智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于MCCA的痤疮宏基因组数据辅助分析

孙梦茹，王瑜，何聪芬，贾焱，高学义

引用本文:
孙梦茹, 王瑜, 何聪芬, 等. 基于MCCA的痤疮宏基因组数据辅助分析[J]. 智能系统学报, 2020, 15(5): 972–977.
SUN Mengru, WANG Yu, HE Congfen, et al. Assisted analysis of acne metagenomic sequencing data using multi–set canonical correlation analysis methods[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(5): 972–977.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201810005

DOI: 10.11992/tis.201810005
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190516.2353.002.html

基于 MCCA 的痤疮宏基因组数据辅助分析

孙梦茹[1], 王瑜[1], 何聪芬[2], 贾焱[2], 高学义[1]
(1. 北京工商大学 计算机与信息工程学院 食品安全大数据技术北京市重点实验室, 北京 100048; 2. 北京工商大学 理学院 中国轻工业化妆品重点实验室, 北京 100048)

摘 要: 痤疮作为常见皮肤病之一, 发病机制复杂, 其中微生物定植在痤疮发病中的作用是一个热点研究问题。本文从宏基因组学的角度, 利用机器学习方法分析痤疮宏基因组数据, 包括痤疮患者的患病皮肤 (diseased skin, DS) 样本集和健康皮肤 (healthy skin, HS) 样本集, 以及正常对照组 (normal control, NC) 样本集。为了同时分析 3 组样本集以获得可以区分不同样本集的脂质, 使用多重集典型相关分析 (multi-set canonical correlation analysis, MCCA) 方法进行研究。实验结果可得到仅对某一样本集有显著影响的脂质, 以及同时对 3 个样本集影响程度不同的脂质, 这些脂质可以作为判别皮肤状态的指标, 用于辅助指导皮肤痤疮疾病的诊断、预后和治疗。

关键词: 痤疮; 宏基因组学; 面部皮肤; 脂质; 机器学习; 多重集典型相关分析
中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2020)05−0972−06

中文引用格式: 孙梦茹, 王瑜, 何聪芬, 等. 基于 MCCA 的痤疮宏基因组数据辅助分析 [J]. 智能系统学报, 2020, 15(5): 972–977.
英文引用格式: SUN Mengru, WANG Yu, HE Congfen, et al. Assisted analysis of acne metagenomic sequencing data using multi-set canonical correlation analysis methods[J]. CAAI transactions on intelligent systems, 2020, 15(5): 972–977.

Assisted analysis of acne metagenomic sequencing data using multi-set canonical correlation analysis methods

SUN Mengru[1], WANG Yu[1], HE Congfen[2], JIA Yan[2], GAO Xueyi[1]
(1. Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China; 2. Key Laboratory of Cosmetic of China National Light Industry, School of Science, Beijing Technology and Business University, Beijing 100048, China)

Abstract: As one of the common skin diseases, the pathogenesis of acne is very complicated. The role of microbial colonization in the pathogenesis of acne is an active research area. The purpose of this paper is to analyze acne metagenomic data, including sample sets of acne diseased skin (DS) and healthy skin (HS) as well as normal control (NC) by using machine learning from the perspective of macrogenomics. Multi-set canonical correlation analysis (MCCA) method is used to analyze the above three sample sets at the same time and to confirm the lipids that can distinguish these three sample sets. The results show that lipids that had a significant impact on only one set and those that had different impacts on the three sample sets respectively can be used as indicators to determine the skin status. Moreover, these lipids can be used to guide diagnosis, prognosis, and treatment of skin acne diseases.

Keywords: acne; macrogenomics; facial skin; lipids; machine learning; multi-set canonical correlation analysis

痤疮是世界上最常见的皮肤病之一, 表现为一种毛囊皮脂腺的慢性炎症性, 主要发生于面部、胸背部等皮脂溢出区, 患者表现为粉刺、丘疹、脓疱、囊肿、结节及萎缩性瘢痕等皮损, 大约会影响 80% 的青少年和青壮年[1]。痤疮普遍而且错误的被概括为只是患者经历的一个阶段, 但对于一些人来说, 痤疮可以持续多年, 不仅影响患

收稿日期: 2019−10−09. **网络出版日期:** 2019−05−17.
基金项目: 国家自然科学基金面上项目 (61671028); 北京市自然科学基金面上项目 (4162018).
通信作者: 王瑜. E-mail: wangyu@btbu.edu.cn.

