Metagenomics 测序总结报告

客 户：DEMO

客户单位：DEMO

项目编码：DEMO

项目员：DEMO

您值得信赖的基因组学与蛋白质组学技术合作伙伴

宏基因组测序简介

从地球生命产生至今，微生物在很多方面一直占据着主导地位[1]。它们存储着地球上最多的 C、N、P 营养元素。它们几乎分布在所有的生态群落中，如大气、土壤、海洋，甚至一些极端条件的环境，如温度高达 340 摄氏度的深海热泉、地表以下 6km 的岩石中[2]。它们还在其他很多方面起着重要作用，如生态循环、食品生产、动植物健康。

传统的微生物研究主要采用纯培养的方法，但是自然界中 99%的微生物是无法培养的[3]，这极大地限制了对微生物多样性的研究及利用。

宏基因组学（metagenomics）的概念是由 Handelsman[4]等在 1998 年提出的，他们认为可以把微生物群体的所有基因组当成一个单一的基因组去处理。随后，Kevin[5]等对宏基因组学进行了定义，即"无需对微生物个体进行分离培养，直接应用基因组学技术对自然环境中的微生物群落进行研究"的学科。它规避了对样品中的微生物进行分离培养，提供了对无法分离培养的微生物进行研究的途径，避免了实验过程中由环境改变引起的微生物序列变化所带来的偏差[6]。

宏基因组测序按照具体的研究策略又可分为扩增子测序（Amplicon Sequencing）和鸟枪法测序（Shotgun Sequencing）两大类。扩增子测序利用基因组中的某些特定区段进行研究，如细菌的 16S rRNA 基因[7]、真菌的 ITS 区[8]等。鸟枪法测序[9]则是将提取的微生物基因组 DNA 打断成一系列片段分别进行测序。相比于扩增子测序，鸟枪法测序得到的序列信息更多，能够真实的反应样本中微生物多样性情况，同时又能在分子水平对其代谢通路、基因功能进行研究。

项目信息

1.样本信息

样品类型：DNA

研究物种：细菌/古菌/真菌/病毒测序区域：全基因组

2.测序信息

测序平台：HiSeq 4000 测序策略：

PE150 3.数据库信息

名称	版本/日期	链接
NR	2016.07.12	ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz
KEGG Database	2016.05.10	http://www.genome.jp/kegg/
eggNOG	4.5	http://eggnogdb.embl.de/download/
GO Database	2016.07.02	http://geneontology.org/
CARD database	2016.06.16	https://card.mcmaster.ca/download
CAZy database	2016.06.16	http://www.cazy.org/ http://www.phi-
PHI database	4.1	base.org/
Customized database	N/A	N/A

4.数据分析程序

分析项目	名称	版本
去除测序接头	cutadapt	1.9
去除低质量碱基	fqtrim	0.94
数据质控统计	FastQC[10]	0.10.1
去除宿主污染	bowtie2[11]	2.2.0
基因组组装	IDBA-UD[12]	1.1.1
组装结果统计评估	QUAST[13]	3.2
数据库比对	DIAMOND[14]	0.7.12
物种注释及统计比较	ACGT101_metagenome	2.0
功能注释及统计比较	ACGT101_metagenome	2.0

技术方法与实验流程

1.实验流程

采用自主研发的试剂盒进行 paired end 文库构建，文库质检合格后用 HiSeq4000 进行高通量测序，测序模式为 PE150。文库构建主要步骤如下图所示，包括：基因组 DNA 用超声打断、DNA 片段末端修复、加 'A' 碱基至 DNA 片段的 3’末端、加测序接头、片段选择、PCR 扩增、文库质检、上机测序。

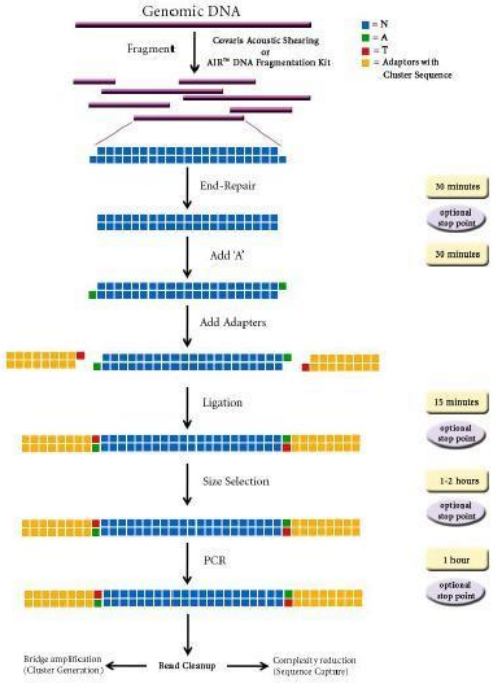


图 1: 测序文库构建原理图

2.信息分析流程具体步骤如下：

（1） 数据预处理：将测序得到的原始数据进行过滤低质量数据处理，保证后续信息分析结果的准确性；

（2） 宏基因组组装：得到有效数据后，按照样本分别进行组装，然后将未比上的 reads 进行混合组装，尽可能得到样本中的所有物种信息；

（3） 基因预测：将组装好的 contigs 进行 CDS 预测，随后根据预测结果进行过滤和去冗余；并进行相应的丰度计算；过滤低丰度表达后获得 Unigenes；

(4) 物种注释：将 Unigenes 与 NR_mate 库进行比对，获得物种注释信息；

(5) 功能注释：将 Unigenes 与 GO、KEGG、eggNOG、CAZy、CARD、PHI 数据库比对，进行功能注释和丰度分析；

(6) 统计及比较分析：在物种、功能、基因水平进行丰度统计分析及差异比较分析。

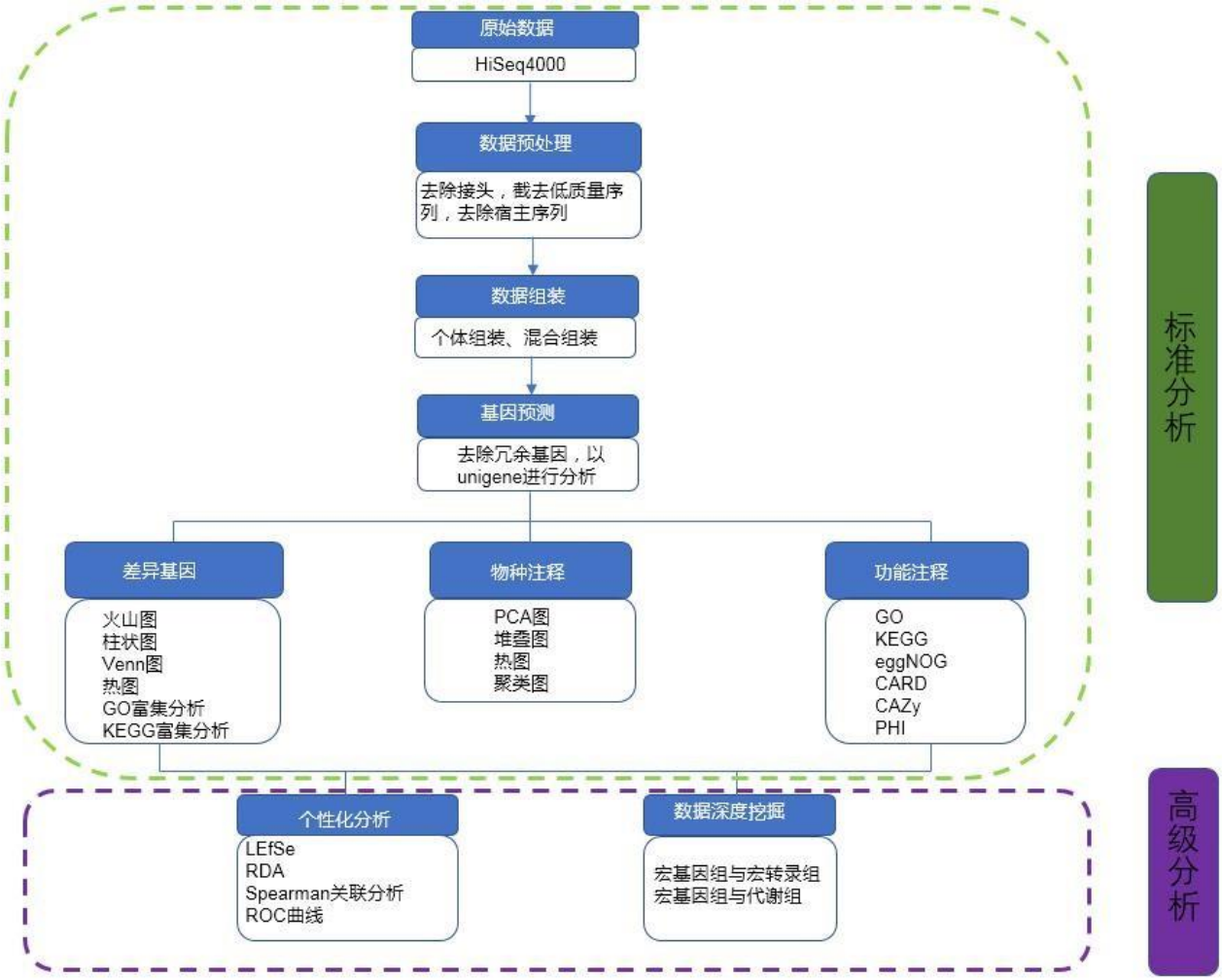


图 2：宏基因组信息分析流程

结果与分析

1.样品名称与分组信息

样品名称与分组信息表

#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.测序数据预处理

由于原始下机数据可能存在含有带有接头的（建库过程引入）和一定比例低质量数据（测序读取产生），为了保证后续分析的结果准确可靠，我们会对 Rawdata 进行预处理，从而获得用于后续分析的 Clean 数据。具体预处理步骤如下：

（1）去除测序 reads 中的接头序列

（2）对测序 reads 进行窗口法质量扫描，扫描窗口默认为 6bp，当窗口内平均质量值低于 20 时，将 read 从窗口起始到 3’终止的部分截掉

（3） 去除截短后长度小于 60bp 的序列

（4） 去除截短后 N 的含量在 5%以上的序列

（5） 去除宿主序列污染有效数据统计表

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/*_data_stat.xlsx

各列意义如下：

Term	**意义**
Sample	测序样本名称
Raw Data/Read	测序原始数据，以四行为一单位，统计每个文件的拼接后测序序列个数
Raw Data/Base	测序序列的个数乘以测序序列的长度，并以单位 G 表示
Valid Data/Read	预处理后，以四行为一个单位，统计每个文件的拼接后测序序列个数
Valid Data/Base	预处理后，测序序列的个数乘以测序序列的长度，并以单位 G 表示
Valid Ratio%	有效数据（Valid）与原始数据（Raw）比值，百分比表示
Q20%	有效数据中数据质量≥Q20 的数据比例
Q30%	有效数据中数据质量≥Q30 的数据比例 GC% 有效数据中数据 GC 含量

3.基因组组装

3.1 数据组装我们使用软件 IDBA-UD 对每个样品的测序数据进行基因组组装。 IDBA-UD（http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/）是一个基于交互式 De Bruijin 作图原理针对不同测序深度短 reads 进行从头拼接的组装软件。该软件可以从小的 k-mer 开始到大的 k-mer 不断迭代处理，形成 contigs，根据阈值过滤掉短的和低深度的 contigs，从而逐步完成低深度测序和高深度测序短 reads 的组装。组装结果为 FASTA 格式文件，其中的一条序列，也称为一个 contig，我们将长度大于 500bp 的 contigs 进行保留，用于后续聚类分析。部分序列如下所示：

>contig-80_0 length_11653 read_count_20759

AACCTAAAACTCCTGTAAGAAGAAGCGCCCCTAGAAATTGAGTCATCTTATCACCTACTTTTTATTTTA
TTATAACACATTTAGTACCTAGTACTAAATT

ACGGGTAGCCCGACTACCCTTATTATTTTTTAATATTTTATAGAACATACGTTCTTGCAGGAGGTATAAA
CATGTGGATTGAAAAATTTAAAAACAAAAA

TAACGAAACTAAATACAGATACTACGAGAAGTATAAAGAT...

其中 contig-80_0 代表 contig 编号（80 表示的是 kmer，0 表示第 n+1 个 contig），11653 代表序列碱基数，20759 为 reads count。

示例结果文件路径：summary/3_Assembly/CON001/contigs.min500.fasta

3.2 组装结果评估根据组装结果，我们用 QUAST[5]程序进行组装结果评估。一般来说，组装的结果的优劣主要看组装后的 contig 数目多少、N50 大小、组装总长度，组装后 contig 数目越少、 N50 越大、组装总长度越大，组装片段越完整，组装效果越好，反之，组装后 contig 数目越多、N50 越小、组装总长度越小，组装片段越不完整，组装效果越差。关于 N50 等指标的解释请看下文。

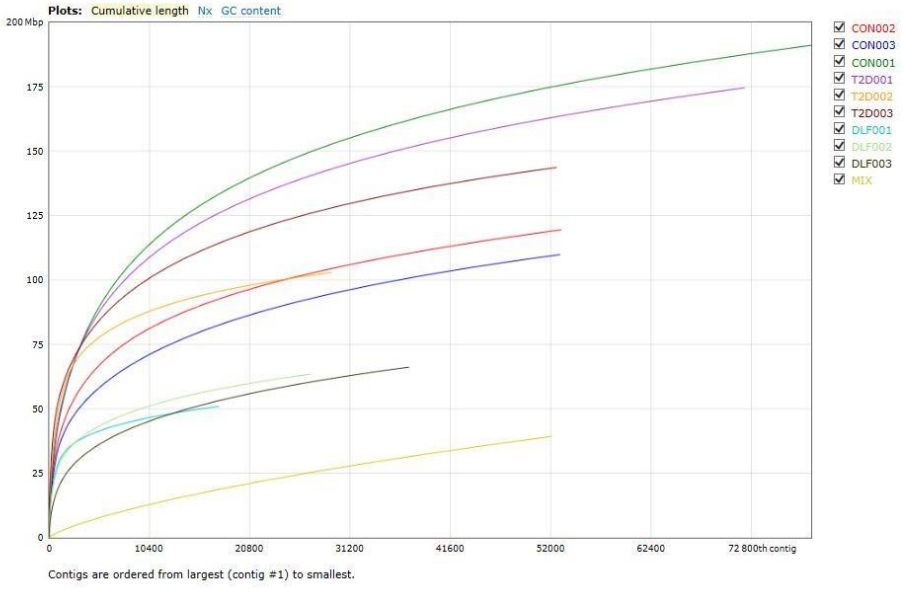
QUAST report

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



注：图中标签 Cumulative length 代表长度累计分布图，横轴代表 contig 从长到短排序的编号，纵轴代表累计长度；标签 Nx 代表 Nx 统计图，横轴代表 x 的数值，纵轴代表具体 Nx 数值，例如 x 为 50 时，Nx 就代表 N50；标签 GC content 代表 GC 含量分布图，横轴代表 GC 含量百分比，纵轴代表概率密度。

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

4.基因预测及丰度分析

4.1 基因预测及丰度统计

我们利用 MetaGeneMark 对各个样本及混合组装的 contigs（>=500bp）进行 CDS（Coding Region）预测，并且根据预测结果将 CDS 长度小于 100nt 的序列进行过滤。随后根据 CDS 预测结果，利用 CD-HIT 软件进行去冗余，获得非冗余的 gene，同时以 identity 95%, coverage 90%进行聚类，并选取最长的序列为代表性序列；利用 Bowtie2 将各样本的 Clean Data 比对到 gene 序列上，计算各个基因在各样品中比对上的 reads 数目；过滤掉在所有样品中比对上 reads 数目<=2 的基因，获得最终用于后续分析的 Unigenes；

每个样本中 unigene 比对上 reads 数目统计

Unigene_ID	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
Unigene1	134	0	0	0	0	11	0	2	0
Unigene2	227	0	5	0	0	20	0	0	0
Unigene3	291	4	52	0	0	0	0	0	0
Unigene4	917	2	0	2	1	8	3	1	0
Unigene5	50	0	13	0	0	0	2	1	0
Unigene6	10	0	4	0	0	0	0	0	0
Unigene7	356	1	35	0	2	5	0	1	0
Unigene8	101	0	22	0	0	0	0	1	0
Unigene9	149	2	28	0	0	1	0	0	0
Unigene10	65	0	8	0	0	2	1	0	0
Unigene11	109	65	26	1402	225	0	66	140	6
Unigene12	226	0	0	0	0	7	0	0	0
Unigene13	283	242	123	573	96	5	43	37	2488
Unigene14	39	2	6	0	3	1	1	0	0
Unigene15	40	0	7	0	6	2	0	0	0

Unigene16	136	2	15	0	22	6	0	0	0
Unigene17	117	1	14	0	12	3	0	0	0
Unigene18	107	0	23	0	0	5	1	0	2
Unigene19	144	0	25	0	0	9	0	0	0
Unigene20	55	2	7	0	0	2	0	3	1
Unigene21	161	0	0	0	0	3	0	1	0
Unigene22	85	0	19	0	0	0	0	0	0
Unigene23	14	0	0	0	0	0	0	0	0
Unigene24	105	0	0	4	0	2	0	0	0
Unigene25	102	0	0	0	0	0	0	0	0
Unigene26	67	0	0	0	152	0	0	0	0
Unigene27	85	0	0	0	218	2	0	0	0
Unigene28	69	0	0	0	236	0	0	0	0
Unigene29	47	0	1	0	140	1	0	0	0
Unigene30	18	3	0	0	0	0	0	0	0
Unigene31	169	0	0	0	0	4	0	0	0
Unigene32	28	0	0	0	0	0	0	0	0
Unigene33	16	0	0	0	0	0	0	33	0
Unigene34	77	0	0	0	0	1	0	256	0
Unigene35	67	0	0	0	0	4	0	201	0
Unigene36	101	1	3	1	0	3	0	0	1
Unigene37	53	0	0	0	0	0	0	0	0
Unigene38	89	0	0	0	0	0	2	0	0
Unigene39	51	0	0	0	0	0	0	0	0

