

细菌基因组重测序变异检测报告

SRR2584866 · 大肠杆菌 REL606 长期进化实验克隆 REL11365 (公开数据示例)

项目 DEMO-DNA-01

数据来源 NCBI SRA SRR2584866

分析日期 2026-10-01

出具 HST GENOMICS 生物信息分
析

项目概况

本报告是公开数据示例交付物，用于展示 HST GENOMICS 微生物基因组重测序变异检测的交付形式。数据来自 NCBI SRA 公开数据 SRR2584866：大肠杆菌 B 株 REL606 长期进化实验（研究题为“E. coli genome evolution over 50,000 generations”）中的克隆 REL11365，Illumina HiSeq 2500 双端 150 bp 测序；以祖先株 REL606 的参考基因组（NCBI GCF_000017985.1）为参照检测变异。

4.95 M

过滤后读段

99.99%

比对率

143.9x

平均深度

789

过滤后变异

分析方法

- 读段质控：fastp 1.3.7，默认参数（去接头、低质量过滤）。
- 比对：BWA-MEM 0.7.19-r1273 比对到 REL606 参考基因组，samtools 1.24 排序与建索引。
- 变异检测：bcftools 1.24 mpileup + call（单倍体模式 --ploidy 1）。
- 过滤：QUAL $\geq$ 30 且位点深度 DP $\geq$ 10。
- 注释：snpEff 5.4c，用 NCBI 提供的 REL606 基因注释（GFF3）自建数据库。

质量控制

指标

数值

原始读段	5,536,796
过滤后读段	4,951,018
过滤后 Q30 比例	91.6%
比对率	99.99%
平均测序深度	143.9x
深度 $\geq 10x$ 的基因组比例	93.1%

表 1 测序与比对质控指标。

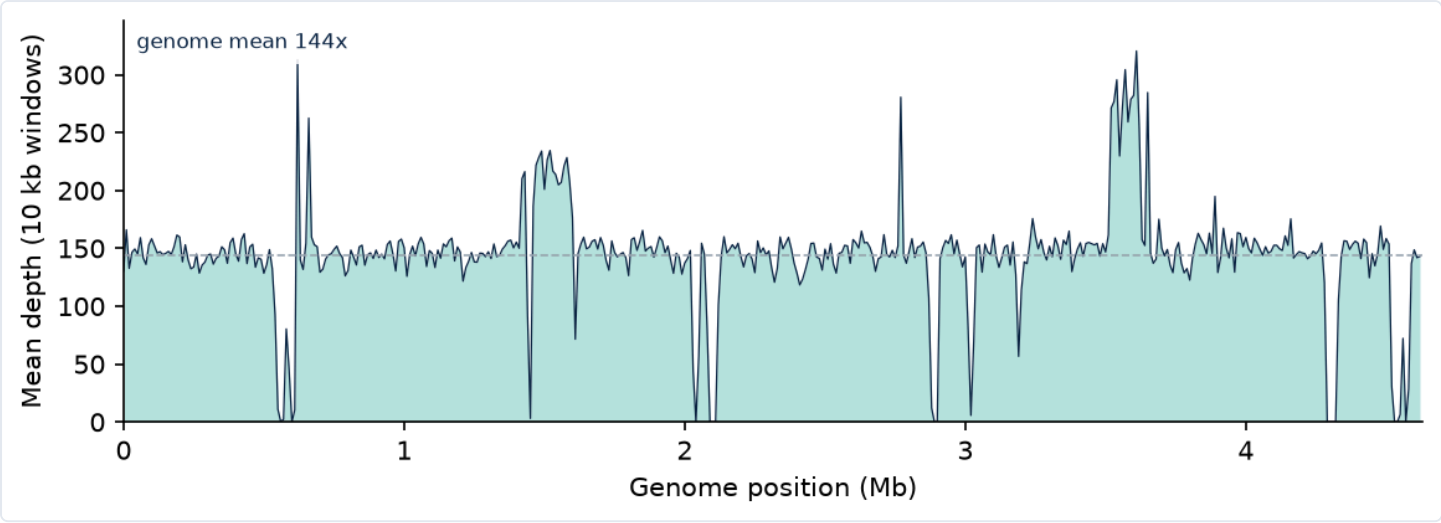


图 1 沿基因组的覆盖深度（10 kb 窗口平均）。约 6.9% 的参考基因组位置深度低于 10x；覆盖接近 0 的连续区段可能对应该克隆相对祖先株的缺失，深度约为平均值两倍的区段可能对应扩增，均需结合读段证据（如断点处的分裂比对）进一步确认。

结果

过滤后共 789 个变异：SNP 683 个、插入 65 个、缺失 41 个。按 snpEff 预测影响：HIGH 76 个、MODERATE 362 个、LOW 208 个、MODIFIER 143 个。

