宏转录组测序结题报告

客　户：demo

客户单位：demo

项目编码：demo

销　　售：demo

技术支持：demo

您值得信赖的基因组学与蛋白质组学技术合作伙伴

前言

1. 微生物学宏时代

2016 年 5 月，时任美国总统奥巴马宣布，花费数亿美元启动“国家微生物组计划”，目的旨在研究人体及各种不同生态系统中的微生物组。因为对微生物组的研究不仅仅局限于人类的身体，地球上的土壤、海洋和环境等各处都有自己的微生物“居民”，科学家们将要尝试对各种形式的微生物群落进行研究。这项庞大的计划预计会完成三个目标：支持研究、开创新技术、让更多的人投入到微生物组研究中来。

曾经于 1998 年率先提出宏基因组（Metagenome）概念的 Jo Handelsman 博士，目前任职于白宫科学与技术政策办公室并担任副主任。这位著名的耶鲁大学微生物学家 Microbiologists 兴奋地说：“我们期盼‘国家微生物组计划’促进这个重要领域取得进展，给我们的星球及其居住者带来可观效益。如果有一件事我们可以确信，那就是微生物虽小，但它们的影响巨大。”

环境修复、人体肥胖、糖尿病、细菌耐药性、基因编辑等一系列问题本质上都与微生物息息相关。这些毫不起眼的小家伙们在我们的日常生活中扮演着极其重要的作用，而我们才对它们的了解不到 2%。

你能想象在几万米的高空中细菌是如何生存的吗？几千米深的海洋以及火山口的细菌又是如何在这种极端的环境中度过的？石油中的微生物又是如何发挥它们的作用？这些问题都说明，微生物在我们的地球上无处不在，而我们对它们的了解还仅仅只是开始。

大家对于青霉素的故事已经耳熟能详，但是接下来的故事则大大推进了现代分子生物学的发展，那就是 TAQ 酶的发现。TAQ 酶和 PCR 技术并称上世纪 80 年代分子生物学双子星，改写了生物研究的历史。这种具有热稳定性的 DNA 聚合酶，起初是一位台湾籍科学家钱嘉韵从黄石公园的嗜热菌中获得灵感，最终成功分离出了 TAQ 聚合酶。

时间再向后推进 10 年，日本科学家在大肠杆菌的基因序列中发现了短回文重复序列，同时西班牙科学家 Mojica 博士在研究地中海嗜盐菌的时候也发现了这种现象，并把这种现象命名为 clustered regularly interspaced short palindromic repeat sequences 即成簇的规律性间隔的短回文重复序列，简称 CRISPR。

让我们再把时间的指针拨向 2013 年，当年 Mojica 博士可能也没想到，CRISPR 如今已经风靡全球。在 George Church、张锋、Jennifer Doudna 等人的推动下，这项强大的基因编辑工具可以以低廉的成本和简便的操作对任何生物的基因组进行高效编辑。韩春雨更是从嗜盐杆菌性中发现了 NgAgo 蛋白，开发了全新的 gDNA 基因编辑技术。

关于微生物的故事还有很多，只不过许多微生物学家苦于工作环境的简陋和研究工具的匮乏，他们的研究进程缓慢而又低效。研究机构之间各自为阵造成了信息不畅，最终产生不必要的浪费。许多国家在意识到这一点后，纷纷启动了各式各样的微生物计划如人类肠道微生物组计划等。例如微生物学家通过对人类肠道的菌群进行宏基因组测序（Metagenomics Sequencing）和宏转录组测序（Metatranscriptome Sequencing）后，发现人类的肥胖与这些菌群有着密切的关系。以上种种庞大的微生物组学计划，不仅有利于研究者们之间的信息交流，也为微生物学家们提供了大量的新工具和新方法。

所以微生物学在过去的几十年内发生了翻天覆地的变化。迈入 21 世纪后，生物芯片技术（microarray）的发展，大大加快了微生物学家的研究进程。不过生物芯片还是存在一定的技术瓶颈，曾经以芯片业务起家的 Illumina 于 2006 年收购 Solexa 平台后加速了高通量测序的产业布局，其他测序巨头紧随其后推出了自己的测序平台，如罗氏的 454 测序平台、华大基因的 BGI-CG 测序平台和 Life Technology 的 Ion Torrent 等。这些技术的出现可以帮助微生物学家更好地去研究对于那些无法在实验室分离培养的微生物。也就是说，生物芯片和高通量测序技术避开了微生物分离培养的过程，扩展了微生物资源的利用，为微生物的研究提供了有效的工具。微生物学家们迎来了一个全新的黄金时代

2. 宏转录组横空出世

二代测序技术（Next Generation Sequencing, NGS）也被称作高通量测序技术，这是相对一代测序技术（Sanger Sequencing）而言的。目前最主流的二代测序平台是 Illumina 所生产的测序仪，包括 MiSeq 系列、HiSeq 系列、NextSeq 系列等。另外的还包括罗氏公司的 454 测序仪（已关闭）、华大基因的 BGI-CG 测序仪以及 Life Technology（已被 Thermo Fisher 收购）的 Ion Torrent 等。

但是 NGS 本身存在一定的技术缺陷，虽然测序通量比较高，但是只能读取几十个到几百个碱基长度的序列。所以较长的序列会被打断成小的片段连接上接头序列，再进行上机测序。之后下机的数据一般也是几十个到几百个碱基长度的序列片段，也叫 reads。之后这些 reads 通过与参考基因组的比对或是 De Novo 拼装，组装成更长的 contigs，之后是 scaffold。这三者之间的关系是：

（1）Reads：测序仪产出的序列就是 reads

（2）Contigs：将很多 read 根据 reads 之间的 overlap 区拼接在一起拼出的片段称作 contigs

（3）Scaffold：根据先后顺序的 contigs 组成的长序列称作 scaffold

所以说 NGS 技术已经对功能基因组学等基础学研究产生了巨大影响。微生物学家们在 NGS 技术的帮助下相继发表了许多高质量的研究成果，包括但不仅限于宏基因组（metagenomics）、16s rDNA 以及宏转录组（metatranscriptome）等。

目前任职于白宫科学与技术政策办公室的 Jo Handelsman 博士，曾经于 1998 年首次提出了宏基因组（Metagenome）的概念。宏基因组泛指野生样品中所有微小生物总的 DNA，主要研究对象包括了细菌、真菌、病毒等，以研究特定生境中微生物群落结构、潜在代谢功能、进化关系及对环境因子响应机制为主。

宏基因组概念的出现颠覆了传统的微生物学研究，配合 16s rDNA 的测序结果，宏基因组测序为微生物学家们提供了大量的宝贵信息，包括微生物群落（colony）的遗传信息和每个种群（population）的相对丰度等。

然而，在一个相对复杂的微生物生态系统（complex microbial ecosystem）中，如何研究微生物群落中不同种群之间在功能上的相互作用，宏基因组测序显得束手无措。因为当外界的环境发生变化时，这些微生物群体对环境的应答也会发生变化，随后其基因在 mRNA 水平的表达也会大有不同。

Velculescu 博士曾于 1997 年首次提出了转录组（Transcriptome）的概念，即特定细胞在特定时间内表达所有的 mRNA 的总和。后来转录组的概念逐渐延伸到动植物乃至原核生物。当转录组的概念被应用到微生物群落时，衍生出了宏转录组（Metatranscriptome）这一新的概念。

所以宏基因组测序是基于 DNA 水平，而宏转录组是基于 RNA 水平也就是转录层面。对于研究基因调控的科学家来说，转录水平往往能够更直观地了解整个微生物群落内不同种群基因的表达水平。

很多微生物无法在实验室分离培养，而宏转录组测序无需对微生物分离培养就可以对样品中的微生物群落进行全面检测。另外宏转录组测序的出现填补了宏基因组测序的研究空白。近年来，RNA-Seq 技术的发展大大推动了转录组学的研究，pubmed 和 scholar 上关于转录组学的学术论文每年呈指数型增长。基于单物种（a single species）的转录组测序研究对象主要为动植物，而宏转录组的研究对象则是自然界相对复杂的微生物群体。

由于自然环境下，不同样本中微生物种群在数量上差异巨大，分析这些复杂的转录数据为生物信息分析（bioinformatics analysis）提出了新的挑战。首先在上机测序前，真核生物的 mRNA 由于富含 PolyA 可以用磁珠法进行富集，而细菌等原核生物由于不含 PolyA，一般会设计相应的探针去除 total RNA 中的多余序列如宿主 RNA、rRNA 等。之后富集的 mRNA 被打断成只有几十 bp 的小片段，这是由于 illumina 二代测序平台技术瓶颈造成的，只能读取 75bp 到 500bp 左右长度的序列。这些几十 bp 到几百 bp 的序列被称作 reads。生物信息分析人员会把下机数据中的 reads 比对到相关的基因组参考序列上或者是注释完整的宏基因组序列。但是不同样品和不同参考序列本身存在遗传距离大（larger evolutionary distance）和序列多样性（extensive sequence diversity）等问题，这会大大降低比对效率，产生许多错误的比对结果。

基于 de Brujin 算法的 De Novo 组装策略可以克服上述的问题。这个算法命名规则来自于荷兰数学家 Nicolaas Govert de Bruijn。De Novo 组装可以把 reads 拼成更长的 contigs，因此能够注释更多有用的基因信息。这种方法尤其适用于未知的微生物群落中那些表达基因的注释。接下来，这些新拼接完的 contigs 作为目标序列（也可以理解为参考序列），用于 mRNA-reads 比对，然后得到基因（转录本）的表达量。最后得到拼接完的转录本后，后续可以对这些转录本进行 GO 和 KEGG 注释，阐释其生物学功能。

当然，受限于二代测序本身的技术缺陷和 De Novo 拼装的结果误差，分别验证后期感兴趣的单个基因必不可少，荧光定量 PCR 成为了验证测序结果是否可靠的常规方法之一。

项目信息

1. 样品信息

样本来源：RNA;

样本形式：Total RNA

2. 数据库信息

使用数据库	网页链接	版本/日期
NCBI NR	ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz	2016.07.12
Gene Orthology (GO)	http://geneontology.org/download	2016.06
KEGG	http://www.kegg.jp/kegg/download/	2016.05
CAZy	http://www.cazy.org/	2016.04

CARD	https://card.mcmaster.ca/download	2016.06.16
eggNOG	http://eggnogdb.embl.de/	2015.10

3. 数据分析程序

分析项目	软件名称	版本或日期
去除接头	Cutadapt	1.9
去除低质量序列	Fqtrim	0.95
数据质控	FastQC	0.10.1
序列组装与拼接	Trinity	2.2.0
序列比对与功能注释	blast	2.25

项目整体流程

1. 实验流程

（1）样本收集与准备

从土壤、肠道、海水等环境中采集实验样本，进行一些前期处理后，用于后续的 RNA 提取工作。具体的提取方案根据老师的样本来源出具个性化的样本提取方案。

（2）测序文库构建

总 RNA 提取使用的是 TRizol Reagent Kit。由于样本中存在大量的核糖体 RNA（rRNA），使用去核糖体试剂盒 Ribo-Zero rRNA Removal Kit 去除总 RNA 中的 rRNA 以及其他宿主 RNA 序列。之后将 mRNA 打断成段片段，构建上机的 cDNA 文库。对 cDNA 文库 PCR 扩增后，进行上机测序，测序平台为 Illumina HiSeq™ 4000，测序模式为双端测序 pair-end。