Unigene40	306	0	0	0	0	12	3	0	0
Unigene41	26	0	1	0	0	0	0	0	0
Unigene42	107	2	7	0	4	3	1	0	0
Unigene43	78	4	8	0	8	6	1	4	0
Unigene44	131	15	35	2	20	10	0	0	0
Unigene45	21	2	1	0	3	1	0	0	0
Unigene46	48	0	0	1	0	7	0	1	0
Unigene47	74	0	3	0	0	15	1	0	0
Unigene48	303	5	11	1	1	29	1	0	0
Unigene49	83	0	0	0	0	8	0	0	0
Unigene50	41	2	4	0	8	3	0	0	0
Unigene51	514	0	0	0	1	8	0	0	0
Unigene52	33	1	4	0	3	1	0	0	0
Unigene53	44	0	2	0	5	2	1	1	0
Unigene54	102	8	8	0	7	1	0	0	1
Unigene55	48	16	6	0	28	18	42	0	11
Unigene56	60	21	6	0	0	17	1	0	0
Unigene57	431	0	0	0	0	13	0	0	0
Unigene58	80	0	41	0	0	0	0	1	0
Unigene59	53	0	0	2	67	0	0	0	18
Unigene60	75	0	0	2	143	1	0	0	31
Unigene61	40	0	0	0	136	0	0	0	35
Unigene62	219	1	1	0	14	0	0	0	3
Unigene63	50	0	0	0	0	0	0	0	0
Unigene64	73	40	7	0	0	31	51	0	21
Unigene65	137	0	0	0	0	4	0	0	0
Unigene66	12	4	0	0	0	0	11	0	0
Unigene67	248	32	121	0	0	85	1	0	0
Unigene68	177	15	69	0	0	45	0	0	0
Unigene69	191	3	42	0	1	29	1	0	1
Unigene70	92	40	48	23	8	20	49	165	0
Unigene71	13	11	9	4	0	3	6	27	0
Unigene72	199	0	0	0	1	10	0	0	0

Unigene73	97	16	0	1	0	1	0	0	0
Unigene74	51	7	14	0	6	2	15	13	0

Unigene75	60	8	5	0	16	0	13	10	0
Unigene76	67	14	14	1	6	11	0	0	3
Unigene77	108	12	17	0	5	21	14	0	0
Unigene78	60	6	10	1	0	21	3	0	0
Unigene79	105	15	14	0	4	30	7	0	0
Unigene80	226	8	33	1	2	14	577	2	0
Unigene81	71	29	15	1	8	29	18	0	3
Unigene82	104	7	14	0	0	24	3	0	2
Unigene83	57	11	7	0	1	18	6	0	3
Unigene84	31	1	6	0	1	13	1	0	0
Unigene85	131	6	13	0	5	3	2	0	0
Unigene86	47	3	2	0	1	0	0	0	0
Unigene87	52	1	6	0	4	1	0	0	0
Unigene88	17	0	3	0	0	0	0	0	0
Unigene89	117	8	17	2	5	4	3	2	2
Unigene90	200	0	29	0	0	4	0	0	0
Unigene91	137	0	28	1	0	0	0	0	0
Unigene92	56	0	13	0	0	0	0	0	1
Unigene93	90	0	0	0	0	18	0	63	0
Unigene94	217	0	0	0	0	34	2	155	0
Unigene95	13	0	0	0	0	3	0	12	0
Unigene96	220	0	7	0	0	19	0	0	0
Unigene97	161	0	9	0	1	14	0	0	0
Unigene98	55	0	2	0	0	6	0	0	0
Unigene99	103	0	0	0	0	6	0	0	0

全部结果请查看以下路径：summary/4_GenePredict/4_Unigenes_count.txt

根据比对上的 reads 数目及各个基因长度，计算各基因在每个样品中的丰度信息（即 TPM 值），计算公式如下所示：

$$G_k = \frac{r_k}{L_k} * \frac{1}{\sum_{i=1}^{n} \frac{r_i}{L_i}} * 10^6$$

注：r 为比对上基因的 reads 数目，L 为基因的长度

每个样本中 unigene 丰度统计

Unigene_ID	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
Unigene1	8.12	0.00	0.00	0.00	0.00	1.26	0.00	0.11	0.00
Unigene2	5.10	0.00	0.12	0.00	0.00	0.85	0.00	0.00	0.00
Unigene3	5.97	0.08	1.12	0.00	0.00	0.00	0.00	0.00	0.00
Unigene4	8.37	0.02	0.00	0.03	0.02	0.14	0.03	0.01	0.00
Unigene5	1.86	0.00	0.51	0.00	0.00	0.00	0.07	0.04	0.00
Unigene6	1.02	0.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00
Unigene7	4.44	0.01	0.46	0.00	0.06	0.12	0.00	0.01	0.00
Unigene8	5.68	0.00	1.30	0.00	0.00	0.00	0.00	0.05	0.00
Unigene9	5.33	0.07	1.05	0.00	0.00	0.07	0.00	0.00	0.00
Unigene10	3.58	0.00	0.46	0.00	0.00	0.21	0.05	0.00	0.00
Unigene11	2.95	1.77	0.74	56.3885	14.8032	0.00	1.68	3.58	0.14
Unigene12	8.00	0.00	0.00	0.00	0.00	0.47	0.00	0.00	0.00
Unigene13	3.29	2.84	1.50	9.92	2.72	0.11	0.47	0.41	24.2512
Unigene14	2.33	0.12	0.38	0.00	0.44	0.11	0.06	0.00	0.00
Unigene15	2.37	0.00	0.44	0.00	0.86	0.22	0.00	0.00	0.00
Unigene16	2.46	0.04	0.29	0.00	0.97	0.21	0.00	0.00	0.00
Unigene17	2.21	0.02	0.28	0.00	0.55	0.11	0.00	0.00	0.00
Unigene18	3.58	0.00	0.81	0.00	0.00	0.32	0.03	0.00	0.06
Unigene19	3.16	0.00	0.58	0.00	0.00	0.37	0.00	0.00	0.00
Unigene20	2.70	0.10	0.36	0.00	0.00	0.19	0.00	0.14	0.04
Unigene21	10.4071	0.00	0.00	0.00	0.00	0.37	0.00	0.06	0.00
Unigene22	3.28	0.00	0.77	0.00	0.00	0.00	0.00	0.00	0.00
Unigene23	1.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Unigene24	1.70	0.00	0.00	0.10	0.00	0.06	0.00	0.00	0.00
Unigene25	2.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Unigene26	3.15	0.00	0.00	0.00	17.3946	0.00	0.00	0.00	0.00
Unigene27	2.88	0.00	0.00	0.00	17.9753	0.13	0.00	0.00	0.00
Unigene28	2.38	0.00	0.00	0.00	19.8054	0.00	0.00	0.00	0.00
Unigene29	2.74	0.00	0.06	0.00	19.8761	0.11	0.00	0.00	0.00
Unigene30	2.91	0.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Unigene31	9.36	0.00	0.00	0.00	0.00	0.42	0.00	0.00	0.00
Unigene32	3.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Unigene33	1.51	0.00	0.00	0.00	0.00	0.00	0.00	2.95	0.00
Unigene34	1.47	0.00	0.00	0.00	0.00	0.04	0.00	4.62	0.00
Unigene35	1.64	0.00	0.00	0.00	0.00	0.19	0.00	4.66	0.00
Unigene36	1.99	0.02	0.06	0.03	0.00	0.11	0.00	0.00	0.02
Unigene37	2.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Unigene38	2.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00
Unigene39	2.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Unigene40	6.98	0.00	0.00	0.00	0.00	0.52	0.06	0.00	0.00
Unigene41	2.08	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00
Unigene42	2.38	0.04	0.16	0.00	0.22	0.13	0.02	0.00	0.00
Unigene43	2.35	0.12	0.25	0.00	0.59	0.34	0.03	0.11	0.00
Unigene44	2.40	0.28	0.67	0.05	0.89	0.35	0.00	0.00	0.00
Unigene45	2.04	0.20	0.10	0.00	0.71	0.18	0.00	0.00	0.00
Unigene46	2.23	0.00	0.00	0.07	0.00	0.61	0.00	0.04	0.00
Unigene47	3.40	0.00	0.14	0.00	0.00	1.30	0.04	0.00	0.00
Unigene48	4.00	0.07	0.15	0.02	0.03	0.72	0.01	0.00	0.00
Unigene49	3.28	0.00	0.00	0.00	0.00	0.60	0.00	0.00	0.00
Unigene50	1.26	0.06	0.13	0.00	0.60	0.17	0.00	0.00	0.00
Unigene51	8.12	0.00	0.00	0.00	0.04	0.24	0.00	0.00	0.00
Unigene52	1.71	0.05	0.22	0.00	0.38	0.10	0.00	0.00	0.00
Unigene53	1.15	0.00	0.06	0.00	0.32	0.10	0.02	0.02	0.00
Unigene54	1.91	0.15	0.16	0.00	0.32	0.04	0.00	0.00	0.02
Unigene55	1.98	0.67	0.26	0.00	2.81	1.40	1.63	0.00	0.38
Unigene56	2.60	0.92	0.27	0.00	0.00	1.39	0.04	0.00	0.00
Unigene57	7.83	0.00	0.00	0.00	0.00	0.45	0.00	0.00	0.00
Unigene58	2.53	0.00	1.37	0.00	0.00	0.00	0.00	0.03	0.00
Unigene59	1.45	0.00	0.00	0.08	4.47	0.00	0.00	0.00	0.41
Unigene60	1.53	0.00	0.00	0.06	7.09	0.04	0.00	0.00	0.53
Unigene61	0.97	0.00	0.00	0.00	8.05	0.00	0.00	0.00	0.71
Unigene62	1.63	0.01	0.01	0.00	0.25	0.00	0.00	0.00	0.02
Unigene63	1.63	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Unigene64	1.24	0.69	0.13	0.00	0.00	1.00	0.82	0.00	0.30

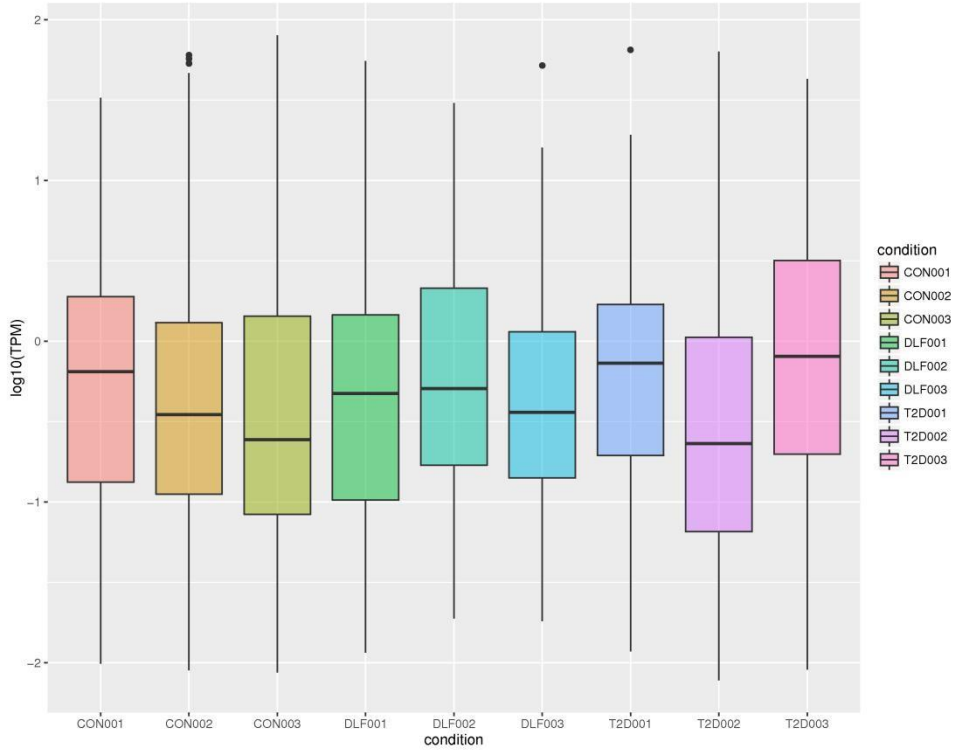
Unigene65	7.81	0.00	0.00	0.00	0.00	0.43	0.00	0.00	0.00
Unigene66	1.58	0.53	0.00	0.00	0.00	0.00	1.36	0.00	0.00
Unigene67	4.89	0.64	2.51	0.00	0.00	3.17	0.02	0.00	0.00
Unigene68	3.54	0.30	1.45	0.00	0.00	1.70	0.00	0.00	0.00

Unigene69	3.11	0.05	0.72	0.00	0.04	0.89	0.02	0.00	0.01
Unigene70	1.33	0.58	0.73	0.49	0.28	0.55	0.67	2.25	0.00
Unigene71	2.29	1.95	1.67	1.05	0.00	1.00	1.00	4.50	0.00
Unigene72	7.64	0.00	0.00	0.00	0.09	0.73	0.00	0.00	0.00
Unigene73	1.54	0.26	0.00	0.02	0.00	0.03	0.00	0.00	0.00
Unigene74	1.15	0.16	0.33	0.00	0.33	0.09	0.32	0.28	0.00
Unigene75	1.34	0.18	0.12	0.00	0.87	0.00	0.27	0.21	0.00
Unigene76	5.91	1.24	1.30	0.13	1.29	1.83	0.00	0.00	0.22
Unigene77	6.81	0.76	1.13	0.00	0.77	2.51	0.83	0.00	0.00
Unigene78	4.43	0.45	0.78	0.11	0.00	2.93	0.21	0.00	0.00
Unigene79	4.52	0.65	0.63	0.00	0.42	2.45	0.28	0.00	0.00
Unigene80	4.44	0.16	0.68	0.03	0.10	0.52	10.6804	0.04	0.00
Unigene81	4.14	1.70	0.92	0.09	1.14	3.20	0.99	0.00	0.15
Unigene82	4.58	0.31	0.65	0.00	0.00	2.00	0.12	0.00	0.07
Unigene83	3.68	0.72	0.48	0.00	0.16	2.20	0.37	0.00	0.16
Unigene84	2.31	0.08	0.47	0.00	0.18	1.83	0.07	0.00	0.00
Unigene85	1.59	0.07	0.17	0.00	0.15	0.07	0.02	0.00	0.00
Unigene86	1.24	0.08	0.06	0.00	0.06	0.00	0.00	0.00	0.00
Unigene87	1.12	0.02	0.14	0.00	0.21	0.04	0.00	0.00	0.00
Unigene88	2.81	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.00
Unigene89	2.17	0.15	0.33	0.06	0.23	0.14	0.05	0.04	0.03
Unigene90	5.99	0.00	0.91	0.00	0.00	0.23	0.00	0.00	0.00
Unigene91	1.80	0.00	0.39	0.02	0.00	0.00	0.00	0.00	0.00
Unigene92	1.80	0.00	0.44	0.00	0.00	0.00	0.00	0.00	0.03

Unigene93	2.01	0.00	0.00	0.00	0.00	0.76	0.00	1.33	0.00
Unigene94	2.44	0.00	0.00	0.00	0.00	0.72	0.02	1.65	0.00
Unigene95	0.86	0.00	0.00	0.00	0.00	0.38	0.00	0.75	0.00
Unigene96	4.04	0.00	0.14	0.00	0.00	0.66	0.00	0.00	0.00
Unigene97	4.26	0.00	0.25	0.00	0.06	0.70	0.00	0.00	0.00
Unigene98	3.39	0.00	0.13	0.00	0.00	0.70	0.00	0.00	0.00
Unigene99	4.27	0.00	0.00	0.00	0.00	0.47	0.00	0.00	0.00

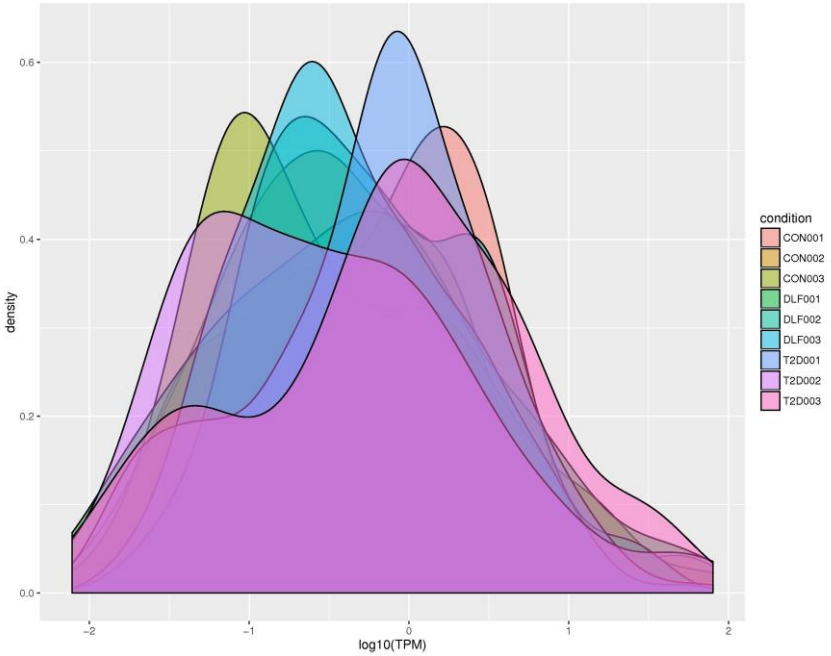
全部结果请查看以下路径：summary/4_GenePredict/4_Unigenes_abund.txt

4.2 基因表达量统计



注：横坐标为各个样本名称；纵坐标为各个样本 TPM 值的 log 值。每个区域的盒形图对五个统计量（每个盒形图至上而下分别对应该样本表达量的最大值、上四分位数、中值、下四分位数和最小值）。

结果路径：summary/4_GenePredict/4_Unigenes_abund.boxplot.png



注： 横坐标为 log10TPM 值，纵坐标为基因密度。**结果路径：** summary/4_GenePredict/4_Unigenes_abund.density.png

4.3 预测结果基本信息统计

Unigene 基本信息统计表

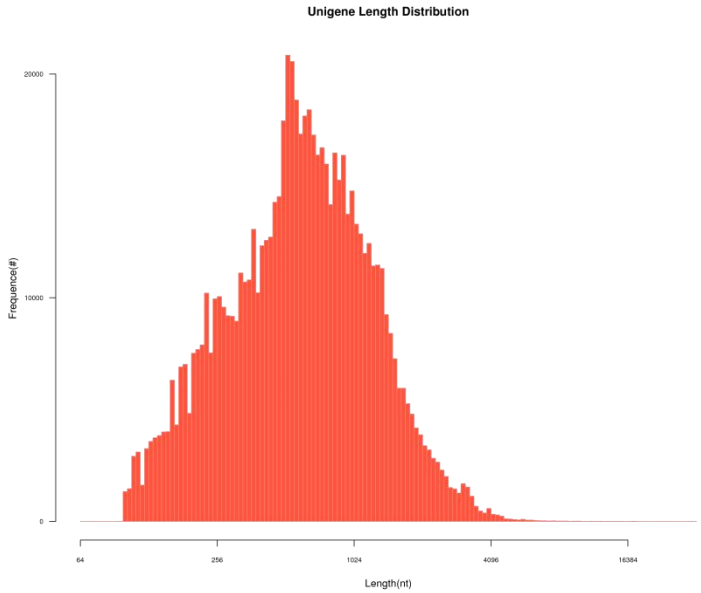
ORFs NO.	**739705**
integrity:start	139358(18.84%)
integrity:end	108767(14.70%)
integrity:all	442333(59.80%)
integrity:none	49247(6.66%)
Total Len.(Mbp)	560.34
Average Len.(bp)	757.52
GC percent	45.88%

