GENESKY 天昊生物 遗传分析整体解决方案专家

版本号 V1.0

MethylTarget™ 目标区域甲基化测序标准项目报告

上海市浦东新区康桥路787号9号楼
400-065-6886
http://biotech.geneskies.com/

目录

1. 项目信息

项目名称	上海王明华－42 个样品 16B1212A 项目 MethylTarget 富集测序		
项目编号	16B1212A-2		
客户单位	苏州大学		
项目样品信息			
物种信息	Human	参考基因组信息	Hg19
项目类型	MethylTarget 目标区域甲基化测序	测序信息	Illumina2×150bp
备注			
天昊生物联系人信息			
实验技术支持	黄泽斌	电话	021-50802060-131
		邮箱	huangzb@geneskies.com
数据分析支持	李才华	电话	021-50802060-117
		邮箱	lch@geneskies.com
项目整体支持	项目专家组	电话	021-50802060-142
		邮箱	
项目售后支持	熊伟明	电话	021-50802060-135
		邮箱	xiongwm@geneskies.com
项目审批			
确认报告显示的内容与合同要求完全一致，同意项目结束，批准本项目总结报告发送。 签名：李才华 2017 年 12 月 28 日			

2. 实验结果快速导航

本次实验的数据分析内容有哪些？
总表位于本报告【3.5.2】中的项目分析内容
本次实验获得测序原始数据的质量如何？
每个样本的原始数据测序质量见文件夹：Data_statistics.xlsx，结果解释见本报告【0】
本次实验各样本甲基化水平分析内容有哪些？
每个样本甲基化数据统计见文件夹：methylation.xlsx，结果解释见本报告【4.3】
本次实验甲基化相似性分析内容有哪些？
分析结果见：pca.pdf 和 heatmap.pdf，结果解释见本报告【4.4】
本次实验 SNV 分析内容有哪些？
每个样本 SNV 分析结果见 SNV_bwa.xlsx，结果解释见本报告【4.5】
本次实验结果都包含了什么数据？
详细结果列表见【6.1】的结果目录文件列表

3. 实验流程

3.1. MethylTarget™ 简介

作为一种重要的表观遗传学标记，DNA 甲基化修饰参与调控基因表达、染色质稳定性等最基础的生命活动。甲基化多态性是个体表型差异的重要原因，甲基化变异也可能导致个体表型异常。已有大量研究表明甲基化异常和癌症，糖尿病，帕金森等复杂疾病密切相关，甲基化检测对生物学、转化医学等研究具有重要意义。基于芯片以及二代测序平台，已有多种实验方法能够用于全基因组甲基化检测，包括 Illumina 850K 甲基化芯片，WGBS（Whole-genome bisulfite sequencing，全基因组重亚硫酸盐测序），RRBS（Reduced representation bisulfite sequencing，简化基因组重亚硫酸盐测序）等。而针对特定位点/区域的甲基化检测，目前仍多通过传统方法，如 MSP（Methylation-specific PCR，甲基化特异 PCR）和 BSP（Bisulfite-Sequencing PCR，甲基化 PCR 测序）的方法实现，通量小，成本高，且无法准确计算位点/区域的甲基化水平。天昊生物依托公司多重 PCR 的专利技术，结合二代测序平台开发的 MethylTarget 技术，实现对多个特定 CpG 岛同时捕获测序，并凭借高深度测序数据，能够准确计算每个 CpG 位点的甲基化水平。该技术准确性高，灵活性强，性价比优，为 DNA 甲基化研究提供有效的实验工具。

3.2. 实验流程图

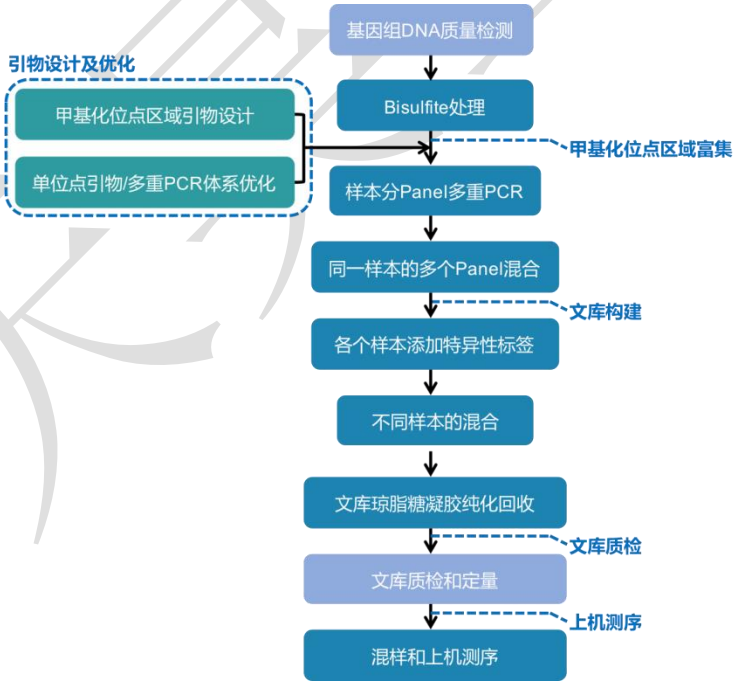

图 3-1 MethylTarget 实验流程图

3.3. 实验试剂与仪器

- 主要骑器材

器材名称	器材供应商
ABI 2720 Thermal Cycler	Applied Biosystems, Waltham, MA, USA
Eppendorf 5810R Centrifuge	Eppendorf, Hamburg, Germany
XiangYi H1650-W	XiangYi, Hunan, China
EP600 Gel electrophoresis	上海玉柏实业有限公司
NanoDrop	NanoDrop technologies, Wilmington, DE, USA
Invitrogen Qubit Spectrophotometer	Invitrogen, Carlsbad, CA, USA