2. 生物信息分析流程

（1）De novo 从头拼装、功能注释和物种分类

刚下机的原始数据称为 Raw Data，在分析之前先进行预处理。首先使用 Cutadpter 和 Fqtrim 过滤下机原始数据中不合格的 reads，包括接头污染序列、低质量序列等。使用 FastQC 对过滤完的 CleanData 进行质检。使用 Trinity 分别对每个样本的 CleanData 进行 de novo 拼装，并把每个样本拼装完的结果进行整合。整合完的拼装结果用 CD-HIT-EST 去冗余，得到 Unigene 集合。

拼接获得的 Unigene 集合，与 5 个数据库（NR、GO、KEGG、CAZy、CARD）中的序列进行比对，取阈值 e≤1e-5，通过序列相似性进行功能注释。序列相似性主要使用 Blast2.25 进行比对运算。另外我们对 blast 在本地算法上进行了优化，比传统的比对速度提高了几千倍，可以大大缩短比对时间。

（2）差异表达分析

在宏转录组测序项目中，使用 TPM（TranscriptsPer Million）来统计基因在不同样本中的表达丰度。将 Trinity 组装完的 Unigene 作为参考基因序列，把每个样本中的有效数据比对到参考序列得到 TPM 值。同一个基因或转录本在不同样本之间的表达量是否有差异一般使用自主研发的软件包 ACGT101_Metatrascriptome 进行计算，之后对差异表达的基因进行 GO 和 KEGG 聚类富集分析与注释，使用的序列比对注释软件为 ACGT101_Metatrascriptome 中的 Functional_Enrichment。

数据处理

1. 样本采集和分组信息

#SampleID	Group
CON001	control
CON002	control
CON003	control
DLF001	stage1
DLF002	stage1
DLF003	stage1
T2D001	stage2
T2D002	stage2
T2D003	stage2

结果路径：summary/1_RawData/sample_map.xlsx

2. 测序序列统计与质控

测序产生的下机原始数据（raw data）需要进行预处理，使用 Cutadpter 和 Fqtrim 过滤掉不需要的序列后得到有效数据（clean data），才能进行下一步分析，具体处理步骤如下：

（1）去除质量值≤10 的低质量 reads（Q≤10 的碱基数量占整个 read 的 20%以上）；

（2）去除含 N 的碱基比例超过 5%的 reads；

（3）去除接头污染序列（默认 15 bp 的 overlap，设置为 15）；

（4）宿主序列（默认超过 90%序列相似）；

Sample	Raw Data		Valid Data		Valid%	Q20%	Q30%	GC%
	Read	Base	Read	Base				
CON001	44142248	3.97G	42363994	3.72G	95.97	98.62	94.91	46.92
CON002	47793804	4.30G	44667528	3.50G	93.46	96.18	88.40	44.07

CON003	45268384	4.07G	42374960	3.32G	93.61	96.06	88.24	45.62
DLF001	33230154	2.46G	30099172	2.18G	90.58	97.74	93.16	40.95
DLF002	20447208	1.51G	18137438	1.30G	88.70	96.95	91.76	43.91
DLF003	26964088	2.00G	24433606	1.76G	90.62	97.28	93.38	44.82
T2D001	48251116	4.34G	45573628	3.98G	94.45	98.24	92.97	46.03
T2D002	51845286	4.67G	48497638	4.24G	93.54	98.25	92.08	43.84
T2D003	58333542	5.25G	54462066	4.74G	93.36	98.13	91.56	44.15

结果路径：summary/2_CleanData/demo_data_stat.xlsx

表格参数说明：

Term	**注释**
Sample	样本名
Raw Data Reads	原始下机数据的 reads 数
Valid Data Reads	有效数据的 reads 数
Valid Ratio%	有效 reads 所占比例
Base	数据量大小
Valid Ratio%	处理后数据（Valid）与原始数据（Raw）比值，百分比表示
Q20%	质量值≥20 的碱基所占比例（测序错误率小于 0.01）
Q30%	质量值≥30 的碱基所占比例（测序错误率小于 0.001）
GC content%	GC 含量所占的比例

3. 序列组装

获得有效数据（Clean Data）后，使用 Trinity 分别对每个样本进行序列的组装拼接。拼装前先对 Clean Data 中的 rRNA、tRNA 等序列进行过滤。之后对多个样本的拼装结果进行聚类整合，使用 CD-HIT-EST 构建非冗余 Unigene 集合（默认设置值为 0.95）。

3.1 单个样本组装及多个样本聚类

样本组装结果统计报告：

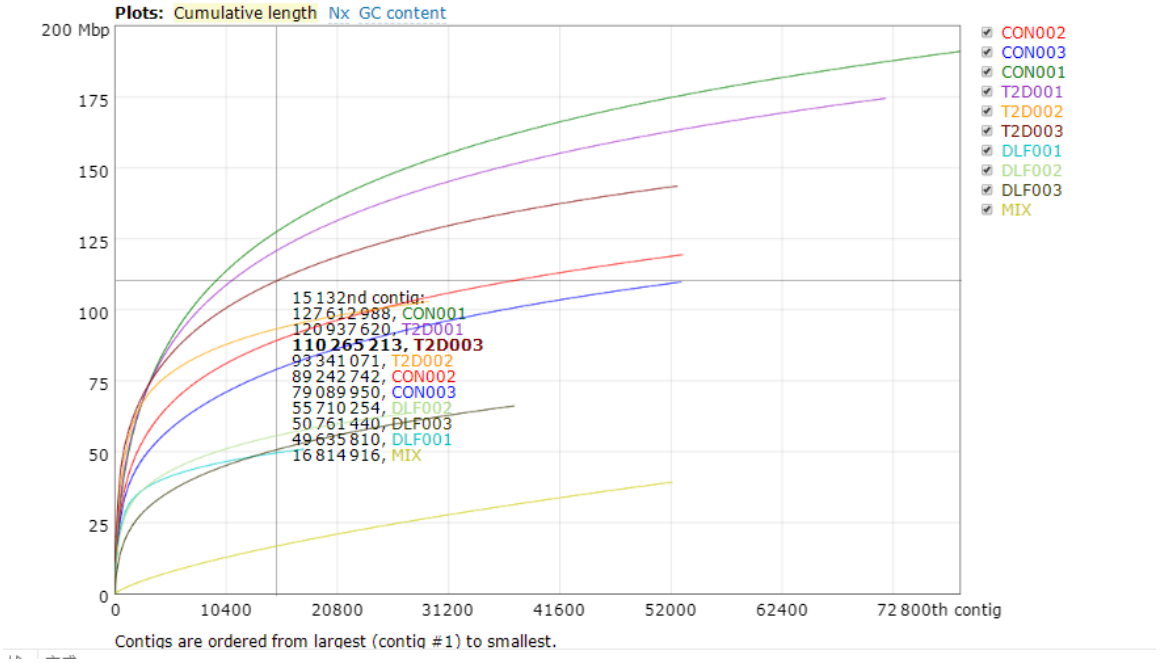
QUAST report

17 October 2016, Monday, 18:22:02

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs.)

Worst Median Best ☑ Show heatmap

Statistics without reference	CON002	CON003	CON001	T2D001	T2D002	T2D003	DLF001	DLF002	DLF003	MIX
# contigs	53 126	52 974	79 055	72 076	29 286	52 597	17 588	27 148	37 291	52 139
# contigs (>= 0 bp)	53 126	52 974	79 055	72 076	29 286	52 597	17 588	27 148	37 291	52 139
# contigs (>= 1000 bp)	23 283	24 119	38 656	34 466	14 619	27 149	7477	12 324	15 544	7274
# contigs (>= 5000 bp)	3776	2806	6943	5989	3011	4257	1491	1810	1897	3
# contigs (>= 10000 bp)	1604	1265	2919	2583	1744	1721	909	927	791	0
# contigs (>= 25000 bp)	509	489	755	729	747	638	365	313	210	0
# contigs (>= 50000 bp)	195	161	224	245	310	320	135	115	53	0
Largest contig	390 458	336 082	412 492	512 987	533 794	1 011 255	281 712	301 536	188 330	5529
Total length	119 434 244	109 823 904	191 041 218	174 512 026	102 907 377	143 614 858	50 938 783	63 429 621	66 148 880	39 306 468
Total length (>= 0 bp)	119 434 244	109 823 904	191 041 218	174 512 026	102 907 377	143 614 858	50 938 783	63 429 621	66 148 880	39 306 468
Total length (>= 1000 bp)	98 997 930	89 909 653	163 384 221	148 637 576	92 837 178	125 915 666	44 041 977	53 139 240	51 177 385	9 914 453
Total length (>= 5000 bp)	60 740 459	48 756 680	99 433 565	91 844 821	69 860 165	79 956 110	32 646 827	33 363 917	25 550 762	15 816
Total length (>= 10000 bp)	45 814 989	38 190 333	71 586 671	68 228 548	61 022 049	62 631 953	28 533 756	27 281 828	17 931 012	0
Total length (>= 25000 bp)	29 276 532	26 068 184	39 407 618	40 399 560	45 190 344	46 978 106	19 995 525	17 783 149	9 191 155	0
Total length (>= 50000 bp)	18 429 879	14 466 138	21 344 012	23 917 611	29 665 677	35 883 663	12 208 854	10 806 749	3 805 223	0
N50	5223	3731	5480	5699	18 417	6804	14 471	6022	2805	730
N75	1436	1309	1667	1639	2866	1877	2091	1470	1080	589
L50	3576	4247	6196	5127	1041	2853	654	1509	3957	18 880
L75	15 363	17 500	22 910	20 427	4989	13 714	3278	7690	14 036	33 969
GC (%)	43.68	44.61	47.13	46.34	44.38	45.06	42.85	43.95	43.69	45.65
Mismatches										
# N's	0	0	0	0	0	0	0	0	0	0
# N's per 100 kbp	0	0	0	0	0	0	0	0	0	0



结果路径：summary/3_Assembly/assembly_stats.html

3.2 非冗余序列集合构建

使用 CD-HIT-EST 对多样本聚类结果构建非冗余 Unigene 集合（默认设置值为 0.95）。

结果路径：summary/4_GenePredict/1_Unigenes_CDS.fa

4. 基因表达水平分析

在宏转录组测序中，我们使用 TPM（TranscriptsPer Million）来统计基因在不同样本中的表达丰度。将 Trinity 组装完的 Unigene 作为参考基因序列，把每个样本中的有效数据比对到参考序列得到 TPM 值。同一个基因或转录本在不同样本之间的表达量是否有差异我们使用我们自主研发的软件包

ACGT101_Metatrascriptome 进行计算，之后对差异表达的基因进行 GO 和 KEGG 聚类富集分析与注释，使用的序列比对注释软件也是 ACGT101_Metatrascriptome。

