

Bacterial genome resequencing: variant report

SRR2584866 · E. coli REL606 long-term evolution clone REL11365 (public-data example)

Project DEMO-DNA-01

Data source NCBI SRA
SRR2584866

Analysis date 2026-10-01

Issued by HST GENOMICS
bioinformatics

Project overview

This report is a sample deliverable built on public data to show how HST GENOMICS delivers a microbial genome resequencing and variant calling project. Data: NCBI SRA run SRR2584866, clone REL11365 from the E. coli B REL606 long-term evolution experiment (study "E. coli genome evolution over 50,000 generations"), Illumina HiSeq 2500, 2 × 150 bp. Variants are called against the ancestral REL606 reference genome (NCBI GCF_000017985.1).

4.95 M

reads after QC

99.99%

mapped

143.9x

mean depth

789

variants after
filtering

Methods

- Read QC: fastp 1.3.7 with default settings (adapter trimming, quality filtering).
- Alignment: BWA-MEM 0.7.19-r1273 to the REL606 reference; sorting and indexing with samtools 1.24.
- Variant calling: bcftools 1.24 mpileup + call in haploid mode (--ploidy 1).
- Filtering: QUAL $\geq$ 30 and site depth DP $\geq$ 10.
- Annotation: snpEff 5.4c with a database built from the NCBI REL606 annotation (GFF3).

Quality control

Metric	Value
Raw reads	5,536,796
Reads after QC	4,951,018
Q30 after QC	91.6%
Mapping rate	99.99%
Mean depth	143.9x
Genome at ≥10x	93.1%

Table 1. Sequencing and alignment QC.

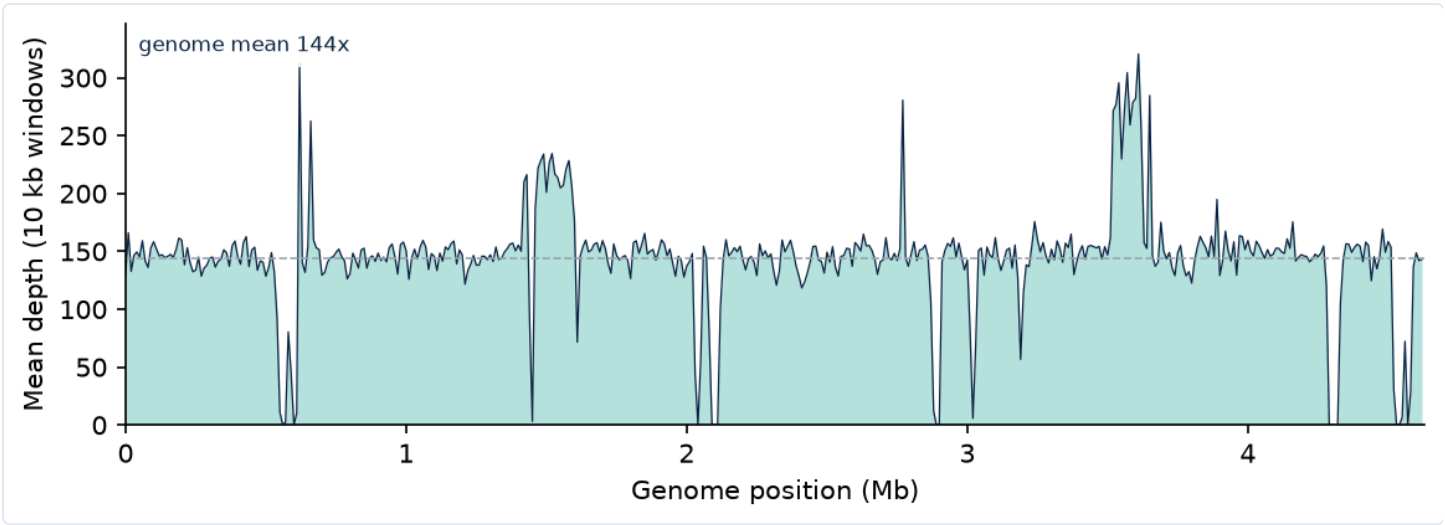


Figure 1. Read depth along the genome (10 kb window means). About 6.9% of reference positions have depth below 10x; contiguous stretches near zero may reflect deletions in this clone relative to the ancestor, and stretches at roughly twice the mean may reflect amplifications; both need confirmation from read-level evidence such as split reads at breakpoints.

