

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2018 年 5 月 17 日 (17.05.2018)



(10) 国际公布号
WO 2018/086045 A1

(51) 国际专利分类号:
G06F 19/20 (2011.01) **G06F 19/22** (2011.01)

(21) 国际申请号: PCT/CN2016/105372

(22) 国际申请日: 2016 年 11 月 10 日 (10.11.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 深圳华大基因研究院(BGI SHENZHEN)
[CN/CN]; 中国广东省深圳市盐田区北山工业
区综合楼, Guangdong 518083 (CN)。

(72) 发明人: 彭也(PENG, Ye); 中国广东省深圳市盐田
区北山工业区综合楼, Guangdong 518083 (CN)。
李俊桦(LL, Junhua); 中国广东省深圳市盐田区北
山工业区综合楼, Guangdong 518083 (CN)。 唐
珊媚(TANG, Shanmei); 中国广东省深圳市盐田
区北山工业区综合楼, Guangdong 518083 (CN)。

张慧(ZHANG, Hui); 中国广东省深圳市盐田区北
山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: 上海一平知识产权代理有限公司
(XU & PARTNERS, LLC.); 中国上海市普陀
区真北路 958 号天地科技广场 1 号楼 106
室, Shanghai 200333 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家
保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,
BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,
CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP,
KE, KG, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU,
LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ,
NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA,
RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST,
SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ,
VC, VN, ZA, ZM, ZW。

(54) Title: METHOD FOR PERFORMING QUANTITATIVE ANALYSIS ON SUBGROUP IN SPECIFIC GROUP

(54) 发明名称: 一种对特定群中的亚群进行定量分析的方法

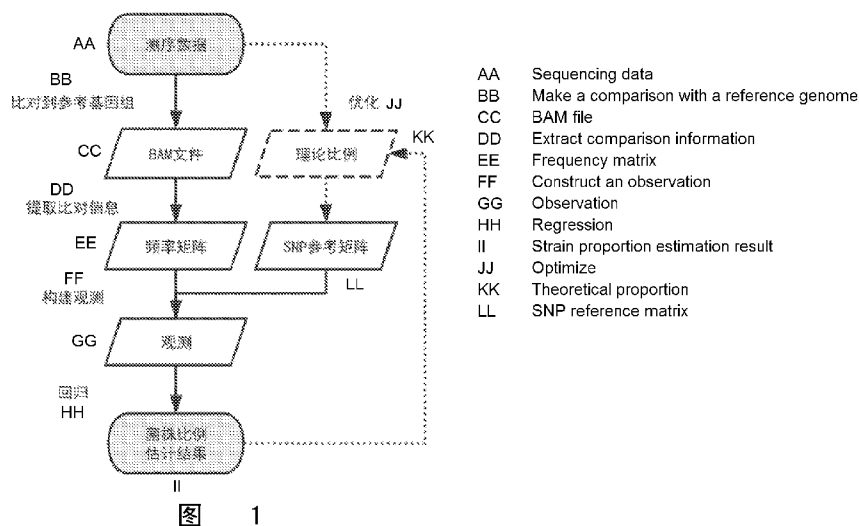


图 1

(57) Abstract: A method for performing quantitative analysis on a subgroup in a specific group. Specifically, the method comprises: (1) providing reference genome sequence data (a), a reference SNP matrix (b) and metagenome sequencing data of a specific group; (2) comparing the metagenome sequencing data with the reference genome sequence data of the specific group, so as to obtain the comparison result; (3) constructing a frequency matrix according to site information of the reference SNP matrix; (4) performing binarization processing on the reference SNP matrix according to the frequency matrix, so as to obtain a binarized SNP matrix; and (5) obtaining a relative abundance of each subgroup in the specific group by means of a restricted linear model and according to the frequency matrix, the binarized SNP matrix, a theoretical base frequency vector (x) and an observation base frequency vector (y), so as to obtain the quantitative analysis result of the subgroup in the specific group.

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种对特定群中的亚群进行定量分析的方法。具体地, 包括以下步骤: (1) 提供对应于所述特定群的 (a) 参考基因组序列数据、(b) 参考SNP矩阵和 (c) 宏基因组测序数据; (2) 将所述宏基因组测序数据比对到对应于所述特定群的参考基因组数据上, 以便获得比对结果; (3) 根据所述参考SNP矩阵的位点信息, 构建频率矩阵; (4) 根据频率矩阵对参考SNP矩阵做二值化处理, 得到二值化SNP矩阵; 和 (5) 基于所述的频率矩阵、所述的二值化SNP矩阵、理论碱基频率向量 $f(x)$ 和观测碱基频率向量 y , 通过有约束的线性模型, 得出所述特定群中各亚群的相对丰度, 从而获得对所述特定群中的亚群的定量分析结果。

一种对特定群中的亚群进行定量分析的方法

技术领域

本发明属于生物信息学领域，更具体地涉及一种对特定群中的亚群进行定量分析的方法。

背景技术

宏基因组学 (metagenomics)，其通过从微生态样品中提取全部微生物的 DNA，构建宏基因组文库，进而利用基因组学的研究策略分析样品中全部微生物的遗传组成及其群落的潜在功能。该组学技术不依赖于单细菌的分离和纯培养技术，在很大程度上解决了大部分微生物因不能分离培养而难于研究的问题，同时也能反映研究生态环境中微生物组成的真实情况。而在关于人类健康的微生态研究中，基于宏基因组学测序数据对微生物进行定量 (quantification)，是研究其群落构成、物种间相互作用，探索其与疾病发生发展的关系等相关规律的基础。随着科学研究的进展，越来越多的研究显示，对特定物种较低级别的分类单元，即菌株 (strains) 进行准确的注释愈发重要。如果简单地在一个较高的分类级别研究细菌与疾病的关系，则很可能把与疾病的发展呈正相关、无关甚至负相关的类别都加到一起，这无论在生物学还是统计学上都有明显谬误；而现有研究成果也亟待通过提高微生物定量精度来校正，或进行更深入的机制研究。

现有宏基因组物种鉴定、定量方法中，分辨率较高，即能注释到菌种或菌株水平的有基于全基因组序列比对和基于标记序列比对的方法，代表的流程分别为 WG-FAST 和 Constrains。其中，WG-FAST 可针对特定菌种，将其菌株与参考基因组进行全基因组序列比对，构建该菌种的 SNP 参考矩阵，进而利用这个矩阵，对宏基因组学样品中该菌种所含菌株进行鉴定；Constrains 则是通过一个菌种定量流程 (MetaPhlAn) 得到菌种丰度和菌种所含菌株的 SNP，然后对每一个样品的 SNP 进行聚类得到每一个样品的 SNP 模式，再对这些 SNP 模式进行聚类，通过模型学则得到最佳的 SNP 模式，根据这个 SNP 模式对每个样品中，该菌种可能含有的菌株的个数和相对丰度做出推断。这两个流程的实现步骤分别如下：

1. WG-FAST

1) 确定需要鉴定菌株的菌种；
2) 通过与参考基因组的序列比对，鉴定出各菌株的 SNP，构建 SNP 参考矩阵；

- 3) 基于实际样品与参考基因组的比对结果，生成该样品的 SNP 矩阵；
5 4) 结果实际样品的 SNP 矩阵和参考 SNP 矩阵，生成系统发生树。

2. Constrains

- 1) 利用 MetaPhlAn 流程进行菌种定量；
2) 通过标记序列比对，鉴定某个菌种内各菌株的 SNP；
3) 对上述 SNP 进行聚类，由聚类结果得到每个样品中的 SNP 模式；
10 4) 对多个样品的 SNP 模式进行聚类，通过模型选择得到最佳的 SNP 模式；
5) 根据最佳的 SNP 模式，推断出每个样品中该菌种各菌株的相对丰度。

现有技术的上述流程中，WG-FAST 虽然能鉴定出特定菌种下的菌株，但不能给出各菌株的丰度(比例)信息，且所依赖的软件和软件库较多，包括
15 BWA-MEM, GATK, Picard-tools, DendroPy, RAxML, BioPython, Trimmomatic 和 SAMtools 等；而 Constrains 只能给出菌种下菌株的一个粗略分类和相应丰度，并不能确定具体菌株信息，并且要求比到菌种水平的读序的覆盖层数在 10 层(10×)以上。

综上所述，本领域尚缺乏基于宏基因组对特定群(如菌群)进行定量分析的方法，现有的方法无法准确测定具体亚群(如菌株)的丰度。此外，即使有些
20 方法即使采用了高时空复杂度的计算，但是仍无法获得的理想的结果。

因此，本领域迫切需要开发能够对特定群中的亚群进行有效定量分析的方法。

25 发明内容

本发明的目的在于提供一种对特定群中的亚群进行有效定量分析的方法。

在本发明的第一方面，提供了一种非诊断性的对特定群中的亚群进行定量分析的方法，包括以下步骤：

- 30 (1) 提供对应于所述特定群的(a)参考基因组序列数据、(b)参考 SNP 矩阵和(c)宏基因组测序数据，其中所述的特定群包括 N 个亚群，N 为 ≥ 2 的正整数；

(2) 将所述宏基因组测序数据比对到对应于所述特定群的参考基因组数据上，以便获得比对结果；

(3) 根据所述参考 SNP 矩阵的位点信息，从所述比对结果中提取出各分析位点上 4 种碱基 A、T、C、和 G 各自的分布频率，构建频率矩阵；

5 (4) 根据频率矩阵对参考 SNP 矩阵做二值化处理，得到二值化 SNP 矩阵，其中在所述二值化处理时，对于某一分析位点而言，若检出的碱基与参考矩阵上的碱基相同则记 1，否则记为 0；并且如果某一分析位点未被覆盖，则该分析位点不参与观测模型的构建；和

