联川生物细菌重测序分析总结报告

客　户：Demo

客户单位：Demo

项目编码：Demo

报告出具日期：Demo

项目经理：Demo　　Demo　　Demo

技术支持：　　Demo　0571-87662413　support@lc-bio.com

您值得信赖的基因组学与蛋白质组学技术合作伙伴

1 细菌基因组测序简介

从 1995 年第一个细菌基因组流感嗜血杆菌（*Haemophilus influenzae*）发表开始，细菌基因组研究已经历了 20 多年。基因组测序技术也在不断革新，从第一代 sanger 测序，发展到第二代高通量测序，再到第三代单分子测序，被测序的微生物基因组越来越多，人们可以研究微生物的功能、进化、自身之间及与环境间的相互作用。测序技术让人们对微生物多样性的了解达到了前所未有的高度，帮助人们研究疾病的传播，开发药物和疫苗。

一般来说，细菌基因组研究可以分为重测序和 *de novo* 两大类，两者的侧重点不同，但也可以相互结合。

细菌基因组重测序是指对基因组序列已知的细菌个体进行基因组测序，通过与已知的参考基因组比对，获得该细菌个体或群体的差异的测序方法。这些差异主要包括：SNP (单核苷酸多态性位点, Single Nucleotide Polymorphism) 、InDel (插入缺失位点, Insertion & Deletion) 、SV (结构变异位点, Structural Variation) 。目前微生物基因组重测序被广泛应用于病原微生物的检测及鉴定、病原菌演变及起源、致病菌种群结构及种群结构的进化等众多方面。

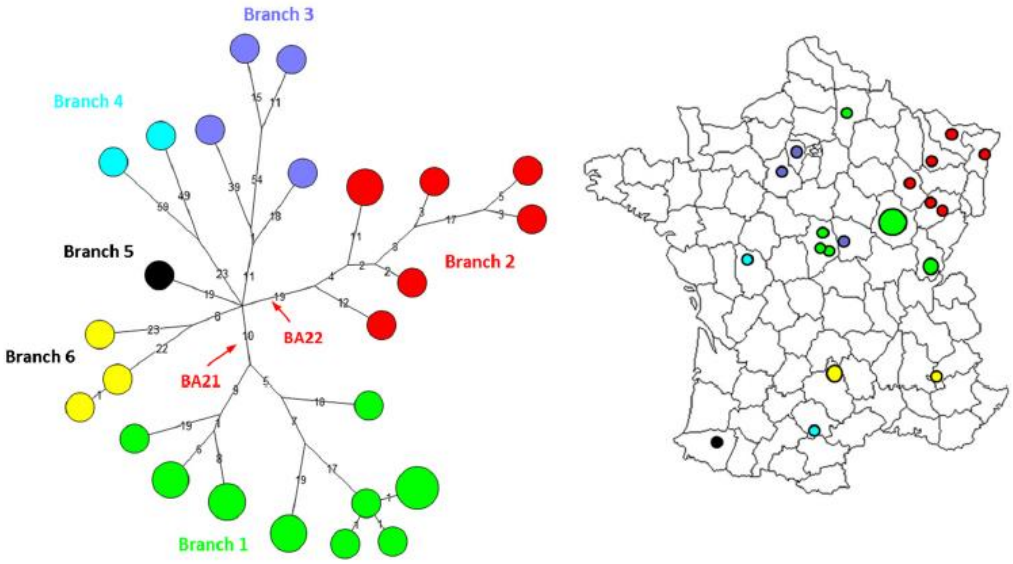
技术优势

1、与 SNP 芯片技术相比，基因组重测序能够发现未知的遗传变异信息。

2、与传统测序技术相比，基因组重测序数据通量更高、速度更快、成本更低。

3、基因组重测序可以检测 SNP、InDel、SV 等多种遗传变异类型。

细菌基因组 *de novo* 测序则不依赖参考基因组，直接利用测序数据对细菌基因组进行从头组装。基于组装结果，我们可以预测细菌基因组中所包含的基因，并通过功能数据库比对获得基因的功能信息。



2 项目信息

2.1 样品信息

物种名：鲍曼不动杆菌

物种拉丁名：*Acinetobacter baumannii*

2.2 参考基因组信息

链接参考基因组序列

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/016/805/GCF_000016805.1_ASM1680v1/GCF_000016805.1_ASM1680v1_genomic.fna.gz

参考基因组注释

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/016/805/GCF_000016805.1_ASM1680v1/GCF_000016805.1_ASM1680v1_genomic.gff.gz

2.3 数据分析程序

分析项目	名称	版本	说明
数据质控	cutadapt[1]	1.9	去除序列接头
	fqtrim[2]	0.94	去除低质量碱基
	FastQC[3]	0.10.1	数据质控统计
基因组比对	BWA[4]	0.7.10	测序 reads 参考基因组进行比对
	SAMtools[5]	0.1.19	比对文件排序
	Picard[6]	1.119	比对文件标记 PCR 重复及计算深度
SNP/InDel 检测	FreeBayes[7]	v1.0.2-29-g41c1313	根据 reads 比对结果检测 SNP 及 InDel
SV 检测	lumpy-sv[8]	0.2.11	根据 reads 比对结果检测 SV
进化分析	RAxML[9]	8.2.4	构建进化树

3 技术方法与流程

3.1 实验流程

联川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 HiSeq 4000 进行高通量测序，测序模式为 PE 150。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。