结果路径：summary/4_GenePredict/2_Unigenes_CDS_stats.xlsx

各列意义如下：

ORFs NO.	**unigene 中基因的数目**
integrity:start	包含起始密码子的基因数目及百分比

integrity:end	包含终止密码子的基因数目及百分比
integrity:all	完整基因（既有起始密码子也有终止密码子）数目的百分比
integrity:none	没有起始密码子也没有终止密码子的基因数目及百分比
Total Len.(Mbp)	unigene 中基因的总长，单位是百万
Average Len.(bp)	unigene 中基因的平均长度
GC percent	预测的 unigene 中基因的整体 GC 含量值

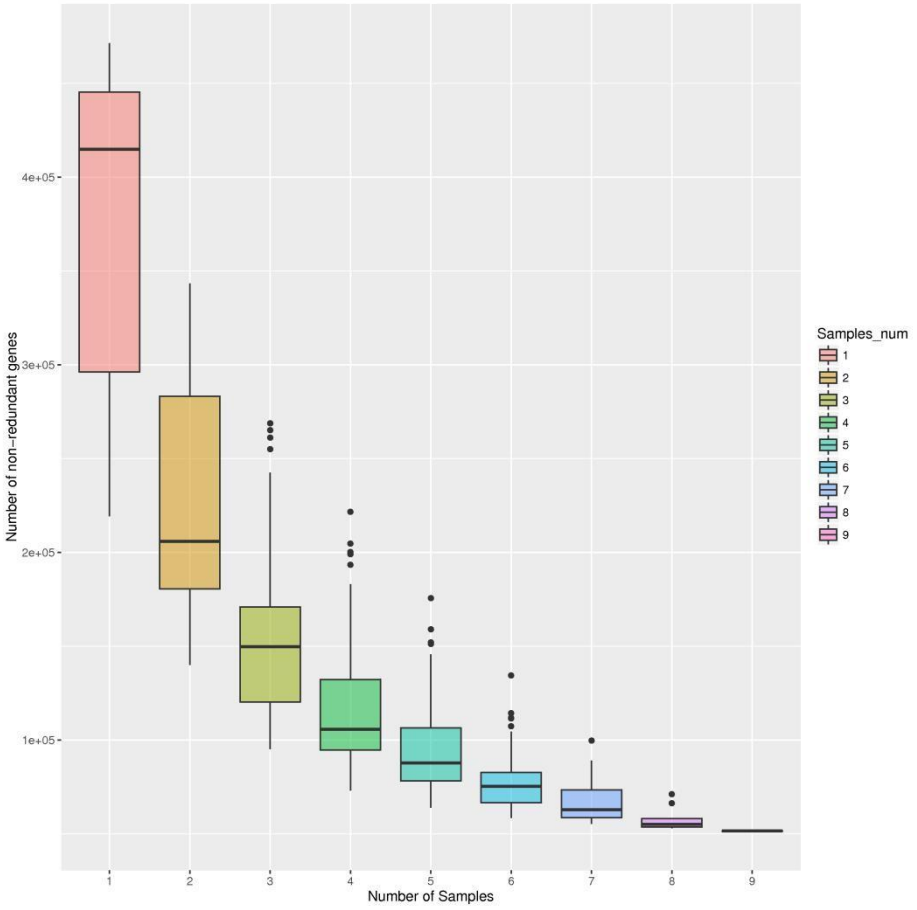


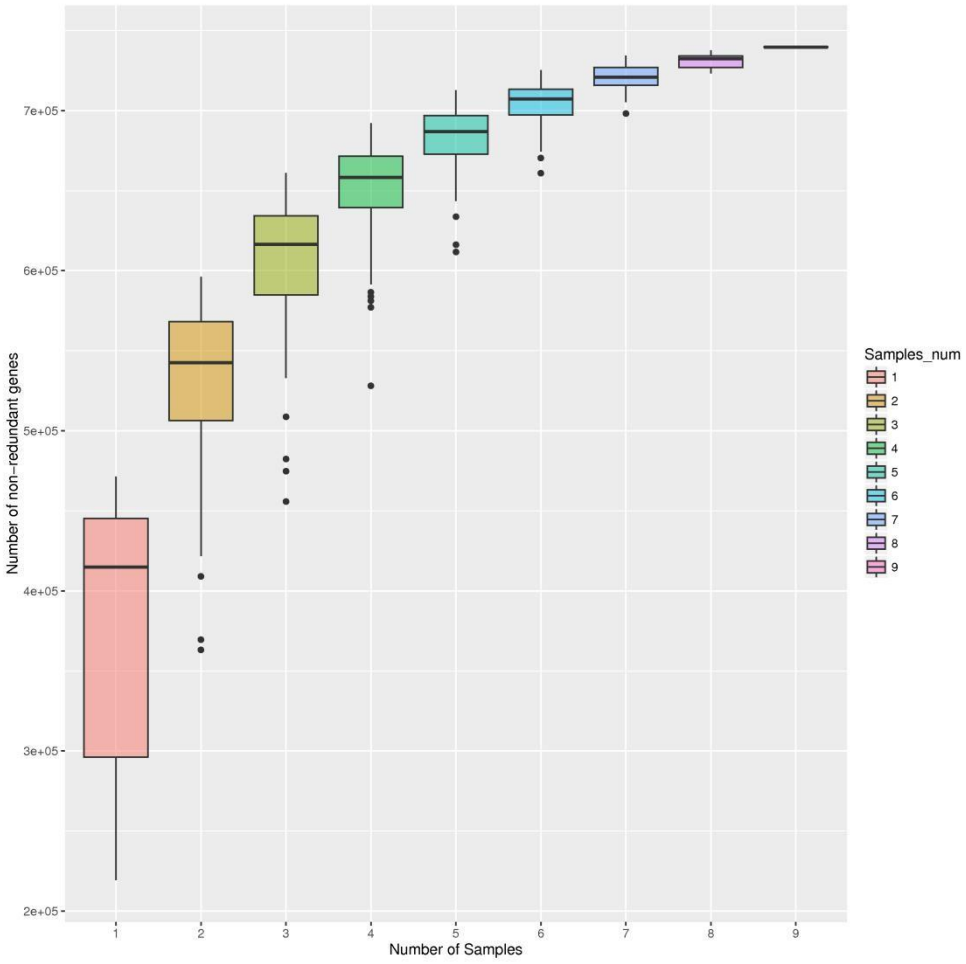
注：纵轴 Frequence(#) 表示 Unigene 的数目；横轴表示 Unigene 的长度。

结果路径：summary/4_GenePredict/2_Unigenes_len.png

4.4 core-pan 基因分析

此外，我们根据基因在各样品中的丰度表，可以获得各样品的基因数目信息，通过随机抽取不同数目的样品，可以获得不同数目样品组合间的基因交集（core genes）的数目和基因并集(pan genes)的数目，由此我们构建和绘制了 Core 和 Pan 基因的稀释曲线，结果如下：


Number of non−redundant genes
4e+05
3e+05
2e+05
1e+05
1
2
3
4
5
6
7
8
9
Number of Samples
Samples_num
1
2
3
4
5
6
7
8
9

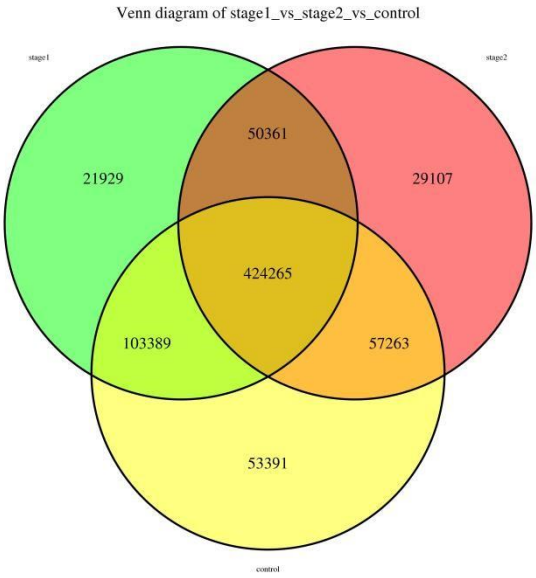


core 基因-pan 基因稀释曲线

结果路径： summary/4_GenePredict/5_core_gene.png **结果路径：** summary/4_GenePredict/5_pan_gene.png

4.5 基因数目韦恩图分析

Venn 图可用于统计多组样品中所共有和独有的 Unigene 数目，可以比较直观的表现环境样品的 Unigene 数目组成相似性及特异性。



5. 物种注释

5.1 注释数据库介绍

NR 数据库是 NCBI 中的一个非冗余的蛋白数据库。它包含从 GeneBank 核酸序列翻译而来的非冗余序列，并且还收录了其他蛋白数据库的非冗余序列，包括 RefSeq、PDB、SwissProt、PIR 和 PRF。从 NR 数据库（版本 2016.07.12）中提取属于微生物（Bacteria,Archaea,Viruses,Fungi）的序列，序列数目为 52375954，这些序列的构成 NR_meta 子库。

NR 数据库链接：ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz

5.2 物种注释方法

利用 DIAMOND 软件将 Unigenes 与 NR_meta 库进行比对（blastp，evalue ≤ 1e5）；比对后对每个 Unigene 的比对结果，选取 evalue <= 最小 evalue*10 的比对结果进行物种分类。结合 NCBI 的物种分类系统（它描述了序列 gi 号与物种分类号 taxid 的一种对应关系），通过与 MEGAN 软件相似的 LCA（Lowest Common Ancestor）算法，可以得到序列的各个分类等级（SuperKingdom、Phylum、Class、Order、Family、Genus、Species）的具体物种注释信息。

将物种分类信息与 Unigene 的丰度信息结合，可以得到各个分类等级的丰度信息，某个分类等级的丰度等于注释到该分类等级的 Unigene 丰度的加和。

上文的 LCA 算法是指：对于一条比对到 NCBI 分类系统中多个分类节点即多个 taxid 的序列，序列的分类名称最终确定为这几个分类节点的最低共同祖先的分类名称。图 9 为 LCA 的简单示意，图中深绿色的结点代表 x,y 两个序列分类名称的最低共同祖先分类名称。

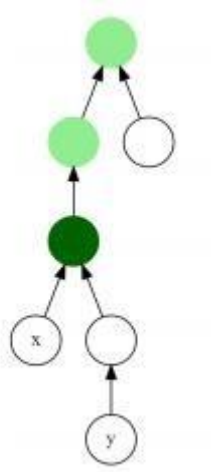


图 4：LCA 示意图

5.2.1 物种注释汇总表

Unigene 相对丰度及物种注释

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Taxonomy
[illegible]	8.12	0.00	0.00	0.00	0.00	1.26	0.00	0.11	0.00	d__Bacteria;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Akkermansiaceae;g__Akkermansia;s__Akkermansia_muciniphila
[illegible]	5.10	0.00	0.12	0.00	0.00	0.85	0.00	0.00	0.00	d__Bacteria;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Akkermansiaceae;g__Akkermansia;s__Akkermansia_muciniphila
[illegible]	5.97	0.08	1.12	0.00	0.00	0.00	0.00	0.00	0.00	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes;s__Alistipes_indistinctus
[illegible]	8.37	0.02	0.00	0.03	0.02	0.14	0.03	0.01	0.00	d__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium;s__Clostridium_sp._CAG:226
[illegible]	1.86	0.00	0.51	0.00	0.00	0.00	0.07	0.04	0.00	d__unclassified;p__unclassified;c__unclassified;o__unclassified;f__unclassified;g__unclassified;s__unclassified
[illegible]	1.02	0.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00	d__unclassified;p__unclassified;c__unclassified;o__unclassified;f__unclassified;g__unclassified;s__unclassified

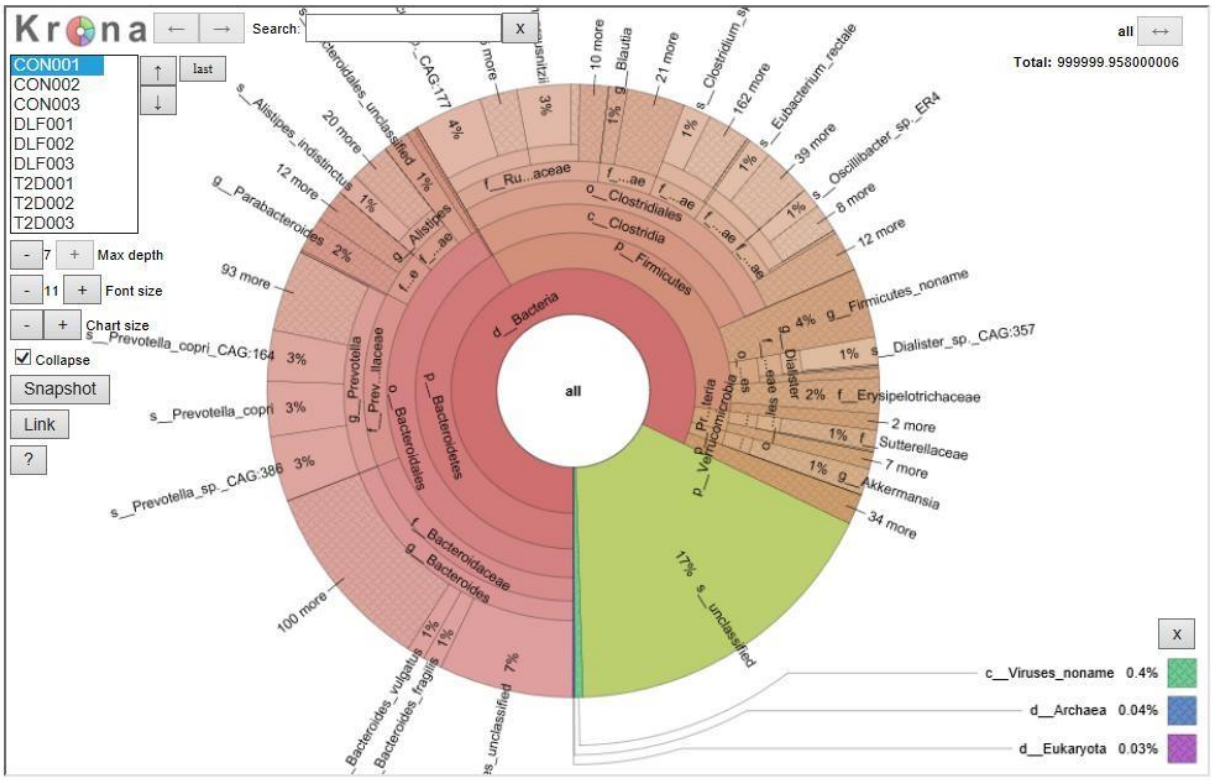
	4.44	0.01	0.46	0.00	0.06	0.12	0.00	0.01	0.00	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes;s__Alistipes_indistinctus
	5.68	0.00	1.30	0.00	0.00	0.00	0.00	0.05	0.00	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes;s__Alistipes_indistinctus
	5.33	0.07	1.05	0.00	0.00	0.07	0.00	0.00	0.00	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes;s__Alistipes_indistinctus
	3.58	0.00	0.46	0.00	0.00	0.21	0.05	0.00	0.00	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes;s__Alistipes_indistinctus

全部结果请查看以下路：

summary/5_TaxonomicProfiling/Unigenes_abund_taxonomy.txt

5.2.2 Krona 分析

根据各个样本不同分类层级（域门纲目科属种）的不同丰度，我们利用 Krona 对物种注释结果进行可视化展示。



Krona 物种注释结果**注：**圆圈从内到外依次代表不同的分类级别（域门纲目科属种）；扇形的大小代表不同物种的相对比例。

结果路径：summary/5_TaxonomicProfiling/krona.html

5.3 物种丰度分析

5.3.1 物种丰度统计

将每个样本在某个分类水平上的丰度合在一起，得到某水平的样本丰度表，根据该表我们可以对某个分类层级上各个样品中的丰度进行比较观察。表：样品丰度表部分数据(以 Taxonomy 的 Phylum 水平为例)