- 主要试剂

试剂名称	试剂供应商
EZ DNA Methylation-Gold Kit	ZYMO, CA, USA
TIANGEN Gel Extraction kit	TIANGEN, Beijing，China
10× Reaction buffer	TaKaRa, Dalian, China
HotStarTaq polymerase	TaKaRa, Dalian, China

3.4. 实验步骤

！以下实验步骤为标准步骤说明，限于理解实验过程。

1) 样品质控

a. 琼脂糖凝胶电泳检测基因组 DNA 完整性：电泳条带清晰可见，无明显降解，且无 RNA 污染。

b. Nanodrop 2000 检测基因组 DNA 质量：浓度≥20 ng/μL，总量≥1 μg（可供 10 个多重 PCR Panel 的检测），OD260/280 =1.7~2.0，OD260/230 ≥1.8；

2) 引物设计与单位点 PCR 条件优化

基于公司的专利软件，针对客户提供的目标区域设计高质量的测序引物。挑选能够以经重亚硫酸盐处理的人基因组为模板，扩增获得清晰单一条带的引物用于后续实验。

3) 多重 PCR 引物 panel 优化

将经步骤 2）优化后的引物混合为多重 PCR 引物 panel。并使用公司多重 PCR 的专利技术，以标准人基因组为模板进行扩增。基于毛细管电泳的特殊方法，我们能够判断多重体系中每对引物是否高效、特异地进行扩增，并以此调整，优化多重 PCR panel 中的引物组成及浓度。

4) 重亚硫酸盐处理

使用 EZ DNA Methylation-Gold Kit 进行样本处理，将基因组 DNA 未被甲基化修饰的胞嘧啶 C 转化为胸腺嘧啶 U。

5) 样本目标片段多重 PCR 反应

使用优化后的多重 PCR 引物 panel，以转化后的样品基因组为模板，进行多重 PCR 扩增。经质控后，将以同一个样品基因组 DNA 为模板的所有多重 PCR 引物 panel 的扩增产物混合，并确保每个位点引物扩增产物的量相当。

6) 样本添加特异性标签序列

利用带有 Index 序列的引物，通过 PCR 扩增向文库末端引入和 illumina 平台兼容的特异性标签序列。反应采用 11 个循环数的 PCR 程序，尽可能降低 PCR 的倾向性。

7) 定量后上机测序

将所有样品 Index PCR 扩增产物等量混合，并经割胶回收获得最终的 MethylTarget 测序文库，文库的片段长度分布经 Agilent 2100 Bioanalyzer 验证。文库摩尔浓度精确定量后，最终于 Illumina Hiseq/Miseq 平台，以 2×150 bp/2×250 bp 的双端测序模式进行高通量测序，获得 FastQ 数据。

3.5. 数据分析流程

3.5.1. 数据分析流程图

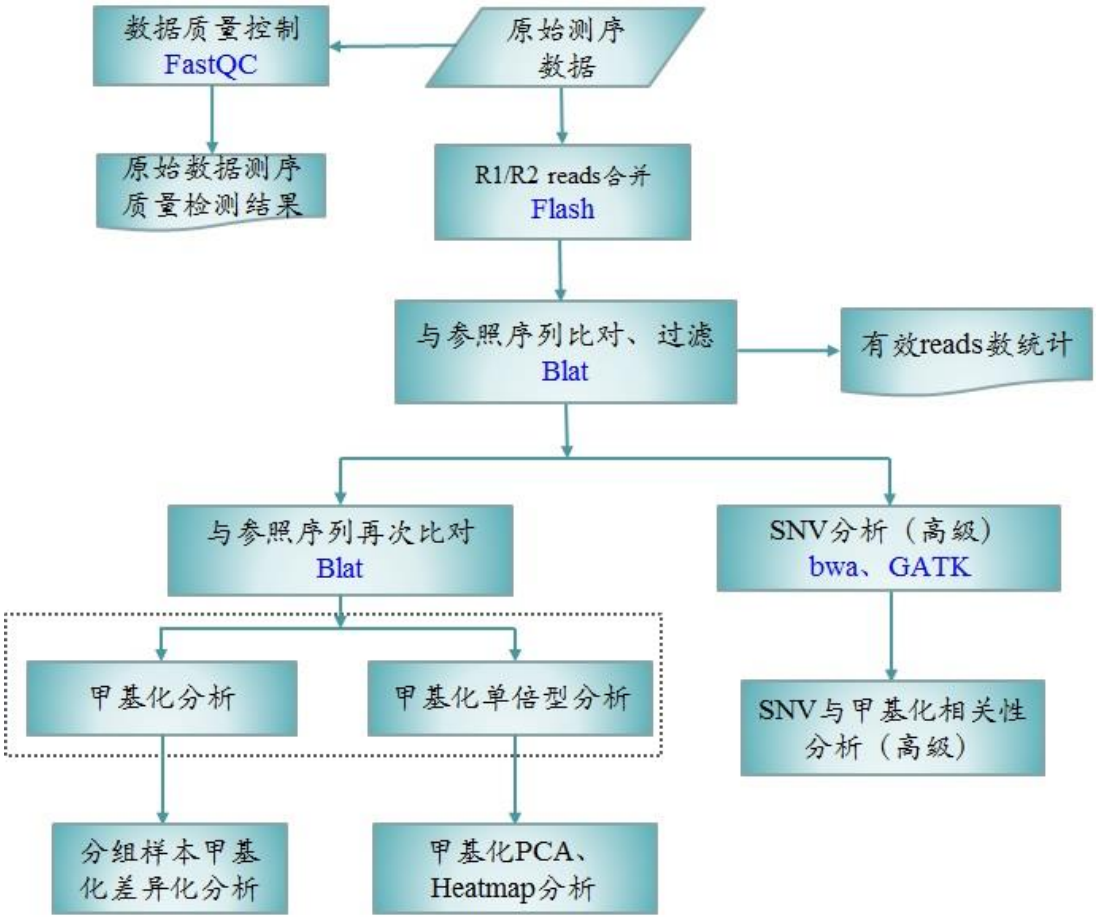

图 3-2 MethylTarget 数据分析流程图

3.5.2. 生物信息分析内容

本项目生物信息学分析内容见下表（√打勾部分）：

生物信息分析内容

基本分析	
引物设计与 CpG 岛注释	√
原始数据整理及质量评估	√
胞嘧啶转**化效率**统计	√
CpG 位点甲基化水平计算	√
CpG 位点甲基化单倍型分析	√
甲基化水平相似性分析（PCA 分析，聚类分析）	√
甲基化水平差异分析	
高级分析	
SNV 分析	√
SNV 与甲基化相关性分析	
个性化分析	

