, 这样的话, 我们才可以进一步找到实际上能够从局部推测全局基因型的固定长度窗口。

2.4.3.3 HBV 基因组编码阅读框的重定位

通过 (HBV 基因组示意图 1.1) 的讨论我们可以知道, HBV 基因组上的每一个碱基, 都是被归属到至少一个蛋白对应的开放阅读框内, 所以, 实际上, HBV

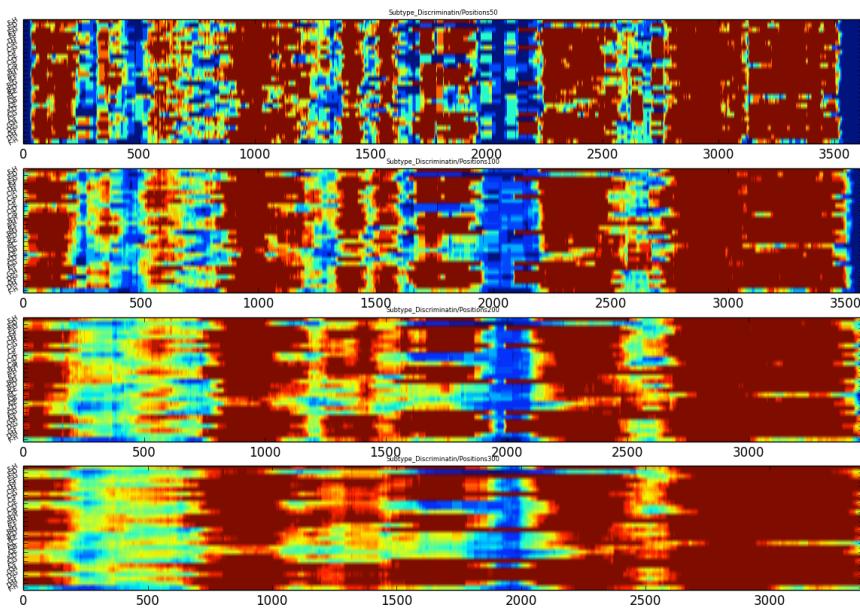


图 2.34 不同片段长度带来不同的分型能力：上图分别设定的基因组上的滑动窗口长度为:50,100,200,300bp 下, 全局的可分性窗口的分布情况, 我们可以通过 $Ax = b$ 的矩阵操作, 判断出每一个窗口是否包含了所有 HBV 亚型的亚型检查信息, 同时, 当滑动窗口长度上升至 100bp 以上是, 除了坐标 2000bp 左右 (对应 core protein 和 e 表面抗原的 ORF 区) 之外, 短片断在全局的分型能力均匀且健壮 (Robust)。

基因组上每一个位点的进化很显然需要受到开放阅读况蛋白质功能的约束。同时, 同一个开放阅读框下的碱基由于功能的限制, 在复制过程中的一致性, 会显示出高于不同开放阅读框碱基的相关联性。不同的开放阅读况对应着不同的功能, 也会导致整个 HBV 基因组的亚型倾向于发生分化。例如, 我们之前讨论的 HBV 基因的阿德福韦酯抗性是由于基因组上的 RT 区对应氨基酸发生了替换突变:

344 个氨基酸序列上的 236 位出现天冬酰胺 \Rightarrow 苏氨酸, [AAT, AAC \Rightarrow ACT, ACC, ACA, ACG], 这里可能的突变是: A \rightarrow C。

另外还有 181 位的丙氨酸 \Rightarrow 缬氨酸突变, [GCT, GCC, GCA, GCG \Rightarrow GTT, GTC, GTA, GTG], 可能的突变: C \rightarrow T。

使用数据库定义 HBV refseq(NC_003977.1) 的开放阅读框坐标信息, 我使用 EMBOSS::waterman 局部序列对齐算法, 重新定义每一个 HBV 蛋白开放阅读框在我的使用标准序列集合多重序列重新定义的坐标系下的坐标信息 [图:2.36]。

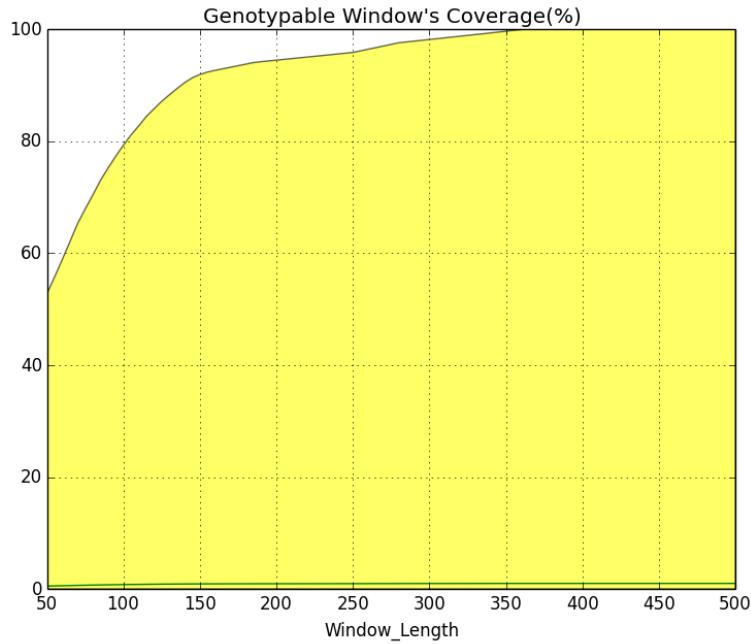


图 2.35 上图为不同长度的滑动窗口下, $Ax = b$ 判别下的可分型窗口在 HBV 基因组全长上面所占据的比例; 当窗口长度 $\geq 100\text{bp}$, 超过 80% 的窗口可以完全覆盖 HBV 基因组上所有亚型间的差异特征。

如此, 基于开放阅读框的定义, 我基于不同的子阅读框不存在相互重叠的原则, 以及实际上表达的开放阅读框归属, 可以把 HBV 基因组序列完全的正交划分为 11 个互不重叠的区域, 每个区域对应的编码阅读框具体信息为 (2.38a 和 2.38b).

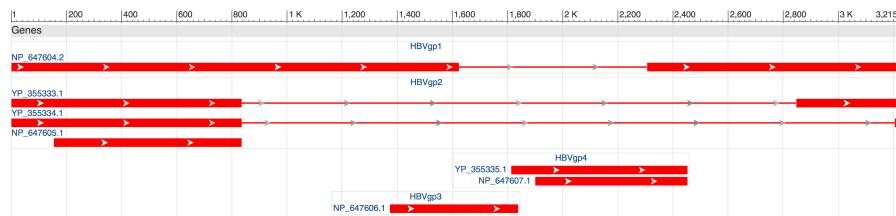


图 2.36 HBV 基因组开放阅读框坐标分布; 数据来源:(NCBI,HBV Genome Structure Diagram)

基于 (2.38a 和 2.38b) 定义的区间, 我们对每一个区间的 HBV 分型能力和开放阅读框内碱基保守性进行分析;

同时, 我们还发现, 不同的非重合开放阅读框内, 都存在着我们之前基于

HBV Gene	Rescaled Coordination
HBVgp1 (Polymerase)	(77, 1733); (2621, 3536)
HBVgp2 (Pre-S)	(77, 936); (3170, 3546)
HBVgp2 (Pre-S)	(77, 936); (3536, 3546)
HBVgp2 (S)	(232, 936)
HBVgp4 (Pre-c)	(2033, 2772)
HBVgp4 (core)	(2178, 2772)
HBVgp3 (X)	(1477, 2125)

图 2.37 基于 HBV refseq(*NC_003977.1*), 和 HBV 标准序列集合标准化之后的各个开放阅读框的实际坐标。

Pre-S+Polymerase	[77, 232]	
Pre-S+Polymerase+S	[232, 936]	
Polymerase	[936, 1477]	[2772, 3170]
Polymerase+X	[1477, 1733]	
X	[1733, 2033]	
X+pre-C	[2033, 2125]	
Pre-C	[2125, 2178]	
core+pre-C	[2178, 2621]	
core+pre-C+Polymerase	[2621, 2772]	
Pre-S1	[3170, 3536]	
Pre-S1+Pre-S2	[3536, 3546]	

Pre-S+Polymerase	[0, 154]	
Pre-S+Polymerase+S	[154, 834]	
Polymerase	[834, 1373]	[2445, 2841]
Polymerase+X	[1373, 1622]	
X	[1622, 1807]	
X+pre-C	[1807, 1841]	
Pre-C	[1841, 1894]	
core+pre-C	[1894, 2300]	
core+pre-C+Polymerase	[2300, 2445]	
Pre-S1	[2445, 3194]	
Pre-S1+Pre-S2	[3194, 3204]	

(a) 11 个独立区域描述及其标准序列集合重定义坐标 (b) 11 个独立区域去除高频 gap 区域后坐标

图 2.38 左图, 因为 HBV 基因组上存在开放阅读框重合的现象, 所以, 我们实际上使用了不重合的方法把重合的开放阅读框区域单独的描述出来, 所以, 我们得到了 11 个互相不重合的唯一定义的开放阅读框的组合阅读框, 考虑到阅读框重叠元素之后, 不同的重组阅读框内的碱基在功能上各不相同。我这里把这种阅读框定义为‘幂阅读框’

位点相关矩阵找到的高度关联分量, 这也暗示我们虽然 HBV 的位点演化在不同的蛋白开放阅读框内仍然呈现倾向性, 但是显著的位点间偶联不止出现在同一个 HBV 基因组编码阅读框内, 也就是说, 我们可以通过一个阅读框的序列具体内容, 耦联分析其他 ORF 的具体情况。

基于 HBV 基因组上开放阅读框的分析可以让我们针对所处理问题的不同情况, 选择不同的坐标范围下的连续片段集合, 这样的话, HBV 的开放阅读框框架提供了一个更加 rational 的分型片段选择评价标准, 用户可以更好的使用对药物动力学特性选择最合适的窗口。例如, 我们如果希望研究阿德福韦酯的抗药性突变, 我们就特异性的去扩增测序阿德福韦酯抗性突变区间。另外, 如果我们希望能够得到关于 HBV 基因组更为准确的分型信息, 我们就应当选择包含可分型位点足够多的区域。所以, 对于基于片段的分型, HBV 基因组上的

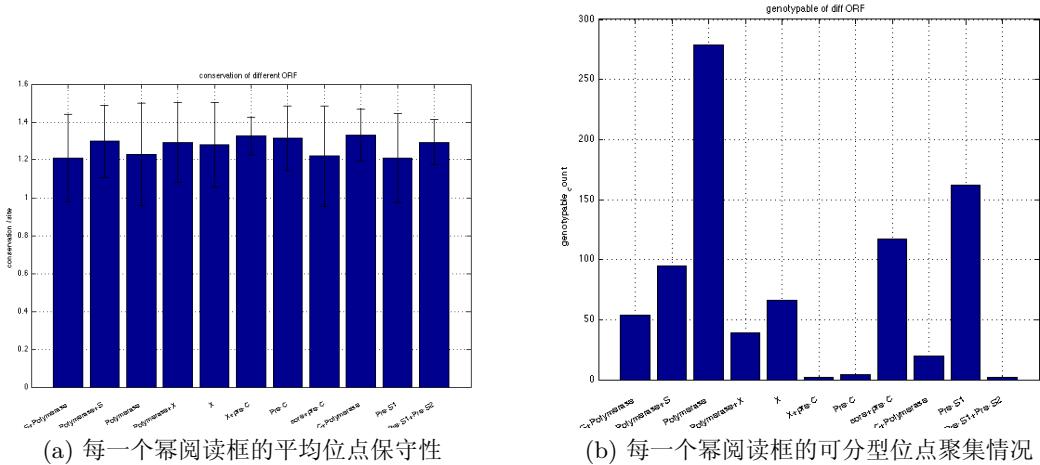


图 2.39 左图, 平均位点保守性在非重叠阅读框间是保守的, 不同非重叠阅读框内的位点保守性没有显著差异; 右图, 不同非重叠阅读框内可分型位点的分布情况, 我发现, 在 RT 区内分型位点相对更多, 而 HBV 的核心区 (core) 以其前前提区可分型位点的数目相对很少, 所以, 从基因组结构的角度考虑, 如果我们希望找到可分型窗口, 应当倾向性的在 HBV 的 RT 区间被选择