Unigene 基本信息统计表

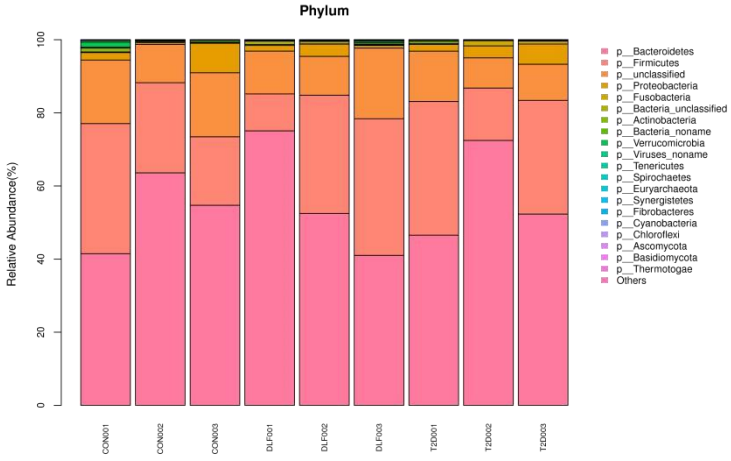
Phylum	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
p__Bacteroidetes	414760.18	635855.80	547111.64	750758.79	524851.07	410170.15	465413.82	724467.24	523149.01
p__Firmicutes	355851.13	246482.12	187447.46	100911.55	323426.14	373590.09	365696.85	143405.70	310832.46
p__unclassified	173674.07	104536.04	175052.41	117089.84	106136.55	193277.14	137458.07	82485.64	98899.05
p__Proteobacteria	20519.89	5185.64	80774.99	15326.71	33488.35	7078.77	18706.30	32703.76	55138.26
p__Fusobacteria	252.08	187.38	237.51	2758.56	580.65	404.22	200.57	14267.13	8206.12
p__Bacteria_unclassified	1444.01	2757.06	1421.73	9143.35	6847.53	1189.42	1765.69	588.74	782.86
p__Actinobacteria	10641.16	807.38	1209.26	592.70	2075.99	2005.75	5913.36	554.09	1591.68
p__Bacteria_noname	2346.14	2189.37	4968.94	1377.72	1567.41	4402.53	3472.80	649.87	478.48
p__Verrucomicrobia	14109.93	27.55	366.73	7.71	23.08	3958.15	14.86	27.56	3.13
p__Viruses_noname	4446.83	187.34	259.84	1309.14	272.93	261.58	109.52	134.67	117.77
p__Tenericutes	370.68	457.41	561.93	229.50	256.51	2583.95	650.97	563.87	396.00
p__Spirochaetes	420.65	234.90	259.86	106.97	263.03	410.89	264.16	71.24	256.68
p__Euryarchaeota	373.54	880.64	79.15	95.82	46.98	400.86	38.33	29.15	55.69
p__Synergistetes	339.46	103.62	68.70	72.77	100.06	109.42	155.43	20.48	48.10
p__Fibrobacteres	7.42	2.04	51.42	154.20	0.56	9.09	1.81	0	3.47
p__Cyanobacteria	61.24	25.00	43.75	2.16	11.88	21.72	40.25	0.03	1.55
p__Chloroflexi	27.57	20.62	7.12	32.70	2.36	18.29	14.94	1.09	14.43
p__Ascomycota	94.30	2.07	6.83	0.99	3.89	10.96	3.50	9.59	6.50
p__Basidiomycota	64.15	1.19	0.74	0.32	0.32	5.77	0	0	0.03
p__Thermotogae	32.79	2.81	21.58	0.20	2.00	0.91	1.27	3.02	0.27
p__Eukaryota_noname	58.53	0.02	0	0	0.07	0.26	0	0	0
p__Glomeromycota	11.74	0.02	0	0	0.18	43.70	1.68	1.34	0
p__Nitrospirae	4.34	3.06	1.89	0	0.84	2.41	43.60	0	1.53
p__Gemmatimonadetes	1.52	0	22.27	3.74	9.70	0.48	2.17	0.20	4.96
p__Planctomycetes	12.57	4.17	12.91	4.47	0.85	2.43	3.09	0.11	0.49
p__Acidobacteria	4.63	1.91	1.73	1.40	14.45	5.86	0.05	1.36	0.05
p__Chlorobi	9.25	3.63	5.13	3.93	0.08	0.67	0.08	3.33	3.43
p__Poribacteria	1.91	27.37	0	0	0	0.11	0	0	0
p__Marinimicrobia	2.26	2.94	0.10	0.15	7.17	4.83	6.29	0	0
p__Atribacteria	3.69	0.13	0	7.46	0.24	6.97	3.47	0	0.03
p__Deinococcus-Thermus	2.50	2.98	0.41	0.27	0.22	3.53	1.18	4.02	4.43
p__Thermodesulfobacteria	1.54	0.16	1.95	4.90	0	0.58	5.51	0.14	0.04

p__Chlamydiae	6.89	0	0.08	0	3.32	3.91	0	0	0
p__Aquificae	2.70	2.14	0	0.06	4.87	2.22	0.64	0.02	0.03
p__Cryptomycota	10.87	0	0.06	0	0	0	0	0	0
p__Aminicenantes	2.69	1.12	1.02	0	0.17	3.40	0	0	0
p__Microsporidia	7.18	0	0	0	0	0.52	0	0.69	0
p__Chytridiomycota	6.94	0.02	0	0	0	0.04	0	0	0
p__Armatimonadetes	2.78	1.72	0.16	0	0	0.97	0.03	0	0
p__Elusimicrobia	0.18	0.25	0.38	0.05	0	0.05	4.61	0.16	0
p__Cloacimonetes	0.02	0	0	0	0	3.82	0	0.74	0
p__Deferribacteres	1.00	0	0	1.77	0	0	0	1.65	0
p__Parcubacteria	1.43	0	0	0	0	2.82	0	0	0
p__Latescibacteria	0	0	0.02	0.02	0.04	0.03	0	1.55	2.50
p__Dictyoglomi	0.40	0.13	0	0	0	0.11	1.68	1.61	0
p__Ignavibacteriae	1.00	0.05	0.19	0.16	0	0.55	1.35	0	0
p__Crenarchaeota	1.69	0.24	0	0	0.14	0	1.03	0	0
p__Chrysiogenetes	0	2.36	0	0	0.05	0	0	0	0
p__Thaumarchaeota	0	0.07	0.18	0	0	0.13	0.43	0.20	1.02
p__Omnitrophica	1.14	0.03	0	0	0	0	0	0	0
p__Acetothermia	0.17	0	0.05	0	0.11	0	0.69	0	0
p__Blastocladiomycota	0.63	0	0	0	0	0	0	0	0
p__Lentisphaerae	0.52	0	0	0	0	0	0	0	0

结果路径：summary/5_TaxonomicProfiling/2_Phylum/taxonomy_Phylum_abund.xlsx

5.3.2 样品柱状图分析

得到样品丰度表之后，默认选取丰度最高的 20 个物种分类或功能分类，并将其余的物种设置为 Others，进行相对丰度计算，即计算每个分类的百分比，得到相对丰度文件，绘制样品丰度比较的柱状图，该柱状图以堆叠柱状图(stacked bar chart)形式展现，便于更直观地进行样品丰度的比较。



图：样品丰度柱状图

注：横轴为样品名称，纵轴代表某分类的相对丰度

结果路径：summary/5_TaxonomicProfiling/*/taxonomy_*_stacked_bar.png 样品丰度分析数据(以 Taxonomy 的 Phylum 水平为例)

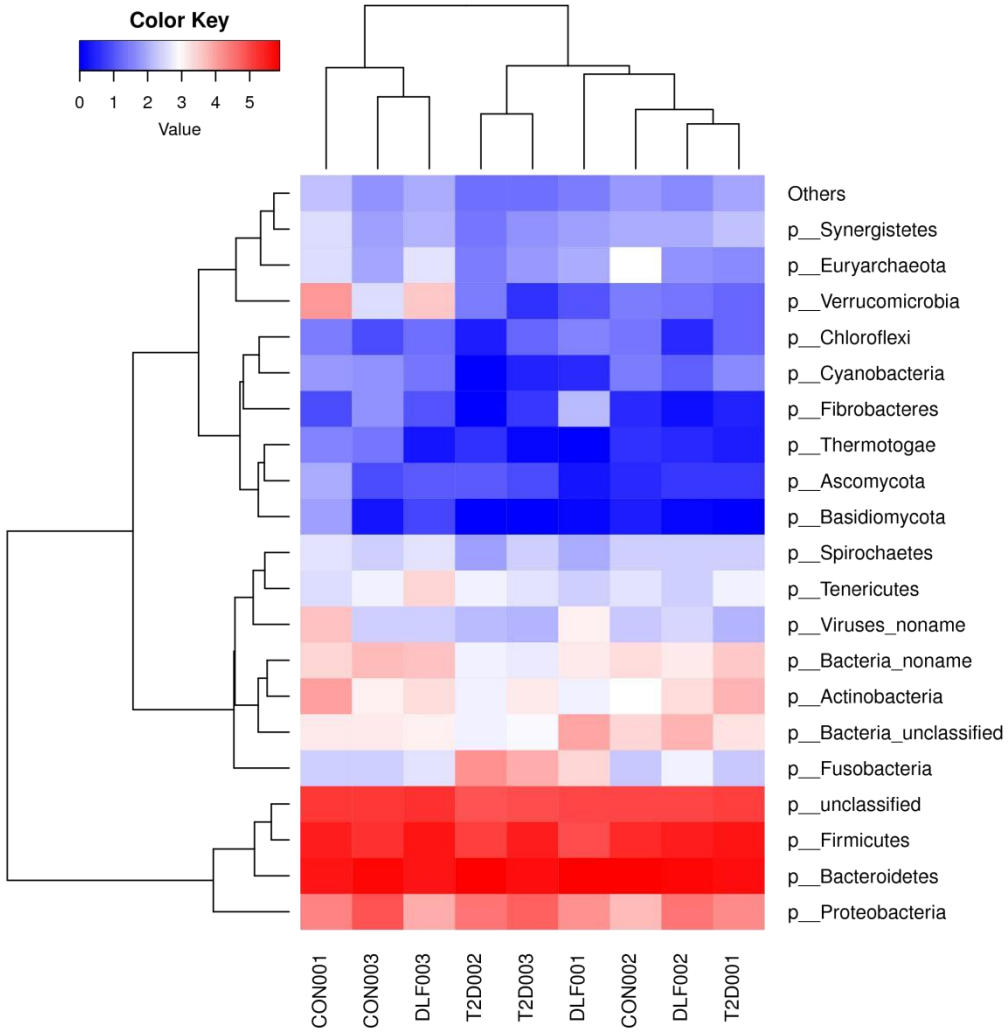
Phylum	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
p__Bacteroidetes	414760.18	635855.80	547111.64	750758.79	524851.07	410170.15	465413.82	724467.24	523149.01
p__Firmicutes	355851.13	246482.12	187447.46	100911.55	323426.14	373590.09	365696.85	143405.70	310832.46
p__unclassified	173674.07	104536.04	175052.41	117089.84	106136.55	193277.14	137458.07	82485.64	98899.05
p__Proteobacteria	20519.89	5185.64	80774.99	15326.71	33488.35	7078.77	18706.30	32703.76	55138.26
p__Fusobacteria	252.08	187.38	237.51	2758.56	580.65	404.22	200.57	14267.13	8206.12
p__Bacteria_unclassified	1444.01	2757.06	1421.73	9143.35	6847.53	1189.42	1765.69	588.74	782.86
p__Actinobacteria	10641.16	807.38	1209.26	592.70	2075.99	2005.75	5913.36	554.09	1591.68
p__Bacteria_noname	2346.14	2189.37	4968.94	1377.72	1567.41	4402.53	3472.80	649.87	478.48
p__Verrucomicrobia	14109.93	27.55	366.73	7.71	23.08	3958.15	14.86	27.56	3.13
p__Viruses_noname	4446.83	187.34	259.84	1309.14	272.93	261.58	109.52	134.67	117.77
p__Tenericutes	370.68	457.41	561.93	229.50	256.51	2583.95	650.97	563.87	396.00
p__Spirochaetes	420.65	234.90	259.86	106.97	263.03	410.89	264.16	71.24	256.68
p__Euryarchaeota	373.54	880.64	79.15	95.82	46.98	400.86	38.33	29.15	55.69
p__Synergistetes	339.46	103.62	68.70	72.77	100.06	109.42	155.43	20.48	48.10
p__Fibrobacteres	7.42	2.04	51.42	154.20	0.56	9.09	1.81	0	3.47
p__Cyanobacteria	61.24	25.00	43.75	2.16	11.88	21.72	40.25	0.03	1.55
p__Chloroflexi	27.57	20.62	7.12	32.70	2.36	18.29	14.94	1.09	14.43
p__Ascomycota	94.30	2.07	6.83	0.99	3.89	10.96	3.50	9.59	6.50
p__Basidiomycota	64.15	1.19	0.74	0.32	0.32	5.77	0	0	0.03
p__Thermotogae	32.79	2.81	21.58	0.20	2.00	0.91	1.27	3.02	0.27
Others	162.73	54.50	48.51	28.38	42.48	90.40	77.59	17.13	18.50

结果路径：

summary/5_TaxonomicProfiling/2_Phylum/taxonomy_Phylum_stacked_bar.xlsx

5.3.3 样品热图分析

根据样品相对丰度表，选取一定数据量的高丰度物种或功能分类，默认选取 top20，进行样品热图（Heatmap）绘制。Heatmap 图用颜色变化来反映二维矩阵或表格中的数据信息，它可以直观地将数据值的大小以定义的颜色深浅表示出来。常根据需要对各物种分类、功能分类或样品进行相似性聚类，以聚类树的形式展现，相近丰度模式的物种分类、功能分类或样品会被聚集到一起。通过颜色梯度及相似程度来反映多个样品在各分类水平上组成的相似性和差异性。



图：样品热图

注：横轴为样品名称，纵轴代表某分类的相对丰度

结果路径：summary/5_TaxonomicProfiling/*/taxonomy_*_heatmap.png

样品热图分析数据(以 Taxonomy 的 Phylum 水平为例)

Phylum	CON001	CON002	CON003	DLF001	DLF002	DLF003	T2D001	T2D002	T2D003
p__Bacteroidetes	414760.18	635855.80	547111.64	750758.79	524851.07	410170.15	465413.82	724467.24	523149.01
p__Firmicutes	355851.13	246482.12	187447.46	100911.55	323426.14	373590.09	365696.85	143405.70	310832.46
p__unclassified	173674.07	104536.04	175052.41	117089.84	106136.55	193277.14	137458.07	82485.64	98899.05
p__Proteobacteria	20519.89	5185.64	80774.99	15326.71	33488.35	7078.77	18706.30	32703.76	55138.26
p__Fusobacteria	252.08	187.38	237.51	2758.56	580.65	404.22	200.57	14267.13	8206.12
p__Bacteria_unclassified	1444.01	2757.06	1421.73	9143.35	6847.53	1189.42	1765.69	588.74	782.86
p__Actinobacteria	10641.16	807.38	1209.26	592.70	2075.99	2005.75	5913.36	554.09	1591.68

p__Bacteria_noname 2346.14 2189.37 4968.94 1377.72 1567.41 4402.53 3472.80 649.87 478.48 p__Verrucomicrobia 14109.93 27.55 366.73 7.71 23.08 3958.15 14.86 27.56 3.13 p__Viruses_noname 4446.83 187.34 259.84 1309.14 272.93 261.58 109.52 134.67 117.77

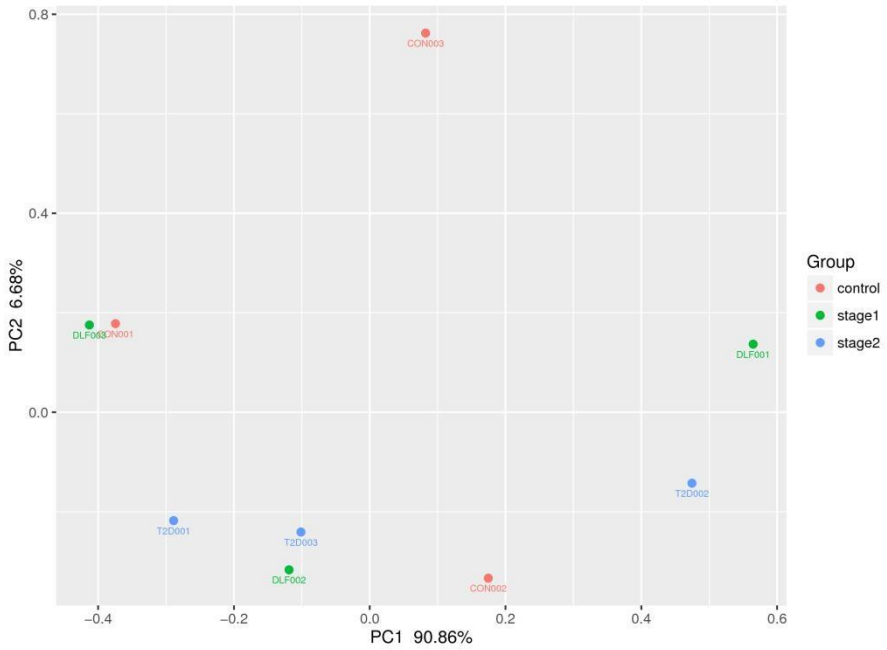
p__Tenericutes	370.68	457.41	561.93	229.50	256.51	2583.95	650.97	563.87	396.00
p__Spirochaetes	420.65	234.90	259.86	106.97	263.03	410.89	264.16	71.24	256.68
p__Euryarchaeota	373.54	880.64	79.15	95.82	46.98	400.86	38.33	29.15	55.69
p__Synergistetes	339.46	103.62	68.70	72.77	100.06	109.42	155.43	20.48	48.10
p__Fibrobacteres	7.42	2.04	51.42	154.20	0.56	9.09	1.81	0	3.47
p__Cyanobacteria	61.24	25.00	43.75	2.16	11.88	21.72	40.25	0.03	1.55
p__Chloroflexi	27.57	20.62	7.12	32.70	2.36	18.29	14.94	1.09	14.43
p__Ascomycota	94.30	2.07	6.83	0.99	3.89	10.96	3.50	9.59	6.50
p__Basidiomycota	64.15	1.19	0.74	0.32	0.32	5.77	0	0	0.03
p__Thermotogae	32.79	2.81	21.58	0.20	2.00	0.91	1.27	3.02	0.27
Others	162.73	54.50	48.51	28.38	42.48	90.40	77.59	17.13	18.50

结果路径：

summary/5_TaxonomicProfiling/2_Phylum/taxonomy_Phylum_heatmap.xlsx

5.3.4 物种主成分分析(PCA)

主成分分析 (PCA，Principal Component Analysis)通过寻找矩阵的线性无关向量实现数据降维，这些线性无关的向量就称为主成分，第一主成分解释了数据最大部分的方差或波动性。PCA 能够提取出最大程度反映样品间差异的两个坐标轴，从而将多维数据的差异反映在二维坐标图上，进而揭示复杂数据背景下的简单规律。基于不同分类层级的物种丰度表，我们进行了 PCA 分析，如果样品的物种组成越相似，则它们在 PCA 图中的距离越接近。



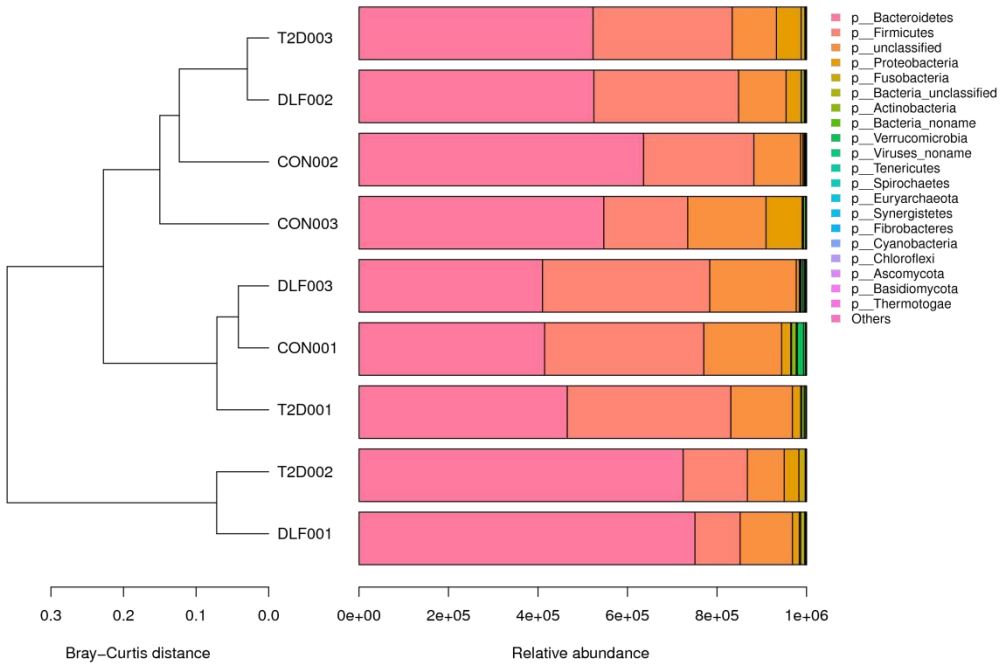
图：物种 PCA 结果展示**注：**横坐标表示第一主成分，百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，百分比表示第二主成分对样品差异的贡献值；图中的每个点表示一个样品，同一个组的样品使用同一种颜色表示。