4. 数据分析结果

4.1. 引物信息

该部分分析内容给出：

- ✓ 针对目标区域设计的引物序列
- ✓ 引物扩增的参考序列在参考基因组上的注释信息

4.1.1. 引物

采用 primer3 进行对 Bisulfite 处理之后的序列进行引物设计，软件 http://primer3.ut.ee/

表 4-1 引物信息

Primer Name	seq
ADCY2_F	GTATTTTTAAAGTGTGATTGAAGTTGAG
ADCY2_R	CCCATCCTCTCCACCCTCT
AHRR_F	TTTTGTAATGGGGTTTGGTTGT
AHRR_R	TCTCCCAAATACAAACCCAAAAA
ARRB2_F	TTGTTGATTTGTGGGAGGTTTTT
ARRB2_R	AAACAAATATCCTTTCACACTTCCTAC
BCAN_F	GTGGTTTTGGAGGATTTAGAGG
BCAN_R	CCTCTATAAACRAATAAAACCRACTAAAA
DLX5_F	GTAGTTTATTTAATAGAGTGTTTYGGAGGTT
DLX5_R	TCAACCACCACCCTCATACC

注：Primer Name：引物名称； seq：引物序列。

***仅最多展示 10 条引物，引物信息目录：methylation_info.xlsx

4.1.2. 参考序列信息

对甲基化引物扩增的参考序列使用 megablast 比对到参考基因组，同时根据 refGene 信息获取基因的转录本信息，对参考序列进行注释（表 4-2）。

表 4-2 参考序列信息

Target	Chr	Gene	mRNA	mRNA_Strand	TSS	Start	End	Length	Target_Strand	Distance2TSS
ADCY2_	5	ADCY2	NM_020546	+	7396342	7395281	7395481	201	+	-1061
AHRR_	5	AHRR	NM_001242412	+	304290	374201	374465	265	+	69911
ARRB2_	17	ARRB2	NM_001257328	+	4613788	4615035	4615197	163	+	1247
BCAN_	1	BCAN	NM_021948	+	156611739	156616408	156616595	188	+	4669
DLX5_	7	DLX5	NM_005221	-	96654143	96650039	96650220	182	+	3923
DMRTA2_	1	DMRTA2	NM_032110	-	50889119	50888416	50888644	229	+	475
FGF8_	10	FGF8	NM_006119	-	103535759	103536294	103536553	260	+	-794
GRIN1_	9	GRIN1	NM_000832	+	140033608	140051074	140051260	187	+	17466
GSC_	14	GSC	NM_173849	-	95236499	95235313	95235470	158	+	1029
HIST1H3G_	6	HIST1H3G	NM_003534	-	26271612	26271577	26271773	197	+	-161

注："Target"： 目标片段名称；Chr：染色体；Gene：基因；mRNA：距离产物较近的 mRNA(基因存在多种转录本的情况)；mRNA_Strand：mRNA 方向；TSS：mRNA 转录起始位点；Start：产物在参照基因组上的起始位置；End：产物在参照基因组上的终止位置；Length：产物长度；Target_Strand：产物方向；Distance2TSS：产物相对 TSS 的距离；注意：当不存在参考基因组或设计的参考序列与参考基因组不一致时，该表格基因组信息为空。

***仅最多展示 10 个扩增子的 SNV 分析，结果目录：methylation_info.xlsx

4.2. 原始数据质量评估

该部分分析内容给出：

- ✓ 各样本原始测序数据量及数据质量评估；
- ✓ 胞嘧啶转化效率统计
- ✓ 有效 reads 筛选统计；

4.2.1. 测序数据量统计

对于原始测序数据，统计各样品的 reads 数目及测序质量。

表 4-3 原始测序数据统计

Sample	R1				R2			
	Total Reads	Reads length	Q20 (%)	Q30 (%)	Total Reads	Reads length	Q20 (%)	Q30 (%)
10	70594	35-151	99.99%	97.90%	70594	35-151	99.36%	95.81%
11	92439	35-151	99.99%	98.14%	92439	35-151	99.35%	95.46%
12	84609	35-151	99.99%	97.95%	84609	35-151	99.31%	95.61%
13	80667	35-151	99.99%	97.94%	80667	35-151	99.21%	95.23%
14	86708	35-151	99.99%	97.89%	86708	35-151	99.09%	95.16%
15	69596	35-151	100.00%	98.02%	69596	35-151	99.29%	95.88%
16	76334	35-151	100.00%	98.05%	76334	35-151	99.44%	95.82%
17	82974	35-151	99.99%	97.86%	82974	35-151	99.38%	95.70%
18	93099	35-151	99.99%	97.88%	93099	35-151	99.07%	95.29%
19	99608	35-151	99.99%	98.04%	99608	35-151	99.31%	95.30%

注：“Total Reads”和“Read Length”代表原始序列数量和测序读长，“Q20 (%)”和“Q30 (%)”分别表示原始数据中测序质量值 Q 不低于 20 和 30 的序列比例。

***仅最多展示 10 个样本，剩余结果目录：Data_statistics.xlsx/QC_RawData sheet

4.2.2. 原始数据碱基质量分布

碱基质量表述方法见第 5.3 小节。单个样品整体的测序质量可由每个位置碱基的质量分布直观展示（图 4-1），同时展示每个位置碱基组成分布图，用于评估测序结果是否均衡（图 4-2），