者的面容外观, 还常伴有疼痛、瘙痒等躯体感觉, 甚至会引起自卑、焦虑、抑郁等心理疾病, 严重影响患者的身心健康[2-3]。因此在皮肤病学领域, 痤疮的研究和治疗是一个广泛研究的问题。

痤疮的发病机制复杂, 目前国内外公认的包括毛囊导管角化异常、微生物定植、皮脂分泌增加以及炎症反应等[4-5]。其中, 微生物定植在痤疮发病中的作用是一个热点研究问题, 大多工作是针对单一微生物的研究, 如痤疮丙酸杆菌、金黄色葡萄球菌及表皮葡萄球菌等被认为和痤疮发病有一定的关联性[6-9]。

随着人类微生物组计划的开展, 人们逐渐意识到, 人的健康状况可以通过对人类微生物组的研究分析而评估, 与宿主生活在一起的微生物在大部分情况下是作为一个整体发挥着重要作用[10]。微生物组是指存在于微生物群中的基因组和基因的集合。然而, 对单个微生物基因组的研究存在着一定的限制, 自然界中 99% 的微生物不能通过分离和培养进行研究, 而且微生物更倾向于作为微生物群这样的整体发挥作用, 因此, 研究人员提出宏基因组学, 即环境微生物中所有物种基因组信息的总和[11]。

在高通量测序技术迅速发展的推动下, 宏基因组学吸引了大量研究人员, 通过挖掘不同部位宏基因组的微生物群落结构, 以及分析不同健康状态的宏基因组样本的差异, 去探索人体健康与其寄宿的微生物之间玄妙的相互关系[12]。研究人员通过肠道微生物序列分析, 发现肠道菌群紊乱与儿童孤独症发生有相关性[12]。此外, 也发现许多了许多其他人类疾病, 包括癌症、糖尿病, 甚至神经发育障碍均与微生物组有关[13-14]。近年来, 研究人员开始利用机器学习方法进行宏基因组学的研究工作。Huang 等[15] 使用主成分分析方法分析牙龈炎和健康牙龈的数据, 获得主要影响牙龈炎的细菌。Wisittipanit 等[16] 利用支持向量机对炎症性肠炎的病人和健康人群的肠道微生物样本进行分类。Qin 等[17] 使用相关分析方法研究 Ⅱ 型糖尿病患者和健康人群的肠道宏基因组研究, 发现可以区分样本的基因簇。

目前, 关于痤疮宏基因组数据的研究比较缺乏, 而基于机器学习方法在宏基因组数据上的有效应用已经有目共睹, 本文尝试使用多重集典型相关分析 (multi-set canonical correlation analysis, MCCA) 方法分析痤疮的宏基因组测序数据, 具体包括健康皮肤数据、痤疮患者的健康皮肤数据和患病皮肤数据。获得对不同样本集有不同影响的脂质, 以及仅对其中一个样本集有显著影响的脂质, 这些脂质可以有效地区分不同的皮肤状态, 可用于指导痤疮的预防、诊断和治疗过程。

1　样本与方法

1.1　样本采集

本次实验收集 35 名痤疮患者面部皮肤的感染细胞和健康细胞, 同时收集没有患痤疮的 35 名志愿者的面部皮肤细胞作为正常对照组 (NC)。使用色谱设备 (Waters ACQUITY UPLC I-Class (Waters Corporation, Milford, Massachusetts, USA)), 保持流速为 0.3 mL/min, 注射量为 2.0 μL。使用流动相洗涤注射器针头在超高效液相色谱 (ultra performance liquid chromatography, UPLC) 运行期间, 洗脱液出口连接到 QTOF-MS 来用于实体检测和表征。高分辨率质量测量使用设备 (Waters Xevo G2-XS QTOF-MS (Waters Corporation, Milford, Massachusetts, USA)), 该设备配有以正离子模式操作的电喷雾电离 (electrospray ionization, ESI) 界面。在操作色谱流动流速下, 将 UPLC 系统洗脱物引入装置 QTOF-MS, 使用氮气作为雾化和脱溶剂化气体, 通过系统 (Masslynx 4.1 (Waters Corporation, Milford, Massachusetts, USA)) 收集 UPLC-QTOF-MS 数据作为质心原始数据。最终获得痤疮患者的患病皮肤 (diseased skin, DS) 样本集, 健康皮肤 (healthy skin, HS) 样本集, 以及正常对照组 (normal control, NC) 样本集, 其中每个样本集包括有 35 名志愿者, 每名志愿者收集 2 520 个序列。