结果路径：summary/5_TaxonomicProfiling/*/PCA/PCA_Group.png

5.3.5 样品聚类分析

为了研究不同样品之间的相似性，还可以通过对样品进行聚类分析，构建聚类树。Bray-Curtis 距离是系统聚类法中使用最普遍的一个距离指标，它主要用来刻画样品间的相近程度，它的大小是进行样品分类的主要依据。

根据各物种在各样品中的丰度表出发，以 Bray-Curtis 距离矩阵进行样品间聚类分析，并将聚类结果与各样品的物种相对丰度整合进行展示。以门水平为例：



注：左侧是 Bray-Curtis 距离聚类树结构；右侧的是各样品在门水平上的物种相对丰度分布图。

结果路径：summary/5_TaxonomicProfiling/*/taxonomy_*_cluster.png

5.4 样品差异分析

根据样品丰度表，我们可以进行样品或样品组间的物种或功能差异分析。差异检验方法方法是 Fisher's exact test 和 Kruskal-Wallis test，其中 Fisher's exact test 适用于无生物学重复样本，而 Kruskal-Wallis test 是一种非参数检验方法，适用于有生物学重复样本的多组比较（2 组及以上）。

样品差异分析结果(以 Taxonomy 的 Phylum 水平为例)

Phylum	DLF001	DLF002	DLF003	CON001	CON002	CON003	mean_stan	mean_age1	mean_control	log2FC	regulation	p_value	q_value	significance
p__Latescibacteria	0.02	0.04	0.03	0	0	0.02	0.02	0.03	0.01	2.48	up	0.05	0.52	yes

p__Fusobacteria	2758.56	580.65	404.22	252.08	187.38	237.57	36.73	1247.81	225.66	2.47	up	0.05	0.52	yes
p__Cyanobacteria	2.16	11.88	21.72	61.24	25.00	43.75	27.63	11.92	43.33	-1.86	down	0.05	0.52	yes
p__Thermotogae	0.20	2.00	0.91	32.79	2.81	21.58	10.05	1.03	19.06	-4.20	down	0.05	0.52	yes
p__Elusimicrobia	0.05	0	0.05	0.18	0.25	0.38	0.15	0.03	0.27	-3.04	down	0.05	0.52	yes
p__Armatimonadetes	0	0	0.97	2.78	1.72	0.16	0.94	0.32	1.55	-2.27	down	0.12	0.56	no
p__Cryptomycota	0	0	0	10.87	0	0.06	1.82	0	3.64	-Inf	down	0.12	0.56	no

---

结果路径：

summary/5_TaxonomicProfiling/2_Phylum/diff_abundance/Group.stage1_vs_control/

Group.stage1_vs_control.diff.xlsx

6. 功能注释

6.1 GO 数据库注释

6.1.1 数据库简介

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。GO 数据库链接：http://geneontology.org/

6.1.2 GO 注释

Unigene GO 注释总表

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

结果路径：summary/6_FunctionalProfiling/GO/Unigenes_GO.txt **各列意义如下：**

Term	意义
Query	Unigene 名称
GO_hit	该 Unigene 对应的 GO 名称
Gene	该 GO 对应的 NCBI 基因 ID
Symbol	该 GO 对应的基因名称

GO_ID	gene ontology ID
GO_Term	该 GO 信息
GO_Function	GO ontology 的功能分类
GO_Level	该 Unigene 对应的 GO 层次

6.1.3 GO 丰度分析

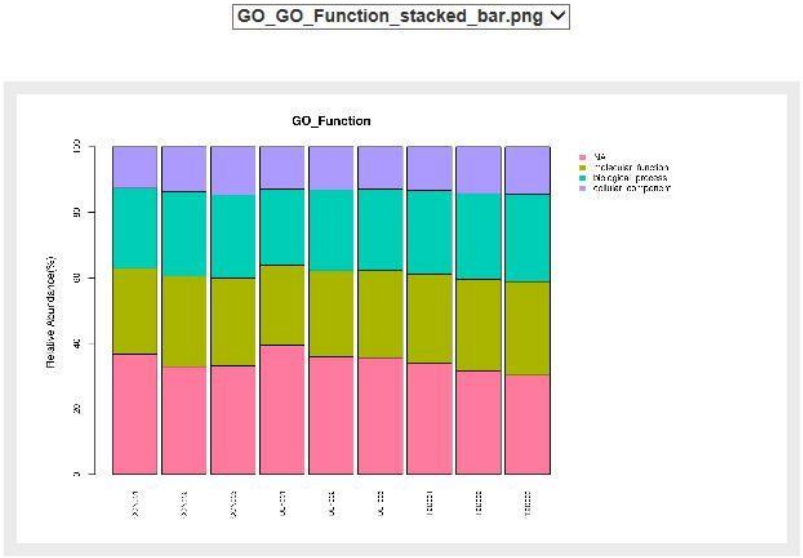


图: GO 丰度柱状图

注：横轴为样品名称，纵轴代表各个分类水平的相对丰度

结果路径：summary/6_FunctionalProfiling/GO/*/GO_*_stacked_bar.png

6.1.4 GO 热图

GO_GO_Function_heatmap.png

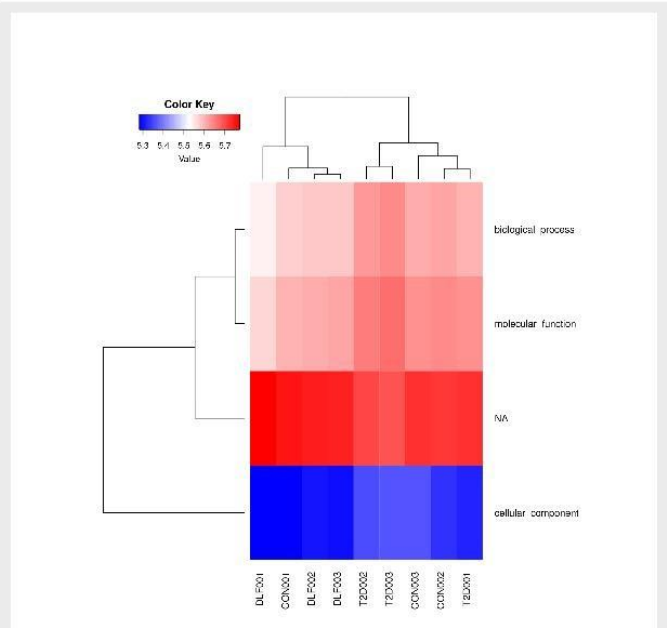


图: GO 热图分析

注：图中用蓝色到红色的渐变色反映丰度由低到高的变化，越趋近于蓝色，丰度越低，越趋近于红色，丰度越高。

结果路径：summary/6_FunctionalProfiling/GO/*/GO_*_heatmap.png

6.1.5 PCA 分析

1_GO_Function_PCA_Group.png

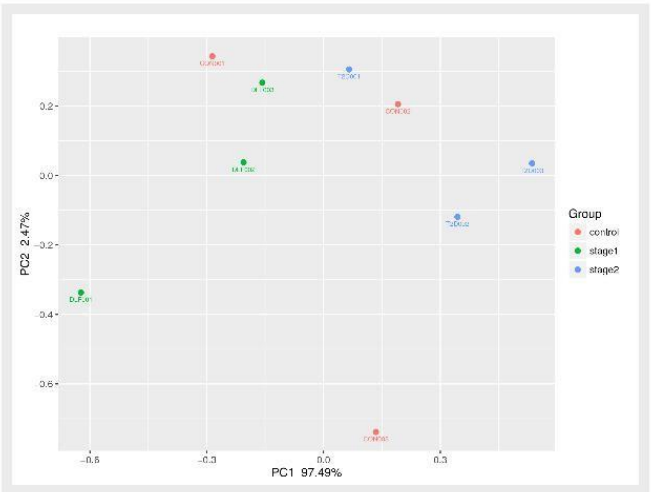


图: 物种 PCA 结果展示

注：横坐标表示第一主成分，百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，百分比表示第二主成分对样品差异的贡献值；图中的每个点表示一个样品，同一个组的样品使用同一种颜色表示。

结果路径：summary/6_FunctionalProfiling/GO/*/PCA/PCA_Group.png

6.1.6 GO 聚类分析

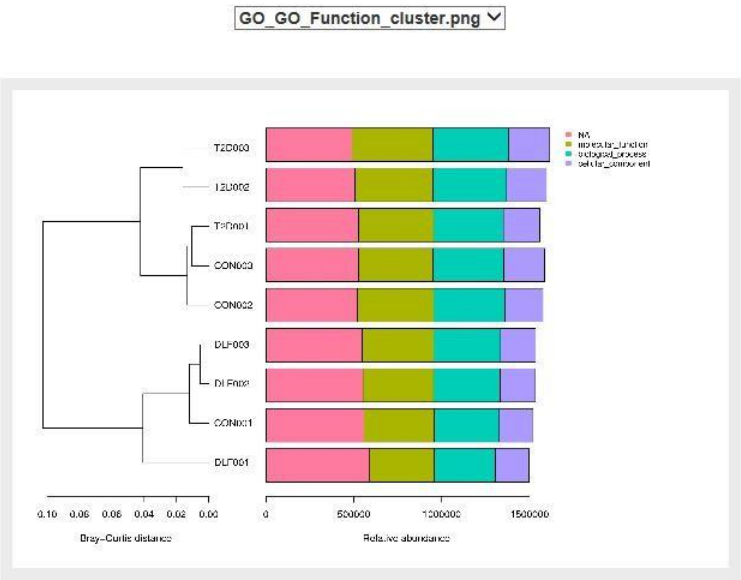


图: GO 聚类分析**注：**左侧是 Bray-Curtis 距离聚类树结构；右侧的是物种相对丰度分布图。

结果路径：summary/6_FunctionalProfiling/GO/*/GO_*_cluster.png

6.2 KEGG 数据库注释

6.2.1 数据库简介

KEGG[15、16]（Kyoto Encyclopedia of Genes and Genomes, 京都基因和基因组百科全书）是联系基因组分子水平的信息与高层次生物系统功能信息，包括细胞层次、生物体层次、生态环境层次的数据库。它是生物系统的一种计算机表示形式，包括作为整个系统架构单元的各种基因和蛋白（基因组信息），各种参与相互作用、化学反应关系网络（系统信息）的化学物质（化学信息），还包括对生物系统产生扰动作用的各种疾病和药物信息（健康信息）。基因组分子水平的信息主要来自于基因组测序及其他的高通量测序技术产生的数据。

KEGG 数据库是 Kanehisa 实验室于 1995 年开始开发的一个数据库，现在已经成为著名的进行基因组和高通量测序数据注释和整合的参考数据库。

根据系统信息、基因组信息、化学信息和健康信息这几个大类，KEGG 目前可分为 17 个主要的数据库。其中的 KEGG PATHWAY 数据库尤其广泛地被用于基因组和高通量数据的注释。KEGG 数据库链接：http://www.kegg.jp/kegg/download

6.2.2 KEGG 注释

KEGG PATHWAY 数据库常用于序列的注释，它是一系列手工绘制的代谢通路的集合，代表了人们目前的对于分子间相互作用和反应网络的知识。KEGG PATHWAY 可以分为四个层次或水平。第一层可分为 7 个方面：1.新陈代谢(Metabolism) 2.遗传信息加工(Genetic Information Processing) 3. 环境信息加工(Environmental Information Processing) 4.细胞过程(Cellular Processes) 5.生物体系统(Organismal Systems) 6.人类疾病(Human Diseases) 7.药物开发(Drug Development)。第二层进一步细分，第三层则是具体到各个代谢通路图，第四层为每个代谢通路图的具体注释信息。

利用 DIAMOND 软件将 Unigenes 与 KEGG PATHWAY 库的蛋白序列进行比对（blastp，evalue ≤ 1e-5），取分值最高的 KEGG 比对结果序列（Hits）的注释作为该 Unigene 序列的注释。结合 KEGG PATHWAY 数据库的层次结构，我们可以统计出 KEGG PATHWAY 各个层级或水平的信息。

Query	KEGG_hit	GeneName	GeneDescription	KOEntry	KODescription	EC	PathwayEntry	PathwayDefinition	KEGGLevel1	KEGGLevel2
Unigene3	bcel:BcellWH2_00928	K01785	aldose 1-epimerase precursor	EC:5.1.3.3	Metabolism	mro_2	epimerase	bcel00010	Glycolysis / Gluconeogenesis	Carbohydrate smmetabolism
Unigene3	bcel:BcellWH2_00928	K01785	aldose 1-epimerase precursor	EC:5.1.3.3	GalactoseMetabolism	mro_2	epimerase	bcel00052	Carbohydrate smmetabolism	

结果路径：summary/6_FunctionalProfiling/KEGG/Unigenes_KEGG.txt

各列意义如下：

Term	意义
Query	Unigene 名称

KEGG_hit	该 Unigene 对应的蛋白名称
GeneName	该蛋白对应的基因名称
GeneDescription	该基因的功能注释
KOEntry	KEGG 直系同源簇 ID
KODescription	该 KO 的功能注释
EC	该 KO 对应的 KEGG 酶 ID
PathwayEntry	KEGG 数据库 Pathway 信息
PathwayDefinition	KEGG 数据库 Pathway 注释信息
KEGGLevel1	KEGG 数据库功能范畴分类 1
KEGGLevel2	KEGG 数据库功能范畴分类 2

根据 KEGG 注释总表，可以对 KEGG 注释的总体情况进行 Unigene 数目的统计。

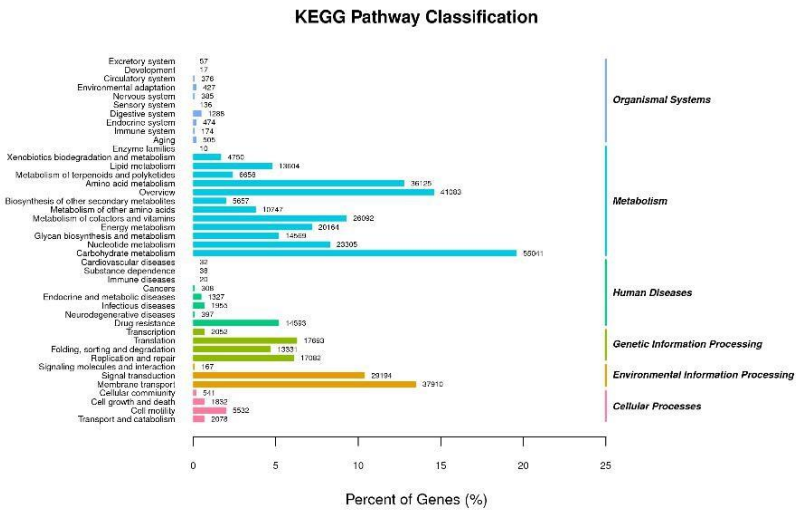


图: KEGG 注释统计图

注：左边纵轴为 KEGG PATHWAY 的二级分类信息，右边纵轴为 KEGG PATHWAY 的一级分类信息，横轴代表注释到的 Unigene 的百分比。

结果路径：summary/6_FunctionalProfiling/KEGG/KEGG_category.png

6.2.3 KEGG 丰度分析

KEGG PATHWAY 可以分为四个层次或水平，针对每一个层次进行丰度统计分析。

6.2.3.1 KEGG 柱状图

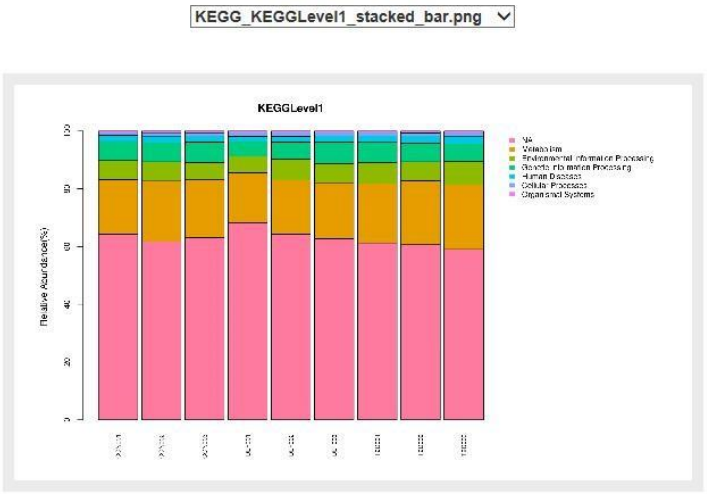


图: KEGG 丰度柱状图

注：横轴为样品名称，纵轴代表各个分类水平的相对丰度

结果路径：summary/6_FunctionalProfiling/KEGG/*/KEGG_*_stacked_bar.png

6.2.3.2 KEGG 热图

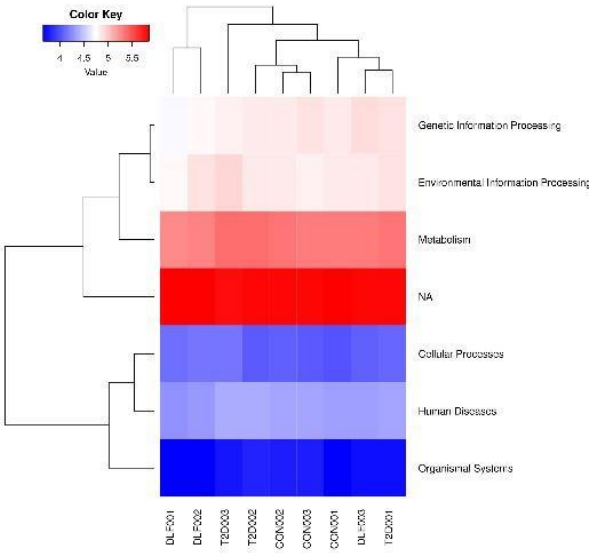


图: KEGG 热图分析**注：**图中用蓝色到红色的渐变色反映丰度由低到高的变化，越趋近于蓝色，丰度越低，越趋近于红色，丰度越高。

结果路径：summary/6_FunctionalProfiling/KEGG/*/KEGG_*_heatmap.png

6.2.3.3 PCA 分析

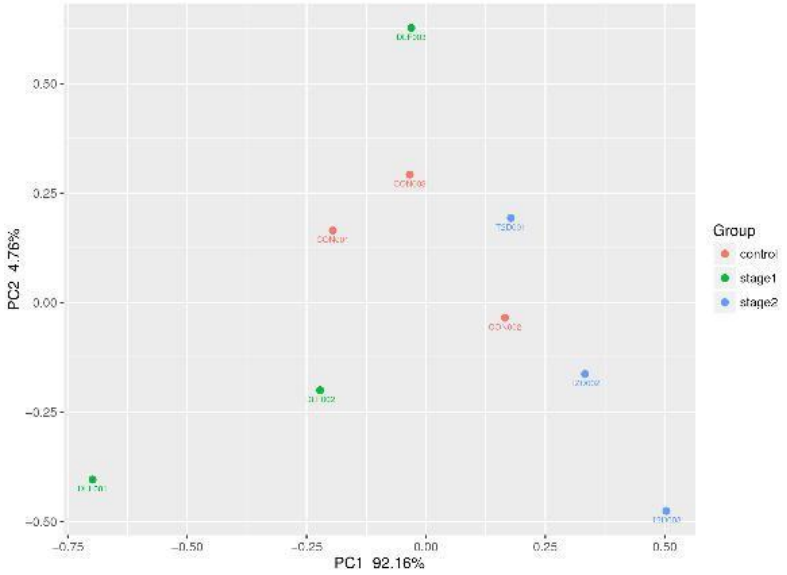
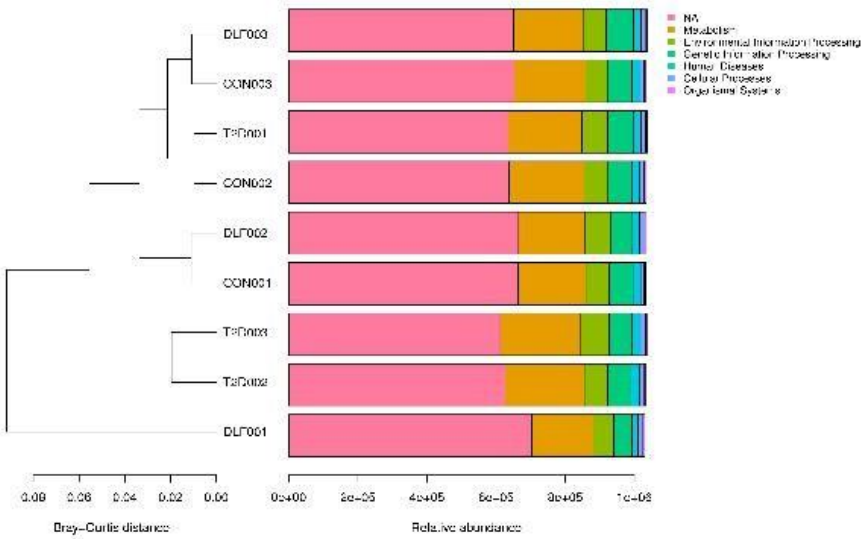