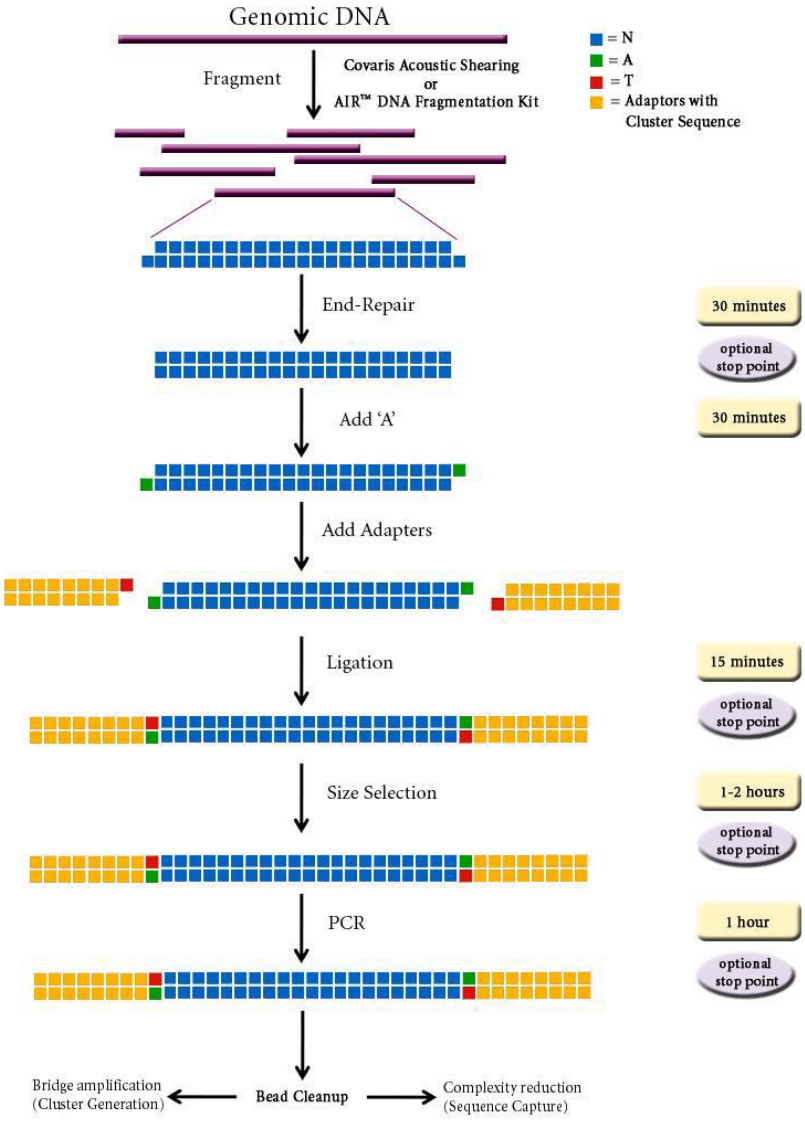


图 1：Paired end 文库构建流程

3.2 信息分析流程

在获得原始测序数据后，我们对数据进行生物信息学分析。首先，对原始数据进行数据预处理，去除测序接头和低质量的碱基，得到有效数据，并对数据量和数据质量进行统计。接着，将有效数据(reads)比对至参考基因组，标记重复序列，并进行比对率和覆盖度的统计。然后，根据比对结果进行变异检测，

包括 SNP、InDel、SV，并对变异的具体类型和区域进行注释统计。最后，根据得到的 SNP 信息进一步构建物种进化树。

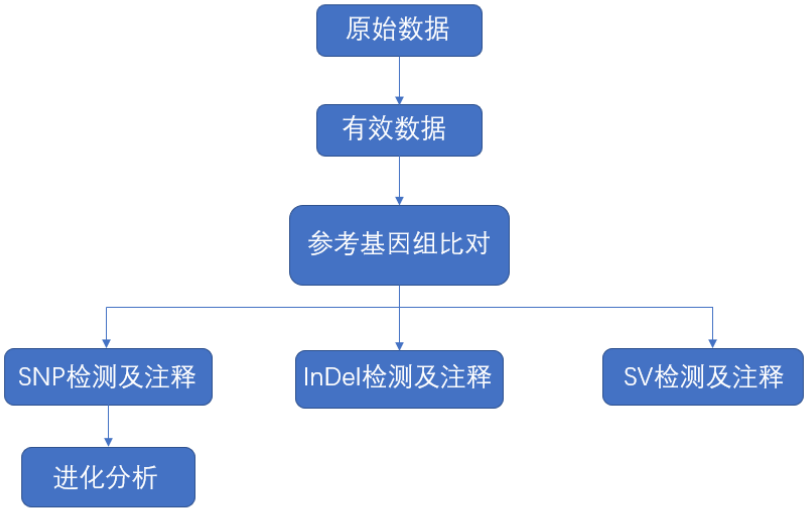


图 2：细菌重测序信息分析流程

4 样品信息表

#SampleID	Group
A	A
B	B
C	C

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

5 原始数据预处理

通过高通量测序仪获得的 paired-end 原始数据，其中可能含有带有接头的(建库过程引入）和低质量的测序数据(测序读取产生)。为了确保准确、可信

的分析结果，需要对于原始数据进行预处理，得到有效数据(Valid Data) ，以用于后续的信息分析。

数据预处理的步骤如下：

(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 100bp 的序列

(4) 去除截短后 N 的含量在 5%以上的序列

各样品经优化处理的测序数据统计部分结果见下表。

表 1: 测序数据质量预处理结果一览表

Sample	Raw Data		Valid Data		Valid%	Q20%	Q30%	GC%
	Read	Base	Read	Base				
A	14438500	2.18G	14284010	2.13G	98.93	99.06	97.25	32.76
B	12385000	1.86G	12095720	1.78G	97.66	99.16	97.24	32.86
C	13440000	2.02G	13170678	1.96G	98.00	99.18	97.39	33.54