4.1 各样本中基因的表达量 TPM 值信息表

Unigene_ID	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
Unigene1	8.12	0	0	0	0	1.26	0	0.11	0
Unigene2	5.10	0	0.12	0	0	0.85	0	0	0
Unigene3	5.97	0.08	1.12	0	0	0	0	0	0
Unigene4	8.37	0.02	0	0.03	0.02	0.14	0.03	0.01	0
Unigene5	1.86	0	0.51	0	0	0	0.07	0.04	0
Unigene6	1.02	0	0.43	0	0	0	0	0	0
Unigene7	4.44	0.01	0.46	0	0.06	0.12	0	0.01	0
Unigene8	5.68	0	1.30	0	0	0	0	0.05	0
Unigene9	5.33	0.07	1.05	0	0	0.07	0	0	0
Unigene10	3.58	0	0.46	0	0	0.21	0.05	0	0
Unigene11	2.95	1.77	0.74	56.39	14.80	0	1.68	3.58	0.14
Unigene12	8.00	0	0	0	0	0.47	0	0	0
Unigene13	3.29	2.84	1.50	9.92	2.72	0.11	0.47	0.41	24.25
Unigene14	2.33	0.12	0.38	0	0.44	0.11	0.06	0	0
Unigene15	2.37	0	0.44	0	0.86	0.22	0	0	0
Unigene16	2.46	0.04	0.29	0	0.97	0.21	0	0	0
Unigene17	2.21	0.02	0.28	0	0.55	0.11	0	0	0
Unigene18	3.58	0	0.81	0	0	0.32	0.03	0	0.06
Unigene19	3.16	0	0.58	0	0	0.37	0	0	0
Unigene20	2.70	0.10	0.36	0	0	0.19	0	0.14	0.04
Unigene21	10.41	0	0	0	0	0.37	0	0.06	0
Unigene22	3.28	0	0.77	0	0	0	0	0	0
Unigene23	1.19	0	0	0	0	0	0	0	0
Unigene24	1.70	0	0	0.10	0	0.06	0	0	0
Unigene25	2.26	0	0	0	0	0	0	0	0
Unigene26	3.15	0	0	0	17.39	0	0	0	0
Unigene27	2.88	0	0	0	17.98	0.13	0	0	0
Unigene28	2.38	0	0	0	19.81	0	0	0	0
Unigene29	2.74	0	0.06	0	19.88	0.11	0	0	0
Unigene30	2.91	0.49	0	0	0	0	0	0	0
Unigene31	9.36	0	0	0	0	0.42	0	0	0

Unigene32	3.24	0	0	0	0	0	0	0	0
Unigene33	1.51	0	0	0	0	0	0	2.95	0
Unigene34	1.47	0	0	0	0	0.04	0	4.62	0
Unigene35	1.64	0	0	0	0	0.19	0	4.66	0
Unigene36	1.99	0.02	0.06	0.03	0	0.11	0	0	0.02
Unigene37	2.74	0	0	0	0	0	0	0	0
Unigene38	2.00	0	0	0	0	0	0.04	0	0
Unigene39	2.31	0	0	0	0	0	0	0	0
Unigene40	6.98	0	0	0	0	0.52	0.06	0	0
Unigene41	2.08	0	0.08	0	0	0	0	0	0
Unigene42	2.38	0.04	0.16	0	0.22	0.13	0.02	0	0
Unigene43	2.35	0.12	0.25	0	0.59	0.34	0.03	0.11	0
Unigene44	2.40	0.28	0.67	0.05	0.89	0.35	0	0	0
Unigene45	2.04	0.20	0.10	0	0.71	0.18	0	0	0
Unigene46	2.23	0	0	0.07	0	0.61	0	0.04	0
Unigene47	3.40	0	0.14	0	0	1.30	0.04	0	0
Unigene48	4.00	0.07	0.15	0.02	0.03	0.72	0.01	0	0
Unigene49	3.28	0	0	0	0	0.60	0	0	0
Unigene50	1.26	0.06	0.13	0	0.60	0.17	0	0	0
Unigene51	8.12	0	0	0	0.04	0.24	0	0	0
Unigene52	1.71	0.05	0.22	0	0.38	0.10	0	0	0
Unigene53	1.15	0	0.06	0	0.32	0.10	0.02	0.02	0
Unigene54	1.91	0.15	0.16	0	0.32	0.04	0	0	0.02
Unigene55	1.98	0.67	0.26	0	2.81	1.40	1.63	0	0.38
Unigene56	2.60	0.92	0.27	0	0	1.39	0.04	0	0
Unigene57	7.83	0	0	0	0	0.45	0	0	0
Unigene58	2.53	0	1.37	0	0	0	0	0.03	0
Unigene59	1.45	0	0	0.08	4.47	0	0	0	0.41
Unigene60	1.53	0	0	0.06	7.09	0.04	0	0	0.53
Unigene61	0.97	0	0	0	8.05	0	0	0	0.71
Unigene62	1.63	0.01	0.01	0	0.25	0	0	0	0.02
Unigene63	1.63	0	0	0	0	0	0	0	0
Unigene64	1.24	0.69	0.13	0	0	1.00	0.82	0	0.30
Unigene65	7.81	0	0	0	0	0.43	0	0	0
Unigene66	1.58	0.53	0	0	0	0	1.36	0	0
Unigene67	4.89	0.64	2.51	0	0	3.17	0.02	0	0
Unigene68	3.54	0.30	1.45	0	0	1.70	0	0	0

Unigene69	3.11	0.05	0.72	0	0.04	0.89	0.02	0	0.01
Unigene70	1.33	0.58	0.73	0.49	0.28	0.55	0.67	2.25	0
Unigene71	2.29	1.95	1.67	1.05	0	1.00	1.00	4.50	0
Unigene72	7.64	0	0	0	0.09	0.73	0	0	0
Unigene73	1.54	0.26	0	0.02	0	0.03	0	0	0
Unigene74	1.15	0.16	0.33	0	0.33	0.09	0.32	0.28	0
Unigene75	1.34	0.18	0.12	0	0.87	0	0.27	0.21	0
Unigene76	5.91	1.24	1.30	0.13	1.29	1.83	0	0	0.22
Unigene77	6.81	0.76	1.13	0	0.77	2.51	0.83	0	0
Unigene78	4.43	0.45	0.78	0.11	0	2.93	0.21	0	0
Unigene79	4.52	0.65	0.63	0	0.42	2.45	0.28	0	0
Unigene80	4.44	0.16	0.68	0.03	0.10	0.52	10.68	0.04	0
Unigene81	4.14	1.70	0.92	0.09	1.14	3.20	0.99	0	0.15
Unigene82	4.58	0.31	0.65	0	0	2.00	0.12	0	0.07
Unigene83	3.68	0.72	0.48	0	0.16	2.20	0.37	0	0.16
Unigene84	2.31	0.08	0.47	0	0.18	1.83	0.07	0	0
Unigene85	1.59	0.07	0.17	0	0.15	0.07	0.02	0	0
Unigene86	1.24	0.08	0.06	0	0.06	0	0	0	0
Unigene87	1.12	0.02	0.14	0	0.21	0.04	0	0	0
Unigene88	2.81	0	0.52	0	0	0	0	0	0
Unigene89	2.17	0.15	0.33	0.06	0.23	0.14	0.05	0.04	0.03
Unigene90	5.99	0	0.91	0	0	0.23	0	0	0
Unigene91	1.80	0	0.39	0.02	0	0	0	0	0
Unigene92	1.80	0	0.44	0	0	0	0	0	0.03
Unigene93	2.01	0	0	0	0	0.76	0	1.33	0
Unigene94	2.44	0	0	0	0	0.72	0.02	1.65	0
Unigene95	0.86	0	0	0	0	0.38	0	0.75	0
Unigene96	4.04	0	0.14	0	0	0.66	0	0	0
Unigene97	4.26	0	0.25	0	0.06	0.70	0	0	0
Unigene98	3.39	0	0.13	0	0	0.70	0	0	0
Unigene99	4.27	0	0	0	0	0.47	0	0	0
Unigene100	5.50	0	0.24	0	0	0	9.44	0	0

结果路径： summary/4_GenePredict/3_Unigenes_abund.xlsx

结果路径： summary/4_GenePredict/3_Unigenes_abund.boxplot.png **表格参数说明：**

Term	Description

Sample	样本名
Exp gen	各样本所有 unigenes 表达数
Min	各样本所有 unigenes 表达量 TPM 值的最小值
1st Qu	各样本所有 unigenes 表达量 TPM 值的 1/4 值
Median	各样本所有 unigenes 表达量 TPM 值的中值
Mean	各样本所有 unigenes 表达量 TPM 值的平均值
3rd Qu	各样本所有 unigenes 表达量 TPM 值的 3/4 值
Max	各样本所有 unigenes 表达量 TPM 值的最大值
Sd.	各样本所有 unigenes 表达量 TPM 值的方差
Sum	各样本所有 unigenes 表达量 TPM 值的总和

4.2 基因表达密度值

结果路径：summary/4_GenePredict/3_Unigenes_abund.density.png

结果与分析

1. 物种与功能注释

1.1 物种分类注释

使用序列比对注释软件 blast2.25 将得到的 Unigene 集合与 NR 数据库进行比对，并通过 NR 库中对应的物种分类信息进行物种注释。接着使用物种对应的 Unigene 丰度总和计算该物种的丰度，并在界门纲目科属种各个分类学层次水平上统计物种在每个样品中的丰度。NR 数据库包含 PDB、SwissProt、GeneBank 等蛋白数据库中所有的非冗余蛋白序列信息。

（1）各样本基因表达量和分类信息

Unigene_ID	CON001	DLF001	T2D002	Taxonomy
Unigene1	8.12	0	0.11	d__Bacteria;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Akkermansiaceae;g__Akkermansia;s__Akkermansia_muciniphila
Unigene2	5.10	0	0	d__Bacteria;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Akkermansiaceae;g__Akkermansia;s__Akkermansia_muciniphila
Unigene3	5.97	0	0	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes;s__Alistipes_indistinctus
Unigene4	8.37	0.03	0.01	d__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium;s__Clostridium_sp._CAG:226
Unigene5	1.86	0	0.04	d__unclassified;p__unclassified;c__unclassified;o__unclassified;

Unigene6	1.02	0	0	f__unclassified;g__unclassified;s__unclassified d__unclassified;p__unclassified;c__unclassified;o__unclassified; f__unclassified;g__unclassified;s__unclassified

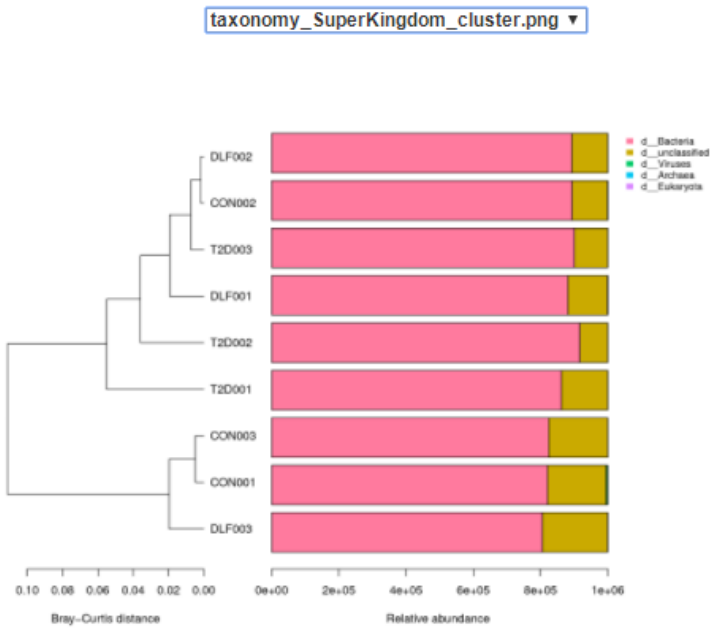
注：第一列是 Unigene 的 ID 号，后面几列分别是 Unigene 在各样本中的表达量和分类信息