10 (5) 基于所述的频率矩阵、所述的二值化 SNP 矩阵、理论碱基频率向量 $f(x)$ 和观测碱基频率向量 y ，通过有约束的线性模型，得出所述特定群中各亚群的相对丰度，从而获得对所述特定群中的亚群的定量分析结果。

在另一优选例中，在步骤(5)中，包括以下步骤：

(5a) 设所述观测碱基频率向量为 y ， y 由所述频率矩阵得到；

15 以及设定理论碱基频率向量为 $f(x)$ ，

$$f(x) = \omega^T x$$

式中， x 为所述二值化 SNP 矩阵；

ω 为列向量，即所述各亚群(1, 2, 3, ...N)的比例值($\omega_1, \omega_2, \omega_3 \dots \omega_N$)构成的列向量 ω ，

20 上标 T 表示转置，行向量转置后为列向量，反之亦然；

(5b) 基于下式 Q1、Q2、Q3、Q4，

$$f(x) \approx y \quad (Q1)$$

$$\omega = (\omega_1, \omega_2, \omega_3 \dots \omega_N)^T \quad (Q2)$$

$$\sum_{i=1}^N \omega_i = 1 \quad (Q3)$$

$$25 \quad 0 \leq \omega_i \leq 1 \quad (Q4)$$

(注：Q3 和 Q4 表示：所有亚群的比例加和为 1，且各亚群比例为 0 到 1 之间的正数)；

运用回归方法，对各亚群比例进行求解，使 ω^* 满足式 Q5 和 Q6，从而得出所述特定群中各亚群的相对丰度，

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (f(x_i) - y_i)^2 \quad (\text{Q5})$$

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (\omega^T x_i - y_i)^2 \quad (\text{Q6})$$

5 各式中，

m 为频率不为零的碱基总个数；

T , x , y , 和 ω 如上定义。

在另一优选例中， N 为 2-50，较佳地 3-20，更佳地 3-10。

10 在另一优选例中，所述的 $y = (q_1, q_2, \dots, q_m)$ ，其中，各 $q_1, q_2, \dots, q_m$ 为频率矩阵中不等于 0 的各分布频率， m 为所述频率矩阵中频率不为零的碱基总个数。

在另一优选例中，所述的特定群为种，而所述的亚群为亚种，或株。

在另一优选例中，在步骤(5b)中，使理论碱基频率($f(x)$)与观测碱基频率(y)的残差平方和最小。

15 在另一优选例中，在步骤(5b)中，使用最小二乘法求解。

在另一优选例中，在步骤(5b)中，使用序贯最小二乘法求解。

在另一优选例中，所述群选自下组：细菌、病毒、肿瘤细胞、真菌、衣原体、支原体。

在另一优选例中，所述亚群选自下组：亚种、异质细胞、菌株。

20 在另一优选例中，所述的宏基因组测序数据的测序深度的下限为 $0.01 \times$ (较佳地 $0.05 \times$ ，更佳地 $0.1 \times$)，而测序深度的上限为至 $10 \times$ (较佳地 $5 \times$ ，更佳地 $2 \times$ ，更佳地 $1 \times$)。

25 在本发明的第二方面，提供了一种非诊断性的对特定群中的亚群进行定量分析的方法，包括以下步骤：

(1) 提供对应于所述特定群的 (a) 参考基因组序列数据或参考转录组序列数据、(b) 参考 SNP 矩阵和 (c) 宏转录组测序数据，其中所述的特定群包括 N 个亚群， N 为 ≥ 2 的正整数；

30 (2) 将所述宏转录组测序数据比对到对应于所述特定群的参考基因组数据或参考转录组序列数据上，以便获得比对结果；

(3) 根据所述参考 SNP 矩阵的位点信息, 从所述比对结果中提取出各分析位点上 4 种碱基 A、T、C、和 G 各自的分布频率, 构建频率矩阵;

(4) 根据频率矩阵对参考 SNP 矩阵做二值化处理, 得到二值化 SNP 矩阵, 其中在所述二值化处理时, 对于某一分析位点而言, 若检出的碱基与参考矩阵上的碱基相同则记 1, 否则记为 0; 并且如果某一分析位点未被覆盖, 则该分析位点不参与观测模型的构建; 和

(5) 基于所述的频率矩阵、所述的二值化 SNP 矩阵、理论碱基频率向量 $f(x)$ 和观测碱基频率向量 y , 通过有约束的线性模型, 得出所述特定群中各亚群的相对丰度, 从而获得对所述特定群中的亚群的定量分析结果。

10 在另一优选例中, 所述的参考 SNP 矩阵包括基于基因组序列的参考 SNP 矩阵、或基于转录组序列的参考 SNP 矩阵。

在另一优选例中, 在步骤 (5) 中, 包括以下步骤:

(5a) 设所述观测碱基频率向量为 y , y 由所述频率矩阵得到;

以及设定理论碱基频率向量为 $f(x)$,

$$15 \quad f(x) = \omega^T x$$

式中, x 为所述二值化 SNP 矩阵;

ω 为列向量, 即所述各亚群 (1, 2, 3, $\dots$ N) 的比例值 ($\omega_1, \omega_2, \omega_3 \dots \omega_N$) 构成的列向量 ω ,

上标 T 表示转置, 行向量转置后为列向量, 反之亦然;

20 (5b) 基于下式 Q1、Q2、Q3、Q4,

$$f(x) \approx y \quad (Q1)$$

$$\omega = (\omega_1, \omega_2, \omega_3 \dots \omega_N)^T \quad (Q2)$$

$$\sum_{i=1}^N \omega_i = 1 \quad (Q3)$$

$$0 \leq \omega_i \leq 1 \quad (Q4)$$

25 (注: Q3 和 Q4 表示: 所有亚群的比例加和为 1, 且各亚群比例为 0 到 1 之间的正数);

运用回归方法, 对各亚群比例进行求解, 使 ω^* 满足式 Q5 和 Q6, 从而得出所述特定群中各亚群的相对丰度,

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (f(x_i) - y_i)^2 \quad (Q5)$$

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (\omega^T x_i - y_i)^2 \quad (\text{Q6})$$

各式中，

m 为频率不为零的碱基总个数；

5 T , x , y , 和 ω 如上定义。

在另一优选例中， N 为 2-50，较佳地 3-20，更佳地 3-10。

在另一优选例中，所述的 $y = (q_1, q_2, \dots, q_m)$ ，其中，各 $q_1, q_2, \dots, q_m$ 为频率矩阵中不等于 0 的各分布频率， m 为所述频率矩阵中频率不为零的碱基总个数。

10 在另一优选例中，所述的特定群为种，而所述的亚群为亚种，或株。

在另一优选例中，在步骤(5b)中，使理论碱基频率($f^{(x)}$)与观测碱基频率(y)的残差平方和最小。

在另一优选例中，在步骤(5b)中，使用最小二乘法求解。

在另一优选例中，在步骤(5b)中，使用序贯最小二乘法求解。

15 在另一优选例中，所述群选自下组：细菌、病毒、肿瘤细胞、真菌、衣原体、支原体。

在另一优选例中，所述亚群选自下组：亚种、异质细胞、菌株。

20 在另一优选例中，所述的宏转录组测序数据的测序深度的下限为 $0.01 \times$ (较佳地 $0.05 \times$ ，更佳地 $0.1 \times$)，而测序深度的上限为至 $10 \times$ (较佳地 $5 \times$ ，更佳地 $2 \times$ ，更佳地 $1 \times$)。

应理解，在本发明范围内中，本发明的上述各技术特征和在下文(如实施例)中具体描述的各技术特征之间都可以互相组合，从而构成新的或优选的技术方案。限于篇幅，在此不再一一累述。

25

附图说明

图 1 显示了对特定菌种内的菌株进行定量的流程图，其中虚线部分为流程构建过程特有数据或步骤。

30 图 2 显示了流程对模拟简单环境样品处理结果的残差，其中各测序深度下有三个平行实验。

图 3 显示了流程对模拟简单环境样品处理结果的正误比例，其中各测序深度下有三个平行实验。

图 4 显示了流程对模拟复杂环境样品处理结果的残差，其中各测序深度下有三个平行实验。

5 图 5 显示了流程对模拟复杂环境样品处理结果的正误比例，其中各测序深度下有三个平行实验。

具体实施方式

10 本发明人经过广泛而深入地研究，首次开发了一种对特定群中的亚群进行有效定量分析的方法。本发明基于参考基因组或转录组序列数据、参考 SNP 矩阵和宏基因组或转录组测序数据，通过对各数据(库)的比对构建频率矩阵，并通过二值化处理、通过理论碱基频率向量 $f(x)$ 和观测碱基频率向量 y 构建有约束的线性模型，从而有效地得出特定群中各亚群的定量检测结果。在此基础上，完成了本发明。

15 具体地，以宏基因组学为例，本发明基于特定群的参考基因组序列数据、参考 SNP 矩阵和宏基因组测序数据，将宏基因组测序数据与参考基因组数据进行比对，获取比对结果；根据上述参考 SNP 矩阵的位点信息，从所述比对结果中提取出各分析位点上 4 种碱基 A、T、C、和 G 各自的分布频率，构建频率矩阵；根据频率矩阵对参考 SNP 矩阵做二值化处理，得到二值化 SNP 矩阵；基于
20 上述频率矩阵、二值化 SNP 矩阵、理论碱基频率向量 $f(x)$ 和观测碱基频率向量 y ，通过有约束的线性模型，得出所述特定群中各亚群的相对丰度，从而获得对所述特定群中的亚群的定量分析结果。