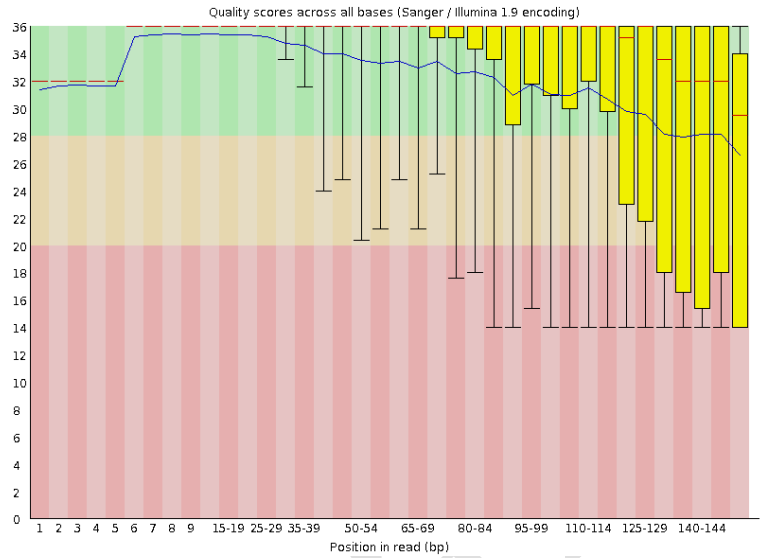

图 4-1 原始数据碱基质量分布图（示意图）

注：示例样品 R1 端测序错误率分布图，横坐标为 reads 的碱基位置，纵坐标为 reads 的碱基质量值 Q 值。采用箱线图的方式展示对应位置的碱基质量分布。背景色根据碱基质量从优到劣分为绿色、黄色、红色三个部分，直观展示碱基质量。

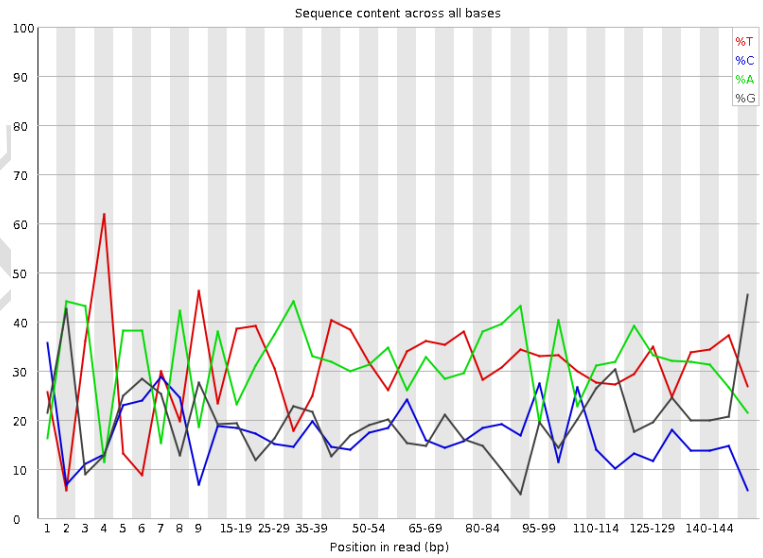

图 4-2 原始数据碱基组成分布图（示意图）

注：示例样品 R1 端序列的碱基含量分布图，横坐标为 reads 的碱基位置，纵坐标以不同颜色展示单碱基 A/T/G/C 所占的比例。

****示例样品 CASE 的原始数据碱基错误率分布图（CASE_ErrorRate.png）、原始数据碱基质量分布图（CASE_R1_per_base_quality.png 和 CASE_R2_per_base_quality.png）、原始数据碱基组成分布图(CASE_R1_per_base_sequence_content.png 和 CASE_R2_per_base_sequence_content.png）。结果目录为 Image/fastqc。

4.2.3. 有效 reads 筛选统计

使用软件 FLASH(FLASH: Fast length adjustment of short reads to improve genome assemblies.)，将经过滤的 R1、R2 reads 拼接。并用 FastX (http://hannonlab.cshl.edu/fastx_toolkit/index.html)处理拼接获得的 fastq 文件，得到 fa 格式序列。继而使用 blast+(Camacho C, (2009) "BLAST+: architecture and applications")将该 fa 中所有 reads 和目的区域参考序列进行比对。筛选能够覆盖其目标序列的 90%，或有 90%碱基能够完整覆盖其目标序列的 reads，记为有效 reads，并对其进行统计，结果见表 4-、表 4-。

表 4-4 目标区域 reads 过滤统计

Sample	Clean Reads	Raw Reads	Clean Reads Ratio
10	63224	70594	0.9
11	84292	92439	0.91
12	73564	84609	0.87
13	71402	80667	0.89
14	76865	86708	0.89
15	61974	69596	0.89
16	67480	76334	0.88
17	73316	82974	0.88
18	82858	93099	0.89
19	89812	99608	0.9

****仅最多展示 10 份样品，其余样品结果见 Data_statistics.xlsx/QC_CleanData sheet。

表 4-5 各样品目标区域测序深度统计

Sample	平均测序深度	大于 10X 所占比例	大于 20X 所占比例	ADCY2_	AHRR_
10	1290.286	1	0.98	2362	101
11	1720.245	1	0.98	2782	160
12	1501.306	0.98	0.98	2684	133
13	1457.184	0.98	0.98	2260	121
14	1568.673	0.98	0.98	2894	132
15	1264.776	0.98	0.98	3512	91
16	1377.143	0.98	0.98	3724	104
17	1496.245	0.98	0.98	4156	157
18	1690.98	0.98	0.98	3382	152
19	1832.898	0.98	0.98	3085	172