Results

After filtering there are 789 variants: 683 SNPs, 65 insertions and 41 deletions. By snpEff predicted impact: HIGH 76, MODERATE 362, LOW 208, MODIFIER 143.

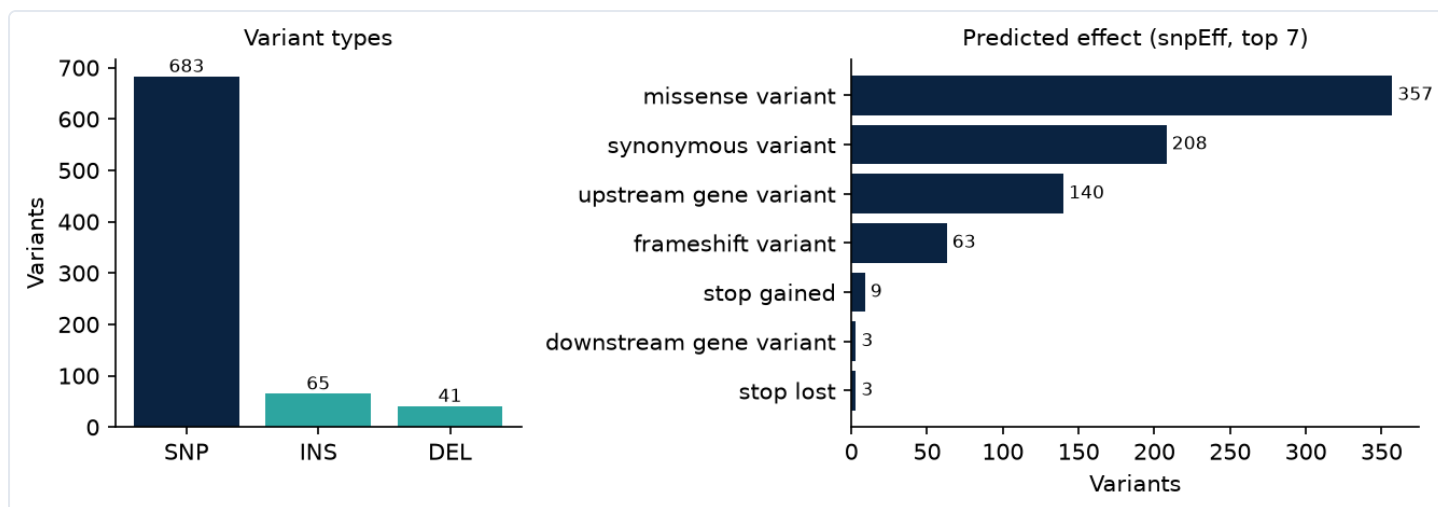


Figure 2. Variant types and predicted functional effects.

