遗传 HEREDITAS (Beijing) 2012 年 6 月, 34(6): 765－772
ISSN 0253-9772 www.chinagene.cn

研究报告

DOI: 10.3724/SP.J.1005.2012.00765

30 株大肠杆菌的泛基因组学特征分析

付静[1,2], 秦启伟[1]

1. 中国科学院南海海洋研究所海洋生物资源可持续利用重点实验室, 广州 510301;
2. 中国科学院研究生院, 北京 100049

摘要: 泛基因组(Pan-genome)是某一物种全部基因的总称, 其中包括核心基因组(该物种所有个体中都存在的基因)和非必须基因组(只在部分个体中存在的基因, 以及某个体特有的基因)。文章从泛基因组学角度比较分析了 30 株已经完成测序的大肠杆菌的基因、基因组成及其进化特征, 结果表明核心基因只占据每株大肠杆菌全部基因数目的 50%左右, 而平均每个菌株有 146 个特有基因, 结果表明随着更多大肠杆菌菌株的基因组被测序, 将会不断有新基因被发现。通过比较分析大肠杆菌不同菌株之间基因的保守性与基因的 GC 含量以及选择压力之间的关系, 发现越保守的基因其 GC 含量变化范围越窄, 同时在进化中受到的选择压力也越大。这些结果将有助于深入了解大肠杆菌基因组的进化特征及其基因组成的动态变化, 并为预防和控制由致病性大肠杆菌引发的流行疾病提供理论依据, 同时也为大规模病原菌基因组数据的分析方法提供借鉴。

关键词: 泛基因组; 大肠杆菌; GC 含量; 选择压力

Pan-genomics analysis of 30 *Escherichia coli* genomes

FU Jing[1, 2], QIN Qi-Wei[1]

1. *Key Laboratory of Marine Bio-resources Sustainable Utilization*, *South China Sea Institute of Oceanology*, *Chinese Academy of Sciences*, *Guangzhou* 510301, *China*;
2. *Graduate University of the Chinese Academy of Sciences*, *Beijing* 100049, *China*

Abstract: A pan-genome describes the full complement of genes in species. It is a superset of all the genes in all the individuals of a species, which is composed of a 'core genome' containing genes present in all individuals, and a 'dispensable genome' containing genes present only in some individuals and individual-specific genes. From pan-genome sight, 30 finished genomes from *Escherichia coli* were employed to analyze their gene and genome compositions and evaluation in this study. The results indicated that the core genes accounted for about 50% of the total number of genes, while about 146 strain-specific genes existed in the each strain tested. The data suggests that the *E. coli* pan-genome is vast, and unique genes will continue to be identified when more *E. coli* genomes are sequenced. After analyzing relationships of the gene conservation, GC content and selection pressure in different strains tested, we found that more conserved genes had a narrow range of GC content, and they also bear more selection pressure. These results will be helpful for better understanding

收稿日期: 2011−10−11; 修回日期: 2012−01−13
基金项目: 国家杰出青年基金项目(编号：30725027)资助
作者简介: 付静, 硕士研究生, 专业方向：海洋生物学。E-mail: fujing871123@gmail.com
通讯作者: 秦启伟, 研究员, 博士生导师, 研究方向：海洋环境微生物功能基因组学。E-mail: qinqw@scsio.ac.cn
网络出版时间: 2012-4-11 10:40:52
URL: http://www.cnki.net/kcms/detail/11.1913.R.20120411.1040.005.html

of the evolution profile of *E. coli* genome, and the dynamic changes of its gene compositions. The *E. coli* pan-genome provides useful information for prevention and control of the diseases caused by pathogenic *E. coli*, and also provides a paradigm for the large-scale analysis of pathogenic bacteria genomes.