结果文件各列意义如下:

Term	意义
Sample	测序文库名称
Raw Data/Read	测序原始数据，以四行为一个单位，统计每个文件的测序序列个数
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

结果路径：summary/2_CleanData/Demo_data_stat_paired-end.xlsx

6 参考基因组比对

将各样品的测序 reads 用 BWA 比对到参考基因组，对比对文件进行排序、标记 PCR 重复序列，最后进行比对结果的统计。

Term	A	B	C
TOTAL_READS	14284010(100.00%)	12095720(100.00%)	13170678(100.00%)
MAPPED_READS	13152311(92.07%)	11307992(93.48%)	10968291(83.28%)
TARGET_TERRITORY	2937129	2937129	2937129
MEAN_TARGET_COVERAGE	541.61	373.37	351.34
PCT_TARGET_BASES_30X	89.24%	89.14%	89.14%
PCT_TARGET_BASES_20X	89.28%	89.19%	89.22%
PCT_TARGET_BASES_10X	89.33%	89.27%	89.31%
PCT_TARGET_BASES_2X	89.48%	89.44%	89.48%
PCT_TARGET_BASES_1X	89.53%	89.49%	89.52%

结果文件各列意义如下:

Term	**注释**
TOTAL_READS	参与比对的总 reads 数目(比例)
DUPLICATE_READS	重复的 reads 数目(比例)
MAPPED_READS	比对到参考基因组上的总 reads 数目(比例)
TARGET_TERRITORY	目标区域的碱基个数
MEAN_TARGET_COVERAGE	目标区域的平均覆盖深度

PCT_TARGET_BASES_30X	目标区域中，覆盖深度不低于 30X 的碱基所占的比例
PCT_TARGET_BASES_20X	目标区域中，覆盖深度不低于 20X 的碱基所占的比例
PCT_TARGET_BASES_10X	目标区域中，覆盖深度不低于 10X 的碱基所占的比例
PCT_TARGET_BASES_2X	目标区域中，覆盖深度不低于 2X 的碱基所占的比例
PCT_TARGET_BASES_1X	目标区域中，覆盖深度不低于 1X 的碱基所占的比例

结果路径：summary/3_Align/map_stat.xlsx

7 SNP 检测及注释

SNP 全称 Single Nucleotide Polymorphisms,是指在基因组上单个核苷酸的变异，形成的遗传标记，其数量很多，多态性丰富。基因组上单个核苷酸的变异包括置换，缺失和插入。根据单核苷酸碱基形态的多样性，可以将置换分为转换(transition, CT,GA，同型碱基替换）和颠换(transversion, CA, GT, CG, AT，异型碱基替换)，转换发生率总是明显高于其它几种变异，具有转换型变异的 SNP 约为 2/3，其它几种变异的发生几率相似。一般而言，SNP 是指变异频率大于 1%的单核苷酸变异。

根据单核苷酸多态性在基因中的位置，可以分为基因非编码区 SNP，基因的间隔区 SNP(基因之间的区域)和基因编码区 SNP。从对生物的遗传性状的影响上来看，cSNP 又可分为 2 种：一种是同义 cSNP (synonymous cSNP)，即 SNP 所致的编码序列的改变并不影响其所翻译的蛋白质的氨基酸序列，突变碱基与未突变碱基的含义相同;另一种是非同义 cSNP(non-synonymous cSNP)，

指碱基序列的改变可使以其为蓝本翻译的蛋白质序列发生改变，从而影响了蛋白质的功能。这种改变常是导致生物性状改变的直接原因。

对检测到的 SNP，我们用 SnpEff 对其进行功能注释。

7.2 SNP 统计结果

表 2: SNP 所在位置分类统计

Sample	Total	DOWNSTREAM	EXON	INTERGENIC	INTRON	NONE	SPLICING	UPSTREAM	UTR_3_PRIME	UTR_5_PRIME
A	19249	684	15960	6	0	32	0	2567	0	0
B	18912	675	15646	6	0	32	0	2553	0	0
C	19482	680	16167	6	0	36	0	2593	0	0

结果文件各列意义如下:

Term	意义
Sample	样本名称
Total	SNP 总数
DOWNSTREAM	转录终止位点下游 500bp 区域
EXON	外显子区域
INTERGENIC	基因间区域
INTRON	内含子区域
NONE	NONE 类型区域
SPLICING	splicing junction 10bp 区域
UPSTREAM	转录起始位点上游 500bp 区域
UTR_3_PRIME	3’ UTR 区域
UTR_5_PRIME	5’ UTR 区域