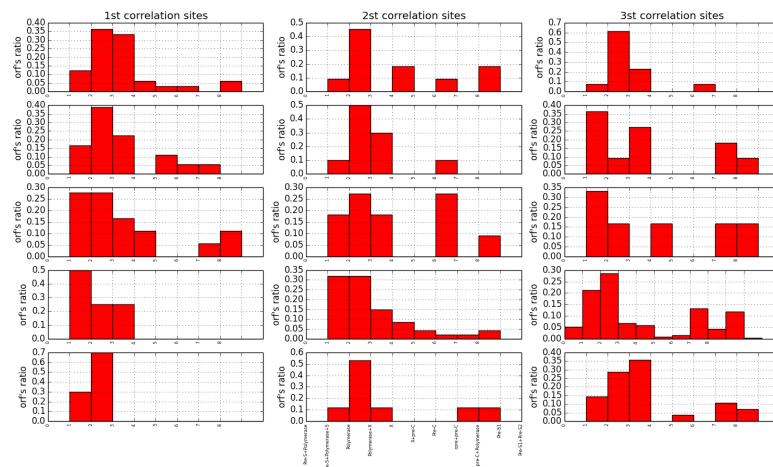


图 2.40 不同开放阅读框内位点在 HBV 基因组全局上的相关情况, 我们通过位点相关矩阵一共找到 15 个显著相关的位点关联联通分量, 1, 每一个开放阅读框内的位点倾向于更高可能落在同一个位点相关联通分量里面; 2, 不同开放阅读框的位点也存在广泛的耦联, 第 11 个联通分量包含了不同的开放阅读框内的位点 (Pre-S+Polymerase+S+X+C+core+pre-C)。

RT 区间是最优的选择。

2.4.3.4 框架总结

HBV 基因组序列大部分是极为保守的，只有少量 24.20%(889/3674) 的序列具有分型能力，这些可分型位点在整个基因组上的分布是均匀的，所以我们可以知道，对于序列采样信号来说，即使我们只保留很少的显著信号位点，也可以准确的恢复实际 HBV 序列集合的位置相关矩阵以及序列相关矩阵，所以，我们可以认为，由于 HBV 序列框架的保守性，以及位点之间的高度关联耦合，我们可以使用短片断来完全推断出全局序列实际上的基因型信息。而且，我们发现对于 HBV 标准序列集合构建的序列相关集合来说，不同的 HBV 亚型刚好占据了不同的信道，所以，我们使用短片断分型的理论基础可以被描述为：由于位点相关信源对应的线性变换矩阵的稀疏，导致了恢复初始的 HBV 位点相关矩阵是可以基于短片断来完成的，另外，由于序列相关矩阵和位点相关矩阵都是来自于 $X3d$ 矩阵的投影阵的奇异值分解，所以，我们使用基于 PSSM 的类似 K means 聚类方法可以基于短片断的恢复 HBV 基因型。

为了验证我们的方法，我们使用 $10\times$ 交叉验证的方法来确定不同子片段的分型准确性以及其泛化能力(图：2.41)，我们发现当选择的窗口长度 $\geq 100\text{bp}$ 时，我们得到的片段分型准确性 $\geq 90\%$ (图:2.42)，所以，我们认为这样的基于 PSSM 矩阵作为不同 HBV 亚型中心的聚类方法时在短序列分型上时可靠的。基于 $10\times$ 交叉验证，我们还可以确定整个 HBV 基因组上各不同长度片段对应的序列的实际分型能力。

所以，我们可以使用这套短序列分型理论框架，来实现实际上的病人 HBV 基因型分型工作。同时，为了验证我们的方法，我们使用了一套完整的慢性乙肝患者分析案例的应用，基于这个案例来证明我们整个框架和方法的可靠性。

2.5 基于短序列窗口 HBV 分型平台构建及其应用

通过 (REGA 分型：2.4)，实际上我们描述出了各个 HBV 基因组亚型的模式信息。对应不同 HBV 亚型，PSSM 矩阵描述了不同位点的 ($A, T, C, G, -$) 碱基

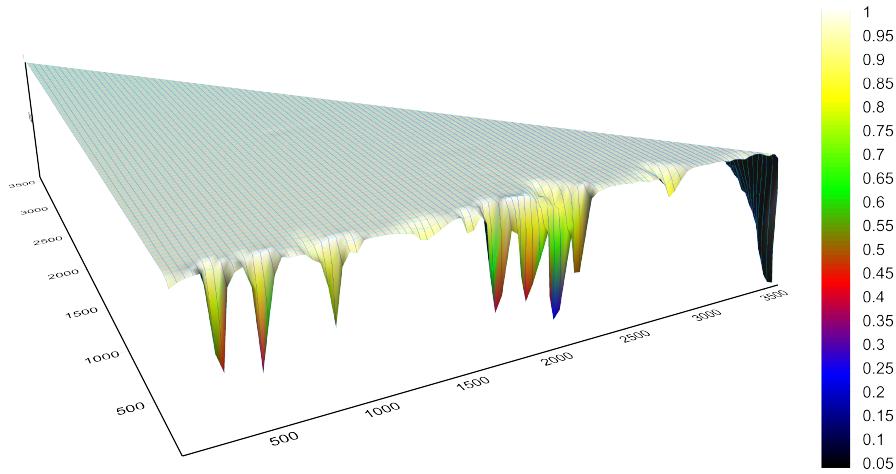


图 2.41 HBV 基因组范围的不同窗口的整体打分; 前方数轴为窗口起始位点, 左侧数轴为选定的窗口的长度, 纵轴为该窗口打分能力评分, 表示了不同窗口长度和位置下, 此窗口的对 HBV 基因组全局的分型准确性和泛化能录打分。

概率分布。因为2.1里我们已经讨论了可以选择 HBV 基因组上的短片断, 进行 HBV 亚型的分类学习。所以, 这里构建平台流程示意图:

不同于普通的 HBV 二代测序, 随机的打断全基因组序列, 进行全局的测序。新构建的 HBV 亚型分类平台目标是找到特定的高可靠性的分型窗口来完成具有准种代表性的特征片段的测序工作。这样做的优点: 检测序列窗口长度限制在了第二代测序技术中可以被直接测完的范围之内, 既保证了分型的准确性, 同时提高了分型效率。有利于个性化医疗过程中测序方法的推广。

HBV 感染的研究主要包括 5 个部分 (图2.43) :

1. HBV 病例样本的采集, 详见2.5.1;
2. Solexa 测序结果还原原始序列比例的探究;
3. HBV 分型窗口的评价体系的建立;
4. Solexa 对 HBV 分型窗口扩增的设计;
5. HBV 测序结果的准种重构。

2.5.1 病人实际样本的获得

采集了 200 例 HBV 慢性感染病人血清, 其中 160 例来自重庆医科大学附属第二医院感染科 2011 年 3 月至 2011 年 6 月门诊病人, 40 例来自重庆医科大

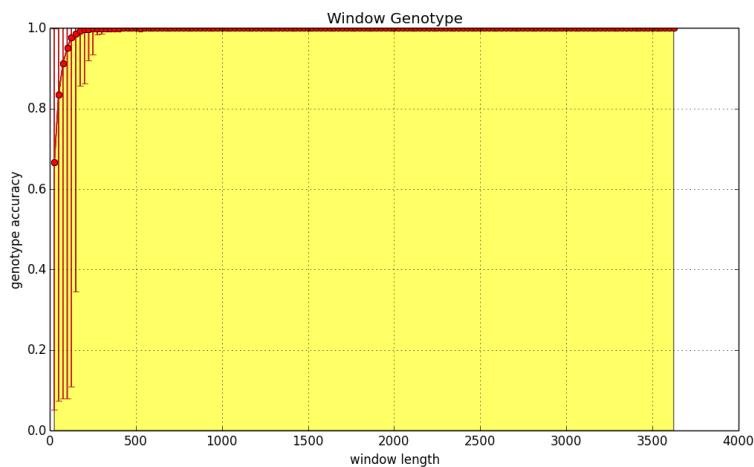


图 2.42 不同长度窗口的平均分型能力(分型准确性); 右图, 不同的片段长度下的分型准确性和泛化能力打分并且表明其相对的分型准确性的标准差, 可以发现, 当序列长度大于 75bp 时, $\geq 95\%$ 的窗口具有准确的 HBV 基因分型能力。

学附属第一医院感染科 2004 年 -2005 年 ADV 临床三期试验受试者。所有标本进行相关临床检验: 转氨酶等肝脏生物化学指标由全自动生化仪检测 (Beckman Coulter, Fullerton, CA) , HBV 血清学标志物包括 HbsAg、HbsAb、HBeAg 、HbeAb、HbcAb 由酶联免疫吸附试验检测 (Abbott, Chicago, IL)。HBV-DNA 检测用荧光定量 PCR(Roche Diagnostics, CA), 检测下限 300copies/mL。所有实验遵守赫尔辛基宣言¹⁶, 签订病人书面知情同意书, 并取得重庆医科大学伦理委员会的同意。

2.5.2 Solexa 测序结果还原原始序列比例的探究

由于混合感染^{*}的存在, 为了探究 Solexa 测序得到的序列的比例是否能够正确反映待测原始序列的比例, 从而还原出病人混合感染的情况, 故人工设计了不同的序列 (图2.44) 并按照不同的比例混合进行 Solexa 测序, 将测序结果与初始混合比例之间进行比较分析。

设计的序列模板来源于 HBV 的 B 基因型的部分序列, 但是在该序列的基础上进行定点突变其中的 10 个位点, 保证不同的人工设计的序列可以区分, 并且满足序列之间的两两差异大于等于 $5/140 = 3.6\%$ 。由于实验的研究对象的

^{*}混合感染, 指同一个病人感染了不同基因型的 HBV

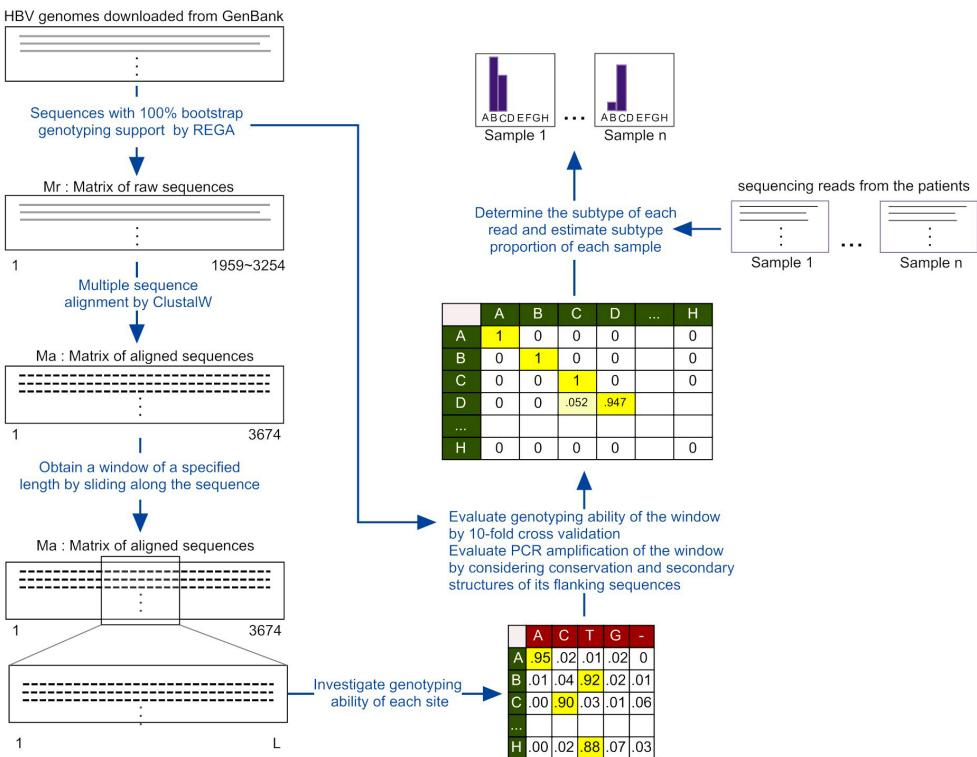


图 2.43 确定亚型窗口的流程示意图



图 2.44 为探究二代测序方法能否还原模板比例而人工设计的 8 条模板

HBV 不同基因型之间的差异约为 7%，远大于 3.6%，所以，如果结果中能够保证正确地通过 Solexa 测序还原原始序列的混合比例，则更大的差异也不会被实验中的系统误差所干扰。

2.5.3 窗口分型能力评价函数的确定

基于 HBV 标准序列集合的亚型中心特征提取 (2.4.3.1)，我们得到了 HBV 标准序列集合中每一个亚型对应的碱基分布概率密度中心，在这个位置特异矩阵中，每一个位点的碱基频率分布会随着亚型的不同而不同，为了确定该窗口

是否存在分型能力, 使用 $10\times$ 交叉验证 (10-fold cross-validation) 来估计不同窗口的实际分型能力。

确定窗口分型能力测试基于 9 份训练集建立起来的, 由于 HBV 基因组上的位点本身保守性高, 在计算每一个目的序列和实际亚型分类中心距离时, 我们使用似然函数来定义每一条目标片段和标准序列集合亚型中心的度量, 同时, 又由于对于一个序列片段来说, 位点之间的耦联相较于实际上的序列差异在较小 (因为实际上的不独立的可分型位点在全基因组上的比例很小), 所以在计算每一条目标序列和亚型 PSSM 中心的似然程度时, 我们可以使用各个碱基对应的位置特异矩阵的 $P(B|G, i)$ 的独立乘积, 来定义实际上的序列 - 中心相似程度, $P(B|G, i)$ 的含义是位点 i 当序列是 G 分型时的碱基 B 出现的频率。所以对于整个窗口 (长度 = k) 的评估可以定义度量为:

$$P(W = B_1B_2...B_k | G) = \prod_{i=1}^k P(B_i | G, i) \quad (2.16)$$

虽然这是窗口分型能力的先验分布, 为了得到 $P(G|W = B_1B_2...B_k)$, 通过贝叶斯 (Bayes) 变换中可以得到:

$$P(G = g | W = B_1B_2...B_k) = \frac{P(W = B_1B_2...B_k | G = g)p(G = g)}{P(W = B_1B_2...B_k)} \quad (2.17)$$