1.2　多重集典型相关分析

当研究两组样本集的数据分析时, 典型相关分析 (canonical correlation analysis, CCA) 可以取得较好的效果, 但是在分析多组样本集 (不小于 3 组) 时, CCA 却很难得到令人满意的效果。为了同时分析 3 组样本集, 本文选用 MCCA 方法, 该方法是一种研究多组样本集之间关系的方法, 给定多个样本集 $X_1, X_2, \cdots, X_n$, 假设每个样本集包括 N 个样本, 定义多重集典型相关分析的准则函数为

$$\boldsymbol{J}_{MCCA}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{S}_{ij}\boldsymbol{\omega}_j}{\sqrt{\sum_{i=1}^{n}\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{S}_{ii}\boldsymbol{\omega}_j}} \tag{1}$$

式中: $\boldsymbol{S}_{ij}=\boldsymbol{X}_i^{\mathrm{T}}\cdot\boldsymbol{X}_j$, 表示随机向量 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 的互协方差矩阵, $\boldsymbol{S}_{ii}=\boldsymbol{X}_i^{\mathrm{T}}\cdot\boldsymbol{X}_i$, 表示的是随机向量 $\boldsymbol{X}_i$ 的协方差矩阵。MCCA 问题可以简化为如下最优化模型的求解:

$$\mathrm{argmax}\beta=\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{S}_{ij}\boldsymbol{\omega}_j \ \ s.t.\sum_{i=1}^{n}\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{S}_{ii}\boldsymbol{\omega}_j=1 \tag{2}$$

即当样本集之间的相关系数 β 最大时，找到对应于每个样本集的典型变量 ω_i。因此，使用拉格朗日乘子法求解式 (2)，得

$$\boldsymbol{L}(\omega_1,\omega_2,\cdots,\omega_n)=\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{S}_{ij}\boldsymbol{\omega}_j-\beta\left(\sum_{i=1}^{n}\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{S}_{ii}\boldsymbol{\omega}_j-1\right) \tag{3}$$

令 $\dfrac{\partial \boldsymbol{L}}{\partial \boldsymbol{\omega}_i}=0, i=1,2,\cdots,n$，得到：

$\sum_{i=1}^{n}\boldsymbol{S}_{ij}\boldsymbol{\omega}_j=\boldsymbol{\beta}\boldsymbol{S}_{ii}\boldsymbol{\omega}_i, i=1,2,\cdots,n$，即可写作：

$$(\boldsymbol{C}-\boldsymbol{D})\boldsymbol{\omega}=\beta\boldsymbol{D}\boldsymbol{\omega} \tag{4}$$

其中，

$$\boldsymbol{C}=\begin{bmatrix}\boldsymbol{x}_1\boldsymbol{x}_1^{\mathrm{T}} & \cdots & \boldsymbol{x}_1\boldsymbol{x}_N^{\mathrm{T}}\\ \vdots & \ddots & \vdots\\ \boldsymbol{x}_N\boldsymbol{x}_1^{\mathrm{T}} & \cdots & \boldsymbol{x}_N\boldsymbol{x}_N^{\mathrm{T}}\end{bmatrix},\ \boldsymbol{D}=\begin{bmatrix}\boldsymbol{x}_1\boldsymbol{x}_1^{\mathrm{T}} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \boldsymbol{x}_N\boldsymbol{x}_N^{\mathrm{T}}\end{bmatrix}$$

得到每个样本集对应的典型变量 ω_i 后，与对应的原始样本集数据结合进行分析，获得对每个样本集有重要影响的脂质，统计这些脂质出现的频率，将其作为对样本集的贡献率，进而根据贡献率的不同获得区分不同样本集的脂质。