Keywords: Pan-genome; *E. coli*; GC-content; selection pressure

近年来由于DNA测序技术的迅速发展，大量的细菌全基因组序列已被逐渐报道。在基因组时代初期，人们认为单个菌株的基因组足以用来描述其菌种的基因复杂性，比较基因组学主要被应用于研究相近微生物间的遗传进化和分类。随着某些细菌越来越多菌株基因组序列的测定，应用抑制消减杂交(Subtractive hybridization)和比较基因组杂交(Comparative genome hybridization, CGH)方法研究发现细菌种内基因组间存在很大差异。造成种间基因组差异的原因是多方面的，包括基因的复制和缺失、基因组重组以及通过基因横向转移获取新基因等[1]。由于细菌基因组存在丰富的多样性，因此一直缺乏准确有效的概念从基因组角度描述细菌物种。2005 年，Tettelin等[2]提出了微生物泛基因组概念(Pan-genome，pan源自希腊语‘παν’，全部的意思)，泛基因组即某一物种全部基因的总称，包括核心基因组和非必须基因组。核心基因组由那些在所有菌株中都存在的基因组成，一般与细菌基本生物学功能和主要表型特征相关；非必须基因组由只在部分菌株中存在的基因以及某个菌株特有的基因组成，非必须基因体现了细菌菌株多样性，一般与某些特别的生物代谢途径和环境适应相关，比如产抗生素、耐受性，毒力等方面。根据细菌泛基因组大小与基因组数目的关系，细菌的泛基因组可以分为开放型泛基因组和闭合型泛基因组。开放型泛基因组是指随着测序的基因组数目不断增加，泛基因组大小无限增加，而闭合型泛基因组则会在测序的基因组数目增加到一定程度后收敛于某一恒定值。随着这一概念的提出，泛基因组已被逐渐应用于描述微生物物种，泛基因组学方法也相继被用于B型无乳链球菌、肺炎链球菌等物种的基因组演化分析[3,4]，以及病原菌疫苗制备[5]等方面的研究。

作为模式生物，大肠杆菌一直被广泛地用于生物学研究。目前已有多个大肠杆菌的不同菌株全基因组序列被报道，这为从基因组学角度分析大肠杆菌的物种演化、基因组结构的遗传变异等提供了有利的资源。Rasko等[6]比较分析了大肠杆菌致病株和非致病株在基因组结构上的差异，Touchon等[7]通过比较 20 株大肠杆菌基因组序列，对大肠杆菌的多样性演化进行了比较分析，Wagied等[8]分析了 5 株大肠杆菌基因组序列间基因横向转移对操纵子以及蛋白相互作用网络的影响。但是，目前国内外对大肠杆菌基因组的研究主要集中于物种的演化、基因横向转移以及不同菌株间基因组结构的差异这几个方面，而大肠杆菌基因组中基因的组成与保守性研究始终是一个空缺。因此，本文将从泛基因组学角度，通过比较研究目前已经测序的 30 株大肠杆菌全基因组序列，来分析大肠杆菌的泛基因组特征及其基因的保守性。

1 材料和方法

1.1 研究对象

30 株大肠杆菌基因组完成图序列以及注释信息数据从 NCBI 的 FTP 中下载，其序列号分别为：AC_000091，NC_004431，NC_009800，NC_010498，NC_011741，NC_011750，NC_012967，NC_013364，NC_000913，NC_007946，NC_009801，NC_011353，NC_011742，NC_011751，NC_013008，NC_013941，NC_002655，NC_008253，NC_010468，NC_011415，NC_011745，NC_012759，NC_013353，NC_002695，NC_008563，NC_010473，NC_011601，NC_011748，NC_012947，NC_013361。

必须基因数据从天津大学database of essential gene[9] (http://tubic.tju. edu.cn/deg/) 下载，然后提取其中*E.coli*的数据集。

1.2 方法

1.2.1 构建同源基因类

提取 30 个细菌中的全部蛋白质序列存为FASTA格式后，利用BLAST中的blast all程序进行all

vs all的蛋白序列比对，其中 e-value设为 1e-10，比对结果中选取score≥50, coverage≥50%以及identity≥50%的匹配记录，然后将过滤后的结果利用mcl程序进行聚类分析[4,10,11,12]。

聚类结果中，根据每一个同源基因类里面基因覆盖的菌株数，定义每个同源基因类的保守性值(CV)，如：某个同源基因类里面包含了来自 20 个菌株的基因，则此同源基因类的 CV 为 20，其他依次类推。

1.2.2　泛基因组分析

为了推断大肠杆菌中全部同源基因类数目(p)以及核心同源基因类数目(c)与菌株数目(n)之间的关系，分别计算n= 1, 2 … 30 时的p与c的值。为了避免抽样引起的偏差，对于某一特定n值，分别计算C(30,n)种组合下的p值与c值，并取其平均值[4]。

1.2.3　同源基因类的 COG 分析

同源基因类的分析过程中，首先需要确定每一个基因类的COG分类[13]。在细菌基因组注释时有一些基因缺少COG分类信息，在NCBI数据库中被标记为“–”，后续分析算法中认定这样的分类为“不明确的COG分类”。有些基因的COG分类号码最后一位为R或者S，根据COG数据库定义，这两类属于功能不能被准确注释的分类(COG数据库中原词为“Poorly characterized”)，后续分析算法中认定这样的分类为“不可靠的COG分类”。