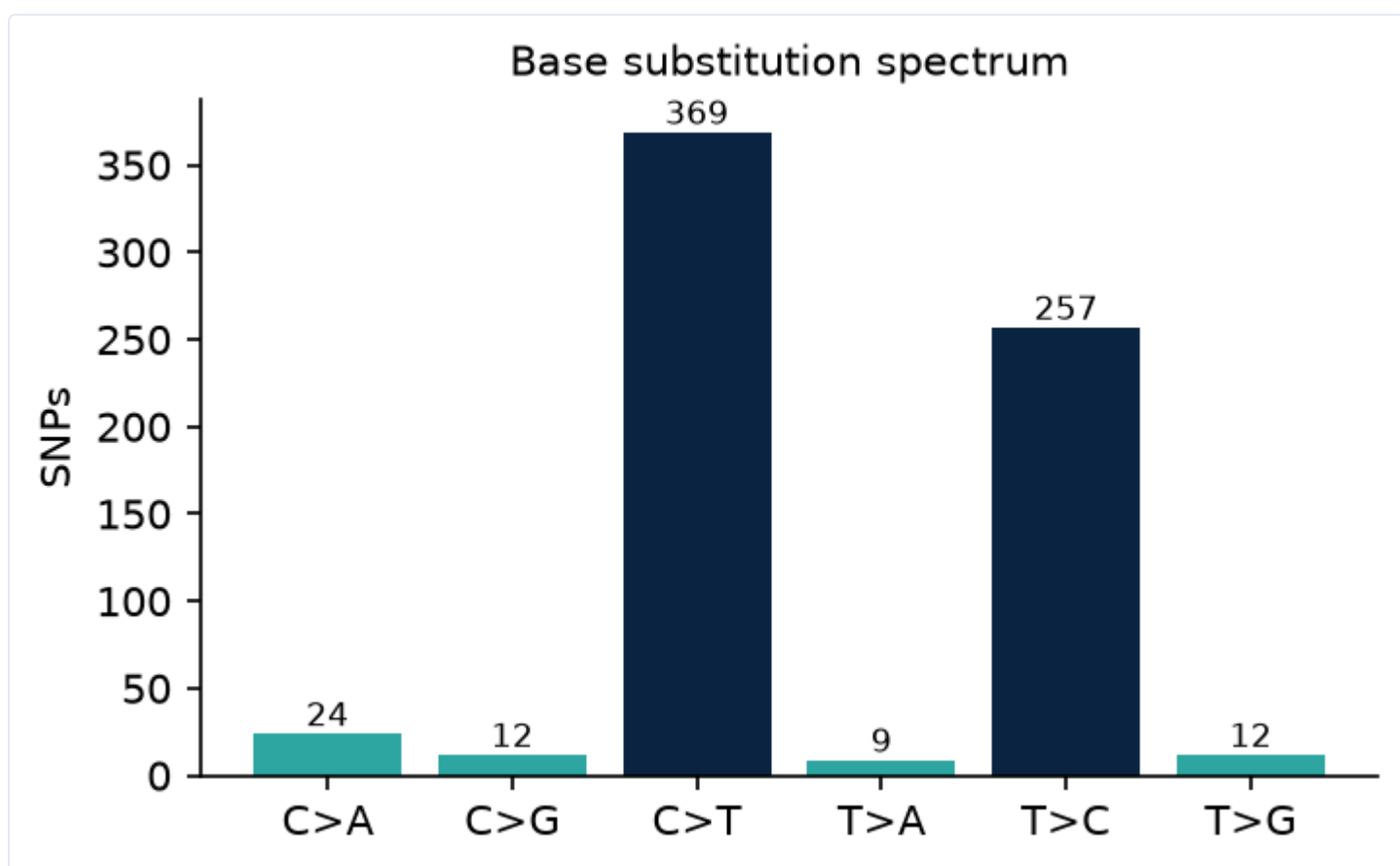


Figure 3. Base substitution spectrum (collapsed to six pyrimidine-referenced classes). Transitions 626, transversions 57.