**** 仅最多展示 10 份样品以及 2 个扩增子区域，其余样品结果见 Data_statistics.xlsx/TargetCoverage sheet。

4.2.4. 胞嘧啶转化效率统计

统计每个样本的有效测序数据中，经重亚硫酸盐处理后，碱基 C 转变为 T 的效率（表 4-）。

表 4-6 CT 转化效率统计

Sample	Transferred(C->T)	All Base(C)	Efficiency
10	2017047	2033015	99.21%
11	2733669	2755483	99.21%
12	2374594	2392964	99.23%
13	2350781	2369009	99.23%
14	2488473	2507761	99.23%
15	1987693	2003597	99.21%
16	2164285	2181167	99.23%
17	2382624	2401170	99.23%
18	2680924	2703027	99.18%
19	2901809	2925075	99.20%

备注：Transferred(C->T)：重亚硫酸盐处理后**非 CpG 位点**碱基 C 转变为 T 的数量；All Base(C)：参照序列中**非 CpG 位点**碱基 C 总数；Efficiency：CT 转化效率。

***仅最多展示 10 份样品，结果目录：methylation.xlsx/Efficiency sheet

4.3. CpG 位点甲基化分析

该部分分析内容给出：

- ✓ CpG 位点的甲基化水平
- ✓ 扩增子 CpG 位点甲基化单倍型类型及丰度

4.3.1. CpG 位点甲基化水平计算

计算扩增子中，CpG 位点甲基化水平。结果列于表 4-。

表 4-7 CpG 位点甲基化水平

Target	Position	Chr	GenomePosition	Distance2TSS	Type	10	11
ADCY2_	29	5	7395309	-1033	CG	0.378653	0.562838
ADCY2_	38	5	7395318	-1024	CG	0.441949	0.616853
ADCY2_	43	5	7395323	-1019	CG	0.404661	0.597407
ADCY2_	47	5	7395327	-1015	CG	0.315544	0.525558
ADCY2_	56	5	7395336	-1006	CG	0.277174	0.543439
ADCY2_	58	5	7395338	-1004	CG	0.186377	0.504823
ADCY2_	64	5	7395344	-998	CG	0.326096	0.524116
ADCY2_	69	5	7395349	-993	CG	0.36394	0.56183
ADCY2_	94	5	7395374	-968	CG	0.448664	0.613929
ADCY2_	115	5	7395395	-947	CG/rs35838516	0.424457	0.594652

注：Target：扩增子名称；Position:甲基化位点在所处扩增子上的位置；Chr: 片段所在的染色体；GenomePosition:该位点在参照基因组上的位置；Distance2TSS:该位点在参照基因组上相对转录起始位点的距离,负号表示该位点在转录起始位点的上游;Type：甲基化类型（CG 表示 CG 位点的甲基化），如果检测到 1000g 数据库在该位点存在高频突变（大于 0.01），则添加/SNP 标志；甲基化程度 = 该位点甲基化的 reads 数目（即检测到碱基 C 的 reads 数目）/ 该位点总的 reads 数目。

***仅最多展示 10 个位点以及 2 个样本，结果目录：methylation.xlsx/Methyl Table sheet

4.3.2. CpG 位点甲基化单倍型分析

以扩增子为单位，分析 CpG 位点的单倍型。结果列于表 4-。

表 4-8 单倍型类型统计

Target	Haplotype	Depth	10	11
ADCY2_	ttttttttttttttt	23647	0.22458	0.12528
ADCY2_	ccccccccccccccc	5806	0.034925	0.116331
ADCY2_	cccccccccccccct	2895	0.002653	0.026473
ADCY2_	ttttttttctttttt	1784	0.013705	0.006711
ADCY2_	tttttttctttttt	1538	0.011494	0.010067
ADCY2_	ttctttttttttttt	1208	0.012821	0.006339
ADCY2_	tttttttttttctttt	989	0.009726	0.003356
ADCY2_	tctttttttttttt	945	0.0084	0.006339
ADCY2_	ccccccccctccccc	875	0.005747	0.043251
ADCY2_	cccccccccctcccct	707	0.000884	0.014914

注：Target：扩增子名称。Haplotype：单倍型类型，假设扩增子序列为：ATCATXGATCXGCTAXGCTTTAXGCCTAT，X 可为 c（甲基化修饰）或 t（未甲基化修饰）。其中一条测序的 read 为：ATCATCGATCTGCTACGCTTTATGCCTAT，则该 read 对应的扩增子甲基化单倍型为：CTCT。