具体地，本发明人还提供了一种在待研究菌种测序层数(基因组被测序读序覆盖的次数)较少的情况下，实现对包含于参考矩阵的菌株的比例的定量方
25 法；构建合适的模型，使流程实现过程的时空复杂度较小。

对亚群进行定量分析的方法

以下结合一个具体例子进一步描述本发明的技术方案。在该实施例中，先提供：1)特定菌种的参考基因组，2)该菌种的参考 SNP 矩阵，3)宏基因组测序
30 数据。

流程处理过程包括：1)使用序列比对软件，将宏基因组测序数据比对到参

考基因组上，得到比对结果；2)根据参考 SNP 矩阵的位点信息，从上述比对结果中提取出各位点上 4 种碱基(A、T、C、G)的频率，构建频率矩阵；3)结合参考 SNP 矩阵和频率矩阵，构建观测模型；4)运用回归方法，对各菌株比例进行估计。该流程如附图 1 所示。

- 5 假设某个菌种下有 3 个菌株，其中菌株 I 的基因组为该菌种的参考基因组(reference, ref)，其参考 SNP 矩阵包括 n 个位点，该矩阵示例如下：

表 1 SNP 矩阵示例

矩阵位点序号 菌株	1	2	3	...	n
I (ref)	C	G	A	...	A
II	T	G	G	...	T
III	C	A	G	...	C

流程各步骤的详细情况介绍如下：

- 10 1. 序列比对及结果提取

使用序列比对软件，将宏基因组测序数据比对到参考基因组上，经过碱基质量控制后得到包含比对信息的文件。

2. 频率矩阵构建

- 15 根据参考 SNP 矩阵上的位点信息，从上述 pileup 文件中提取出各个位点上 4 种碱基的频率，若某位点未被覆盖，则将 4 种碱基的频率均设为 0(如位点 2 的情况)。得到如下所示的频率矩阵。

表 2 频率矩阵示例

矩阵位点序号 碱基类型	1	2	3	...	n
A	0	0	0.1	...	0.3
T	0.5	0	0	...	0.3
C	0.5	0	0	...	0.4
G	0	0	0.9	...	0

3. 模型构建

- 20 a)根据频率矩阵对参考 SNP 矩阵做二值化处理，即若某一位置上检出的碱基与参考矩阵上的碱基相同则记 1，否则记为 0；若某位点未被覆盖，则该位

点不参与观测的构建(如位点 2 的情况)。二值化的 SNP 矩阵如下表所示。

表 3 二值化 SNP 矩阵示例

矩阵位点序号 菌株	1		3		...	n		
	T	C	A	G	...	A	T	C
I (ref)	0	1	1	0	...	1	0	0
II	1	0	0	1	...	0	1	0
III	0	1	0	1	...	0	0	1

为了求解菌株在菌种中的比例，现构建有约束的线性模型。

- 5 b) 设观测碱基频率向量(由上述频率矩阵得到)为 y ，理论碱基频率向量(经计算得到)为 $f(x)$ ，二值化 SNP 矩阵为 x ，待估计菌株(1, 2, 3)的比例值($\omega_1, \omega_2, \omega_3$)构成的列向量为 ω ，其中，

$$f(x) = \omega^T x$$

上标 T 表示转置，列向量转置后为行向量，反之亦然。

10 使得，

$$f(x) \approx y$$

其中，

$$\omega^T = (\omega_1, \omega_2, \omega_3)$$

$$\sum_{i=1}^3 \omega_i = 1$$

15 $0 \leq \omega_i \leq 1$

即所有菌株的比例加和为 1，且各菌株比例为 0 到 1 之间的正数。

在本例中，

$$y = (0.5, 0.5, 0.1, 0.9, \dots, 0.3, 0.3, 0.4)$$

$$\omega^T = (\omega_1, \omega_2, \omega_3)$$

20

$$x = \begin{pmatrix} 0110 & \dots & 100 \\ 1001 & \dots & 010 \\ 0101 & \dots & 001 \end{pmatrix}$$

4. 回归求解

如前所述，估计菌株比例的问题，即解有约束的优化问题，亦即求得 ω^* ，使

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (f(x_i) - y_i)^2$$

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (\omega^T x_i - y_i)^2$$

其中，m 为频率不为零的碱基总个数（即表 3-二值化 SNP 矩阵中碱基的总个数）。

- 5 即，使理论碱基频率 ($f(x)$) 与观测碱基频率 (y) 的残差平方和最小。发明人使用序贯最小二乘法求解，并将比例较小的估计结果过滤掉，即认为此类菌株并不存在于样品中。

假设上述示例中仅第 1、3、n 个位点有测序序列覆盖，即

$$y = (0.5, 0.5, 0.1, 0.9, 0.3, 0.3, 0.4),$$

10

$$x = \begin{pmatrix} 0110100 \\ 1001010 \\ 0101001 \end{pmatrix},$$

则，

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^7 (f(x_i) - y_i)^2$$

$$= \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^7 (\omega^T x_i - y_i)^2$$

15

$$\begin{aligned} &= \underset{\omega}{\operatorname{argmin}} \{[(\omega_1, \omega_2, \omega_3)(010) - 0.5]^2 \\ &\quad + [(\omega_1, \omega_2, \omega_3)(101) - 0.5]^2 \\ &\quad + [(\omega_1, \omega_2, \omega_3)(100) - 0.1]^2 \\ &\quad + [(\omega_1, \omega_2, \omega_3)(011) - 0.9]^2 \\ &\quad + [(\omega_1, \omega_2, \omega_3)(100) - 0.3]^2 \\ &\quad + [(\omega_1, \omega_2, \omega_3)(010) - 0.3]^2 \\ &\quad + [(\omega_1, \omega_2, \omega_3)(001) - 0.4]^2\} \end{aligned}$$

$$\begin{aligned}
&= \operatorname{argmin}_{\omega} \{ (\omega_2 - 0.5)^2 \\
&\quad + (\omega_1 + \omega_3 - 0.5)^2 \\
&\quad + (\omega_1 - 0.1)^2 \\
&\quad + (\omega_2 + \omega_3 - 0.9)^2 \\
&\quad + (\omega_1 - 0.3)^2 \\
&\quad + (\omega_2 - 0.3)^2 \\
&\quad + (\omega_3 - 0.4)^2 \}
\end{aligned}$$

经上述回归方法求解，可得

$$\omega^* = (0.2427, 0.3859, 0.3714).$$

5

得到特定物种下菌株的比例后，结合其他流程得到的菌种定量结果，即可得到各菌株在宏基因组样品中的相对丰度。设特定菌种的菌种丰度为 a ，各菌株相对丰度为向量 φ ，即有，

$$\varphi = \omega^* \times a.$$

10

基于宏转录组测序数据对亚群进行定量分析的方法，与上述基于宏基因组测序数据的方法基本类似，不同点主要在于，用宏转录组测序数据替换宏基因组测序数据。此外，由于转录组数据与基因组数据存在映射关系，因此比对时可以采用参考基因组序列数据，也可以参考转录组序列数据。在本发明方法中，参考 SNP 矩阵可以基于基因组序列，也可以基于转录组序列。

15

本发明的主要优点包括：

(a) 本发明可对宏基因组数据中特定菌种的菌株间的比例进行定量。例如用于对某个粪便样品中所含大肠杆菌菌株的比例进行定量，结合现有流程的菌种定量结果得到菌株的丰度，这将有利于提高相关研究的分辨率或疾病诊断、治疗的精准性。

20

(b) 本发明开发了一套全新的、可自定义目标的菌株定量流程，能够帮助有需求的生物研究人员快速而准确地对宏基因组样品中同一菌种下的菌株进行定量。

(c) 本发明提供的分析策略能对序列相似性较高的群组进行细分，因此不

仅可用于多种复杂微生物群落的宏基因组学研究流程，也可拓展至宏转录组学等相关组学研究，甚至可用于探究肿瘤异质性方面的科学问题，这将推动精准医学的发展。

下面结合具体实施例，进一步阐述本发明。应理解，这些实施例仅用于说明本发明而并不用于限制本发明的范围。下列实施例中未注明具体条件的实验方法，通常按照常规条件如 Sambrook 等人，分子克隆：实验室手册 (New York: Cold Spring Harbor Laboratory Press, 1989) 中所述的条件，或按照制造厂商所建议的条件。除非另外说明，否则百分比和份数按重量计算。本发明中所涉及的实验材料如无特殊说明均可从市售渠道获得。

实施例 1 模拟简单环境样品的分析

1. 数据模拟

为了模拟简单环境样品，即仅含同一菌种的不同菌株的样品，本发明人选取基因组资源丰富的大肠杆菌作为研究对象，从 NCBI 基因组数据库中选取了 44 个代表菌株(2015 年 11 月 19 日版本)用于构建参考矩阵，其中大肠杆菌 K-12 的基因组为参考基因组(reference, ref)；随后从中挑选了 5 个菌株，用于模拟简单环境样品测序数据，各菌株信息及其在复杂环境样品中的理论丰度表 4 所示。

20 表 4 模拟简单样品详细信息

菌株名称	所占比例
<i>Escherichia coli</i> O103:H2 str. 12009	0.2
<i>Escherichia coli</i> O104:H4 str. 2011C-3493	0.2
<i>Escherichia coli</i> O127:H6 str. E2348/69	0.2
<i>Escherichia coli</i> O145:H28 str. 4865/96	0.2
<i>Escherichia coli</i> O157:H7 str. Sakai	0.2