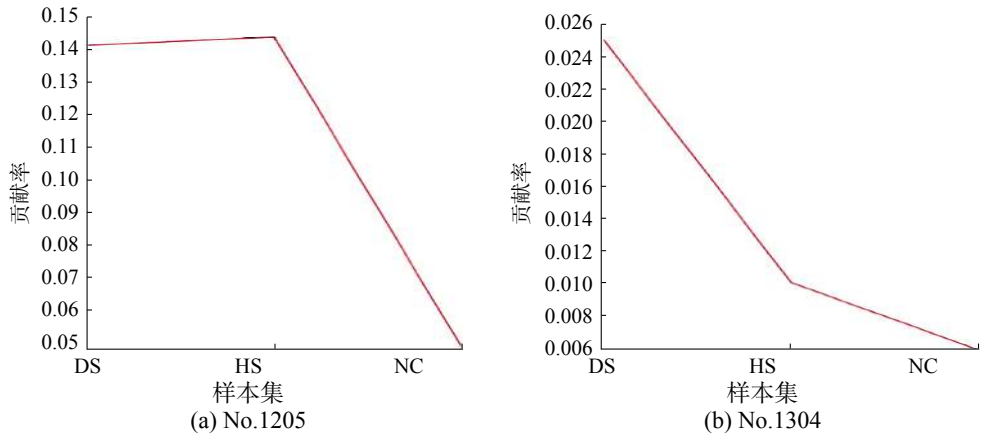
2　实验结果与分析

通过实验发现，使用 MCCA 方法能够有效分析痤疮宏基因组数据，算法复杂度为 $o(n^3)$。其中，有 15 种脂质同时与 3 组样本集有关，表 1 表示这 15 种脂质的具体描述。而这 15 种脂质，有两种脂质对不同样本集的贡献有明显差异，如图 1 所示。

表 1　MCCA 获得同时对 3 组样本有影响的脂质

Table 1　The lipids found by MCCA

编号	描述
1192	FMC-5(d18:1/18:0)
1205	1-(6-[5]-ladderane-hexanoyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerophosphocholine
1219	PG(20:3(8Z,11Z,14Z)/17:0)
1236	1-(8-[3]-ladderane-octanoyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerophosphoethanolamine
1240	PS(22:6(4Z,7Z,10Z,13Z,16Z,19Z)/18:1(9Z))
1244	PS(20:5(5Z,8Z,11Z,14Z,17Z)/20:0)
1245	1-(8-[3]-ladderane-octanoyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerophosphoethanolamine
1264	PS(22:6(4Z,7Z,10Z,13Z,16Z,19Z)/19:1(9Z))
1266	1-(6-[3]-ladderane-hexanoyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerophosphocholine
1279	1-(6-[3]-ladderane-hexanoyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerophosphocholine
1283	PS(18:4(6Z,9Z,12Z,15Z)/22:0)
1302	PS(20:3(8Z,11Z,14Z)/19:0)
1304	(3'-sulfo)Galbeta-Cer(d18:0/18:0(2OH))
1311	PS(22:6(4Z,7Z,10Z,13Z,16Z,19Z)/19:0)
1315	PS(22:2(13Z,16Z)/16:0)



(a) No.1205　　(b) No.1304

图 1　对 DS、HS 和 NC 样本集影响差异较大的脂质

Fig. 1　The effect of the lipids on DS, HS and NC samples

在图 1(a) 中，No.1205 表示的脂质在 NC 样本集中有较低的贡献率，几乎可以忽略不计，但在 DS 和 HS 样本集中显示出较高且相似的影响，因此可以使用这种脂质来区分 NC 样本集和其他两个样本集。图 1(b) 显示出 No.1304 代表的脂质对 DS、HS 和 NC 样本的影响呈单调递减的趋势，可以认为这是有效区分 DS、HS 和 NC 样本集的一个脂质。

除此之外，使用 MCCA 方法还可以获得仅对其中一个样本集有显著影响，但是对其他两组样本集几乎没有影响的脂质，如图 2 所示。图 2 中编号所代表的脂质具体描述如表 2 所示。图 2(a) 和图 2(b) 显示 No.95 和 No.1256 代表的脂质对 DS 样本的影响普遍大于 HS 和 NC 样本。在图 2(c) 和图 2(d) 中，No.608 和 No.2334 表示的脂质明显只会出现在 HS 样本集中，因此可以认为当这两