注：当表格中样品量或列数过多时，只展示部分结果，完整结果请查看以下路径。

结果路径：summary/4_VarCall/reads/SNP_region_stat.xlsx

表 3: SNP 功能分类统计

Sample	Total	missense	start_gained	start_lost	stop_gained	stop_lost	synonymous	other
A	19249	5332	0	6	54	19	10546	3292
B	18912	5216	0	5	53	16	10354	3268
C	19482	5336	0	5	52	17	10755	3317

结果文件各列意义如下:

Term	注释
Sample	样本名称
Total	SNP 总数
missense	错义突变的 SNP 个数
start_gained	起始密码子获得的 SNP 个数
start_lost	起始密码子丢失的 SNP 个数
stop_gained	终止密码子获得的 SNP 个数
stop_lost	终止密码子丢失的 SNP 个数
synonymous	同义突变的 SNP 个数
other	其他类型的 SNP 个数

注：关于变异更细致的功能分类和影响程度可以参考附件 VCF 注释说明中的 Annotations and putative impacts 部分说明。

结果路径：summary/4_VarCall/reads/SNP_type_stat.xlsx

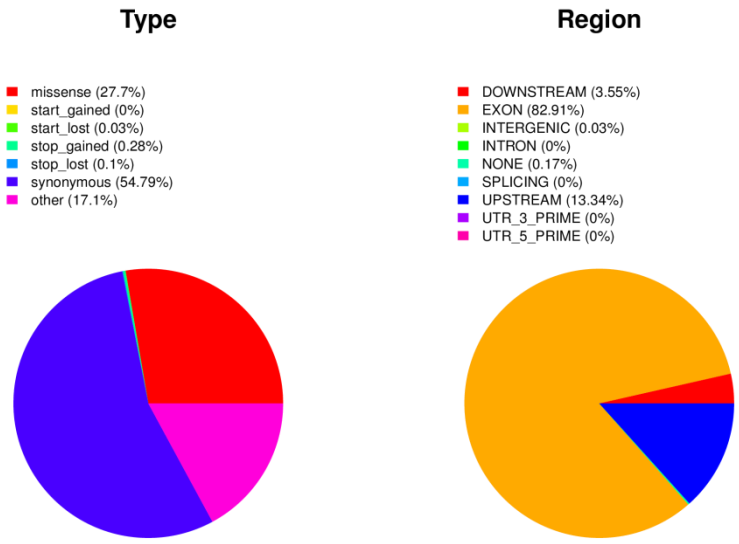


图 3: SNP 功能(左) 及位置(右) 分类统计

结果路径：summary/4_VarCall/reads/A/A.snp.pdf

7.3 SNP SnpEff 注释结果

使用 SnpEff[10]软件对 SNP 进行注释，注释结果采用 VCF4.1 的格式。

结果文件参数说明如下：

Term	Description	注释
CHROM	chromosome	染色体编号
POS	position	染色体位置
ID	identifier	变异 ID
REF	reference base(s)	参考序列碱基
ALT	alternate base(s)	变异碱基
QUAL	quality	变异碱基的质量值
FILTER	filter status	过滤状态

INFO	information	注释信息
DP	Approximate read depth	该位点的覆盖度
GQ	Genotype Quality	基因型的质量值
GT	Genotype	样品的基因型(genotype)
PL	Genotype Quality	指定的三种基因型的质量值

注：更详细的说明可以参看结果文件的注释行(以#号开头) 及附件 VCF 格式说明。

结果路径(以样品 A 为例)：summary/4_VarCall/reads/A/A.snp.ann.vcf

8 InDel 检测及注释

InDel (Insertion-Deletion) 是指相对于参考基因组，样本中发生的小片段的插入缺失，该插入缺失可能含一个或多个碱基。根据 InDel 在基因组中的位置，可以分为编码区序列的 InDel 和非编码区序列的 InDel。编码序列中的 InDel 发生与编码蛋白质的功能和氨基酸位点在结构和功能上的重要性有关。如果在 DNA 编码序列中插入或缺失一个或几个碱基(非 3 的倍数) ，这种突变称之为移码突变(frame shift mutation)。此类突变会造成插入点或缺失点下游的 DNA 编码框架全部改变，其结果是突变点以后的氨基酸序列都发生改变。而发生在非编码区(如：内含子区)的 InDel，将会降低转录的效率和剪切的准确性。

8.1 InDel 检测方法

我们采用 FreeBayes[7]进行对各样品比对结果进行个体 InDel 的检测。为了降低检测的假阳性，InDel 过滤标准如下：

(1) 位点最低测序深度(min-coverage)为 30

(2) 位点最小比对质量(min-mapping-quality)为 30

(3) 位点最低碱基质量(min-base-quality)为 20

(4) 变异最低支持深度(min-alternate-count)为 5

(5) 变异最低频率(min-alternate-fraction)为 0.05

8.2 InDel 统计结果

表 4: InDel 所在位置分类统计

Sample	Total	DOWNSTREAM	EXON	INTERGENIC	INTRON	NONE	SPLICING	UPSTREAM	UTR_3_PRIME	UTR_5_PRIME
A	628	137	88	0	0	68	0	335	0	0
B	617	136	81	0	0	71	0	329	0	0
C	629	134	87	0	0	75	0	333	0	0

结果文件各列意义如下:

Term	注释
Sample	样本名称
Total	InDel 总数
DOWNSTREAM	转录终止位点下游 500bp 区域
EXON	外显子区域
INTERGENIC	基因间区域
INTRON	内含子区域
NONE	NONE 类型区域