Depth：支持此单倍型的测序数量（所有样本该分型的 reads 总和）。

***仅最多展示 10 种单倍型以及两个样本，结果目录：methylation.xlsx/Haplotype sheet

4.4. 甲基化水平相似性分析

该部分分析内容给出：

- ✓ 基于每个 CpG 位点的甲基化水平，各样本的 PCA 分析
- ✓ 基于每个 CpG 位点的甲基化水平，各样本的聚类分析

基于扩增子中 CpG 位点的甲基化水平，对所有样本进行 PCA 分析（图 4-3）与聚类分析（图 4-4）。

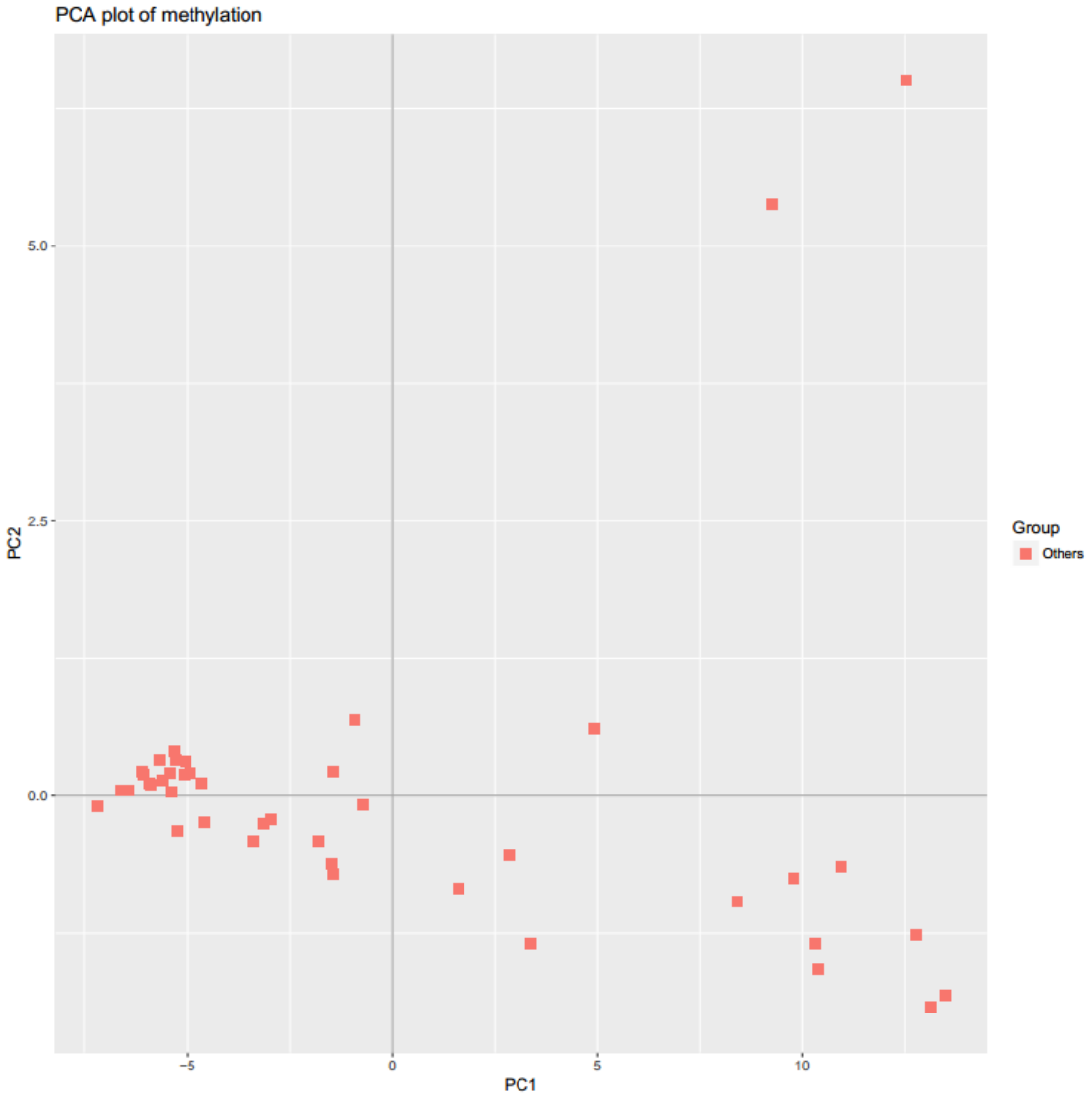

图 4-3 PCA 图颜色区分样本组；X、Y 轴分别对应最能反映样本真实物种组成的指标（由软件计算给出）

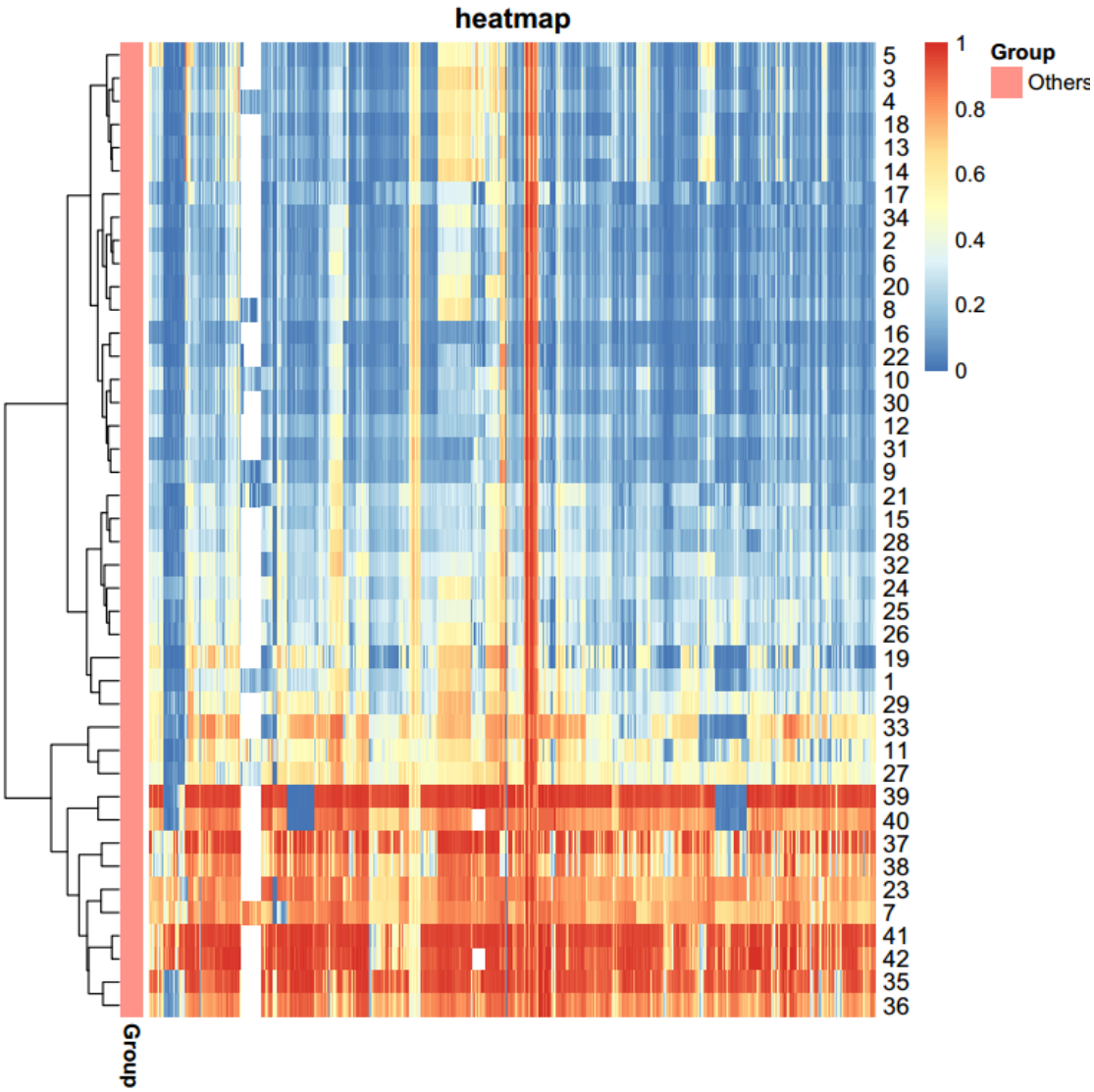

图 4-4 Heatmap 每个单元格表示对应行样品的 CpG 位点的相对甲基化水平，并以颜色梯度反映甲基化水平变化，趋向于蓝色则甲基化水平越低，反之趋向于红色则甲基化水平越高。样本甲基化水平的相似性由行的排布顺序展示，越相邻的行表示其代表的样本甲基化水平整体上相似度越高，图左侧的树状图则系统地描述这种相似程度。

4.5. SNV 分析

使用 BWA 进行比对后，再利用 GATK 进行 SNV calling，输出 SNV calling 结果（表 4-12）。

表 4-2 SNV 分析

Target	Position	Original Ref	Ref	Alt	Warning	Genome Chr	Genome Position	Genome Strand	Geno(0|1|2)	Alt Allele Freq	CallRate	HWE	1
AHRR_	184-184	T	T	-		5	374384-374384	+	24|18|0	0.214286	1	0.164029	T/T
ARRB2_	55	G	G	A		17	4615089	+	0|0|42	1	1	1	A/A
ARRB2_	64	A	A	G		17	4615098	+	2|18|22	0.738095	1	0.697189	G/G

***仅最多展示 3 个扩增子的 SNV 分析，结果目录： SNV_bwa.xlsx

表头注释

标识	注释
Target	目标片段名称
Position	突变在参考序列上的位置
Original Ref	参考等位基因
Ref	参考等位基因（C 转化为 T，因为 DNA 经过重亚硫酸盐处理）
Alt	突变等位基因
Warning	突变警告信息
GenomeChr	突变在参考基因组的染色体号（如果没有参考基因组，设为空）
GenomePosition	突变在参考基因组对应染色体上的位置（如果没有参考基因组，设为空）
GenomeStrand	突变在参考基因组上的方向
Geno(0|1|2)	突变类型数量统计，0 为野生型，1 为杂合突变，2 为纯合突变
Alt Allele Freq	突变等位基因频率
CallRate	召回率
HWE	HWE 检验

5. FAQ

5.1. *.fastq 文件如何打开

高通量测序（如 Illumina HiSeq/MiSeq 等测序平台）得到的原始图像数据文件经碱基识别（Base Calling）分析转化为原始测序序列（Sequencing Reads），我们称之为 Raw Data 或 Raw Reads，结果以 FASTQ（简称为 fq）文件格式存储。原始测序数据文件一般较大，建议使用性能较好的计算机或者使用更适合处理大量数据的 Unix/Linux 系统打开。