种脂质出现时, 痤疮患者的皮肤状态正在好转或者健康者的皮肤正在恶化。从图 2(e) 中明显看出, 当 No.2374 表示的脂质在 NC 样本集有明显的增高, 区别于 HS 和 DS 两个样本集, 它可以反映受试者的皮肤状态是健康的, 可以认为痤疮患者的治疗效果是显著的。

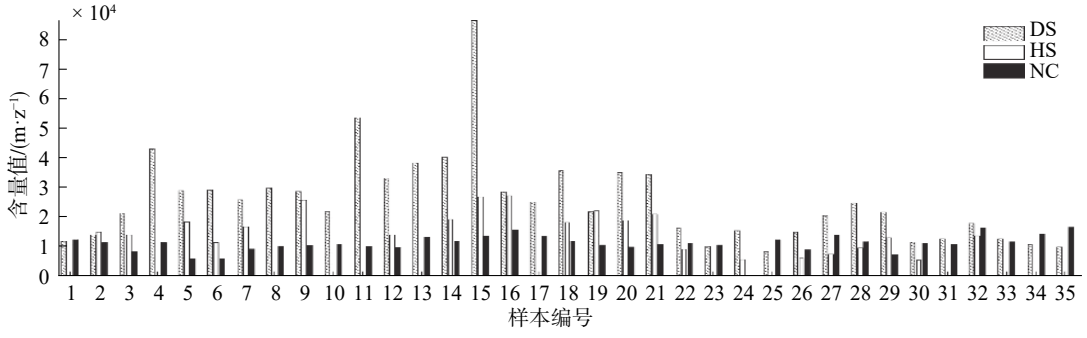
表 2　图 2 中脂质编号的具体描述

Table 2　The lipids represented in Figure 2

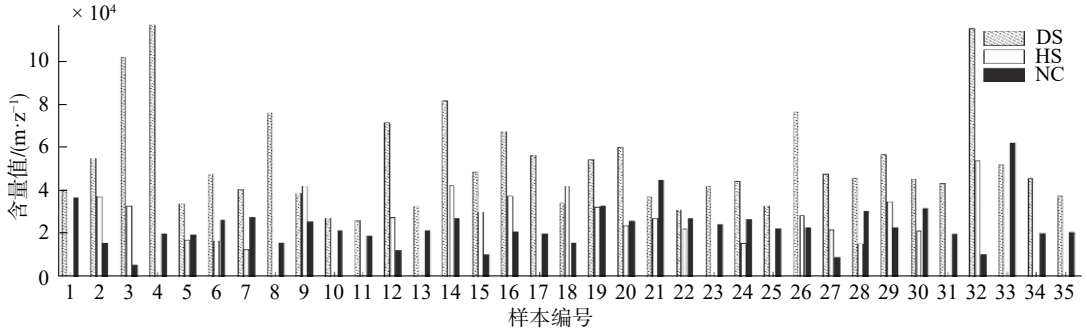
Label	描述
95	Prodelphinidin B6
608	PC(20:0/26:0)
1256	PI(20:2(11Z,14Z)/15:0)
2334	GlcAbeta-Cer(d18:1/18:0)
2374	Phoenicoxanthin/ Adonirubin/3-Hydroxycanthaxanthin