本发明人模拟了一系列的测序数据，各系列中，上述 5 个菌株的模拟的测序深度(层数，×)分别为 0.01×、0.1×、和 1×，各个系列均平行生成 3 次。模拟的单端测序片段的长度为 50bp，插入片段的大小为 200bp，测序错误率为 0。

2. 序列比对及结果提取

2.1 本发明人选择大肠杆菌 K-12 菌株的基因组作为参考序列, 使用 BWA 序列比对软件, 将宏基因组测序数据比对到参考基因组上, 得到包含原始比对结果的 BAM 文件;

5 2.2 使用 samtools 工具, 将 BAM 文件按位点排序后, 把每个位点上比对质量值大于 10, 即比对正确率大于 90% 的碱基输出, 得到包含排序、过滤处理后的比对信息的 pileup 文件;

3. 频率矩阵构建

10 根据参考 SNP 矩阵上的位点信息, 从上述 pileup 文件中提取出各个位点上 4 种碱基的频率, 若某位点未被覆盖, 则将 4 种碱基的频率均设为 0, 得到频率矩阵。

4. 模型构建

15 4.1 根据频率矩阵对参考 SNP 矩阵做二值化处理, 即若某一位置上检出的碱基与参考矩阵上的碱基相同则记 1, 否则记为 0; 若某位点未被覆盖, 则该位点不参与观测的构建。

4.2 为了求解菌株在菌种中的比例, 现构建有约束的线性模型。设碱基频率向量为 $f(x)$, 二值化 SNP 矩阵为 x , 待估计比例值构成的向量为 ω , 三者满足,

$$f(x) = \omega^T \times x$$

20 其中,

$$\omega = (\omega_1, \omega_2, \omega_3, \dots, \omega_n)$$

其中,

$$\sum_{i=1}^N \omega_i = 1$$

$$0 \leq \omega_i \leq 1$$

25 即所有菌株(在本例中 N 为 44 株)的比例加和为 1, 且各菌株比例为 0 到 1 之间的正数。

5. 回归求解

30 如前所述, 估计菌株比例的问题, 即解有约束的优化问题, 亦即求得 ω^* , 使,

$$\begin{aligned}\omega^* &= \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (f(x_i) - y_i)^2 \\ &= \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (\omega^T \times x_i - y_i)^2\end{aligned}$$

发明人使用序贯最小二乘法求解，并将比例小于 0.0001 的估计结果过滤掉，结果如表 5-1、表 5-2、表 5-3 及附图 2 所示。

5

6. 流程估计结果分析

对流程估计结果的分析显示，各测序深度下的估计值与理论值间的残差分布随测序深度的增加逐渐收敛，在 0.01×、0.1×及 1×的深度下，平均残差平方和分别为 0.0578、0.0119 和 0.0048，模拟数据所选菌株的平均总比例分别为 0.7580、0.8496 和 0.9235。另一方面，随着测序深度的增加，各系列平行实验的估计值更加接近，即精确度更高，结果如表 5-1、表 5-2、表 5-3、表 6 及附图 2、附图 3 所示。

表 5-1 模拟简单环境样品菌株估计结果

菌株名称	理论值	0.01×		
		1	2	3
<i>Escherichia coli</i> K-12 (ref)	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 102606_aEPEC	0.0000	0.0000	0.0000	0.0088
<i>Escherichia coli</i> 1-182-04_S3_C2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 1-250-04_S3_C1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 138B_L1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 207A_L2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 2-222-05_S3_C2	0.0000	0.1368	0.0000	0.0000
<i>Escherichia coli</i> 400791	0.0000	0.0000	0.0688	0.0000
<i>Escherichia coli</i> 401140	0.0000	0.0674	0.0173	0.0000
<i>Escherichia coli</i> 402403	0.0000	0.0000	0.0498	0.0000
<i>Escherichia coli</i> 48	0.0000	0.0100	0.0000	0.0000
<i>Escherichia coli</i> 53638	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 541-15	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 80B_L1	0.0000	0.0000	0.0000	0.0022
<i>Escherichia coli</i> B185	0.0000	0.0000	0.0000	0.0104
<i>Escherichia coli</i> B354	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> B7C5	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> B str. REL606	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> CFT073	0.0000	0.0000	0.0364	0.0117
<i>Escherichia coli</i> ED1a	0.0000	0.0000	0.0000	0.0000

<i>Escherichia coli</i> H299	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> HS	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> IAI39	0.0000	0.0000	0.0000	0.0464
<i>Escherichia coli</i> KTE102	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> KTE156	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> KTE33	0.0000	0.0000	0.0028	0.0000
<i>Escherichia coli</i> M863	0.0000	0.0000	0.0000	0.0246
<i>Escherichia coli</i> MS 69-1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> MS 84-1	0.0000	0.0341	0.0000	0.0000
<i>Escherichia coli</i> NA114	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O103:H2 str. 12009	0.2000	0.2566	0.2369	0.3020
<i>Escherichia coli</i> O104:H4 str. 2011C-3493	0.2000	0.0380	0.0279	0.0210
<i>Escherichia coli</i> O127:H6 str. E2348/69	0.2000	0.1632	0.1612	0.1428
<i>Escherichia coli</i> O145:H28 str. 4865/96	0.2000	0.1112	0.1402	0.2129
<i>Escherichia coli</i> O157:H7 str. Sakai	0.2000	0.1618	0.1355	0.1629
<i>Escherichia coli</i> O83:H1 str. NRG 857C	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> S6400	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> SCD2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> SE11	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> UMEA 3318-1	0.0000	0.0000	0.0743	0.0197
<i>Escherichia coli</i> UMN026	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> upec-182	0.0000	0.0208	0.0000	0.0225
<i>Escherichia coli</i> upec-55	0.0000	0.0000	0.0492	0.0122
<i>Escherichia coli</i> upec-64	0.0000	0.0000	0.0000	0.0000

表 5-2 模拟简单环境样品菌株估计结果

菌株名称	理论值	0.1×		
		1	2	3
<i>Escherichia coli</i> K-12 (ref)	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 102606_aEPEC	0.0000	0.0035	0.0129	0.0000
<i>Escherichia coli</i> 1-182-04_S3_C2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 1-250-04_S3_C1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 138B_L1	0.0000	0.0123	0.0000	0.0000
<i>Escherichia coli</i> 207A_L2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 2-222-05_S3_C2	0.0000	0.0000	0.0407	0.0000
<i>Escherichia coli</i> 400791	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 401140	0.0000	0.0000	0.0000	0.0005
<i>Escherichia coli</i> 402403	0.0000	0.0000	0.0000	0.0129
<i>Escherichia coli</i> 48	0.0000	0.0165	0.0000	0.0000
<i>Escherichia coli</i> 53638	0.0000	0.0000	0.0234	0.0000
<i>Escherichia coli</i> 541-15	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 80B_L1	0.0000	0.0000	0.0000	0.0039
<i>Escherichia coli</i> B185	0.0000	0.0000	0.0000	0.0124

<i>Escherichia coli</i> B354	0.0000	0.0000	0.0000	0.0096
<i>Escherichia coli</i> B7C5	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> B str. REL606	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> CFT073	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> ED1a	0.0000	0.0084	0.0166	0.0000
<i>Escherichia coli</i> H299	0.0000	0.0000	0.0356	0.0000
<i>Escherichia coli</i> HS	0.0000	0.0000	0.0000	0.0020
<i>Escherichia coli</i> IAI39	0.0000	0.0174	0.0000	0.0791
<i>Escherichia coli</i> KTE102	0.0000	0.0209	0.0000	0.0000
<i>Escherichia coli</i> KTE156	0.0000	0.0000	0.0474	0.0000
<i>Escherichia coli</i> KTE33	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> M863	0.0000	0.0015	0.0008	0.0000
<i>Escherichia coli</i> MS 69-1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> MS 84-1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> NA114	0.0000	0.0036	0.0039	0.0000
<i>Escherichia coli</i> O103:H2 str. 12009	0.2000	0.1932	0.1516	0.1974
<i>Escherichia coli</i> O104:H4 str. 2011C-3493	0.2000	0.1496	0.1477	0.1532
<i>Escherichia coli</i> O127:H6 str. E2348/69	0.2000	0.1615	0.1393	0.1543
<i>Escherichia coli</i> O145:H28 str. 4865/96	0.2000	0.1856	0.1732	0.1938
<i>Escherichia coli</i> O157:H7 str. Sakai	0.2000	0.2063	0.1700	0.1723
<i>Escherichia coli</i> O83:H1 str. NRG 857C	0.0000	0.0197	0.0000	0.0000
<i>Escherichia coli</i> S6400	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> SCD2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> SE11	0.0000	0.0000	0.0368	0.0000
<i>Escherichia coli</i> UMEA 3318-1	0.0000	0.0000	0.0000	0.0086
<i>Escherichia coli</i> UMN026	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> upec-182	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> upec-55	0.0000	0.0000	0.0001	0.0000
<i>Escherichia coli</i> upec-64	0.0000	0.0000	0.0000	0.0000

表 5-3 模拟简单环境样品菌株估计结果

菌株名称	理论值	1×		
		1	2	3
<i>Escherichia coli</i> K-12 (ref)	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 102606_aEPEC	0.0000	0.0014	0.0000	0.0000
<i>Escherichia coli</i> 1-182-04_S3_C2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 1-250-04_S3_C1	0.0000	0.0000	0.0057	0.0000
<i>Escherichia coli</i> 138B_L1	0.0000	0.0000	0.0004	0.0000
<i>Escherichia coli</i> 207A_L2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 2-222-05_S3_C2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 400791	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 401140	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 402403	0.0000	0.0000	0.0000	0.0000