根据全概率公式, 可以计算出对应的窗口长度为 K 且序列已确定的情况下, 该序列为任意基因型的概率。取概率最大的基因型, 即对一个待分型序列的分型结果:

$$P(G = g_s | W = B_1B_2...B_k) = \frac{P(W = B_1B_2...B_k | G = g_s)p(G = g_s)}{\sum_{s=A\sim H} P(W = B_1B_2...B_k | G = g_s)p(G = g_s)} \quad (2.18)$$

假设序列进化到不同基因型是等概率的, 由此窗口的分型能力就可以完全由不同基因型假设下的窗口出现几率所决定。基于 PSSM 矩阵计算出 $10\times$ 交叉验证中的测试集的分型效果。对分型效果主要考虑分型的准确性 (accuracy), 即

实际测试序列被分型正确比例：利用 9 份标准序列建立起来的测试去测试剩余的 1 份序列并进行分型，并根据已知的 REGA 分型结果来判定分型的准确性。

接下来对 HBV 整个基因组进行了窗口长度步长为 50nt 的多起点、多长度打分。从打分结果的热图（见图2.41）来看，在窗口长度取得比较长的时候，分型能力基本稳定在 1.00 左右，而当窗口长度较短的时候，则会随着窗口起点选择的不同，得到不同的分型能力的窗口。对于同一起点的不同长度的窗口，窗口的分型能力是随着窗口的长度而增大的。

统计窗口分型能力的打分情况，发现仅当窗口长度 $> 75\text{nt}$ 时，窗口的分型能力 > 0.95 ，而当窗口长度 $> 200\text{nt}$ 时，窗口的分型能力基本稳定在 0.99 以上。其统计结果图见图2.42。

为了更加精细地研究窗口的分型能力，并根据 Solexa 和 454 测序长度的特点，找到适合于 Solexa 测序和 454 测序的窗口，所以又进行窗口起点与长度步长均为 1nt 的、 $75\text{nt} < \text{窗口长度} < 250\text{nt}$ 的部分进行细致打分。两次打分的结果热图见（图2.45）。后续的实验会根据窗口长度 =100nt 的不同起点的窗口分型能力的打分情况，找到适合 Solexa 测序并进行 HBV 分型的窗口。

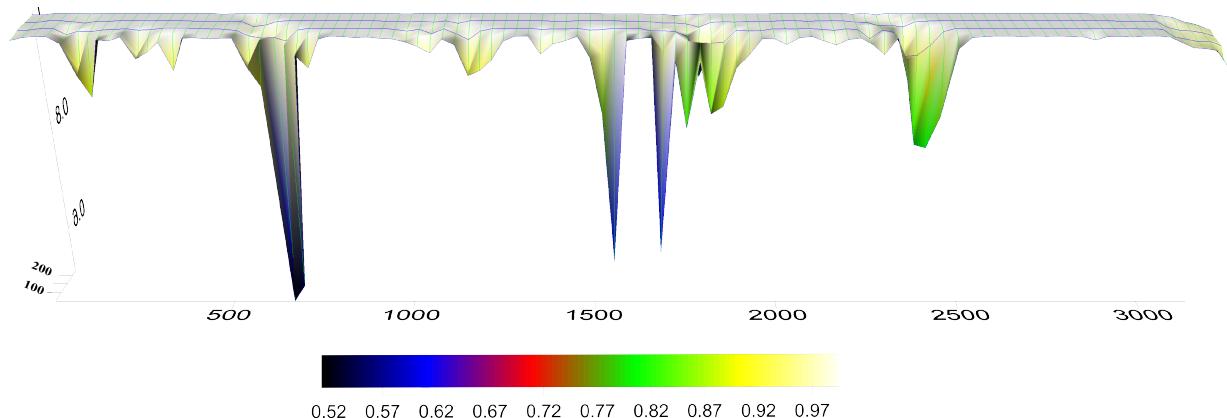


图 2.45 HBV 基因组窗口的打分情况 (75nt, 250nt): 前方数轴为窗口起始位点, 左侧数轴为选定的窗口的长度, 纵轴为该窗口打分能力评分

2.5.4 利用 RT 区进行确定窗口分型能力的参数

根据文献报道，很多研究者曾经通过对 RT 区的分析进行分型，且分型结果较为准确¹⁷⁻²¹。

由于 HBV 的 RT 区是一个功能上很重要的区段^{17,20}, 所以在 HBV 的基因组上映射到 RT 区窗口的实际位置 [716-981]nt, 对该窗口进行窗口分型能力的评价, 以确定 RT 区实际的分型能力, 结果 RT 区窗口的分型准确率为 0.999。

通过对 RT 区分型的评价, 另外考虑到对分型准确度的高要求, 确定后续寻找 Solexa 测序窗口的参数至少准确率 >0.99。

2.5.5 对 454RT 区测序结果进行荧光强度的校正

由于 454 测序所获得的原始荧光强度数据的特点 (图2.46):

1. 由于 454 荧光读数并不总是理想的整数, 如何从原始的荧光流数据得到序列信息需要合适地设定阈值;
2. 当待测序列有大于 5 个同一种核苷酸连续多聚 (Polyhomer) 的情况时, 难以通过荧光强度的读数准确地估计核苷酸的数目;
3. 实际的焦磷酸测序引入了背景噪音的同时, 不同 Polyhomer(不同的连续核苷酸数目 [1, 2, ..., 8]) 在测序过程中的荧光强度分布满足 Log normal Distribution[图2.46].

所以, 实验中试图建立了一套对荧光数据进行校正的方法来克服这些问题。

2.5.6 消除 polyhomer 读数的随机波动带来的实际判断误差

在 454 测序中, 例如对序列 “CTAAAG” 进行测序, 按照 T,A,C,G 的 NTP 流流入进行焦磷酸荧光测序, 会获得 (0.1,0.02,0.8,0.2)(1.3,3.6,0.1,1.4) 类似的流序列数据。这样的话, 在转换为碱基序列用简单的四舍五入就会形成 “CTAAAAG”, polyhomer 位置出现了插入现象, 这样并不能正确的体现实际序列的序列信息。

基于焦磷酸测序中的问题, 实验中提出一种新的流数据处理方法, 就是基于机器学习中的 EM/Gibbs 采样算法对序列的荧光强度读数进行分类, 把实际的荧光强度矫正回对应的实际的碱基数目。

Balzer 等人的研究认为, 焦磷酸测序中的实际荧光强度是一个在对应碱基数目为均值的正态分布或者对数正态分布。²² 而在 Quince 关于焦磷酸测序矫正

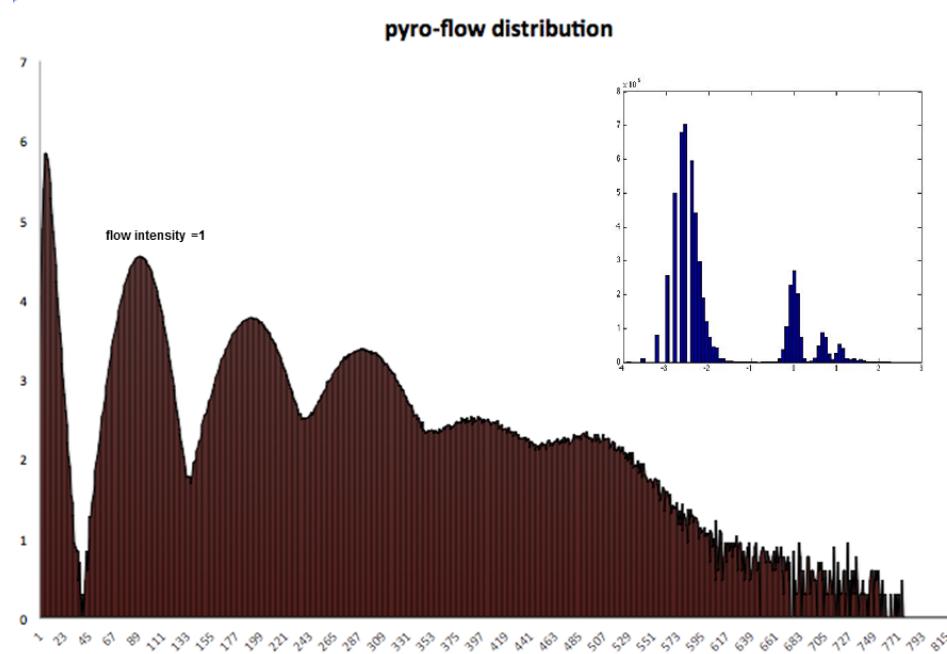


图 2.46 焦磷酸测序所得荧光强度的分布; 在测序过程中不同的 Polyhomer 的荧光强度分布满足 Log normal Distribution

的文章里,²³ 设定该分布是一个正态分布，并且满足条件：

$$\text{均值} = \text{实际的碱基数目} \quad (2.19)$$

$$\text{标准差} = 0.03 + 0.04 \times \text{实际的碱基数目} \quad (2.20)$$

并且以此来进行他们后续的研究，当时这里本身就是存在一个互悖的证明，即正态分布应当满足均值与标准差相互独立，而在他们的假设中，均值与标准差并不独立。

从实际的 123 个样本的所有的荧光强度的读数分布来看，不同样本的荧光强度分布是一致的，但是到当荧光强度大于等于 5 之后的荧光强度可以看出，数据量严重不足，导致在函数拟合的时候带来偏差。

基于对 Quince 文章²³ 的修改，首先设定满足这是一个正态分布，并且均值和方差是独立的，又因为实际的荧光强度是体现实际的碱基读数。所以，在这里按照 EM 算法假定这是一个有 N 个正态分布线性组合形成的高斯混合模型。N 代表实际序列中的 polyhomer 类型的数目。

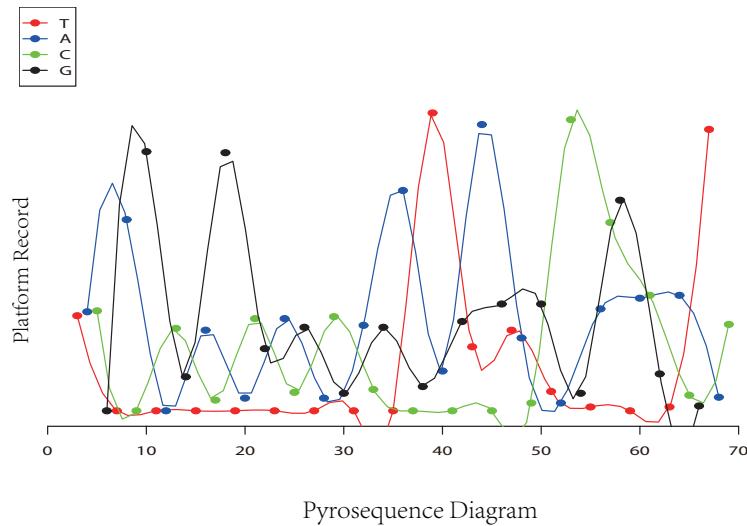


图 2.47 焦磷酸测序所得原始数据示意图

对每一个样本进行 EM 算法聚类之后，确定了实际荧光读数与实际碱基数目的映射关系。在基于这个映射关系，获得矫正后的序列。

2.5.7 原始荧光数据的校正

测序原始数据为序列集合中每一个样本对应的焦磷酸聚合反应的荧光强度结果，并且在对比不同样本数据发现，在 Polyhomer 数 ≤ 5 时，碱基的荧光强度的分布是一致的。

所以基于已有的观察，可以对实际的荧光强度读数进行一个矫正，给每一个荧光强度的读数一个对应的碱基数目。

矫正方法：当已知荧光强度的分布是一个 N 种 polyhomer 组合形成的混合分布，当样本量足够时，可以把它当作是 $N(NE[1, \text{maximum}(\text{polyhomer})])$ 个高斯分布线性组合形成的混合模型。

这里假定存在 N 种 polyhomer 即 N 个高斯模型，而每一个荧光强度实际上都会被标记上一个标签，表明该荧光强度属于第几个高斯模型，从而基于归属的标签来确定实际上荧光强度与实际碱基的数目之间的关系。

基于这个工作思路，每一个荧光强度的标签实际上是实际荧光强度对应的一个隐含变量，把这个隐含变量表示为一个 N 维向量，每一维对应了这个荧光

强度属于对应高斯模型的几率。基于引入的隐变量，引入 EM 算法的思路：

2.5.7.1 基于荧光强度的计数范围 ($\max(\text{raw intensity})$)

确定高斯混合模型中的模型数目为：

$$N = \max_{\text{flowintensity}} + 1 \quad (2.21)$$

2.5.7.2 确定初值

已知 EM 算法中初值的确定是很重要的一个步骤，但是在该实验中，由于没有更多的先验知识，所以，初值的设定遵循：Model_N 的 (E = 碱基的数目, $sd = 0.5$)。