图: 物种 PCA 结果展示**注：**横坐标表示第一主成分，百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，百分比表示第二主成分对样品差异的贡献值；图中的每个点表示一个样品，同一个组的样品使用同一种颜色表示。

结果路径：summary/6_FunctionalProfiling/KEGG/*/PCA/PCA_Group.png

6.2.3.4 KEGG 聚类分析



图：KEGG 聚类分析**注：**左侧是 Bray-Curtis 距离聚类树结构；右侧的是物种相对丰度分布图

结果路径：summary/6_FunctionalProfiling/KEGG/*/KEGG_*_cluster.png

6.2.4 KEGG 差异分析

样品差异 KEGG 分析结果（以 1_KEGGLevel1 为例）

KEGGLevel1	DLF001	DLF002	DLF003	CON001	CON002	CON003	mean_stage1	mean_control	log2FC	regulation	p_value	significance	mean value
Human Diseases	1887.31	2105.09	2196.37	2322.63	2241.36	2312.76	2213.25	20629.5	23198.9	-0.17	down	0.05	0.30	yes
Metabolism	17777.62	19302.34	20100.58	19575.63	21696.28	20673.13	19854.13	190602.18	206485.25	-0.12	down	0.13	0.30	no
Cellular Processes	13931.18	14146.65	11623.84	10330.48	12023.58	11160.52	12202.71	13233.8	11171.5	0.24	up	0.13	0.30	no
NA	70342.34	76504.96	50276.07	63686.63	78365.22	49662.12	66212.63	672917.22	651336.63	0.05	up	0.28	0.39	no
Organismal System	4756.15	4477.21	5390.05	4671.19	5868.44	5884.73	5174.63	4874.47	5474.78	-0.17	down	0.28	0.39	no

结果路径：

summary/6_FunctionalProfiling/KEGG/1_KEGGLevel1/diff_abundance/Group.stage1_vs_control/Group.stage1_vs_control.diff.xlsx

6.2.5 KEGG 通路注释

基于 KEGG 不同分类层级的功能丰度表，进行了聚类、热图、PCA 分析等，并且对该层级下不同比较组的功能差异进行了统计。

其中，我们在第四层级 KOEntry 中，提供了所有差异 KO（p 值<0.05）所在的通路图链接。**老师后续如有关注的通路信息，直接点击通路对应链接即可**

结果路径：

summary/6_FunctionalProfiling/KEGG/4_KOEntry/diff_abundance/Group.stage1_vs_control/KEGG_pathway_links.xlsx

6.3 eggNOG 数据库注释

6.3.1 数据库简介 eggNOG[17 、 18] （ evolutionary genealogy of genes: Non-supervised Orthologous Groups)是欧洲分子生物学实验室 EMBL 构建的一个基因组直系同源蛋白簇及其功能注释的数据库。截至 4.5 版本，eggNOG 共包含 2031 个物种，14875530 蛋白序列， 190K 直系同源簇。 eggNOG 数据库链接：http://eggnogdb.embl.de/download/

6.3.2 eggNOG 注释

eggNOG 可以分为四个层次。第一层包括：1.信息存储和加工(INFORMATION STORAGE AND PROCESSING) 2.细胞过程和信号转导(CELLULAR PROCESSES AND SIGNALING) 3.新陈代谢(METABOLISM) 4.未确定分类(POORLY CHARACTERIZED)。第二层进一步细分成 25 个分类，每一个分类都可以用一个字母表示。第三层为共同功能描述（Consensus Functional Description)。第四层为具体的直系同源蛋白簇。

利用 DIAMOND 软件将 Unigenes 与 eggNOG 库的蛋白序列进行比对（blastp，evalue ≤ 1e-5)，取分值最高的 eggNOG 比对结果序列（Hits）的注释作为该 Unigene 序列的注释。结合 eggNOG 数据库的层次结构，我们可以统计出 eggNOG 各个层级或水平的信息。

Unigene eggNOG 注释总表

Query	eggNOG_hit	NOG	COGFunctionalCategory	COGFunctionalCategoryDescription	NOGDescription
Unigen 349741.Amuc_129		NA	NA	NA	NA
e1	3				
Unigen 349741.Amuc_142	COG05		T	Signal transduction	Serine Threonine protein kinase

e2	1	15		mechanisms	
					converts alpha-aldose to the beta-anomer. It is active on D-
Unigen 471870.BACINT_02		COG20		Carbohydrate transport	
			G		glucose, L-arabinose, D-xylose,
e3	778	17		and metabolism	
					D-galactose, maltose and lactose (By similarity)
					Aldehyde oxidase and xanthine
Unigen 693746.OBV_1723		COG15		Energy production and	
			C		dehydrogenase, molybdopterin
e4	0	29		conversion	
					binding

结果路径：summary/6_FunctionalProfiling/eggNOG/Unigenes_eggNOG.txt **各列意义如下：**

Term	意义
Query	Unigene 名称
eggNOG_hit	eggNOG 蛋白 ID
NOG	NOG 号
COGFunctionalCategory	COG 功能分类
COGFunctionalCategoryDescription	COG 功能分类描述
NOGDescription	NOG 描述

根据 eggNOG 注释总表，可以对 eggNOG 注释的总体情况进行 Unigene 数目的统计。

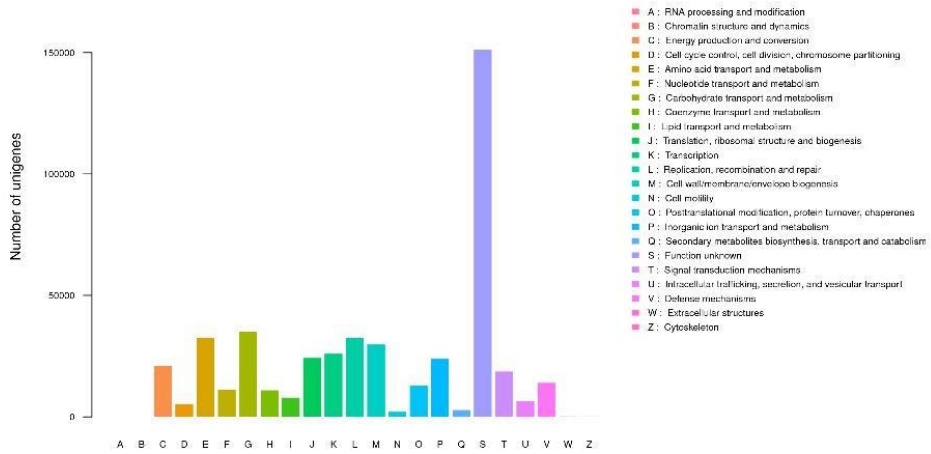


图: eggNOG 分类统计图

注：横轴为 eggNOG 的各个二级分类，纵轴代表注释到的 Unigene 数目。

结果路径： summary/6_FunctionalProfiling/eggNOG/eggNOG_category.png 根据 eggNOG 第二层级（OG 层级）的注释结果，进行每个 Unigene 物种归属统计。

eggNOG 物种归属统计

aciNOG	**96**
acidNOG	41
acoNOG	46
actNOG	6082
agaNOG	10
agarNOG	10
apiNOG	68
aproNOG	736
aquNOG	104
arNOG	1395
arcNOG	14

注：第一列为一大类物种的 OG；第二例为这一物种 OG 对应的 Unigene 数目。

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

6.3.3 eggNOG 丰度分析

eggNOG 可以分为四个层次或水平，针对每一个层次进行丰度统计分析。

6.3.3.1 eggNOG 柱状图

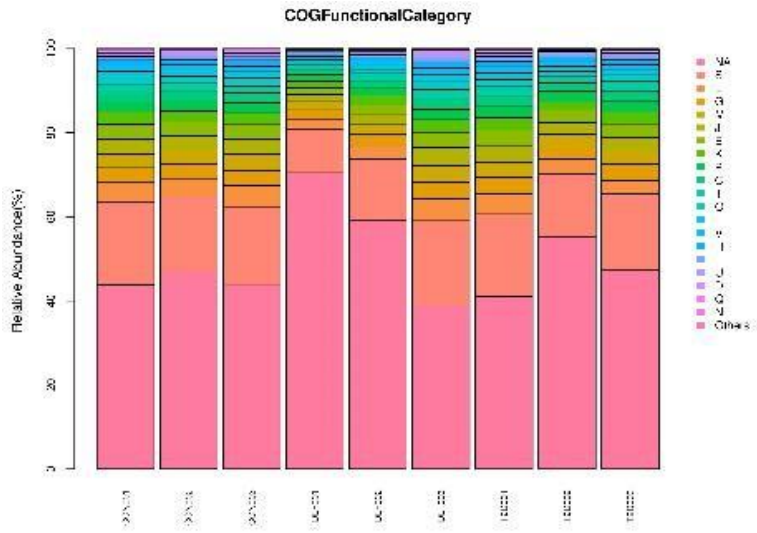
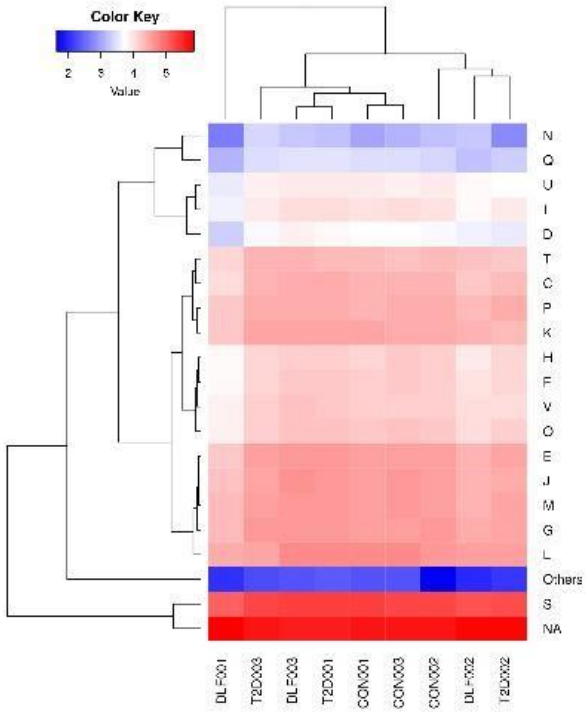


图: eggNOG 丰度柱状图

注：横轴为样品名称，纵轴代表 COGFunctionalCategory 中各个分类水平的相对丰度

结果路径：summary/6_FunctionalProfiling/eggNOG/*/eggNOG_*_stacked_bar.png

6.3.3.2 eggNOG 热图

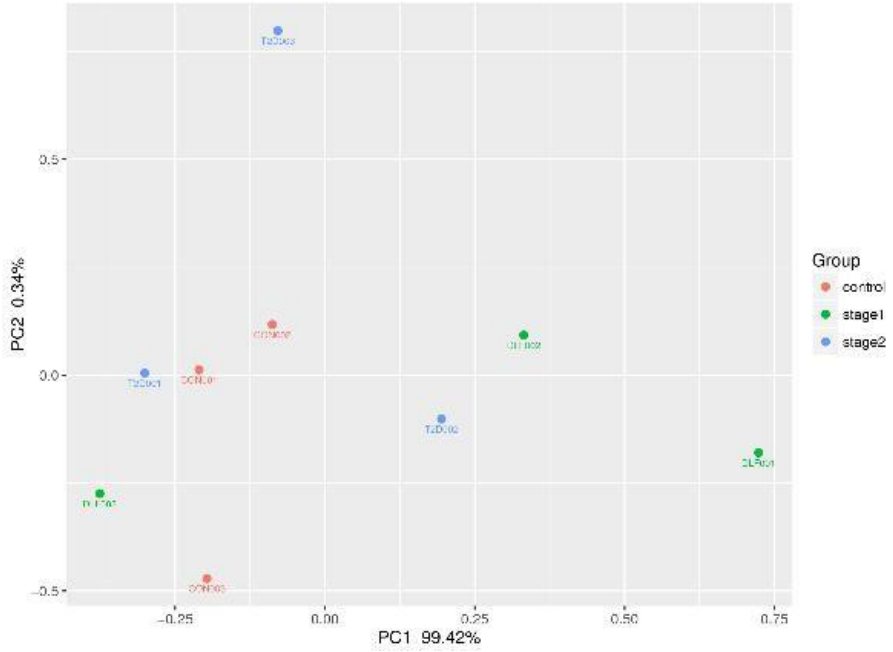


图：eggNOG 热图分析

注：图中用蓝色到红色的渐变色反映丰度由低到高的变化，越趋近于蓝色，丰度越低，越趋近于红色，丰度越高。

结果路径：summary/6_FunctionalProfiling/eggNOG/*/eggNOG_*_heatmap.png

6.3.3.3 PCA 分析

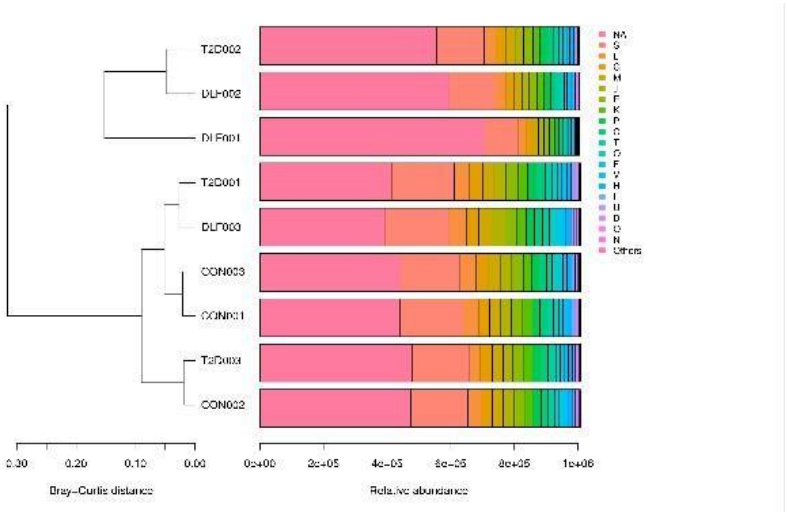


图：PCA 结果展示

注：横坐标表示第一主成分，百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，百分比表示第二主成分对样品差异的贡献值；图中的每个点表示一个样品，同一个组的样品使用同一种颜色表示。

结果路径： summary/6_FunctionalProfiling/eggNOG/*/PCA/PCA_Group.png

6.3.3.4 eggNOG 聚类分析



图：eggNOG 聚类分析

注：左侧是 Bray-Curtis 距离聚类树结构；右侧的是 eggNOG 在 COGFunctionalCategory 水平上的功能相对丰度分布图。

结果路径：summary/6_FunctionalProfiling/eggNOG/*/eggNOG_*_cluster.png

6.3.3.5 eggNOG 差异分析

样品差异 eggNOG 分析结果（以 COGFunctionalCategory 水平为例）

COGFunctionalCategory	DLF001	DLF002	DLF003	CON001	CON002	CON003	mean_stage1	mean_control	log2FC	regulation	p_value	q_value	significance	mean
P	1473	1941	2482	2400	2623	2484	2234	19656.8	25028.6	-0.35	down	0.13	0.56	no
K	1625	2354	2944	3006	2585	2534	2508	23083.1	27091.0	-0.23	down	0.28	0.56	no
H	6283.	8257.	1417	1261	1353	1467	1159	9572.67	13609.8	-0.51	down	0.28	0.56	no
U	3813.	6640.	8545.	8552.	7785.	7454.	7131.	6332.94	7930.98	-0.32	down	0.28	0.56	no
NA	7068	5961	3920	4417	4747	4431	5091	565002.	453207.	0.32	up	0.51	0.56	no
	1035	1460	2040	1975	1791	1865	1694	151200.	187760.	-0.31				

S											down	0.51	0.56	no
	29.68	42.77	29.90	73.36	63.21	44.00	80.49	78	19	0.31				
	2536	3316	5267	4760	4035	5096	4168	37067.2	46308.3	-				
L											down	0.51	0.56	no
	5.24	4.35	2.12	4.09	1.53	9.44	7.79	4	5	0.32				

结果路径：

summary/6_FunctionalProfiling/eggNOG/1_COGFunctionalCategory/diff_abundance/Group.stage1_vs_control/Group.stage1_vs_control.diff.xlsx