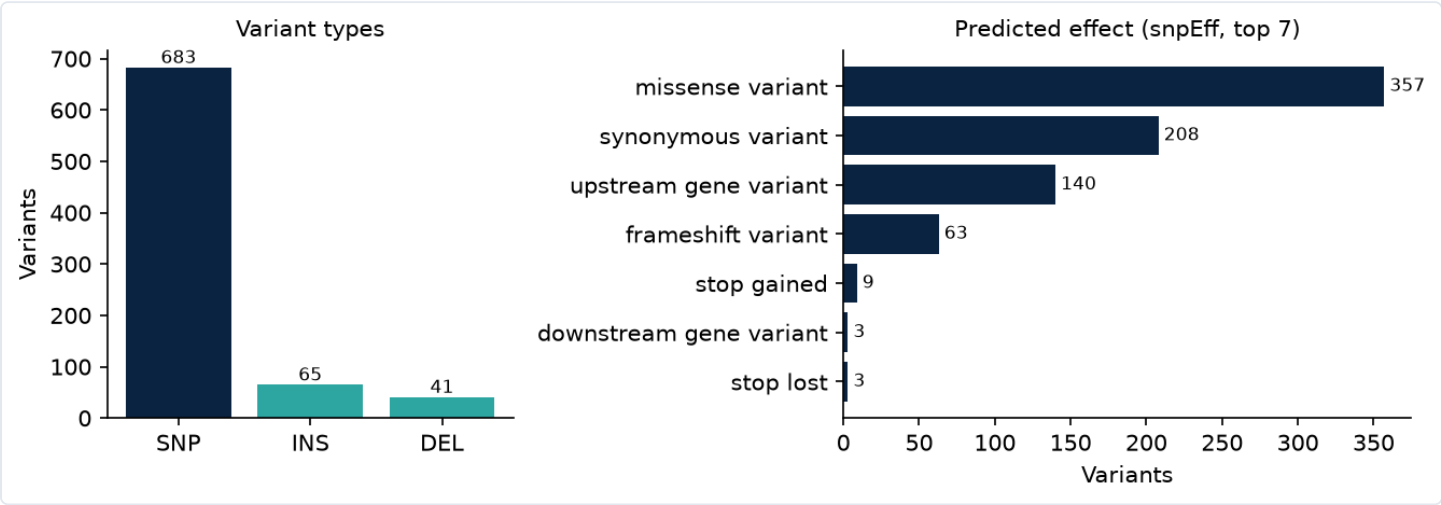


图 2 变异类型与预测功能效应。

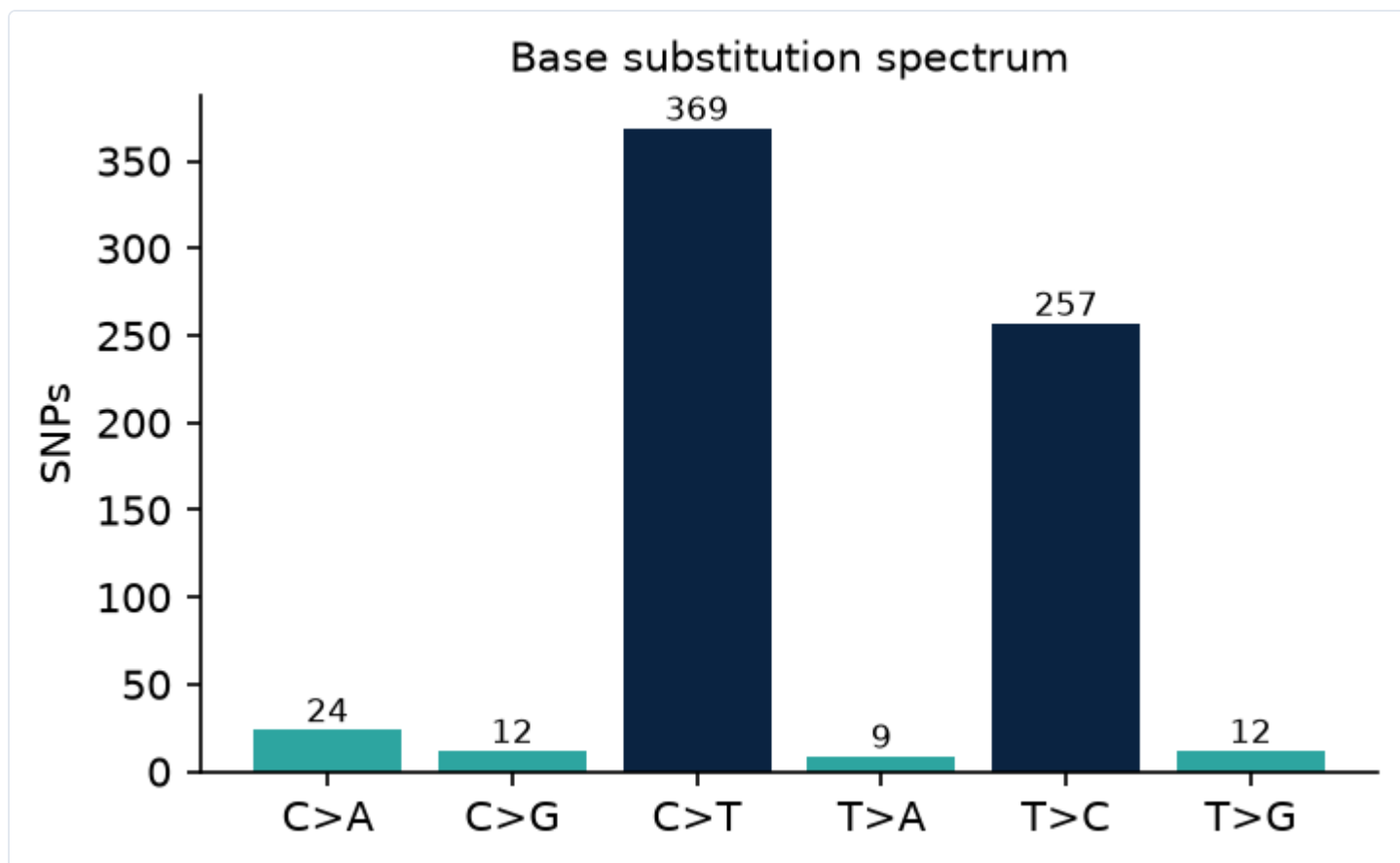


图 3 单碱基替换谱（按嘧啶参考碱基归并为 6 类）。转换 626 个、颠换 57 个。

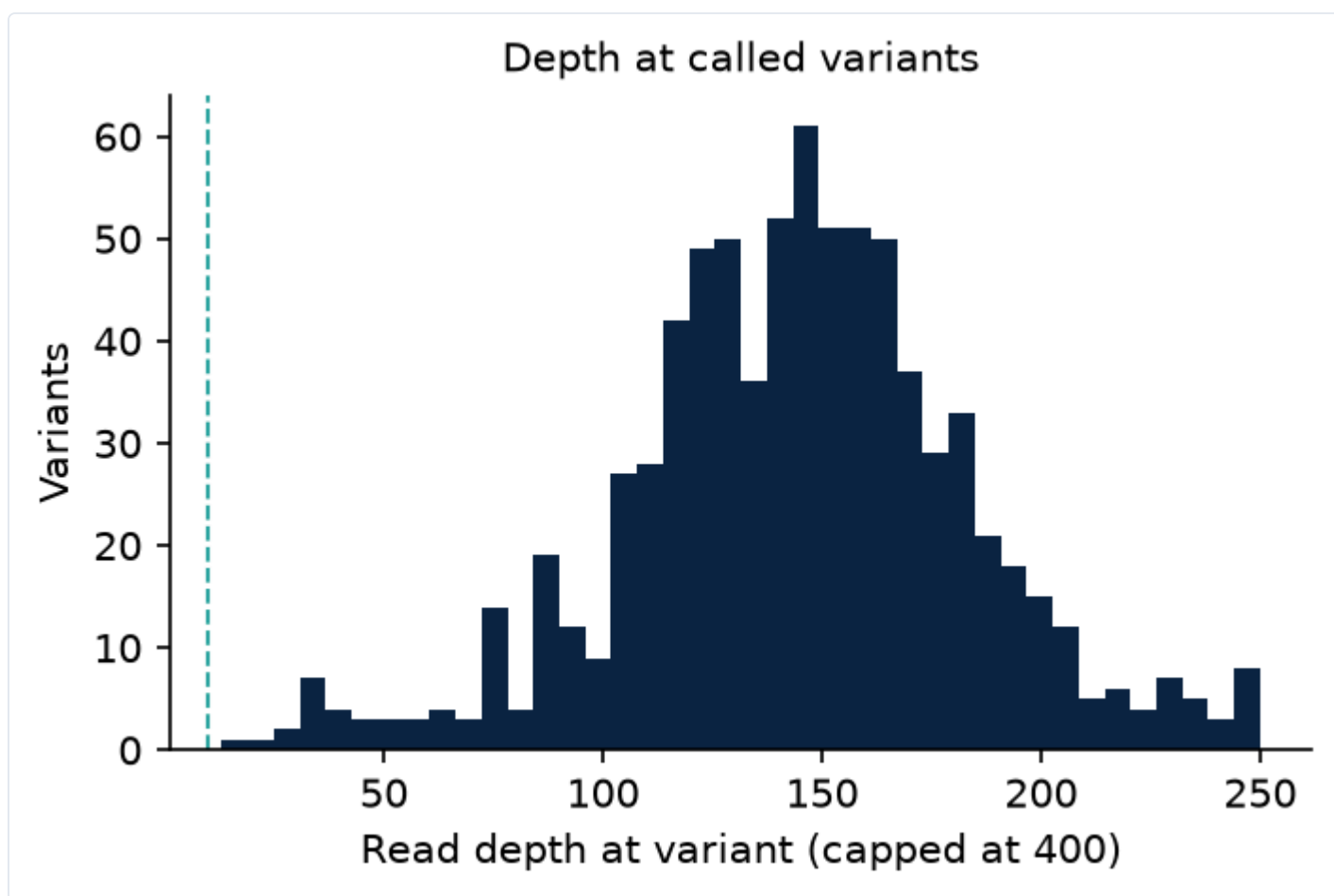


图 4 变异位点的读段深度分布，虚线为过滤阈值 10x。

位置	变化	基因	预测效应	HGVS	深度
16,518	INS	ECB_RS00080	frameshift variant	p.Arg362fs	229
105,581	G>A	ddlB	stop gained	p.Trp182*	138
148,134	INS	yadE	frameshift variant	p.Arg86fs	88
157,998	DEL	htrE	frameshift variant	p.Asn91fs	119
172,553	CAA>CA	fhuA	frameshift variant	p.Thr744fs	186
175,213	GAA>GA	fhuB	frameshift variant	p.Val305fs	144
233,686	INS	yafC	frameshift variant	p.Leu41fs	88
337,563	C>T	lacZ	stop gained	p.Trp433*	194
360,604	INS	ECB_RS01680	frameshift variant	p.Ala485fs	185
377,000	T>G	ECB_RS01790	stop lost	p.Ter154Cysext*?	106
429,704	INS	ECB_RS02045	frameshift variant	p.Arg362fs	217
451,177	INS	ylaC	frameshift variant	p.Ala48fs	136