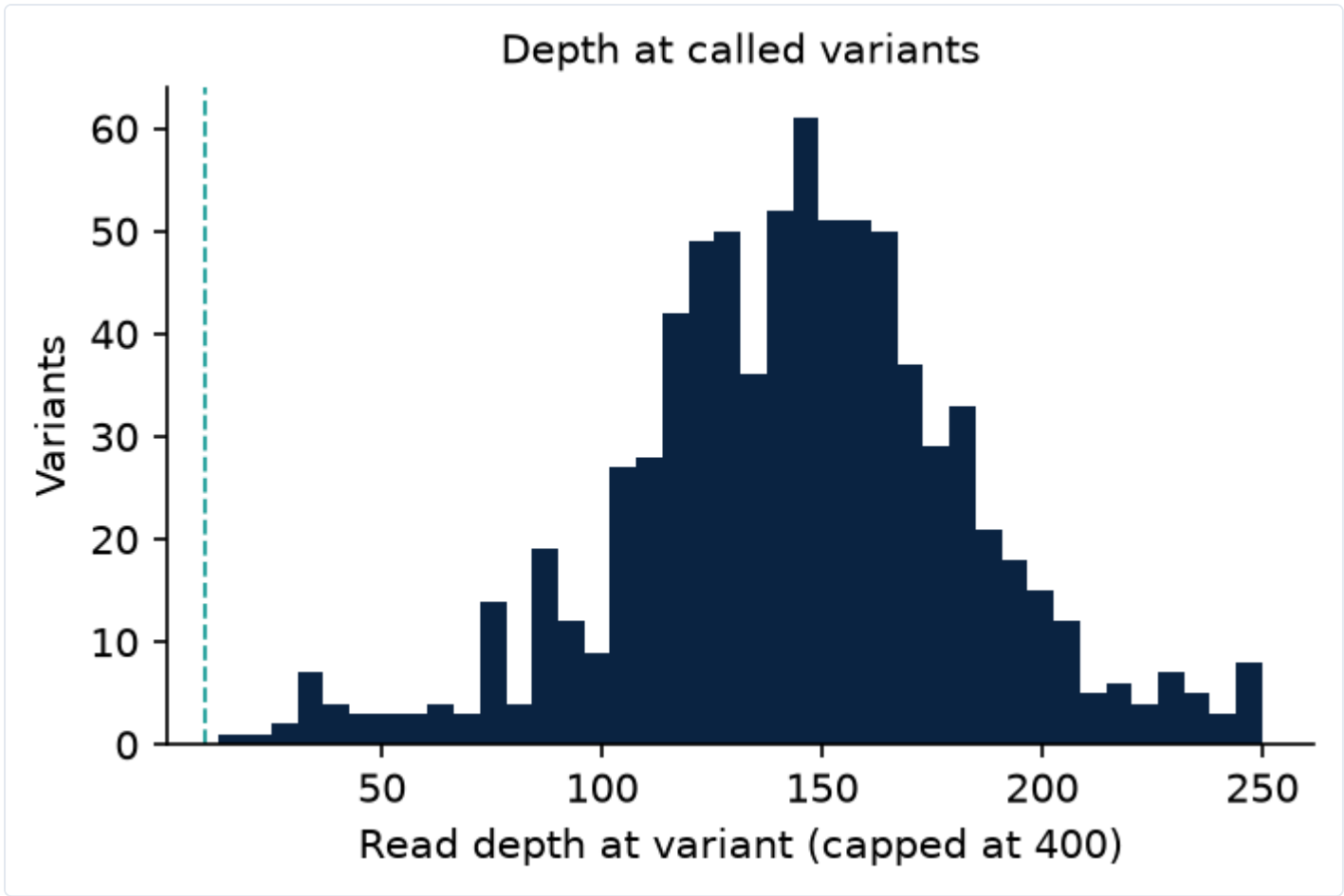


Figure 4. Read depth at called variants; dashed line marks the 10x filter.

Position	Change	Gene	Effect	HGVS	Depth
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

Table 2. Variants with HIGH predicted impact (76 in total, first 30 shown; full list in tables/variants_table.csv).

Interpretation note: predicted effects describe where a variant falls and how it may change coding; they are not functional evidence. In real projects candidates are prioritised for the research question (resistance, adaptation, phenotype) with suggestions for validation.

Deliverables

- tables/variants_table.csv: all filtered variants with annotation (position, type, gene, effect, HGVS, depth)
- tables/variants_filtered.vcf.gz, tables/variants_annotated.vcf: standard VCF files
- tables/coverage_10kb_windows.tsv, depth_summary.tsv, flagstat.txt, fastp.json, bcftools_stats.txt: QC statistics
- figures/*.png: publication figures (PDF/SVG on request)
- run_variant.sh, report_variant.py, report_lib.py, environment.yml, README.txt: complete re-runnable pipeline, plotting code and instructions

Methods paragraph (for the manuscript)

Paired-end reads (SRA SRR2584866) were quality-filtered with fastp 1.3.7 and aligned to the E. col

This example uses public data; every number and figure is recomputed by the attached code. Real projects follow the same structure with content defined by the written plan.