2.5.7.3 开始 EM 算法

E-step，对应着求关于后验概率的期望亦即 $P(M_k | I)$ ；

$$P(M_k | I) = \frac{P(I | M_k)P(M_k)}{\sum_{i=1}^{\max} P(I | M_i)P(M_i)} \quad (2.22)$$

其中， I 表示实际的荧光强度， M_k 表示归属于哪一个模型。

基于上面的贝叶斯公式，可以求出每一个荧光强度对应的 N 维分布变量，基于这个变量，可以计算出每一个荧光强度在属于不同模型的概率

M-step，则对应于接下来的正常的大似然的方法估计相关的参数亦即，每一个高斯模型对应的均值和方差。

2.5.8 HBV 分型 Solexa 窗口扩增的设计

2.5.8.1 确定可用于 Solexa 测序的可分型窗口

选择分型窗口除了需要满足准确性和敏感性的要求外，还需要需要满足：

1. 窗口上下游 20nt 是保守的：上下游错配碱基数目不超过 3bp；

2. 可以设计用四种不同长度 barcode 标记到引物两端, 以保证同一位点上的碱基尽可能的错开。

确定要寻找的窗口长为 100nt 后, 实验对不同起点的窗口进行分型能力的细致打分, 其结果热图表示如图2.48。

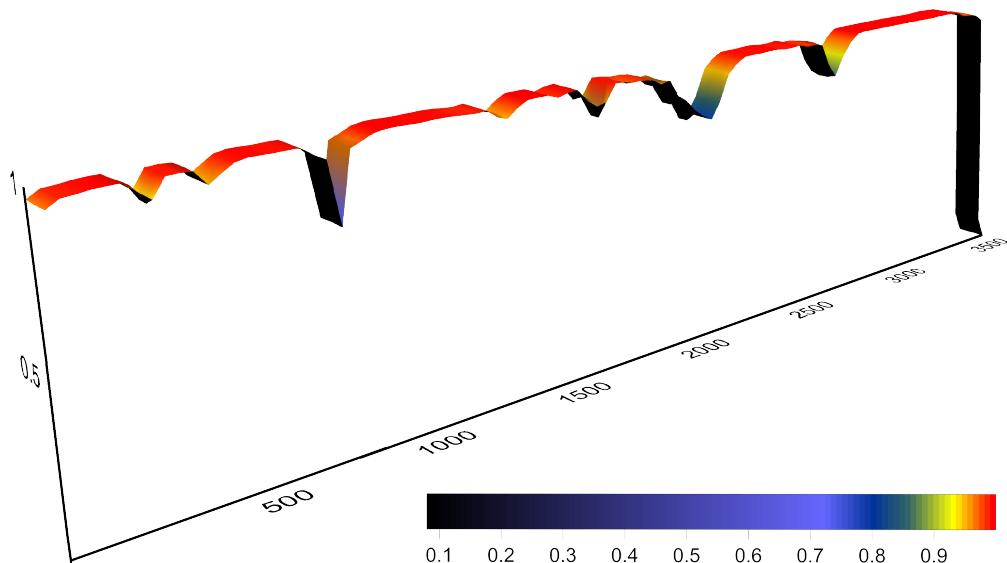


图 2.48 Solexa 窗口 (100nt) 分型能力评价示意: x 轴, 窗口起始位点; y 轴, 窗口分型能力

分别以分型准确率为 95% 和 99.9% 扫描以上评价结果, 分别找到符合要求的窗口 3015 和 518 个。由于测序区段还要考虑两端引物序列的保守性, 结合分型准确性和引物保守性选择合适测序区段。最终选择其中一个 Solexa 的分型窗口 [1550-1649]nt^{*}继续下一步的实验。[1550-1649]nt 区间的分型准确率 99.9%, 两端引物保守性 100%。

2.5.8.2 基于窗口的 barcode 设计

确定出有分型能力的 100nt 窗口之后, 为了增强 Solexa 测序的可靠性, 需要通过不同长度的 barcode[†]把扩增窗口错开来, 以保证 Solexa 测序平台检测序列同一位置时荧光强度可以分散开来, 选定了进行扩增和 Solexa 测序的四种不

^{*}将该区段映射到多重序列比对后的标准序列集, 该窗口包含 gap, 对应 HBV 基因组序列的 [1446-1544]nt

[†]PCR 扩增时在 primer 两端加上的有标签作用的序列

同长度的 barcode 标记，均匀的分配到不同样本同一测序窗口的两侧，以保证测序时检测位置在同一位置时荧光强度能均匀的分散到不同的碱基上。Barcode 长度组合最终选定为 (12:11:6:7) bp。

Barcode 在第一次 PCR 引入之后，执行的角色与 PCR 的引物一致。所以认为 Barcodes+Primer 必须满足 PCR 过程中引物的要求：

1. 为了防止发卡结构等高级结构以及引物二聚的出现，在做局部序列比对和全局序列比对时，均不能含有超过 1 个以上的三连氢键数目；
2. 为了确保一个较为合适的退火温度，GC 的含量尽量控制在 50% ~ 60% 之间；
3. barcode 不能与窗口序列有重复；
4. barcode 里不能含有 CCC 或 GGG 这样的碱基组合
5. 不同长度 barcode 保证检测序列错开，使得测序过程中每一个位置 ATCG 碱基分布倾向于均匀；

最终，实验中共设计 288 个 barcode 的组合，实际 barcode 数目控制在 162 个，下游 *barcode + Primer* = 115 个，上游 *barcode + Primer* = 47 个。

实验中上下游的 Barcode 设计及组合见附件。

2.5.9 HBV 测序结果的分型

基于“标准序列集”的位置特异矩阵，对每一个窗口下的测序序列进行分析，确定于测序窗口序列的模式最接近的基因型。

由于对于基因型 (Genotype) 的定义是相似性至少大于 70% 的序列^{12,24}，所以确定序列相似性 0.70 作为可分型的下界。筛选后的序列可以确定是 HBV 且获得正确的分型结果。并且从公式 2.23 得到序列被确定分型后的后验概率。

对 454 测序结果和 Solexa 测序结果均作相同的处理：

1. 测序结果与“标准序列集” PSSM 生成的共有序列进行多重序列对齐。
2. 对齐好的序列在该测序窗口下计算。

$$P(G = g_s | W = B_1B_2...B_k) = \frac{P(W = B_1B_2...B_k | G = g)P(G = g_s)}{P(W = B_1B_2...B_k)} \quad (2.23)$$

计算每一条序列的 $\underbrace{\operatorname{argmax}_{g_s} P(G = g_s \mid W = B_1B_2...B_k)}$, 若该值大于设定的阈值即 0.70, 则认为分型有效, 并确定为取概率最大的该种基因型。由此得到每条测序序列实际分型结果, 统计每个样本中的 HBV 病毒基因型的混合比例。

2.6 HBV 短序列分型结果分析

短片断数据的获取来源于病人 2.5.1, 我们基于 2.4.1 和每一个特定长度的序列片段上下游 20bp 序列的相对保守性, 同时基于不同蛋白表达开放阅读框中可分型位点的分布情况 (图 2.39b), 我们倾向于在 HBV 基因组上的 RT 区域寻找合适的测序片段。

针对不同的测序技术长度, 我们找到了选择了两个不同的长度下的分型窗口 (100bp 和 300bp) 的分型窗口, 这两个分型窗口的选择时完全基于 (2.5.8.1) 定义的约束条件。我们确定下 [1550-1649] 和 [619-879] 这两个区域的片段用于 HBV 基因组的亚型分类; 这两个片段分别对应不同的蛋白编码开放阅读框, [619-879]nt 对应的是 HBV 的聚合酶逆转录区域; 而 [1550-1649] 则对应的是 HBV 上的 X 蛋白开放阅读框。我们选择这两个区域是因为一方面他们的分型能力在我们的序列扫描过程中都具有很高的分型准确率和泛化能力, 同时, 由于不同的位点区域之间位点耦合的程度相对同一个开放阅读框内较小, 多重片段的采样也可以证明同一个开放阅读框内位点不存在进化耦合过高的区域分型偏差。

表 2.2 [1550-1649] 和 [619-879] 测序窗口信息, 对应的位点保守性和序列

二代测序方法	窗口大小	窗口范围	分型准确性	引物保守度
Solexa	100bp	[1550-1649]bp	99.9%	100.0%
454	261bp	[619-879]bp	100.0%	70.0%

2.6.1 二代测序还原 HBV 基因型亚型的标准曲线

为了确保实验过程中对病人血清基因组亚型分布还原的准确性, 我们还在 (2.44) 中, 设计了人工合成的底参序列集合, 以证明我们的测序方法可以准确

的还原实际片段的正式内容以及完全反映不同准种之间的序列丰度差异。确定序列测序丰度对应的亚型比例与实际上的血清中基因型的亚型比例之间存在的转化关系。所以，我们需要基于测序方法本身建立一个测序片段数比例与实际原始 HBV 基因亚型分布之间的转化关系：

我们基于 HBV 的 B 型基因片段设计了评价差异 $\sim 7\%$ 的 8 条不同人工合成序列 (2.44). 并且，为了使用二代测序方法对不同的序列间差异为 $5\% \sim 10\%$ 的序列的区分能力和丰度还原能力，我们按照 8 类人工合成序列的混合香农熵不同，即不同的病毒序列复杂程度，构建梯度混合序列模版集合：构建的混合样本保证了不同的合成的差异模版的混合比例导致的熵值形成了 $Entropy \in [0, 1]$ 的均匀采样，并且基于不同的差异序列类型数目 (图:2.49,2.50)。8 类人工合成序列组合成的 35 个不同的混合比例的样本集合，作为验证测序片段数与实际序列丰度之间映射的基准。

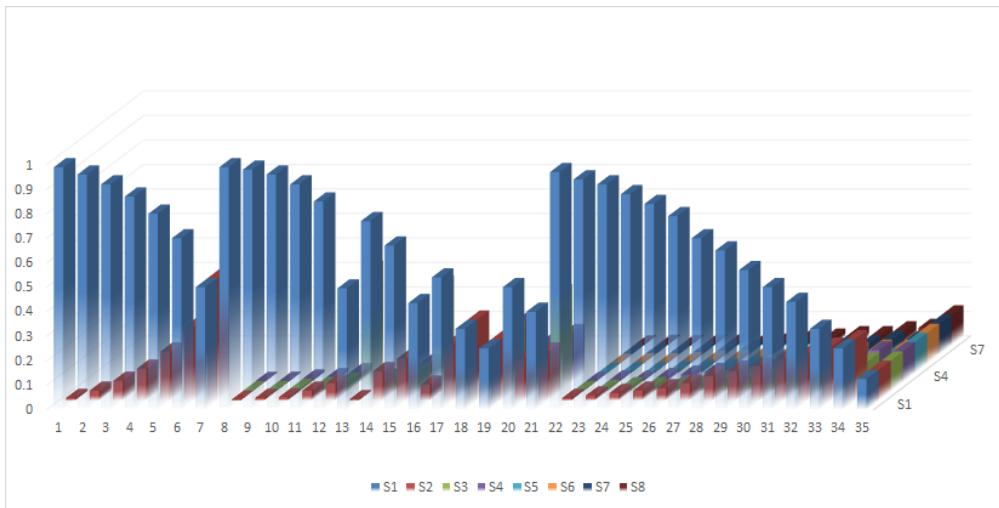


图 2.49 8 个人工合成模版在 35 个混合样本中的实际分布情况

我们使用 solexa 测序完成了这 35 个样本的测序，并且通过 pairwise alignment 的方法完全还原了 35 个不同合成序列混合样本的在测序中的不同的模版读数，基于不同的模版读数我们可以对比模版读数分布和实际设计模版混合比例之间的关系。

标准试验的结果 (2.51a,2.51b):

1. 我们基于模版比例可以监测并还原的最小混合比例是 0.02%，这个结论与

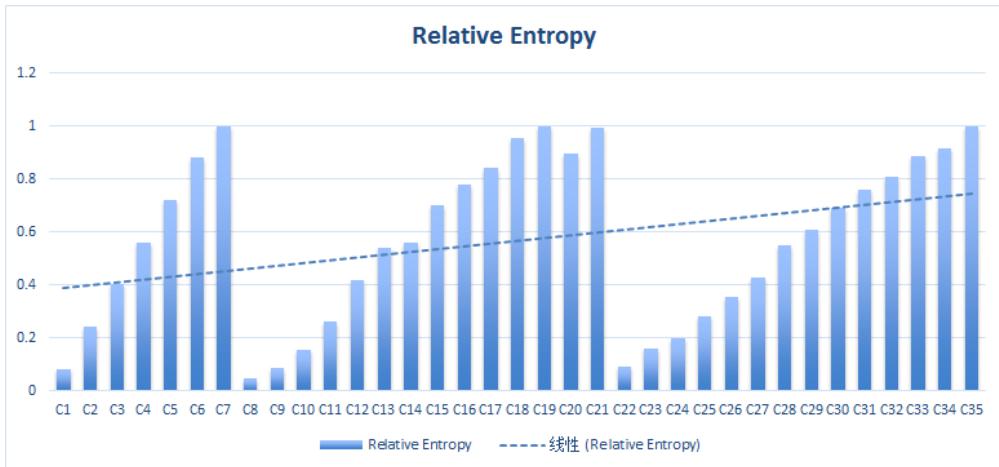


图 2.50 8 个人工合成模版在 35 个混合样本的实际混合熵值是在 [0,maxEntropy] 之间的均匀采样

我们在模版测序中使用小鼠血清作为阴性对照的结果是一致的(小鼠血清中可检测的序列读数是 0.0002142857)。这保证了我们使用二代测序的高分辨率和低污染率。

2. 通过测序片段数目还原的序列丰度 ($\frac{count_{eachtemplate}}{count_{total}}$) 与实际的模版混合比例分布呈现显著的线性以来关系，并且斜率系数为 $observed\ freq = 0.9523 \times real\ freq$.