对于每一个同源基因类，在分析器 COG 分类过程中采取以下算法流程：

(1)若该同源基因类中至少有一个基因的 COG 分类不是属于“不明确的 COG 分类”或“不可靠的 COG 分类”时；去掉那些 COG 分类是“不明确的 COG 分类”和“不可靠的 COG 分类”的基因，剩余的基因 COG 分类即被认定为该基因类的 COG 分类。

(2)若该同源基因类中所有的基因的 COG 分类都属于“不明确的 COG 分类”或“不可靠的 COG 分类”时，根据以下原则判断：若所有的基因的 COG 分类均为“不明确的 COG 分类”，则该同源基因类为“不明确的 COG 分类”；否则归为“不可靠的 COG 分类”。

1.2.4　同源基因类的选择压力分析

首先，利用mafft[14]软件对每个同源基因类中基因的蛋白序列进行多序列比对，然后利用Perl脚本将蛋白比对结果转换成相对应的核酸比对结果，统计每个同源基因类中同义突变数目(ds)和非同义突变数目(dn)，通过dn/ds 的值来评估同源基因类的进化选择压力。

2　结果与分析

2.1　同源基因类分析

根据 1.2.1.中的定义，实验中依据每一个同源基因类里面基因覆盖的菌株数，给予每个同源基因类一个保守性值(CV)。不同的 CV 值体现了每个基因类的基因在菌株中分布的差异性，CV 值越大说明这个基因在细菌群体里面分布越广泛，也说明该基因类在细菌中越保守。

在 30 个大肠杆菌基因组中共提取蛋白编码基因 141 473 个，通过聚类分析被分成 11 670 类，其中每一种CV值的同源基因类数目如图 1 所示。其中CV值为 30 的同源基因类，即核心基因总数为 2 344 个(约占总同源基因类数目的 20%，平均每个菌株中49.7%的基因为核心基因)，而CV值为 1 的同源基因类，即菌株特有的基因数目为 4 379 个(约占总数的37.5%，平均每个菌株有 146 个特有基因)。而在*Haemophilus influenza* (*H.influenza*)的基因组研究中，核心基因数目为 1 461(占总数 2786 的 52.4%，平均每个菌株中 79.9%的基因为核心基因)，菌株特有的基因数目为 539(占总同源基因类数 2786 的 19.3%，平均每个菌株特有基因数目为 39)[15]。在 17 个*Streptococcus pneumoniae* (*S.pneumoniae*)基因组的研究中，核心基因组数目为 1 454(占全部同源基因类总数 46%，平均每个菌株中约 73%的基因为核心基因)，菌株特有基因数目约为 576(占总同源基因数目的 1%，平均每个菌株特有基因数目为 34)[16]。菌株特有基因通常是细菌在进化过程中新产生的基因，对于原核生物来说基因横向转移(HGT)是细菌获取新基因的主要动力[17~19]。从泛基因组学角度分析，随着被测序的基因组数目的增加，一些之前被鉴定为菌株特有的基因可能会在新测序的基因组中找到同源基因，因而这些菌株特有的基因在基因组数目增加之后，将可能会被重新定义为非必须基因，因此随着基因组的增加菌株特有基因的数目会逐渐减

少。在以上的比较中发现，即使大肠杆菌的菌株基因组数目达到 30，平均每个菌株中特有基因数目远大于其他两种病原菌。由此推断，与 *H.influenza* 和 *S.pneumoniae* 相比，*E.coli* 中基因横向转移的频率更高。

根据图 1 所示，非必须基因中(CV 值为 2~29)CV 值从 4 到 28 的基因类数目比较少，为了避免某些 CV 值下的基因类数目太少而在分析中产生偏差,我们根据 CV 值的高低将非必须基因类等分成 4 组，每组由连续的 7 个 CV 值组成，即 2~8、9~15、16~22、23~29 各为一组。另外,菌株特有基因类(CV 值为 1)和核心基因类(CV 为 30)各为一组，由此我们将这 30 个菌株的 CV 值分成 6 组，分别为 1、2~8、9~15、16~22、23~29 以及 30。

分别统计 6 个分组中基因类的 COG 分类，比较分析了不同保守程度大肠杆菌基因类的功能富集分布(图 2)。由图 2 可以观察到，随着基因保守性增加，代谢功能相关的基因数目占的比例逐渐升高，而细胞内的过程与信号传递相关基因的比例逐渐降低，遗传信息存储与处理相关的基因数目比例变化不大。在菌株特异性基因中，除去分类功能未知的基因外，细胞内的过程与信号传递相关基因占主要部分约 30%，其代谢功能相关的基因占 20%。