5.2. FASTQ 文件格式特点

FASTQ 文件，包含了测序序列的序列信息及其对应的测序质量信息。测序样品中真实数据随机截取结果如下图所示：

```
@ST-E00211:153:HL7L2CCXX:8:1101:26362:2153 1:N:0:ATGTCAGA
CTCCTTTTCTTTGTCTATGTTGTACATTTTCATGATATAACTTTTAACTATGTCT
AGAGAAGGCAGGCTCTGCAAGAGAGGTGCCCTTTCAACCCGCTCAGTGCCC
+
A<FFFKFKAKKKKKKKKKKKKKKKKKFKKFKKKKKKKKKKKKKKKKKKK
KKKKKKKKKKKKKFKKFF7KFFKKKKKKFKK<FFKKKKKKKKKKKKK
```

FASTQ 格式文件中每条 read 由四行描述，其中第一行以“@”开头，随后为 Illumina 测序标识符（Sequence Identifiers）和描述文字（选择性部分）；第二行是碱基序列；第三行以“+”开头，随后为 Illumina 测序标识符（选择性部分）；第四行是对应序列的测序质量。

Illumina 测序标识符来源：Illumina 测序仪一个 run 包含 2 个 flowcell；一个 flowcell 中包含 8 个 lane；每个 lane 包含 2 列；每一列又包含 60 个 tile，每一个 tile 又会种下不同的 cluster，其产生的测序文件识别标志（Sequence Identifiers）中的详细信息如下表所示：

表 5-1 测序文件识别标志（Sequence Identifiers）详细信息对照表

@ST-E00211	Unique instrument name
153	Run ID
HL7L2CCXX	Flowcell ID
8	Flowcell lane
1101	Tile number within the flowcell lane
26362	'x'-coordinate of the cluster within the tile
2153	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read fails filter (read is bad), N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
ATGTCAGA	Index sequence

5.3. FASTQ 文件碱基质量表示

Read 的质量分数以不同的大写英文字符来表示，其中每个字符对应的 ASCII 值减去 33，即为对应的测序质量值。一般地，碱基质量从 0~40，即对应的 ASCII 码为从“！”（0+33）到“I”(40+33）。如果测序错误率用 E 表示，Illumina HiSeq/MiSeq 的碱基质量值用 Q 表示，则有下列关系：

$$\text{公式 1：} Q = -10\log_{10}(E)$$

测序错误率与测序质量值简明对应关系如下表：

表 5-2 测序错误率与测序质量值简明对应表

测序错误率 (E)	测序质量值 (Q)	对应 ASCII 码
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