SPLICING	splicing junction 10bp 区域
UPSTREAM	转录起始位点上游 500bp 区域
UTR_3_PRIME	3’ UTR 区域
UTR_5_PRIME	5’ UTR 区域

注：当表格中样品量或列数过多时，只展示部分结果，完整结果请查看以下路径。

结果路径：summary/4_VarCall/reads/INDEL_region_stat.xlsx

表 5: InDel 功能分类统计

Sample	Total	exon_loss	frameshift	inframe_deletion	inframe_insertion	other
A	628	0	57	16	15	540
B	617	0	52	14	15	536
C	629	0	56	16	15	542

结果文件各列意义如下:

Term	注释
Sample	样本名称
Total	InDel 总数
exon loss	外显子删除的 InDel 个数
frameshift	发生移码的 InDel 个数
inframe_deletion	编码删除的 InDel 个数
inframe_insertion	编码插入的 InDel 个数
other	其他类型的 InDel 个数

注：关于变异更细致的功能分类和影响程度可以参考附件 VCF 注释说明中的 Annotations and putative impacts 部分说明。

结果路径：summary/4_VarCall/reads/INDEL_type_stat.xlsx

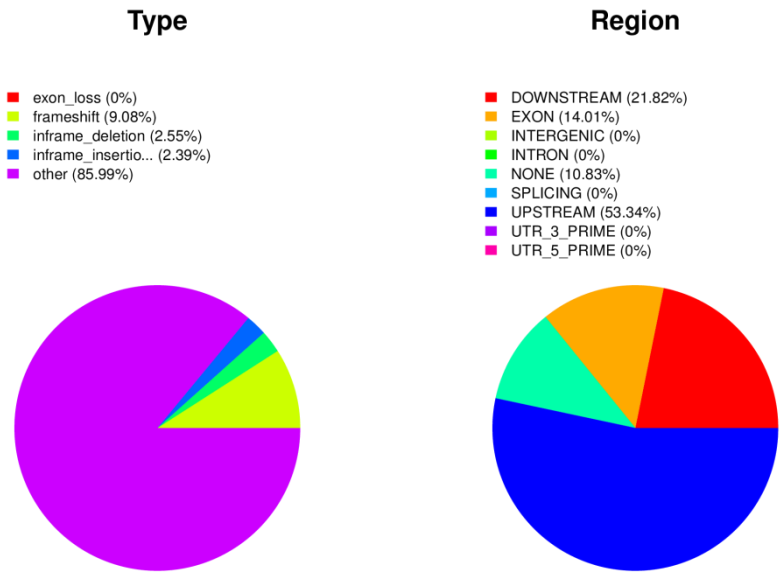


图 4: InDel 功能(左) 及位置(右) 分类统计

结果路径：summary/4_VarCall/reads/A/A.indel.pdf

8.3 InDel SnpEff 注释结果

使用 SnpEff[10]软件对 SNP 进行注释，注释结果采用 VCF4.1 的格式。结果文件参数说明如下：

Term	Description	注释
CHROM	chromosome	染色体编号
POS	position	染色体位置
ID	identifier	变异 ID
REF	reference base(s)	参考序列碱基

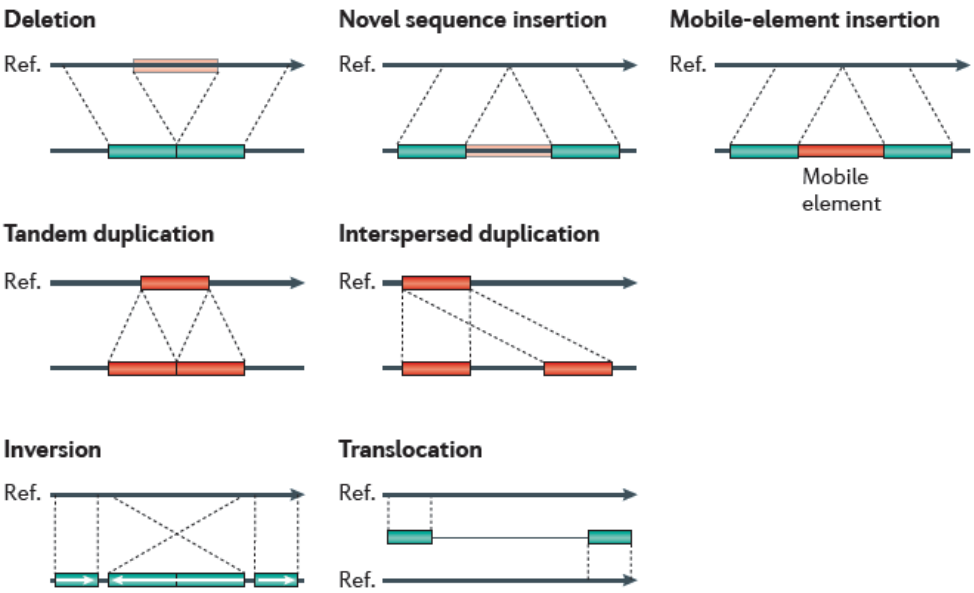
ALT	alternate base(s)	变异碱基
QUAL	quality	变异碱基的质量值
FILTER	filter status	过滤状态
INFO	information	注释信息
DP	Approximate read depth	该位点的覆盖度
GQ	Genotype Quality	基因型的质量值
GT	Genotype	样品的基因型(genotype)
PL	Genotype Quality	指定的三种基因型的质量值