6.4 CAZy 数据库注释

6.4.1 数据库简介

CAZy 是一个特殊的数据库，致力于分析碳水化合物活性酶（CAZymes）的基因组，结构和生化信息等。CAZy 数据库主要涵盖[1]大功能类：糖苷水解酶（Glycoside Hydrolases，GHs），糖基转移酶（Glycosyl Transferases，GTs），多糖裂合酶（Polysaccharide Lyases，PLs），碳水化合物酯酶（Carbohydrate Esterases，CEs），辅助氧化还原酶(Auxiliary Activities，AAs)和碳水化合物结合模块（Carbohydrate-Binding Modules，CBMs)。6 个功能大类又可以进一步分成各功能小类，整个数据库具有明显的两个层次的结构。在过去的几年里，CAZy 分类家族不断扩展，

[1] .4.2 CAZy 注释利用 DIAMOND 软件将 Unigenes 与 CAZy 库的蛋白序列进行比对（blastp，evalue ≤ 1e-5），取分值最高的 CAZy 比对结果序列（Hits）的注释作为该 Unigene 序列的注释。结合 CAZy 数据库的层次结构，我们可以统计出 CAZy 各个层级或水平的信息。

Unigene CAZy 注释总表

为基因组和宏基因组的分析提供更完备的数据。CAZy 数据库链接：http://www.cazy.org/

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	NA	GlycosylTransf	GT51

		family [Pelosinus fermentans	erases JBW45]		GH13
		TonB family protein [Granulicella		Glycoside	
Unigene11	ADW67276.1		NA		GH13
		tundricola MP5ACTX9]		Hydrolases	
		glycoside hydrolase GH13 family		Glycoside	GH23
Unigene47	ALU14107.1		NA		
		[Eubacterium limosum]		Hydrolases	CBM50
		hypothetical protein ACP90_18880		Glycoside	
Unigene58	AMN56057.1		NA		
		[Labrenzia sp. CP4]		Hydrolases	
		teichoic acid ABC transporter ATP-		Carbohydrate-	
Unigene70	AJA56769.1	binding protein [Lactococcus lactis	NA	Binding	
		subsp. lactis]		Modules	

结果路径：summary/6_FunctionalProfiling/CAZy/Unigenes_CAZy.txt **各列意义如下：**

Term	意义
Query	Unigene 名称
CAZy_hit	CAZy 蛋白 ID
NCBIAnno	NCBI 注释
EC	EC 编号
CAZyLevel1	CAZy 层次一注释
CAZyLevel2	CAZy 层次二注释

根据 CAZy 分类统计信息，绘制 CAZy 分类统计图，如下所示：

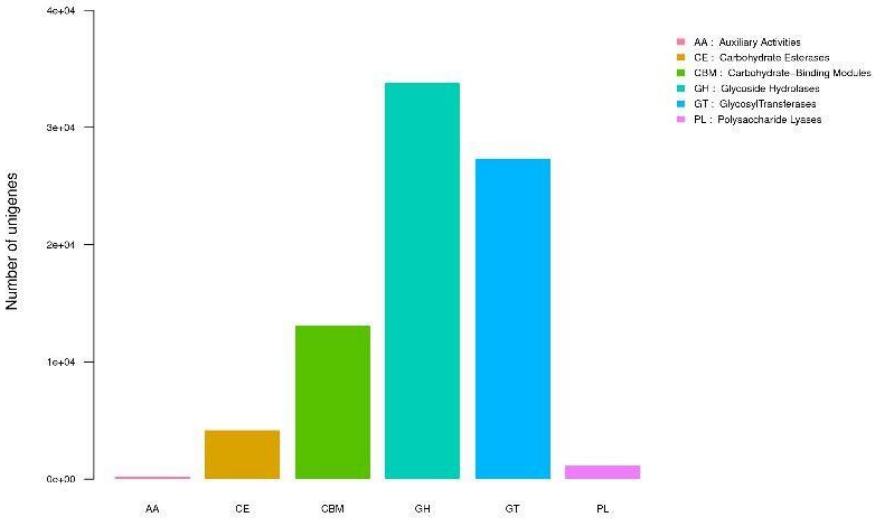


图: CAZy 分类统计图**注：**横轴为 CAZy 的各个二级分类，纵轴代表注释到的 Unigene 数目**结果路径：**summary/6_FunctionalProfiling/CAZy/CAZy_category.png

6.4.3 CAZy 丰度分析

CAZy 可以分为两个明显的层次或水平，针对每一个层次进行丰度统计分析。

6.4.3.1 CAZy 柱状图

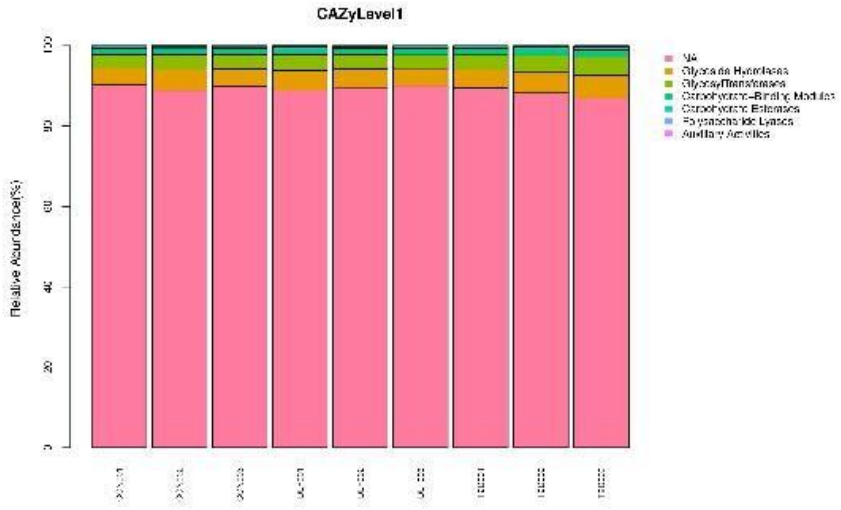
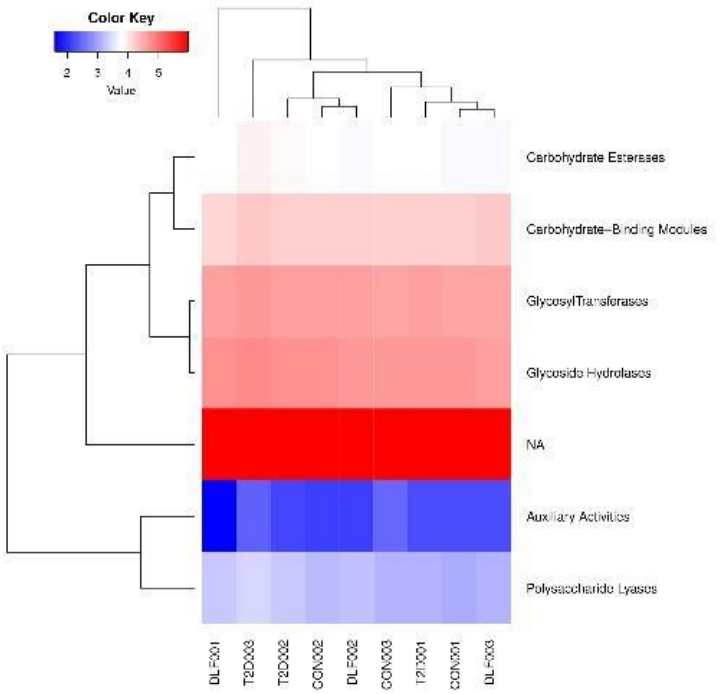


图: CAZy 丰度柱状图

注：横轴为样品名称，纵轴代表各个分类水平的相对丰度

结果路径：summary/6_FunctionalProfiling/CAZy/*/CAZy_*_stacked_bar.png

6.4.3.2 CAZy 热图分析

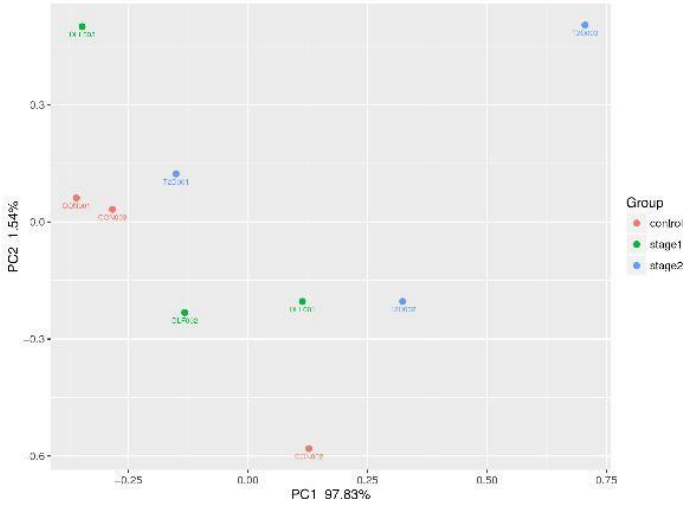


图：CAZy 热图分析

注：图中用蓝色到红色的渐变色反映丰度由低到高的变化，越趋近于蓝色，丰度越低，越趋近于红色，丰度越高。

结 果 路 径 ： summary/6_FunctionalProfiling/CAZy/*/CAZy_*_heatmap.png

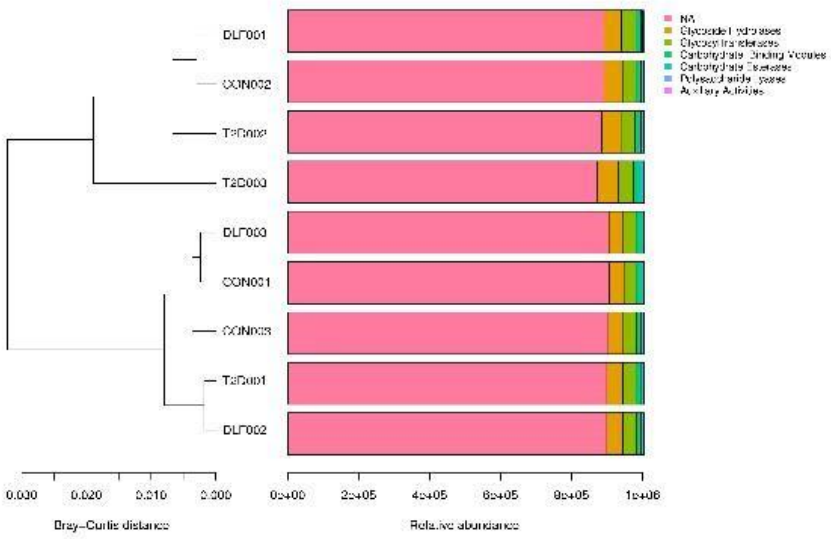
6.4.3.3 PCA 分析



图：PCA 结果展示**注：**横坐标表示第一主成分，百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，百分比表示第二主成分对样品差异的贡献值；图中的每个点表示一个样品，同一个组的样品使用同一种颜色表示。

结 果 路 径 ： summary/6_FunctionalProfiling/CAZy/*/PCA/PCA_Group.png

6.4.3.4 CAZy 聚类分析



图：CAZy 聚类分析

注：左侧是 Bray-Curtis 距离聚类树结构；右侧的是物种相对丰度分布图。

结果路径：summary/6_FunctionalProfiling/CAZy/*/CAZy_*_cluster.png

6.4.3.5 CAZy 差异分析

样品差异 CAZy 分析结果（以 CAZyLevel1 水平为例）

CAZyLevel1	DLF001	DLF002	DLF003	CON001	CON002	CON003	mean	mean_stage1	mean_control	log2FC	regulation	p_value	q_value	significance
Polysaccharide Lyases	2136.12	1751.74	1319.23	1189.63	1672.80	1367.54	1572.84	1735.70	1409.99	0.30	up	0.28	0.72	no
Auxiliary Activities	38.70	147.65	179.76	192.72	132.72	302.94	165.75	122.04	209.46	-0.78	down	0.28	0.72	no
NA	891956.00	899443.15	905287.06	905902.66	892071.92	903648.00	899718.13	898895.41	900540.86	0	down	0.51	0.72	no
GlycosylTransferases	39643.99	36799.87	35693.89	33803.21	37535.06	35902.28	36563.05	37379.25	35746.85	0.06	up	0.51	0.72	no
Carbohydrate-Binding Modules	14191.22	14561.95	16944.97	16292.55	14794.58	14884.62	15278.32	15232.71	15323.92	-0.01	down	0.51	0.72	no
Glycoside Hydrolases	49222.21	45833.89	40672.50	42004.74	51661.56	42439.18	45305.68	45242.87	45368.50	0	down	0.83	0.83	no

结果路径：

summary/6_FunctionalProfiling/CAZy/1_CAZyLevel1/diff_abundance/Group.stage1_vs_control/Group.stage1_vs_control.diff.xlsx

6.5 CARD 数据库注释

6.5.1 数据库简介

CARD（Comprehensive Antibiotic Resistance Database）寻找包括抗生素相关数据及其抗生素抗性基因的靶标蛋白。CARD 集合了不同的分子和序列数据，提供了一个独特的以 ARO（Antibiotic Resistance Ontology）形式的组织原则，并能快速识别新的未注释的基因组序列，并推测抗生素抗性基因。CARD 数据库链接：https://card.mcmaster.ca/download/

6.5.2 CARD 注释

Unigene CARD 注释总表

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

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

各列意义如下：

Term	意义

Query	Unigene 名称
CARD_hit	CARD 蛋白 ID
ARO_id	ARO 编号
ARO_name	ARO 名称
ARO_definition	ARO 具体注释

6.6 PHI 数据库注释

6.6.1 数据库简介

PHI （Pathogen Host Interactions Database）是一个病原与宿主互作数据库，收录了实验验证的真菌、卵菌和细菌病原的致病性、毒力和致病基因，感染的宿主包括动物、植物、真菌以及昆虫。因此 PHI 是发现在医学上新的基因和农学中重要病原体的宝贵资源，这可能成为化学干预的潜在目标。PHI 数据库链接：http://www.phi-base.org/

6.6.2 PHI 分类统计

Unigene PHI 注释总表

Gene_id	Species	IdentE_val ue	PHI_Acce ssion	Gene_n ame	Protein_Acc ession	NCBI_TA X_ID	Phenotype	
Unigen e36	Magnaporthe_or yzae	2.5e-118	PHI:877 0383	MGG_0	EDK03005	318829	Reduced_virule nce	
Unigen e70	Magnaporthe_or yzae	1.7e-105	PHI:1018	ABC3	AAZ81480	318829	Loss_of_patho genicity	
Unigen e89	Cryptococcus_g attii	1.3e-64	PHI:3332	snf7	A0A095CFR8	552467	loss_of_pathog enicity	
Unigen e103	Xanthomonas_h ortorum	2.8e-14	PHI:3591	rpfF	G3K6H9	56454	reduced_virule nce	
Unigen	Edwardsiella_ict	1.6e-55.9	PHI:2962	glyA	C5BEV2	67780	Reduced_virule	

e123	126	aluri	nce

结果路径：summary/6_FunctionalProfiling/PHI/Unigenes_PHI.txt

各列意义如下：

Term	意义
Gene_id	Unigene 名称
Identity	比对相似度
E_value	比对
PHI_Accession	PHI 编号
Gene_name	基因名称
Protein_Accession	蛋白编号
NCBI_TAX_ID	NCBI 物种编号
Species	物种信息

7. 基因表达差异及富集分析

差异表达的基因是宏基因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

7.1 差异基因表达分析

差异基因分析结果展示

Unigene_ID	LF0102	DLF03	DLF001	CON002	CON003	mean_stanage1	mean_control	log2FC	regulation	q_value	p_value	significance
Unigene272	0	0	0	0.01	0.01	3.42	0.57	1.14	-Inf	down	0.03	0.65	yes
Unigene6440	0	0	0	0.01	0.01	4.79	0.80	1.60	-Inf	down	0.03	0.65	yes

Unigene28							0.7						
	0	0	0	0.01	0.01	4.56		0	1.52	-Inf	down	0.03	0.65 yes
0211							6						
Unigene58							0.0						
	0	0	0	0.01	0.01	0.01		0	0.01	-Inf	down	0.03	0.65 yes
8018							0						
							1.2						
Unigene3	0	0	0	5.97	0.08	1.12		0	2.39	-Inf	down	0.04	0.65 yes
							0						

差异表达的基因是宏基因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)

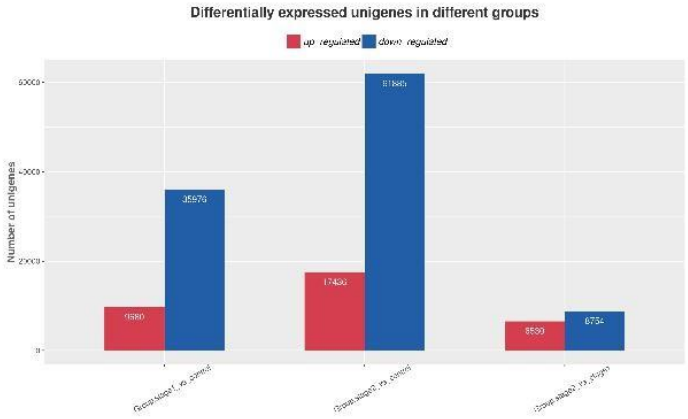
结果路径：

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

各列意义如下：

Term	Description
Unigene_ID	Unigene 的 ID 号
mean	平均基因表达量
log2FC	log2FC，即对 foldChange 值取 log2
p_value	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
q_value	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
regulation	基因上调或者下调
significance	差异显著或不显著

7.2 差异显著基因上调下调个数统计图

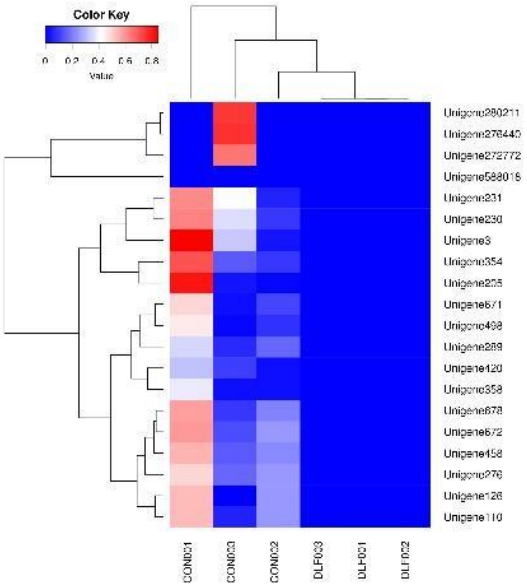


图：差异显著基因个数统计图

结果路径：summary/7_DifferentialExpression/diff_regulation.png

7.3 差异基因表达水平聚类分析

差异基因聚类分析用于判断基因在不同的试验条件下调控模式的聚类模式。根据样品基因表达谱的相似程度，将基因进行聚类分析，直观地展示了基因在不同样本之间（或不同处理之间）的表达情况，由此获得生物学相关信息。为了更好直观反映聚类表达模式，我们采用了 log10(TPM+1)进行基因表达展示。同时也将差异基因的 TPM 值通过 Z 值的方式进行基因表达展示。