测序 reads 错误率会随着测序的进行而升高，是由于测序过程中化学试剂的消耗造成，这是 Illumina 高通量测序平台的通病。

6. 附录 I

6.1. 结果文件目录列表

项目号-分析结果

| Data_statistics.xlsx

| methylation.xlsx

| methylation_info.xlsx

| SNV_bwa.xlsx

|

|--fastqc

| *_ErrorRate.png

| *_R1_per_base_quality.png

| *_R1_per_base_sequence_content.png

| *_R2_per_base_quality.png

| *_R2_per_base_sequence_content.png

|

|--相似性分析

| heatmap.pdf

| pca.pdf

| 样本分组说明.txt

6.2. 软件信息列表

Software	URL
Fastqc-v0.11.5	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FLASH-v1.2.11	http://www.cbcb.umd.edu/software/flash
blat-v.36	http://genome.ucsc.edu/FAQ/FAQblat.html
Bwa- v0.7.15	http://bio-bwa.sourceforge.net/
GATK-v3.5	https://software.broadinstitute.org/gatk/

7. 附录 II

7.1. 上海天昊生物科技有限公司简介

上海天昊生物科技有限公司，2008 年 4 月创建于上海浦东张江高科技园区。天昊生物目前拥有员工近 100 名，2500 平方的办公与实验场地，建立了整套完备的实验室管理体系和标准流程，主要目标是构建一个具备国际一流水准的标准化多层次分子生物学研究分析平台，为国内外基因生物领域科研机构、医学院校以及生物制药企业提供精确、高效的基因检测分析服务。

2013 年我们在北京和南京分别成立的两个办事处，力求为全国客户提供更全面、更优质的服务。

2014 年天昊生物被评为“国家高新技术企业”，公司总经理姜正文博士获得：国家“千人计划”，江苏省“双创人才”、苏州工业园区“领军人才”等荣誉。

2015 年 7 月天昊基因与谱润、山蓝和盛宇 3 家知名基金投资公司，完成了人民币 8000 万元的首轮融资。

自天昊生物成立以来，我们已经建立并完善了包括 ABI 3130xl、ABI 3730xl 测序仪、实时定量 PCR 仪、Illumina GAIIx、Illumina MiSeq、Illumina NextSeq 二代测序平台和细胞生物学分析平台在内的多个分子生物学分析研究平台。不仅先后开展了疾病基因定位、基因突变及多态性测序分析、各种通量 SNV 基因分型分析、基因转录水平分析、甲基化水平分析等服务项目，同时还利用目前世界先进的二代测序技术开展了全基因组测序、全外显子组测序、目的区域富集测序、RNA 测序、micro-RNA 测序、甲基化测序等成熟的科研分析服务。此外，上海天昊在进行高质量技术服务的同时还积极集中优势力量自主创新，独立研发了 AccuCopy®多重拷贝数检测技术、CNVplex®高通量拷贝数检测技术、SNVscan®高通量 SNV 分型方法等具有国际水平的专利技术，另外利用二代测序技术不断开发出大量新的基因分析技术，例如多种策略方法的目的区域富集大样品测序技术 EasyTargetTM、超高通量 SNV 分型技术 SNVseq®、超高通量目的区域 CNV 检测技术 CNVseq®等。截至 2015 年 10 月底，公司申请了 4 项 PCT 专利和 24 项国内发明专利，其中 3 项专利已经授权，另 3 项正在实审阶段，另外获得 5 项软件著作权，申请了 40 商标著作权，已获 14 商标著作权。目前公司可以提供的分子生物学或基因组学相关科研服务项目超过 70 类，迄今已经为国内外近 550 多家科研院校、医疗单位和生物公司提供了超过 6000 项科研技术服务。

在为广大新老客户提供科研技术实验服务的同时，上海天昊集服务与产品于一身，利用自主研发的 AccuCopy®、CNVplex®、SNVscan®、EasyTarget®、SNVseq®和 CNVseq®等全新技术为客户提供灵活定制科研项目检测试剂盒服务，为具备检测实力的大中型科研单位提供更为先进高效的检测方法和试剂。

上海天昊遗传分析中心借助多名长期从事基因及遗传分析的领域专家构成的专家咨询团队，结合准确、高效、经济的科研技术服务体系，致力于长期为分子生物学及医学遗传学领域的研究者提供高质量的科研策略咨询、实验技术服务和遗传数据分析，帮助广大科研人员获得更为优质

的科研成果。

上海天昊生物科技有限公司

地址：上海市浦东新区康桥路 787 号 9 号楼

服务热线：400-065-6886

电话：021-50802060/021-50807380；传真：021-50802059

网址：

7.2. 权责声明

第 1 条

本报告及其附件中披露的由上海天昊遗传分析中心负责设计优化的检测体系（合同中对权利归属已做说明的除外）的知识产权归上海天昊生物科技有限公司（以下简称公司）所有，任何其它单位及个人未经公司许可不得对其进行专利申请或用于商业用途。上述检测体系包括设计合成的各种核酸序列、检测反应组成、反应条件及检测流程等。

第 2 条

本报告及其附件中披露的项目内容、实验数据及结果，公司必须严格保密，未经客户单位负责人允许，严谨公司任何人对外宣传。

第 3 条

为了保证项目成果的安全，本报告及其附件中披露的实验数据及结果已通过邮件发送，而本报告及其附件中披露的实验及数据分析相关信息、原始实验数据则需客户单位的项目负责人或其联系人通过公司授权的帐号在公司项目管理系统中下载。

第 4 条

本报告及其附件内容仅限于项目团队成员、客户单位的项目负责人及其联系人查阅或拷贝，公司内任何其他人员未经项目总监允许不得查阅或拷贝，公司外任何其他人员未经客户单位的项目负责人许可不可查阅或拷贝。

7.3. 论文引用或致谢

如果您认可我们的工作，在发表文章时愿意将我们放到致谢名单中，我们将非常感谢，文章中天昊的中英文名称如下：

上海天昊生物科技有限公司（中文）

Genesky Biotechnologies Inc., Shanghai, 201315 (English)