结果路径：summary/5_TaxonomicProfiling/Unigenes_abund_taxonomy.xlsx

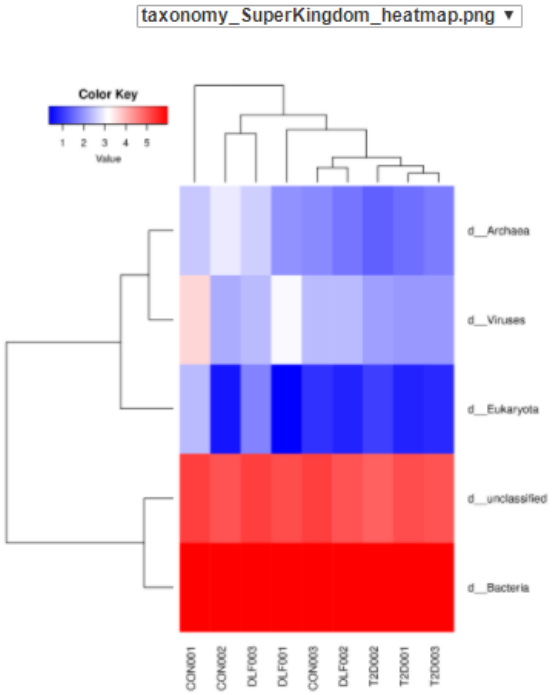
2）样本物种分布图

所有样本中，按照界门纲目科属种，一共 7 个层次对不同的物种分布作图。

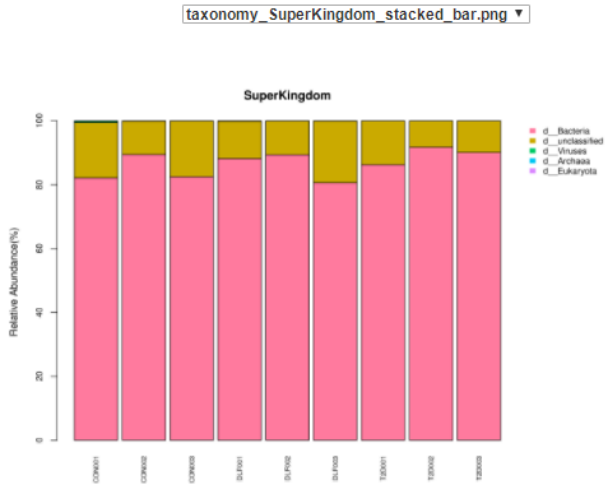
taxonomy_cluster 图：



taxonomy_heatmap 图：



taxonomy_stacked_bar 图：



1.2 GO 功能注释

GO（gene ontology）是基因本体联合会所建立的数据库，旨在建立一个适用于各物种的标准。这项语言词汇标准随着人们对基因的不断深入了解随时保持更新，用于对基因和蛋白质的功能进行限定和描述。

GO 提供了三类描述的系统定义方式，用于描述基因产物的功能，GO 的结构包括三个方面：Molecular Function、Biological Process 和 Cellar Component。在实际应用中存在同一个基因会存在于多个功能注释的情况，Molecular Function 描述的是在分子生物学上的活性如催化活性或结合活性，Biological Process 是由分子功能有序地组成的，具有多个步骤的一个过程，Cellar Component 则是指基因产物位于细胞器或基因产物组件中如核糖体蛋白酶体等。GO 对基因和蛋白注释阐明了基因产物和用

于定义它们的 GO 术语之间的关系，注释需要反映在正常情况下基因产物的功能、生物途径、定位等，而不包括其在突变或病理状态下的情况。

（1）基因 GO 功能注释信息表

Query	GO_hit	Gene	Symbol	GO_ID	GO_Term	GO_Function	GO_Level
Unigene3	UNIPROTKB|Q882J1	1184288	galM	GO:0004034	aldose 1-epimerase activity	molecular_function	7
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0004854	xanthine dehydrogenase activity	molecular_function	7
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0005829	cytosol	cellular_component	9
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0009115	xanthine catabolic process	biological_process	11
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0016903	oxidoreductase activity, acting on the aldehyde or oxo group of donors	molecular_function	5
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0050660	flavin adenine dinucleotide binding	molecular_function	6
Unigene10	UNIPROTKB|Q8A382	1074372	BT_3073	GO:0005829	cytosol	cellular_component	9
Unigene10	UNIPROTKB|Q8A382	1074372	BT_3073	GO:0017150	tRNA dihydrouridine synthase activity	molecular_function	6
Unigene12	UNIPROTKB|C4M5C6	3406093	EHI_121800	GO:0004139	deoxyribose-phosphate aldolase activity	molecular_function	7
Unigene12	UNIPROTKB|C4M5C6	3406093	EHI_121800	GO:0009264	deoxyribonucleotide catabolic process		

结果路径：summary/6_FunctionalProfiling/GO/GO.anno.xlsx

表格参数说明：

Term	Description
Query	Unigene 的 ID 号
GO_hit	Unigene 在 GO 数据库中比对上的注释信息
Gene	Unigene 在 GO 数据库中比对上的基因 ID
Symbol	Unigene 在 GO 数据库中比对上的基因名
GO_ID	Unigene 在 GO 数据库中对应的 GO ID
GO_Term	GO ID 对应的结构
GO_Function	GO ID 对应详细的注释信息

GO_Level	GO ID 对应的层次关系

（2）GO 各层级在各样本中基因表达丰度统计信息

a：GO Level 统计表部分结果

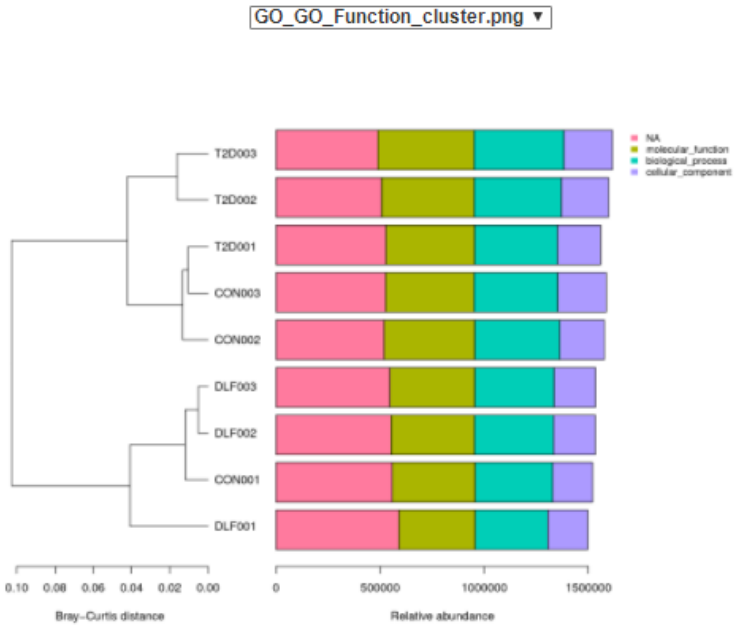
GO_Function	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
NA	558076.00	518763.58	526280.47	590087.74	551966.08	546769.90	527980.04	506058.52	489077.11
molecular_function	398257.07	435711.02	424434.02	366196.55	400993.07	407864.79	425565.78	445273.14	461554.33
biological_process	371932.83	407559.31	402452.75	348480.06	379252.60	380650.89	397955.58	418069.75	432148.93
cellular_component	191774.97	214180.43	232283.42	192358.34	201214.46	198757.60	206857.93	227464.48	231970.61

结果路径：summary/6_FunctionalProfiling/GO/1_GO_Level/GO_GO_Level_abund.xlsx

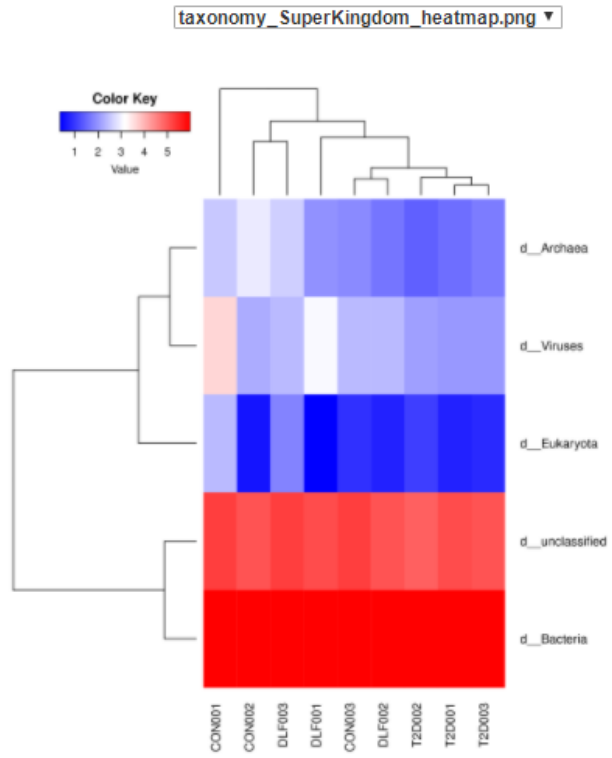
注：其他层次各样本中基因表达丰度部分统计信息见（summary/6_FunctionalProfiling/GO/）相应路径下

b：GO Level 统计图

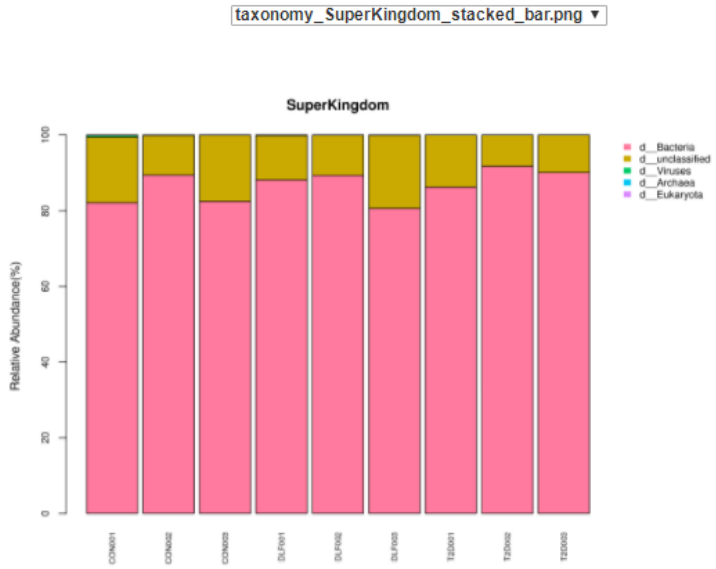
GO_taxonomy_cluster 图：



GO_taxonomy_heatmap 图：



taxonomy_stacked_bar 图:



1.2 GO 功能注释

GO（gene ontology）是基因本体联合会所建立的数据库，旨在建立一个适用于各物种的标准。这项语言词汇标准随着人们对基因的不断深入了解随时保持更新，用于对基因和蛋白质的功能进行限定和描述。

GO 提供了三类描述的系统定义方式，用于描述基因产物的功能，GO 的结构包括三个方面：Molecular Function、Biological Process 和 Cellar Component。在实际应用中存在同一个基因会存在于多个功能

注释的情况，Molecular Function 描述的是在分子生物学上的活性如催化活性或结合活性，Biological Process 是由分子功能有序地组成的，具有多个步骤的一个过程，Cellar Component 则是指基因产物位于细胞器或基因产物组件中如核糖体蛋白酶体等。GO 对基因和蛋白注释阐明了基因产物和用于定义它们的 GO 术语之间的关系，注释需要反映在正常情况下基因产物的功能、生物途径、定位等，而不包括其在突变或病理状态下的情况。