<i>Escherichia coli</i> 48	0.0000	0.0011	0.0057	0.0000
<i>Escherichia coli</i> 53638	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 541-15	0.0000	0.0000	0.0000	0.0019
<i>Escherichia coli</i> 80B_L1	0.0000	0.0032	0.0000	0.0000
<i>Escherichia coli</i> B185	0.0000	0.0003	0.0000	0.0036
<i>Escherichia coli</i> B354	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> B7C5	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> B str. REL606	0.0000	0.0010	0.0000	0.0000
<i>Escherichia coli</i> CFT073	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> ED1a	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> H299	0.0000	0.0056	0.0000	0.0000
<i>Escherichia coli</i> HS	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> IAI39	0.0000	0.0711	0.0482	0.0516
<i>Escherichia coli</i> KTE102	0.0000	0.0000	0.0000	0.0044
<i>Escherichia coli</i> KTE156	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> KTE33	0.0000	0.0003	0.0000	0.0000
<i>Escherichia coli</i> M863	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> MS 69-1	0.0000	0.0000	0.0004	0.0000
<i>Escherichia coli</i> MS 84-1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> NA114	0.0000	0.0000	0.0000	0.0013
<i>Escherichia coli</i> O103:H2 str. 12009	0.2000	0.1804	0.1839	0.1844
<i>Escherichia coli</i> O104:H4 str. 2011C-3493	0.2000	0.1764	0.1808	0.1859
<i>Escherichia coli</i> O127:H6 str. E2348/69	0.2000	0.1828	0.1896	0.1918
<i>Escherichia coli</i> O145:H28 str. 4865/96	0.2000	0.1907	0.1857	0.1895
<i>Escherichia coli</i> O157:H7 str. Sakai	0.2000	0.1858	0.1832	0.1795
<i>Escherichia coli</i> O83:H1 str. NRG 857C	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> S6400	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> SCD2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> SE11	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> UMEA 3318-1	0.0000	0.0000	0.0164	0.0000
<i>Escherichia coli</i> UMN026	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> upec-182	0.0000	0.0000	0.0000	0.0013
<i>Escherichia coli</i> upec-55	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> upec-64	0.0000	0.0000	0.0000	0.0048

表 6 模拟简单环境样品菌株估计的正误比例

类别	理论值	0.01×			0.1×			1×		
		1	2	3	1	2	3	1	2	3
所选大肠杆菌菌株	1.0000	0.7309	0.7016	0.8415	0.8961	0.7818	0.8710	0.9161	0.9232	0.9312
其他大肠杆菌菌株	0.0000	0.2691	0.2984	0.1585	0.1039	0.2182	0.1290	0.0839	0.0768	0.0688

上述结果表明，对于由 5 个亚群构成的特定群，当测序深度为 0.01×或更高时，本发明方法就可以非常有效地对特定群中的各亚群进行定量分析，并提供可靠的检测结果。其中，当测序深度为 0.01×时，检测值与理论值的误差小于约 30%；当测序深度为 0.1×时，检测值与理论值的误差小于约 20%；当测序深度为 1×时，检测值与理论值的误差小于约 10%。

实施例 2 模拟复杂环境样品的分析

1. 数据模拟

为了模拟复杂环境样品，即包含多个菌种和大肠杆菌多个菌株的样品。本发明人在常见的拟杆菌门、厚壁菌门、变形菌门和放线菌门中选取了 10 个物种，即长双歧杆菌(*Bifidobacterium longum*)、迟缓埃格特菌 (*Eggerthella lenta*)、吉氏副拟杆菌 (*Parabacteroides distasonis*)、活泼瘤胃球菌 (*Ruminococcus gnavus*)、艰难梭状芽胞杆菌(*Peptoclostridium difficile*)、韦荣球菌 HPA0037(*Veillonella* sp. HPA0037)、链球菌 I-P16(*Streptococcus* sp. I-P16)、柯氏柠檬酸杆菌(*Citrobacter koseri*)、埃希氏菌 艾伯替埃希氏菌(*Escherichia albertii*)和大肠杆菌(*Escherichia coli*)，其中大肠杆菌从前述参考矩阵中选取了 5 个菌株，各物种选取的菌株信息及其在复杂环境样品中的理论丰度如下表所示。

表 7 复杂微生物群落详细信息

菌株名称	所占比例
<i>Escherichia coli</i> UMN026	1/14
<i>Escherichia coli</i> UMEA_3318_1	1/14
<i>Escherichia coli</i> NA114	1/14
<i>Escherichia coli</i> IAI39	1/14
<i>Escherichia coli</i> CFT073	1/14
<i>Escherichia albertii</i> KF1	1/14
<i>Citrobacter_koseri</i> ATCC BAA-895	1/14
<i>Streptococcus</i> sp. I-P16	1/14
<i>Veillonella</i> sp. HPA0037	1/14
<i>Peptoclostridium difficile</i> 630	1/14
<i>Ruminococcus gnavus</i> AGR2154	1/14
<i>Parabacteroides distasonis</i> ATCC 8503	1/14
<i>Eggerthella lenta</i> DSM 2243	1/14
<i>Bifidobacterium longum</i> NCC2705	1/14

与前述方法类似，根据表 7 所示的物种构成，生成了系列模拟测序数据，各系列中各菌株测序深度分别为 $0.01\times$ 、 $0.1\times$ 和 $1\times$ ，每个系列平行重复三次。

2. 流程处理过程

本实施例的序列比对及结果提取、频率矩阵构建、模型构建和回归求解过程均与实施例 1 一致。

3. 流程处理结果分析

运行本发明处理后，对结果的分析发现，与前述探究结果类似，随着测序深度的增加，估计值与理论值的残差平方和总体上呈减小趋势，在 $0.01\times$ 、 $0.1\times$ 和 $1\times$ 的深度下，平均残差平方和分别为 0.0744、0.0304 和 0.0278，模拟数据所选菌株的平均总比例分别为 0.5569、0.7797 和 0.7700，且对应的平行实验的结果趋于一致。由于受到大肠杆菌近缘物种测序序列的干扰，该估计效果相比简单环境样品略有下降，但在菌株测序深度达到 $0.1\times$ 及以上时，仍有较高的准确性和精确性，结果如表 8-1、表 8-2、表 8-3、表 9 及附图 4、附图 5 所示。

表 8-1 模拟复杂环境样品菌株估计结果

菌株名称	理论值	$0.01\times$		
		1	2	3
<i>Escherichia coli</i> K-12 (ref)	0.0000	0.0224	0.0703	0.0011
<i>Escherichia coli</i> 102606_aEPEC	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 1-182-04_S3_C2	0.0000	0.0360	0.0169	0.0000
<i>Escherichia coli</i> 1-250-04_S3_C1	0.0000	0.0000	0.0000	0.0402
<i>Escherichia coli</i> 138B_L1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 207A_L2	0.0000	0.0744	0.0019	0.0000
<i>Escherichia coli</i> 2-222-05_S3_C2	0.0000	0.0252	0.0000	0.0014
<i>Escherichia coli</i> 400791	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 401140	0.0000	0.0000	0.0318	0.0190
<i>Escherichia coli</i> 402403	0.0000	0.0000	0.0229	0.0000
<i>Escherichia coli</i> 48	0.0000	0.0000	0.0441	0.0000
<i>Escherichia coli</i> 53638	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 541-15	0.0000	0.0964	0.0361	0.0887
<i>Escherichia coli</i> 80B_L1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> B185	0.0000	0.0000	0.0134	0.0304
<i>Escherichia coli</i> B354	0.0000	0.0320	0.0000	0.0000
<i>Escherichia coli</i> B7C5	0.0000	0.0040	0.0000	0.0310
<i>Escherichia coli</i> B str. REL606	0.2000	0.1724	0.2182	0.1332
<i>Escherichia coli</i> CFT073	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> ED1a	0.0000	0.0212	0.0000	0.0000
<i>Escherichia coli</i> H299	0.0000	0.0000	0.0294	0.0000

<i>Escherichia coli</i> HS	0.2000	0.0952	0.0788	0.0565
<i>Escherichia coli</i> IAI39	0.0000	0.0000	0.0887	0.1202
<i>Escherichia coli</i> KTE102	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> KTE156	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> KTE33	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> M863	0.0000	0.0000	0.0089	0.0000
<i>Escherichia coli</i> MS 69-1	0.0000	0.0000	0.0150	0.0000
<i>Escherichia coli</i> MS 84-1	0.0000	0.0099	0.0000	0.0260
<i>Escherichia coli</i> NA114	0.2000	0.0659	0.0846	0.1540
<i>Escherichia coli</i> O103:H2 str. 12009	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O104:H4 str. 2011C-3493	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O127:H6 str. E2348/69	0.0000	0.0000	0.0172	0.0000
<i>Escherichia coli</i> O145:H28 str. 4865/96	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O157:H7 str. Sakai	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O83:H1 str. NRG 857C	0.0000	0.0000	0.0000	0.0391
<i>Escherichia coli</i> S6400	0.0000	0.0803	0.0000	0.0301
<i>Escherichia coli</i> SCD2	0.0000	0.0000	0.0012	0.0287
<i>Escherichia coli</i> SE11	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> UMEA 3318-1	0.2000	0.0809	0.0921	0.0930
<i>Escherichia coli</i> UMN026	0.2000	0.1393	0.1287	0.0780
<i>Escherichia coli</i> upec-182	0.0000	0.0000	0.0000	0.0294
<i>Escherichia coli</i> upec-55	0.0000	0.0125	0.0000	0.0000
<i>Escherichia coli</i> upec-64	0.0000	0.0320	0.0000	0.0000