注：更详细的说明可以参看结果文件的注释行(以#号开头) 及附件 VCF 格式说明。

结果路径（以样品 A 为例）： summary/4_VarCall/reads/A/A.indel.ann.vcf

9 SV 检测及注释

相对于 SNP 和 InDel 这种长度较短(<50bp) 的变异类型，SV (Structural Variation，结构变异)是指长度≥50bp 的长片段变异类型，主要包括插入(insertion) 、删除(deletion) 、染色体倒位(inversion) 、染色体内部或染色体之间的序列易位(translocation) (如图 5)。



目前主要有 4 种检测基因组上结构性变异的策略(如图 6)，分别为：(1) Read Pair (也称为 Pair-End Mapping，简称 PEM)；(2) Split Read (简称 SR)；(3) Read Depth (简称 RD)和(4) 基于 *de novo* 组装的方法。同时生物信息研究人员也已开发了众多根据以上 4 中策略中一种或者多种的软件用于结构性变异的检测。

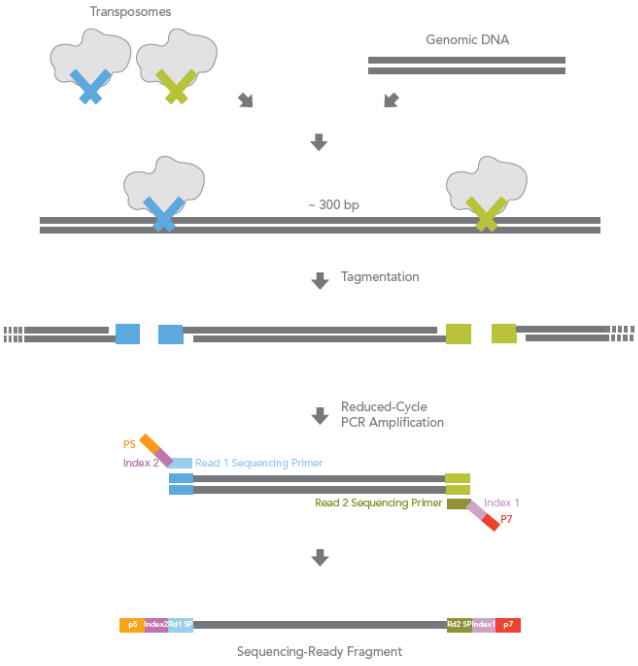


图 6: 结构变异检测策略

9.1 SV 检测方法

根据 reads 与参考基因组比对结果，我们采用软件 lumpy-sv[8]检测 SV，lumpy-sv 同时考虑 Read Pair 和 Split Read 的信息，可以将 SV 检测的精度提升到单碱基的水平。对于检测到 SV，我们将 lumpy-sv 标记为 IMPRECISE 的 SV 过滤掉，并保留 SV 质量值(QUAL) 大于 100 的 SV 结果。

9.2 SV 统计结果

表 6: SV 类型统计

Sample	Total	Deletion	Insertion	Inversion	Translocation
A	156	86	36	0	34
B	159	88	32	0	39
C	159	88	34	0	37

结果文件各列意义如下:

Term	注释
Sample	样本名称
Total	SV 总数
Deletion	缺失类型数目
Insertion	插入类型数目
Inversion	倒位类型数目
Translocation	易位类型数目

注：当表格中样品量或列数过多时，只展示部分结果，完整结果请查看以下路径。

结果路径：summary/4_VarCall/reads/SV_stat.xlsx

9.3 SV SnpEff 注释结果

相对于 SNP/InDel, SV 的结构更加复杂，SV 的注释也一直还在研究中，没有成熟的工具。一般来说，SV 可以用一系列的断点(breakpoints)来表征，此处，我们使用 SnpEff[10]软件对 SV 的断点进行注释，由于 lumpy-sv 检测到的 SV 结果为 VCF4.2 格式，我们在此基础上添加注释，注释结果也是 VCF4.2 的格式。

结果文件参数说明如下：

Term	Description	注释
CHROM	chromosome	染色体编号
POS	position	染色体位置
ID	identifier	变异 ID
REF	reference base(s)	参考序列碱基
ALT	alternate base(s)	变异碱基
QUAL	quality	变异碱基的质量值
FILTER	filter status	过滤状态
INFO	information	注释信息
SVTYPE	SV type	SV 细分类型
SVLEN	SV length	SV 长度
SU	supporting reads	SV 支持 reads 数目
PE	supporting paired-end reads	SV 支持 PE reads 数目

SR	supporting split reads	SV 支持 split reads 数目
ANNPOS	POS annotation	SV 起始位置注释
ANNEND	END annotation	SV 终止位置注释
GQ	Genotype Quality	基因型的质量值
GT	Genotype	样品的基因型(genotype)

注：更详细的说明可以参看结果文件的注释行(以#号开头）及附件 VCF 格式说明。

结果路径：summary/4_VarCall/reads/A/A.sv.ann.vcf

10 进化分析

在生物学中，进化分析是指根据遗传或者表型的差异研究推断不同物种或者个体间的进化关系的一门学科。进化分析的结果通常用进化树的形式展示，在进化树上每个叶子结点代表一个物种或个体，那么两个叶子结点之间的最短距离就表示相应的两个物种之间的差异程度。构建系统发育树的方法主要有：基于距离的算法(UPGMA 和 NJ 算法）、最大简约算法(MP）、最大似然值法(ML)。我们使用比较流行的软件 RAxML[9]构建系统进化树，RAxML 是用极大似然法建立进化树的软件之一，可以处理超大规模的序列数据，包括上千至上万个物种。