2.6.2 二代测序结果和患者混合感染情况分析

实验中利用了多种测序方法对病人的样本的不同基因区间进行 HBV 分型，并按分型流程对所测序列基因分型并计算其各自比例。结果 200 例 Solexa 检测标本中基本都是 B 型和 C 型混合，根据劣势基因型所占比例统计： $>0.02\%$ 的混合感染有 191/200(占 95.5%)， $>0.05\%$ 的有 160/200(占 80%)， $>0.1\%$ 的 97/200(48.5%)， $>1\%$ 的 24/200(12%)。根据空白对照设定的混合感染阈值 =0.02%，是混合感染和单基因型感染的分界点，结果显示混合感染是普遍存在的，其中低比例的混合感染占大多数。

为了保证实验结果的准确性，实验中的每一个细节都非常注意防止污染。同时，实验中用小鼠血清做空白对照。结果空白对照只有 20 条 HBV 序列，而

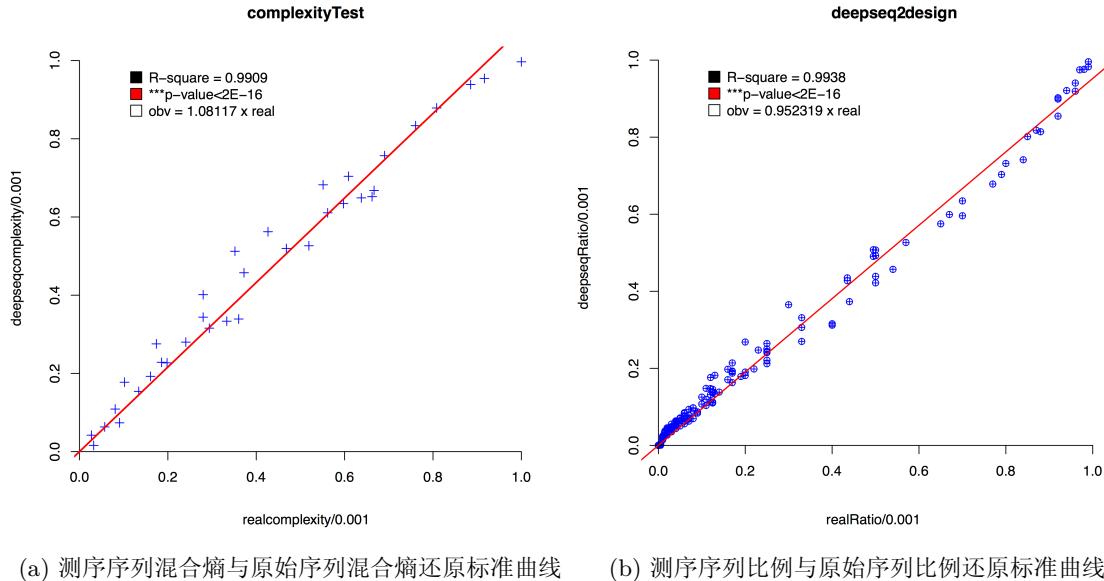


图 2.51 Solexa 测序序列比例与原始序列比例以及混合熵呈较显著的等价线性关系：斜率系数分别为：1.082 和 0.9523

所有标本的平均读数约为 10 万条，说明实验中防污染措施是成功的，Solexa 的结果是可靠的。

我们使用不同的方法 solexa 和 454 针对我们找到的不同的序列区间片段，并且通过不同样本测序的序列数目我们可以确定每一个病人基因型分布推断的健壮性

我们发现在监测的 38 个病人在 0 周的混合感染情况，使用选择出的不同对应开放阅读框的序列片段对应的序列片段都可以还原出完全一致的 HBV 基因型混合比例，这个例子也佐证了我们建立的短序列分型框架是健壮稳定的 2.52。

2.6.2.1 混合感染慢性乙肝患者在阿德福韦治疗过程中的基因型反转现象

使用深度测序方法，我们从 200 个病人患者中的血清样本中，分别在使用阿德福韦治疗之前和治疗之后分别检测慢性乙肝患者的血清病毒基因型亚型的分布，由于在 2.6.1 的实验中，我们确认可检测的序列混合比例的下确界是 0.02%，所以，如果我们发现混合比例中 B,C 两个亚型的比例均在 0.02% 之上，我们就可以认为实际上病患在用药之前本身就是混合感染。

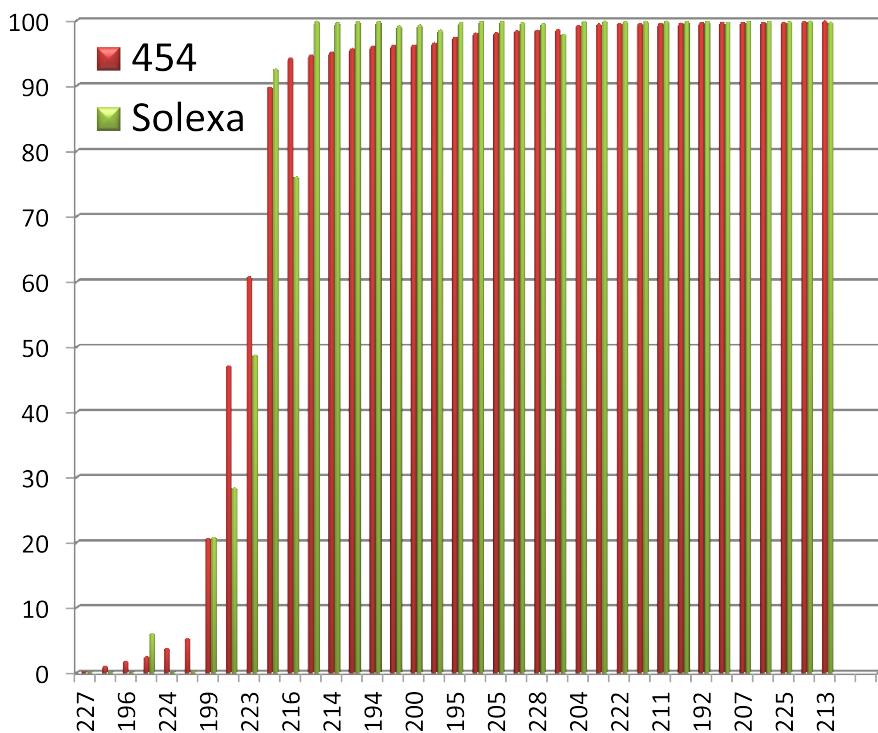


图 2.52 比较 38 个病人样本使用不同的基因组区间下还原病毒亚型分布的一致性：相关系数为：0.994。由于实验中用到的病人样本全部为 HBV 的 B、C 基因型的混合感染，检测到的其他类型的序列极少，故在此仅讨论 B、C 基因型，而两种基因型序列比例相加越 100%，故作图时仅做 B 型感染比例来表示病人混合感染情况

所以，基于对混合感染的定义，我们确定了 200 个病人患者中存在 38 个存在不同 HBV 亚型混合感染的患者，同时，在这 38 个病患中找到 13 个患者在用药过程中 [0~48] 周发生显著 HBV 基因型亚型 ($B \leftrightarrow C$) 转换现象??。

leftrightarrow 所以，基于 HBV 基因型反转现象，我们发现在阿德福韦的治疗下出现了显著的 $B \leftrightarrow C$ 亚型转化现象，这是我们继 Jardi 等人发现 HBV 亚型 A 对于核算类似物的 HBV 抗病毒药物具有低敏感性之后又发现的新的阿德福韦治疗下出现的基因型倾向性。从我们的试验结果分析来看，HBV 基因型在阿德福韦抗病毒药物的治疗过程中，出现了阿德福韦酯对 HBV 基因组 C 亚型的偏好性 (2.55)，使用 ADV 治疗之后，患者的 B 亚型丰度比例出现显著上升，而 C 亚型下降。所以，阿德福韦对于 C 亚型更加敏感。

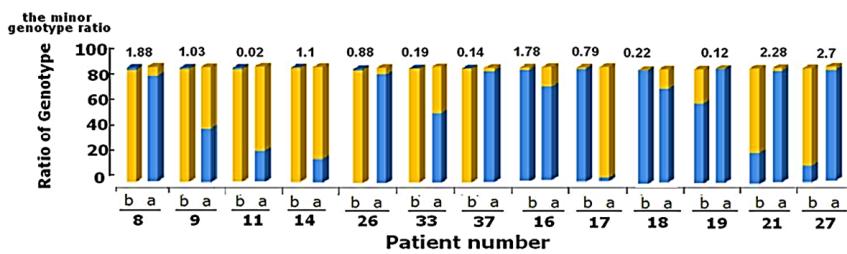


图 2.53 使用基于片段的二代测序方法确定的发生混合感染基因型反转的患者; 图中,a 表示治疗之前, b 表示治疗之后; 黄色是对应 HBV C 亚型, 蓝色对应 HBV B 亚型。

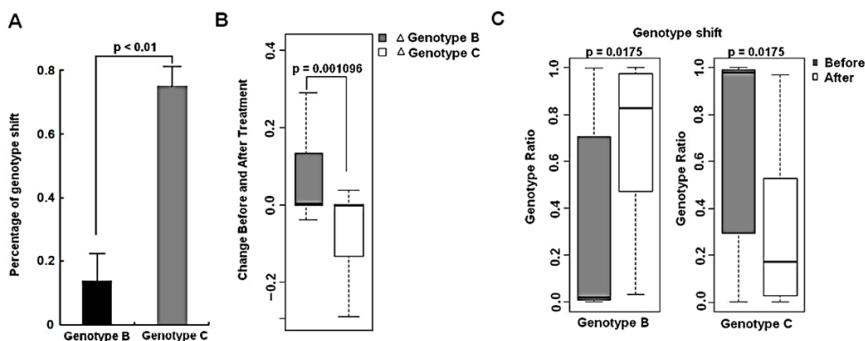


图 2.54 用药前后，发生基因型反转的患者的 B 型和 C 型偏好明显；A: 在阿德福韦治疗之前 B 型主导和 C 型主导的患者在治疗后发生亚型反转的比例。B: 不同患者不同亚型在治疗前后发生丰度变化的情况。C: 不同亚型在治疗前后发生丰度变化的情况

2.6.3 基因型反转现象的解释

对于阿德福韦酯而言，在??里面我们讨论了当发生 (rtN236T, rtA181V) 突变时，阿德福韦药物对于变异株的病毒失效。所以，我通过分析不同 HBV 病毒 B, C 亚型之间的差异位点，来解释 B 型相对与 C 型 HBV 病毒对于阿德福韦酯更加敏感。

2.6.3.1 B,C 亚型的基因组差异

通过比较 HBV 基因组标准序列集合对应的 PSSM 矩阵在 B, C 亚型不同位点的差异；

因为，对应到 HBV 标准序列集合 3674nt 坐标系下，实际上对应的 344aa HBV RT 区聚合酶在实际的坐标是 [220, 1280]nt, 找到实际上的 236th 和 181th 个氨基酸实际上对应的 HBV genome 上的坐标。我们找到实际上这写位点对应

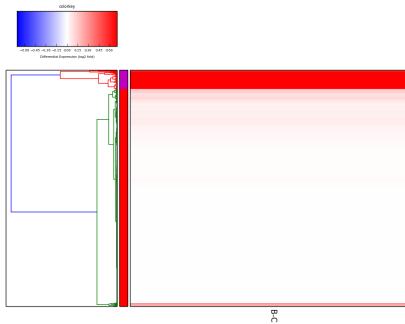
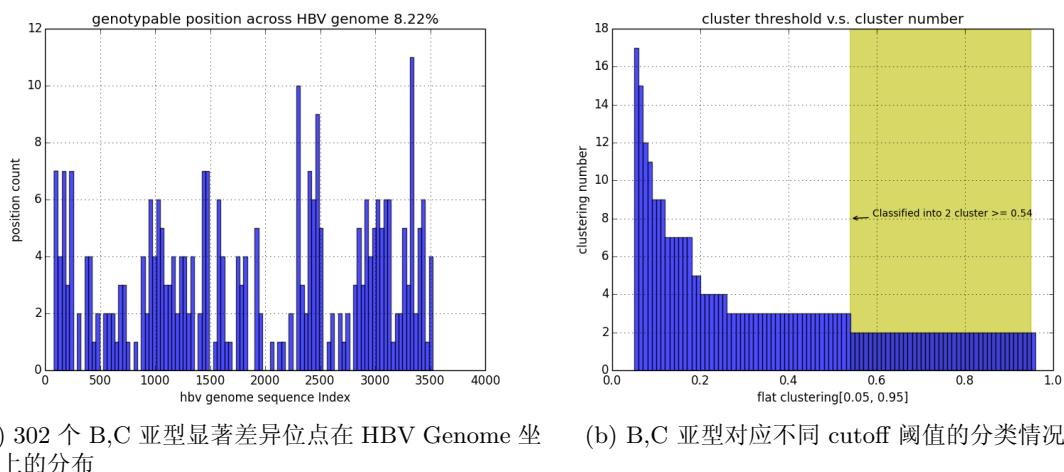