表 8-2 模拟复杂环境样品菌株估计结果

菌株名称	理论值	0.1×		
		1	2	3
<i>Escherichia coli</i> K-12 (ref)	0.0000	0.0495	0.0359	0.0240
<i>Escherichia coli</i> 102606_aEPEC	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 1-182-04_S3_C2	0.0000	0.0033	0.0000	0.0000
<i>Escherichia coli</i> 1-250-04_S3_C1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 138B_L1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 207A_L2	0.0000	0.0040	0.0046	0.0326
<i>Escherichia coli</i> 2-222-05_S3_C2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 400791	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 401140	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 402403	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 48	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 53638	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 541-15	0.0000	0.0693	0.0687	0.0764
<i>Escherichia coli</i> 80B_L1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> B185	0.0000	0.0151	0.0000	0.0000
<i>Escherichia coli</i> B354	0.0000	0.0000	0.0000	0.0112

<i>Escherichia coli</i> B7C5	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> B str. REL606	0.2000	0.2543	0.2343	0.2234
<i>Escherichia coli</i> CFT073	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> ED1a	0.0000	0.0000	0.0111	0.0000
<i>Escherichia coli</i> H299	0.0000	0.0000	0.0000	0.0250
<i>Escherichia coli</i> HS	0.2000	0.1339	0.1503	0.1473
<i>Escherichia coli</i> IAI39	0.0000	0.0711	0.0827	0.0008
<i>Escherichia coli</i> KTE102	0.0000	0.0000	0.0017	0.0000
<i>Escherichia coli</i> KTE156	0.0000	0.0000	0.0000	0.0319
<i>Escherichia coli</i> KTE33	0.0000	0.0023	0.0000	0.0074
<i>Escherichia coli</i> M863	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> MS 69-1	0.0000	0.0045	0.0000	0.0000
<i>Escherichia coli</i> MS 84-1	0.0000	0.0028	0.0000	0.0000
<i>Escherichia coli</i> NA114	0.2000	0.1167	0.1264	0.1411
<i>Escherichia coli</i> O103:H2 str. 12009	0.0000	0.0000	0.0000	0.0074
<i>Escherichia coli</i> O104:H4 str. 2011C-3493	0.0000	0.0000	0.0008	0.0000
<i>Escherichia coli</i> O127:H6 str. E2348/69	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O145:H28 str. 4865/96	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O157:H7 str. Sakai	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O83:H1 str. NRG 857C	0.0000	0.0000	0.0070	0.0022
<i>Escherichia coli</i> S6400	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> SCD2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> SE11	0.0000	0.0000	0.0000	0.0077
<i>Escherichia coli</i> UMEA 3318-1	0.2000	0.1345	0.1399	0.1107
<i>Escherichia coli</i> UMN026	0.2000	0.1388	0.1366	0.1509
<i>Escherichia coli</i> upec-182	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> upec-55	0.0000	0.0000	0.0001	0.0000
<i>Escherichia coli</i> upec-64	0.0000	0.0000	0.0000	0.0000

表 8-3 模拟复杂环境样品菌株估计结果

菌株名称	理论值	1×		
		1	2	3
<i>Escherichia coli</i> K-12 (ref)	0.0000	0.0440	0.0483	0.0527
<i>Escherichia coli</i> 102606_aEPEC	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 1-182-04_S3_C2	0.0000	0.0000	0.0049	0.0000
<i>Escherichia coli</i> 1-250-04_S3_C1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 138B_L1	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 207A_L2	0.0000	0.0075	0.0082	0.0000
<i>Escherichia coli</i> 2-222-05_S3_C2	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 400791	0.0000	0.0000	0.0052	0.0000
<i>Escherichia coli</i> 401140	0.0000	0.0000	0.0032	0.0014
<i>Escherichia coli</i> 402403	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> 48	0.0000	0.0000	0.0000	0.0000

<i>Escherichia coli</i> 53638	0.0000	0.0042	0.0000	0.0000
<i>Escherichia coli</i> 541-15	0.0000	0.0828	0.0778	0.0736
<i>Escherichia coli</i> 80B_L1	0.0000	0.0000	0.0000	0.0049
<i>Escherichia coli</i> B185	0.0000	0.0003	0.0000	0.0006
<i>Escherichia coli</i> B354	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> B7C5	0.0000	0.0018	0.0000	0.0093
<i>Escherichia coli</i> B str. REL606	0.2000	0.2203	0.2065	0.2211
<i>Escherichia coli</i> CFT073	0.0000	0.0000	0.0019	0.0000
<i>Escherichia coli</i> ED1a	0.0000	0.0008	0.0000	0.0000
<i>Escherichia coli</i> H299	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> HS	0.2000	0.1292	0.1279	0.1190
<i>Escherichia coli</i> IAI39	0.0000	0.0589	0.0498	0.0542
<i>Escherichia coli</i> KTE102	0.0000	0.0015	0.0000	0.0028
<i>Escherichia coli</i> KTE156	0.0000	0.0000	0.0000	0.0048
<i>Escherichia coli</i> KTE33	0.0000	0.0059	0.0084	0.0087
<i>Escherichia coli</i> M863	0.0000	0.0000	0.0000	0.0010
<i>Escherichia coli</i> MS 69-1	0.0000	0.0000	0.0027	0.0009
<i>Escherichia coli</i> MS 84-1	0.0000	0.0017	0.0000	0.0000
<i>Escherichia coli</i> NA114	0.2000	0.1403	0.1329	0.1275
<i>Escherichia coli</i> O103:H2 str. 12009	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O104:H4 str. 2011C-3493	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O127:H6 str. E2348/69	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> O145:H28 str. 4865/96	0.0000	0.0060	0.0029	0.0000
<i>Escherichia coli</i> O157:H7 str. Sakai	0.0000	0.0019	0.0000	0.0007
<i>Escherichia coli</i> O83:H1 str. NRG 857C	0.0000	0.0000	0.0118	0.0000
<i>Escherichia coli</i> S6400	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> SCD2	0.0000	0.0080	0.0046	0.0000
<i>Escherichia coli</i> SE11	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> UMEA 3318-1	0.2000	0.1463	0.1543	0.1631
<i>Escherichia coli</i> UMN026	0.2000	0.1342	0.1405	0.1466
<i>Escherichia coli</i> upec-182	0.0000	0.0043	0.0083	0.0070
<i>Escherichia coli</i> upec-55	0.0000	0.0000	0.0000	0.0000
<i>Escherichia coli</i> upec-64	0.0000	0.0000	0.0000	0.0000

表 9 模拟复杂环境样品菌株估计的正误比例

类别	理论值	0.01×			0.1×			1×		
		1	2	3	1	2	3	1	2	3
所选大肠杆菌菌株	1.0000	0.5536	0.6023	0.5147	0.7782	0.7874	0.7734	0.7703	0.7622	0.7774
其他大肠杆菌菌株	0.0000	0.4464	0.3977	0.4853	0.2218	0.2126	0.2266	0.2297	0.2378	0.2226

上述结果表明，即使是对于 14 个亚群构成的特定群，当测序深度为 0.1

×或更高时，本发明方法可以非常有效地对特定群中的各亚群进行定量分析，并提供可靠的检测结果。当测序深度为 0.1×和 1×时，检测值与理论值的误差均小于 25%。

- 5 在本发明提及的所有文献都在本申请中引用作为参考，就如同每一篇文献被单独引用作为参考那样。此外应理解，在阅读了本发明的上述讲授内容之后，本领域技术人员可以对本发明作各种改动或修改，这些等价形式同样落于本申请所附权利要求书所限定的范围。

权 利 要 求

1. 一种非诊断性的对特定群中的亚群进行定量分析的方法，其特征在于，包括以下步骤：

5 (1) 提供对应于所述特定群的 (a) 参考基因组序列数据、(b) 参考 SNP 矩阵和 (c) 宏基因组测序数据，其中所述的特定群包括 N 个亚群，N 为 ≥ 2 的正整数；

(2) 将所述宏基因组测序数据比对到对应于所述特定群的参考基因组数据上，以便获得比对结果；

10 (3) 根据所述参考 SNP 矩阵的位点信息，从所述比对结果中提取出各分析位点上 4 种碱基 A、T、C、和 G 各自的分布频率，构建频率矩阵；

(4) 根据频率矩阵对参考 SNP 矩阵做二值化处理，得到二值化 SNP 矩阵，其中在所述二值化处理时，对于某一分析位点而言，若检出的碱基与参考矩阵上的碱基相同则记 1，否则记为 0；并且如果某一分析位点未被覆盖，则该分析位点不参与观测模型的构建；和

15 (5) 基于所述的频率矩阵、所述的二值化 SNP 矩阵、理论碱基频率向量 $f(x)$ 和观测碱基频率向量 y ，通过有约束的线性模型，得出所述特定群中各亚群的相对丰度，从而获得对所述特定群中的亚群的定量分析结果。

2. 如权利要求 1 所述的方法，其特征在于，在步骤 (5) 中，包括以下步骤：

(5a) 设所述观测碱基频率向量为 y ， y 由所述频率矩阵得到；

20 以及设定理论碱基频率向量为 $f(x)$ ，

$$f(x) = \omega^T x$$

式中， x 为所述二值化 SNP 矩阵；

ω 为列向量，即所述各亚群 (1, 2, 3, $\dots$ N) 的比例值 ($\omega_1, \omega_2, \omega_3 \dots \omega_N$) 构成的列向量 ω ，