（1）基因 GO 功能注释信息表

Query	GO_hit	Gene	Symbol	GO_ID	GO_Term	GO_Function	GO_Level
Unigene3	UNIPROTKB|Q882J1	1184288	galM	GO:0004034	aldose 1-epimerase activity	molecular_function	7
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0004854	xanthine dehydrogenase activity	molecular_function	7
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0005829	cytosol	cellular_component	9
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0009115	xanthine catabolic process	biological_process	11
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0016903	oxidoreductase activity, acting on the aldehyde or oxo group of donors	molecular_function	5
Unigene4	UNIPROTKB|A5I5W5	5184562	xdhAC	GO:0050660	flavin adenine dinucleotide binding	molecular_function	6
Unigene10	UNIPROTKB|Q8A382	1074372	BT_3073	GO:0005829	cytosol	cellular_component	9
Unigene10	UNIPROTKB|Q8A382	1074372	BT_3073	GO:0017150	tRNA dihydrouridine synthase activity	molecular_function	6
Unigene12	UNIPROTKB|C4M5C6	3406093	EHI_121800	GO:0004139	deoxyribose-phosphate aldolase acti		

结果路径：summary/6_FunctionalProfiling/GO/GO.anno.xlsx

表格参数说明：

Term	Description
Query	Unigene 的 ID 号
GO_hit	Unigene 在 GO 数据库中比对上的注释信息
Gene	Unigene 在 GO 数据库中比对上的基因 ID
Symbol	Unigene 在 GO 数据库中比对上的基因名
GO_ID	Unigene 在 GO 数据库中对应的 GO ID
GO_Term	GO ID 对应的结构

GO_Function	GO ID 对应详细的注释信息
GO_Level	GO ID 对应的层次关系

（2）GO 各层级在各样本中基因表达丰度统计信息

a：GO Level 统计表部分结果

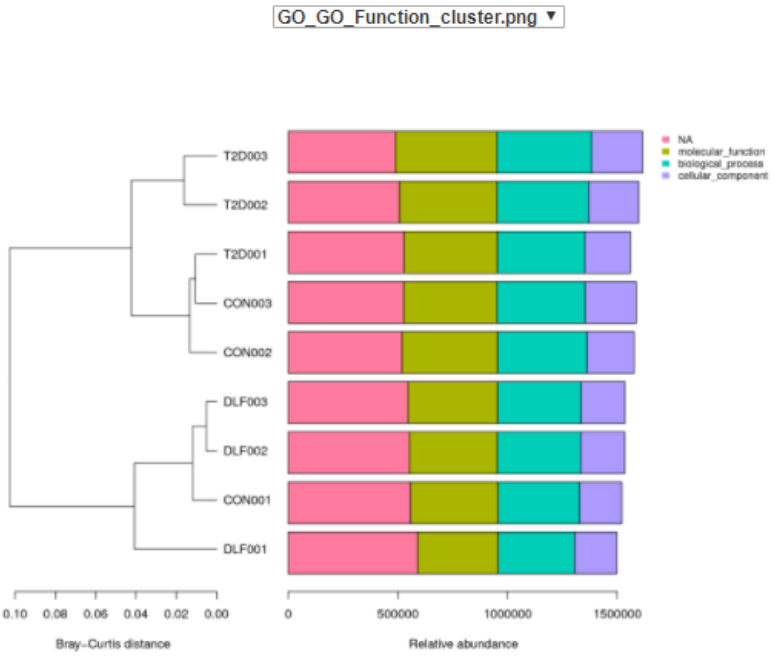
GO_Function	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
NA	558076.00	518763.58	526280.47	590087.74	551966.08	546769.90	527980.04	506058.52	489077.11
molecular_function	398257.07	435711.02	424434.02	366196.55	400993.07	407864.79	425565.78	445273.14	461554.33
biological_process	371932.83	407559.31	402452.75	348480.06	379252.60	380650.89	397955.58	418069.75	432148.93
cellular_component	191774.97	214180.43	232283.42	192358.34	201214.46	198757.60	206857.93	227464.48	231970.61

结果路径：summary/6_FunctionalProfiling/GO/1_GO_Level/GO_GO_Level_abund.xlsx

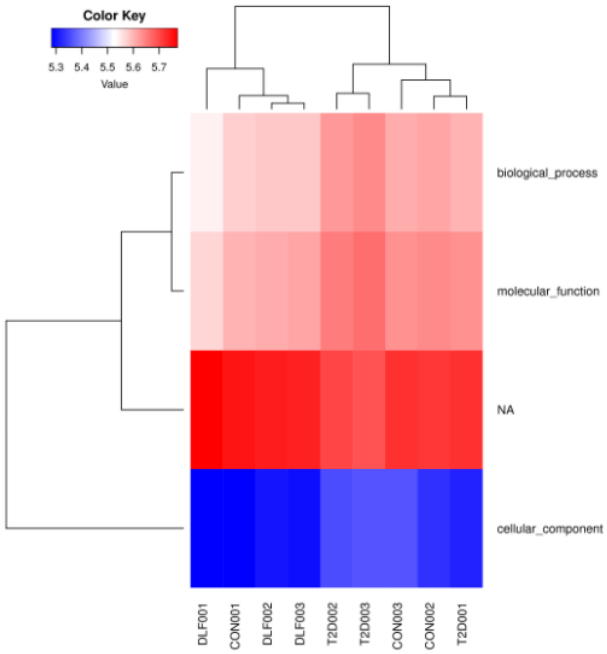
注：其他层次各样本中基因表达丰度部分统计信息见（summary/6_FunctionalProfiling/GO/）相应路径下

b：GO Level 统计图

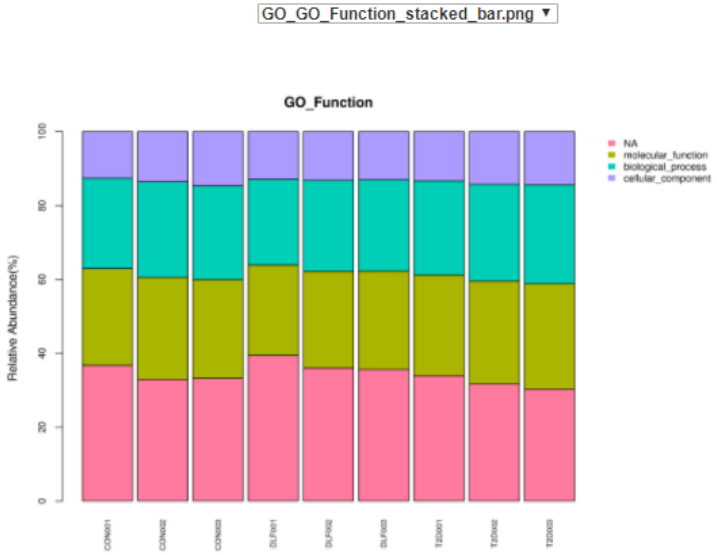
GO_taxonomy_cluster 图:



GO_taxonomy_heatmap 图:



GO_taxonomy_stacked_bar 图:



1.3 KEGG 注释

KEGG（Kyoto Encyclopedia of Genes and Genomes）全称是京都基因和基因组百科全书，是基因组破译方面的公共数据库，网址是 www.genome.jp/kegg。

基因组信息存储在 GENES 数据库里，包括完整和部分测序的基因组序列；更高级的功能存储在 pathway 数据库里，包括图解的细胞生物化学过程如代谢、膜转运、信号传递、细胞周期；KEGG 另一个数据库是 ligand，包含关于化学物质、酶分子、酶反应等信息。KEGG 提供的整合代谢途径（pathway）查询十分出色，包括碳水化合物、核苷、氨基酸等的代谢及有机物的生物降解，不仅提供了所有可能的代谢途径，而且对催化各步反应的酶进行了全面的注解，包含有氨基酸序列、PDB 库等。

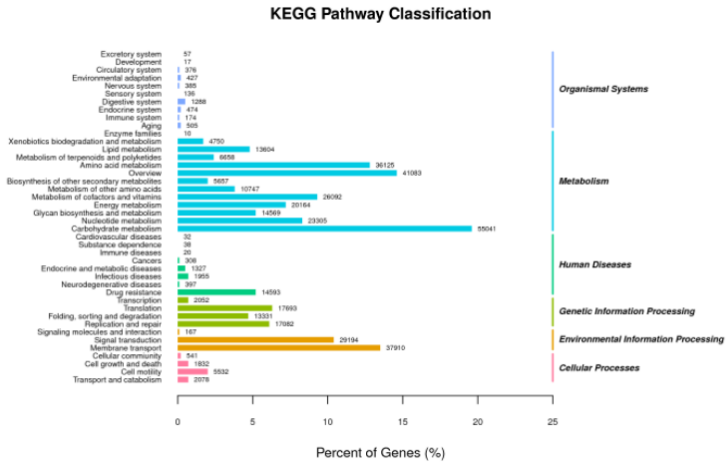
KEGG 中的 pathway 是根据相关知识手绘的，这里的手绘的意思可能是指人工以特定的语言格式来确定通路 各组件的联系;基因组信息主要是从 NCBI 等数据库中获得，除了有完整的基因序列外，还有没完成的草图;另外 KEGG 中有一个“专有名词”KO(KEGG Orthology)，它是蛋白质(酶)的一个分类体系，序列高度相似，并且在同一条通路上有相似功能的蛋白质被归为一组，然后打上 KO(或 K)标签。

（1）基因 KEGG 注释信息表

Query	KEGG_hit	GeneName	GeneDescription	KOEntry	KODescription	EC	PathwayEntry	PathwayDefinition	KEGGLevel1	KEGGLevel2
Unigene3	bcel:BcellWH2_00928	mro_2	Aldose 1-epimerase precursor	K01785	aldose 1-epimerase	EC:5.1.3.3	bcel00010	Glycolysis / Gluconeogenesis	Metabolism	Carbohydrate metabolism
Unigene3	bcel:BcellWH2_00928	mro_2	Aldose 1-epimerase precursor	K01785	aldose 1-epimerase	EC:5.1.3.3	bcel00052	Galactose metabolism	Metabolism	Carbohydrate metabolism
Unigene4	mta:Moth_1999	-	xanthine dehydrogenase	K00087	xanthine dehydrogenase molybdenum-binding subunit	EC:1.17.1.4	mta00230	Purine metabolism	Metabolism	Nucleotide metabolism
Unigene7	pep:AQ505_04660	-	cell division protein	K03587	cell division protein FtsI (penicillin-binding protein 3)	-	pep00550	Peptidoglycan biosynthesis	Metabolism	Glycan biosynthesis and metabolism
Unigene7	pep:AQ505_04660	-	cell division protein	K03587	cell division protein FtsI (penicillin-binding protein 3)	-	pep01501	beta-Lactam resistance	Human Diseases	Drug resistance

结果路径：summary/6_FunctionalProfiling/KEGG/KEGG.anno.xlsx

（2）KEGG 通路分类图



（3）KEGG 在 level1 层次各样本中基因表达丰度部分统计信息结果示例

a：KEGG Level 统计表部分结果

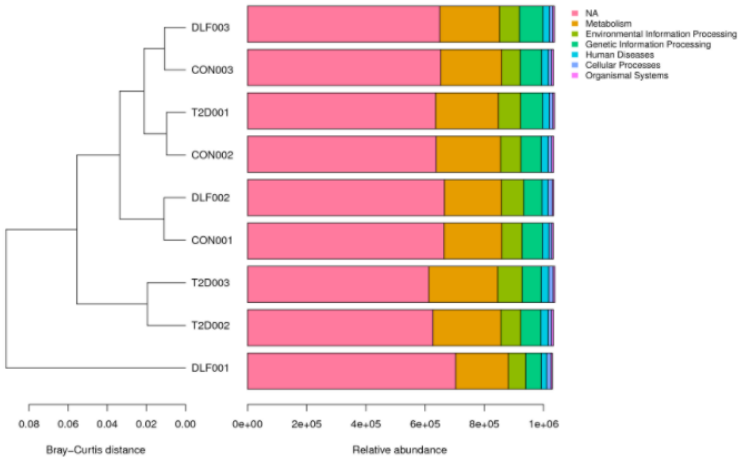
KEGGLevel1	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
NA	663681.05	637833.61	652495.23	703427.34	665048.37	650275.96	635357.91	626752.12	612588.46
Metabolism	195756.34	216969.28	206730.13	177773.62	193025.34	201007.58	212544.55	228831.92	232479.75
Environmental Information Processing	67914.22	69039.86	62287.70	58776.02	74653.46	67401.10	75016.56	67858.02	82723.59
Genetic Information Processing	68391.64	68170.81	72038.51	52325.62	61868.75	79105.84	73409.46	67207.47	64416.18
Human Diseases	22322.63	24141.36	23132.76	18875.31	21053.09	21960.37	23040.06	25592.27	25365.90
Cellular Processes	10330.48	12023.58	11160.52	13931.18	14146.65	11623.84	12380.52	10871.43	14866.83
Organismal Systems	4671.19	5868.44	5884.73	4756.15	4477.21	5390.05	5227.20	6412.98	5534.74