473,901	INS	ybaL	frameshift variant	p.Ala469fs	141
495,902	INS	rhsD	frameshift variant	p.Ala161fs	73
619,368	T>G	ECB_RS02980	stop lost	p.Ter154Cysext*?	76
636,294	INS	dpiB	frameshift variant	p.Val493fs	157
708,824	DEL	kdpB	frameshift variant	p.Thr103fs	122
830,946	INS	ybil	frameshift variant	p.Ile46fs	155
901,822	INS	ECB_RS04470	frameshift variant	p.Ala460fs	114
917,183	INS	artJ	frameshift variant	p.Phe34fs	108
998,936	INS	ldtD	frameshift variant	p.Pro269fs	142
1,018,106	DEL	elfC	frameshift variant	p.Asn603fs	181
1,019,308	DEL	elfG	frameshift variant	p.Asp140fs	64
1,046,822	DEL	rlmI	frameshift variant	p.Ser80fs	102
1,062,285	INS	gfcE	frameshift variant	p.Leu204fs	143
1,407,377	C>T	dbpA	stop gained	p.Gln233*	157
1,424,975	INS	pinR	frameshift variant	p.Ile103fs	169
1,426,401	DEL	ECB_RS25730	frameshift variant	p.Leu20fs	175
1,439,576	INS	feaB	frameshift variant	p.Ala153fs	215

1,470,490	DEL	ydck	frameshift variant	p.Asn23fs	224
-----------	-----	------	--------------------	-----------	-----

表 2 预测影响为 HIGH 的变异（共 76 个，此处列前 30 个；完整表见 tables/variants_table.csv）。

解读提示：预测效应只说明变异落在基因的什么位置、可能怎样改变编码，不等于功能已被证实。真实项目中会结合研究问题（例如抗性、适应性、表型差异）筛选候选变异，并说明验证建议。

交付物清单

- tables/variants_table.csv：全部过滤后变异及注释（位置、类型、基因、效应、HGVS、深度）
- tables/variants_filtered.vcf.gz、tables/variants_annotated.vcf：标准 VCF
- tables/coverage_10kb_windows.tsv、depth_summary.tsv、flagstat.txt、fastp.json、bcftools_stats.txt：质控统计
- figures/*.png：出版级图（可按期刊要求改为 PDF/SVG）
- run_variant.sh、report_variant.py、report_lib.py、environment.yml、README.txt：可复跑的完整流程、作图代码与运行说明

方法学段落（可直接改写进论文）

双端测序读段（SRA SRR2584866）经 fastp 1.3.7 质控后，用 BWA-MEM 0.7.19-r1273 比对到大肠杆菌 B REL606 参考基因组。

本示例使用公开数据，所有数字与图均由随附代码重新计算得到；真实项目交付结构与此一致，内容按书面方案定制。