2.2 大肠杆菌的泛基因组特征

为了进一步分析大肠杆菌群体中基因类数目与

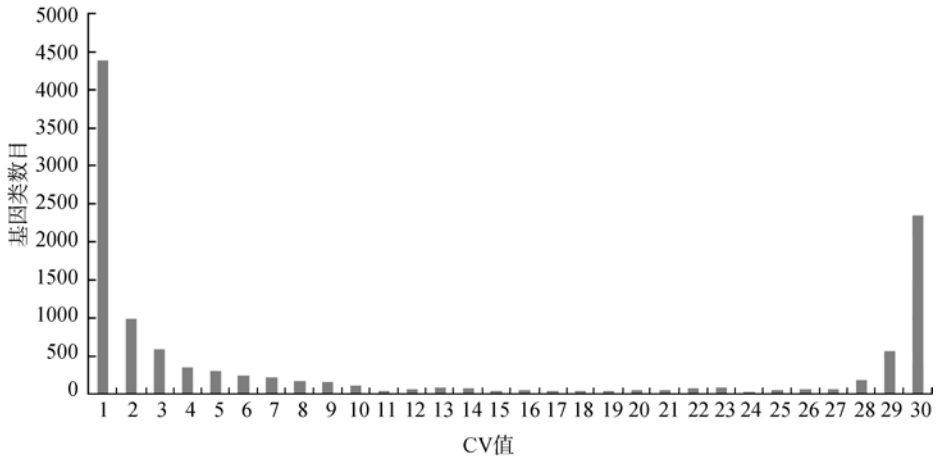

图 1 不同 CV 值的同源基因类数目分布

基因类的 CV 值由该基因类中基因的保守性决定。某个基因类中的基因在 n 个菌株中存在，则该基因类的 CV 值为 n，由于实验中总共利用了 30 个菌株的数据，因此 CV 值有 1 到 30，共 30 个等级。图中 CV 值为 30 的基因是核心基因，CV 值为 2～29 的基因是非必须基因，CV 值为 1 则表示该基因是菌株特有基因。

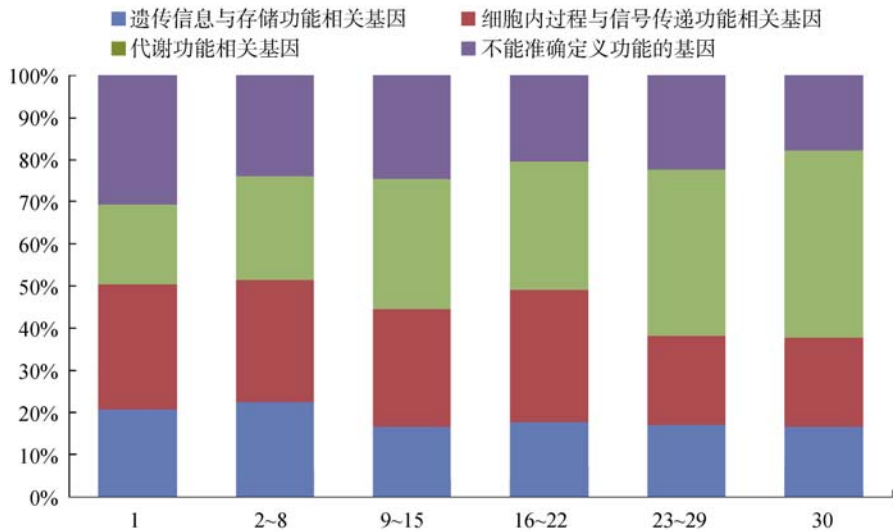

图 2 不同保守程度基因的 COG 功能分类

基因组数目之间的关系，本文分别计算基因组数目为 n(1...30)时存在的基因类数目。为了避免抽样引起的误差，在抽样时采取遍历所有组合的方式，即当 $n=n_0$ 的时候，分别计算 $C(30,n_0)$种 n_0 个菌株组合的泛基因组大小和核心基因数目，再计算平均值。

利用 30 株大肠杆菌中已经鉴定的全部基因类，我们计算得到了大肠杆菌泛基因组大小、核心基因数目与菌株数目的关系(图 3)。同时基于Heaps定律[20~23]，拟合可以得到泛基因组大小(p)与基因组数目(n)的函数关系如下：

$$P=2022.5n^{0.447}+2334.5\ (R^2=0.999967)$$

从泛基因组曲线特征观察，随着测序的基因组数目不断增加，泛基因组大小基本呈现线性增加，故由此可以推断大肠杆菌具有开放型的泛基因组。根据其他病原菌的泛基因组研究报道显示，*Bacillus cereus* (*B.cereus*)和*S.pneumoniae*也具有开放型的泛基因组，而*Bacillus anthracis* (*B.anthrac*)和*Ureaplasma urealyticum* (*U.urealyticum*)等具有闭合型的泛基因组[21]。不同的泛基因组特征一方面可以体现不同细菌的生活环境之间的差异性，另一方面也反映了不同菌株的基因得失差异性。同时从大肠杆菌泛基因组拟合曲线可以推断，每当一个新的大肠杆菌基因组序列被测序，将有 150 个左右的新基因被发现。在原核基因组进化过程中，基因横向迁移是细菌进化的主要动力[24~26]，由此进一步暗示大肠杆菌从外界获取基因的能力非常强。