图 2.55 B, C 不同亚型在标准序列集合下得到的 PSSM 矩阵显示的亚型间差异位点聚类, 当模糊阈值大于 $0.54 \max distance$ 时, 我们找到的亚型间显著 PSSM 差异位点一共 302 个



(a) 302 个 B,C 亚型显著差异位点在 HBV Genome 坐标上的分布 (b) B,C 亚型对应不同 cutoff 阈值的分类情况

图 2.56 B, C 亚型在不同的位点上的差异

的 HBV 标准序列集合为 [952, 953, 954] 和 [786, 787, 788]。我们发现, 显著的可分型位点 954 刚好被包含在了阿德福韦的敏感位点中, 此时对应的 B 亚型和 C 亚型在此处的氨基酸密码子为: B 亚型'AAA', C 亚型是'AAC'; 所以我们可以得到 C 亚型的 236 位不变, 而 B 亚型的 236 位对应的氨基酸从天冬酰胺转换为赖氨酸。

我们通过 (2.57a 和 2.57b) 描述整个 HBV 基因组上 [950 ~ 955] 的 HBV 序列在不同亚型上的密码字分布 Logo:

所以, 这就解释了为什么 HBV 的 B 亚型对于阿德福韦酯产生了不敏感,

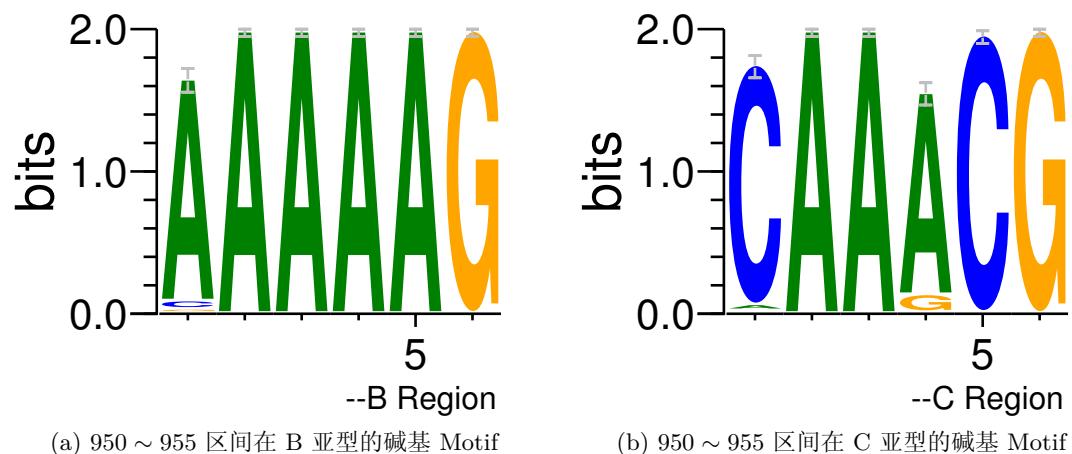


图 2.57 不同 HBV 亚型带来了基因组上的位点碱基偏好性, 导致了 B 型更容易产生抗药性
在药物治疗过程中出现了亚型的倾向性进化。

参考文献

- [1] Dane D S, Cameron C H, Briggs M. Virus-Like Particles in Serum of Patients with Australia-Antigen-Associated Hepatitis. *Lancet*, 1970, 1(7649):695-&. F8933Times Cited:780Cited References Count:9.
- [2] Liu Y, Wang C, Zhong Y, et al. Genotypic resistance profile of hepatitis B virus (HBV) in a large cohort of nucleos (t) ide analogue-experienced Chinese patients with chronic HBV infection. *Journal of Viral Hepatitis*, 2011, 18(4):e29–e39.
- [3] Mirandola S, Sebastiani G, Rossi C, et al. Genotype-specific mutations in the polymerase gene of hepatitis B virus potentially associated with resistance to oral antiviral therapy. *Antiviral research*, 2012..
- [4] Ganem D, Varmus H E. The Molecular Biology of the Hepatitis B viruses. *Annual Reviews Biochemistry*, 1987, 56:651–693.
- [5] Richard Myers¹ A K P K, Tedder² R. Genotyping Hepatitis B virus from whole- and sub-genomic fragments using position-specific scoring matrices in HBV STAR. *Journal of General Virology*, 2006, 87(6):1459–1464.
- [6] Oliveira T. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, 2005, 21(19):3797–3800.
- [7] BRADLEY EFRON E H, HOLMES S. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A*, 1996, 93(7):7085–7090.
- [8] Bell A J, Sejnowski T J. An information-maximization approach to blind separation and blind deconvolution. *NEURAL COMPUTATION*, 1995, 7:1129–1159.
- [9] Ronaghi M. Pyrosequencing sheds light on DNA sequencing. *Genome Res*, 2001, 11(1):3–11. Ronaghi, M*Genome Res*. 2001 Jan;11(1):3-11..
- [10] Kullback S, Leibler R A. On Information and Sufficiency. *Annals of Mathematical Statistics*, 1951, 22(1):79–86. Um018Times Cited:2380Cited References Count:21.

参考文献

- [11] Clarke J, Wu H C, Jayasinghe L, et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 2009, 4(4):265–270. 437AGTimes Cited:228Cited References Count:32.
- [12] Alcantara L C J, Cassol S, Libin P, et al. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Research*, 2009, 37:W634–W642. 469IGTimes Cited:11Cited References Count:28.
- [13] Batuwita R, Palade V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, 2009, 25(8):989–995.
- [14] Lockless S W, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 1999, 286(5438):295–299.
- [15] Bell A J, Sejnowski T J. A new learning algorithm for blind signal separation. *Advances in neural information processing systems*, 1996, 8:757–763. In D. Touretzky, M. Mozer, and M. Hasselmo, editors.
- [16] Williams J R. The declaration of Helsinki and public health. *Bulletin of the World Health Organization*, 2008, 86(8):650–651. 334YNTimes Cited:38Cited References Count:6.
- [17] Liu F, Chen L, Yu D M, et al. Evolutionary patterns of hepatitis B virus quasispecies under different selective pressures: correlation with antiviral efficacy. *Gut*, 2011, 60(9):1269–1277. 803NSTimes Cited:2Cited References Count:35.
- [18] Teshale E H, Ramachandran S, Xia G L, et al. Genotypic Distribution of Hepatitis B Virus (HBV) Among Acute Cases of HBV Infection, Selected United States Counties, 1999–2005. *Clinical Infectious Diseases*, 2011, 53(8):751–756. 821KITimes Cited:0Cited References Count:21.
- [19] Kiesslich D, Crispim M A, Santos C, et al. Influence of Hepatitis B Virus (HBV) Genotype on the Clinical Course of Disease in Patients Coinfected with HBV and Hepatitis Delta Virus. *Journal of Infectious Diseases*, 2009, 199(11):1608–1611. 444SLTimes Cited:9Cited References Count:10.
- [20] Zhao Y, Zhang X Y, Guo J J, et al. Simultaneous Genotyping and Quantification of Hepatitis B Virus for Genotypes B and C by Real-Time PCR Assay. *Journal of Clinical Microbiology*, 2010, 48(10):3690–3697. 659CNTimes Cited:2Cited References Count:36.
- [21] Echevarria J M, Avellon A. Hepatitis B virus genetic diversity. *Journal of Medical Virology*, 2006, 78:S36–S42. Suppl. 1061DXTimes Cited:37Cited References Count:41.

- [22] Balzer S, Malde K, Lanzen A, et al. Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics*, 2010, 26(18):i420–5. Balzer, SusanneMalde, Ketil-Lanzen, AndersSharma, AnimeshJonassen, IngeEnglandOxford, EnglandBioinformatics. 2010 Sep 15;26(18):i420-5..
- [23] Quince C, Lanzen A, Curtis T P, et al. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, 2009, 6(9):639–U27. 488XFTimes Cited:153Cited References Count:20.
- [24] Zagordi O, Klein R, Daumer M, et al. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Research*, 2010, 38(21):7400–7409. 689TKTimes Cited:23Cited References Count:41.
- [25] Bains W, Smith G C. A novel method for nucleic acid sequence determination. *J Theor Biol*, 1988, 135(3):303–7. Bains, WSmith, G CENGLANDJ Theor Biol. 1988 Dec 7;135(3):303-7..
- [26] Beck J, Nassal M. Hepatitis B virus replication. *World Journal of Gastroenterology*, 2007, 13(1):48–64. 126TUTimes Cited:72Cited References Count:162.
- [27] Braslavsky I, Hebert B, Kartalov E, et al. Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A*, 2003, 100(7):3960–3964. 664JRTimes Cited:189Cited References Count:36.
- [28] Brenner S, Williams S R, Vermaas E H, et al. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc Natl Acad Sci U S A*, 2000, 97(4):1665–70. Brenner, SWilliams, S RVermaas, E HStorck, TMoon, KMcCollum, CMao, J ILuo, SKirchner, J JEletr, SDuBridge, R BBurcham, TAlbrecht, GProc Natl Acad Sci U S A. 2000 Feb 15;97(4):1665-70..
- [29] Bruss V. Hepatitis B virus morphogenesis. *World J Gastroenterol*, 2007, 13(1):65–73. Bruss, VolkerChinaWorld J Gastroenterol. 2007 Jan 7;13(1):65-73..
- [30] Chan E Y. Advances in sequencing technology. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis*, 2005, 573(1-2):13–40. 923CXTimes Cited:77Cited References Count:142.
- [31] Oliveira T, Deforche K, Cassol S, et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, 2005, 21(19):3797–3800. 974MOTimes Cited:154Cited References Count:13.
- [32] Ding C, He X. K-means clustering via principal component analysis. *Proceedings of Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004. 29.

参考文献

- [33] Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 1996, 93(23):13429–13429.
- [34] Fedurco M, Romieu A, Williams S, et al. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research*, 2006, 34(3). 016FNTimes Cited:42Cited References Count:32.
- [35] Guenther U P, Yandek L E, Niland C N, et al. Hidden specificity in an apparently nonspecific RNA-binding protein. *Nature*, 2013..
- [36] Harris T D, Buzby P R, Babcock H, et al. Single-molecule DNA sequencing of a viral genome. *Science*, 2008, 320(5872):106–109. 283JRTimes Cited:218Cited References Count:14.
- [37] Higgins D G, Sharp P M. Clustal - a Package for Performing Multiple Sequence Alignment on a Microcomputer. *Gene*, 1988, 73(1):237–244. R7614Times Cited:2738Cited References Count:18.
- [38] Higgins D G, Sharp P M. Fast and Sensitive Multiple Sequence Alignments on a Microcomputer. *Computer Applications in the Biosciences*, 1989, 5(2):151–153. U3780Times Cited:1427Cited References Count:14.
- [39] Howard C R. The Biology of Hepadnaviruses. *Journal of General Virology*, 1986, 67:1215–1235. Part 7D2948Times Cited:29Cited References Count:114.
- [40] Katoh K, Misawa K, Kuma K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 2002, 30(14):3059–3066. 579AYTimes Cited:1005Cited References Count:38.
- [41] Kay A, Zoulim F. Hepatitis B virus genetic variability and evolution. *Virus Research*, 2007, 127(2):164–176. 198YQTimes Cited:50Cited References Count:117.
- [42] Kramvis A, Kew M, Francois G. Hepatitis B virus genotypes. *Vaccine*, 2005, 23(19):2409–23. Kramvis, AnnaKew, MichaelFrancois, GuidoNetherlandsVaccine. 2005 Mar 31;23(19):2409-23..
- [43] Kurbanov F, Tanaka Y, Kramvis A, et al. When should "I" consider a new hepatitis B virus genotype? *Journal of Virology*, 2008, 82(16):8241–8242. 333QFTimes Cited:24Cited References Count:13.
- [44] Li W H, Miao X H, Qi Z T, et al. Hepatitis B virus X protein upregulates HSP90alpha expression via activation of c-Myc in human hepatocarcinoma cell line, HepG2. *Virology Journal*, 2010, 7. 571BPTimes Cited:4Cited References Count:35.