3　结束语

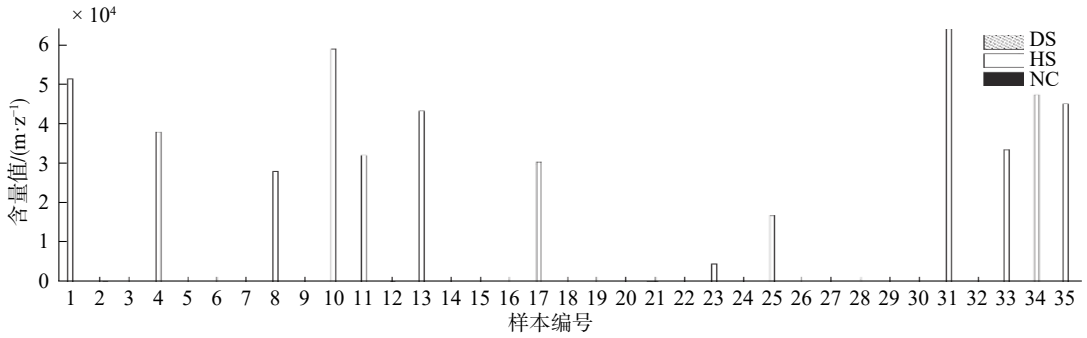
痤疮作为世界上最常见的皮肤疾病之一, 患



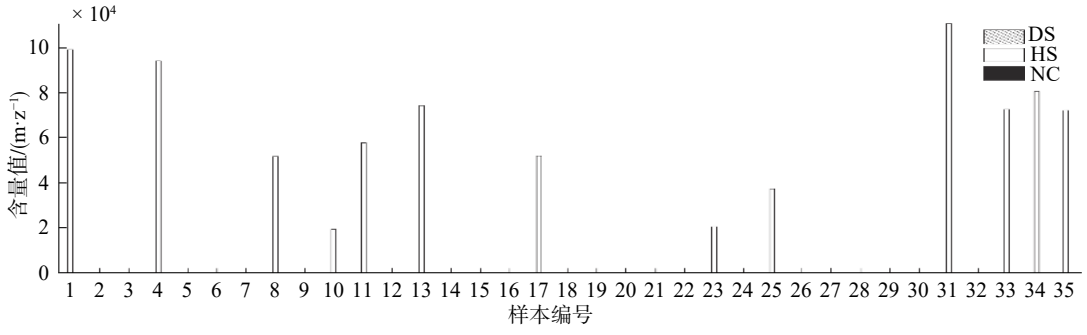
(a) No.95



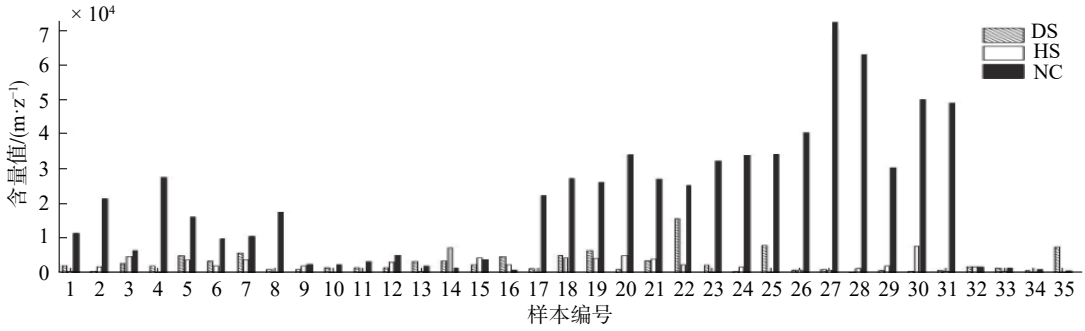
(b) No.1256



(c) No.608



(d) No.2334



(e) No.2374

图 2　区分 DS、HS 和 NC 样本集的脂质

Fig. 2　The lipids which can distinguish the DS, HS and NC samples.

病人数多、病因复杂，并且复发率高，虽然部分患者在青春期之后，其症状会有所缓解，但是对于大部分患者来说，痤疮症状会持续很长时间，对患者的生活质量造成很大的影响，因此对于痤疮的研究和治疗是一个具有重要意义的课题。本文从宏基因组学的角度分析引起痤疮发病的脂质，并尝试使用 MCCA 方法分析 DS、HS 和 NC3 个样本集，可以得到仅对某一样本集有显著影响的脂质，以及同时有效区分 3 个样本集的脂质。实验结果显示，MCCA 方法分析获得的脂质可以有效的区分 3 种不同的皮肤状态，并且对痤疮的预防、诊断和治疗有一定的辅助指导意义。在痤疮发病过程中，也许存在某种脂质的数值虽然很小，但却对痤疮有一定影响，而本文使用 MCCA 方法获取脂质对样本集的贡献时，会一定程度上弱化对这些脂质的分析，对于这些脂质还需要进一步的研究。

参考文献：

[1] MARONI G, ERMIDORO M, PREVIDI F, et al. Automated detection, extraction and counting of acne lesions for automatic evaluation and tracking of acne severity[C]// Proceedings of 2017 IEEE Symposium Series on Computational Intelligence. Honolulu, USA, 2017: 1−6.

[2] LUCUT S, SMITH M R. Dermatological tracking of chronic acne treatment effectiveness[C]//Proceedings of 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Orlando, USA, 2016: 5421−5426.