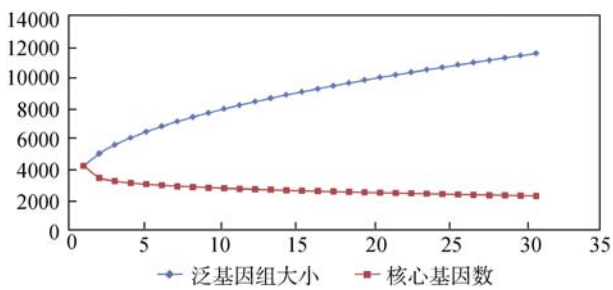

图 3　大肠杆菌的泛基因组特征拟合曲线

横坐标为大肠杆菌菌株数目，纵坐标为基因类数目。图中蓝色菱形块构成的曲线(上面的曲线)表示，随着分析的大肠杆菌菌株数目的增加，大肠杆菌的泛基因组大小(大肠杆菌群体内所有的基因类数目)变化趋势；核红色方块连接而成的曲线(下面的曲线)表示，随着分析的大肠杆菌菌株数目的增加，大肠杆菌的核心基因数目(在所有菌株中都存在基因)的变化趋势。

同时通过对核心基因数目(c)与基因组数目(n)的关系进行拟合可以得到如下函数关系：

$$c=1776.7e^{-0.172n}+2459.7\ (R^2=0.93)$$

由图 3 中核心基因数目的变化曲线可以观察到，随着基因组数目的增加，当菌株数目达到 20 后，核心基因数目趋于稳定。根据拟合的方程可以推断，当大肠杆菌菌株基因组数目达到 40 时，核心基因数目基本稳定在 2 460 左右，即所有大肠杆菌共同包含的基因数目约为 2 460 左右。

2.3　同源基因类的保守性与基因 GC 含量的关系

将基因按照保守性差异分成 6 组，分别计算每一组内基因的 GC 含量的分布(图 4)。由图 4 可以看出，随着同源基因类的保守性的增加，GC 含量变化范围越来越窄，而且 GC 含量的中位数也逐渐升高。由于低保守性的基因一般都是通过基因横向转移获得，其基因来源的差异较大，因此基因的 GC 含量变化也较大，而相对保守的基因在进化过程中由于选择压力的作用，GC 含量逐渐变得越来越集中，更趋向于全基因组的 GC 含量。

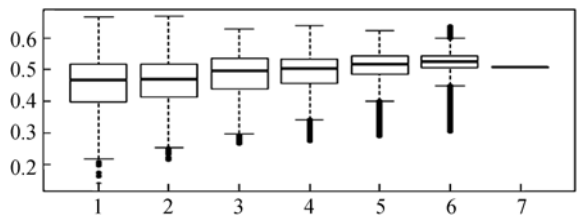

图 4　GC 含量与同源基因类的保守性的关系

横坐标中的 1、2、3、4、5、6 分别对应 CV 值分组为 1、(2~8)、(9~15)、(16~22)、(23~29)、30, 7 为 30 个菌株全基因组的 GC 含量；纵坐标为 GC 含量值。图中 1~6 分别是各组内基因 GC 含量分布的箱线图, 7 是 30 株细菌全基因组 GC 含量分布的箱线图。每个箱线图中间矩形盒上下两条边分别代表该组 GC 含量值的第 1 四分位数(Q1)和第 3 四分位数(Q3)，中间的粗实线代表该组 GC 含量值的中位数，箱线图最上方的横线代表该组 GC 含量值有效数据的上限，数值大小为 Q1+1.5(Q1–Q3)，箱线图最下方的横线代表该组数据有效数据的下限，数值大小为 Q3–1.5(Q1–Q3)，其他实心点表示奇异数值(outlier)。