图：差异基因表达水平聚类分析

结果路径：

summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_diff_heatmap.pn

7.4 差异基因 GO 富集分析

7.4.1 GO 富集结果

差异基因分析结果展示

GO_ID	GO_Term	GO_Function	GO_Level	SUnigene number	TSUnigene number	BUnigene number	TBUnigene 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 activity	molecular_function	9	528	23940	6293	355487	0.00
GO:0005615	extracellular space	cellular_component	5	576	23940	6998	355487	0.00
GO:0004181	metallocarboxypeptidase activity	molecular_function	9	528	23940	6364	355487	0.00

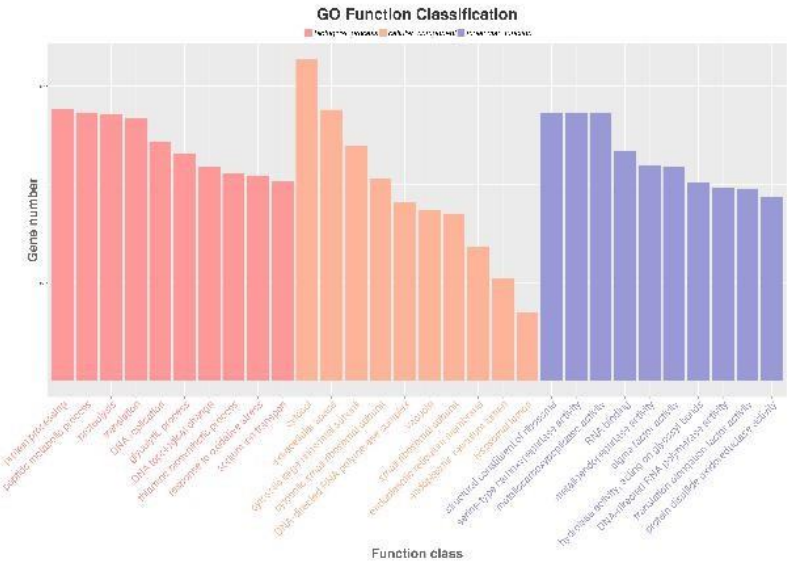
GO:0004222	metalloendopeptidase activity	molecular_function	8	154	23940	1604	355487	0.00
GO:2000234	positive regulation of rRNA processing	biological_process	13	11	23940	39	355487	0.00
GO:0006096	glycolytic process	biological_process	7	205	23940	2315	355487	0.00

结果路径：

summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_GO_enrichment.xlsx **各列意义如下：**

Term	**意义**
GO_ID	Gene ontology ID
GO_Term	GO 信息
GO_function	GO ontology 的功能分类
GO_class	GO 的层级
S gene number	注释为特定 GO 的具有显著性差异的基因数
TS gene number	显著差异的基因数
B gene number	注释为特定 GO 的基因数
TB gene number	基因总数
Pvalue	p 值

7.4.2 GO 富集柱状图



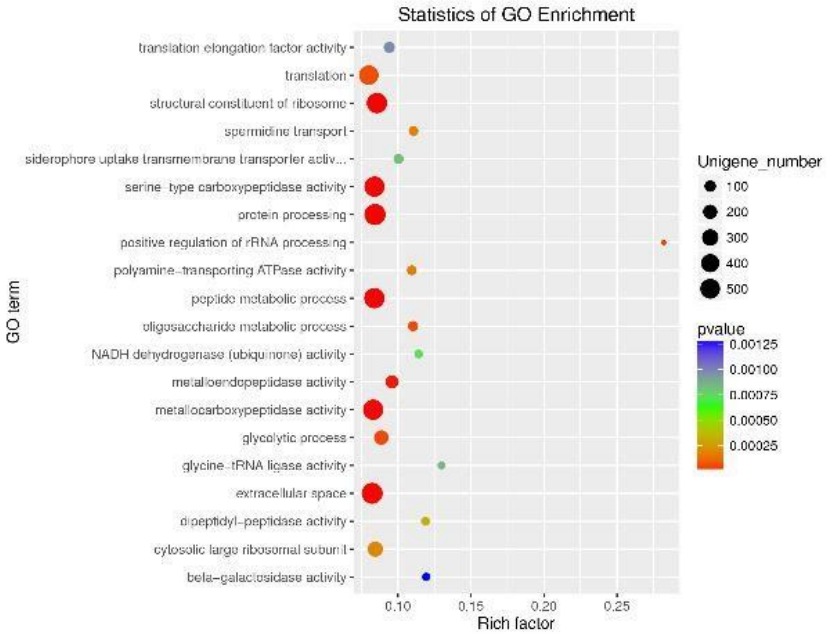
图：GO富集柱状图

结果路径：

summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_GO_enrichment_histogram.png

7.4.3 差异基因 GO 富集散点图

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



图：GO富集柱状图

结果路径：

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

7.5 差异基因 KEGG 富集分析

7.5.1 KEGG 富集结果

KEGG 富集结果

PathwayEntry	PathwayDefinition	TS 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 biosynthesis	545	19130	7119	281285	0.00
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

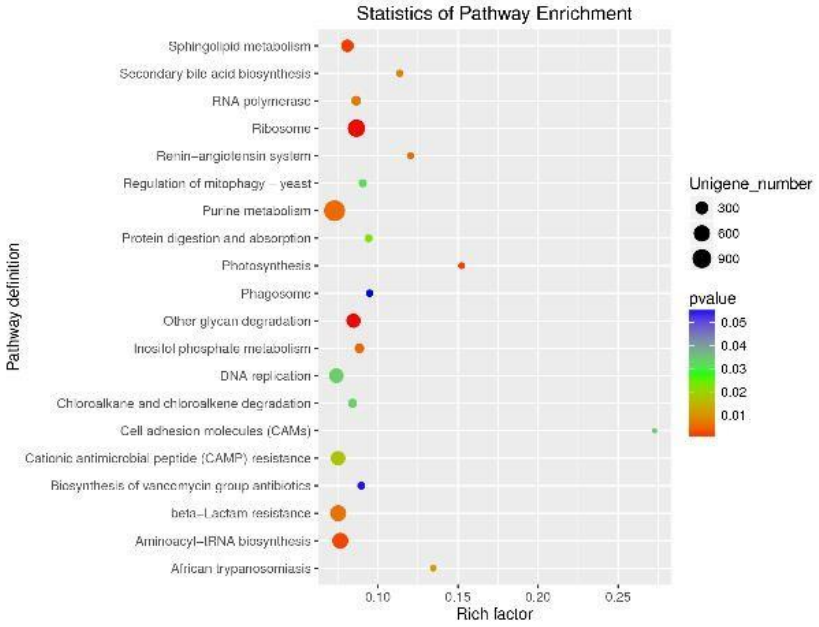
结果路径：

summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_KEGG_enrichment.xlsx **各列意义如下：**

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

7.5.2 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。



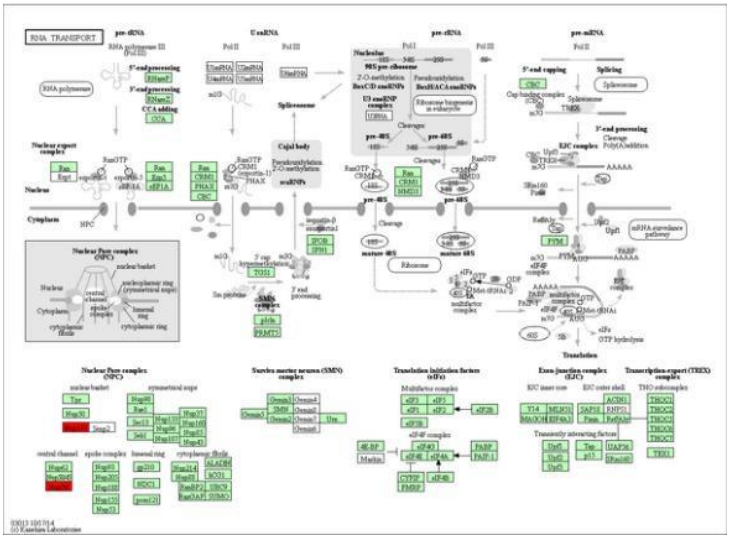
图：差异KEGG富集图

结果路径：

summary/7_DifferentialExpression/Group.stage1_vs_control/Group.stage1_vs_control_KEGG_enrichment_scatterplot.png

7.5.3 KEGG 通路图示例

Pathway 通路图展示与说明：红色红代表注释到某个 ko 上且上调的显著差异表达基因；蓝色代表注释到某个 ko 节点且下调的显著差异表达基因；橙色代表注释到某个 ko 节点不仅有上调又有下调的显著差异表达基因。方框内的 4 位数字表示各种酶的 EC 编号；空心圆圈表示小分子化合物；实心箭头表示生化反应的方向；虚线箭头连接其他的相关代谢途径。



8.1 FASTQC 测序质量控制

测序得到的原始数据为图像文件，经过 base calling，我们得到以 FASTQ 格式存储的结果文件，称之为 rawdata 或 raw reads。FASTQ 文件存储 reads 的序列及测序质量信息。每个 read 由四行描述，例如：

@HISEQ04:76:000000000-ABHVH:1:1101:17453:3799 1:N:0:ATCTCGTA

TTCGCTCCCCTAGCTTTCGTCTCTCAGTGTCAGTGTCGGCCCAGCAGAGTGCTTTCGCCGTTG

GTGTT

+

GGFDFEGGFGFF<ffgggggg8ffggff9ffgggfgggdgbeggggggffggefgggggggggggg

< p="">

第 1 行和第 3 行是序列名称；第 2 行是序列；第 4 行是序列的测序质量，其中每个 ASCII 编码字母与第 2 行每个碱基相对应。

从 Illumina 测序仪 base calling 软件 CASAVA 升级到 1.8 版本开始，测序得到的数据质量都是 Sanger 格式，Sanger 格式可以用 ASCII 表中的 33 到 126 号的字母代表由 0 到 93 的测序质量值。测序质量值越高代表碱基测序错误率越小。表 8 为 Illumina 测序错误率与测序质量值简明对应关系（CASAVA 1.8+）。具体地，如果测序错误率用 p 表示，Illumina 碱基质量值用 Qsanger 表示，则有下列关系：

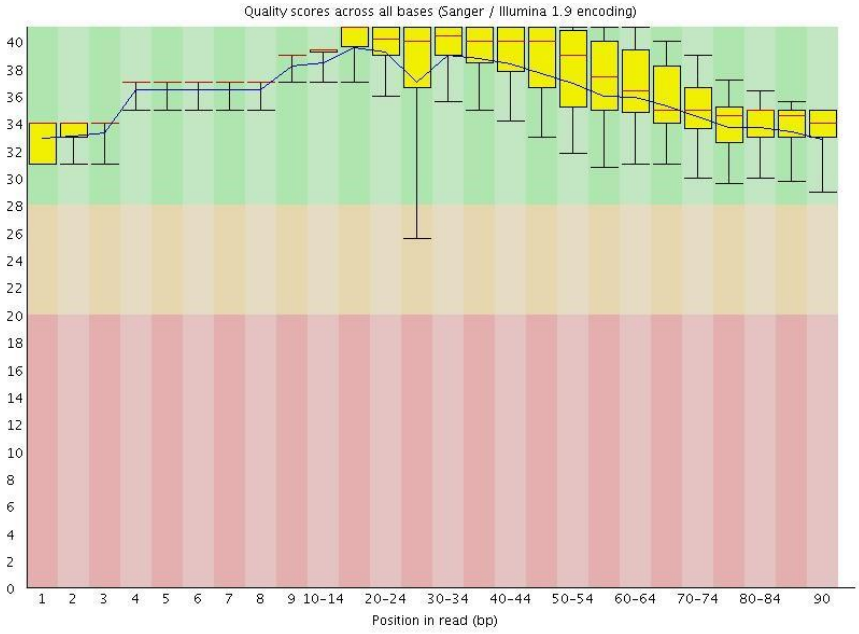
$$Q_{sanger} = -10 \log 10p$$

Illumina测序错误率与测序质量值简明对应关系

各列意义如下：

Sequencing error rate(p)	Sequencing quality(Qsanger)	Corresponding ASCII character
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

对 Read 1 和 Read 2 进行以上操作后，进行有效数据质量评估。由于高通量测序错误率对结果的影响，我们要对优化后的数据进行质量评估。如图所示，横坐标表示测序序列的碱基位置，纵坐标表示碱基质量值（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。

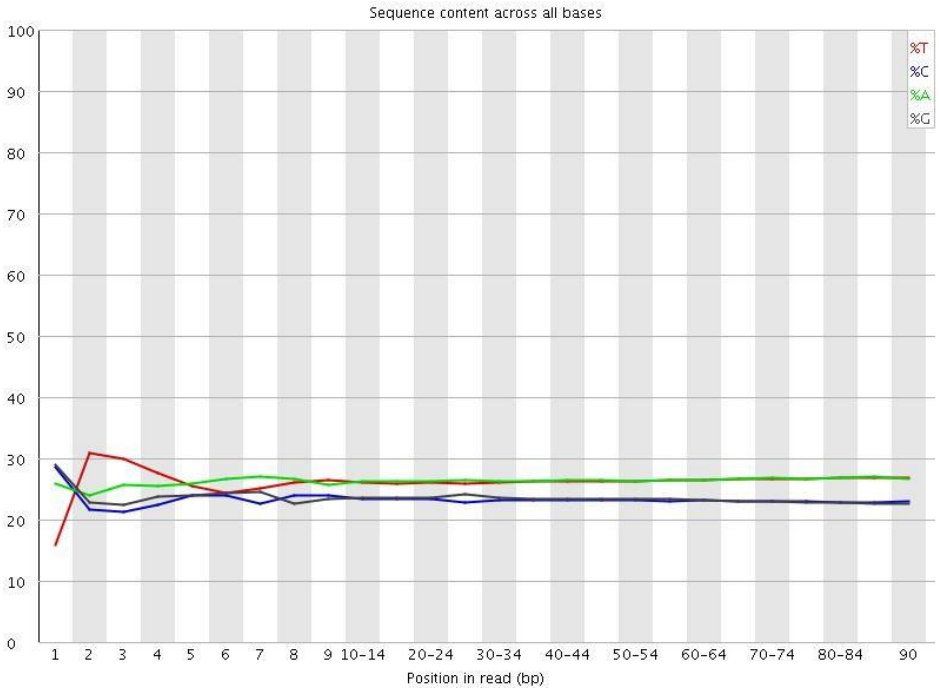


结果路径：

summary/2_CleanData/FastQC_result/*/*_*_fastqc/Images/per_sequence_quality.png

序列组成分布用于分析是否因测序或建库所带来的碱基含量分离现象，以影响后续样品的定

量分析。如图所示，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。 正常情况下 A 与 T、C 与 G 出现的频率应该是接近的，而且没有位置差异。



结果路径：

summary/2_CleanData/FastQC_result/*/*_*_fastqc/Images/per_base_sequence_content.png

8.2 样品间相关性分析

为了实验的严谨性，在实验设计中一般会设置生物学重复。样品间基因丰度相关性是检验实验可靠性和样本选择是否合理性的重要指标。相关系数越接近 1，表明样品之间基因丰度模式的相似度越高。

图：样品间相关系数热图

注：不同颜色代表相关系数的高低；相关系数与颜色间的关系见右侧图例说明；颜色越浅代表样品间相关系数越大。

结果路径：summary/4_GenePredict/7_correlation_heatmap.png

参考文献

[1] William B Whitman, David C Coleman, and William J Wiebe. Prokaryotes: the unseen majority. Proceedings of the National Academy of Sciences, 95(12):6578–6583, 1998.

[2] John C Wooley, Adam Godzik, and Iddo Friedberg. A primer on metagenomics. PLoS Comput Biol, 6(2):e1000667, 2010.

[3] Michael S Rapp´e and Stephen J Giovannoni. The uncultured microbial majority. Annual Reviews in Microbiology, 57(1):369–394, 2003.

[4] Jo Handelsman, Michelle R Rondon, Sean F Brady, Jon Clardy, and Robert M Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & biology, 5(10):R245–R249, 1998.

[5] Kevin Chen and Lior Pachter. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol, 1(2):106–112, 2005.

[6] Susannah Green Tringe and Edward M Rubin. Metagenomics: Dna sequencing of environmental samples. Nature reviews genetics, 6(11):805–814, 2005.

[7] Justin Kuczynski, Christian L Lauber, William A Walters, Laura Wegener Parfrey, Jos´e C Clemente, Dirk Gevers, and Rob Knight. Experimental and analytical tools for studying the human microbiome. Nature Reviews Genetics, 13(1):47–58, 2012.

[8] Eva Bellemain, Tor Carlsen, Christian Brochmann, Eric Coissac, Pierre Taberlet, and H ˚ avard Kauserud. Its as an environmental dna barcode for fungi: an in silico approach reveals potential pcr biases. Bmc Microbiology, 10(1):189, 2010.

[9] Thomas J Sharpton. An introduction to the analysis of shotgun metagenomic data. Frontiers in plant science, 5, 2014.

[10] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. Nucleic acids research, 42(D1):D199–D205,2014.

[11] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research,28(1):27–30, 2000.

[12]Peng Y, Leung H C M, Yiu S M, et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth[J]. Bioinformatics, 2012, 28(11): 1420-1428.

[13] Alexey Gurevich, Vladislav Saveliev and Nikolay Vyahhi,: quality assessment tool for genome assemblies. Bioinformatics, 29(8):1072–1075, 2013.

[14] Benjamin Buchfink, Chao Xie and Daniel H Huson. Fast and sensitive protein alignment using diamond. Nature methods, 12(1):59–60, 2015.

[15] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. Nucleic acids research, 42(D1):D199–D205,2014.

[16] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1):27–30, 2000.

[17] Sean Powell, Kristoffer Forslund, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Jaime Huerta-Cepas, Toni Gabald ´ on, Thomas Rattei, Chris Creevey, Michael Kuhn, et al. eggnog v4. 0: nested orthology inference across 3686 organisms. Nucleic acids research, page gkt1253, 2013.

[18] Lars Juhl Jensen, Philippe Julien, Michael Kuhn, Christian von Mering, Jean Muller, Tobias Doerks, and Peer Bork. eggnog: automated construction and annotation of orthologous groups of genes. Nucleic acids research, 36(suppl 1):D250–D254, 2008.