10.1 SNP 多序列比对结果

根据每个个体检测到的 SNP 位点，过滤掉彼此距离较近的 SNP 位点(<20bp），因为这些位点可能是由基因组重组产生，并不能真正反映物种间的

进化关系。将每个个体非重组的 SNP 位点碱基串联成一条序列，得到个体多序列比对的结果，据此进行后续的进化分析。

```
 10 234
Cow          MAYPMQLGFQ DATSPIMEEL LHFHDHTLMI VFLISSLVLY IISLMLTTKL
Carp         MAHPTQLGFK DAAMPVMEEL LHFHDHALMI VLLISTLVLY IITAMVSTKL
Chicken      MANHSQLGFQ DASSPIMEEL VEFHDHALMV ALAICSLVLY LLTLMLMEKL
Human        MAHAAQVGLQ DATSPIMEEL ITFHDHALMI IFLICFLVLY ALFLTLTTKL
Loach        MAHPTQLGFQ DAASPVMEEL LHFHDHALMI VFLISALVLY VIITTVSTKL
Mouse        MAYPFQLGLQ DATSPIMEEL MNFHDHTLMI VFLISSLVLY IISLMLTTKL
Rat          MAYPFQLGLQ DATSPIMEEL TNFHDHTLMI VFLISSLVLY IISLMLTTKL
Seal         MAYPLQMGLQ DATSPIMEEL LHFHDHTLMI VFLISSLVLY IISLMLTTKL
Whale        MAYPFQLGFQ DAASPIMEEL LHFHDHTLMI VFLISSLVLY IITLMLTTKL
Frog         MAHPSQLGFQ DAASPIMEEL LHFHDHTLMA VFLISTLVLY IITIMMTTKL

             THTSTMDAQE VETIWTILPA IILILIALPS LRILYMMDEI NNPSLTVKTM
             TNKYILDSQE IEIVWTILPA VILVLIALPS LRILYLMDEI NDPHLTIKAM
             S-SNTVDAQE VELIWTILPA IVLVLLALPS LQILYMMDEI DEPDLTLKAI
             TNTNISDAQE METVWTILPA IILVLIALPS LRILYMTDEV NDPSLTIKSI
             TNMYILDSQE IEIVWTVLPA LILILIALPS LRILYLMDEI NDPHLTIKAM
```

图 7：多序列比对结果示意图(phylip 格式)

注：第一行的两个数字分别代表序列数目和比对长度，第二行开始为每条序列的名称和比对序列。

结果路径：summary/5_Phylogeny/aligned_snps.phy

10.2 SNP 进化树构建

我们使用软件 RAxML[9]构建最大似然进化树，输入文件为 SNP 多序列比对结果，使用的模型为 GTRGAMMA，为了检验进化树分支的可信度，我们采用了 bootstrap 方法，设置 bootstrap 的次数为 100 次，从而得到每个分支的 bootstrap 值，该值越接近 100，表示分支的可信度越高。最终得到的进化树是 newick 格式的，类似于

((raccoon:19.19959,bear:6.80041)50:0.84600,((sealion:11.99700,seal:12.0

0300)100:7.52973,((monkey:100.85930,cat:47.14069)80:20.59201,weasel:18.87953)75:2.09460)50:3.87382,dog:25.46154)这种形式，其中被同一个括号起来的两单元是节点上分支，冒号后面数字表示分歧度(平均每位碱基替换次），括号后面的数字是分支可信度(即 bootstrap 值）。