2.4　同源基因类的保守性与同源基因类的选择压力的关系

在进化上，dn/ds<1 意味着这个基因受到强的纯化选择作用(稳定性选择)，而dn/ds>1 意味着这个基因受到强的多样性选择作用(正向选择)[27]。因此通

过比对同源基因间序列而得到的dn/ds比值，体现了作用在这个基因上的选择压力类型，dn/ds比值在理论上和实际中都具有重要意义[28]。近年来，dn/ds被广泛地应用于群体遗传学、蛋白编码序列以及基因组进化方面的研究[29]。本研究比对了每一个基因类的核酸序列，计算每一个基因类的dn/ds，结合基因类在菌株中分布的广度，来分析二者的关系。

在统计每个同源基因类的dn/ds过程，由于每个CV值为 1 的基因类仅有一个基因而无法进行序列比对，因此这一部分不做统计分析。同时对于一些ds=0的基因类，由于dn/ds值无参考价值，因此这部分的数据也被排除。最终共提取 5 743 个同源基因类用于选择压力分析。通过关联同源基因类的dn/ds的值，可以发现同源基因类的dn/ds分布与其保守性关系(图 5)。由图 5 可以发现：越保守的同源基因类，其dn/ds的值越小。在进化中，一般认为dn/ds值越小，说明同源基因类受到的选择压力越大[30,31]，由此可以推断在大肠杆菌中，核心基因相对于非必须基因，在进化过程中受到更大的选择压力，序列更加保守。

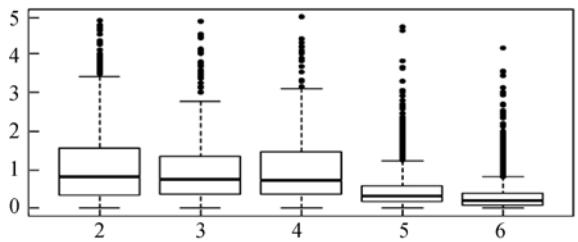

图 5 同源基因类的 **dn/ds** 与保守性的关系

横坐标中的 2、3、4、5、6 分别对应 CV 值分组为(2~8)、(9~15)、(16~22)、(23~29)、30，纵坐标为基因类的 dn/ds 值，图中显示的是不同分组内基因类的 dn/ds 值的分布的箱线图。每个箱线图中间矩形盒上下两条边分别代表该组 dn/ds 值的第 1 四分位数(Q1)和第 3 四分位数(Q3)，中间的粗实线代表该组 dn/ds 值的中位数，箱线图最上方的横线代表该组 dn/ds 值有效数据的上限，数值大小为 Q1+1.5(Q1-Q3)，箱线图最下方的横线代表该组数据有效数据的下限，数值大小为 Q3-1.5(Q1-Q3),其他实心点表示奇异数值(outlier)。

3 讨 论

传统的比较基因组学主要侧重于具体细菌和菌株之间基因组结构、基因组成或特定序列上的差异[32,33]。作为比较基因组学的一个分支，泛基因组学概念自从 2005 年被提出，近年来已被广泛应用到各种病原菌比较基因组学研究中，如*S.agalactiae*[4]、*H.influenza*[15]和*S. pneumoniae*[16]。泛基因组学，从群体角度将细菌中所有的基因分为核心基因、非必须基因以及菌株特有基因，为了解细菌中基因组成、特征以及某特定细菌群体内基因数量的动态变化研究提供了一个新的视角[2,4,34]。

本文首次从泛基因组学角度对大肠杆菌基因组进行大规模的比较分析，研究了大肠杆菌群体的基因组成及基因组分特征。通过比较分析 30 株大肠杆菌基因组，结果发现大肠杆菌中约 49.7%的基因为核心基因。Thomas等[35]利用基因芯片对 47 株*Streptococcus thermophilus*进行比较基因组杂交分析，结果发现*S.thermophilus*中 58%的基因为核心基因。同时，Justin等[15]通过对 13 株*H.influenza*基因组进行比较分析，结果发现平均每个菌株中 79.9%的基因为核心基因。对 17 株*S.pneumoniae*基因组的研究结果显示[16]，平均每个*S.pneumoniae*菌株基因组中约 73%的基因为核心基因。由此可见,大肠杆菌基因组中保守的基因在其全部基因中的比例非常小，体现了大肠杆菌基因组成上丰富的多样性。同时，比较这些菌株中的特有基因发现，大肠杆菌中平均每个菌株里面大约有 146 个菌株特有基因，而*S.pneumoniae*中每个菌株的特有基因数目为 34，*H.influenza*中每个菌株特有基因数目为 39。由于细菌的泛基因组由三部分组成(核心基因、非必须基因、菌株特有基因)，其在定义和分类上具有一定的相对性。随着新的菌株基因组被测序，一些已经被鉴定为菌株特有的基因在后来测序的基因组中可能会找到同源基因，因而这些菌株特有基因将会被归类为非必须基因。这一趋势意味着随着测序的菌株数目增加，每个菌株中的特有基因数目会逐渐减少。但是比较大肠杆菌与*S.pneumoniae*和*H.influenza*可以发现，虽然大肠杆菌的菌株数目要远大于另外两个细菌，但是其菌株特有基因数目比*S.pneumoniae*和*H.influenza*中的菌株特有基因数目要大的多。菌株特有基因通常是细菌在进化过程中新产生的基因，基因横向转移(HGT)是细菌获取新基因的主要动力[17~19]。大肠杆菌中新基因出现的频率比较高，这些高频出现的新基因体现了大肠杆菌通过基因横向转移从外界获取新基因的能力比较强。之前的一些研究已经多次表明，大肠杆菌基因组具有丰富的多样性[6,36,37]，本文通过泛基因组学角度进一步证实了频繁的基因横向迁移为大肠杆菌基因组的多样性提