结果路径：summary/6_FunctionalProfiling/1_KEGGLevel1/KEGG_KEGGLevel1_abund.xlsx

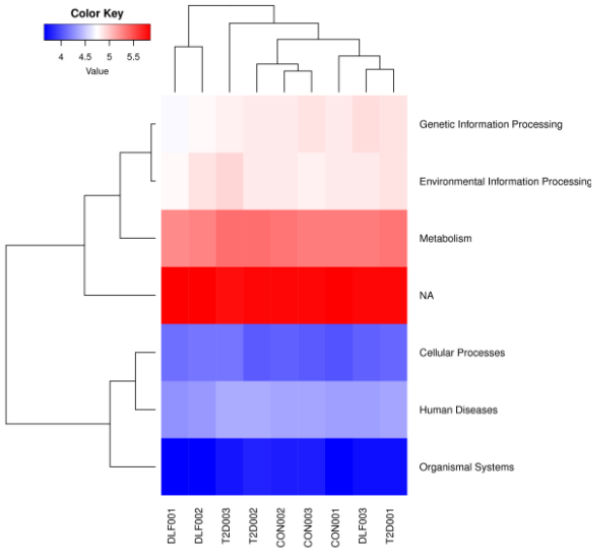
注：其他层次各样本中基因表达丰度部分统计信息见（summary/6_FunctionalProfiling/KEGG/）相应路径下

b：KEGG Level 统计图

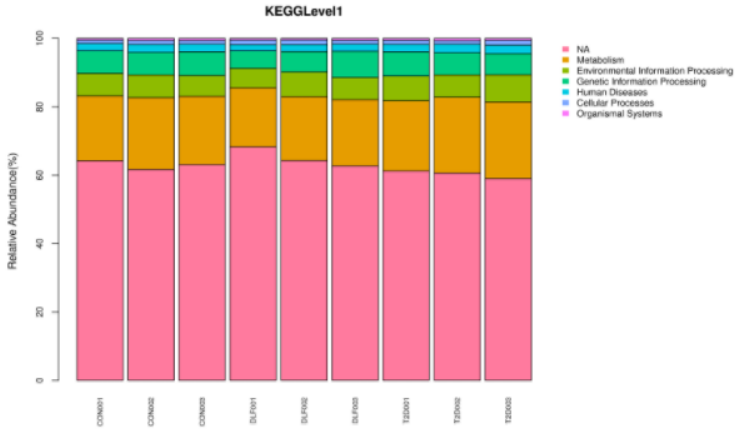
KEGG_taxonomy_cluster 图：



KEGG_taxonomy_heatmap 图:



KEGG_taxonomy_stacked_bar 图:



1.4 抗生素耐药基因注释

抗生素耐药基因数据库的先驱是 ARDB，整合了 NCBI 和 SwissProt 的一万多个细菌耐药基因信息，然而由于时间的关系，ARDB 自 2009 年后就再也没有进行过更新了，这时候 CARD 数据库横空出世接过了 ARDB 的接力棒，成为目前主流的细菌耐药性基因注释数据库之一。ARDB 官网也强烈推荐科研工作者使用 CARD。

CARD 全称是 Comprehensive Antibiotic Resistance Database，网址是 https://card.mcmaster.ca/ 。这个数据库目前仍然在不断地更新，所有数据均来自于志愿者的贡献。该数据库以 ARO（Antibiotic Resistance Ontology）为核心对耐性数据进行整理，以达到数据实时更新的效果。

使用序列比对注释软件 blast2.25 将基因序列与 CARD 数据库进行比对（E-value=0.0001），对 Unigene 进行对应的抗生素耐药性注释，然后使用抗生素耐性功能对应的 Unigene 在各样本的表达丰度，计算该抗生素耐药功能的丰度。

（1）基因抗生素耐药注释信息表

Query	CARD_hit	ARO_id	ARO_name	ARO_definition
Unigene70	gi|2769708|ARO:3000118|vgaB	ARO:3000118	vgaB	Vga(B) is an efflux protein expressed in staphylococci that confers resistance to streptogramin A antibiotics and related compounds. It is associated with plasmid DNA." [PMID:9427556, PMID:15728891]
Unigene75	gi|7110136|ARO:3002945|vanHF	ARO:3002945	vanHF	vanHF is a vanH variant in the vanF gene cluster" [PMID:15980329]
Unigene103	gi|AEX49906.1|ARO:3003583|PmrB	ARO:3003583	PmrB	Histidine protein kinase sensor Lipid A modification gene; part of a two-component system involved in polymyxin resistance that senses high extracellular Fe(2+)" [PMID:14507375, PMID:25006521]
Unigene255	gi|437315|ARO:3002987|bcrA	ARO:3002987	bcrA	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance" [PMID:7476193]
Unigene324	gi|99079563|ARO:3002881|lmrC	ARO:3002881	lmrC	lmrC is a chromosomally-encoded efflux pump that confers resistance to lincosamides in Streptomyces lincolnensis and Lactococcus lactis. It can dimerize with lmrD" [PMID:16958846]
Unigene381	gi|27461218|ARO:3002944|vanHD	ARO:3002944	vanHD	vanHD is a vanH variant in the vanD gene cluster" [PMID:16323116, PMID:12499162]
Unigene	gi|169633158|A	ARO:300077	adeF	AdeF is the membrane fusion protein of the

389	RO:3000777|adeF	7		multidrug efflux complex AdeFGH." [PMID:20696879]
Unigene390	gi|152938410|ARO:3000784|cmeB	ARO:3000784	cmeB	

结果路径：summary/6_FunctionalProfiling/CARD/Unigenes_CARD.xlsx

表格参数说明：

Term	Description
Query	Unigene 的 ID 号
CARD_hit	基因在 CARD 数据库中比对上的注释信息
ARO_id	ARO（Antibiotic Resistance Ontology）的 ID 号
ARO_name	ARO（Antibiotic Resistance Ontology）名
ARO_definition	ARO（Antibiotic Resistance Ontology）的解释

1.5 碳水化合物活性酶注释（CAZy）

碳水化合物活性酶（Carbohydrate-active enzymes，CAZyme）对地球上所有碳水化合物的合成、降解与修饰起重要作用，因此深入研究 CAZyme，对于了解微生物碳水化合物的代谢机制非常重要。

碳水化合物活性酶数据库（CAZy, www.cazy.org）是关于能够合成或者分解复杂碳水化合物和糖复合物的酶类的一个数据库资源，其基于蛋白质结构域中的氨基酸序列相似性，将碳水化合物活性酶类归入不同蛋白质家族。CAZy 数据库中包含了碳水化合物酶类的物种来源、酶功能 EC 分类、基因序列、蛋白质序列及其结构等信息。而随着宏基因组学测序和宏转录组测序的快速发展，CAZy 数据库中家族内序列数据量剧增，这为家族内进一步进行亚家族分类奠定了基础；而蛋白质家族内新一层精细分类的引入可提高亚家族中酶分子功能预测的准确度，进而可指导酶分子理性设计来提高特定功能酶组分设计的成功概率。使用序列比对注释软件 blast2.25 将基因序列与 CAZy 数据库进行比对（E-value=0.0001），获得 Unigene 对应的碳水化合物活性酶注释信息，然后使用碳水化合物活性酶对应 Unigene 的表达丰度总和，计算该碳水化合物活性酶的基因表达丰度。

（1）基因碳水化合物活性酶注释信息表

Query	CAZy_hit	NCBIAnno	EC	CAZyLevel1	CAZyLevel2
Unigene3	AEA20065.1	aldose 1-epimerase [Prevotella denticola F0289]	NA	Glycoside Hydrolases	GH43
Unigene7	AJQ27844.1	penicillin-binding protein, 1A family [Pelosinus fermentans JBW45]	NA	GlycosylTransferases	GT51
Unigene11	ADW67276.1	TonB family protein [Granulicella tundricola MP5ACTX9]	NA	Glycoside Hydrolases	GH13
Unigene47	ALU14107.1	glycoside hydrolase GH13 family [Eubacterium limosum]	NA	Glycoside Hydrolases	GH13

Unigene58	AMN56057.1	hypothetical protein ACP90_18880 [Labrenzia sp. CP4]	NA	Glycoside Hydrolases	GH23
Unigene70	AJA56769.1	teichoic acid ABC transporter ATP-binding protein [Lactococcus lactis subsp. lactis]	NA	Carbohydrate-Binding Modules	CBM50

结果路径：summary/6_FunctionalProfiling/CAZy/CAZy.anno.xlsx

表格参数说明：

Term	Description
Query	Unigene 的 ID 号
CAZy_hit	基因在 CAZy 数据库中比对上的注释信息
NCBIAnno	基因在 NCBI 中的注释信息
EC	碳水化合物活性酶 EC 的 ID 号
CAZyLevel1	CAZy 上的某一类碳水化合物活性酶的名称
CAZyLevel2	某一类碳水化合物活性酶下的具体分类

（2）CAZy Level1 和 Level2 在各样本中的基因表达丰度图

a：CAZy 统计表部分结果

CAZyLevel1	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
NA	905902.66	892071.92	903648.00	891956.00	899443.15	905287.06	899718.68	885868.13	873656.85
Glycoside Hydrolases	42004.74	51661.56	42439.18	49222.21	45833.89	40672.50	44828.93	53189.69	57615.25
GlycosylTransferases	33803.21	37535.06	35902.28	39643.99	36799.87	35693.89	36765.00	40953.78	44417.40
Carbohydrate-Binding Modules	16292.55	14794.58	14884.62	14191.22	14561.95	16944.97	15970.74	15146.25	18195.68
Carbohydrate Esterases	5139.98	6143.81	5950.92	6123.38	5470.03	5337.36	6025.87	6990.87	7816.65
Polysaccharide Lyases	1189.63	1672.80	1367.54	2136.12	1751.74	1319.23	1435.78	2053.09	2827.06
Auxiliary Activities	192.72	132.72	302.94	38.70	147.65	179.76	172.62	154.67	260.78