结果路径：summary/5_Phylogeny/RAxML bipartitions.tre

10.3 SNP 进化树展示

用 FigTree[11]软件对构建好的进化树进行展示，结果如下：

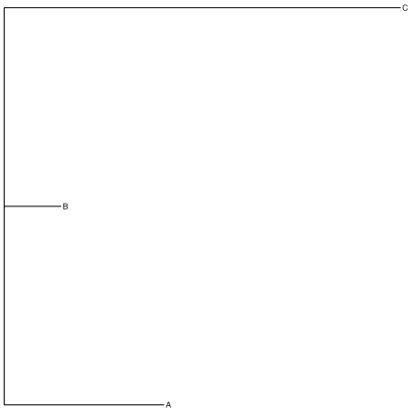


图 8: SNP 进化树

注：将 newick 格式的进化树文件导入 FigTree，可以根据需要调整进化树的拓扑结构，以及标注不同的字体、颜色、分歧值等。

结果路径：summary/5_Phylogeny/reads/RAxML_bipartitions.pdf

11 质量控制

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。

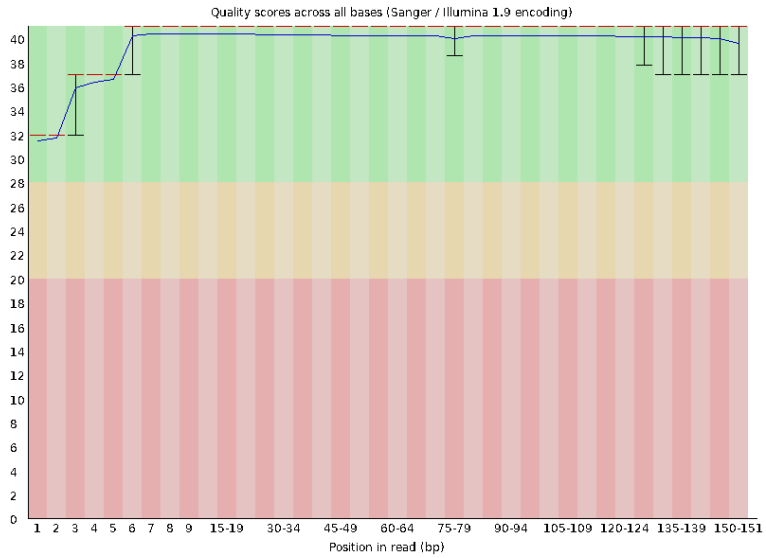


图 9：样品 A 序列测序质量分布图

结果路径：

summary/2_CleanData/paired-end/A/A_R1_fastqc/Images/per_base_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。

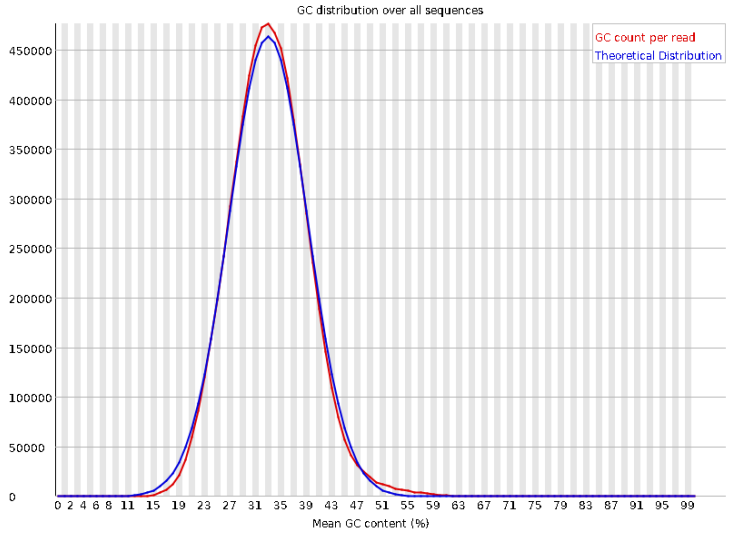


图 10: 样品 A 序列组成分布图

结果路径:

summary/2_CleanData/paired-end/A/A_R1_fastqc/Images/per_base_sequence_content.png

12 参考文献

[1] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[2] Geo Pertea. fqtrim: v0.9.4 release, July 2015.

[3] Ravi K. Patel and Mukesh Jain. Ngs qc toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7(2):e30619, 02 2012.

[4] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform.*Bioinformatics*, 25(14):1754–1760, 2009.

[5] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[6] Alec Wysoker, Kathleen Tibbetts, and Tim Fennell. Picard tools version 1.90, April 2013.

[7] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing.*arXiv preprint arXiv*:1207.3907, 2012.

[8] Ryan M. Layer, Colby Chiang, Aaron R. Quinlan, and Ira M. Hall. Lumpy: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):1–19, 2014.

[9] Alexandros Stamatakis. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 2014.

[10] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden.A program for annotating and predicting the effects of single nucleotide polymorphisms,snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*,6(2):80–92, 2012.

[11] Martha I Nelson, C´ecile Viboud, Amy L Vincent, Marie R Culhane, Susan E Detmer,David E Wentworth, Andrew Rambaut, Marc A Suchard, Edward C Holmes, and PhilippeLemey. Global migration of influenza a viruses in swine. *Nature communications*, 6, 2015.

13 辅助材料

为了每个客户方便查看数据以及完成一些个性化分析目的，我们提供了一些软件及其使用说明， 以及一些相关附件。

软件：

1、大容量数据浏览软件 VIM

测序下机数据量庞大，因此需要大容量数据浏览软件查看数据。VIM 是一款非常强大的编辑器，可以用于打开较大的文本格式的文件。

2、Notepad

对于客户来讲，我们之所以提供 Notepad++ ，是由于测序及分析的结果文件中，很多文件并不是常规的文本文档和 excel 文档，查看这些文档内的信息需要一款方便简洁并且能够编辑信息的编辑器。而 Notepad++ 是一款非常有特色的编辑器，足够支持信息查看、编辑。

3、谷歌 Chrome 浏览器

Chrome 浏览器用于打开结果文件中一些 html 格式的文件，由于 Chrome 支持 WebGL 是一种 3D 绘图标准，对于一些图形的展示，它可以得到更好的效果。

附件：

附件 1、FASTQ 格式详解

二代测序下机数据以 FASTQ 格式展现，该附件详细的介绍了 FASTQ 格式数据信息。

附件 2、VCF 格式说明

基于 reads 比对的变异检测结果以 VCF 格式展现，该附件详细的介绍了 VCF 格式数据信息。

附件 3、VCF 注释说明

使用 SNPEff 对 VCF 文件中变异位点进行注释，该附件详细的介绍了 SNPEff 注释的相关信息。

辅助材料路径： summary/6_SupplementalMaterial

14 联系我们

地址：浙江省杭州市下沙经济开发区 6 号大街 260 号

邮政编码：310018

联系电话：0571-87662413

传真：0571-81951905

邮箱地址：support@lc-bio.com

网址：www.lc-bio.com