[3] THIBOUTOT D M, DRéNO B, ABANMI A, et al. Practical management of acne for clinicians: an international consensus from the global alliance to improve outcomes in acne[J]. Journal of the American academy of dermatology, 2018, 78(2, Suppl 1): S1–S23.e1.

[4] PAUGAM C, CORVEC S, SAINT-JEAN M, et al. Propionibacterium acnes phylotypes and acne severity: an observational prospective study[J]. Journal of the European academy of dermatology and venereology, 2017, 31(9): e398–e399.

[5] 王鸿. 寻常型痤疮发病机制研究进展 [J]. 西南医科大学学报, 2018, 41(4): 385–388.
WANG Hong. Research progress on the pathogenesis of acne vulgaris[J]. Journal of Southwest Medical University, 2018, 41(4): 385–388.

[6] FITZ-GIBBON S, TOMIDA S, CHIU B H, et al. Propionibacterium acnes strain populations in the human skin microbiome associated with acne[J]. Journal of investigative dermatology, 2013, 133(9): 2152–2160.

[7] DAGNELIE M, CORVEC S, SAINT-JEAN M, et al. 461 Diversity of Propionibacterium acnes phylotypes according to body localization in acne patients versus healthy controls[J]. Journal of investigative dermatology, 2017, 137(10, Suppl 2): S271.

[8] ZOUBOULIS C C, JOURDAN E, PICARDO M. Acne is an inflammatory disease and alterations of sebum composition initiate acne lesions[J]. Journal of the European academy of dermatology and venereology, 2014, 28(5): 527–532.

[9] 吴赟, 吉杰, 张玲琳, 等. 微生物在痤疮发病中的作用 [J]. 中国皮肤性病学杂志, 2016, 30(3): 311–314.
WU Yun, JI Jie, ZHANG Linglin, et al. Roles of microorganisms in the pathogenesis of acne[J]. The Chinese journal of dermatovenereology, 2016, 30(3): 311–314.

[10] ZHANG Xuegong, LIU Shansong, CUI Hongfei, et al. Reading the underlying information from massive metagenomic sequencing data[J]. Proceedings of the IEEE, 2017, 105(3): 459–473.

[11] VAN OPSTAL E J, BORDENSTEIN S R. Rethinking heritability of the microbiome[J]. Science, 2015, 349(6253): 1172–1173.

[12] KANG D W, PARK J G, ILHAN Z E, et al. Reduced in-

cidence of Prevotella and other fermenters in intestinal microflora of autistic children[J]. PLoS one, 2013, 8(7): e68322.

[13] SEARS C L, GARRETT W S. Microbes, microbiota, and colon cancer[J]. Cell host & microbe, 2014, 15(3): 317–328.

[14] HSIAO E Y, MCBRIDE S W, HSIEN S, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders[J]. Cell, 2013, 155(7): 1451–1463.

[15] HUANG Shi, LI Rui, ZENG Xiaowei, et al. Predictive modeling of gingivitis severity and susceptibility via oral microbiota[J]. The ISME journal, 2014, 8(9): 1768–1780.

[16] WISITTIPANIT N, RANGWALA H, GILLEVET P, et al. SVM-based classification and feature selection methods for the analysis of Inflammatory Bowel disease microbiome data[C]//Proceedings of the 9th International Workshop on Data Mining in Bioinformatics. Washington, USA, 2010: 1−8.

[17] QIN Junjie, LI Yingrui, CAI Zhiming, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes[J]. Nature, 2012, 490(7418): 55–60.

作者简介：

孙梦茹，硕士研究生，主要研究方向为图像处理、模式识别。

王瑜，副教授，博士，主要研究方向为图像处理、模式识别。申请国家发明专利 15 项。主持国家自然科学基金面上项目 2 项、北京市自然科学基金面上项目。出版学术专著 2 部，发表学术论文 30 余篇。

何聪芬，教授，博士，主要研究方向为皮肤分子生态学与化妆品生物技术。主持纵向科研项目 6 项，作为主研人参加并完成国家自然科学基金项目 1 项。主持北京市教委纵向课题和参加 973 计划、863 计划课题。合作主编著作 2 部，获批国家专利 8 项，国外专利 1 项。在美国国立生物技术信息中心 (The National Center for Biotechnology Information (NCBI)) 注册新基因 6 个。参编著作 2 部。发表学术论文 40 余篇。