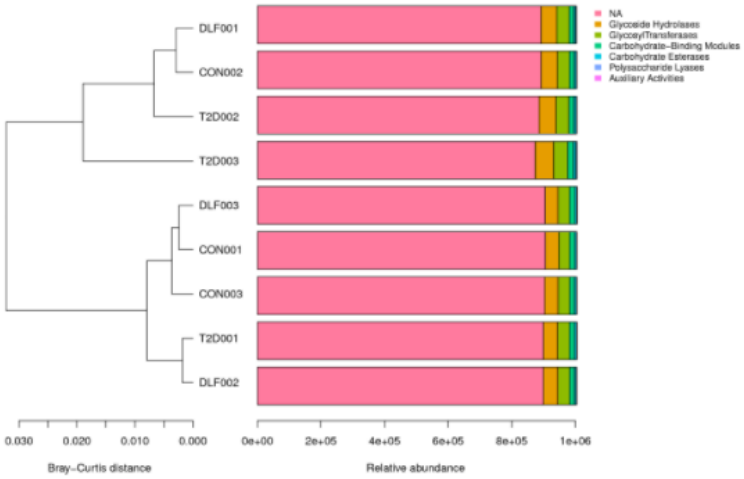
结果路径：summary/6_FunctionalProfiling/CAZy/1_CAZyLevel1/CAZy_CAZyLevel1_abund.xlsx

注：其他层次各样本中基因表达丰度部分统计信息见（summary/6_FunctionalProfiling/CAZy）相应路径下

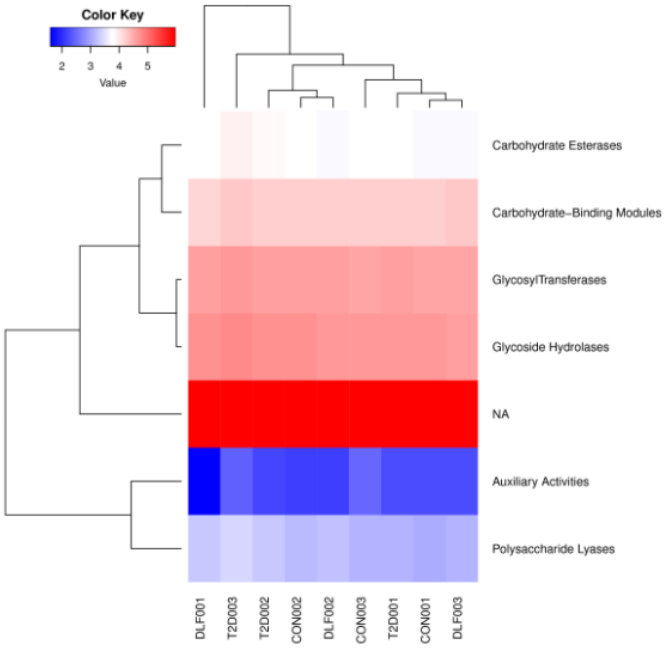
b：CAZy 统计图

CAZy_cluster 图：

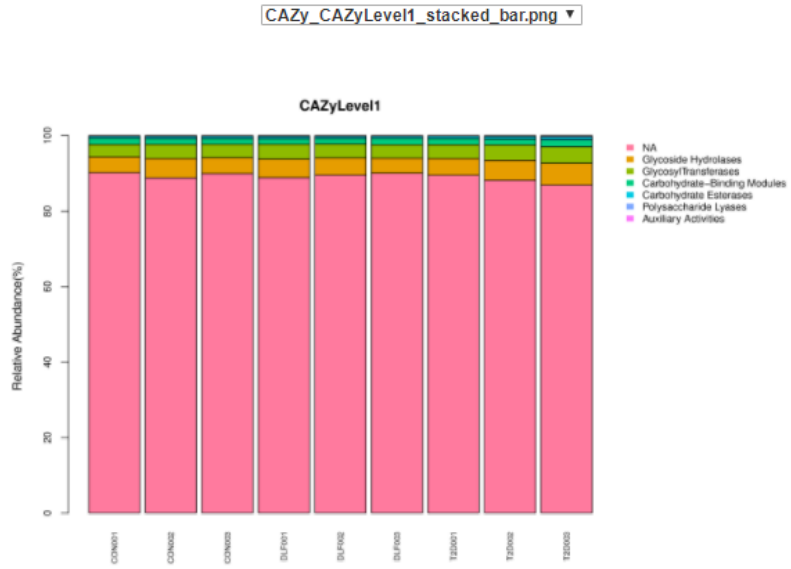
CAZy_CAZyLevel1_cluster.png ▾



CAZy_heatmap 图:



CAZy_stacked_bar 图:



1.6 eggNOG 功能注释

eggNOG（evolutionary genealogy of genes: Non-supervised Orthologous Groups）是一个用于同源聚类基因群注释的专业数据库，由 EMBL（European Molecular Biology Laboratory）维护，网址是 eggnogdb.embl.de。

该数据库最初版本始于 2007 年，2015 年 10 月已经更新到了最新的 4.5 版本。eggNOG 包括来自原始 COG/KOG 的功能分类，以及基于分类学的功能注释。目前该数据库包含 170 万个直系同源类群，覆盖了 3686 个物种，给定了 107 个不同的分类级别的同源群。

（1）基因 eggNOG 功能注释信息表

Query	eggNOG_hit	NOG	COGFunctional Category	COGFunctionalCategoryDescription	NOGDescription
Unigene1	349741.Amuc_1293	NA	NA	NA	NA
Unigene2	349741.Amuc_1421	COG0515	T	Signal transduction mechanisms	Serine Threonine protein kinase
Unigene3	471870.BACINT_02778	COG2017	G	Carbohydrate transport and metabolism	converts alpha-aldose to the beta-anomer. It is active on D-glucose, L-arabinose, D-xylose, D-galactose, maltose and lactose (By similarity)
Unigene4	693746.OBV_17230	COG1529	C	Energy production and conversion	Aldehyde oxidase and xanthine dehydrogenase, molybdopterin binding
Unigene4	693746.OBV_17230	COG2080	C	Energy production and conversion	(2Fe-2S)-binding domain protein
Unigene6	697281.Mahau_0023	ENOG410ZHQ3	S	Function unknown	nhl repeat containing protein
Unigene7	743722.Sph	COG07	M	Cell wall/membrane/envelope	penicillin-binding protein

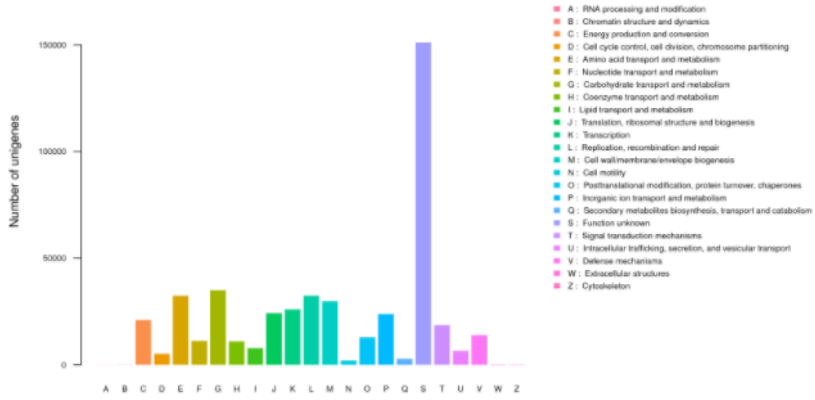
	21_0061	68		biogenesis
Unigene8	694427.Palpr_1505	ENOG41126 5D	S	Function unknown

结果路径：summary/6_FunctionalProfiling/eggNOG/eggNOG.anno.xlsx

表格参数说明：

Term	Description
Query	Unigene 的 ID 号
eggNOG_hit	基因在 eggNOG 数据库中比对上的注释信息
NOG	NOG ID（宏转录组中一般是 COG 开头）
COGFunctionalCategory	COG 注释对应的功能分类
COGFunctionalCategoryDescription	COG 注释对应的功能分类具体解释
NOGDescription	NOG 层面的具体解释

（2）COG 功能分类与 COG 基因表达丰度统计



结果路径：summary/6_FunctionalProfiling/eggNOG/eggNOG_category.png

2. 差异基因表达分析

差异表达的基因是宏转录组测序中最值得关注的结果，这些结果能够完整展现出不同处理或不同样本之间基因差异表达的情况。通常情况下差异显著的基因默认阈值为：log2(fold_change)≥1, (p<0.05)

2.1 差异基因表达分析

部分结果展示：

Unigene_ID	DLF001	DLF002	DLF003	CON001	CON002	CON003	mean	mean_stage1	mean_control	log2FC	regulation	p_value	q_value	significance
Unigene272772	0	0	0	0.01	0.01	3.42	0.57	0	1.14	-Inf	down	0.03	0.65	yes
Unigene276440	0	0	0	0.01	0.01	4.79	0.80	0	1.60	-Inf	down	0.03	0.65	yes
Unigene280211	0	0	0	0.01	0.01	4.56	0.76	0	1.52	-Inf	down	0.03	0.65	yes

Unigene588018	0	0	0	0.01	0.01	0.01	0.00	0	0.01	-Inf	down	0.03	0.65	yes
Unigene3	0	0	0	5.97	0.08	1.12	1.20	0	2.39	-Inf	down	0.04	0.65	yes
Unigene110	0	0	0	2.41	0.81	0.14	0.56	0	1.12	-Inf	down	0.04	0.65	yes
Unigene126	0	0	0	2.41	0.79	0.01	0.53	0	1.07	-Inf	down	0.04	0.65	yes
Unigene205	0	0	0	5.33	0.04	0.10	0.91	0	1.82	-Inf	down	0.04	0.65	yes
Unigene230	0	0	0	3.20	0.24	1.31	0.79	0	1.58	-Inf	down	0.04	0.65	yes

结果路径：

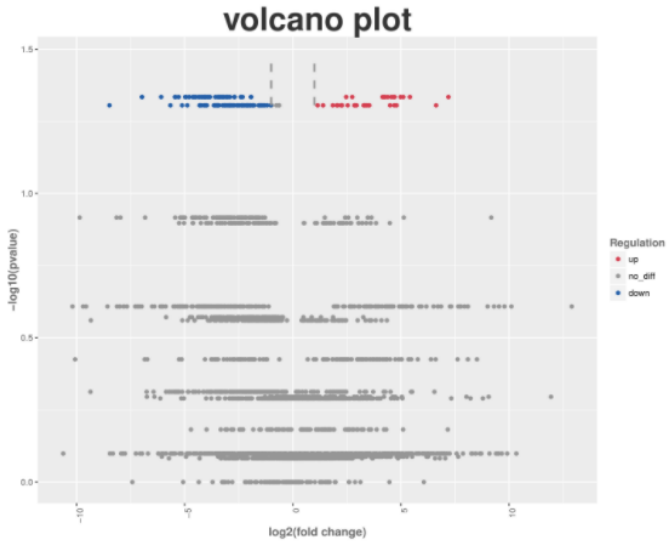
summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_diff_exp.xlsx

表格参数说明：

Term	Description
id	Unigene 的 ID 号
Sample A	样本 A
Sample B	样本 B
baseMean	平均基因表达量
baseMeanA	处理 A 的平均表达量
baseMeanB	处理 B 的平均表达量
foldChange	(处理 B)/(处理 A)的倍数的平均数，或者是(处理组)/(对照组)的倍数的平均数
log2(fold_change)	Log2(fold_change)，即对 foldChange 值取 log2
pvalue	即 p 值，用统计学方法得到的 p 值用于判断基因在样本之间是否存在差异
FDR	FDR 校正，即 false discovery rate 错误发现率，也就是错误拒绝的个数占所有被拒绝的
regulation	原假设个数的比例的期望值，是比较新的一种统计学检验方法，阈值设置比较灵活 基因上调或者下调
significant	差异显著或不显著
GO	GO 注释
KEGG	KEGG 注释