供了丰富的来源，大肠杆菌基因组成上的多样性可能为其适应不同的生活环境提供了保障。从大肠杆菌中含有较多的菌株特有基因可以推测，一方面大肠杆菌具有极强的整合外来遗传物质的能力；另一方面大肠杆菌的生存环境中存在大量游离的遗传物质，为大肠杆菌基因的多样性和变异性提供了来源。

较多的菌株特有基因在一定的程度上增加了致病性大肠杆菌的预防和控制难度，泛基因组学为我们提供了控制和预防病原菌引发的疾病的新思路，即可以通过针对大肠杆菌参与代谢或其他生命活动途径的 2460 个核心基因设计有效的药物作用靶点或者疫苗[1,5,38]，起到对所有致病大肠杆菌进行预防和控制的作用。另外，由于生物学家对细菌传统的表型鉴定法和分子遗传学鉴定法存在质疑，通过判断核心基因的存在与否，可从基因组学角度为大肠杆菌的分类鉴定提供新的参考依据[39]。非必须基因的比较研究可以进一步深入了解致病大肠杆菌的致病性、抗生素耐受性以及环境适应性等方面的特征。

同时，通过分析不同保守程度(基因类在细菌群体中的分布广度)的基因类与其GC含量以及选择压力之间的关系，可以发现，核心基因的GC含量的变化范围相对于其他非核心基因要窄。同时，从核心基因和其他保守性差的基因的比较可知，保守性强的基因受到更强的选择压力，这一方面体现了核心基因在大肠杆菌的生存中起到重要作用[2]，这些序列稳定保守的基因更适合作为疫苗制备的候选基因；另一方面则体现选择压力驱使外来基因的GC含量与整个基因组水平趋于一致，GC含量的改变在一定程度上改变基因的功能，使新获得的基因逐步适应在大肠杆菌体内的代谢与调控网络。因此从基因的GC含量角度分析大肠杆菌，可以使我们从一个新的角度去了解大肠杆菌基因的动态变化及其生命活动。

参考文献**(References)**:

[1] Muzzi A, Masignani V, Rappuoli R. The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. *Drug Discov Today*, 2007, 12(11–12): 429–439.DOI

[2] Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. *Curr Opin Genet Dev*, 2005, 15(6): 589–594. DOI

[3] Lefebure T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol*, 2007, 8(5): R71. DOI

[4] Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarity Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou LW, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial pan-genome. *Proc Natl Acad Sci*, 2005, 102(39): 13950–13955. DOI

[5] Bambini S, Rappuoli R. The use of genomics in microbial vaccine development. *Drug Discov Today*, 2009, 14(5-6): 252–260. DOI

[6] Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol*, 2008, 190(20): 6881–6893. DOI

[7] Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EP, Denamur E. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet*, 2009, 5(1): 1000344. DOI

[8] Davids W, Zhang ZL. The impact of horizontal gene transfer in shaping operons and protein interaction networks--direct evidence of preferential attachment. *BMC Evol Biol*, 2008, 8: 23. DOI

[9] Zhang R, Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res*, 2009, 37(1): 455–458. DOI

[10] Remm M, Storm CEV, Sonnhammer ELL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 2001, 314(5): 1041–1052. DOI

[11] Enright AJ, Dongen SV, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 2002, 30(7): 1575–1584. DOI

[12] Shi GQ, Zhang LQ, Jiang T. MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome

rearrangement. *BMC Bioinformatics*, 2010, 11: 10. DOI

[13] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, 2000, 28(1): 33–36. DOI

[14] Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, 2010, 26(15): 1899–1900. DOI

[15] Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol*, 2007, 8(6): R103. DOI

[16] Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J, Barbadora K, Klimke W, Dernovoy D, Tatusova T, Parkhill J, Bentley SD, Post JC, Ehrlich GD, Hu FZ. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J Bacteriol*, 2007, 189(22): 8186–8195. DOI

[17] Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol*, 2005, 3: 679–687. DOI

[18] Lawrence JG. Horizontal and vertical gene transfer: the life history of pathogens. *Contrib Microbiol*, 2005, 12: 255–271. DOI

[19] Kurland CG, Canback B, Berg OG. Horizontal gene transfer: a critical view. *Proc Natl Acad Sci USA*, 2003, 100(17): 9658–9662. DOI

[20] Heaps HS. Information Retrieval—Computational and Theoretical Aspects. New York, NY: Academic Press, 1978.