25 上标 T 表示转置，行向量转置后为列向量，反之亦然；

(5b) 基于下式 Q1、Q2、Q3、Q4，

$$f(x) \approx y \quad (Q1)$$

$$\omega = (\omega_1, \omega_2, \omega_3 \dots \omega_N)^T \quad (Q2)$$

$$\sum_{i=1}^N \omega_i = 1 \quad (Q3)$$

$$30 \quad 0 \leq \omega_i \leq 1 \quad (Q4)$$

运用回归方法，对各亚群比例进行求解，使 ω^* 满足式 Q5 和 Q6，从而得出所述特定群中各亚群的相对丰度，

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (f(x_i) - y_i)^2 \quad (\text{Q5})$$

5

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^m (\omega^T x_i - y_i)^2 \quad (\text{Q6})$$

各式中，

m 为频率不为零的碱基总个数；

T ， x ， y ，和 ω 如上定义。

10

3. 如权利要求 1 所述的方法，其特征在于， N 为 2-50，较佳地 3-20，更佳地 3-10。

4. 如权利要求 1 所述的方法，其特征在于，所述的 $y = (q_1, q_2, \dots, q_m)$ ，其中，各 $q_1, q_2, \dots, q_m$ 为频率矩阵中不等于 0 的各分布频率， m 为所述频率矩阵中频率不为零的碱基总个数。

15

5. 如权利要求 1 所述的方法，其特征在于，所述的特定群为种，而所述的亚群为亚种，或株。

6. 如权利要求 2 所述的方法，其特征在于，在步骤(5b)中，使理论碱基频率($f(x)$)与观测碱基频率(y)的残差平方和最小。

20

7. 如权利要求 2 所述的方法，其特征在于，在步骤(5b)中，使用最小二乘法求解。

8. 如权利要求 2 所述的方法，其特征在于，在步骤(5b)中，使用序贯最小二乘法求解。

9. 如权利要求 1 或 2 所述的方法，其特征在于，所述群选自下组：细菌、病毒、肿瘤细胞、真菌、衣原体、支原体；

25

和/或所述亚群选自下组：亚种、异质细胞、菌株。

10. 一种非诊断性的对特定群中的亚群进行定量分析的方法，其特征在于，包括以下步骤：

(1) 提供对应于所述特定群的 (a) 参考基因组序列数据或参考转录组序列数据、(b) 参考 SNP 矩阵和 (c) 宏转录组测序数据，其中所述的特定群包括 N 个亚群， N 为 ≥ 2 的正整数；

30

(2) 将所述宏转录组测序数据比对到对应于所述特定群的参考基因组数据或参考转录组序列数据上，以便获得比对结果；

(3) 根据所述参考 SNP 矩阵的位点信息，从所述比对结果中提取出各分析位点上 4 种碱基 A、T、C、和 G 各自的分布频率，构建频率矩阵；

5 (4) 根据频率矩阵对参考 SNP 矩阵做二值化处理，得到二值化 SNP 矩阵，其中在所述二值化处理时，对于某一分析位点而言，若检出的碱基与参考矩阵上的碱基相同则记 1，否则记为 0；并且如果某一分析位点未被覆盖，则该分析位点不参与观测模型的构建；和

10 (5) 基于所述的频率矩阵、所述的二值化 SNP 矩阵、理论碱基频率向量 $f(x)$ 和观测碱基频率向量 y ，通过有约束的线性模型，得出所述特定群中各亚群的相对丰度，从而获得对所述特定群中的亚群的定量分析结果。