2.2 差异基因表达水平火山图

通过绘制火山图可以了解差异基因的整体分布情况。以 log2(fold_change)为横坐标，log(pvalue)为纵坐标，对差异表达的所有基因进行火山图绘制。红色的点代表上调的差异显著基因，蓝色的点代表下调的差异显著基因，灰色的点代表差异不显著的基因。



结果路径：

summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_diff_exp_volcano.png

2.3 差异基因 GO 富集分析

GO （gene ontology）是基因本体联合会所建立的数据库，旨在建立一个适用于各物种的标准。GO 提供了三类描述的系统定义方式，用于描述基因产物的功能，GO 的结构包括三个方面：Molecular Function、Biological Process 和 Cellar Component。GO 的基本单位是 term（词条、节点），每个 term 都对应一个属性。GO 功能富集分析首先把所有差异显著的基因比对到 GO 数据库，计算每个 term 对应的基因数目，然后使用超几何检验，以整个基因组作为背景，在显著性差异表达基因中找出显著富集的 GO term 个数。通过 GO 功能显著性富集分析能确定差异表达基因行使的主要生物学功能。

（1）GO 富集结果

GO_ID	GO_Term	GO_Function	GO_Level	S Unigene number	TS Unigene number	B Unigene number	TB Unigene number	pvalue
GO:0003735	structural constituent of ribosome	molecular_function	4	535	23940	6251	355487	0.00
GO:0016485	protein processing	biological_process	7	593	23940	7038	355487	0.00
GO:0006518	peptide metabolic process	biological_process	7	539	23940	6428	355487	0.00
GO:0004185	serine-type carboxypeptidase	molecular_function	9	528	23940	6293	355487	0.00

	activity							
GO:0005615	extracellular space	cellular_component	5	576	23940	6998	355487	0.00
GO:0004181	metallocarboxypeptidase activity							

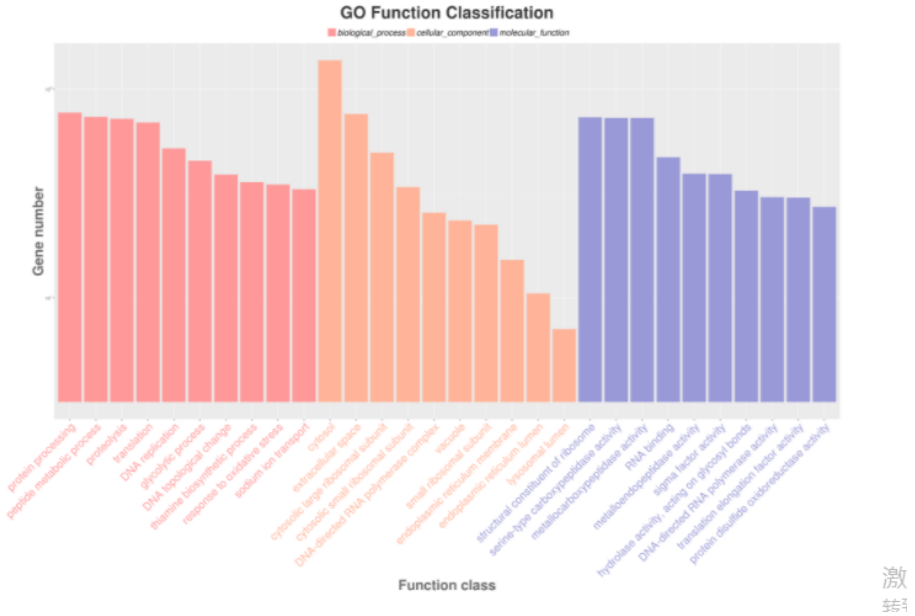
结果路径：

summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_GO_enrichment.xlsx

表格参数说明：

Term	Description
GO_ID	Gene ontology ID
GO_Term	GO 信息
GO_function	GO 功能分类
GO_class	GO 层级
S gene number	具有 GO 注释信息的差异显著基因个数
TS gene number	所有差异显著的基因个数
B gene number	具有 GO 注释信息的所有基因个数
TB gene number	基因总数
pvalue	p 值

（2）差异基因 GO 富集柱状图

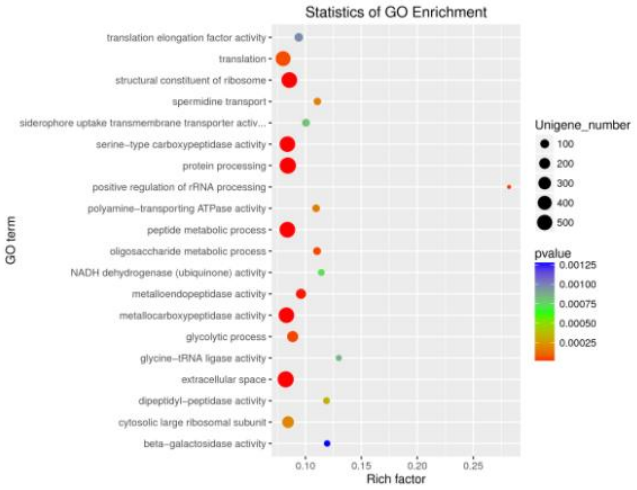


结果路径：

summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_GO_enrichment_barplot.png

（2）差异基因 GO 富集柱状图

采用 R 语言中的 ggplot2 对 GO 富集分析结果以散点图形化方式展示，Rich factor：位于该 GO 的差异基因个数/位于该 GO 的总基因数。Rich factor 越大，GO 富集程度越高



结果路径：

summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_GO_enrichment_scatterplot.png

2.7 差异基因 KEGG 富集分析

在生物体内，不同基因组相互协调行使生物学功能，基于 pathway 的分析有助于更进一步了解基因的生物学功能。KEGG 是有关 pathway 的主要公共数据库，pathway 显著性富集分析以 KEGG pathway 为单位，应用超几何检验找出与整个基因组背景相比，在显著性差异表达基因中显著富集的 pathway。

（1）KEGG 富集结果

Pathway Entry	PathwayDefinition	S Unigene number	TS Unigene number	B Unigene number	TB Unigene number	pvalue
map03010	Ribosome	705	19130	8136	281285	0.00
map00511	Other glycan degradation	393	19130	4635	281285	0.00
map00600	Sphingolipid metabolism	269	19130	3319	281285	0.00
map00195	Photosynthesis	16	19130	105	281285	0.00
map00970	Aminoacyl-tRNA	545	19130	7119	281285	0.00

	biosynthesis					
map00562	Inositol phosphate metabolism	98	19130	1107	281285	0.01
map00230	Purine metabolism	1176	19130	16118	281285	0.01
map04614	Renin-angiotensin system	23	19130	191	281285	0.01

结果路径：

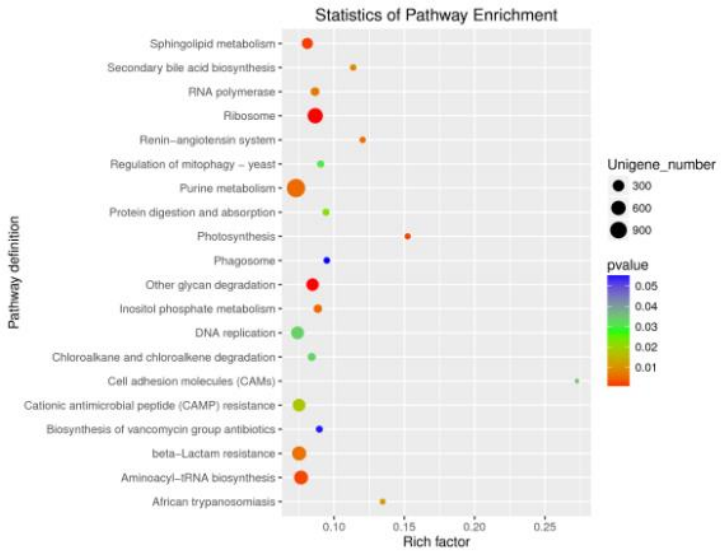
summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_KEGG_enrichment.xlsx

表格参数说明：

Term	Description
pathway_id	KEGG 的 ID
pathway_name	Pathway 具体信息
S gene number	具有 GO 注释信息的差异显著基因个数
TS gene number	所有差异显著的基因个数
B gene number	具有 GO 注释信息的所有基因个数
TB gene number	基因总数
pvalue	p 值

（2）差异基因 KEGG 富集性散点图

采用 R 语言中的 ggplot2 对 KEGG 富集分析结果以散点图形化方式展示，Rich factor：位于该 KEGG 的差异基因个数/位于该 KEGG 的总基因数。Rich factor 越大，KEGG 富集程度越高



结果路径：
src/summary_report/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_KEGG_enrichment_scatterplot.png

质量控制

1. 测序质量控制

高通量测序（Illumina HisSeq 4000 测序平台）得到的原始数据经碱基识别（Base Calling）转化为原始测序序列（Sequenced Reads），我们称之为 Raw Data 或者 Raw Reads，结果以 FASTQ 格式存储，其中包含测序序列（Reads）信息以及对应的测序质量信息。对于 FASTQ 格式文件的具体解释请参考辅助材料文件夹。鉴于高通量测序错误率对结果影响，我们需要对原始数据进行质量评估。测序的错误率会随着测序序列（Reads）的读长的增加而升高，这是由于测序过程中化学试剂的消耗而导致的，并且这是 Illumina 高通量测序平台的共有特征。

如上图所示，横坐标表示测序序列的碱基位置，纵坐标表示碱基质量测序错误率(Q 值)。Q20 反映了数据的质量，表示测序结果中，由于测序仪器造成的某个位置的碱基错误概率小于 1%。Q30 表示测序结果中，由于测序仪器造成的某位置碱基错误概率小于 0.1%。在测序的起始，测序质量都很高，随着反应进行，测序质量有所下降。GC 含量的分布用于分析是否因 测序或建库所带来的 GC 分离现象，以影响后续样品的定量分析。

正常情况下四种碱基的出现频率应该是接近的。而且没有位置差异。因此好的样本中四条线应该平等且接近。如下图所示，横轴为 reads 的碱基位置；纵轴为碱基所占的百分比；不同的颜色代表不同的碱基类型。

2. 生物学重复相关性验证

样品之间基因表达水平相关性是检验实验可靠性和样本合理性的重要指标。相关系数越接近 1，表明样品之间表达模式的相似度越高，其离散系数就越小。

结果路径：summary/4_GenePredict/5_correlation_heatmap.png

附录

1. 常用分析软件

为了每个客户方便查看数据以及完成一些简单的个性化分析，我们会额外提供一些软件和使用说明。部分软件由于占用空间较大，如果需要可以单独向对应的销售和项目经理索要。我们会为老师提供单独的下载链接和使用说明。

聚类分析软件 MeV

差异基因聚类分析用于判断基因在不同实验条件下调控的聚类模式。根据样本基因表达谱的相似程度，将基因进行聚类分析，直观地展示基因在不同样本（或者是不同处理）中的表达情况，由此获得生物学相关信息。

网络图绘制软件 Cytoscape

该软件可以构建可视化的分子相互作用网格，并将已有的的基因表达信息整合进该网络，可以非常方便地观察分子之间或者是蛋白之间的关联性

Notepad++

Nodepad++不仅是一款简单实用的文档查看软件，也是一款十分强大的代码编写软件。对于一些较大的文档，不管是 txt、fasta 还是 gtf，都可以用 notepad++打开进行任意编辑

2. 常用数据库说明

对于 FASTQ 格式详解，KEGG 官网通路查询，GEO 数据上传等问题，可以通过发送邮件 support@lc-bio.com 或向对应的销售和项目经理索要详细的步骤说明，对应的技术人员会发送对应的资料到您的邮箱，感谢您的理解。