参考文献

- [45] Locarnini S. Molecular virology of hepatitis B virus. *Seminars in Liver Disease*, 2004, 24:3–10. Suppl. 1829XDTimes Cited:62Cited References Count:43.
- [46] Mahtab M A, Rahman S, Khan M, et al. Hepatitis B virus genotypes: an overview. *Hepatobiliary & Pancreatic Diseases International*, 2008, 7(5):457–464. 359JRTTimes Cited:12Cited References Count:52.
- [47] Mardis E R. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 2008, 9:387–402. 354HGTimes Cited:360Cited References Count:55Annual Review of Genomics and Human Genetics.
- [48] NeedlemaSb, Wunsch C D. A General Method Applicable to Search for Similarities in Amino Acid Sequence of 2 Proteins. *Journal of Molecular Biology*, 1970, 48(3):443–&. F9611Times Cited:4223Cited References Count:11.
- [49] Needleman S B, Wunsch C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 1970, 48(3):443–453.
- [50] Notredame C. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, 2002, 3(1):131–144. 557BPTimes Cited:115Cited References Count:84.
- [51] Notredame C, Higgins D G, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 2000, 302(1):205–217. 354WETimes Cited:2323Cited References Count:38.
- [52] Notredame C, Holm L, Higgins D G. COFFEE: An objective function for multiple sequence alignments. *Bioinformatics*, 1998, 14(5):407–422. 106MMTimes Cited:97Cited References Count:44.
- [53] Ronaghi M, Karamohamed S, Pettersson B, et al. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 1996, 242(1):84–89. Vt729Times Cited:370Cited References Count:24.
- [54] Rusk N. Cheap third-generation sequencing. *Nature Methods*, 2009, 6(4):244–245. 426WKTTimes Cited:15Cited References Count:1.
- [55] Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 1977, 74(12):5463–7. Sanger, FNicklen, SCoulson, A RProc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7..
- [56] Shendure J, Porreca G J, Reppas N B, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 2005, 309(5741):1728–1732. 963WKTTimes Cited:479Cited References Count:26.

参考文献

- [57] Dando T M, Plosker G L. Adefovir Dipivoxil: A Review of its Use in Chronic Hepatitis B. *ADIS DRUG EVALUATION*, 2003, 63(20).
- [58] Turcatti G, Romieu A, Fedurco M, et al. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research*, 2008, 36(4). 272KDTimes Cited:24Cited References Count:44.
- [59] Whitfeld P R. A Method for the Determination of Nucleotide Sequence in Polyribonucleotides. *Biochemical Journal*, 1954, 58(3):390–396. Ub568Times Cited:114Cited References Count:24.
- [60] Zeng G, Wang Z, Wen S, et al. Geographic distribution, virologic and clinical characteristics of hepatitis B virus genotypes in China. *Journal of Viral Hepatitis*, 2005, 12(6):609–617. 975JXTimes Cited:50Cited References Count:33.

附录

2.7 454 测序和 Solexa 测序后分型结果的对比数据

Table 4 solexa, 454 分型结果比较 (n=37)

Patient	Solexa(reads)			454(reads)		
	Total(reads)	B%	C%	Total(reads)	B%	C%
barcode224	113144	0.02 (22)	99.98 (113122)	1969	3.86 (76)	96.14 (1893)
barcode196	110094	0.02 (22)	99.98 (110072)	532	1.88 (10)	98.12 (522)
barcode202	93881	0.08 (77)	99.92 (93804)	1539	1.10 (17)	98.90 (1522)
barcode227	343512	0.11 (373)	99.89 (343139)	4205	0.14 (6)	99.86 (4199)
barcode217	109385	0.24 (267)	99.76 (109118)	4006	5.32 (213)	94.68 (3793)
barcode218	25494	6.10 (1554)	93.90 (23940)	2807	2.57 (72)	97.43 (2735)
barcode199	34666	20.85 (7228)	79.15 (27438)	3716	20.67 (768)	79.33 (2948)
barcode197	26715	28.44 (7597)	71.56 (19118)	4749	47.23 (2243)	52.77 (2506)
barcode223	85028	48.88 (41562)	51.12 (43466)	5950	60.82 (3619)	39.1 (2331)
barcode216	44424	76.12 (33816)	23.88 (10608)	3707	94.36 (3498)	5.64 (209)
barcode206	9356	92.63 (8666)	7.37 (690)	4344	89.89 (3905)	10.11 (439)
barcode201	93738	98.00 (91859)	2.00 (1879)	4321	98.68 (4264)	1.32 (57)
barcode226	17734	98.64 (17492)	1.36 (242)	4158	96.66 (4019)	3.34 (139)
barcode190	51206	99.25 (50823)	0.75 (383)	5468	96.27 (5264)	3.73 (204)
barcode200	63965	99.39 (63576)	0.61 (389)	5537	96.33 (5334)	3.67 (203)
barcode228	81174	99.57 (80825)	0.43 (349)	5056	98.62 (4986)	1.38 (70)
barcode214	120545	99.73 (120219)	0.27 (326)	5961	95.24 (5677)	4.76 (284)
barcode198	104472	99.75 (104210)	0.25 (262)	2375	98.53 (2340)	1.47 (35)
barcode195	49331	99.76 (49211)	0.24 (120)	1134	97.53 (1106)	2.47 (28)
barcode220	126116	99.80 (125869)	0.20 (247)	4112	99.78 (4103)	0.22 (9)
barcode213	104059	99.82 (103872)	0.18 (187)	5329	99.98 (5328)	0.02 (1)
barcode194	135279	99.91 (135156)	0.09 (123)	518	96.14 (498)	3.86 (20)
barcode222	57846	99.92 (57799)	0.08 (47)	5946	99.63 (5924)	0.37 (22)
barcode221	91231	99.92 (91160)	0.08 (71)	5840	98.17 (5733)	1.83 (107)
barcode212	87205	99.92 (87138)	0.08 (67)	3322	99.67 (3311)	0.33 (11)
barcode205	121044	99.92 (120953)	0.08 (91)	1795	98.22 (1763)	1.78 (32)
barcode225	163025	99.93 (162905)	0.07 (120)	3478	99.80 (3471)	0.20 (7)
barcode215	98235	99.93 (98164)	0.07 (71)	4442	95.79 (4255)	4.21 (187)
barcode192	114509	99.93 (114431)	0.07 (78)	4127	99.76 (4117)	0.24 (10)
barcode207	129382	99.94 (129307)	0.06 (75)	1383	99.78 (1380)	0.22 (3)
barcode209	125853	99.95 (125784)	0.05 (69)	957	99.79 (955)	0.21 (2)
barcode191	113179	99.95 (113117)	0.05 (62)	2744	99.64 (2734)	0.36 (10)
barcode204	91504	99.95 (91458)	0.05 (46)	2113	99.29 (2098)	0.71 (15)
barcode189	102007	99.96 (101962)	0.04 (45)	5329	94.78 (5051)	5.22 (278)
barcode203	111701	99.97 (111662)	0.03 (39)	2358	99.53 (2347)	0.47 (11)
barcode211	104756	99.97 (104720)	0.03 (36)	3073	99.64 (3062)	0.36 (11)
barcode219	110252	99.97 (110217)	0.03 (35)	4547	99.93 (4544)	0.07 (3)

2.8 Solexa 测序 Barcode 及组合设计

表 2.3 Solexa 测序 Barcode 及组合设计

barcodeID	上游 barcode 序列	下游 barcode 序列
BC0011_83	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTTGATAGTGGT
BC0011_88	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTGGAGTATATT
BC0011_89	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTAGTTGATGAT
BC0011_91	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCATAGTAGTGGT
BC0011_102	CACCTATCACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTTGATAGTGGT
BC0011_110	CACCTATCACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCATAGTAGTGGT
BC0011_137	CTATTCTAATCTACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTGGAGTATATT
BC0011_138	CTATTCTAATCTACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCATAGTAGTGGT
BC0011_145	CTACTTCTCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTGGAGTATATT
BC0011_150	CTACTTCTCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCAGGAATTGTGG
BC0011_75	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGATTGGATTGG
BC0011_79	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGTAATGTGTGG
BC0011_81	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCAAGTAGTTATG
BC0011_82	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTAGGATGGAAG
BC0011_84	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGATAGTAATGG
BC0011_87	CACTTATCCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCAGGAATTGTGG
BC0011_92	CACTTATCACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGAATTGGATTGG
BC0011_97	CACTTATCACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGAATTGGATTGG
BC0011_99	CACTTATCACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGGATGAGAGTG
BC0011_100	CACTTATCACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGTAATGTGTGG
BC0011_11	ACCTTACTACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTTGATAGTGGT
BC0011_16	ACCTTACTACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTGGAGTATATT
BC0011_17	ACCTTACTACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTAGTTGATGAT
BC0011_19	ACCTTACTACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCATAGTAGTGGT
BC0011_30	ACTATCTACTTACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTTGATAGTGGT
BC0011_50	TTATCATTATTACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTGAATTAAATAG
BC0011_57	TCCCTCTCCAATACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCAGGAATTGTGG
BC0011_62	TCCCTCTCCAATACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGAATTGGATTGG
BC0011_63	TCCCTCTCCAATACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGGATGAGAGTG
BC0011_64	TCCCTCTCCAATACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGTAATGTGTGG
BC0011_67	TCCCTCTCCAATACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCTAGGATGGAAG
BC0011_70	TCCCTCTCCAATACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGATAGTAATGG
BC0011_111	TACCTCCTCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCAGGAATTGTGG
BC0011_115	TACCTCCTCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGAATTGGATTGG
BC0011_117	TACCTCCTCTATAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCAAGTAGTTATG
BC0011_1	ACCTTACTACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCAGGAATTGTGG
BC0011_6	ACCTTACTACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGAATTGGATTGG
BC0011_9	ACCTTACTACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCGTAATGTGTGG
BC0011_10	ACCTTACTACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCAAGTAGTTATG
BC0011_12	ACCTTACTACTAACGTCCCGTCGGCGCTGAA	CTCCCCGTCTGTGCCTCTCATAGGATGGAAG

附录

表 2.4 Solexa 测序 Barcode 及组合设计

barcodeID	上游 barcode 序列	下游 barcode 序列
BC0011_15	ACCTTACTACTTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGATAGTAATGG
BC0011_29	ACTATCTACTTTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCAAGTAGTTATG
BC0011_32	ACTATCTACTTTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGAATTAAATAG
BC0011_34	ACTATCTACTTTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGATAGTAATGG
BC0011_177	ACATCCTCCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCAGGAATTGTGG
BC0011_51	TTATCATTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCAAGTAATATAA
BC0011_60	TCCCTCTCCAATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTAGTGGTTAGA
BC0011_61	TACCTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGGGATAAA
BC0011_112	TACCTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCATTAAGTGGTA
BC0011_113	TACCTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCATATGTGGATAAA
BC0011_76	CACTTATCCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCATTAAGTGGTA
BC0011_77	CACTTATCCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTAGTGGTTAGA
BC0011_78	CACTTATCCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTAGTGGTTAGA
BC0011_95	CACCTATCACTTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTAGTGGTTAGA
BC0011_96	CACCTATCACTTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTAGTGGAGATGA
BC0011_98	CACCTATCACTTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTATGTGGATAAA
BC0011_127	CTATTCTAATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCATTAAGTGGTA
BC0011_128	CTATTCTAATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTGGTTAGA
BC0011_129	CTATTCTAATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCATTAAGTGGTA
BC0011_140	CTATTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTGGTTAGA
BC0012_3	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTGGAGGT
BC0012_4	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTGTGAAGTAAGT
BC0012_8	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCAAGACTGGTAGT
BC0012_10	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCAAGGATGGTGGT
BC0012_23	CCTCTATCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTGGAGGT
BC0012_24	CCTCTATCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGTGTGAAGTAAGT
BC0012_27	CCTCTATCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCAAGACTGGTAGT
BC0012_29	CCTCTATCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCAAGGATGGTGGT
BC0012_38	CTTAECTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTGGAGGT
BC0012_39	CTTAECTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTGTGAAGTAAGT
BC0012_2	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTAGGATATAAG
BC0012_7	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGAATGATGGAAG
BC0012_9	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTTGTTGAGTG
BC0012_17	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTAGTAGTG
BC0012_19	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTTGAGTAGGAG
BC0012_22	CCTCTATCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTGTAGGATATAAG
BC0012_26	CCTCTATCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGAATGATGGAAG
BC0012_28	CCTCTATCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTTGTTGAGTG
BC0012_34	CCTCTATCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTAGGATATAAG
BC0012_37	CTTAECTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTAGTAGTG
BC0012_94	ATTATCTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAAGTAGTATGT
BC0012_142	AATCTTCCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTAGGAGGT
BC0012_143	AATCTTCCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGTGTAGGATATAAG
BC0012_144	AATCTTCCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGAATGATGGAAG
BC0012_152	AATCTTCCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTTGTTGAGTG
BC0012_71	TCACACTTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTAGTAGTG
BC0012_76	TCACACTTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTTGAGTAGGAG
BC0012_78	TCACACTTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGAATGATGGAAG
BC0012_79	TCACACTTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTTGTTGAGTG
BC0012_88	TCACACTTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTAGTAGTG
BC0012_90	TCACACTTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTTGAGTAGGAG
BC0012_118	TACTTCTTCCCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTGAATGATGGAAG
BC0012_290	TCCTCTCTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGAATGATGGAAG
BC0012_296	TCACCATTAACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGGATATGTGGTG
BC0012_301	TCACCATTAACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTTGTTGAGTG
BC0012_93	ATTATCTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGAGTATAAG
BC0012_102	ATTATCTCTCTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTGAATGATGGAAG
BC0012_141	AATCTTCCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCGAATGATGGAAG
BC0012_157	AATCTTCCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTGTAGGATATAAG
BC0012_237	ACCTTCCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCTAGTAGTAGTG
BC0012_246	ACCTTCCCTTATACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCCTCTCATAGTAGTAGTG

附录

表 2.5 Solexa 测序 Barcode 及组合设计

barcodeID	上游 barcode 序列	下游 barcode 序列
BC0012_248	ACCTTCCTTATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTTGTAGTGGTAGG
BC0012_251	ACCTCATCTTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCTGTAGGATATAAG
BC0012_256	ACCTCATCTTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCTGTGTTGAGTG
BC0012_262	ACCTCATCTTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCATAGTAGTAGTG
BC0012_81	TCACACTTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCGTAGGATGGTGA
BC0012_85	TCACACTTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGGTTAGGATGA
BC0012_87	TCACACTTACACTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGGATGAGGTGA
BC0012_120	TACTTCTTCTTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAAGAGTGTATAAA
BC0012_123	TCATAATCTTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCTAGGATATGTAA
BC0012_14	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGGTTAGGATGA
BC0012_16	CACACACTTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGGATGAGGTGA
BC0012_54	CTTATCTACCTATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCTAGGATATGTAA
BC0012_59	CTTATCTACCTATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGTGAATAGTA
BC0012_65	CTTATCTACCTATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCATGGTTATAGA
BC0012_66	CTTATCTACCTATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAAGAGTGTATAAA
BC0012_163	CCAATATCCTTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAAGAGTGTATAAA
BC0012_166	CCTCTAATCTTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGGATATGTAA
BC0012_169	CCTCTAATCTTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGTGAATAGTA
BC0012_173	CCTCTAATCTTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAAGTGGTAATAAA
BC006_7	CTATATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAATAAT
BC006_22	CTATATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCGTTAGT
BC006_274	CATACATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGTGGT
BC006_278	CATACATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCGTTAGT
BC006_364	CCTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGTGGT
BC006_368	CCTATCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCGTTAGT
BC006_389	CCTTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGTGGT
BC006_395	CCTTATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCGTTAGT
BC006_487	CAACTCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGTGGT
BC006_499	CAACTCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCGTTAGT
BC006_6	CTATATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCATTGTG
BC006_8	CTATATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCATGTAG
BC006_10	CTATATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCGTTAAG
BC006_19	CTATATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCTTAGTGT
BC006_25	CTATATTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCTATAAG
BC006_264	CATACATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGGTTG
BC006_267	CATACATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCATTGTG
BC006_270	CATACATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGTGGT
BC006_275	CATACATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCGGTATG
BC006_276	CATACATACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGTTGG
BC006_94	ATCTCCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCAGTGGT
BC006_106	ATCTCCTACGTCCCGTCGGCGCTGAA	CTCCCCGCTGTGCCTCTCGTTAGT

附录

表 2.6 Solexa 测序 Barcode 及组合设计

barcodeID	上游 barcode 序列	下游 barcode 序列
BC006_134	AACCTCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGTGGT
BC006_146	AACCTCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGTGGT
BC006_304	ATCATCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGTGGT
BC006_31	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGTGGT
BC006_32	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGTGGT
BC006_38	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCATTGTG
BC006_40	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGTGGT
BC006_42	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTTAAG
BC006_47	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGTAGG
BC006_51	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTATAG
BC006_53	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAATGG
BC006_54	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAAGATG
BC006_57	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAATTAG
BC006_75	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTAGTAG
BC006_76	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGGTG
BC006_79	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTTGG
BC006_81	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCATTGTG
BC006_83	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCATGTAG
BC006_85	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTTAAG
BC006_90	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGTAGG
BC006_91	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTGTGG
BC006_93	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTATAG
BC006_95	ATCTCCTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAAGATG
BC006_30	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTTATA
BC006_34	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGGTA
BC006_35	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGGTA
BC006_37	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGTATA
BC006_41	TCACAATACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGATTAA
BC006_1	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTTATA
BC006_3	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGGTA
BC006_4	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGGTA
BC006_9	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGATTAA
BC006_11	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGATAAA
BC006_12	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCATAGTA
BC006_13	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCATGTGA
BC006_15	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTGTAA
BC006_17	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTAGTTA
BC006_26	CTATATTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTTATGA
BC007_45	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGTAAAGT
BC007_48	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGATGTAT
BC007_51	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCATAGGAT
BC007_53	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGGATAT
BC007_58	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGGATGT
BC007_60	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGGTAGT
BC007_65	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGGAATT
BC007_71	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAAGTAAT
BC007_73	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGAATAGT
BC007_74	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGTATAT
BC007_44	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGTAAATG
BC007_47	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCATAGGTG
BC007_50	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCATATAGG
BC007_54	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAATGTAG
BC007_57	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGGATTG
BC007_63	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGAATGTG
BC007_67	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTGTATAG
BC007_68	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGTATAG
BC007_69	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGTTGAG
BC007_75	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGATGAG
BC007_3	ACCATTATACTGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCGATGTAT
BC007_4	ACCATTATACTGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCATAGGAT
BC007_10	ACCATTATACTGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCTGGATGT
BC007_12	ACCATTATACTGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTCAGGTAGT
BC007_24	ACCATTATACTGTCCCCGTCGGCGCTGAA	CTCCCCGCTCTGTGCCTTCTAGTGAT

表 2.7 Solexa 测序 Barcode 及组合设计

barcodeID	上游 barcode 序列	下游 barcode 序列
BC007_200	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTGTAATG
BC007_203	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCATAGGTG
BC007_206	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCATATAAGG
BC007_210	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCAATGTAG
BC007_213	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCAGGATTG
BC007_219	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGAATGTG
BC007_223	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGTATAG
BC007_224	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTGATATG
BC007_225	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGTTGTAG
BC007_231	TACCTTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTGATGAG
BC007_6	ACCATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCAATGTAG
BC007_9	ACCATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCAGGATTG
BC007_17	ACCATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGTATAG
BC007_18	ACCATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGTTGTAG
BC007_20	ACCATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTGATGAG
BC007_32	ACCATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGATAATG
BC007_33	ACCATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCAAGAGTG
BC007_40	ACCATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTGTATGG
BC007_42	ACCATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTGTTAGG
BC007_452	ACTACCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTGTAATG
BC007_101	TAATATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGAATATA
BC007_106	TAATATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTGATTAA
BC007_110	TAATATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCATATTAA
BC007_115	TAATATTATAACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCATTGTAA
BC007_122	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTATATGA
BC007_43	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGTGGTAA
BC007_46	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGAATATA
BC007_49	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCAAGGTA
BC007_55	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTGATTAA
BC007_56	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGAATTGA
BC007_62	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCATGGATA
BC007_66	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCATTGTAA
BC007_70	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCTAGTATA
BC007_72	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCAGGTTAA
BC007_76	CTTATCTTACGTCCCCGTCGGCGCTGAA	CTCCCCCGTCTGTGCCTTCTCGTTATGA

致 谢

这是我在中国科学技术大学的 7 年，也是我进入吴家睿教授实验组的 3 年，在这段时间里，我所从事的学习和研究工作，都是在导师以及系里其他老师和同学的指导和帮助和影响下进行的。在完成论文之际，请容许我对他们表达诚挚的谢意。

首先，我应该感谢的我本科和研究生期间的实验室的导师，吴家睿教授和梁治老师，在这三年多的时间里，两位教授导师都在我整个研究生生涯中作出了十分重要的指导，纠正惰性和愚蠢，偏执和狂妄。也在我迷茫的时候给我指明方向，培养我良好的科研态度和习惯，帮助我从阅读文献，讲解报告，编写代码等一系列细节开始，逐渐走上正途。与此同时，老师们在科研工作中严谨的态度，扎实的逻辑推理，同时又不失大胆假设的做法，让我逐渐能够明白很多研究工作和思考问题本身的平衡和本质。逻辑是一切的出发点，但逻辑不等价与不反常规的理念，是我时刻提醒着我自己的，虽然自己现在仍然常常徘徊于胡想便不顾逻辑，推理就钻牛角尖的两个极端之间，但是，我希望自己能够在振荡的过程中，收敛到导师们呈现给我面前的平衡点上。这是我应该继续学习和思考的地方，也是老师们给我的最重要的启示。

然后，我还应当感谢所有和我讨论过以及我请教过的同学，老师。我深知自己的愚昧和局限，但是，同学和老师们仍然能够在工作问题上给我很多的指点和启发，看他们的工作，能够让我学到很多东西。听他们给我讲解他们对问题本身的理解，即能够让我学会更多的知识，也能够帮助我摆脱自己时常因为错误经验带来的偏执和短视，无论我曾经的心态如何，现在，我都必须感谢他们。同时，感谢 Lewis Wang 在我入门 coding 以及系统生物学实验室过程中给我的建议以及给我树立的榜样。

同时，我必须再次感谢亲自指导我写代码的梁治老师，周宏师兄以，周斌老师和 Rama。让我重新的认识和了解 coding，了解逻辑体系。了解 perl, python, C++, cuda 以及 matlab，重新理解智能。以至于彻底改变了我看待世界的方式。

致 谢

法，虽然我现在在编程上仍然很弱，但是，有了老师和师兄的作为榜样，有他们的代码对比来帮我培养良好习惯，我会继续努力进步。

这里还要再次感谢周宏师兄曾经借给我的两本书；一本是漫画版的罗素的故事，一本是纽曼先生编写的哥德尔证明。谢谢师兄的这两本书，以及他推荐给我的所有论文。虽然至今我仍然没能完全的理解哥德尔的思想，但是，他的想法和模式表达的技巧会一直影响着伴随着我思考。正如系统生物学本身就是面对问题的构造本身一样，深刻的理解和表达生物体模式是我们的工作目标之一。

我还要感谢在这个 HBV 分型工作中与我合作过的郭志昂师弟，在他的影响下，我开始使用 LaTex，没有他的帮助，今天的这篇论文还是很可能会来自我蹩脚的 Microsoft office 的编辑，充斥着没有规范的排版了乱糟糟的文献索引。同时，感谢中国科学技术大学的 google code 开源项目提供的经过多位前辈修改的 LaTex 模版。

我还要感谢我一直以来所有的同学们，作为我整个 7 年里对应的所有的扰动项，你们的一举一动都决定了过去几年里的我在这里的一切，感谢你们让我明白我还有很多的不足，也能够在这些年中，越来越觉得充满激情和动力。

最后，我还要感谢我的母亲，是她一直以来对我的无私支持，我才能有今天。

周鑫

2013 年 11 月 10 日

在读期间发表的学术论文与取得的研究成果

研究工作:

High prevalence of mixed HBV infection contributes to genotype shift
during antiviral therapy

2011-2012

Background implementation of web server (Two-step pipeline of this work
have run on web server: <http://netalign.ustc.edu.cn/ShwinGen/>)

NGS deep sequencing data correction and processing

HBV segment subtype assignment via Bayesian inference

Barcode Designment for NGS.

Integrating the environmental factor into the strategy updating rule
to promote cooperation in evolutionary games

2011

Data statistic and data visualization.

Complex Network Analysis.(Degree distribution, Betweenness, Clossness)

Small RNAjs regulation modeling and evolutionary selection

2011-2012

ODE construction via control motif, parameter search, parameter sampling
in ODE simulation.

RegulonDB's transcriptional regulation network analysis

GPGPU based smith-waterman algorithm acceleration

2011-2012

Cuda kernel algorithm optimization, program interface implementation.

Transplant Smith-Waterman Algorithm to CUDA

The study of life science collaboration network of China

2011-2012

Data preprocessing, network visualization, network analysis, network cliques partition

Processing raw context by NLP technology

Driver node detection on E.coli's gene regulation network from controllability perspective

2011-2012

Data retrieve, driver nodes detection algorithm implement, data analysis.

Protein-Protein interaction interface evolution analysis

2013-Now

PDB Crawler, PDB and PFam and Uniprot Database intergrating

Evolutionary analysis on primary sequence

Structure Alignment

已发表论文:

1. Integrating the environmental factor into the strategy updating rule to promote cooperation in evolutionary games

[Zhao Lin et al 2012 Chinese Phys. B 21 018701 doi: 10.1088/1674-1056/21/1/018701]

Zhao Lin, Zhou Xin, Liang Zhi and

Wu Jia-Rui

待发表论文:

1. High prevalence of mixed HBV infection contributes to genotype shift during antiviral therapy

[Wang, Yu-Wei, Zhi Liang, Xin Zhou et al] (under review)