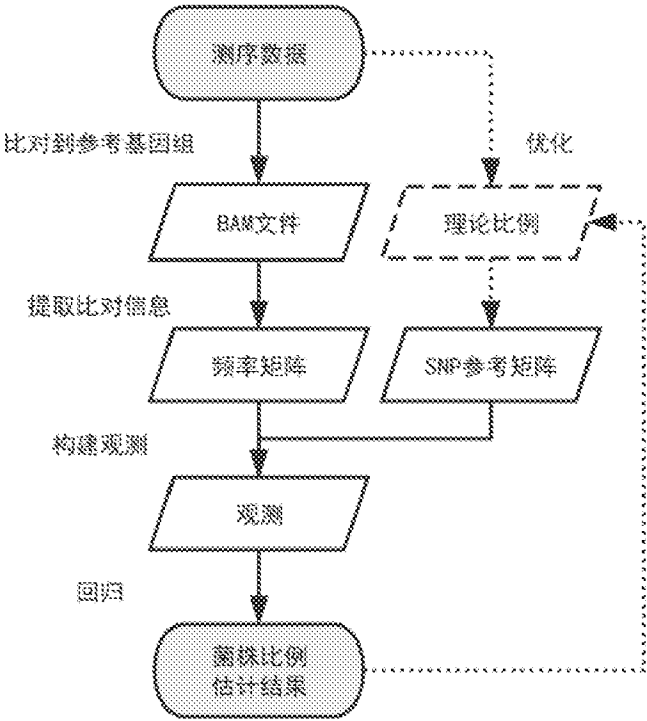


图 1

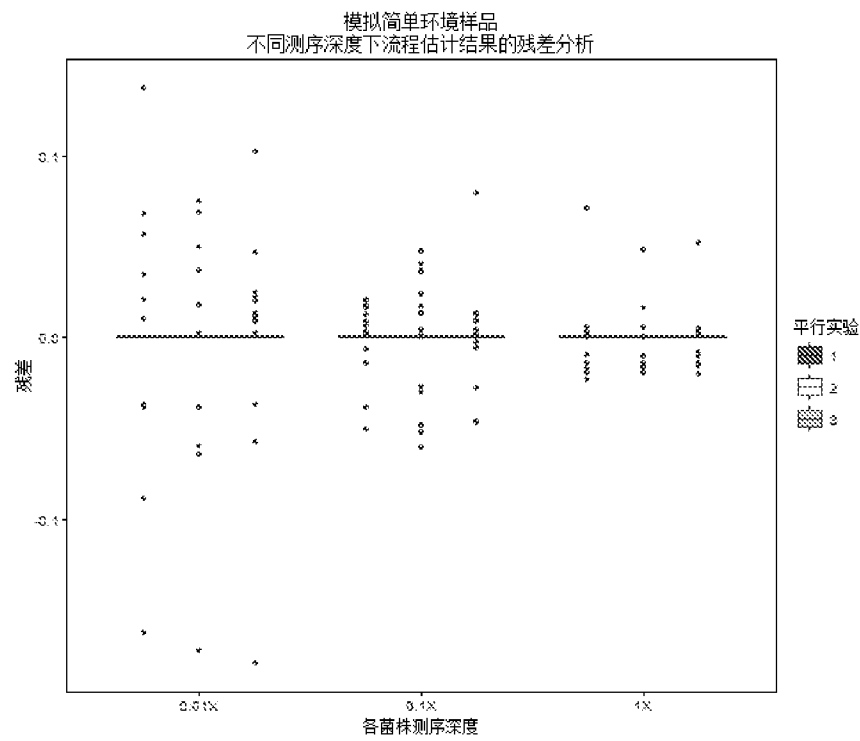


图 2

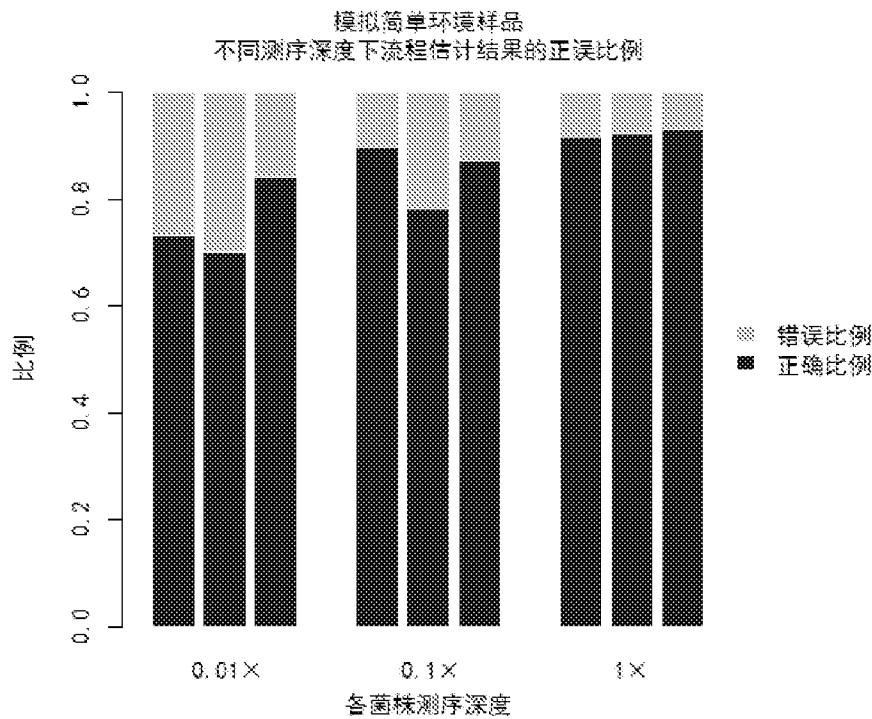


图 3

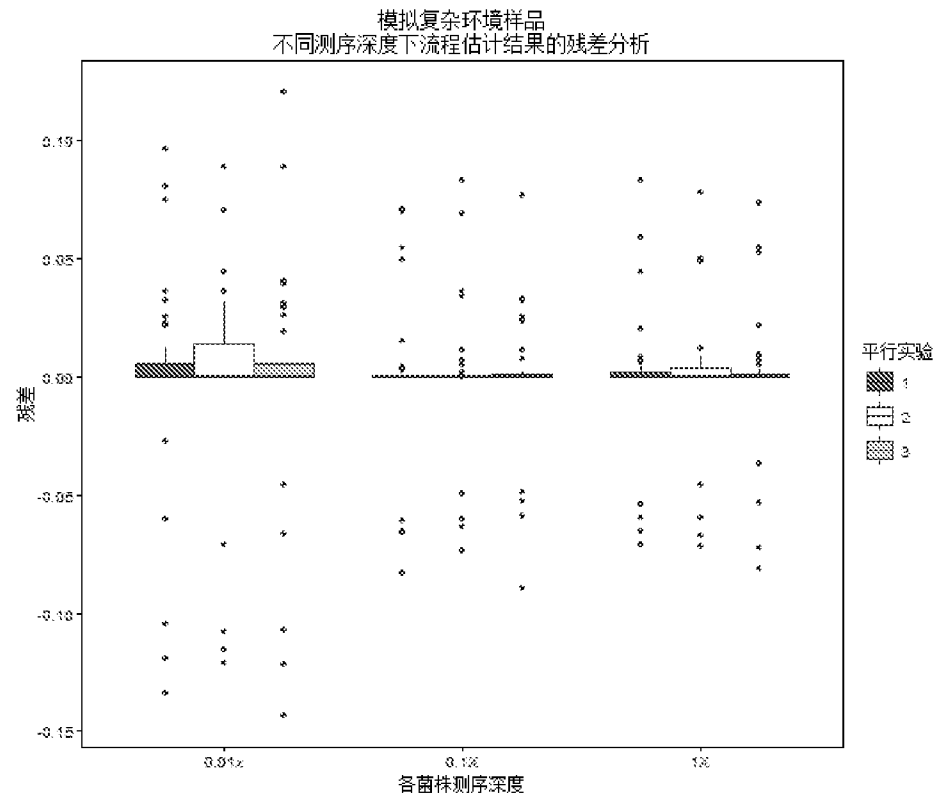


图 4

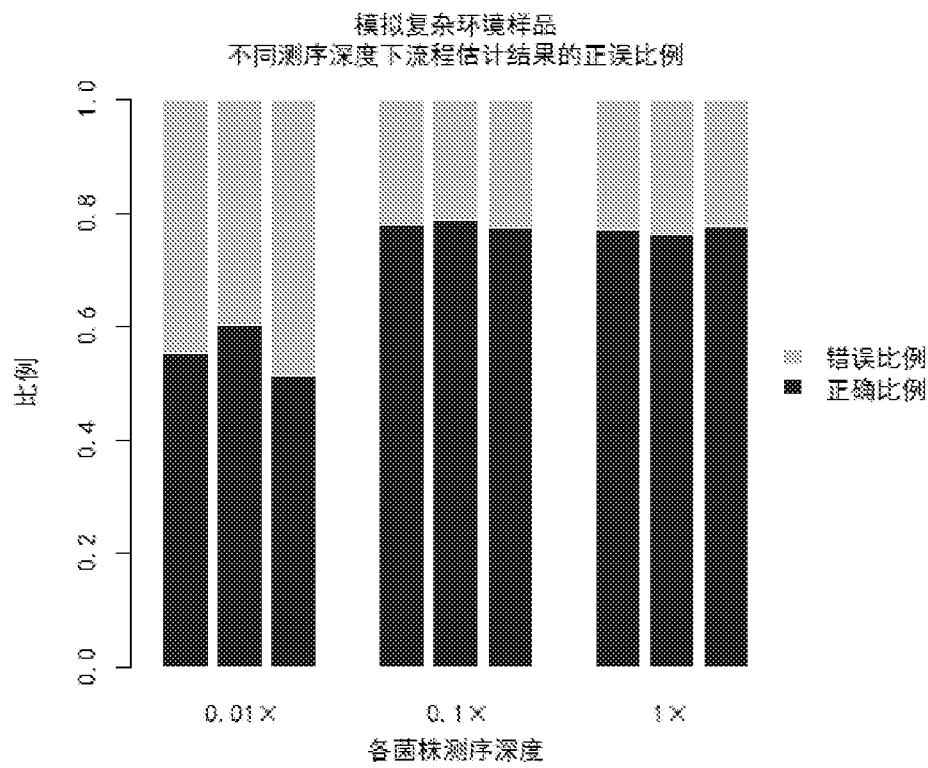


图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/105372

A. CLASSIFICATION OF SUBJECT MATTER

G06F 19/20 (2011.01) i; G06F 19/22 (2011.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, DWPI, VEN, CNKI, PubMed, ISI Web of Knowledge, CNTXT, EPTXT, USTXT: 华大基因, 彭也, 李俊桦, 唐珊媚, 张慧, 群, 亚群, 亚种, 菌株, 细菌, 宏基因组, 丰度, 定量分析, 频率, SNP, 二值化, 碱基频率, metagenomic, binning, typing, bacterial strains, genotype+, constrains, WG-FAST, group, subgroup, strain+, microbial strains, strain profiling, abundance, frequency, quantitative analysis, binaryzation, base, matrix

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 103955629 A (JILIN UNIVERSITY), 30 July 2014 (30.07.2014), abstract, and claims 1-3	1-10
A	陈波等, “宏基因组分类问题中的特征提取及其降维研究”, 计算机系统应用, 24(11), 31 December 2015 (31.12.2015), pages 31-37, (CHEN, Bo et al., “Features Extraction and Dimensions Reduction in Metagenomic Binning Problem”, Computer Systems & Applications)	1-10
A	SAHL, J.W. et al., “Phylogenetically Typing Bacterial Strains from Partial SNP Genotypes Observed from Direct Sequencing of Clinical Specimen Metagenomic Data.”, Genome Medicine., 7(52), 09 June 2015 (09.06.2015), pages 1-13	1-10
A	LUO, C.W. et al., “ConStrains Identifies Microbial Strains in Metagenomic Datasets.”, Nat. Biotechnol., 33(10), 07 September 2015 (07.09.2015), pages 1045-1052	1-10

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	
“E” earlier application or patent but published on or after the international filing date	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	“&” document member of the same patent family

Date of the actual completion of the international search 18 July 2017	Date of mailing of the international search report 07 August 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer LIAO, Wenyong Telephone No. (86-10) 62413867

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/105372

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106055924 A (PERFECT (CHINA) CO., LTD.), 26 October 2016 (26.10.2016), entire document	1-10
A	ALEXEEV, D. et al., "Bacterial Rose Garden for Metagenomic SNP-Based Phylogeny Visualization.", Biodata Mining., 8(10), 21 March 2015 (21.03.2015), pages 1-12	1-10
A	NAYFACH, S. et al., "An Integrated Metagenomics Pipeline for Strain Profiling Reveals Novel Patterns of Bacterial Transmission and Biogeography.", Genome Res., vol. 26, 18 October 2016 (18.10.2016), pages 1-14	1-10

International application No.
PCT/CN2016/105372

Form PCT/ISA/210 (patent family annex) (July 2009)

国际检索报告

国际申请号

PCT/CN2016/105372

A. 主题的分类

G06F 19/20(2011.01)i; G06F 19/22(2011.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNABS, DWPI, VEN, CNKI, PubMed, ISI Web of Knowledge, CNTXT, EPTXT, USTXT, 华大基因, 彭也, 李俊桦, 唐珊媚, 张慧, 群, 亚群, 亚种, 菌株, 细菌, 宏基因组, 丰度, 定量分析, 频率, SNP, 二值化, 碱基频率, metagenomic, binning, typing, bacterial strains, genotype+, constrains, WG-FAST, group, subgroup, strain+, microbial strains, strain profiling, abundance, frequency, quantitative analysis, binaryzation, base, matrix

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 103955629 A (吉林大学) 2014年 7月 30日 (2014 - 07 - 30) 摘要, 权利要求1-3	1-10
A	陈波 等. "宏基因组分类问题中的特征提取及其降维研究." 计算机系统应用., 第24卷, 第11期, 2015年 12月 31日 (2015 - 12 - 31), 第31-37页	1-10
A	SAHL, J. W.等. "Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data." GENOME MEDICINE., 第7卷, 第52期, 2015年 6月 9日 (2015 - 06 - 09), 第1-13页	1-10
A	LUO, C. W.等. "ConStrains identifies microbial strains in metagenomic data-sets." NAT. BIOTECHNOL., 第33卷, 第10期, 2015年 9月 7日 (2015 - 09 - 07), 第1045-1052页	1-10

☒ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2017年 7月 18日

国际检索报告邮寄日期

2017年 8月 7日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

廖文勇

电话号码 (86-10)62413867

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 106055924 A (完美中国有限公司) 2016年 10月 26日 (2016 - 10 - 26) 全文	1-10
A	ALEXEEV, D. 等. "Bacterial rose garden for metagenomic SNP-based phylogeny visualization." BIODATA MINING., 第8卷, 第10期, 2015年 3月 21日 (2015 - 03 - 21), 第1-12页	1-10
A	NAYFACH, S. 等. "An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography." GENOME RES., 第26卷, 2016年 10月 18日 (2016 - 10 - 18), 第1-14页	1-10

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/105372

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	103955629	A	2014年 7月 30日	无	
CN	106055924	A	2016年 10月 26日	无	