[21] Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*, 2008, 11(5): 472–477. DOI

[22] Deng XY, Phillippy AM, Li ZX, Salzberg SL, Zhang W. Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics*, 2010, 11: 500. DOI

[23] Bottacini F, Medini D, Pavesi A, Turroni F, Foroni E, Riley D, Giubellini V, Tettelin H, van Sinderen D, Ventura M. Comparative genomics of the genus *Bifidobacterium*. *Microbiology*, 2010, 156(11): 3243–3254. DOI

[24] Henryk U, Ast JC, Kaeding AJ, Oliver JD, Dunlap PV. Phylogenetic analysis of the incidence of *lux* gene horizontal transfer in *Vibrionaceae*. *J Bacteriol*, 2008, 190(10): 3494–3504. DOI

[25] Dutta C, Pan A. Horizontal gene transfer and bacterial diversity. *J Biosci*, 2002, 27(1): 27–33. DOI

[26] Jain R, Rivera MC, Moore JE, Lake JA. Horizontal gene transfer in microbial genome evolution. *Theor Popul Biol*, 2002, 61(4): 489–495. DOI

[27] Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. *PLoS Genet*, 2008, 4(12): e1000304. DOI

[28] Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. Comparisons of d*N*/d*S* are time dependent for closely related bacterial genomes. *J Theor Biol*, 2006, 239(2): 226–235. DOI

[29] Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, Jiang L, Holtzapple E, Busch, JD, Smith KL, Schupp JM, Solomon D, Keim P, Fraser CM. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science (Washington D C)*, 2002, 296(5575): 2028–2033. DOI

[30] Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol*, 2006, 239(2): 226–235. DOI

[31] Hurst LD. The *Ka*/*Ks* ratio: diagnosing the form of sequence evolution. *Trends Genet*, 2002, 18(9): 486–487. DOI

[32] McClelland M, Sanderson KE, Clifton SW, Latreille P, Porwollik S, Sabo A, Meyer R, Bieri T, Ozersky P, McLellan M, Harkins CR, Wang CY, Nguyen C, Berghoff A, Elliott G, Kohlerg S, Strong C, Du FY, Carter J, Kremizki C, Layman D, Leonard S, Sun H, Fulton L, Nash W, Miner T, Minx P, Delehaunty K, Fronick C, Magrini V, Nhan M, Warren W, Florea L, Spieth J, Wilson RK. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nature Genetics*, 2004, 36(12): 1268– 1274. DOI

[33] Deng W, Liou SR, Plunkett G, III, Mayhew GF, Rose DJ, Burland V, Kodoyianni V, Schwartz DC, Blattner FR. Comparative genomics of *Salmonellaenterica* serovar Typhi strains Ty2 and CT18. *J Bacteriol*, 2003, 185(7): 2330–2337. DOI

[34] Lefébure T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*. *Genome Biol*, 2007, 8(5): R71. DOI

[35] Rasmussen TB, Danielsen M, Valina O, Garrigues C, Johansen E, Pedersen MB. Streptococcus *thermophilus* core genome: Comparative genome hybridization study of 47 strains. *Appl Environ Microbiol*, 2008, 74(15): 4703–4710. DOI

[36] Caugant DA, Levin BR, Selander RK. Genetic diversity and temporal variation in the *E. coli* population of a human host. *Genetics*, 1981, 98(3): 467–490. DOI

[37] Martinez-Medina M, Aldeguer X, Lopez-Siles M, González-Huix F, López-Oliu C, Dahbi G, Blanco JE, Blanco J, Garcia-Gil LJ, Darfeuille-Michaud A. Molecular diversity of *Escherichia coli* in the human gut: new ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease. *Inflamm Bowel Dis*, 2009, 15(6): 872–882. DOI

[38] Pallen MJ, Wren BW. Bacterial pathogenomics. *Nature*,

2007, 449(7164): 835–842. DOI

[39] Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci*, 2006, 361(1475): 1929–1940. DOI